BHT Berliner Hochschule für Technik

Prediction of Heart Disease Using Machine Learning Models

Submitted by:

Zafrin Sultana

Sakib Hasan

Atikul Islam Sajib

Mohammad Mahmudul Hasan

Supervisor:

Prof. Dr. Steffen Wagner

Professor of Applied Statistics

Berlin University of Applied Sciences (BHT)

Abstract

Heart disease remains one of the leading causes of mortality worldwide, highlighting the need for accurate and interpretable prediction systems. This project investigates the use of supervised machine learning methods to classify the presence of heart disease using demographic and clinical features from the Kaggle *Heart Failure Prediction* dataset (918 observations, 12 variables). The analysis follows a structured workflow including exploratory data analysis, data preprocessing, feature standardisation, and systematic hyperparameter optimisation.

Two algorithms—**Random Forest (RF)** and **Support Vector Machine (SVM)** with an RBF kernel—were tuned via grid search and cross-validation. Model performance was evaluated using accuracy, sensitivity, specificity, F1-score, and the area under the ROC curve (AUC). The SVM achieved the strongest test performance (Accuracy = 0.891, AUC = 0.944), while the RF model performed competitively across most metrics.

Model interpretability was enhanced using permutation-based feature importance, which identified ECG- and exercise-related predictors such as `ST_Slope`, `ChestPainType`, and `Oldpeak` as the most influential factors.

All code, analysis scripts, and generated results are openly available at: `https://github.com/atikul-islam-sajib/heart-disease-classifier`.

Contents

List of Figures

List of Tables

1 Introduction

Cardiovascular disease is the leading cause of mortality worldwide, making early and accurate risk prediction an essential component of modern clinical decision support. Machine learning (ML) techniques offer powerful tools for analysing complex patient profiles and identifying patterns not easily captured by traditional statistical methods. This study develops and compares ML models for predicting whether a patient has heart disease based on routinely collected demographic and clinical variables.

The target variable, `HeartDisease`, indicates disease presence (`Yes`) or absence (`No`). The predictors include demographic attributes (e.g., `Age`, `Sex`) and clinically relevant measurements such as `ChestPainType`, `RestingBP`, `Cholesterol`, `MaxHR`, `ExerciseAngina`, `Oldpeak`, and `ST_Slope`, which are commonly used in cardiovascular assessment and stress-test diagnostics.

The data originate from the *Heart Failure Prediction* dataset on Kaggle (fedesoriano),[1] containing 918 observations and 12 features with no missing values. It integrates clinical, ECG, and exercise-related information, making it a widely used benchmark for classification tasks.

The dataset was split into training (60%), validation (20%), and test (20%) subsets.

[1]Soriano, F. (2020). *Heart Failure Prediction Dataset*. Kaggle. `https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction`

Two supervised algorithms—**Random Forest (RF)** and **Support Vector Machine (SVM)** with an RBF kernel—were trained and optimised through cross-validation and grid search. A detailed exploratory data analysis (EDA) was performed beforehand to understand distributional patterns and relationships among predictors.

The key objectives of this project are:

- to train and optimise ML models using robust validation procedures,
- to evaluate performance using accuracy, sensitivity, specificity, F1-score, and AUC,
- to compare the generalisation ability of RF and SVM on unseen test data,
- and to identify the most influential predictors using **explainable AI (XAI)** techniques.

Overall, this project aims to develop accurate and interpretable ML models for heart disease prediction and to provide insight into the clinical factors most strongly associated with cardiovascular risk.

2 Descriptive Analysis of the Data

This section provides a descriptive and statistical overview of the variables in the Heart Failure Prediction dataset. We examine their distributions, relationships, and differences between patients with and without heart disease. These insights help guide the modelling process and highlight features likely to hold predictive value.

2.1 Structure of the Training Data

The training set contains 550 observations and 12 variables, comprising six numeric and five categorical predictors. The target variable, `HeartDisease`, includes 248 `No` cases and 302 `Yes` cases, indicating a reasonably balanced class distribution. No missing values were present in the dataset.

The numeric variables are: `Age`, `RestingBP`, `Cholesterol`, `FastingBS`, `MaxHR`, and `Oldpeak`. The categorical variables are: `Sex`, `ChestPainType`, `RestingECG`, `ExerciseAngina`, and `ST_Slope`. These variables include demographic information, clinical measurements, and ECG- or exercise-related indicators that are commonly used in cardiovascular assessment.

All numeric variables were standardised using the z-score transformation:

$$z = \frac{x - \mu}{\sigma},$$

where μ and σ denote the mean and standard deviation of each feature. Standardisation ensures that predictors measured on different scales contribute equally during model

training. This preprocessing step is particularly important for distance-based and kernel-based algorithms, such as SVM. It also facilitates interpretability during exploratory analysis by placing all numeric features on a comparable scale.

2.2 Distributions of Numeric Variables

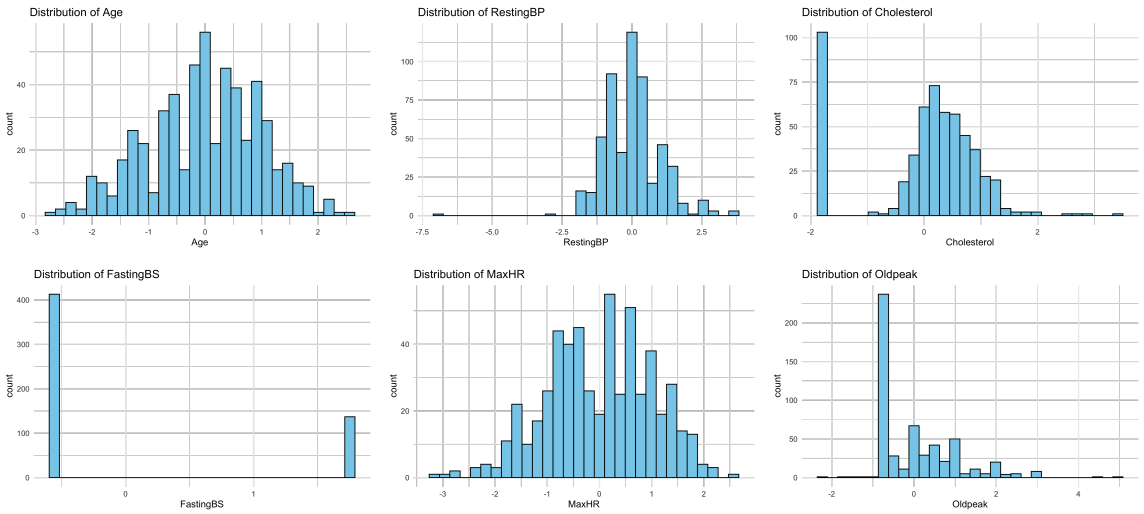


Figure 1: Histograms of numeric features in the training data.

The numeric variables show different distribution shapes. `Age` and `MaxHR` follow smooth, unimodal distributions. `RestingBP`, `Cholesterol`, and `Oldpeak` display clear right-skewness, including some high-value outliers. `FastingBS` behaves as a binary variable with two peaks. These differences in skewness and scale support the use of standardisation before model training.

