



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85206>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Comparative Evaluation of Machine Learning Algorithms for Binary Classification of Heart Disease: A Cross-Validated Analysis of a Clinical Dataset

Taniya Mondal¹, Durjodhan Mondal²

Department of IT, Techno Main Salt Lake, WB

Abstract: Machine learning (ML) algorithms are increasingly used to classify heart disease using routinely collected clinical variables. This study compared nine supervised classifiers to identify the best-performing model for binary heart disease classification. A secondary analysis was conducted using a dataset of 303 observations, 13 predictors, and a binary outcome. Data completeness and outcome distribution were assessed, and no missing values were identified. The dataset included 165 positive cases (54.5%) and 138 negative cases (45.5%). Logistic Regression, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, radial-basis-function Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Decision Tree were evaluated using five-fold stratified cross-validation. Numerical predictors were standardized, and categorical predictors were one-hot encoded within preprocessing pipelines. Performance was assessed using accuracy, precision, sensitivity, specificity, F1-score, and receiver operating characteristic area under the curve (ROC-AUC). Logistic Regression achieved the highest reported accuracy (86.5%), precision (85.2%), sensitivity (90.9%), specificity (81.2%), F1-score (88.0%), and ROC-AUC (0.914). Random Forest and AdaBoost achieved ROC-AUC values of 0.901, while Extra Trees achieved 0.896. Decision Tree performed least effectively, with an accuracy of 74.3% and ROC-AUC of 0.739. Logistic Regression emerged as the strongest-performing classifier in this comparison and represents a reasonable baseline for further investigation. However, the findings remain preliminary because confidence intervals, formal paired comparisons, nested hyperparameter tuning, calibration assessment, and independent external validation were unavailable. Further research should reproduce the analysis using the raw dataset and evaluate model generalizability before considering clinical applications.

Keywords: Heart disease; machine learning; Logistic Regression; Random Forest; ROC-AUC; cross-validation; clinical prediction.

I. INTRODUCTION

Cardiovascular diseases are a major public health challenge and are associated with substantial morbidity, mortality, and healthcare utilization. Early identification of people who may have heart disease can support timely clinical evaluation and appropriate diagnostic investigations. Prediction models can assist risk stratification, but their performance and usefulness depend on the population, predictor definitions, outcome ascertainment, and intended clinical role.

Conventional regression methods remain valuable in clinical prediction because they are relatively interpretable and can quantify associations between predictors and outcomes. Machine learning (ML) methods provide complementary tools for detecting complex patterns in structured clinical data. Algorithms such as Random Forest, Extra Trees, Support Vector Machine (SVM), and boosting methods can model nonlinear relationships, while Logistic Regression offers a useful and interpretable baseline.

A robust comparison should not rely on accuracy alone. Sensitivity measures the proportion of positive cases correctly identified, whereas specificity measures the proportion of negative cases correctly classified. Precision and F1-score provide additional information about positive predictions, and the area under the receiver operating characteristic curve (ROC-AUC) measures discrimination across classification thresholds [1]. However, discrimination alone does not establish calibration or clinical utility.

Validation and transparent reporting are central to prediction-model research. Cross-validation can estimate performance across multiple partitions of a dataset, while TRIPOD+AI provides updated guidance for reporting prediction models developed using regression or ML methods [2,3]. The present study compared nine commonly used classifiers on a small clinical dataset, with the objective of identifying the strongest-performing tested configuration and describing the implications for further research.

II. METHODS

A. Study design and data source

This was a secondary analysis of a structured clinical dataset supplied as heart.xls. The file was identified as comma-separated text despite its .xls extension and was imported for analysis accordingly. It contained 303 observations, 13 predictor variables, and one binary outcome. The structure resembles a commonly used processed heart disease dataset, but the exact provenance of the supplied file was not independently verified. The UCI Machine Learning Repository documents a Heart Disease collection that includes the processed Cleveland subset [4].

B. Variables and outcome

Predictors included age and sex; resting blood pressure, serum cholesterol, maximum heart rate, and ST-segment depression (oldpeak); fasting blood sugar, resting electrocardiographic findings, and exercise-induced angina; and coded clinical categories including chest pain type, ST-segment slope, number of major vessels (ca), and thal. The dependent variable was target, coded 0 or 1. Target = 1 was treated as the positive class and target = 0 as the negative class for the reported metrics. Clinical interpretation of the coding should be checked against the original dataset documentation.

