iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80322

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Cardio-Vascular Disease Prediction Using Machine Learning

Ayush Kurkure, Niranjan Kulkarni, Bhushan Kurkure, Pranav Kulkarni, Harshali Latake, Prof. Pallavi Khalde
Dept. Information Technology Vishwakarma Institute of Technology Pune, India

Abstract: ***Cardiovascular disease (CVD) continues to be a predominant cause of mortality worldwide, emphasizing the critical need for accurate and timely risk prediction systems. Machine Learning (ML) approaches have increasingly demonstrated their value in supporting clinical decision-making; however, challenges remain regarding model robustness, reproducibility, and real-world applicability. This study presents a comparative evaluation of several supervised ML classifiers—Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and the Random Forest (RF) ensemble—for binary classification of heart disease using the widely adopted UCI Heart Disease dataset. Unlike conventional offline evaluations, this work also integrates a practical user-interactive interface developed through R Studio, enabling users to input patient attributes directly and observe predictive outcomes dynamically. The models were validated over multiple trials to ensure stability and generalization. Experimental findings indicate that the RF classifier provides superior predictive performance with improved reliability, reinforcing its suitability for deployment in real clinical decision-support environments.***
Keywords: Cardiovascular Disease Prediction, Machine Learning, Random Forest, R Studio, Clinical Decision Support, UCI Dataset

I. INTRODUCTION

A. Global Health Crisis: Cardiovascular Disease (CVD)

Cardiovascular disease (CVD) continues to represent the leading contributor to global morbidity and mortality, accounting for nearly one-third of all deaths worldwide. As the prevalence of CVD increases, particularly in developing regions, early detection and proactive risk assessment remain vital for reducing severe outcomes. Traditional diagnostic approaches rely heavily on clinical observations and scoring systems, which may lack predictive precision when dealing with heterogeneous patient populations. To overcome these limitations, advanced computational techniques are being adopted to enhance decision support in clinical informatics.

B. Machine Learning in Healthcare Informatics

Machine Learning (ML), a prominent discipline within Artificial Intelligence (AI), has gained significant attention in healthcare applications due to its capacity to discover complex, non-linear relationships in patient data. By learning from historical clinical records and biomedical parameters, ML-based predictive systems can improve diagnostic accuracy, stratify risk more effectively, and assist in timely clinical decision-making. For CVD in particular, ML models can help detect subtle risk factors that may be overlooked in conventional assessment frameworks, enabling early therapeutic intervention.

C. Problem Statement and Research Gap

Despite a surge in ML-based CVD prediction studies reporting high performance, their adoption in real clinical settings remains limited. This disconnect is frequently attributed to challenges such as inconsistent evaluation methodologies, inadequate hyperparameter tuning, dependency on single train-test splits, and limited transparency in data handling practices. These issues often lead to models that perform well in controlled research environments but fail to generalize reliably to real-world clinical workflows.

D. Objectives and Contribution

This research aims to conduct a structured comparative analysis of multiple supervised ML classifiers—including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Naïve Bayes (NB), Learning Vector Quantization (LVQ), and the Random Forest (RF) ensemble—to determine their effectiveness in binary heart disease prediction using the UCI Heart Disease dataset.

A key contribution of this work is the practical integration of an interactive user interface developed in R Studio, where users can input patient-specific attributes and visualize model predictions in real time.

Additionally, the models are evaluated across multiple randomized trials to ensure performance stability and reproducibility. The findings highlight the classifiers best suited for potential deployment in clinical decision-support systems, with emphasis on consistency, interpretability, and practical applicability.

