ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25sep429

Machine Learning Models for Predicting Heart Disease from Patient Data

Ezzelddin Shoary[1]

Publication Date: 2025/10/09

Abstract: Heart disease is among the foremost causes of mortality and morbidity worldwide, claiming an estimated 18 million lives annually. With the growing volume of healthcare data generated from clinical examinations, laboratory reports, and electronic health records, machine learning (ML) has emerged as a transformative approach for early disease prediction and risk stratification. This research investigates six supervised ML algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN)—applied to the Cleveland Heart Disease dataset. A comprehensive pipeline encompassing data preprocessing, model optimization, and cross-validation was implemented. Performance was measured using multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Results indicate that ensemble and deep learning approaches substantially outperform linear models, with XGBoost achieving the highest overall predictive power (accuracy = 90.2%, ROC-AUC = 0.94). Beyond raw performance, the study emphasizes the ethical imperatives of interpretability, fairness, and clinical trust in deploying ML systems in healthcare. Findings support the integration of ML-based tools into clinical practice for early cardiovascular diagnosis and patient-specific risk management.

Keywords: *Heart Disease Prediction, Machine Learning, Artificial Intelligence, Healthcare Analytics, Cardiovascular Diagnosis, Explainable AI.*

How to Cite: Ezzelddin Shoary (2025). Machine Learning Models for Predicting Heart Disease from Patient Data. *International Journal of Innovative Science and Research Technology*,10(9), 2770-2772. https://doi.org/10.38124/ijisrt/25sep429

I. INTRODUCTION

Cardiovascular diseases (CVDs) continue to represent the single largest contributor to global mortality. According to the World Health Organization (2024), CVDs account for roughly 32% of all global deaths each year, with coronary artery disease and myocardial infarction comprising the most prevalent subtypes. The growing burden of cardiovascular morbidity underscores the necessity of early detection and prevention. Traditional diagnostic tools, including risk scoring systems such as the Framingham Risk Score and multivariate regression models, offer valuable but limited predictive insight due to their dependence on linear assumptions and small feature sets.

Machine learning, a subfield of artificial intelligence, provides an alternative by automatically identifying complex nonlinear relationships in large datasets. In medicine, ML algorithms can process diverse data—from laboratory values and ECG signals to imaging scans—to generate accurate predictions that may surpass traditional diagnostic reasoning. In the context of cardiology, ML can identify latent patterns among variables such as age, cholesterol, chest pain type, and blood pressure, offering physicians new avenues for data-driven decision-making.

The goal of this research is to evaluate multiple ML algorithms for heart disease prediction using the Cleveland dataset and to compare their relative strengths in predictive performance, interpretability, and practical feasibility. The findings not only highlight algorithmic performance but also address the broader implications of integrating artificial intelligence into clinical workflows.

II. BACKGROUND AND LITERATURE REVIEW

Heart disease is a multifactorial condition influenced by both genetic and environmental variables. The rise of sedentary lifestyles, high-fat diets, and stress has intensified its prevalence in both developed and developing countries. Early detection can dramatically improve outcomes by allowing for timely medical intervention, yet conventional diagnostic pathways—such as angiography or echocardiography—are costly, invasive, and time-consuming. Consequently, computational models capable of predicting heart disease risk based on routine clinical data hold significant promise for preventive medicine.

For decades, cardiologists relied on regression-based models, where the probability of disease was estimated as a weighted combination of independent risk factors. While interpretable, these models assume linearity and independence among predictors, which are rarely valid in complex biological systems. The advent of ML has shifted focus toward nonparametric approaches that learn directly from data without predefining relationships. Ensemble and deep learning models, in particular, have demonstrated the

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25sep429

ability to capture nonlinear interactions and hierarchical structures among features, substantially improving predictive accuracy (Zhang, Li, & Zhou, 2022).

Recent studies have leveraged ML to enhance cardiovascular diagnostics. Logistic regression remains the baseline for interpretability, while tree-based models such as Random Forests and Gradient Boosting Machines deliver higher predictive capability (Chen & Guestrin, 2016). Neural networks have also shown success in cardiac imaging and ECG analysis due to their adaptability to high-dimensional data. Despite progress, challenges persist—namely overfitting in small datasets, lack of interpretability, and demographic bias (Doshi-Velez & Kim, 2017). These gaps justify the present study's comparative and ethical focus.

III. METHODOLOGY

The Cleveland Heart Disease Dataset, sourced from the UCI Machine Learning Repository, remains one of the most widely used datasets for cardiovascular risk prediction. It contains 303 records and 14 attributes, including both numerical and categorical features. The dependent variable denotes the presence (1) or absence (0) of heart disease. Key variables include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, slope, ca, and thal. Data preprocessing included handling missing values, normalization, encoding categorical variables, outlier removal, and 10-fold cross-validation. Six algorithms were trained: Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost, and ANN. Performance metrics included accuracy, precision, recall, F1-score, and ROC-AUC.

IV. RESULTS

The models demonstrated varying predictive power. XGBoost achieved the highest accuracy (90.2%) and ROC-AUC (0.94), followed by Random Forest and ANN. Logistic Regression maintained interpretability despite slightly lower accuracy. Ensemble methods outperformed single estimators, highlighting their capacity to manage bias-variance tradeoffs.

Table 1 Summarizes the Model Results

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	84.2	0.83	0.82	0.82	0.88
Decision Tree	78.6	0.77	0.75	0.76	0.80
Random Forest	89.1	0.88	0.87	0.87	0.93
SVM	85.7	0.84	0.83	0.83	0.89
XGBoost	90.2	0.89	0.88	0.88	0.94
ANN	88.0	0.86	0.85	0.85	0.91

V. ANALYSIS AND INTERPRETATION

The findings show that ensemble-based methods like XGBoost and Random Forest yield robust predictions, while neural networks offer flexibility for complex feature interactions. Logistic Regression remains essential for clinical interpretation. Feature importance analysis revealed that chest pain type, ST depression, and maximum heart rate were the most significant predictors. Explainable AI tools such as SHAP provide transparency into model behavior, illustrating how variables influence patient-level outcomes.

- *Ethical, Legal, and Social Implications*

Ethical considerations include potential bias from underrepresented demographic groups, patient privacy under HIPAA and GDPR, and model transparency for physician trust. Federated learning and differential privacy can mitigate these challenges, enabling secure, distributed model training without exposing sensitive data.

VI. LIMITATIONS

The study's limitations include small dataset size (n=303), lack of temporal or imaging data, and absence of external validation. Consequently, real-world generalization remains limited until larger, more diverse datasets are employed.

VII. FUTURE WORK

Future research should integrate multimodal datasets (genetic, imaging, wearable data), employ longitudinal modeling with RNNs or transformers, and prioritize model interpretability and fairness. Collaboration between clinicians and data scientists will be vital for translating ML models into deployable clinical tools.

VIII. CONCLUSION

Machine learning, particularly ensemble and deep learning models, has demonstrated high potential for early detection of heart disease. While XGBoost provided the highest accuracy, logistic regression remains valuable for interpretability. The future of clinical ML lies in balancing predictive performance with ethical accountability and transparency.

REFERENCES

[1]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 785–794.
[2]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25sep429

[3]. World Health Organization. (2024). Cardiovascular diseases (CVDs): Key facts. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
[4]. Zhang, Y., Li, H., & Zhou, M. (2022). Deep learning approaches for automated cardiac diagnosis: A comprehensive review. Computer Methods and Programs in Biomedicine, 215, 106636.