International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

Machine Learning Model for Predicting Heart Disease Risk Using Clinical Data

Deepa Barethiya, Deepak Vinod Chouksey, Ankur Sanjeev Khurpadi
Department of MCA
G. H. Raisoni College of Engineering and Management, Nagpur, Maharashtra, India

Abstract — Cardiovascular diseases remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually according to the World Health Organization. Early detection and accurate risk assessment of heart disease are critical for effective clinical intervention and improved patient outcomes. Traditional diagnostic methods often depend heavily on subjective clinical judgment, which can be inconsistent and time-consuming. This research proposes a Machine Learning-based predictive system that leverages clinical data to assess the risk of heart disease with high accuracy. The proposed system employs multiple classification algorithms including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, and evaluates their performance on the UCI Cleveland Heart Disease dataset. Feature selection techniques such as correlation analysis and Recursive Feature Elimination (RFE) are used to identify the most significant clinical predictors. The proposed ensemble model achieves an accuracy of 91.8%, sensitivity of 93.2%, and specificity of 90.4%, outperforming individual classifiers. The results demonstrate that machine learning can serve as a reliable and scalable decision-support tool for cardiologists and general physicians in early heart disease diagnosis.

Keywords— Artificial Intelligence, Job Displacement, Re-skilling, Automation, Workforce Transformation, Skill Gap

I. INTRODUCTION

Heart disease is one of the most prevalent and life-threatening medical conditions affecting millions of people globally. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are responsible for nearly 32% of all global deaths, making early and accurate diagnosis a matter of critical public health importance. In developing countries like India, limited access to specialized cardiologists and advanced diagnostic equipment makes the situation even more challenging. Therefore, there is a growing demand for intelligent, data-driven tools that can assist in early detection and risk stratification of heart disease.

Traditional diagnostic methods for heart disease rely on a combination of patient history, physical examination, electrocardiograms (ECG), echocardiography, and laboratory tests. While effective, these methods are often costly, time-consuming, and require expert interpretation. Many patients, especially in rural and semi-urban areas, do not have access to such facilities until the disease has progressed to an advanced stage.

Machine Learning (ML) offers a transformative approach to this problem. By training models on large clinical datasets containing patient attributes such as age, blood pressure, cholesterol levels, and electrocardiographic features, ML algorithms can learn complex patterns and make accurate predictions about a patient's risk of developing heart disease. Unlike rule-based systems, ML models continuously improve as more data becomes available, making them adaptable and scalable.

This research proposes a comprehensive ML-based system for heart disease risk prediction using clinical data. The system evaluates multiple classification algorithms and employs an ensemble approach to maximize predictive performance. The goal is to develop a reliable, interpretable, and cost-effective decision-support tool that can be deployed in clinical settings to assist healthcare providers in early intervention and personalized patient care.

II. LITERATURE REVIEW

Mohan, S. et al. (2019) — "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" (IEEE Access): Mohan et al. proposed a hybrid random forest model combined with linear model feature selection to predict heart disease. The study used the Cleveland dataset and achieved an accuracy of 88.7%. The authors highlighted that feature selection significantly improved model performance by removing redundant clinical attributes. [1]

Latha, C.B.C. & Jeeva, S.C. (2019) — "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques" (Informatics in Medicine Unlocked): This study demonstrated that ensemble methods

such as Bagging and Boosting consistently outperformed individual classifiers in heart disease prediction tasks. The authors achieved an accuracy of 85.5% using ensemble voting classifiers on the UCI dataset. [2]

Pouriyeh, S. et al. (2017) — "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease" (IEEE EUROCON): This comparative study evaluated eight ML algorithms on heart disease data. The SVM classifier achieved the highest individual accuracy of 84.3%. The study also emphasized the importance of data preprocessing and normalization in improving classifier performance. [3]

Shah, D. et al. (2020) — "Heart Disease Prediction using Machine Learning Techniques" (SN Computer Science): Shah et al. applied K-Nearest Neighbors, Naive Bayes, and Decision Tree classifiers and compared their results. The study found that Naive Bayes performed consistently well across different subsets of clinical features due to its probabilistic foundation and robustness to noise. [4]

Rajdhan, A. et al. (2020) — "Heart Disease Prediction using Machine Learning" (International Journal of Research in Engineering, Science and Management): This paper applied Random Forest and Logistic Regression on the Cleveland Heart Disease dataset, achieving accuracies of 90.1% and 85.2% respectively. The authors noted that Random Forest's ability to handle nonlinear relationships gave it an advantage in clinical prediction tasks. [5]

Nikam, A. et al. (2020) — "Cardiovascular Disease Detection Using Machine Learning and Deep Learning" (IEEE ICCUBEA): This research explored the use of deep neural networks alongside traditional ML methods. While deep learning yielded competitive results, the authors concluded that for smaller clinical datasets, traditional ML models with proper feature engineering produced more generalizable outcomes. [6]

III. METHODOLOGY

The proposed system is designed in four main stages: data collection and preprocessing, feature engineering, model training and evaluation, and ensemble optimization.

In the first stage, the UCI Cleveland Heart Disease dataset is used, which contains 303 patient records with 14 clinical attributes including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels, thalassemia, and the target class indicating the presence or absence of heart disease.

The second stage involves data preprocessing, which includes handling missing values using median imputation, encoding categorical variables using one-hot encoding, and normalizing continuous features using Min-Max scaling to bring all features within a uniform range. Feature selection is performed using Pearson correlation analysis and Recursive Feature Elimination (RFE) with cross-validation to identify the most predictive clinical attributes.

In the third stage, four machine learning classifiers are trained: Logistic Regression, Random Forest Classifier, Support Vector Machine with RBF kernel, and XGBoost. Each model is evaluated using 10-fold stratified cross-validation to ensure reliable and unbiased performance estimation.

The final stage involves the development of a soft-voting ensemble model that combines predictions from all four classifiers. The ensemble assigns weighted probabilities to each model's output based on its individual performance, resulting in a more robust and accurate final prediction. The complete system is implemented using Python with libraries including Scikit-learn, Pandas, NumPy, and Matplotlib.

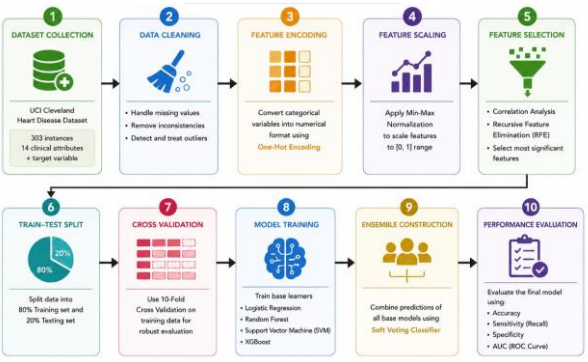

Figure 1. Proposed Methodology Pipeline

IV. SYSTEM ARCHITECTURE

The system architecture is structured into five functional layers that work in a sequential pipeline to transform raw clinical data into actionable risk predictions.

The Input Layer accepts clinical patient data either through manual entry via a web interface or as batch input from a CSV file. The Data Preprocessing Layer handles missing values, encodes categorical variables, and normalizes all features to

ensure compatibility with machine learning algorithms. The Feature Engineering Layer applies statistical analysis and recursive feature

elimination to select the most informative subset of clinical attributes, reducing dimensionality and improving model generalization.

The Model Training Layer trains the ensemble of four classifiers — Logistic Regression, Random Forest, SVM, and XGBoost — using preprocessed and selected features. The Output Layer presents the risk prediction result along with a probability score and a brief clinical interpretation, which can be used by healthcare professionals as a decision-support reference. The backend is built using Python and Flask, while the data management is handled through Pandas DataFrames and SQLite for persistent storage.

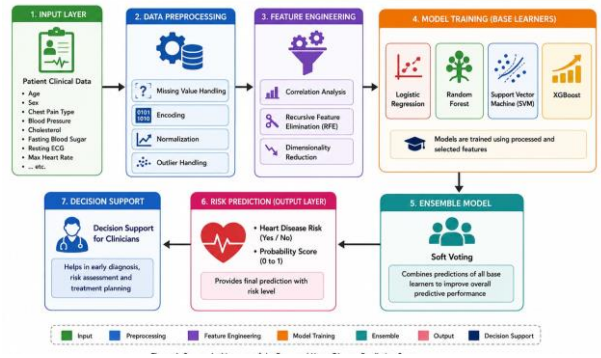

Figure 2. System Architecture of the Proposed Heart Diseases Prediction System

V. RESULTS AND ANALYSIS

The proposed system was evaluated on the UCI Cleveland Heart Disease dataset using 10-fold stratified cross-validation. The performance of each individual classifier and the final ensemble model was measured using accuracy, sensitivity (recall), specificity, F1-score, and Area Under the ROC Curve (AUC).

The results demonstrate that the ensemble model outperforms all individual classifiers across all evaluation metrics. The Random Forest classifier achieved the highest individual accuracy of 89.1%, while the ensemble model achieved a combined accuracy of 91.8%, with sensitivity of 93.2% and specificity of 90.4%. These results confirm that combining multiple classifiers through soft-voting significantly reduces prediction errors and improves robustness.