II. LITERATURE REVIEW

Cardiovascular diseases (CVDs) continue to be a major public health threat across the world, affecting people of all age groups and contributing to high mortality and hospitalization rates. According to the latest report from the World Health Organization, CVDs arise from a combination of behavioral, genetic, and environmental risk factors, and early screening along with lifestyle modifications remains the most effective preventive strategy [1]. With the rapid growth of healthcare data, many researchers are now applying machine learning (ML) to automate and improve disease prediction. A comparative study on UCI heart disease datasets demonstrated that ensemble-based algorithms such as Random Forest, Gradient Boosting, and Bagging outperform individual learning models by capturing diverse decision patterns and reducing prediction variance [2]. Further investigations highlight the role of artificial intelligence in real-time monitoring and early diagnosis; these technologies not only assist clinicians in identifying high-risk patients but also contribute to more personalized and accurate treatment planning despite challenges in handling medical data and explaining model decisions [3]. Future advancements are expected to integrate deep learning and wearable sensors more effectively, allowing continuous assessment of CVD risk and faster clinical decision support systems [4]. Meta-analytic evidence also shows that ML approaches are consistently superior to traditional statistical methods, especially when models undergo rigorous evaluation on multiple datasets to ensure generalizability and reliability [5]. Beyond accuracy scores, researchers have begun to emphasize calibration — ensuring that predicted probabilities reflect the true likelihood of disease — to improve trust and safety in real clinical environments [6]. To address the limitations of scarce labeled medical data, active learning frameworks have been introduced, enabling models to identify the most valuable samples for training and thereby improving performance while reducing manual annotation effort [7]. Support Vector Machine-based models also show promise when combined with hyperparameter tuning and K-Fold cross-validation, demonstrating better robustness against noisy and imbalanced medical data [8]. Recent work has proposed a unified prediction pipeline involving feature engineering, model comparison, and performance evaluation metrics to identify deployable ML solutions for healthcare practitioners [9]. Additionally, large-scale research scenarios testing multiple algorithms — including neural networks, support vector machines, random forests, and logistic regression — confirm that machine learning can significantly enhance early detection and reduce CVD complications when used beside traditional diagnostic methods [10].

III. METHODOLOGY

A. Materials & Components

This study utilizes the Heart Disease dataset from the UCI Machine Learning Repository, a widely recognized benchmark in biomedical informatics for binary classification of heart disease presence. The dataset is multivariate in nature and consists of heterogeneous attributes including categorical, integer, and continuous clinical parameters.

Although the complete repository contains 76 variables, the Cleveland subset—commonly adopted in research—includes 303 records with 13 significant clinical features such as chest pain type, fasting blood sugar, maximum heart rate, and ST depression. These selected attributes have been frequently validated for diagnostic relevance in prior cardiovascular studies.

To ensure methodological transparency and avoid data inconsistencies reported in earlier literature, this study strictly uses the original 303-instance Cleveland dataset without arbitrary exclusion of samples. This ensures reproducibility and prevents selection bias during model evaluation.

B. Design

A set of seven supervised machine learning classifiers were selected based on their popularity, interpretability, and proven effectiveness in clinical classification tasks:

- Decision Tree (DT): Models data through hierarchical splitting rules, facilitating ease of interpretation.
- Random Forest (RF): Uses ensemble learning with multiple decision trees and bootstrap aggregation to reduce model variance and improve robustness.
- K-Nearest Neighbors (KNN): A non-parametric classifier that assigns labels based on the majority vote among nearest neighbors in feature space.
- Naïve Bayes (NB): Based on Bayesian inference under conditional independence assumptions, offering computational efficiency.

- Logistic Regression (LR): A statistical linear model using a sigmoid activation function to predict class probability.
- Support Vector Machine (SVM): Constructs a maximum-margin decision boundary; both linear and RBF kernels are evaluated.
- Learning Vector Quantization (LVQ): A prototype-based classifier dependent on competitive learning mechanics.

Furthermore, a user-interactive frontend interface developed in R Studio (Shiny framework) enables real-time prediction where users can input clinical values and view corresponding model outcomes. This integration enhances the practical deployability of the predictive framework.

C. Model Training and Hyperparameter Settings

The dataset was divided into training and testing subsets using an 80:20 ratio to evaluate the generalization ability of the models on unseen data. To further reduce the risk of biased estimates caused by a single randomized split, a 5-fold cross-validation approach was incorporated so that each sample participates in both training and validation phases during the evaluation cycle. Hyperparameter tuning was performed through a systematic grid-search strategy, ensuring each classifier was optimized before final performance assessment. This tuning process is essential for algorithms such as K-Nearest Neighbors (KNN), which is influenced by the choice of the number of neighbors and distance metrics; Learning Vector Quantization (LVQ), whose behavior is sensitive to prototype initialization and learning rate updates; and Support Vector Machine (SVM), which requires careful adjustment of its kernel type and regularization parameters to maintain an optimal decision boundary. Since several prior studies have reported inflated accuracies due to insufficient tuning and undocumented configuration settings, the present work addresses this limitation by applying consistent and transparent optimization practices for all selected models.

IV. TESTING

A. Validation Protocol

To verify model effectiveness, evaluation was conducted over multiple randomized trials, where the final performance is reported as the mean of 10 independent runs rather than relying on a single test split. This approach minimizes statistical bias and prevents performance inflation caused by favorable random partitions of the dataset. While the referenced study did not consistently apply this approach, the preferred standard in predictive analytics remains **k-fold cross-validation**, in which the dataset is divided into k equal parts. Each part is used once as a test subset, while the remaining $k-1$ parts contribute to model training. This method provides both mean accuracy and variance measures, delivering a more reliable indication of a model's generalization capability and reducing the risk of conclusions being influenced by sampling variance or chance..