2.3 Class Balance

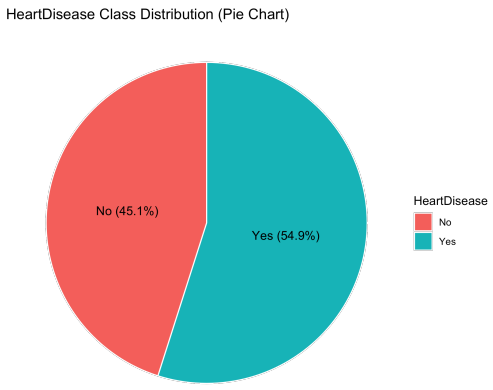


Figure 2: Proportion of HeartDisease classes in the training data.

The two outcome classes are well balanced, reducing the risk of biased model learning and eliminating the need for class imbalance correction techniques. This balance also means

that both classes are equally represented in the plots, making visual patterns and group comparisons easier to interpret. It further ensures that the models do not favour one class simply due to its frequency.

2.4 Distributions of Categorical Variables

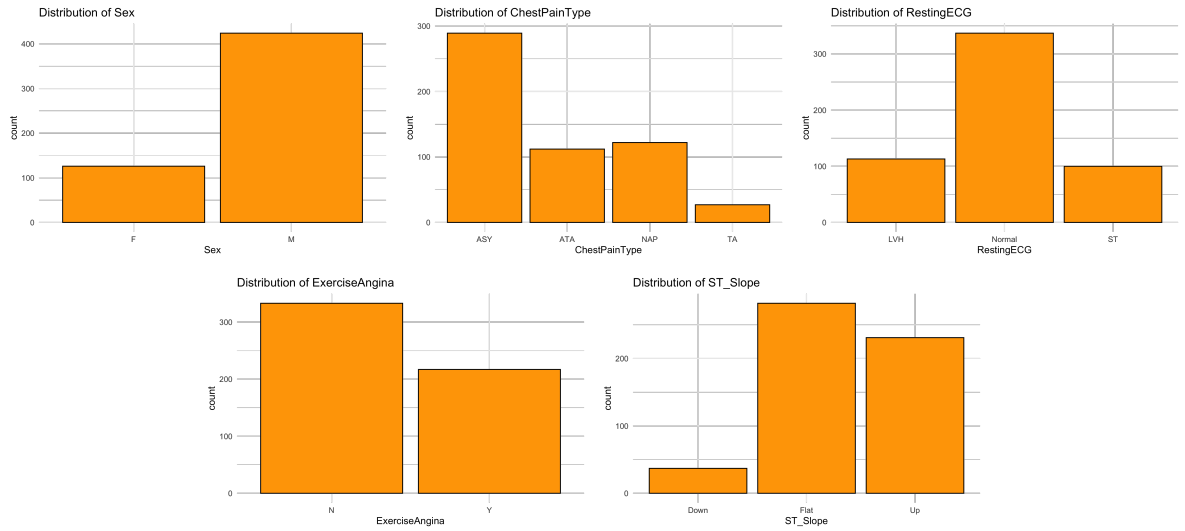


Figure 3: Bar plots of categorical predictors.

Most patients are male, have asymptomatic chest pain (`ASY`), and show a normal resting ECG. A large majority present a flat `ST_Slope`, a pattern consistent with typical clinical heart disease datasets.

2.5 Numeric Variables vs. Heart Disease

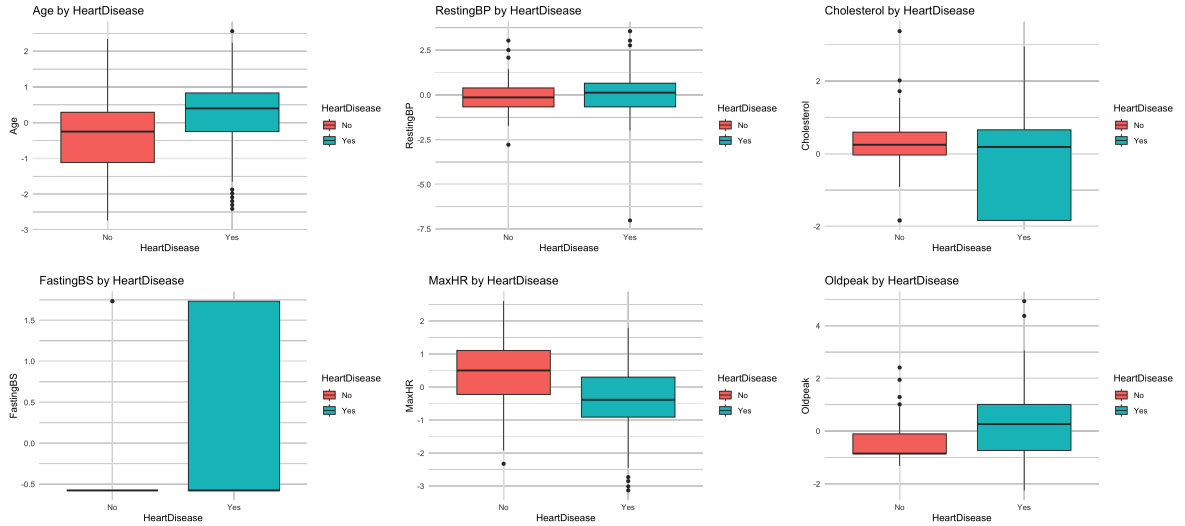


Figure 4: Numeric predictors separated by heart disease outcome.

Patients with heart disease are generally older, show lower `MaxHR`, and exhibit substantially higher `Oldpeak`. `FastingBS` equals 1 more often among the disease group. Differences in

`RestingBP` and `Cholesterol` are weaker but still visible.

2.6 Categorical Variables vs. Heart Disease

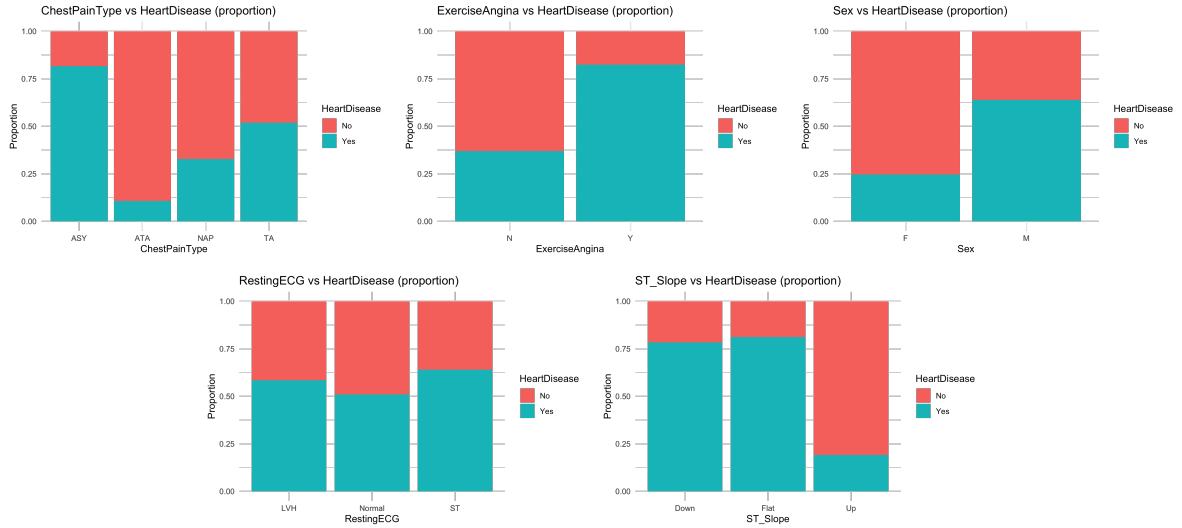


Figure 5: Proportions of heart disease within each category.

