International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Heart Disease Risk Prediction Using Machine Learning

[1]Dr. V. Lavanya, [2]A. Veerabrhmachari,
[3]A. SatyaNarayana, [4]G. Divya Bharathi,
[5]B. Mega Syam

[1]Professor, [2,3,4,5]Student,
[1,2,3,4,5]Electronics and Communication Engineering,
[1,2,3,4,5]Maharaj Vijayaram Gajapathi Raj College of Engineering(Autonomous),
Vizianagaram,India.

Abstract - Heart disease remains one of the leading causes of mortality worldwide, accounting for approximately 17.9 million deaths annually. Traditional diagnostic methods are often costly, time-consuming, and inaccessible for routine screening. This paper proposes a Heart Disease Risk Prediction system built on a weighted soft voting ensemble of Logistic Regression (LR) and Random Forest (RF). The model is trained on a 4,000-record clinical dataset with 13 features. The system provides probability-based outputs suitable for graded clinical risk assessment.

Keywords: Heart Disease, Machine Learning, Logistic Regression, Random Forest, Ensemble Model

1. INTRODUCTION

Cardiovascular disease (CVD) is responsible for approximately 17.9 million deaths annually, representing 32% of all global deaths [1]. Early and accurate detection is critical for reducing mortality and enabling timely intervention. Traditional diagnostic approaches rely on manual clinical examinations, laboratory tests, and expert judgment—methods that can be costly, time-consuming, and subject to variability.

Machine learning (ML) has emerged as a powerful tool for analysing clinical data and identifying patterns associated with disease risk. Various algorithms including Logistic Regression [2], Random Forest [4], Support Vector Machine [3], and deep learning approaches [6] have been applied to heart disease classification. However, single-classifier approaches involve inherent trade-offs between accuracy, interpretability, and generalisation. Ensemble methods that combine multiple models have consistently outperformed individual classifiers in medical tasks.

This paper proposes a weighted soft voting hybrid of LR and RF (weights 1:3) trained on a 4,000-record dataset. Key contributions include: (1) a literature-driven algorithm selection methodology; (2) an empirically weighted ensemble that outperforms individual classifiers; (3) real-time validation with hospital ECG comparison; and (4) probability-based graded risk outputs for clinical use.

2. RELATED WORK

Yousef and Batiha [1] combined Naive Bayes feature selection with SVM, KNN, Decision Tree, and RF, achieving 98% accuracy. Gopalakrishnan et al. [2] applied CNN-based deep learning for automatic feature extraction, achieving high accuracy but requiring substantial GPU resources. Chandrasekhar and Peddakrishna [3] applied GridSearchCV-tuned soft voting over six classifiers, achieving 93.44% accuracy. Bhatt et al. [4] evaluated Decision Tree with k-modes clustering preprocessing. Biswas et al. [5] compared six classifiers under chi-square, ANOVA, and mutual information feature selection. Elsedimy et al. [6] proposed QPSO-SVM for automated hyperparameter tuning. Rimal et al. [7] confirmed LR's robustness via cross-validation. Stonier et al. [8] found RF outperformed all methods including Neural Networks on clinical data. A gap remains in combining LR and RF specifically through empirically weighted soft voting—this paper addresses that gap.

3. DATASET AND FEATURE DESCRIPTION

The dataset used in this study contains 4,000 patient records with 13 clinical input features and one binary target variable (0 = No Heart Disease, 1 = Heart Disease). The dataset contains 2,069 records labelled as heart disease (51.7%) and 1,931 labelled as healthy (48.3%), representing near-balanced class distribution that prevents classifier bias toward the majority class.

Table I summarises the 13 input features, their data types, and clinical significance.

