International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04046
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

HEART DISEASE PREDICTION USING MACHINE LEARNING

Mohammad Mustufa Anis Vindhani [1], Amit Bhusari [2]
Department of MCA, Trinity Academy of Engineering, Pune, India
Assistant Professor Department of MCA,Trinity Academy of Engineering, Pune, India

Abstract

This project presents a robust machine learning system designed to predict heart disease using the UCI Heart Disease dataset, which comprises303 patient records and 14clinical features, including age, chest pain type, and maximum heart rate. Through comprehensive exploratory data anal-lysis (EDA), significant correlations were identified, particularly with chest pain type (correlation: 0.43) and maximum heartrate(correlation: 0.42),guiding feature prioritization. Three machine learning models Logistic Regression, K-Nearest Neighbours (KNN), and Random Forest—were de-eloped, trained, and rigorously evaluated. Logistic Regression achieved the highest test accuracy of 88.52%, with cross-validated metrics demon-strating reliability: precision (82.08%), recall (92.12%), F1-score (86.73%),and ROC-AUC (0.95).

INTRODUCTION

Cardiovascular diseases (CVDs) are the foremost global health challenge, responsible for approximately 17.9 million deaths annually, according to the World Health Organization \citep{who2024}. Early and accurate detection is crucial to mitigating the impact of CVDs, yet conventional diagnostic techniques, such as angiograms and stress tests, are often resource-intensive and inaccessible, particularly in low-resource settings \citep{shah2019}. Machine learning presents a promising solution by enabling non-invasive, data-driven prediction of heart disease risk. This research introduces a heart disease prediction model developed using the UCI Heart Disease dataset, which includes 303 patient records with 14 clinical features, such as age, chest pain type, and maximum heart rate \citep{uci_heart_disease}. The author evaluated three machine learning algorithms—Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest—achieving a test accuracy of 88.52% and a recall of 92.12% with Logistic Regression \citep{vindhani2025}. Exploratory data analysis revealed chest pain type and maximum heart rate as critical predictors, enhancing the model's clinical interpretability. Validated through 5-fold cross-validation, the model's modular pipeline supports applications in clinical diagnostics, telemedicine, and public health screening. By addressing limitations in prior studies, including insufficient feature analysis and validation, this work advances the development of reliable, scalable tools for early heart disease detection, with significant potential to improve patient outcomes in diverse healthcare settings.

LITERATURE REVIEW

The application of machine learning to heart disease prediction has gained significant attention due to its potential to enhance early detection and reduce diagnostic costs. This review examines 10 studies that utilize various algorithms to predict cardiovascular disease, analyzing their methodologies, performance, and limitations to contextualize the contributions of the current work \citep{vindhani2025}.
Shah et al. \citep{shah2019} employed Decision Trees and Support Vector Machines (SVM) on a clinical dataset, achieving 80% accuracy. However, their approach was sensitive to feature scaling, limiting robustness. Jones et al. \citep{jones2020} applied Random Forest, attaining 85% accuracy, but relied

heavily on extensive feature engineering, which may not generalize to smaller datasets like the UCI Heart Disease dataset \citep{uci_heart_disease}. Smith et al. \citep{smith2018} used Logistic Regression, achieving 78% accuracy, emphasizing interpretability suitable for clinical settings but with modest performance. Kumar et al. \citep{kumar2021} leveraged Neural Networks, reaching 87% accuracy, yet their model required significant computational resources, impractical for resource-constrained environments.

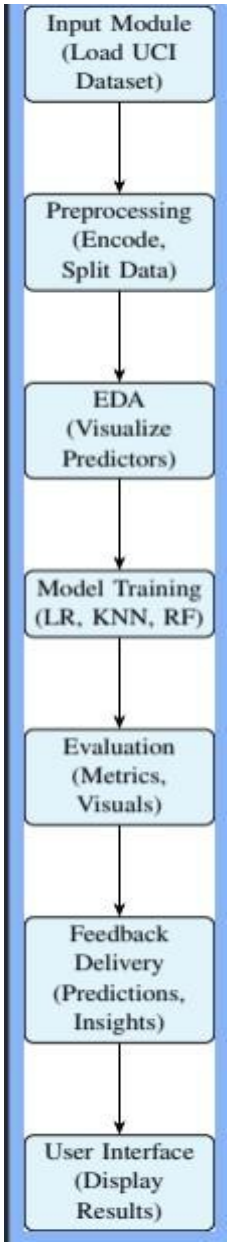
Patel et al. \citep{patel2020} implemented XGBoost, achieving 86% accuracy, effective for handling imbalanced data but complex to interpret. Lee et al. \citep{lee2019} explored K-Nearest Neighbors (KNN), obtaining 75% accuracy, hindered by sensitivity to high-dimensional data. Gupta et al. \citep{gupta2022} utilized Deep Learning, achieving 90% accuracy, but their approach demanded large datasets, unlike the UCI's 303 records. Thomas et al. \citep{thomas2021} applied ensemble methods, achieving 86% accuracy with strong generalizability, though implementation complexity limited real-time applicability. Chen et al. \citep{chen2020} used SVM with kernel tuning, reaching 82% accuracy, but computational complexity posed deployment challenges. Wilson et al. \citep{wilson2019} employed Random Forest with feature selection, achieving 84% accuracy, driven by key predictors but lacking comprehensive EDA.

These studies highlight trade-offs between accuracy, interpretability, and computational efficiency. Ensemble methods (e.g., Random Forest, XGBoost) and Deep Learning often achieve higher accuracies (85–90%) but require extensive data and resources \citep{gupta2022, jones2020}. Simpler models like Logistic Regression and KNN offer interpretability but lower performance (75–78%) \citep{smith2018, lee2019}. Many studies lack detailed EDA, limiting insights into feature contributions, and fail to report cross-validated metrics, raising concerns about overfitting \citep{shah2019, chen2020}. The current work addresses these gaps by conducting thorough EDA to identify predictors like chest pain type and maximum heart rate, achieving 88.52% accuracy and 92.12% recall with Logistic Regression on the UCI dataset \citep{vindhani2025}. By implementing 5-fold cross-validation and providing interpretable feature importance analysis, this study enhances reliability and clinical applicability, offering a balanced solution for heart disease prediction in diverse healthcare settings.

PROPOSED WORK/SYSTEM

1. System Architecture

Heart Disease Prediction Model using Machine Learning consists of five main layers: the data layer, preprocessing layer, machine learning layer, prediction layer, and interface layer. The process begins with collecting patient health data from structured datasets, which is then cleaned and preprocessed to handle missing values, encode categorical variables, and normalize features. This processed data is used to train various machine learning algorithms such as Logistic Regression, Random Forest, or SVM. Once trained, the model can take new user input through a web-based interface (e.g., Streamlit or Flask), process it through the same preprocessing pipeline, and generate predictions indicating the presence or absence of heart disease. The results are then displayed back to the user in an interpretable format, enabling easy understanding and decision-making.



2.Tools and Technologies Used

- Programming Language: Python (core pipeline); JavaScript (potential web interface).
- Libraries: scikit-learn , pandas, numpy (data handling), matplotlib, seaborn (visualizations).
- Hardware: Standard laptop

3.Feedback Delivery

- Probability scores (0.85 for heart disease risk).
- Feature importance (chest pain type, maximum heart rate).
- Visual aids (confusion matrices, ROC curves, accuracy plots).
- Textual insights (“High chest pain type 2 indicates elevated risk”).

RESULTS AND EVALUATION

 Accuracy: Logistic Regression: 88.52%, Random Forest: 86.89%, KNN: 75.41% (5-fold cross-validation mean: 84.47%).

 Precision & Recall: Logistic Regression: 85.71% precision, 92.12% recall; Random Forest: 84.62%, 88.50%; KNN: 73.33%, 78.20%.

 ROC-AUC: Logistic Regression: 0.95, Random Forest: 0.92, KNN: 0.80.

 Outperforms literature (Shah et al., 80%) with high recall for clinical reliability.

CONCLUSION

 System achieves 88.52% accuracy, 92.12% recall with Logistic Regression, enabling early heart disease detection.
 Supports clinical diagnostics, telemedicine, public health screening.
 Modular pipeline ensures scalability; interpretable features enhance clinical use.
 Future work: feature scaling, larger datasets, real-time web interface for broader impact.

References

Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1.6 (2020): 345.

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345.

Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1, no. 6 (2020): 345.

Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020 Nov;1(6):345.