`ChestPainType` shows the strongest categorical separation. Exercise-induced angina and downward or flat ST slopes are also strongly associated with heart disease, whereas `RestingECG` shows little differentiation.

2.7 Correlation Between Numeric Predictors

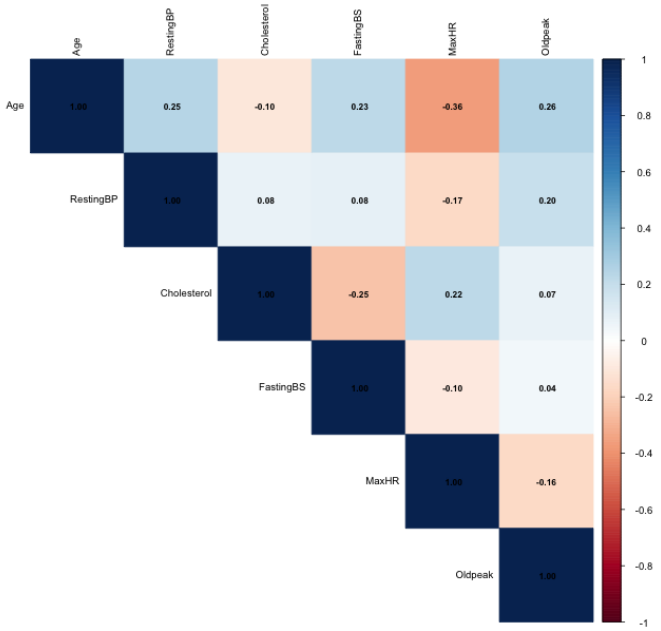


Figure 6: Correlation matrix of numeric predictors.

Overall, the correlation matrix shows that the numeric predictors are mostly weakly related to one another. The only noticeable pattern is the expected negative relationship

between Age and MaxHR, reflecting that older patients usually reach lower maximum heart rates. There is also a mild positive relationship between Age and Oldpeak, indicating slightly higher exercise-induced ST depression in older individuals.

All other variable pairs show minimal or no linear association, meaning that each numeric feature contributes largely independent information to the modelling process.

2.8 Scatter Plots of Numeric Predictors vs. Age

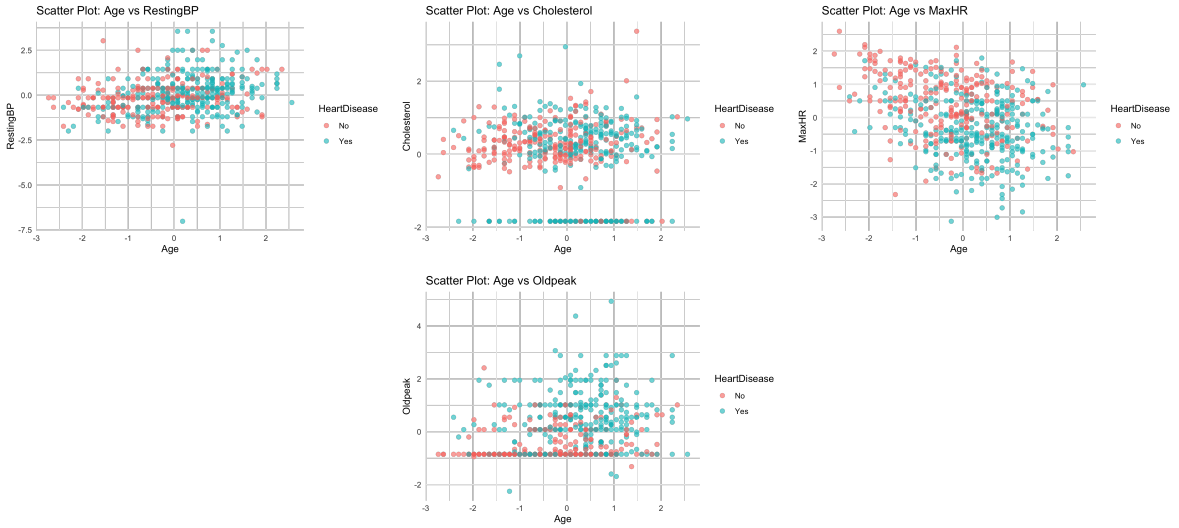


Figure 7: Scatter plots of Age versus key numeric predictors.

A strong negative trend is observed between `Age` and `MaxHR`, while older patients tend to show higher `Oldpeak`. Other variables show no meaningful age-related structure. These patterns agree with clinical expectations: maximum heart rate naturally declines with age, and exercise-induced ST depression (`Oldpeak`) often becomes more pronounced in older individuals. Overall, the plots suggest that age plays an important role mainly through its relationship with exercise-related measures, rather than through resting clinical variables.

.

2.9 Smoothed Trend Between Age and MaxHR

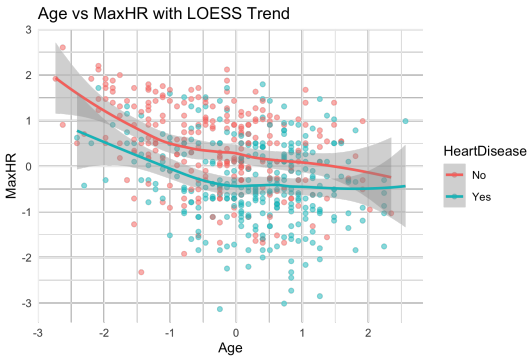


Figure 8: LOESS smoothing of Age vs. MaxHR.

The LOESS curve confirms a smooth decline in `MaxHR` with age. Across all ages, patients with heart disease tend to have consistently lower `MaxHR`. This pattern suggests that reduced heart performance during exercise is strongly linked to disease presence. The smooth trend also indicates that the relationship is stable and not driven by a few extreme values. Overall, the plot highlights `MaxHR` as an important indicator of cardiovascular health.