Table - 1: Clinical Features Used for Heart Disease Prediction

No.	Feature	Type
1	**Age**	Integer
2	**Sex**	Binary
3	**Chest Pain (cp)**	Categorical
4	**Resting BP**	Integer
5	**Cholesterol**	Integer
6	**Fasting BS**	Binary
7	**Resting ECG**	Categorical
8	**Max HR**	Integer
9	**Exang**	Binary
10	**ST Depression**	Float
11	**Slope**	Categorical
12	**Major Vessels**	Integer
13	**Thal**	Categorical

4. PROPOSED METHODOLOGY

A. System Architecture:

The system follows six stages: (1) data loading and preprocessing; (2) target distribution verification; (3) feature-target separation; (4) stratified 80/20 train-test split; (5) feature scaling within the LR pipeline; and (6) hybrid soft voting ensemble training. Implementation uses Python's Scikit-learn library in Google Colab.

B. Preprocessing and Data Splitting:

The dataset is split into 3,200 training and 800 testing records using stratified splitting (stratify=Y, random_state=2) to preserve class proportions. StandardScaler normalises features for Logistic Regression ($x' = (x - \mu) / \sigma$). Random Forest requires no scaling since tree splits are invariant to feature magnitude.

C. Algorithm Selection:

Three algorithms were selected after reviewing 8 published studies [1–8]:

Logistic Regression: Interpretable, probability-calibrated, consistent baseline in medical tasks (all 8 papers).

Support Vector Machine: Non-linear decision boundaries via kernel trick; robust generalisation (6 of 8 papers).

Random Forest: Highest standalone accuracy; ensemble robustness via bagging; native probability output. Stonier et al. [8] found RF outperforms Neural Networks on clinical datasets.

CNN, KNN, Gradient Boosting, and AdaBoost were excluded due to interpretability constraints, inference cost, or no advantage over RF on this dataset.

D. Hybrid Soft Voting Ensembler:

The proposed model uses a VotingClassifier (voting='soft') that averages class probabilities. The weight assignment LR:1, RF:3 reflects the 11.9 percentage-point standalone accuracy gap (RF = 99.20% vs. LR = 87.30%). The final probability is:

$$P_{final} = (1 \times P_{LR} + 3 \times P_{RF}) / (1 + 3)$$

If P_final(disease) ≥ 0.5, the patient is classified as heart disease. LR produces P_LR via the sigmoid function; RF averages probabilities across 100 decision trees to produce P_RF.

4. EXPERMENTAL RESULTS AND ANALYSIS

A. Performance Comparision:

Table II presents accuracy, precision, and recall of all models evaluated on the 800-record test set.

Table - 2: Perfoemance Comaparion on ML Models

Model	Accuracy	Precision	Recall
Logistic Regression	75.12%	73.63%	80.92%
Support Vector Machine	94.38%	94.03%	95.17%
Random Forest	99.38%	99.52%	99.28%
LR + SVM	92.75%	91.59%	94.69%
LR + RF (Proposed)	**98.75%**	**98.10%**	**99.52%**
LR + SVM + RF	98.75%	98.56%	99.03%

The proposed LR+RF model achieves 98.75% accuracy—23.65 points above LR alone and 6.0 points above LR+SVM. Although RF alone reaches 99.38%, the hybrid is preferred clinically for combining RF's predictive power with LR's probability calibration. The 99.52% recall is especially critical: it minimises false negatives, where a diseased patient is incorrectly labelled healthy—the costliest error in cardiovascular screening. Adding SVM to the ensemble (LR+SVM+RF) yields no improvement over LR+RF, confirming the optimal two-classifier design.

B. Training vs Test Accuracy Analysis:

The hybrid model achieved 100.00% training accuracy and 98.75% test accuracy—a gap of just 1.25 percentage points—indicating a well-fitted model with minimal overfitting.

5. REAL-TIME PATIENT VALIDATION

The trained hybrid model was validated using clinical data from four actual patients to assess its practical applicability. Table III presents the clinical features and model outputs for each patient case.

TABLE – 3: Real-Time Patient Validation Results

Parameter	Patient 1	Patient 2	Patient 3	Patient 4
Age	45	49	61	69
Sex	F	M	F	F
Chest Pain	Atypical	Non-anginal	Asymptomatic	Non-anginal
Resting BP (mmHg)	110	140	100	150
Cholesterol (mg/dl)	195	177	233	286
Fasting BS (mg/dl)	101	99	183 (high)	114
ECG	Normal	Normal	Normal	LVH (2)
Max HR (bpm)	177	173	61 (critical)	98

Exang	No	No	No	Yes
ST Depression	0.8	1.2	0.0	1.5
Slope	Flat	Flat	Flat	Downsloping
Major Vessels	0	0	0	2
Blood Disorder	Normal	Normal	Normal	Normal
Actual Target	0 (Healthy)	0 (Healthy)	1 (Disease)	1 (Disease)
Healthy Prob.	69.51%	60.90%	29.23%	25.34%
Disease Prob.	30.49%	39.10%	70.77%	74.66%
Prediction	Correct ✓	Correct ✓	Correct ✓	Correct ✓

All four predictions were correct. Patient 1 (F, 45): low-risk profile with 30.49% disease probability, correctly classified as healthy. Patient 2 (M, 49): elevated BP and ST depression of 1.2, yet overall profile predicted healthy (39.10% disease), confirmed correct—demonstrating balanced multi-feature assessment over isolated values. Patient 3 (F, 61): critically low max HR (61 bpm) with a normal resting ECG; model predicted 70.77% disease probability, correctly identifying disease where a single ECG test would have missed it. Patient 4 (F, 69): multiple high-risk indicators yielded 74.66% disease probability, correctly classified as disease. Probability-based outputs enable graded risk-stratified clinical decisions beyond binary classification.

6. CONCLUSION

This paper presented a Heart Disease Risk Prediction system based on a weighted soft voting hybrid of Logistic Regression and Random Forest (weights 1:3). The proposed LR+RF model achieved 98.75% accuracy, 98.10% precision, and 99.52% recall on a 4,000-record dataset, outperforming all individual classifiers and hybrid combinations. Real-time validation on four patients demonstrated correct classification in all cases, including a clinically significant scenario where a normal ECG coexisted with high disease probability. The system's probability outputs make it suitable for graded early screening and clinical decision support. Future work should evaluate on external datasets (UCI, Framingham) and incorporate biomarkers, genetic risk scores, and wearable sensor data.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who supported and guided me throughout the completion of this project.

First and foremost, I sincerely thank my mentors for their valuable guidance, continuous support, and encouragement throughout this work. Their insightful suggestions and constructive feedback helped me to understand the concepts more clearly and significantly improved the quality of this project.

REFERENCES

1. M. Yousef and K. Batiha, "Heart disease prediction model using Naive Bayes algorithm and machine learning techniques," International Journal of Engineering and Technology, vol. 10, no. 1, pp. 46-56, 2021. doi: 10.14419/ijet.v10i1.31240.
2. S. Gopalakrishnan, M. S. Sheela, K. Saranya, and J. J. Hephzipah, "A novel deep learning-based heart disease prediction system using convolutional neural networks (CNN) algorithm," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 10s, pp. 516-522, 2023. doi: 10.18201/ijisae.2023special.53.
3. N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," Processes, vol. 11, no. 4, p. 1210, 2023. doi: 10.3390/pr11041210.
4. C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, no. 2, p. 88, 2023. doi: 10.3390/a16020088.
5. S. Biswas, S. Islam, Y. Kazi, S. Hasnat, S. Bhuiyan, and H. Kabir, "Machine learning-based model to predict heart disease in early stage employing different feature selection techniques," BioMed Research International, vol. 2023, Article ID 6864343, pp. 1-17, 2023. doi: 10.1155/2023/6864343.
6. E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," Multimedia Tools and Applications, vol. 83, no. 8, pp. 23901-23928, 2024. doi: 10.1007/s11042-023-16661-3.
7. Y. Rimal, N. Sharma, S. Paudel, A. Alsadoon, M. P. Koirala and S. Gill, "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy," Scientific Reports, vol. 15, no. 13444, pp. 1-14, 2025. doi: 10.1038/s41598-025-93675-1.
8. A. A. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," Healthcare Technology Letters, vol. 11, no. 4, pp. 213-217, 2024. doi: 10.1049/htl2.12053.