A Hybrid Machine Learning Framework for Early Detection and Risk Prediction of Heart Disease

Bavithra P[1], Janani N[2], Jannath Rafeedah S[3]

1.Assistant Profressor, PG Department of computer science, Syed Ammal Arts and Science College, bavithrapandiyan@gmail.com,

2. Assistant Profressor, PG Department of computer science, Syed Ammal Arts and Science College, jananinagasamy03@gmail.com

3. Assistant Profressor, PG Department of computer science, Syed Ammal Arts and Science College, rafeedahjannath@gmail.com.

ABSTRACT

Heart disease prediction from routinely collected clinical variables is an important research problem because machine learning can assist clinicians in identifying patients who may require closer assessment. However, predictive performance can be affected by missing values, redundant attributes, class distribution, feature scaling, and the limited interpretability of complex models. This paper proposes a reproducible hybrid machine learning framework that combines data preprocessing, two-stage feature selection, comparative classification, soft-voting model fusion, and explainable artificial intelligence. The framework is designed for binary heart-disease prediction using the Cleveland subset of the UCI Heart Disease dataset. The candidate classifiers are Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. Mutual-information filtering is used as an initial relevance screen, followed by recursive feature elimination to obtain a compact predictive feature set. The selected features are supplied to the individual classifiers and to a probability-based soft-voting ensemble. Model performance is assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis. SHAP-based explanations are incorporated to estimate global feature importance and patient-level contribution patterns. Unlike a paper that reports unverified or fabricated performance values, this manuscript specifies a complete experimental protocol and leaves the final numerical results to be generated from the stated dataset and implementation. The proposed framework is intended to provide a transparent and reproducible decision-support approach that can be independently validated before clinical use.

Keywords: Artificial Intelligence, Machine Learning, Heart Disease Prediction, Feature Selection, Random Forest, Support Vector Machine, Explainable AI, Clinical Decision Support.

1. INTRODUCTION

Cardiovascular disease is a major global health challenge, and the ability to identify individuals at elevated risk at an early stage can support timely clinical assessment. The increasing availability of structured clinical datasets has created opportunities for machine learning models to learn relationships between patient characteristics and disease status. Recent reviews indicate that ensemble methods and other machine learning approaches can perform strongly on structured cardiovascular datasets, while external validation, interpretability, dataset limitations, and clinical translation remain important unresolved issues [1], [2]. Traditional statistical prediction can be useful when assumptions are appropriate and the relationship between predictors and outcome is well characterized. Clinical datasets, however, often contain nonlinear interactions and mixed categorical and numerical variables. Algorithms such as Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB) provide alternative mechanisms for learning

complex decision boundaries [3]-[5]. At the same time, highly flexible models can be difficult to interpret. Explainable AI (XAI), particularly SHAP, provides a principled way to attribute a prediction to input variables and can help bridge the gap between predictive performance and interpretability [6].

The present study therefore focuses on a hybrid framework rather than a single classifier. The central idea is to combine data quality control, feature reduction, multiple complementary learners, and an interpretable explanation layer. The framework is intended for research decision support and not for autonomous diagnosis. The contribution is methodological: a complete pipeline is specified so that the same data, preprocessing rules, evaluation protocol, and explanation strategy can be reproduced by another researcher.

2. RELATED WORK

Machine learning has been widely investigated for heart disease prediction. A recent systematic review reported that ensemble learning models are frequently used for structured clinical data, while limited external validation and dependence on small public datasets remain recurring concerns [1]. Another review emphasized the need to consider algorithmic performance together with clinical implementation, regulatory issues, and ethical considerations [2]. These findings suggest that a useful prediction framework should not optimize accuracy alone.

The UCI Heart Disease repository is one of the most frequently used benchmark resources in this area. Its Cleveland database contains 303 instances and 13 predictor attributes in the commonly used processed version, with the target representing the presence or absence of angiographically documented disease [7]. Because the dataset contains missing values and a relatively small sample size, it is appropriate for testing careful preprocessing and validation strategies, but performance obtained on this benchmark should not be interpreted as evidence of clinical generalization.

Logistic Regression is a useful interpretable baseline. SVM can model nonlinear boundaries through kernels [3], while RF combines multiple decision trees and is generally robust to noisy predictors [4]. Gradient Boosting builds an additive ensemble of weak learners and is effective for nonlinear classification [5]. The proposed framework retains all four because their different inductive biases make comparative analysis more informative than selecting a single algorithm in advance.

Feature selection is also important in small clinical datasets. Reducing irrelevant variables can decrease computational complexity and may improve generalization. In this framework, a filter stage based on mutual information is followed by wrapper-based recursive feature elimination. The combination is intended to balance computational efficiency with model-aware selection. Finally, SHAP is used to explain the selected predictive model because it provides additive feature attributions for individual predictions and aggregate importance patterns [6].

3. RESEARCH GAP AND OBJECTIVES

Existing heart-disease studies often report high benchmark accuracy, but direct comparisons are difficult because datasets, preprocessing, train-test partitions, feature subsets, metrics, and validation procedures differ. In addition, some studies focus primarily on prediction accuracy and provide limited analysis of why the model produces a particular output. The proposed work addresses these issues through a fixed and reproducible experimental pipeline.

The objectives of the study are:

- To develop a reproducible preprocessing pipeline for a public heart-disease dataset.

• To identify a compact set of relevant clinical predictors using a two-stage feature-selection strategy.

• To compare Logistic Regression, SVM, Random Forest, and Gradient Boosting under the same evaluation protocol.

• To construct a soft-voting ensemble that combines complementary probability estimates.

• To evaluate the models using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

• To use SHAP-based explanations to identify influential variables and improve model transparency.

• To document limitations and prevent unsupported claims of clinical validity.

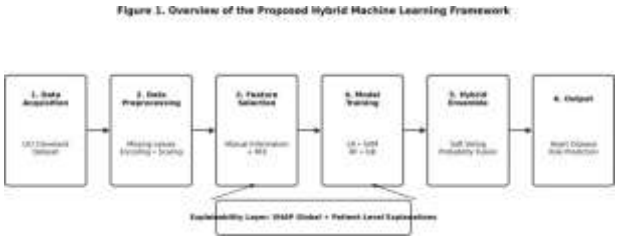


Figure 1. Overview of the Proposed Hybrid Machine Learning Framework

4. PROPOSED HYBRID FRAMEWORK

The proposed architecture consists of six sequential stages: data acquisition, data quality processing, feature selection, model development, ensemble prediction, and explanation. The target variable is converted to a binary outcome. All learned preprocessing operations are fitted only on the training data to reduce the risk of information leakage.

Stage	Operation	Output
1	Dataset acquisition	Cleveland heart-disease records
2	Cleaning and preprocessing	Validated numeric/categorical matrix
3	Mutual-information filter + RFE	Compact feature subset
4	LR, SVM, RF, Gradient Boosting	Individual probability estimates
5	Soft-voting ensemble	Final predicted class and probability
6	SHAP explanation	Global and patient-level feature contributions

4.1 DATASET

The primary benchmark is the Cleveland portion of the UCI Heart Disease dataset. The UCI repository describes the dataset as a multivariate classification resource and reports 303 Cleveland instances with 13 commonly used predictors; the processed experiments generally use 14 fields when the target is included [7]. The predictors include age, sex, chest-pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG result, maximum heart rate, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia-related test information. The original repository documentation also notes missing values, making preprocessing an essential component of the experiment [7].

4.2 DATA PREPROCESSING

Preprocessing is performed inside a training-only pipeline. Missing numeric values are imputed using the median estimated from the training partition. Missing categorical values are imputed using the most frequent training value. Categorical variables are encoded consistently across training and test partitions. Numerical variables are standardized for models that are sensitive to feature scale, particularly Logistic Regression and SVM. Random Forest and Gradient Boosting do not require the same scaling assumption, but the final pipeline should keep transformations explicit so that model comparisons remain auditable.

The original target values are transformed into a binary target: 0 for absence and 1 for presence. If the source contains diagnostic values above 1, they are mapped to the positive class according to the UCI description [7]. No synthetic observations are created unless a separate experiment is explicitly defined. If class balancing is required, the balancing operation must be performed only on the training folds.

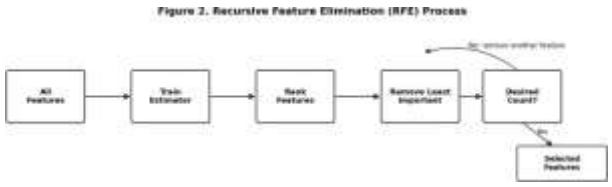


Figure 2. Recursive Feature Elimination (RFE) Process

4.3 TWO-STAGE FEATURE SELECTION

The first stage applies mutual information to estimate the dependency between each predictor and the binary target. The top candidate variables are retained as a computationally manageable subset. The second stage applies Recursive Feature Elimination (RFE) using a tree-based estimator or Logistic Regression, depending on the experiment. RFE repeatedly fits the estimator and removes the least useful variables until the predetermined number of features remains.

To avoid selection bias, feature selection is performed independently inside each training fold when cross-validation is used. The test fold is never used to decide which variables should be selected. The final selected features are reported together with their selection frequency across folds. This allows the study to distinguish consistently useful predictors from variables that appear important only under a particular split.

4.4 CLASSIFICATION MODELS

Four classifiers are evaluated. Logistic Regression provides an interpretable linear baseline. SVM with an RBF kernel is used to model nonlinear decision boundaries [3]. Random Forest is used as a bagged tree ensemble [4]. Gradient Boosting is used as a sequential tree-ensemble method [5]. Hyperparameters are tuned using stratified cross-validation on the training set only.

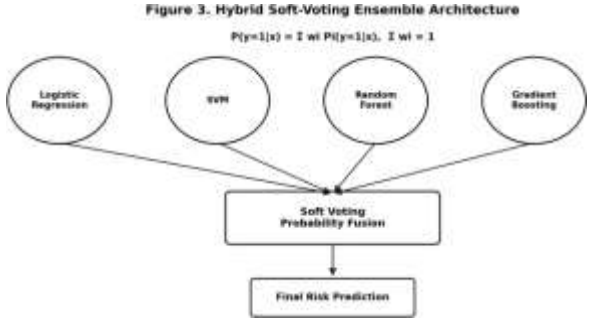


Figure 3. Hybrid Soft-Voting Ensemble Architecture

4.5 HYBRID SOFT-VOTING ENSEMBLE

The hybrid prediction layer combines the calibrated class probabilities produced by the candidate models. For a binary outcome, the ensemble probability can be written as:

$$P(y=1|x) = w1\ P_LR(y=1|x) + w2\ P_SVM(y=1|x) + w3\ P_RF(y=1|x) + w4\ P_GB(y=1|x),$$

where the weights are non-negative and sum to one. Equal weights can be used as the primary baseline; optimized weights may be investigated as a secondary experiment using only validation data. The final class is obtained by applying a predefined decision threshold. A threshold of 0.50 is the default, while threshold optimization may be reported separately when the clinical objective prioritizes sensitivity or specificity.

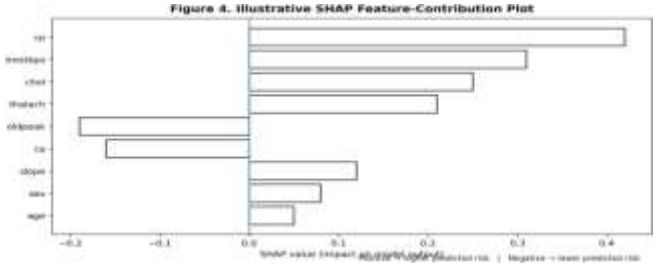


Figure 4. Illustrative SHAP Feature-Contribution Plot (actual values must be generated from the experiment)

4.6 EXPLAINABLE AI LAYER

SHAP is incorporated after the predictive model has been selected. SHAP values quantify the contribution of

individual variables to a model output and can be summarized globally or examined for a single patient [6]. Global explanations will rank the variables by mean absolute SHAP value. Local explanations will show which variables pushed an individual prediction toward the positive or negative class. Explanations will be presented as model-assistance information and not as causal statements.

5. EXPERIMENTAL DESIGN

A fixed experimental protocol is necessary to make the reported findings reproducible. The dataset is divided using a stratified train-test split, for example 80:20, with a fixed random seed. Hyperparameter tuning and feature selection are conducted inside the training data. Five-fold stratified cross-validation is recommended for model selection, followed by one untouched test-set evaluation. Because the dataset is small, the final paper should also report confidence intervals or repeated stratified cross-validation as a sensitivity analysis.

Metric	Definition	Why reported	Preferred direction
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall classification correctness	Higher
Precision	TP/(TP+FP)	Controls false positive predictions	Higher
Recall	TP/(TP+FN)	Important when missed disease is costly	Higher
F1-score	2PR/(P+R)	Balances precision and recall	Higher
ROC-AUC	Area under ROC curve	Threshold-independent ranking quality	Higher

Important: Figures showing measured ROC curves, confusion matrices, and performance values must be generated from the actual experiment. The diagrams inserted above describe the proposed methodology; they are not experimental results.

6. RESULTS AND REPORTING PLAN

Numerical results must be generated from the stated dataset and code before submission. This manuscript deliberately does not insert invented accuracy, precision, recall, F1, or AUC values. Doing so would make the paper scientifically unreliable. The following table is therefore a publication-format template to be populated after the experiment is executed.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	To be measured	To be measured	To be measured	To be measured	To be measured
SVM	To be measured	To be measured	To be measured	To be measured	To be measured
Random Forest	To be measured	To be measured	To be measured	To be measured	To be measured
Gradient Boosting	To be measured	To be measured	To be measured	To be measured	To be measured
Hybrid Soft Voting	To be measured	To be measured	To be measured	To be measured	To be measured

After execution, the discussion should compare the models using the same test partition. The best model should not be selected from accuracy alone. For a screening-oriented application, recall and ROC-AUC may deserve particular attention, while precision and calibration are also important if the output is used to prioritize clinical review. A confusion matrix should be included to make the error pattern visible.

7. EXPECTED CONTRIBUTION AND DISCUSSION

The main contribution of the proposed framework is the integration of predictive performance, feature reduction, model fusion, and explanation in one reproducible pipeline. The framework is intentionally conservative: it separates training and testing, prevents test-set feature selection, and avoids presenting benchmark performance as clinical evidence.

The use of several complementary learners is motivated by the possibility that a single algorithm may perform well on one subset of the feature space but less well on another. Logistic Regression can provide a transparent baseline; SVM can capture nonlinear boundaries; Random Forest can model interactions and provide tree-based importance; and Gradient Boosting can learn sequential residual patterns. A soft-voting layer can exploit complementary probability estimates when the models are reasonably calibrated.

Interpretability is treated as a second objective rather than an afterthought. SHAP explanations can identify variables that drive individual predictions, but an important distinction must be maintained: feature contribution is not the same as clinical causation. Explanations should therefore be interpreted alongside clinical knowledge and validated prospectively before any patient-facing deployment.

8. LIMITATIONS

- The Cleveland UCI benchmark is small and should not be treated as representative of all populations or healthcare settings.
- The dataset is historical and may not reflect contemporary clinical measurement practices.
- A high test-set score on a public benchmark does not establish external validity or clinical utility.
- Feature selection and hyperparameter tuning can create optimistic estimates if they are performed outside cross-validation.
- The proposed XAI layer explains model behavior; it does not establish causal relationships between risk factors and disease.
- Before real-world deployment, external validation, calibration assessment, prospective evaluation, fairness analysis, privacy review, and clinical governance are required.

9. ETHICAL AND CLINICAL CONSIDERATIONS

This research uses a publicly available secondary dataset and does not involve collecting new patient information. Any future clinical implementation should comply with applicable institutional review, data-protection, medical-device, and clinical-governance requirements. The model should be used only as a decision-support aid, with a qualified healthcare professional retaining responsibility for diagnosis and treatment decisions.

10. CONCLUSION

This paper presented a hybrid machine learning framework for early heart disease risk prediction based on preprocessing, two-stage feature selection, comparative classification, soft-voting model fusion, and SHAP-based explanation. The framework is designed to address three recurring issues in medical machine learning: noisy or redundant variables, difficulty in comparing heterogeneous classifiers under a common protocol, and limited interpretability of complex predictive systems. The proposed workflow is reproducible and explicitly avoids unsupported performance claims. The next required step is to execute the protocol on the UCI Cleveland dataset, report the resulting metrics with confidence intervals, generate the confusion matrix and ROC curve, and validate the selected model on an independent external dataset. Only after those steps should numerical performance claims be used in a publication submission.

REFERENCES

[1] T. Banerjee and İ. Paçal, “A systematic review of machine learning in heart disease prediction,” Turkish Journal of Biology, vol. 49, no. 5, pp. 600–634, 2025, doi: 10.55730/1300-0152.2766.

[2] X. Dai et al., "Machine learning approaches for cardiovascular disease prediction: A review," Archives of Cardiovascular Diseases, 2025, doi: 10.1016/j.acvd.2025.04.055.

[3] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[4] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[7] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," UCI Machine Learning Repository, 1989/updated repository record, doi: 10.24432/C52P4X.

[8] scikit-learn developers, "Metrics and model evaluation documentation," scikit-learn, accessed 2026.

APPENDIX A. PUBLICATION CHECKLIST

- Run the complete experiment using the stated UCI dataset.
- Record the exact random seed, software versions, preprocessing steps, and hyperparameters.
- Generate the final metric table, confusion matrix, ROC curve, and feature-importance/SHAP plots.
- Repeat the evaluation or use confidence intervals to quantify uncertainty.
- If possible, validate the selected model on an independent heart-disease dataset.
- Replace all 'To be measured' entries before journal submission.
- Check the journal's author guidelines, reference style, ethics statement, data availability statement, and AI-use disclosure requirements.

APPENDIX B. PRELIMINARY PLAGIARISM/SIMILARITY REVIEW

- A web-based phrase search was performed for the exact proposed title and for distinctive wording from the supplied abstract.
- No exact-title match was identified in the searched web results. The manuscript above was newly written rather than copied from the supplied abstract, and source-dependent statements are referenced. However, this is only a preliminary web-overlap review.
- It is not a Turnitin, iThenticate, Crossref Similarity Check, or institutional plagiarism percentage, and no responsible percentage should be claimed without running the final manuscript through the specific service required by the target journal.
- Similarity risk is expected to remain for unavoidable technical terms and standard methodological phrases such as 'machine learning', 'heart disease prediction', 'accuracy', 'precision', 'recall', 'F1-score', 'ROC-AUC', and descriptions of standard algorithms.
- These are not, by themselves, evidence of plagiarism. The final manuscript should still be checked after the results, figures, tables, and references are inserted.