C. Data quality and preprocessing

The dataset was assessed for missing values, variable structure, and target distribution. No missing values were detected. Five continuous predictors—age, resting blood pressure, cholesterol, maximum heart rate, and oldpeak—were standardized. The remaining eight predictors were treated as categorical and one-hot encoded. Transformations were included in preprocessing pipelines and fitted within the training portion of each cross-validation fold to reduce information leakage. Median and most-frequent imputation were specified in the numerical and categorical pipelines, respectively, although imputation was not required for the supplied data.

D. Classifiers

Nine classifiers were compared: Logistic Regression, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, SVM with a radial basis function kernel, K-Nearest Neighbors, Gaussian Naive Bayes, and Decision Tree. These methods represent linear, tree-based, ensemble, distance-based, and probabilistic classification approaches. The comparison used specified configurations; exhaustive hyper parameter optimization and nested cross-validation were not performed.

E. Validation and performance measures

Five-fold stratified cross-validation with shuffling and a fixed random seed was used. Out-of-fold predictions were pooled to calculate accuracy, precision, sensitivity, specificity, and F1-score; predicted probabilities were used for ROC-AUC. The metrics were defined as follows: accuracy = $(TP + TN)/(TP + TN + FP + FN)$; precision = $TP/(TP + FP)$; sensitivity = $TP/(TP + FN)$; specificity = $TN/(TN + FP)$; and $F1 = 2 \times (\text{precision} \times \text{sensitivity})/(\text{precision} + \text{sensitivity})$, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. ROC-AUC was interpreted as a measure of discrimination [1,5].

The analysis was descriptive. Confidence intervals, formal paired statistical tests, calibration assessment, decision-curve analysis, and external validation were not performed. Consequently, differences in observed performance should not be interpreted as statistically significant or as evidence of clinical readiness.

III. RESULTS

A. Dataset profile

The dataset contained 303 observations and 14 columns, comprising 13 predictors and one outcome. No missing values were identified. There were 165 positive observations (54.5%) and 138 negative observations (45.5%), indicating a moderately balanced outcome distribution. The combination of demographic, physiological, and coded clinical variables provided a basis for comparing the selected supervised classifiers.

B. Overall model comparison

Table 1 summarizes the reported out-of-fold performance of all nine classifiers. Logistic Regression achieved the highest accuracy (86.5%), precision (85.2%), sensitivity (90.9%), specificity (81.2%), F1-score (88.0%), and ROC-AUC (0.914). Random Forest and AdaBoost each achieved a ROC-AUC of 0.901. Extra Trees achieved 82.8% accuracy and 0.896 ROC-AUC. Decision Tree had the lowest reported accuracy (74.3%) and ROC-AUC (0.739).

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	ROC-AUC
Logistic Regression	86.5	85.2	90.9	81.2	88.0	0.914
Random Forest	83.5	82.9	87.9	78.3	85.3	0.901
AdaBoost	82.2	82.1	86.1	77.5	84.0	0.901
Extra Trees	82.8	83.0	86.1	79.0	84.5	0.896
SVM (RBF)	81.8	81.6	86.1	76.8	83.8	0.895
Gradient Boosting	81.2	81.0	85.5	76.1	83.2	0.881
KNN	82.2	81.7	86.7	76.8	84.1	0.879
Gaussian Naive Bayes	77.6	79.0	80.0	74.6	79.5	0.847
Decision Tree	74.3	75.7	77.6	70.3	76.6	0.739

Table 1. Comparative performance of nine classifiers using the reported five-fold cross-validation results.

C. Sensitivity and specificity

Sensitivity and specificity varied across the algorithms. Logistic Regression showed the highest reported sensitivity (90.9%) and specificity (81.2%). Random Forest showed sensitivity of 87.9% and specificity of 78.3%. The Decision Tree classifier had the lowest sensitivity (77.6%) and specificity (70.3%) among the evaluated configurations. These metrics describe the observed cross-validated classifications at the thresholds used in the analysis.