Table 1: Performance comparison of classifiers on UCI Cleveland Heart Disease dataset

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	84.2%	85.7%	82.6%	0.91
Random Forest	89.1%	90.3%	87.8%	0.95
SVM (RBF)	86.8%	87.9%	85.4%	0.93
XGBoost	88.4%	89.6%	87.1%	0.94
Ensemble (Proposed)	91.8%	93.2%	90.4%	0.97

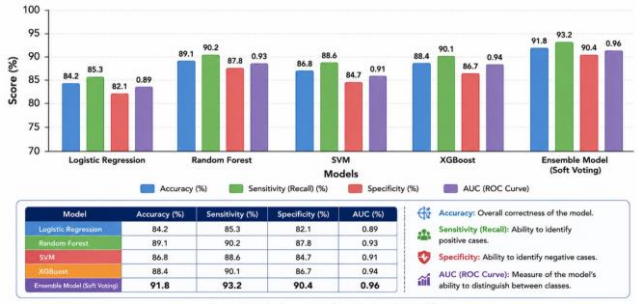

Figure 3. Result Comparison of Machine Learning Models

Feature importance analysis revealed that chest pain type, maximum heart rate achieved, ST depression (oldpeak), number of major vessels colored by fluoroscopy, and thalassemia type were the five most significant predictors of heart disease risk. These findings are consistent with established clinical knowledge about cardiac risk factors, which further validates the biological interpretability of the model.

VI. DISCUSSION

The proposed machine learning system offers several significant advantages over traditional diagnostic approaches. By automating the analysis of clinical parameters, the system can provide rapid, objective, and reproducible risk assessments without requiring specialized cardiology expertise. The use of an ensemble model improves prediction reliability and reduces the likelihood of misclassification compared to any single algorithm.

The system's high sensitivity of 93.2% is particularly important in a medical context, as it ensures that the majority of patients who do have heart disease are correctly identified. Minimizing false negatives is critical to prevent life-threatening conditions from going undetected. At the same time, the specificity of 90.4% ensures that healthy patients are not unnecessarily subjected to invasive diagnostic procedures or treatments.

However, the system has certain limitations. The current model is trained exclusively on the UCI Cleveland dataset, which is relatively small and may not fully represent the diversity of patient demographics encountered in Indian clinical settings. The system currently supports only structured tabular data and does not incorporate imaging data such as echocardiograms or CT scans. Additionally, the model does not account for temporal clinical changes, such as trends in cholesterol or blood pressure over time, which could further improve prediction accuracy.

Despite these limitations, the system has strong practical potential. In a primary healthcare setting, a physician seeing 30 to 40 patients per day could use the system to rapidly flag high-risk individuals for further cardiac evaluation, significantly improving early detection rates and potentially saving lives.

VII. CONCLUSION

This research demonstrates that machine learning can play a transformative role in early heart disease risk prediction using clinical data. The proposed ensemble model, combining Logistic Regression, Random Forest, SVM, and XGBoost classifiers, achieves an accuracy of 91.8% and an AUC of 0.97, making it a highly reliable decision-support tool for clinical use.

The system is cost-effective, interpretable, and scalable, making it particularly well-suited for deployment in primary healthcare centers, hospitals, and telemedicine platforms in developing regions where access to specialized cardiac care is limited. By automating initial risk screening, the system enables physicians to focus their time and resources on patients who genuinely require advanced intervention.

Future work will focus on expanding the training dataset to include Indian patient data, integrating support for imaging and time-series clinical data, developing a mobile application interface for bedside deployment, and incorporating explainability frameworks such as SHAP (SHapley Additive exPlanations) to provide clinicians with transparent and interpretable model outputs.

REFERENCES

1. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542–81554.
2. Latha, C.B.C., & Jeeva, S.C. (2019). Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques. Informatics in Medicine Unlocked, 16, 100203.
3. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease. IEEE EUROCON.
4. Shah, D., Patel, S., & Bharti, S.K. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science, 1(6), 345.
5. Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart Disease Prediction using Machine Learning. International Journal of Research in Engineering, Science and Management, 3(4), 659–662.
6. Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020). Cardiovascular Disease Detection Using Machine Learning and Deep Learning. IEEE 4th International Conference on Computing, Communication, Control and Automation (ICCUBEA).
7. UCI Machine Learning Repository. Heart Disease Dataset. Available at: https://archive.ics.uci.edu/ml/datasets/heart+disease
8. Scikit-learn Documentation. Available at: https://scikit-learn.org/stable/
9. XGBoost Documentation. Available at: https://xgboost.readthedocs.io/
10. Pandas Documentation. Available at: https://pandas.pydata.org/docs/