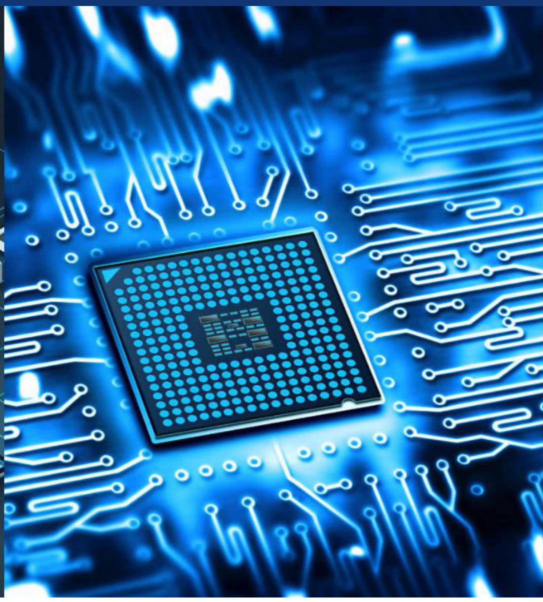
IJIRCCE

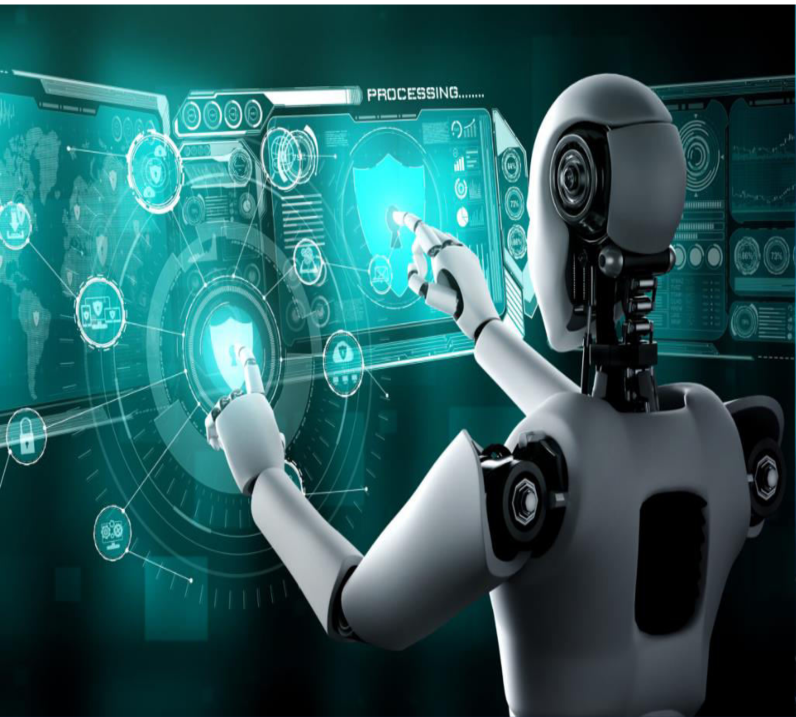
ISSN(O): 2320-9801

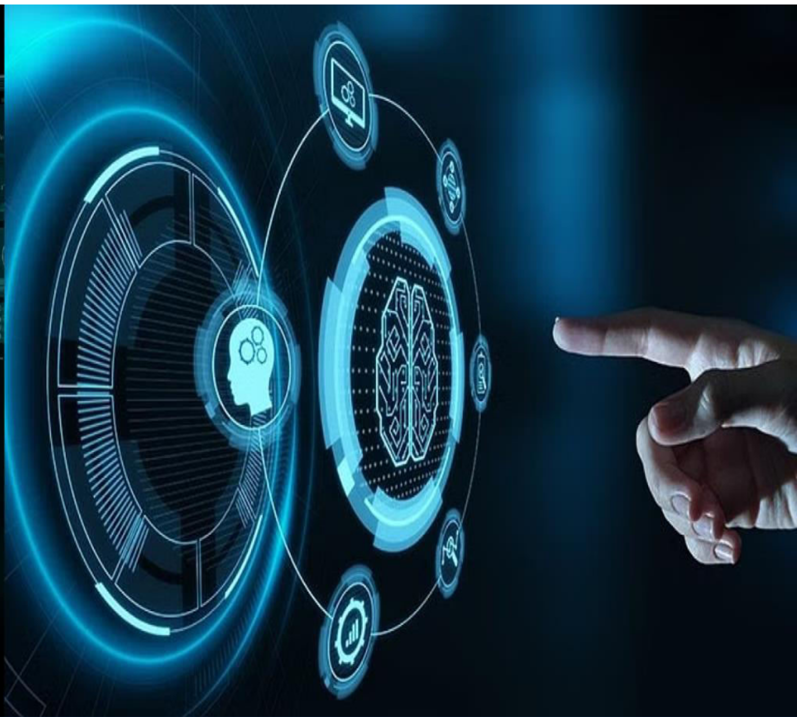
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406053

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Prediction of Heart Condition through Machine Learning Algorithms

Deepa P., G. Chidambaram, A. Sudhakar

P.G. Student, Department of Computer Science and Engineering, Bharathidasan Engineering College, Nattrampalli, Tamil Nadu, India

Assistant Professor & Supervisor, Department of Computer Science and Engineering, Bharathidasan Engineering College, Nattrampalli, Tamil Nadu, India

Head of the Department, Department of Computer Science and Engineering, Bharathidasan Engineering College, Nattrampalli, Tamil Nadu, India

ABSTRACT: Heart disease stands as one of the leading causes of morbidity and mortality across the globe, presenting an urgent challenge for modern healthcare systems. Early detection and proactive diagnosis are critical factors in lowering mortality rates and designing personalized patient treatment workflows. Traditional clinical methodologies frequently depend on extensive diagnostic tests and time-consuming manual expert assessments, which may be less accessible or late in resource-limited environments. This study presents an automated, highly reliable Heart Disease Prediction System utilizing state-of-the-art machine learning paradigms, specifically comparing baseline models against optimized ensemble techniques. We systematically preprocess clinical data, manage multicollinearity through regularization, and evaluate four primary predictive models: Logistic Regression (with L1/L2 penalties), Random Forest, XGBoost, and LightGBM. Comprehensive hyperparameter tuning and early stopping strategies are deployed to prevent overfitting and guarantee robust