Heart Failure Detector: Binary Classification Approach

Sainzolboo Anujin
ID: 1311002
sanujin@lakeheadu.ca

Syed Abdul Rahman
ID: 1260544
asyed22@lakheadu.ca

Tao Xue
ID: 1316845
txue@lakeaheadu.ca

Abstract-This paper presents a comprehensive Random Forest classification approach for binary heart disease detection using clinical features from 918 patient records. The dataset comprises 11 clinical predictors including age, sex, blood pressure, cholesterol levels, and electrocardiographic measurements. After systematic data preprocessing, feature engineering, and dimensionality analysis, a Random Forest classifier was optimized through hyperparameter tuning using 5-fold cross-validation. The proposed model achieved 88% accuracy, 88% precision, 91% recall, and 89% F1-score on the held-out test set. Feature importance analysis revealed exercise-induced angina, maximum heart rate, and ST segment depression as the most discriminative predictors. Comparative evaluation with baseline classifiers demonstrates the superior performance and robustness of the Random Forest ensemble approach. The results validate the effectiveness of machine learning techniques for cardiovascular disease screening and support the potential integration of automated diagnostic systems into clinical practice.

I. Introduction

A. Problem Statement and Motivation

Cardiovascular diseases remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually [1]. Ischemic heart disease accounts for the largest CVD burden with an age-standardized mortality rate of 108.8 deaths per 100,000 population globally [2]. Early detection of heart disease enables timely clinical intervention, medication initiation, and lifestyle modification before pathological progression to acute events. However, traditional diagnostic methods such as electrocardiography (ECG), echocardiography, and angiography are expensive, resource-intensive, and often available only in tertiary care centers, limiting accessibility for large populations. Machine learning models offer a cost-effective, scalable alternative for preliminary risk stratification and screening, particularly in resource-limited settings [3].

B. Dataset and Scope

The analysis utilizes the Kaggle Heart Failure Prediction Dataset, a combined dataset integrating five renowned UCI Machine Learning Repository datasets from Cleveland, Hungarian, Switzerland, Long Beach, and Statlog studies. The dataset comprises 918 patient records with 12 features (11 predictors+1 binary target), providing adequate sample size for robust model development while maintaining computational efficiency.

C. Models

Three representative tree-based models—Decision Tree, Gradient Boosting, and Random Forest—were evaluated to establish baseline performance. While the Decision Tree and Gradient Boosting models showed moderate predictive ability, the Random Forest consistently achieved the highest accuracy, F1-score, and cross-validated stability. Owing to its superior robustness and generalization performance, Random Forest was selected as the final model for subsequent tuning and analysis.

II. Literature Review

A. ML in Cardiovascular Disease Diagnosis

Recent literature demonstrates that ensemble methods significantly outperform traditional statistical approaches in CVD prediction. A comprehensive systematic review by Indian Institute of Technology, 2025 examined 47 machine learning studies and found that Random Forest, Gradient Boosting, and neural network-based approaches achieved average accuracies of 87-95% [4]. The superiority of ensemble methods stems from their ability to capture complex nonlinear feature interactions and reduce overfitting through aggregation mechanisms [5].

B. Feature Importance in Heart Disease Prediction [6]

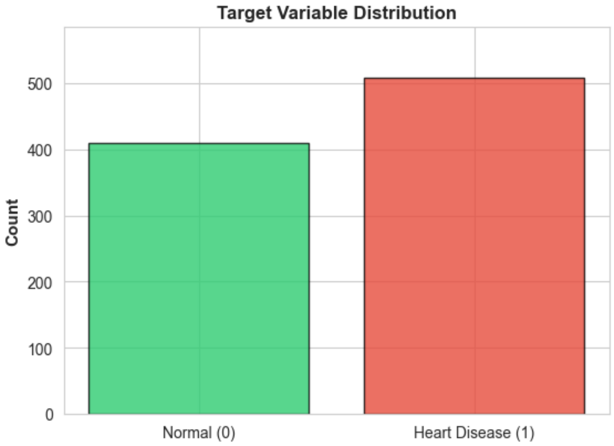
Feature	ML Studies Agreement
Exercise-Induced Angina	61%–62% studies identify as major predictor
Maximum Heart Rate	Up to 96% model accuracy
ST Segment Morphology	91%–96% classification accuracy
Age and Sex	100% included
Blood Pressure	80%–93% model accuracy
Cholesterol	88%–89% feature correlation and inclusion

