© Journal of the Technical University - Sofia
Plovdiv branch, Bulgaria
"Fundamental Sciences and Applications "
Vol. 32 2026

Machine Learning-Based Heart Disease Prediction: A Comparative Analysis with Feature Importance Evaluation

M.Kemal Tezcan [1], Umit Kursun [2], Burak Dursun [3]

[1] Istanbul Beykent University 1; kemaltezcan@beykent.edu.tr
[2] Istanbul Beykent University 2; umitkursun@beykent.edu.tr
[3] Istanbul Beykent University 3; burakdursun@beykent.edu.tr

Presented at the 15th International Scientific Conference TechSys 2026, Plovdiv, Bulgaria, 14–16 May 2026.

Abstract: Heart disease continues to rank as the foremost contributor to premature death across the globe, and developing reliable automated screening tools has become a pressing priority in clinical informatics. This study presents a comparative analysis of three machine learning classifiers — Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) — for binary heart disease prediction using the UCI Cleveland Heart Disease dataset. The dataset comprises 297 patient records with 13 clinical features following removal of missing values. Models were evaluated using an 80/20 stratified train-test split with Min-Max normalisation. Logistic Regression achieved the highest AUC-ROC of 0.9554 and accuracy of 83.33%, followed by Random Forest (AUC-ROC = 0.9375, accuracy = 85.00%) and Gradient Boosting (AUC-ROC = 0.8828, accuracy = 76.67%). Feature importance analysis identified chest pain type (cp), thalassemia (thal), and maximum heart rate (thalach) as the most influential predictors. Results indicate that well-calibrated linear classifiers can match or exceed ensemble methods in structured clinical settings, a practically significant finding for risk-stratification tools where probability estimates must be trusted.

Keywords: heart disease prediction; machine learning; logistic regression; random forest; gradient boosting; feature importance; UCI Cleveland dataset; clinical decision support

1. Introduction

According to WHO estimates, approximately 17.9 million people die from cardiovascular causes annually, placing cardiac conditions at the top of global disease burden rankings. Identifying at-risk patients before symptoms become severe allows physicians to intervene early, dramatically altering disease trajectory and reducing hospitalisation rates. Yet translating routine clinical measurements — blood pressure, cholesterol, ECG readings — into reliable diagnoses is non-trivial, as these variables interact in ways that resist simple rule-based interpretation.

Data-driven classification methods have shown considerable potential in this context, extracting latent predictive structure from tabular patient records without requiring manually engineered diagnostic rules. In contrast to traditional expert systems, such methods adapt directly to observed data patterns, making them well-suited to capturing the multivariate complexity of cardiac risk profiles. Within the field of Intelligent Systems and Monitoring of Complex Systems, such data-driven approaches represent a natural extension of automated monitoring and control paradigms to biomedical domains.

 Online ISSN 2603-459X

Among publicly available cardiac datasets, the Cleveland collection assembled by Detrano et al. [1] has become the standard testbed for comparative classification studies in this domain. Despite extensive prior work, systematic comparisons of interpretable classifiers under a unified, reproducible preprocessing framework — with explicit feature importance analysis — remain valuable for guiding deployment decisions in clinical settings where model transparency and probabilistic calibration are paramount.

This paper contributes: (i) a reproducible preprocessing pipeline for the UCI Cleveland dataset; (ii) a fair comparative evaluation of three ML classifiers under identical experimental conditions; (iii) quantitative performance analysis using Accuracy, F1-Score, and AUC-ROC; and (iv) a feature importance analysis identifying the most clinically significant predictors.

2. Related Work

A substantial body of work has applied supervised learning to the Cleveland cardiac records over the past two decades. The dataset originates from a prospective clinical study by Detrano et al. [1], who showed that invasive angiographic outcomes could be approximated from non-invasive physiological measurements alone. Building on this foundation, researchers have tested a broad array of supervised algorithms on these records. Mohan et al. [2] proposed a hybrid random forest model achieving 88.7% accuracy, while Gavhane et al. [3] demonstrated that support vector machines and neural networks attain competitive results on this benchmark.

Shah et al. [4] benchmarked several algorithms including Naive Bayes, k-nearest neighbours and decision trees, concluding that ensemble approaches yield higher accuracy on average yet do not consistently dominate on discriminability metrics such as AUC-ROC. A recurring observation in this literature is that Logistic Regression, despite its parametric simplicity, produces well-calibrated probability scores that translate directly into reliable risk rankings — a property particularly valued in clinical scoring contexts [5]. The present work contributes a transparent, reproducible head-to-head evaluation of three representative classifiers, augmented by a quantitative feature importance analysis to illuminate which clinical variables drive predictive performance.