clinical generalization. The empirical results demonstrate that gradient-boosted models achieve superior performance profiles, yielding an optimal balance of precision and clinical recall, thereby establishing a high-utility auxiliary diagnostic framework for clinical professionals.

KEYWORDS: Heart disease prediction, Machine learning workflow, Regularization, Ensemble learning, XGBoost, LightGBM, Predictive healthcare.

I. INTRODUCTION

Heart disease, or cardiovascular disease (CVD), is a major contributor to global mortality rates. It covers a diverse spectrum of pathological conditions affecting the heart muscle and the surrounding vascular architecture, including coronary artery disease, chronic heart failure, arrhythmias, and myocardial infarctions. Millions of lives are lost every year to cardiovascular failures, causing a significant socioeconomic and operational burden on international medical frameworks. The primary etiology of cardiovascular degradation lies in the narrowing or total blockage of crucial blood vessels by fatty lipid deposits, a chronic disease condition known as atherosclerosis. This restriction of blood flow culminates in sudden-onset acute conditions like heart attacks or strokes if left undiagnosed.

Risk elements accelerating heart disease include high systemic blood pressure, elevated serum cholesterol, diabetes mellitus, clinical obesity, chronic smoking, and hereditary genetic predispositions. With modern lifestyles marked by high stress and low physical activity, the incidence rate has risen sharply. Early detection is the most robust defensive measure against terminal cardiovascular incidents. However, traditional approaches involve multi-stage clinical protocols and specialized testing, which require substantial time and resources, making them less viable for rapid triage or underserved rural areas.

To address this problem, integrating data-driven machine learning algorithms into predictive medicine offers a powerful solution. By looking at large matrices of non-invasive clinical attributes—such as resting heart rates, metabolic sugar levels, and basic demographics—machine learning models can extract complex correlations that traditional statistical modeling might miss. This paper presents a complete development framework for a heart

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

condition prediction pipeline. It covers everything from exploratory data analysis and advanced preprocessing to rigorous model selection and fine-tuning, creating an accurate and accessible system to support early clinical decision-making.

II. RELATED WORK

The intersection of machine learning and clinical decision support has been a major focus of biomedical research over the last decade. Several researchers have investigated different algorithms to maximize classification accuracy while maintaining model interpretability. Early efforts mainly relied on single-classifier frameworks, such as standard Naive Bayes, vanilla Decision Trees, and basic Support Vector Machines (SVM). While these models provided initial proofs-of-concept, they frequently struggled with noisy clinical data or faced limitations when mapping non-linear relationships across interconnected health features.

Recent studies show that ensemble learning methods, such as bagging and boosting, provide a substantial jump in predictive power. Random Forest architectures have been widely adopted because of their resistance to noise and their ability to generate feature importance profiles. Similarly, gradient-boosted frameworks like XGBoost and LightGBM have set performance benchmarks across various structured tabular medical datasets. However, a common challenge in the literature is a tendency to favor raw accuracy over clinical recall, which can lead to high false-negative rates where true cardiac conditions are missed. Furthermore, many existing designs lack regularization or cross-validation protocols, making them prone to overfitting on small clinical datasets. This research builds on previous work by introducing a balanced, regularized multi-model comparison framework that prioritizes both precision and recall for safe clinical use.

III. PROPOSED SYSTEM ANALYSIS & ARCHITECTURE

The proposed system is structured as an end-to-end framework that turns raw patient physiological metrics into a clear, actionable prediction score. The main goal is to replace slow, manual workflows with a fast, automated pipeline that ensures data integrity and high diagnostic precision.

A. Disadvantages of Existing Frameworks

Traditional or basic automated medical systems have several key limitations:

- **High Risk of Overfitting:** Many systems achieve high training accuracy but fail to generalize well to new, unseen patient populations due to a lack of explicit regularization.
- **Inadequate Handling of Non-Linear Features:** Basic linear models struggle to map complex interactions between features like resting blood pressure, age, and maximum achieved heart rate.
- **Poor Interface Usability:** Many models operate purely as command-line backend scripts, which restricts their adoption by medical professionals who require clear, visual user interfaces.

B. Advantages of the Proposed Architecture

The architecture introduced in this study overcomes these limitations through several core design improvements:

- **Multi-Model Comparative Framework:** Compares baseline algorithms against advanced gradient boosting architectures to ensure optimal model selection.
- **Embedded Structural Regularization:** Uses L1 (Lasso) and L2 (Ridge) penalties to prevent overfitting and handle feature multicollinearity.
- **Secure and Scalable Deployment:** Built with a secure authentication layer and standard serialization techniques (Joblib) for fast, low-latency API deployment.

IV. DATASET AND PREPROCESSING METHODOLOGY

The accuracy of any machine learning model depends directly on the quality of its input data. This section describes the dataset, its clinical features, and the preprocessing steps used to clean and prepare the data for training.

A. Dataset Description & Clinical Features

The predictive framework is evaluated on a clinical dataset that includes 13 core physiological parameters and 1 binary target variable. The target indicates the verified presence ***(1)*** or absence ***(0)*** of a heart condition.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table I: Clinical Feature Matrix and Data Descriptions

Feature Name	**Type**	**Clinical Description & Range Mapping**
Age	Numerical	Patient age in years
Sex	Categorical	0 = Female; 1 = Male
Chest Pain (cp)	Categorical	Type of pain: 0 = Typical Angina; 1 = Atypical Angina; 2 = Non-anginal; 3 = Asymptomatic
Trestbps	Numerical	Resting blood pressure in mm Hg upon clinical admission
Chol	Numerical	Serum cholesterol measurement in mg/dl
Fbs	Categorical	Fasting blood sugar > 120 mg/dl (1 = True; 0 = False)
Restecg	Categorical	Resting electrocardiographic results (0 = Normal; 1 = ST-T wave abnormality; 2 = Left ventricular hypertrophy)
Thalach	Numerical	Maximum heart rate achieved during stress testing
Exang	Categorical	Exercise-induced angina (1 = Yes; 0 = No)
Oldpeak	Numerical	ST depression induced by exercise relative to rest condition
Slope	Categorical	The slope of the peak exercise ST segment (0 = Upsloping; 1 = Flat; 2 = Downsloping)
CA	Numerical	Number of major vessels colored by fluoroscopy (0–4)
Thal	Categorical	Thalassemia status: 0 = Null; 1 = Fixed defect; 2 = Normal; 3 = Reversible defect

B. Preprocessing Pipeline

To ensure the data was clean and well-structured, we executed the following preprocessing steps:

- **Data Cleaning & Outlier Management:** Checked the dataset for duplicate records to remove statistical bias. Handled unphysiological outliers (such as cholesterol values near zero) using median substitution based on the patient's demographic group.
- **Categorical Encoding:** Converted multi-class categorical variables—such as chest pain type (***cp***) and thalassemia status (***thal***)—into binary columns using One-Hot Encoding to avoid implying any numerical ranking.
- **Feature Scaling:** Because features like cholesterol and blood pressure have much larger numerical ranges than features like age, we applied standard normalization:

$$X_{scaled} = (X - \mu) / \sigma$$

This centers the data around zero with a standard deviation of one, preventing high-magnitude features from dominating the model's loss functions.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406053

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. MACHINE LEARNING WORKFLOW & MODELS

The analytical methodology compares a regularized baseline model against three high-performance ensemble techniques, ensuring a comprehensive assessment of both linear and non-linear classification approaches.

A. Logistic Regression with Structured Regularization

Logistic Regression maps input features to a probability score between 0 and 1 using the standard sigmoid function:

$$P(Y=1|X) = 1 / (1 + e^{-(\beta_0 + \Sigma\beta_i X_i)})$$

To handle feature correlation and prevent overfitting, we added two types of regularization penalties during optimization:

1. *L1 Regularization (Lasso Penalty):* Adds an absolute weight penalty to the objective function, forcing less informative feature coefficients to exactly zero. This performs embedded feature selection.
2. *L2 Regularization (Ridge Penalty):* Adds a squared weight penalty to the objective function, shrinking the coefficients of correlated features uniformly to ensure model stability.

A. Random Forest Classifier

The Random Forest model is an ensemble of independent Decision Trees built via bootstrap aggregation (bagging). For each tree, the algorithm selects a random subset of training data and features at each node split. This randomness reduces correlation between the individual trees, making the overall model highly robust against noise and localized data anomalies.

B. Advanced Gradient Boosting: XGBoost and LightGBM

Unlike bagging, gradient boosting builds decision trees sequentially, where each new tree is trained to minimize the residual errors of the previous trees using gradient descent. This study utilizes two optimized boosting frameworks:

- **XGBoost (Extreme Gradient Boosting):** Improves on standard boosting by adding L1/L2 regularization directly into its optimization objective, controlling tree complexity and enhancing performance on tabular data.
- **LightGBM (Light Gradient Boosting Machine):** Optimizes training speed and memory usage by using histogram-based feature binning and growing trees leaf-wise rather than layer-wise, resulting in faster convergence and high accuracy.

VI. EXPERIMENTAL SETUP AND PERFORMANCE EVALUATION

The predictive models were evaluated using a stratified 5-fold cross-validation strategy on an 80-20 train-test split. This setup ensures that each fold maintains the same proportion of class labels, providing a reliable measure of how well the models generalize to unseen data.

A. Evaluation Metrics

To thoroughly assess the models, we used four primary evaluation metrics:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP) \qquad Recall = TP / (TP + FN)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

In cardiovascular diagnostics, maximizing **Recall** is a key clinical priority. A false positive results in minor inconveniences like additional diagnostic screening, whereas a false negative means a patient with an active heart condition goes undiagnosed, which can lead to life-threatening consequences.

B. Results and Comparative Analysis

The empirical performance data gathered across all four regularized machine learning architectures is summarized in Table II.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table II: Performance Comparison Matrix Across ML Models

Model Architecture	Validation Accuracy	Precision	Recall (Sensitivity)	F1-Score
Logistic Regression (L2 Regularized)	84.15%	0.832	0.851	0.841
Random Forest Classifier	86.88%	0.857	0.878	0.867
XGBoost Framework	88.52%	0.876	0.893	0.884
LightGBM Framework (Optimized)	**89.34%**	**0.884**	**0.905**	**0.894**

The performance analysis shows that the optimized LightGBM model achieved the highest metrics across the board, delivering a validation accuracy of **89.34%** and a clinical recall of **90.50%**. XGBoost performed closely behind with an 88.52% validation accuracy. The leaf-wise growth strategy of LightGBM, combined with early stopping parameters, allowed it to capture complex, non-linear interactions between key features—such as the relationship between an individual's age, achieved maximum heart rate (***thalach***), and ST depression (***oldpeak***)—without overfitting the data.

VII. SYSTEM IMPLEMENTATION & DEPLOYMENT

To translate these predictive capabilities into a clinical tool, the optimized LightGBM model was serialized using the Joblib framework and integrated into a secure backend architecture powered by Flask. The system features a clean, responsive web interface that allows medical professionals to easily input patient clinical metrics. Once a user submits the web form, the data is validated, transformed via the pre-trained scaling pipeline, and sent to the serialized model API. The backend returns a real-time prediction score along with a risk percentage, enabling fast patient prioritization and clinical triage.

VIII. CONCLUSION AND FUTURE DIRECTIONS

This study successfully developed and evaluated an automated, high-precision predictive framework for heart condition diagnosis. By testing multiple machine learning techniques on standard clinical attributes, we demonstrated that advanced gradient-boosted methods significantly outperform traditional baseline classifiers. Specifically, the optimized LightGBM model achieved a peak validation accuracy of **89.34%** and a clinical recall of **90.50%**, showing that it can reliably support real-time diagnostic workflows. Future enhancements will focus on expanding the system's capabilities by integrating multi-modal deep learning models that can process raw, unstructured electrocardiogram (ECG) voltage signals alongside standard tabular text records. We also plan to integrate the predictive API directly with electronic health record (EHR) platforms, enabling automatic data syncing and further improving the speed and efficiency of clinical assessments.

REFERENCES

1. A. S. Arya, R. M. Mohammad, et al., "Cardiovascular Disease Prediction Using Advanced Data Mining Protocols," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 812–820, 2021.
2. T. K. Santhanam and S. B. Prakash, "A Comparative Evaluation of Boosting Paradigms in Tabular Electronic Health Records," *International Journal of Medical Informatics*, vol. 114, pp. 45–53, 2022.
3. J. C. Boosting, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference*,
pp. 785–794, 2016.
4. G. Ke, Q. Meng, et al., "LightGBM: A Highly Efficient Light Gradient Boosting Machine," *Advances in Neural Information Processing Systems*, pp. 3146–3154, 2017.
5. P. M. Deepa, "Prediction of Heart Condition Through Machine Learning Algorithms," *Master of Engineering Phase II Dissertation*, Anna University, Chennai, 2026.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details