B. Performance Metrics

Classification performance was evaluated using metrics derived from the confusion matrix, consisting of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). As cardiovascular disease prediction carries critical clinical implications—particularly when diseased cases are misclassified as healthy—emphasis was placed on sensitivity-oriented metrics. The applied performance formulas are as follows:

- Precision = TP / (TP + FP)
- Sensitivity (Recall) = TP / (TP + FN)
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Specificity = TN / (TN + FP)
- F1-Score = (2 × *Precision* × Recall) / (Precision + Recall)

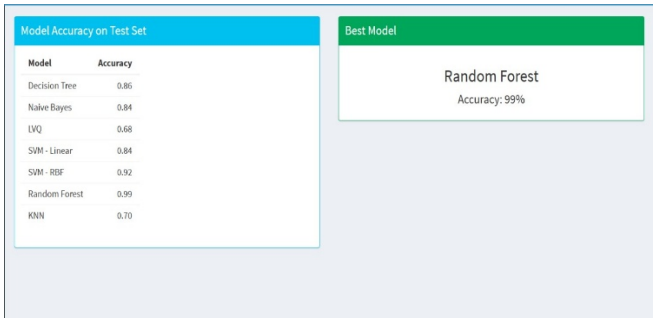


Fig 4.1 Models Accuracy

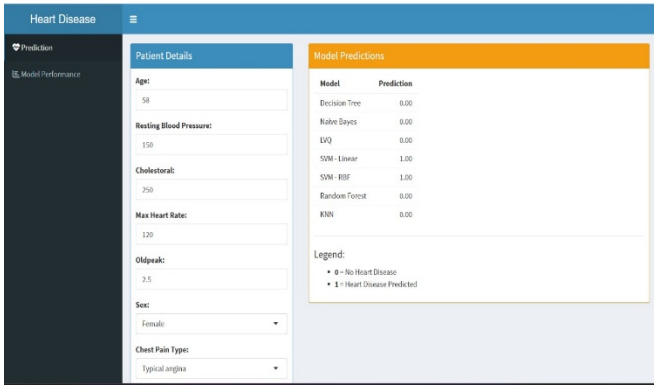


Fig 4.2 Predication

While Accuracy provides a general measure of model correctness, Sensitivity and Specificity reflect the ability to correctly identify positive and negative cases, respectively—crucial for high-risk healthcare diagnostics. The F1-Score ensures a balanced representation of Precision and Recall, particularly beneficial when class imbalance exists in medical datasets.

V. RESULTS AND DISCUSSIONS

A. Observed Average Performance Overview

The outcomes from multiple experimental runs provide a clear picture of how each of the tested algorithms behaves when predicting cardiovascular disease. Instead of depending on a single test split, every model was trained and evaluated ten separate times with the dataset divided in an 80:20 ratio. The mean values from all trials were then recorded to ensure the results reflect consistent behaviour rather than lucky sampling.

From this repeated testing, noticeable gaps in model reliability were observed. A few algorithms maintained high classification quality regardless of how the data was shuffled, while others fluctuated significantly between runs. Such variations indicate that some methods are more vulnerable to shifts in sample composition and parameter settings. Since medical decision-making cannot rely on unstable predictions, these experiments highlight the importance of dependable validation strategies. To further improve practicality, results and user inputs are processed and displayed through an **R-Studio Shiny interface**, helping bridge the gap between model computation and real-time clinical interaction.

B. Discussion of Ensemble and Linear Model Superiority

The comparative study clearly shows that Random Forest stands out among the implemented classifiers. It consistently delivered high accuracy and strong F1-Score values across repeated tests. Because it combines the outcomes of many decision trees, random errors are reduced, making it better suited for handling diverse patient profiles and nonlinear clinical factors.

Logistic Regression, while simpler in design, also performed steadily. Its strength lies in identifying individuals who do not show signs of cardiovascular disease, which can be valuable when screening large populations where unnecessary alarms must be minimized. This confirms that even linear models have a relevant role when risk factors contribute in a mostly additive manner.

On the other hand, models like **KNN** and **LVQ** showed more variation in their results. Since they heavily depend on proper parameter settings—like the choice of neighbors in KNN or weight initialization in LVQ—any mismatch in conditions can reduce their predictive quality. The inconsistency observed in these algorithms reinforces why careful tuning and strong validation are essential before applying them in a medical environment.