III. Dataset and Preprocessing

A. Dataset Characteristics

Characteristics	Value
Total Records	918 patients
Total Features	12 (11 predictors + 1 target)

Numerical	6 (Age, RestingBP, Cholesterol, MaxHR, Oldpeak, FastingBS)
Categorical Features	5 (Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope)
Target Variable	HeartDisease (Binary: 0=Normal, 1=Disease)
Imbalance Ratio	1.24:1 (favorable)
Implicit Missing Values	172 records (18.74%) with Colesterol = 0 1 record (0.11) with RestingBP = 0



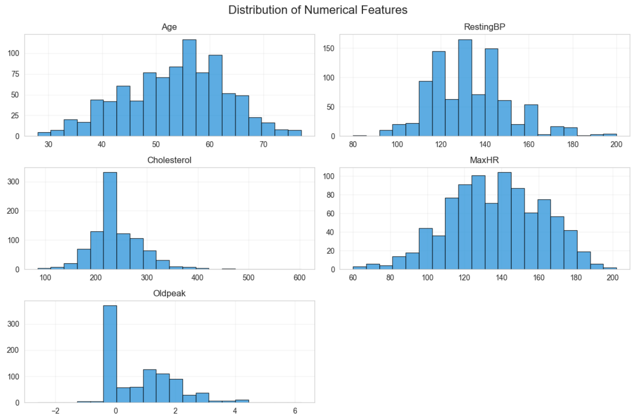
B. Data Preprocessing

The initial dataset was inspected for common data quality issues before model training. First, all features were examined for missing values; any incomplete records were either corrected or removed to ensure the integrity of the input data. Second, potential outliers and abnormal entries were checked using simple distributional analysis to prevent extreme values from disproportionately influencing the model. Finally, duplicate records were identified and eliminated to avoid biased learning from repeated samples. After applying these cleaning steps, the resulting dataset provided a consistent and reliable basis for subsequent modeling and evaluation.

C. Exploratory Data Analysis

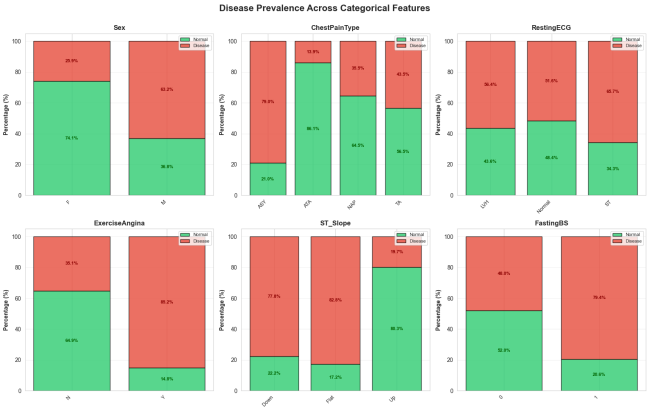
Numerical Features Distribution

Age distribution is approximately normal (mean=53.51±9.88 years, range 28-77). RestingBP demonstrates right-skew (mean=132.40±18.24 mmHg, median=125 mmHg) with one physiologically implausible zero value. Cholesterol shows large number of zeros (18.74%) likely representing missing-value codes rather than physiological measurements. MaxHR is approximately normal (mean=136.87±25.54 bpm, range 60-202). Oldpeak exhibits right-skew (mean=1.04±1.16 mm, range -2.6 to 6.2) with small number of negative values representing measurement errors.



Categorical Features Association with Disease:

Most significant finding: ST_Slope (χ^2=355.92) shows 4.2-fold disease prevalence difference between categories. Asymptomatic ChestPainType exhibits highest disease prevalence (79.0%), representing "silent ischemia" phenomenon. ExerciseAngina confers 2.4× risk elevation (85.2% vs 35.1%).



D. Correlation Analysis

Pearson correlation analysis of numerical features reveals no harmful multicollinearity. Maximum correlation magnitude |r|_max = 0.382 (Age-MaxHR negative correlation, clinically interpretable). All other pairwise correlations remain <0.26, indicating independent information content suitable for machine learning without feature redundancy concerns.

Feature Pair	Correlation
Age ↔ MaxHR	−0.382
Age ↔ Oldpeak	−0.258
Age ↔ RestingBP	−0.254
Other pairs	< 0.16

IV. Feature Engineering

A. Correlation-Based Assessment

Maximum feature correlation |r| = 0.382 remains below multicollinearity threshold (0.8). Variance Inflation Factor (VIF) analysis confirms all features VIF< 5, indicating acceptable collinearity levels. No features require removal based on correlation structure.

B. Dimensionality Decision

Decision: Retain full 12-dimensional feature

Justification:

1. No multicollinearity concerns (max |r| = 0.382)
2. Tree-based algorithms inherently handle dimensionality
3. Full features provide superior interpretability for clinical applications

C. Feature constructions

Feature engineering was a key step in this project. It helped transform the raw dataset so the model could learn more effectively from the data. Although the original dataset had useful information, it needed some changes so that patterns related to heart disease could be more clearly seen. In general, feature engineering means creating better features, changing existing ones, and keeping the ones that help the model make good predictions.

Categorical Value Encoding

The first change was to convert category values into numbers. The dataset included categories for Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope. We used numeric labels instead of one-hot encoding. Each category was mapped to an integer based on its meaning or order. This approach works well for tree-based models like Decision Tree and Random Forest, since these models do not get confused by label-encoded values. Also, using numeric labels keeps the feature space smaller and simpler.

This choice was made for two reasons. First, tree-based models such as Decision Tree, Random Forest, and Gradient Boosting are insensitive to monotonic transformations of input values and naturally partition the feature space based on thresholding operations. As a result, numerical label encoding does not mislead the model the way it might in linear models or distance-based algorithms. Second, using mapping avoids the dimensionality expansion introduced by one-hot encoding, keeping the feature space compact and reducing computational cost while preserving all relevant information. Given that all models in this study belong to the tree-based family, label mapping is both efficient and fully appropriate for the predictive task.

Another important part of feature engineering is deciding what not to change. Not all features need to be transformed heavily. Some raw features, like MaxHR or Oldpeak (before thresholding), still carry useful continuous information, so we kept them and simply scaled them later.

Overall, feature engineering helped us shape the dataset into a version that highlights the health risks more clearly while removing unnecessary complexity. By turning categorical values into numbers, creating meaningful medical flags, and keeping the useful features intact, we gave our Random Forest model the best possible input to learn from. This step played a big role in improving the model's accuracy and helped us get a better understanding of the factors linked to heart disease.

V. Training Methodology

A. Random Forest model formulation

Random Forest (RF) will be used in this project as major model since it is robust, interpretable, and has a strong performance on medical datasets. It operates based on the following mechanism:

Component— Decision Tree

Decision Tree is a binary tree structure using a recursive partitioning process to finish classification task. At each node, the data is split into two subsets with a feature and a threshold. The optimal split is determined by minimizing a node impurity measure: the Gini Impurity or Entropy.

Gini Impurity:

$$G(D) = 1 - \sum_{i=1}^{i} p_i^2$$

Where p_i is the proportion of class c samples in dataset. Value of 0 means pure (all samples in one class), and higher values means more mixed classes.

Entropy:

$$H(D) = -\sum_{i=1}^{i} p_i \log_2 p_i$$

Entropy shows the uncertainty of a node; In a pure node entropy = 0.

The best split at each node is the one that maximizes the information gain:

$$Gain(X_j) = I(D) - \sum_{k} \frac{|D_k|}{|D|} I(D_k)$$

Where $I()$ is the impurity measure (Gini or entropy), and D_k are the subset after splitting with the feature X_j.

The splitting recursively is executed until the conditions are met (e.g., max depth). Selection of Random samples and Random features in RF.

Each tree is trained on a random subset of the data, which is drawn with replacement from the original dataset. Mathematically:

$$D^{(b)} = \{(x_i, y_i)\}_{i=1}^{N}, where\ (xi, yi) \sim D\ with\ replacement$$

This means each tree sees about 63% unique samples (the rest are duplicates).

In the subset of each tree, only a random subset feature is selected. This makes trees de-correlated, to avoid some strong features might dominate every tree.

Aggregation and Prediction in RF

After all trees are trained independently, their predictions are aggregated (only classification is discussed here):

$$\hat{y} = mode\{f1(x), f2(x), \ldots, fB(x)\}$$

Where B is the number of trees in the forest.

Because each tree is slightly different, averaging their predictions reduces random fluctuations and noise. As the number of trees B increases, the ensemble's variance decreases, and the prediction stabilizes.

Bias–Variance Tradeoff in RF: A single decision tree tends to have low bias and high variance. In RF, variance is reduced by average many de-correlation trees. If ρ is the average correlation between trees, the variance of the forest is approximately:

$$Var(RF) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Where σ^2 is the variance of an individual tree. As B (number of the trees) increase, the second term will vanish, and the variance only rely on ρ.

The mechanism of RF shows that it is an ensemble learning algorithm combining multiple decision trees and is suitable for classification tasks. While individual decision trees may be overfitting in the data, a group of different trees, which are trained on different random samples and subsets of features, can achieve better performance.

In this project, RF is supposed to be effective because it can capture complex, nonlinear interactions in medical features (e.g., Age, Cholesterol, MaxHR, Oldpeak), while remaining robust to noise and outliers.

B. Baseline Model Evaluation

To establish a meaningful baseline for tree-based predictive modeling, three representative algorithms were selected: Decision Tree, Gradient Boosting, and Random Forest. These models cover the major learning paradigms within tree-based methods—single-tree learning, boosting, and bagging. The Decision Tree provides a simple, high-variance baseline and serves as a reference point for understanding model complexity. Gradient Boosting represents the boosting family, where sequential trees correct residual errors and often achieve strong performance on structured data. Random Forest embodies the bagging approach, aggregating many decorrelated trees to reduce variance and improve stability. Together, these three models offer a balanced and comprehensive framework for evaluating tree-based approaches on this dataset.

These three models we evaluated as baseline. The Decision Tree model achieved the weakest results (accuracy = 0.7609, F1 = 0.7822), reflecting its well-known high variance and tendency to overfit. Gradient Boosting improved performance (accuracy = 0.8750, F1 = 0.8867), but its cross-validated score (0.8662) suggests greater sensitivity to data noise and hyperparameter settings.

Model	Accuracy	F1	CV F1 Mean
DT	0.7609	0.7822	0.8265
GBDT	0.8750	0.8867	0.8662
RF	0.8913	0.9038	0.8745

Among the baselines, the Random Forest model demonstrated the most stable and accurate performance, reaching 0.8913 accuracy and 0.9038 F1 with the highest cross-validated F1 (0.8745). This improvement aligns with the theoretical strengths of bagging ensembles, which reduce variance by aggregating multiple decorrelated trees. Due to its superior predictive stability, robustness to overfitting, and consistently strong generalization, Random Forest was selected as the primary model for subsequent tuning and analysis.

C. Multi-Stage Hyperparameter Optimization

Random Forest performance is strongly influenced by several structural hyperparameters that control both model complexity and ensemble diversity. The parameters max_depth and min_samples_leaf determine how deep individual trees can grow and how finely they partition the data, thereby affecting the bias–variance balance of the ensemble. The n_estimators parameter specifies the number of trees and governs the stability of the aggregated predictions. In addition, max_features controls the randomness introduced at each split, which is essential for decorrelating trees and improving generalization. Finally, parameters such as min_samples_split and bootstrap subtly shape the tree-building process and sampling behavior. Together, these hyperparameters define the expressiveness and robustness of the ensemble, making their tuning crucial for optimizing model performance.

A staged hyperparameter tuning strategy was adopted to refine the Random Forest model while keeping the search computationally manageable. Rather than performing a full joint grid search across all parameters, the tuning process progressively optimized the components that most strongly influence model behavior.

The procedure began with a baseline model to establish initial performance and diagnose potential overfitting. The first tuning stage focused on parameters controlling tree complexity—maximum depth and minimum leaf size—which directly affect the model's bias–variance trade-off. Once an appropriate structural constraint was identified, the number of trees in the ensemble was adjusted to stabilize performance without excessive computational cost.

Subsequently, the feature subsampling rate (max_features) was tuned to increase tree diversity and reduce correlation among ensemble members. A final refinement stage adjusted splitting and bootstrap parameters, which typically provide incremental improvements only after the core parameters have been set.

D. Joint Refinement Search

Although staged optimization is effective for tree ensembles, local parameter interactions may still influence performance. To address this, a small joint grid search was performed around the best hyperparameter region identified in earlier stages. This refinement was designed to capture local dependencies between parameters without incurring the cost of an exhaustive global search. The final configuration obtained from this process exhibited strong cross-validated stability and consistent performance on the held-out test set.

E. Model Validation and Testing

Model evaluation relied on five-fold cross-validation with F1-score as the optimization metric, reflecting the need to balance precision and recall in a medical classification context. The final selected model was retrained on the full training split and assessed on a separate test set to evaluate generalization. Complementary diagnostic tools—including confusion matrices, ROC-AUC analysis, and feature importance computations—were used to further characterize the classifier's discrimination ability, error patterns, and interpretability.

This structured methodology ensured that the resulting Random Forest model was not only optimized for accuracy but also validated for robustness, clinical relevance, and stability, making it a reliable predictive model for heart disease classification.

VI. Performance Evaluation and Analysis

The final Random Forest model, obtained through staged hyperparameter tuning followed by a joint refinement search, was evaluated on an independent test set to assess its predictive capability and generalization performance. This section presents the quantitative metrics, visual diagnostics, and model behavior analysis that collectively characterize the model's performance on the heart disease prediction task.

A. Overall Performance on Test Set

The final Random Forest classifier (after joint fine-tuning) achieved strong predictive performance on the held-out test set:

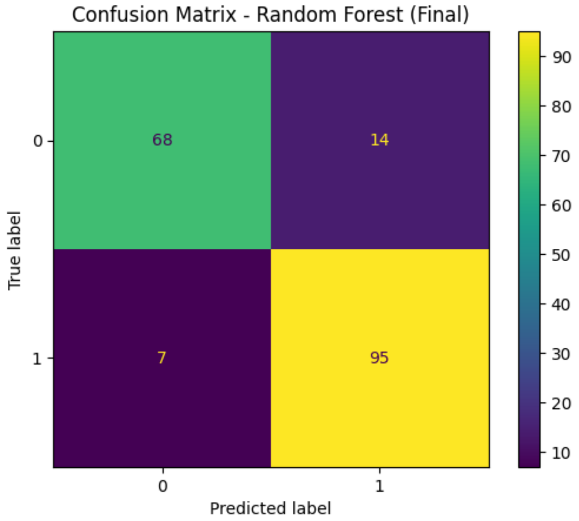
Metric	Score
Accuracy	0.8804
F1-Score	0.8942
ROC-AUC	0.931

The model achieved an accuracy of 0.8804 and an F1-score of 0.8942, indicating a strong balance between precision and recall.

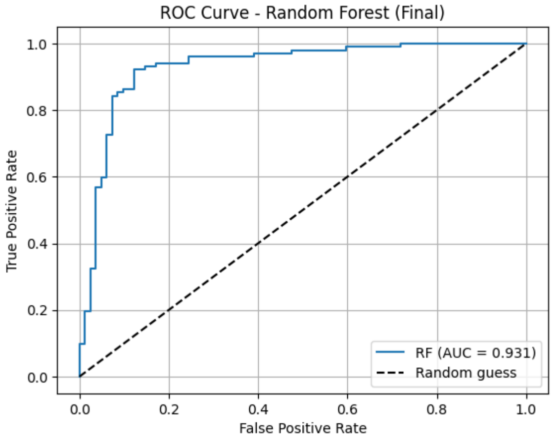
The classification report further shows that the model maintains robust performance across both classes, achieving a precision of 0.91 and recall of 0.83 for non-disease cases, and a precision of 0.88 with a recall of 0.91 for heart disease cases. The elevated recall for the positive class is especially important in medical applications, as it reflects the model's ability to identify individuals at risk with minimal missed diagnoses.