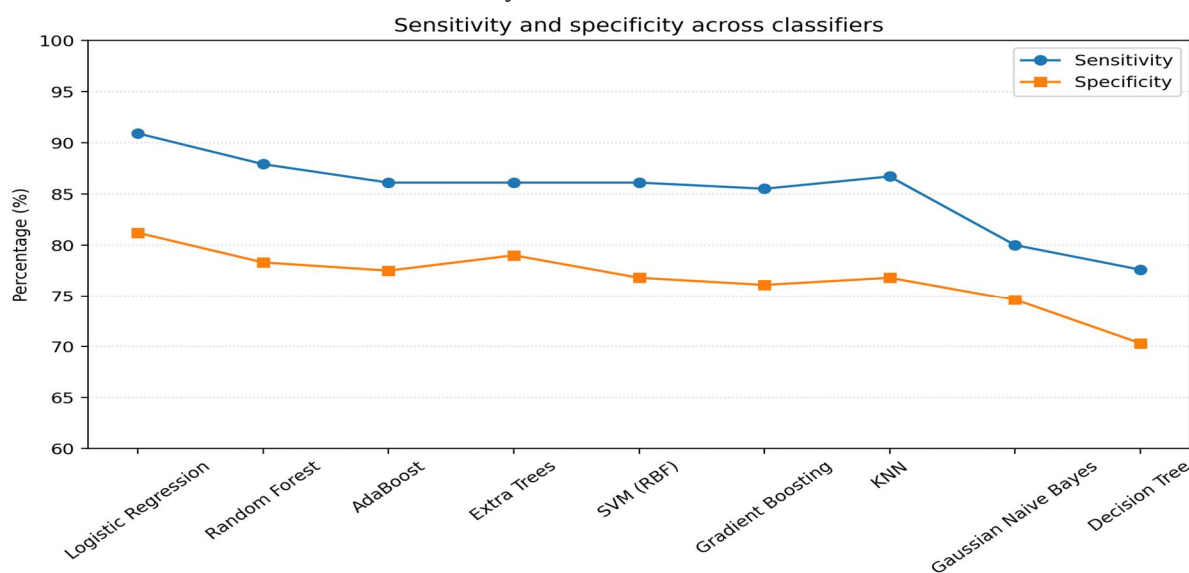


Figure 1. Sensitivity and specificity by classifier. Values are reported percentages at the evaluated classification thresholds.

D. Confusion matrix for Logistic Regression

The pooled out-of-fold confusion matrix reported for Logistic Regression contained 112 true negatives, 26 false positives, 15 false negatives, and 150 true positives. Thus, 262 of 303 observations were classified correctly, corresponding to 86.5% accuracy. The 15 false-negative classifications are particularly relevant when considering potential screening applications, while the 26 false positives illustrate the additional assessment burden that may arise from incorrect positive classifications.

Observed class	Predicted negative (0)	Predicted positive (1)
Negative (0)	112	26
Positive (1)	15	150

Table 2. Logistic Regression confusion matrix based on the reported pooled out-of-fold predictions.

IV. DISCUSSION

A. Principal Findings

In this comparative evaluation, Logistic Regression was the leading classifier across the reported metrics. Its ROC-AUC of 0.914 suggests good apparent discrimination within the dataset. Random Forest and AdaBoost also performed favourably, whereas Decision Tree and Gaussian Naive Bayes showed lower performance. The findings reinforce that greater algorithmic complexity does not automatically yield superior performance on a particular dataset.

Logistic Regression may have performed well because the available predictors and their relationship to the outcome were adequately represented by the model's decision function. However, this study did not examine the functional form of each predictor, interactions, or alternative specifications in sufficient depth to explain the ranking. The result should therefore be understood as an empirical comparison of the configurations tested, not a universal hierarchy of algorithms.

B. Regression-based and ensemble approaches

Logistic Regression is computationally efficient and relatively interpretable [6-9]. When appropriately specified, it can provide coefficient estimates and odds ratios with confidence intervals, supporting assessment of adjusted associations. These properties make it a useful baseline for clinical prediction research. Ensemble methods such as Random Forest, Extra Trees, and boosting algorithms can represent nonlinear relationships and interactions, but their performance depends on sample size, feature representation, and hyper parameter settings.

Random Forest and AdaBoost achieved ROC-AUC values of 0.901, close to the value reported for Logistic Regression. Such differences may be within sampling variability; because confidence intervals and paired comparisons were not calculated, the results do not establish statistically significant superiority. Nested cross-validation and external validation would provide a more robust comparison [2,3,5].

C. Potential clinical interpretation

At the evaluated threshold, Logistic Regression's sensitivity of 90.9% indicates that most positive observations were identified, while specificity of 81.2% indicates that a substantial proportion of negative observations were correctly classified [10-11]. The balance between false negatives and false positives should be selected according to the intended clinical use, the consequences of errors, and the availability of confirmatory testing.

ROC-AUC evaluates discrimination across thresholds but does not establish whether predicted probabilities are accurate. Calibration plots, calibration intercept and slope, and decision-curve analysis should be considered in future work. Predictive values may also vary with disease prevalence in the target population. Accordingly, the current estimates should not be used as clinical performance claims without validation in a relevant independent cohort.

D. Strengths

The analysis compared nine algorithms using a common framework, employed stratified cross-validation, incorporated preprocessing into pipelines, and reported several complementary performance measures. This is more informative than relying solely on training accuracy or a single performance metric. The use of established reporting guidance such as TRIPOD+AI can further improve transparency and reproducibility [2].

E. Limitations

The study has important limitations. First, the sample comprised only 303 observations, so estimates may be sensitive to individual observations and fold composition. Second, the exact source and representativeness of the uploaded file were not independently verified. Third, no independent external validation was performed. Fourth, exhaustive hyperparameter tuning and nested cross-validation were not undertaken. Fifth, confidence intervals and formal paired model comparisons were unavailable. Sixth, calibration, subgroup performance, fairness, and decision-curve analysis were not evaluated. Finally, predictive associations do not establish causal relationships.

F. Implications for future research

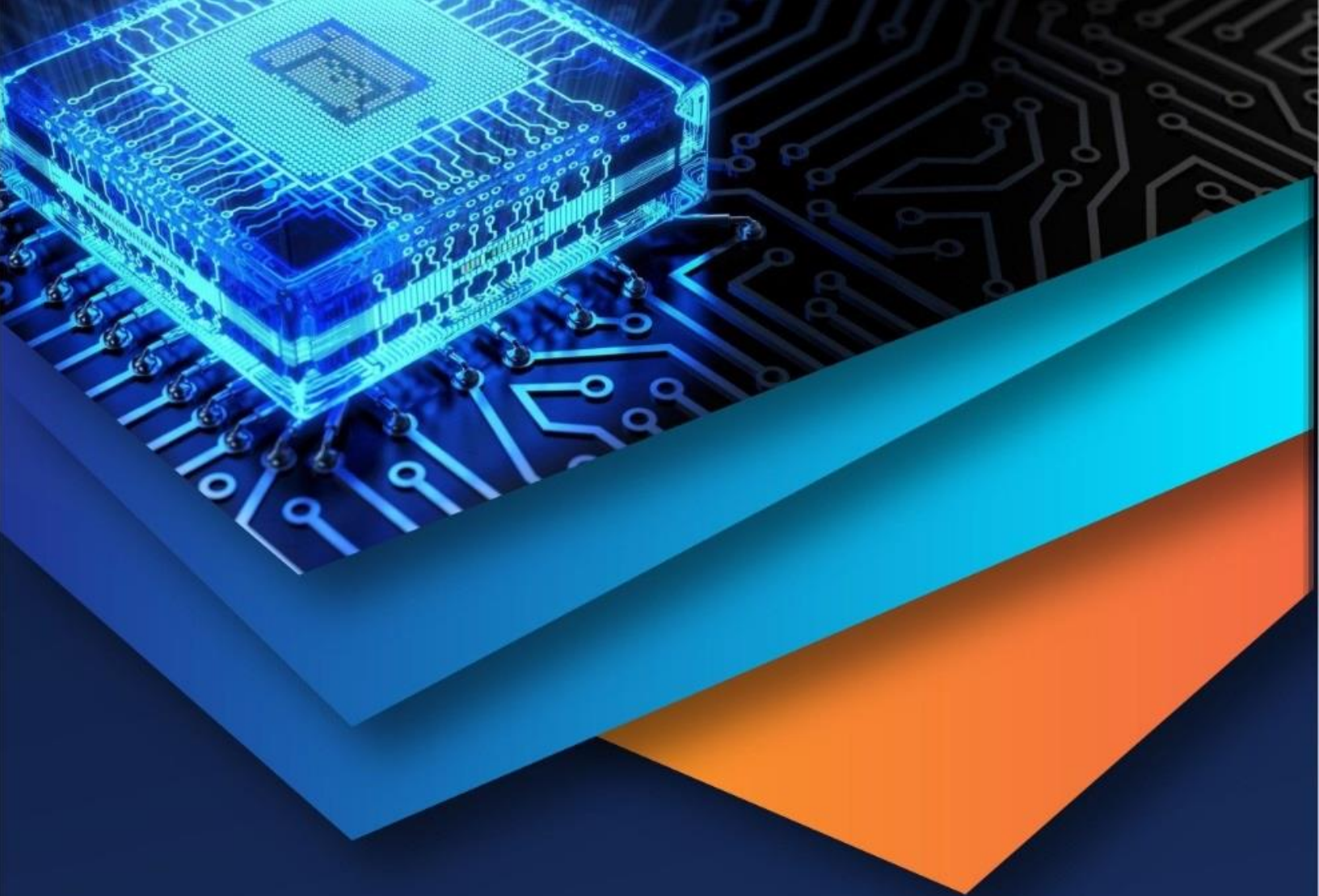
Future studies should reproduce the complete pipeline from the raw dataset, document all preprocessing and model settings, and conduct nested cross-validation for model selection. Confidence intervals should be calculated for discrimination and threshold-dependent metrics, and paired comparisons should be used where appropriate. The leading models should be evaluated for calibration, subgroup performance, and clinical utility. External validation using independently collected data from a different clinical setting is essential before clinical translation. Reporting should follow current prediction-model recommendations, including TRIPOD+AI [2].

V. CONCLUSION

Logistic Regression demonstrated the most favourable reported performance among the nine evaluated classifiers, with accuracy of 86.5%, sensitivity of 90.9%, specificity of 81.2%, F1-score of 88.0%, and ROC-AUC of 0.914. Random Forest and AdaBoost were also promising alternatives, while Decision Tree showed the lowest performance. Logistic Regression is therefore a reasonable leading baseline for further investigation using this dataset. Nevertheless, the limited sample size, absence of external validation, lack of uncertainty estimates, and absence of calibration analyses limit the conclusions. The findings are preliminary and should not be interpreted as evidence that the model is ready for clinical use.

REFERENCES

- [1] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36. doi:10.1148/radiology.143.1.7063747.
- [2] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
- [3] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
- [4] UCI Machine Learning Repository. Heart Disease [Internet]. Irvine (CA): University of California, Irvine. Available from: <https://archive.ics.uci.edu/dataset/45/heart+disease>. Accessed 13 October 2026.
- [5] Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006;27(8):861–874. doi:10.1016/j.patrec.2005.10.010.
- [6] Farhadi Z, Feizi-Derakhshi MR, Bevrani H, Kim W, Ijaz MF. ERDeR: The combination of statistical shrinkage methods and ensemble approaches to improve the performance of deep regression. *IEEE Access*. 2024 Feb 21;12:33361–83.
- [7] Kilimci ZH. Ensemble regression-based gold price (XAU/USD) prediction. *Journal of Emerging Computer Technologies*. 2022 Jul 7;2(1):7–12.
- [8] Fong KS, Motani M. SyREC: A Symbolic-Regression-Based Ensemble Combiner. In: 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI) 2024 Oct 28 (pp. 774–781). IEEE.
- [9] Da Silva PM, Lima MN, Soares WL, Silva IR, Fagundes RA, de Souza FF. Ensemble regression models applied to dropout in higher education. In: 2019 8th Brazilian conference on intelligent systems (BRACIS) 2019 Oct 15 (pp. 120–125). IEEE.
- [10] Flores MG, Barone MC, Bachur MR, Lin PH. Errors of medical interpretation and their potential clinical consequences: a comparison of professional versus ad hoc versus no interpreters. *Annals of emergency medicine*. 2012 Nov 1;60(5):545–53.
- [11] Flores MG, Barone MC, Bachur MR, Lin PH. Errors of medical interpretation and their potential clinical consequences: a comparison of professional versus ad hoc versus no interpreters. *Annals of emergency medicine*. 2012 Nov 1;60(5):545–53.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)