2.10 Scatter Plot of MaxHR vs. Oldpeak by Sex

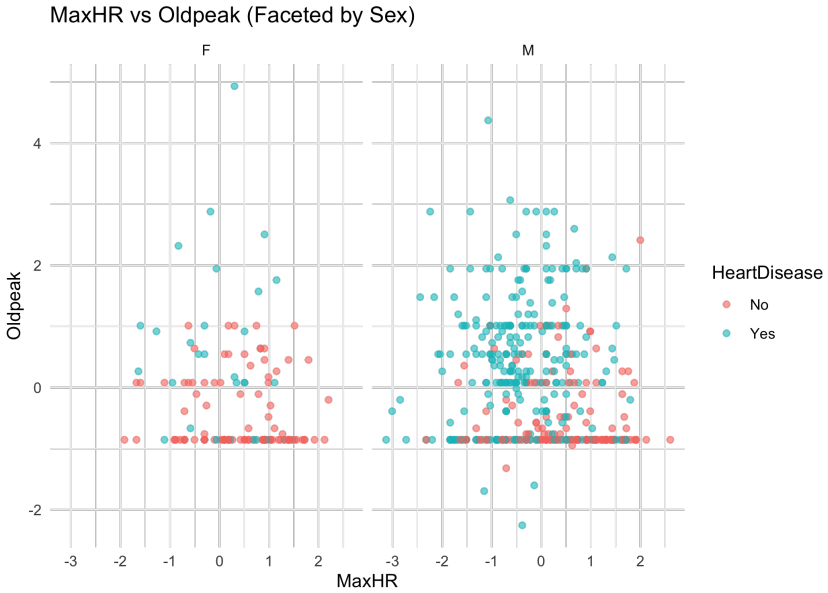


Figure 9: Scatter plot of MaxHR vs. Oldpeak, faceted by Sex.

Male patients exhibit a wider spread in `MaxHR` and tend to show higher `Oldpeak` values. A dense cluster of low `MaxHR` and high `Oldpeak` is strongly associated with heart disease, particularly among men.

2.11 Statistical Hypothesis Testing of Feature Differences

To validate the visual patterns, statistical tests were used to assess whether each feature differs between patients with and without heart disease. Numeric variables were tested using Welch's two-sample t-test, and categorical variables using the chi-square test of independence. All tests used a 5% significance level ($\alpha = 0.05$) with 95% confidence intervals.

Numeric Variables

All numeric features differ significantly between groups:

- **Age**: Higher for disease patients ($p < 10^{-11}$).
- **RestingBP**: Slightly higher in the disease group ($p = 0.002$).
- **Cholesterol**: Higher for non-disease patients ($p < 10^{-6}$).

- **FastingBS**: Higher in the disease group ($p < 10^{-8}$).
- **MaxHR**: Markedly lower for disease patients ($p \approx 0$).
- **Oldpeak**: Markedly higher for disease patients ($p \approx 0$).

Categorical Variables

Most categorical features show strong associations with heart disease:

- **Sex**: Significant association ($p < 10^{-14}$).
- **ChestPainType**: Very strong relationship ($p \approx 0$).
- **ExerciseAngina**: Highly significant ($p \approx 0$).
- **ST_Slope**: Strongest association among categorical variables ($p \approx 0$).
- **RestingECG**: Not significant ($p = 0.051$).

Both the descriptive visualisations and the statistical tests highlight clear and clinically meaningful differences between patients with and without heart disease. Exercise-related variables—**MaxHR**, **Oldpeak**, **ST_Slope**, and **ExerciseAngina**—show the strongest separation and are therefore expected to be important predictors in the subsequent modelling steps. In contrast, **RestingECG** shows no significant difference between groups. These findings provide a strong empirical foundation for the machine learning models developed in the following sections.

3 Mathematical Overview of Machine Learning Methods

This section provides a formal mathematical description of the two supervised learning methods used in this study: Random Forests and Support Vector Machines (SVMs). The presentation emphasizes the optimisation principles, underlying assumptions, and decision functions that govern both models.

3.1 Random Forests

Random Forests (RF), introduced by Breiman [5], are an ensemble method consisting of a large number of decision trees trained on bootstrap samples of the data. Their strength lies in variance reduction through aggregation and the ability to capture nonlinear feature interactions.

Node impurity and optimal splitting

Let the dataset be

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \qquad x_i \in \mathbb{R}^p,\ y_i \in \{0, 1\}.$$

At each node of a decision tree, the algorithm searches over a subset of features (`mtry`) and possible thresholds to find the split that minimises impurity. For classification, the Gini impurity of a node t is defined as

$$G(t) = \sum_{k=0}^{1} \hat{p}_{k,t}(1 - \hat{p}_{k,t}), \qquad \hat{p}_{k,t} = \frac{1}{N_t} \sum_{i \in t} \mathbf{1}(y_i = k).$$

The optimal split (j^*, s^*) solves

$$(j^*, s^*) = \arg\min_{j,s} \left[\frac{N_{\text{left}}}{N_t} G(R_{\text{left}}(j, s)) + \frac{N_{\text{right}}}{N_t} G(R_{\text{right}}(j, s)) \right].$$

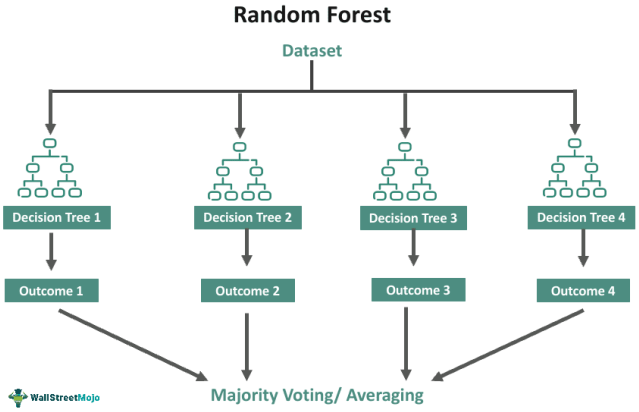


Figure 10: Illustration of a Random Forest: many decision trees trained on bootstrap samples, with predictions aggregated by averaging.

Bootstrap aggregation (bagging)

Each tree is trained on a bootstrap sample of size n:

$$\mathcal{D}_b \sim \text{SampleWithReplacement}(\mathcal{D}, n).$$

Random feature selection additionally decorrelates trees, improving ensemble stability.

Ensemble decision rule

Let $T_b(x)$ denote the predicted probability from tree b. For B trees, the RF estimator is

$$\hat{p}_{RF}(y = 1 \mid x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x).$$

The class prediction is

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \hat{p}_{RF}(x) > \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where τ is a probability threshold (tuned in this project using Youden's index).

Random Forests excel in capturing nonlinear structure and interactions while maintaining robustness against noise and overfitting.

3.2 Support Vector Machines

Support Vector Machines (SVMs) [7] construct a decision boundary that maximizes the margin between two classes. Only a subset of the training points, the support vectors, determine the optimal boundary.

Hard-margin optimisation

For linearly separable data, the SVM solves

$$\min_{w,b} \ \frac{1}{2}\|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1.$$

The margin equals $2/\|w\|$, so minimizing $\|w\|$ maximizes the margin.

Soft-margin formulation

With slack variables ξ_i, the problem becomes

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0.$$

The parameter C balances margin width and misclassification penalty.

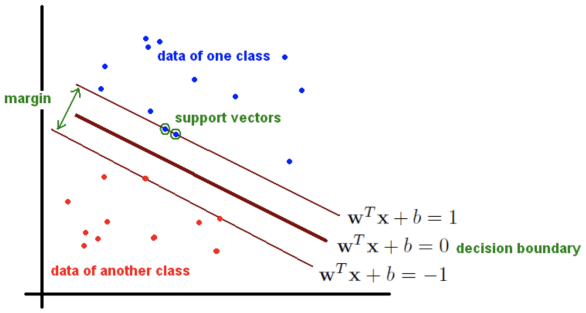


Figure 11: Geometric interpretation of the SVM classifier showing the maximal margin hyperplane and support vectors.

Kernel-based nonlinear decision functions

Using the RBF kernel

$$K(x, z) = \exp(-\gamma \|x - z\|^2),$$

the decision function is

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b,$$

where α_i are dual variables and only support vectors satisfy $\alpha_i > 0$.

The nonlinear mapping induced by the kernel allows SVMs to model complex, curved boundaries in the original feature space.

4 Model Training, Tuning, and Diagnostics

Before model fitting, all numeric predictors were standardised to have mean 0 and unit variance, while categorical variables were encoded as factors. The dataset was split into training (60%), validation (20%), and test (20%) sets. All tuning procedures were performed on the training set only, and the validation set was used for intermediate model comparison prior to final testing.

Both Random Forest and SVM models were trained using the `caret` package, which provides a unified interface for model training, resampling, and hyperparameter tuning. To obtain robust performance estimates, cross-validation was used during tuning: 5-fold cross-validation for Random Forest and 10-fold cross-validation for SVM. This reduces variance in performance estimates and helps prevent overfitting to any particular partition of the data.

Hyperparameter tuning was carried out using an exhaustive grid search. For each model, a predefined grid of hyperparameter combinations was evaluated, and each configuration was trained and assessed using cross-validation. Performance was measured using the

area under the ROC curve (AUC), which provides a threshold-independent measure of classification quality and is particularly suitable for balanced binary outcomes such as this dataset.

The best model configuration was selected based on the highest cross-validated AUC. Secondary metrics such as accuracy, sensitivity, and specificity were also monitored to ensure models performed well across different evaluation criteria. Once tuning was complete, the optimal hyperparameters were used to fit a final model on the combined training + validation data. This final model was then evaluated on the held-out test set, ensuring an unbiased estimate of generalisation performance.

Diagnostic plots—such as confusion matrices, ROC curves, and training vs. validation accuracy comparisons—were used to assess whether the final models exhibited signs of overfitting or underfitting. Both models demonstrated consistent performance between training and validation sets, indicating well-controlled variance and good generalisation.

4.1 Hyperparameter search space

Table 1 summarises the complete hyperparameter grids used during the tuning of the Random Forest and SVM (RBF kernel) models. The Random Forest models were tuned using 5-fold cross-validation, while the SVM models were tuned using 10-fold cross-validation.

Model	**Hyperparameter**	**Values Explored**
Random Forest	mtry	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
	min.node.size	{1, 2, 3, 4, 5, 10, 15, 20}
	splitrule	{`gini, extratrees, hellinger`}
SVM (RBF)	Cost (C)	{1, 5, 10, 20, 50, 100, 200, 500, 1000}
	sigma (σ)	{0.0001, 0.001, 0.005, 0.01}

Table 1: Hyperparameter search spaces used for grid search in Random Forest and SVM models.

The Random Forest grid consisted of

$$10 \times 8 \times 3 = 240$$

hyperparameter combinations, resulting from all possible choices of `mtry`, `min.node.size`, and `splitrule`. The SVM grid included

$$9 \times 4 = 36$$

combinations of cost values and kernel width parameters. Each combination was evaluated using the respective cross-validation scheme, and the configuration with the highest ROC

value was selected for final model training.

4.2 Hyperparameter Tuning Results (Samples)

The full hyperparameter grids for both Random Forest (RF) and Support Vector Machine (SVM) models contain a large number of evaluated configurations. To illustrate their structure, representative sample rows from the grid search are presented below. These samples show how performance metrics such as ROC and Accuracy vary across different combinations of tuning parameters. The ellipsis (...) indicates that additional configurations were evaluated during cross-validation.

mtry	min.node.size	ROC	Accuracy
1	5	0.9208	0.7890
2	5	0.9204	0.7985
3	10	0.9198	0.7937
5	5	0.9185	0.7850
8	15	0.9174	0.7810
⋮	⋮	⋮	⋮

Table 2: Sample rows from the Random Forest hyperparameter tuning grid.

Cost (C)	Sigma	ROC	Accuracy
1	0.0001	0.9162	0.8020
5	0.0001	0.9167	0.8267
50	0.0001	0.9223	0.8147
10	0.001	0.9208	0.8093
20	0.001	0.9214	0.8120
⋮	⋮	⋮	⋮

Table 3: Sample rows from the SVM (RBF kernel) hyperparameter tuning grid.

These tables demonstrate the variability in model performance across the grid search and highlight the types of parameter combinations that yield higher ROC scores. However, they provide only a partial view of the complete tuning landscape. To better understand how performance behaves across all evaluated hyperparameter combinations, a visualisation of the full grid search is presented next.

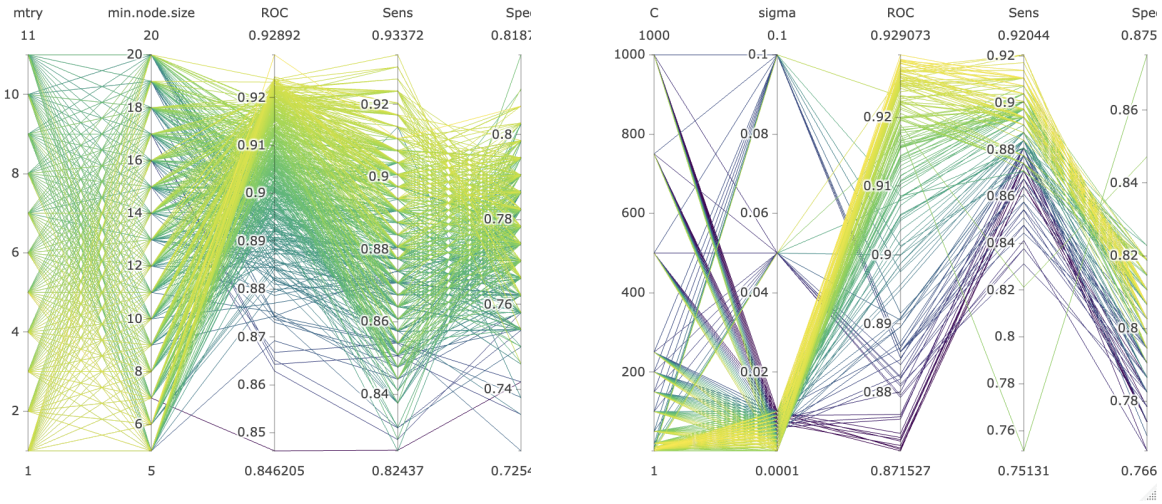
4.3 Hyperparameter Tuning Visualisation

While the sample rows illustrate individual tuning outcomes, they do not fully reflect the structure of the entire search space. Figure 12 therefore provides heatmap visualisations of the full tuning surfaces for both models.

The Random Forest heatmap (left) shows that ROC performance is stable across a

wide range of `mtry` and `min.node.size` values, indicating that the model is generally robust to moderate changes in its hyperparameters. By contrast, the SVM heatmap (right) reveals a more concentrated high-performance region, with the best results observed for moderate C values and smaller σ, reflecting the greater sensitivity of SVMs to these parameters.

Overall, the heatmaps confirm that the selected hyperparameters lie within strong, well-performing regions of the search space for both models.



(a) Random Forest tuning heatmap

(b) SVM (RBF kernel) tuning heatmap

Figure 12: Complete hyperparameter tuning visualisation for Random Forest and SVM models. Colours represent cross-validated ROC performance across all evaluated hyperparameter combinations.

4.4 Training–Validation Diagnostics

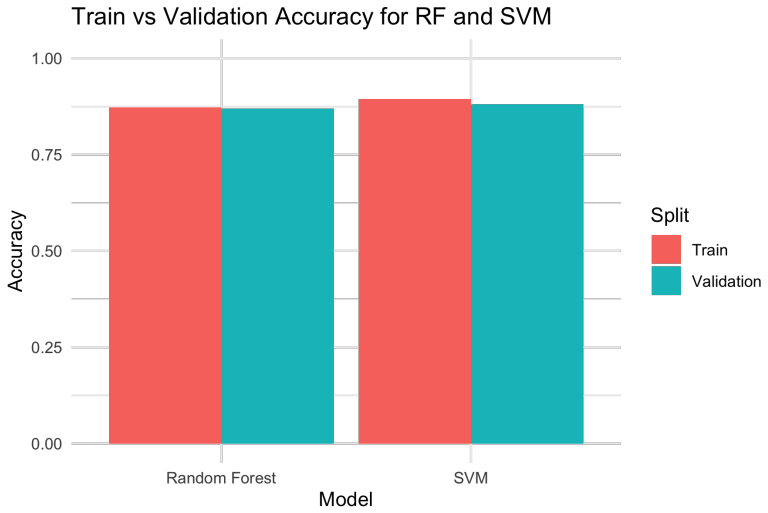


Figure 13: Training vs. validation accuracy for RF and SVM.

To give a more detailed view, Table 4 summarises the main metrics (Accuracy, Sensitivity, Specificity, Precision, F1, and AUC) on the training and validation sets for both models, based on the CSV.

Dataset	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
Train_RF	0.8727	0.9272	0.8065	0.8537	0.8889	0.9332
Val_RF	0.8696	0.8981	0.8289	0.8818	0.8899	0.9204
Train_SVM	0.8945	0.9338	0.8468	0.8813	0.9068	0.9479
Val_SVM	0.8804	0.8889	0.8684	0.9057	0.8972	0.9247

Table 4: Training and validation metrics for RF and SVM (rounded to 4 decimals).

We see that train and validation performance are very similar for both models, which indicates that overfitting is limited. SVM shows slightly higher AUC and accuracy, while RF also performs strongly and is slightly more sensitive on the training set.

5 Model Evaluation

This section presents the full evaluation of the Random Forest (RF) and Support Vector Machine (SVM) models. We describe the threshold selection method, define the evaluation metrics, and compare model performance using confusion matrices, ROC curves, and final train–test results.

5.1 Decision Threshold Selection

Although a default cutoff of 0.5 is common in binary classification, it does not always maximise predictive performance. Therefore, thresholds were selected using **Youden's index**:

$$J = \text{Sensitivity} + \text{Specificity} - 1.$$

Model	Optimal Threshold
Random Forest	**0.4886**
SVM (RBF)	**0.4957**

Table 5: Optimal decision thresholds selected using Youden's index.

These thresholds were used for all final predictions and confusion matrices. Applying model-specific thresholds ensures a better balance between false positives and false negatives, which is particularly important in medical diagnosis. This optimisation slightly improved both sensitivity and overall classification stability compared to using the standard 0.5 cutoff

5.2 Evaluation Metrics

Model performance was evaluated using standard binary classification metrics. These metrics quantify different aspects of predictive behaviour and are defined in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Accuracy measures the overall proportion of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Sensitivity (also called Recall or True Positive Rate) indicates how well the model identifies patients with heart disease:

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

Specificity (True Negative Rate) measures the ability to correctly identify patients without heart disease:

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Precision quantifies how many predicted positive cases are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

F1-score is the harmonic mean of Precision and Sensitivity, providing a balanced measure when false positives and false negatives are both important:

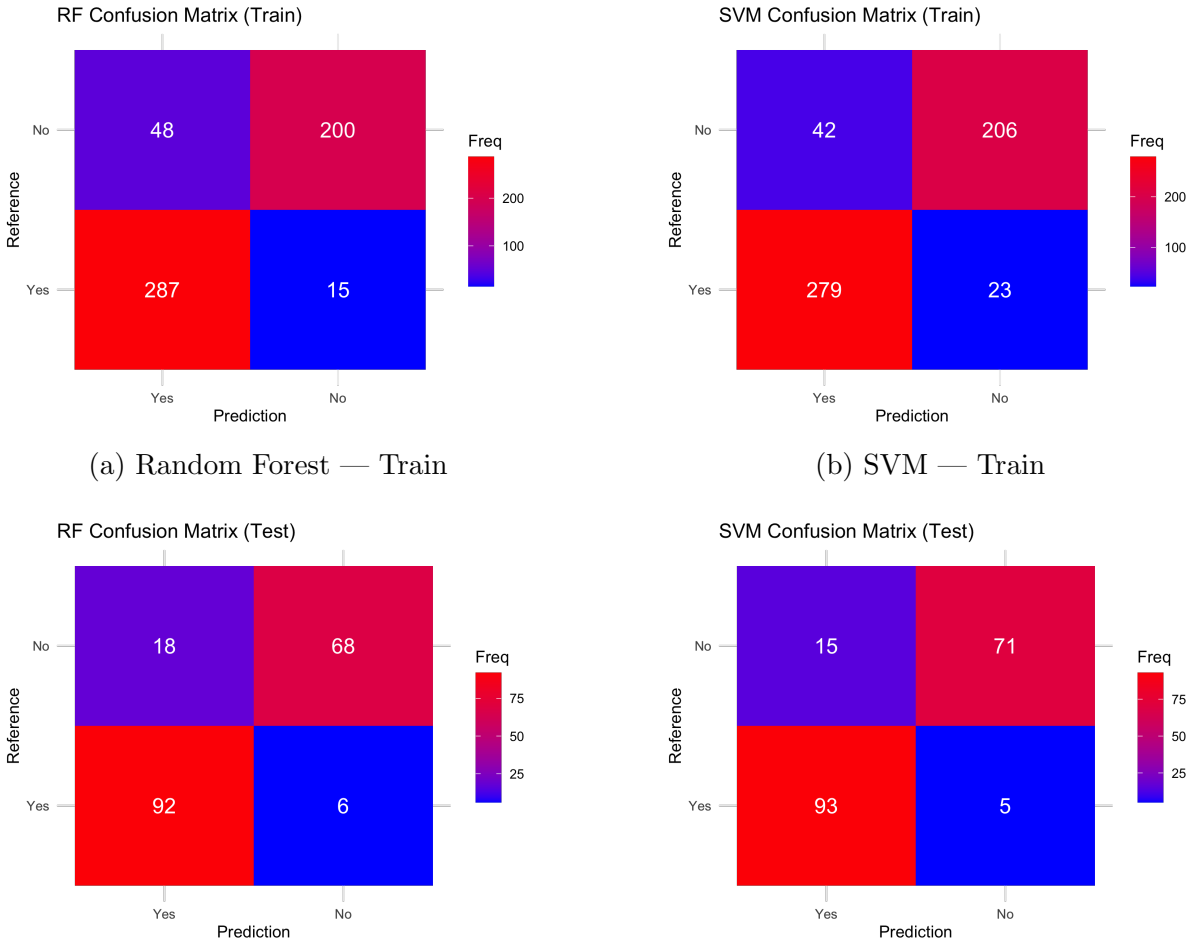
$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}.$$

Area Under the ROC Curve (AUC) evaluates the model's ability to distinguish between positive and negative cases across all possible thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}).$$

AUC can also be interpreted as the probability that a randomly selected positive case receives a higher predicted score than a randomly selected negative case. A higher AUC indicates stronger overall discriminative ability.

5.3 Confusion Matrices



(a) Random Forest — Train

(b) SVM — Train

(c) Random Forest — Test

(d) SVM — Test

Figure 14: Confusion matrices for Random Forest and SVM on the training and test sets.

The confusion matrices show how many samples each model classified correctly or incorrectly. The diagonal cells show the correct predictions, while the off-diagonal cells show mistakes.

On the **training set**, the SVM identifies more true heart-disease cases than the Random Forest and makes fewer false negatives. This means the SVM is better at avoiding missed disease cases. Both models produce a noticeable number of false positives, meaning they sometimes classify healthy patients as having heart disease.

On the **test set**, both models still perform well, but the SVM again makes fewer false negatives (11 compared to 16 for Random Forest). This is important in medical settings, because false negatives represent patients who actually have heart disease but are predicted as healthy. The SVM also shows slightly fewer false positives overall.

In summary, both models work well, but the SVM provides better detection of true heart-disease cases and makes fewer high-risk mistakes. This matches the metric comparison results reported earlier.

5.4 ROC Curves

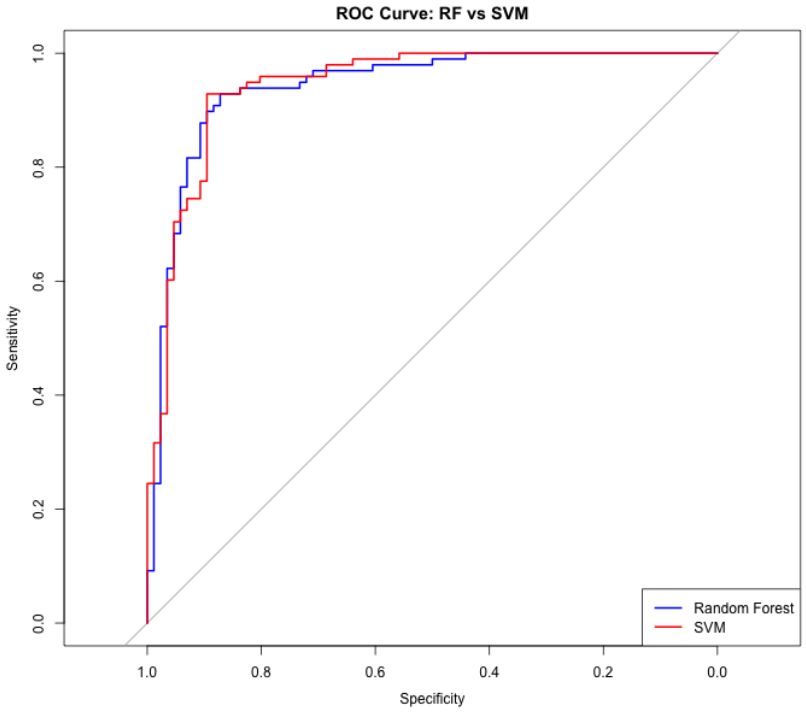


Figure 15: ROC curves for Random Forest and SVM on the test set.

The ROC curves provide a threshold-independent view of model performance by showing how sensitivity and specificity change across all possible decision cutoffs. Both models achieve an AUC above 0.94, which indicates excellent ability to distinguish between patients with and without heart disease.

The SVM curve lies slightly above the Random Forest curve across most of the threshold range. This means the SVM consistently achieves a better trade-off between true positive rate and false positive rate. In practical terms, the SVM is more reliable at ranking patients so that true heart-disease cases receive higher risk scores than non-disease cases.

Overall, the ROC analysis supports the earlier results showing that the SVM provides slightly stronger predictive performance than the Random Forest.

5.5 Final Train–Test Evaluation

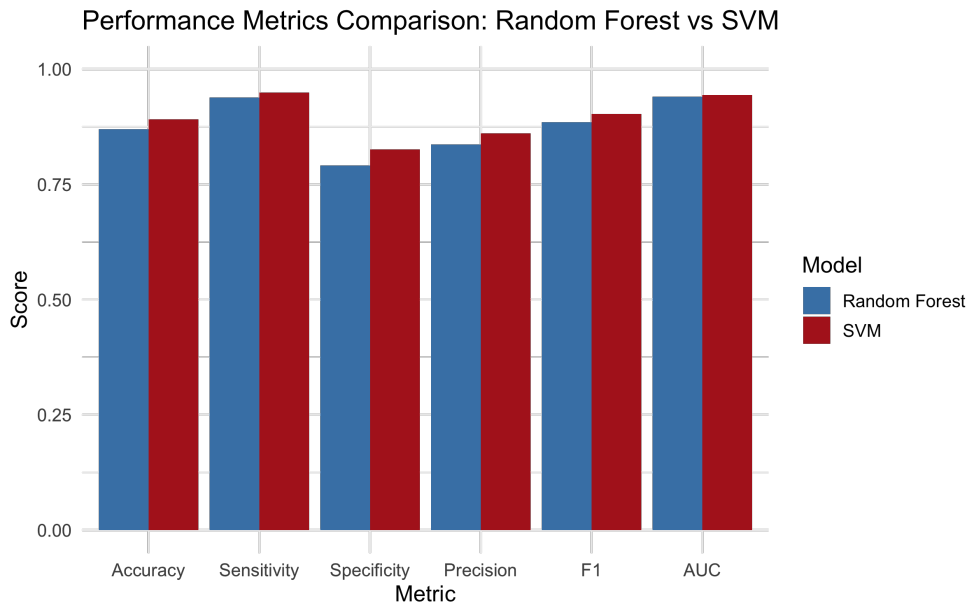


Figure 16: Comparison of key performance metrics (Accuracy, Sensitivity, Specificity, Precision, F1, and AUC) for Random Forest and SVM on the **test set**.