3. Dataset and Preprocessing

3.1. Dataset Description

The dataset employed in this study comprises 303 patient records gathered prospectively at the Cleveland Clinic Foundation and made publicly available through the UCI Machine Learning Repository [1]. Six entries containing missing attribute values were discarded during quality screening, leaving a working sample of 297 complete records. A binary outcome variable was constructed by collapsing the original five-level angiographic severity score: grade 0 was treated as the negative (healthy) class, while grades 1 through 4 were merged into a single positive (disease) class, producing 160 negative and 137 positive cases — a distribution sufficiently balanced to proceed without oversampling.

The 13 clinical features used as predictors are described in Table 1.

Table 1. Clinical Features of the UCI Cleveland Heart Disease Dataset.

Feature	Description	Type	Range
age	Age in years	Continuous	29–77
sex	Sex (1=male, 0=female)	Binary	0, 1
cp	Chest pain type (1–4)	Ordinal	1–4
trestbps	Resting blood pressure (mm Hg)	Continuous	94–200
chol	Serum cholesterol (mg/dl)	Continuous	126–564
fbs	Fasting blood sugar > 120 mg/dl	Binary	0, 1
restecg	Resting ECG results (0–2)	Ordinal	0–2
thalach	Maximum heart rate achieved	Continuous	71–202
exang	Exercise-induced angina (1=yes)	Binary	0, 1
oldpeak	ST depression induced by exercise	Continuous	0–6.2
slope	Slope of peak exercise ST segment	Ordinal	1–3
ca	No. of major vessels coloured by flourosopy	Ordinal	0–3
thal	Thalassemia type (3=normal, 6=fixed, 7=reversable)	Ordinal	3,6,7

3.2. Preprocessing

A uniform preprocessing pipeline was applied prior to model training. Incomplete records (flagged with '?' in the source file) were removed, reducing the dataset to 297 usable instances. The multi-class target was recoded as a binary indicator distinguishing healthy from diseased patients. Feature magnitudes were standardised to the unit interval via Min-Max scaling, with transformation parameters derived exclusively from the training fold to guard against information leakage. Data were split into a 80% training partition and a 20% held-out test set, with stratification applied to maintain consistent class proportions in both subsets.

4. Methodology

4.1. Classifiers

Three algorithms were trained and assessed under matched experimental conditions. Logistic Regression maps feature values to disease probability through a log-odds linear function, yielding inherently calibrated outputs and serving here as both a clinical baseline and an interpretable reference. Random Forest constructs an ensemble of 100 bootstrapped trees whose aggregated votes reduce prediction variance and accommodate complex feature interactions. Gradient Boosting builds 100 trees in an additive, residual-correcting sequence that typically achieves low bias on sufficiently large datasets. All implementations were drawn from scikit-learn v1.3 [6], with the random seed set to 42 throughout to ensure full reproducibility.

4.2. Evaluation Metrics

Three evaluation criteria were selected to capture distinct aspects of classifier behaviour. Overall classification accuracy provides a headline correctness rate across all predictions. The F1-Score harmonises precision and recall into a single figure, penalising both over-prediction and under-prediction of the positive class — an important consideration given that missed diagnoses and unnecessary referrals carry asymmetric clinical costs. The area under the ROC curve (AUC-ROC) quantifies ranking

discrimination across the full range of decision thresholds, making it the primary metric for comparing classifiers intended for continuous risk scoring rather than fixed-threshold binary decisions.

4.3. Feature Importance Analysis

To identify the most influential clinical variables, feature contribution scores were extracted from the fitted Random Forest model using the mean decrease in Gini impurity criterion. Each score represents the average reduction in node impurity attributable to splits on that variable, aggregated across all trees; higher scores indicate stronger discriminative utility within the ensemble.

5. Results and Discussion

5.1. Classification Performance

The performance metrics obtained on the 60-instance held-out test partition are presented in Table 2.

Table 2. Classification Performance on UCI Cleveland Test Set (n = 60).

Model	Accuracy	F1-Score	AUC-ROC ★
Gradient Boosting	0.7667	0.7407	0.8828
Random Forest	0.8500	0.8302	0.9375
Logistic Regression ✦	0.8333	0.8148	0.9554

★ Primary ranking metric. ✦ Best AUC-ROC. ↑ Higher is better.

Logistic Regression recorded the strongest AUC-ROC (0.9554), reflecting its advantage in rank-ordering patients by disease risk across all possible cut-offs. Random Forest led on both accuracy (85.00%) and F1-Score (0.8302), while Gradient Boosting trailed on every criterion.

5.2. Feature Importance Analysis

Table 3 presents the top five most important features identified by the Random Forest model using mean decrease in impurity.

Table 3. Top 5 Feature Importances (Random Forest, Gini Impurity).

Rank	Feature	Description
1	cp	Chest pain type
2	thal	Thalassemia type
3	thalach	Maximum heart rate achieved
4	oldpeak	ST depression (exercise-induced)
5	ca	No. of major vessels (fluoroscopy)

Importance scores: cp=0.1421, thal=0.1216, thalach=0.1213, oldpeak=0.1146, ca=0.0975.

The ranking produced by the Random Forest aligns closely with established cardiology practice. Chest pain classification (cp) is among the first variables assessed in suspected ischaemia. Thalassemia type (thal) captures haematological stress responses

with known cardiac implications. Peak exercise heart rate (thalach) is a widely used index of functional cardiovascular reserve. Exercise-induced ST depression (oldpeak) and fluoroscopy vessel count (ca) are routine components of stress testing protocols. The convergence between algorithmic rankings and clinical convention strengthens the case for deploying this model in screening workflows.

5.3. Discussion

Two factors jointly account for Logistic Regression's AUC-ROC advantage in this experiment. Structural analyses of the Cleveland data have indicated that the healthy and diseased populations are largely separable along linear boundaries in the feature space, an arrangement that inherently favours parametric linear models. Additionally, Logistic Regression minimises log-loss during training, which directly encourages well-calibrated posterior probabilities; Random Forest vote-fraction estimates, by contrast, compress towards the extremes and tend to be miscalibrated near 0 and 1.

Gradient Boosting's comparatively weak performance (AUC-ROC = 0.8828) most plausibly reflects sensitivity to the limited training volume (n = 237) when default regularisation is applied. Sequential boosting algorithms are known to benefit substantially from depth constraints, shrinkage tuning and sub-sampling strategies that were deliberately withheld here in the interest of a level-field comparison. The broader implication is noteworthy: when datasets are small and well-structured, investing in hyperparameter search may yield diminishing returns compared with selecting a model whose inductive bias already matches the data geometry.

6. Conclusions

This work reported a systematic comparison of three supervised classifiers — Logistic Regression, Random Forest, and Gradient Boosting — applied to binary cardiac disease screening on the Cleveland benchmark. Logistic Regression attained the best AUC-ROC (0.9554), confirming its strong rank-ordering capacity — a property directly relevant to clinical triage and risk stratification. Random Forest led on raw accuracy (85.00%) and balanced F1-Score (0.8302). Gini-based feature scoring highlighted chest pain classification, thalassemia type and peak exercise heart rate as the dominant discriminators, a ranking that coheres with standard cardiology assessment protocols.

When data are limited, well-structured and geometrically near-separable, parsimonious linear models can match or outperform more expressive ensemble alternatives. Planned extensions include leave-group-out validation on multi-centre cohorts, systematic hyperparameter search, and incorporation of SHAP-based local explanation to bridge the gap between statistical performance and clinical trust.

Funding: No external funding was received in support of this research.

Institutional Review Board Statement: Not applicable; all analyses were conducted on a de-identified, publicly archived dataset requiring no institutional ethics approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data analysed in this study are drawn from the publicly archived Cleveland Heart Disease collection, accessible at: https://archive.ics.uci.edu/ml/datasets/heart+disease

References

1. Detrano, R.; Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Schmid, J.; Sandhu, S.; Guppy, K.; Lee, S.; Froelicher, *V. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am. J. Cardiol.* **1989**, 64, 304–310.
2. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* **2019**, 7, 81542–81554.

3. Gavhane, A.; Kokkula, G.; Pandya, I.; Devadkar, K. Prediction of heart disease using machine learning. *In Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March* **2018.**
4. Shah, D.; Patel, S.; Bharti, S.K. Heart disease prediction using machine learning techniques. *SN Comput. Sci.* **2020**, 1, 345.
5. Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **2010**, 21, 128–138.
6. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res*. **2011**, 12, 2825–2830.