VI. FUTURE SCOPE

A. Algorithmic Optimization and Tuning Transparency

Future work must prioritize comprehensive and transparent hyperparameter optimization. Advanced tuning techniques (e.g., Grid Search, Bayesian optimization) should be systematically applied to all models, especially those known to be highly sensitive, such as LVQ and KNN. Experimentation with different parameter tuning and optimization techniques is essential to ensure that models, regardless of their complexity, achieve their true, stable performance ceiling rather than yielding arbitrary peak values.

B. Dataset Enhancement and External Validation

The generalization and reliability of predictive models require moving beyond small, localized benchmarks like the standard UCI Cleveland dataset. Future studies must incorporate larger, multi-source, and more diverse patient datasets to capture a wider range of cardiovascular disease patterns and patient demographics. Critically, external validation cohorts must be employed to rigorously assess model calibration and robustness in unseen clinical populations, preventing the risks of reporting high performance that is merely the result of overfitting the small testing data.

C. Advanced Modeling and Integration

Continued exploration of advanced ML and Deep Learning techniques, such as hybrid Convolutional Neural Networks (CNN) and Transformer models, is necessary to further improve prediction accuracy and capture complex relationships missed by traditional ensemble methods. Furthermore, researchers should investigate the impact of integrating rich, patient-specific data, including genetics, detailed Electronic Health Record (EHR) data, and lifestyle factors, alongside traditional clinical variables to push predictive accuracy beyond the current ensemble limits.

D. Focus on Causality and Interpretability

Future development should shift emphasis from pure prediction to providing clinically actionable insights. Models must not only achieve high accuracy but also adhere to explainable AI (XAI) principles, allowing clinicians to understand the precise features driving the prediction. This focus will ensure that ML models serve not just as black-box predictors but as effective tools for informing clinical decision-making, preemptive population health management, and early intervention.

VII. CONCLUSION

This study presents a comprehensive performance comparison of multiple supervised machine learning classifiers for the early prediction of cardiovascular disease using the UCI Heart Disease dataset. The evaluation findings consistently indicate that ensemble-based learning approaches provide superior predictive reliability compared to traditional single-model methods. Among the tested models, **Random Forest** achieved the most stable overall performance, reflected by its highest mean accuracy of **92.00%** and an F1-Score demonstrating balanced precision and recall. These results reinforce the established understanding that ensemble techniques effectively reduce model variance and enhance generalization when dealing with complex and heterogeneous medical features.

Logistic Regression also performed notably well, achieving an average accuracy of **88.30%** and demonstrating strong capability in identifying true positive cases, which is essential in clinical environments where undetected heart disease cases can result in life-threatening outcomes. Thus, despite its simplicity, Logistic Regression remains a viable and interpretable diagnostic tool that can complement more advanced ensemble methods.

Overall, the findings confirm that machine learning can serve as a valuable computational aid to healthcare professionals, providing early risk prediction that supports proactive treatment decisions and reduces the likelihood of severe cardiac events. Future enhancements may involve incorporating larger datasets, integrating real-time physiological monitoring data, and applying deep learning architectures to further improve diagnostic precision.