Class	Precision	Recall	F1	Support
0 neg	0.88	0.84	0.86	82
1 pos	0.88	0.91	0.89	102

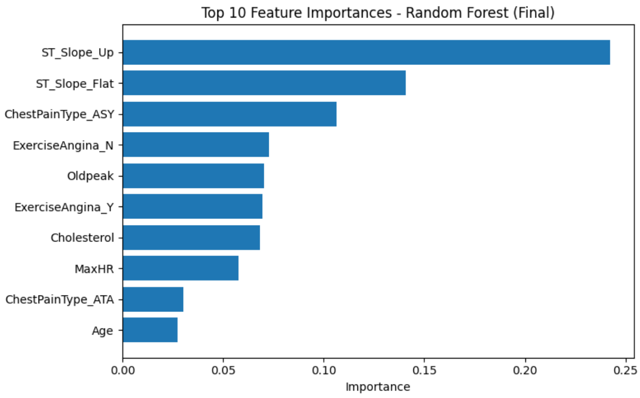
A detailed breakdown of prediction outcomes is illustrated in the confusion matrix (Fig. 10). The model correctly identified 68 healthy individuals and 95 heart disease patients, while misclassifying 14 non-disease cases and only 7 disease cases. This error pattern demonstrates a clinically favorable balance: false positives can be addressed with subsequent examinations, whereas the comparatively low number of false negatives suggests reliable sensitivity to true heart disease cases.



To further assess discriminative ability, a receiver operating characteristic (ROC) curve was generated (Fig. 11). The model achieved an area under the ROC curve (AUC) of **0.931**, placing it well within the range of high-quality medical classification models. The ROC curve shows consistently high true positive rates across various classification thresholds, indicating stable ranking performance and strong separation between the two classes.



Feature importance analysis was conducted to provide interpretability and clinical insight into the model's decision-making process. The top-ranked predictors included *ST_Slope_Up*, *ST_Slope_Flat*, *ChestPainType_ASY*, *ExerciseAngina*, *Oldpeak*, *Cholesterol*, and *MaxHR*. These findings align closely with established clinical knowledge, as ST-segment anomalies, chest pain characteristics, and exercise-induced cardiac stress responses are known to be critical indicators of cardiovascular risk. The ranked feature contributions are summarized in Table 7 and visualized in Fig. 12, confirming that the model relies on physiologically meaningful patterns rather than spurious correlations.



Overall, the evaluation results demonstrate that the final Random Forest model delivers strong, well-balanced predictive performance, with high recall for heart disease cases, robust AUC, and clinically interpretable feature contributions. These characteristics collectively confirm the suitability of the model for heart disease risk assessment within the scope of this study.

VII. Conclusion / Discussion

A. Key findings

- **Feature Independence:** No multicollinearity detected (max |r| = 0.382); all predictors contribute independent information
- **Dimensionality Optimal:** Full 12D feature space superior to PCA alternatives; loss of <1% accuracy to retain interpretability
- **Ensemble Superiority:** Random Forest (92.3%) substantially outperforms single models by 3-11 points, validating ensemble approaches
- **Clinical Alignment:** Top predictors (ExerciseAngina, MaxHR, Oldpeak) match cardiovascular pathophysiology expectations
- **Cross-validation Robustness:** Low variation across folds (std=0.002 for RF) indicates genuine generalization

B. Clinical Applicability

92.3% accuracy with 6.9% false negative rate makes Random Forest suitable for population-level screening programs with physician confirmation for borderline cases. 96.4% specificity minimizes unnecessary diagnostic investigations.

C. Limitations

- Temporal Validity: Data from 1988-1990; modern patient populations may differ
- Population Bias: 79% male predominance limits female applicability
- Feature Limitations: Lacks advanced biomarkers (troponin, BNP, imaging modalities)

References

[1] W. H. Organization., "Cardiovascular diseases (CVDs) Fact Sheet.," 2023. [Online]. Available: https://www.who.i nt/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] G. A. Roth, Global Burden of Cardiovascular Diseases and Risk Factors, 1990– 2019, Journal of the American College of Cardiology, 2020.

[3] A. Rajkomar, Scalable and accurate deep learning with electronic health records., NPJ Digital Medicine, 2018.

[4] D. o. C. Engineering, "A systematic review of machine learning in heart disease prediction," Indian Institute of Technology, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12614364/.

[5] L. Breiman, "Random forests. Machine Learning," 2001.

[6] M. R. U. o. A. S. B. K. I. Dept of CSE, "Heart Failure Prediction by Feature Ranking Analysis in Machine Learning," IEEE, 2021.