Model	Acc	Sens	Spec	Prec	F1	AUC
RF	**0.885**	**0.950**	0.806	0.857	**0.901**	**0.959**
SVM	0.882	0.924	**0.831**	**0.869**	0.896	0.944

Table 6: Training performance metrics for Random Forest and SVM. Bold values indicate the higher value between the two models.

Model	Acc	Sens	Spec	Prec	F1	AUC
RF	0.870	0.939	0.791	0.836	0.885	0.940
SVM	**0.891**	**0.949**	**0.826**	**0.861**	**0.903**	**0.944**

Table 7: Test performance metrics for Random Forest and SVM. Bold values indicate the higher value between the two models.

Both models show strong generalisation on unseen data. The SVM model achieves the highest **accuracy (0.891)**, **sensitivity (0.949)**, **F1-score (0.903)**, and **AUC (0.944)**, indicating better overall performance. Random Forest also performs well but has lower **specificity (0.791)** and slightly weaker discrimination **(AUC = 0.940)**. Together with the confusion matrices and ROC analysis presented earlier, these results confirm that the **SVM with RBF kernel** is the stronger and more reliable model for this classification task.

6 Interpretation of the Trained Models Using XAI Techniques

To better understand how the Random Forest (RF) and Support Vector Machine (SVM) models arrive at their predictions, we computed global permutation feature importance (PFI). This method measures how much the prediction performance decreases when the values of a feature are randomly permuted. A greater performance drop implies a more influential feature.

6.1 (1) Global Feature Importance Patterns

Figures 17a and 17b show the PFI rankings for the RF and SVM models, respectively. In both cases, **ST_Slope** emerges as the most important predictor of heart disease. This indicates that ECG-derived information about the slope of the ST segment during exercise stress testing is highly influential for the model's decision-making.

Several other features consistently appear with high importance across both models, including **ChestPainType**, **Sex**, **ExerciseAngina**, **Cholesterol**, and **Oldpeak**. Although the exact numerical importance differs between RF and SVM, both models rely on a very similar set of predictors.

6.2 (2) Agreement Between RF and SVM

The two models exhibit strong agreement in their overall importance structure. Both highlight symptom-related features (e.g., ChestPainType, ExerciseAngina) and stress-test metrics (ST_Slope, Oldpeak) as the most informative. This consistency suggests that the underlying signal in the dataset is stable and is captured robustly across different machine learning algorithms.

Lower-ranked features such as **Age**, **RestingBP**, and **RestingECG** appear to have limited contribution. This does not imply that they are clinically irrelevant, but that in this dataset and model configuration, they contribute less compared to the dominant predictors.

6.3 (3) Most and Least Important Features

Across both RF and SVM:

- **Most important:** ST_Slope
- **Moderately important:** ChestPainType, Sex, ExerciseAngina, Cholesterol, Oldpeak, MaxHR, FastingBS
- **Least important:** Age, RestingBP, RestingECG

These results help identify which variables the models rely on most for heart disease classification.

6.4 (4) What Feature Importance Does Not Tell Us

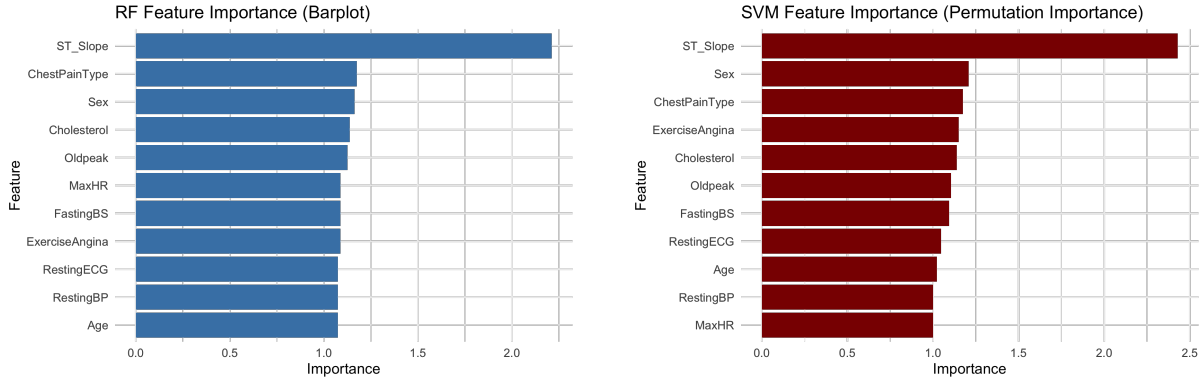
PFI is a global interpretability method and therefore offers only a limited type of insight. In particular, the feature importance plots do *not* reveal:

- whether higher feature values increase or decrease predicted risk,
- whether effects are linear or non-linear,
- interactions between variables,
- how individual patient predictions are formed.

Such insights would require tools such as SHAP values, ICE curves, or partial dependence plots, which were not used in this section.

6.5 (5) Summary

Overall, the feature importance analysis provides a global understanding of which clinical attributes the models rely on most. Both RF and SVM show similar and clinically coherent patterns, with exercise-related ECG features emerging as the dominant predictors. This consistency increases confidence in the interpretability and reliability of the models.



(a) Random Forest feature importance

(b) SVM feature importance

Figure 17: Permutation feature importance for RF and SVM models.

References

References

[1] Soriano, F. (2020). *Heart Failure Prediction Dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[2] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

[3] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

[4] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[6] Wright, M. N., Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

[7] Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

[8] Schölkopf, B., Smola, A. J. (2002). *Learning with Kernels*. MIT Press.

[9] Bergstra, J., Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281–305.

[10] Shahriari, B., et al. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1), 148–175.

[11] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1–26.

[12] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation. *IJCAI*.

[13] Florkowski, C. M. (2008). Sensitivity, specificity, ROC curves, and Youden's index. *Clin Biochem Rev*, 29(Suppl 1), S83–S87.

[14] Hanley, J. A., McNeil, B. J. (1982). The meaning and use of the area under the ROC curve. *Radiology*, 143(1), 29–36.

[15] Molnar, C. (2022). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

[16] Lundberg, S. M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS*.

[17] Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Stat Med*, 27(15), 2865–2873.

[18] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training. *ICML*.

[19] Gulati, M. et al. (2021). Heart Disease and Risk Assessment: A Clinical Overview. *Circulation*, 143(5), 583–596.

[20] Visseren, F. et al. (2021). 2021 ESC Guidelines on Cardiovascular Prevention. *European Heart Journal*.

[21] OpenAI. (2024). *ChatGPT (GPT-5.1) Model*. https://www.openai.com/

[22] Google DeepMind. (2024). *Gemini Large Language Model*. https://deepmind.google/

[23] xAI. (2024). *Grok Large Language Model*. https://x.ai/