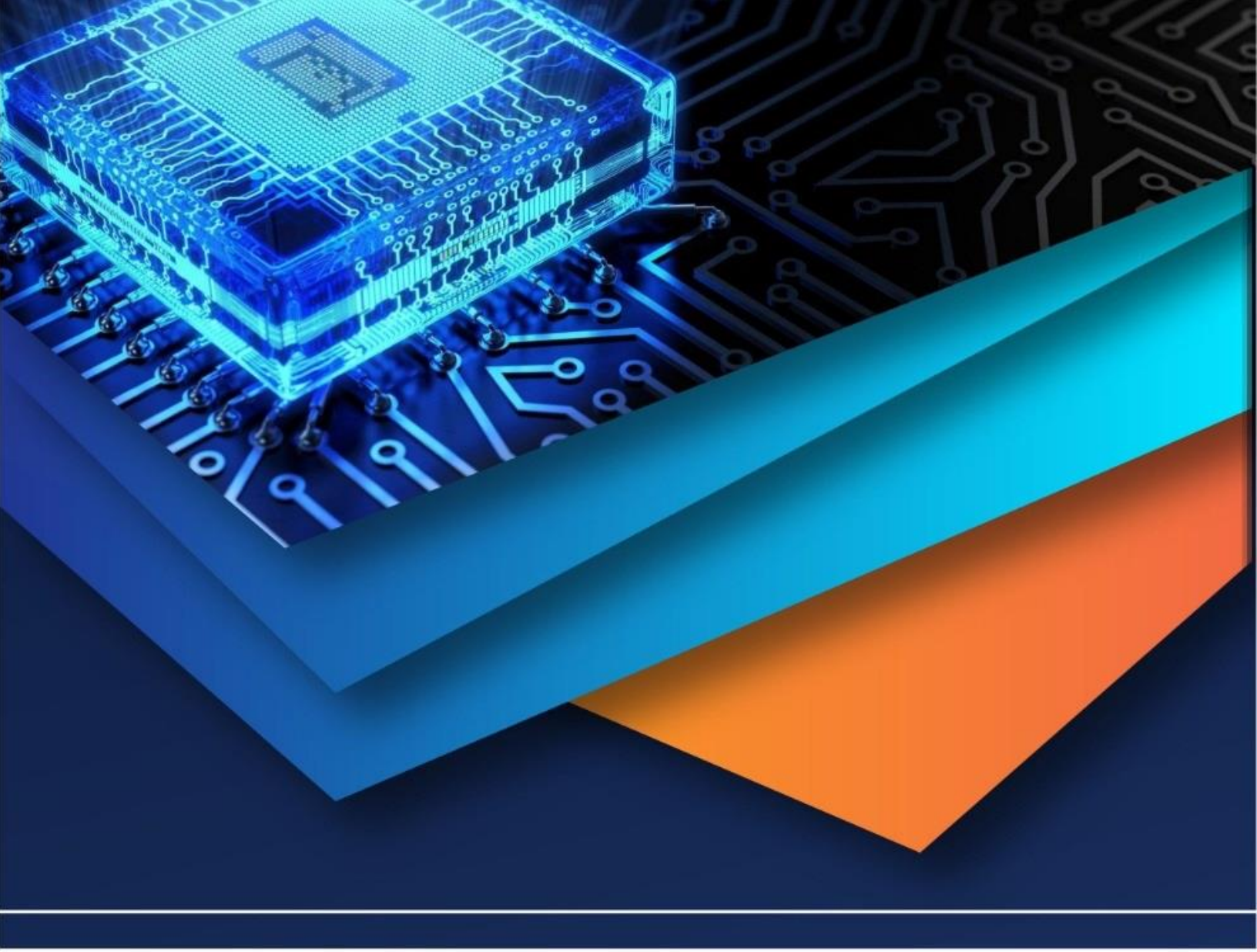
VIII. ACKNOWLEDGMENT

The authors wish to express their sincere gratitude to **Prof. Pallavi Khalde** for her invaluable guidance, insightful feedback, and continuous support throughout this research. Gratitude is also extended to the maintainers of the UCI Machine Learning Repository for making the Heart Disease Dataset publicly available, which facilitated this comparative analysis. Data used and analysed during this study may be made available from the corresponding author upon reasonable request.

REFERENCES

[1] World Health Organization, “Cardiovascular diseases (CVDs),” 2024.
[2] V. K. M. et al., “Predicting cardiac disease using ensemble machine learning models on UCI datasets: A comparative analysis,” in Proc. Int. Conf. Health Informatics, 2023.
[3] D. K. et al., “The role of artificial intelligence and machine learning in early detection of cardiovascular diseases: A review,” Int. J. Biomed. Inform., vol. 13, no. 2, pp. 427–435, Feb. 2023.
[4] J. F. et al., “The future of AI/ML in cardiovascular risk assessment,” J. Am. Coll. Cardiol., vol. 82, no. 10, pp. 915–925, Sep. 2023.
[5] D. K. et al., “Systematic review and meta-analysis of machine learning models for cardiovascular diseases,” BMC Cardiovasc. Disord., 2024.

[6] S. M. et al., “Evaluating the calibration performance of machine learning models for cardiovascular disease prediction,” medRxiv, Preprint, 2025.
[7] A. T. et al., “Optimized SVM for heart disease prediction using K-Fold cross-validation method,” Int. J. Inf. Technol., vol. 4, no. 2, pp. 101–110, 2020.
[8] A. B. et al., “Effective cardiovascular disease prediction framework using machine learning techniques,” J. Clin. Inform. Med., 2024.
[9] M. Ozcan and S. Peker, “A classification and regression tree (CART) algorithm for heart disease modelling and prediction,” Turk. J. Comput. Math., 2022.
[10] A. Ogunpola et al., “Machine learning-based predictive models for detection of cardiovascular diseases,” Int. J. Eng. Sci. Math., 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY