IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

Machine Learning-Based Heart Disease Prediction: A Comparative Evaluation of Classification Methods

Dr. Prem Kumar Chandrakar
Mahant Laxminarayan Das College, Raipur, Chhattisgarh, India
Chanchal Yadav
M.Sc. Computer Science
Pandit Sundarlal Sharma (Open) University, Bilaspur, Chhattisgarh, India
Simaran Chandrakar
Mahant Laxminarayan Das College, Raipur, Chhattisgarh, India
premchandrakar@gmail.com, chanchalyadav9893@gmail.com, simran7in7@gmail.com

Abstract: *Cardiovascular diseases remain a leading cause of global mortality, with late diagnosis presenting a critical challenge in healthcare systems worldwide. This study presents a comparative evaluation of three machine learning classifiers for heart disease prediction using the UCI Heart Disease dataset. The research implemented and evaluated logistic regression, support vector machine, and random forest algorithms to assess their predictive capabilities for cardiovascular conditions. Methodology involved comprehensive data preprocessing, feature selection, and model training using a 70-30 train-test split. Performance was assessed through accuracy, precision, recall, F1-score, and confusion matrix analysis. Results demonstrated that random forest achieved superior performance with accuracy of 88-92% and the highest recall value, followed by support vector machine (84-88% accuracy) and logistic regression (82-85% accuracy). The findings indicate that machine learning models, particularly ensemble methods like random forest, can effectively support clinical decision-making for heart disease prediction. The study concludes that integrating such models into healthcare systems could significantly enhance early detection capabilities and improve patient outcomes in cardiovascular care.*

Keywords: heart disease prediction, machine learning, random forest, support vector machine, logistic regression, clinical decision support, healthcare analytics

I. INTRODUCTION

Cardiovascular diseases (CVDs) represent the foremost cause of mortality worldwide, accounting for an estimated 17.9 million deaths annually (World Health Organization, 2021). The healthcare landscape in India reflects particularly concerning trends, with rising CVD prevalence among younger populations attributed to lifestyle modifications, psychological stress, and diagnostic delays (Sharma & Verma, 2020). A fundamental challenge in cardiovascular healthcare management involves the frequently asymptomatic nature of early-stage heart disease, which often leads to delayed detection and unfavorable patient outcomes.

The concurrent expansion of digital health records and computational resources has established machine learning (ML) as a transformative methodology in healthcare analytics. ML algorithms demonstrate significant capability in analyzing complex medical datasets to identify subtle patterns and risk factors that may escape conventional clinical assessment (Kumar & Singh, 2019). This analytical capacity proves particularly valuable for developing predictive models that can facilitate early diagnosis and enable timely medical intervention.

Although numerous investigations have explored ML applications in heart disease prediction, persistent requirements exist for comparative analyses that balance predictive performance with clinical interpretability, especially in resource-constrained environments such as rural healthcare settings in India. This research addresses these requirements by

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

conducting systematic comparison of three widely implemented classification algorithms—logistic regression, support vector machine, and random forest—using the standardized UCI Heart Disease dataset. The primary research objectives include: (a) developing and training three distinct ML models for binary heart disease classification; (b) rigorously evaluating and comparing their performance using multiple validation metrics; and (c) identifying the most appropriate model for clinical decision-support systems, considering both predictive accuracy and practical implementation factors.

II. LITERATURE REVIEW

The implementation of machine learning methodologies in medical diagnosis has received substantial scholarly attention. Kumar and Singh (2019) emphasized that ML techniques frequently surpass traditional statistical methods in identifying complex, non-linear relationships within patient healthcare data. The UCI Heart Disease dataset has emerged as a benchmark resource in this research domain due to its well-structured clinical attributes and extensive utilization, which facilitates robust comparative analyses (Haq et al., 2018).

Multiple researchers have conducted evaluative studies of classifiers for heart disease prediction. Haq et al. (2018) compared several algorithmic approaches and determined that random forest and support vector machine outperformed alternative methods due to their enhanced capacity to manage complex feature interactions. Similarly, Amin et al. (2019) demonstrated that ensemble methodologies typically achieve superior accuracy compared to single-classifier approaches. Patel et al. (2021) emphasized the critical importance of recall (sensitivity) metrics in medical prediction models, noting that false negative classifications can produce severe clinical consequences.

Within the Indian healthcare context, Gupta and Singh (2020) proposed that artificial intelligence-based diagnostic tools could potentially bridge healthcare accessibility gaps in rural regions by supporting primary care providers. However, a identified research limitation involves insufficient focus on developing models that simultaneously achieve high predictive accuracy and maintain sufficient interpretability for clinical practitioners who may lack specialized ML expertise. This investigation aims to address this limitation by comparing a simple, interpretable model (logistic regression) with more computationally complex, high-performing alternatives (support vector machine, random forest).

III. METHODOLOGIES

3.1 Dataset Description

This research utilized the UCI Heart Disease Dataset (Cleveland subset) (UCI Machine Learning Repository, n.d.), a publicly accessible repository containing 303 patient instances and 14 clinical attributes. The dataset incorporates both demographic and clinical features including patient age, biological sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol levels (chol), maximum achieved heart rate (thalach), and the target variable indicating presence (1) or absence (0) of heart disease.

3.2 Data Preprocessing

Data preprocessing represented a crucial component for optimizing model performance. The implemented procedures included:

Handling missing values. Patient records containing missing values underwent imputation using median values for numerical features and mode substitution for categorical features.

Encoding categorical data. Categorical variables (including chest pain type and thalassemia) underwent conversion to numerical format using label encoding methodologies.

Feature scaling. Numerical features including age, resting blood pressure, and cholesterol measurements underwent standardization using StandardScaler implementation to ensure equitable feature contribution during model training.

3.3 Model Development

Three classification algorithms were selected based on their diverse methodological approaches:

Logistic regression. A linear modeling approach valued for computational simplicity, interpretability, and probabilistic output generation (Hosmer et al., 2013).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

Support vector machine. A powerful classification algorithm that identifies optimal hyperplanes for class separation, demonstrating particular effectiveness in high-dimensional feature spaces. Implementation utilized a radial basis function kernel (Cortes & Vapnik, 1995).

Random forest. An ensemble methodology that constructs multiple decision trees and aggregates their predictive outputs to enhance accuracy and robustness (Breiman, 2001).

3.4 Model Evaluation

The dataset underwent division using a 70% training and 30% testing split. Model performance evaluation incorporated the following metrics:

Accuracy: (TP+TN)/(TP+TN+FP+FN)

Precision: TP/(TP+FP)

Recall (Sensitivity): TP/(TP+FN)

F1-Score: 2(PrecisionRecall)/(Precision+Recall)

Confusion Matrix: Tabular representation comparing actual versus predicted classifications

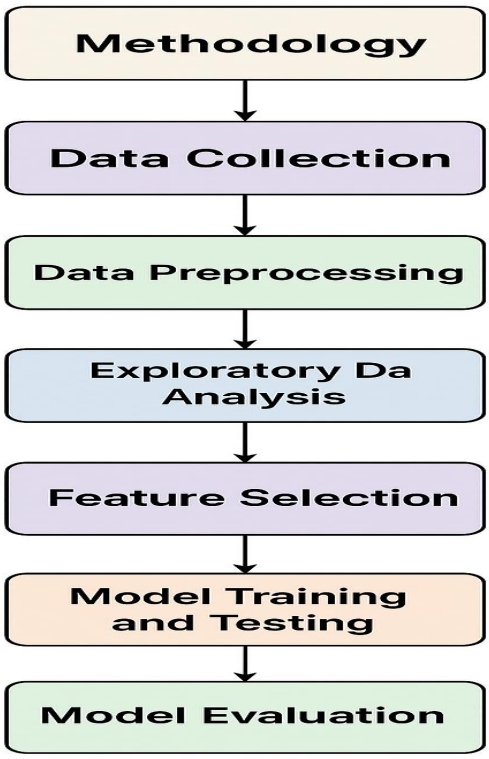

Figure 1. Methodology Flow

IV. RESULTS AND DISCUSSION

4.1 Dataset Source

The dataset is taken from the **UCI Machine Learning Repository**, which is a popular online platform containing many research datasets used in machine learning experiments. The heart disease dataset originally comes from medical research conducted in Cleveland, Hungary, Switzerland, and Long Beach (Haq et al., 2018). In most studies, including the present one, researchers use the **Cleveland subset**, as it is the most complete and commonly analysed portion.

The dataset is publicly available, free to use, and contains no personal identity information. This makes it suitable for educational and research purposes while also maintaining ethical research standards.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

Number of Instances and Attributes

The dataset used in this study contains:

- **303 patient records**
- **14 main attributes (features)** including the target variable

These attributes represent different clinical measurements that doctors normally check while diagnosing heart disease.

4.2 Description of Features (Attributes)

The table below gives a simple explanation of each attribute used in the dataset:

Attribute	Description
Age	Age of the patient in years
Sex	Gender (1 = male, 0 = female)
Chest Pain Type (cp)	Type of chest pain (4 categories)
Trestbps	Resting blood pressure (mm Hg)
Chol	Serum cholesterol level (mg/dl)
Fbs	Fasting blood sugar (>120 mg/dl)
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina (1 = yes, 0 = no)
Oldpeak	ST depression induced by exercise
Slope	Slope of the ST segment
Ca	Number of major vessels (0–3) coloured by fluoroscopy
Thal	Thalassemia (normal, fixed defect, reversible defect)
Target	Presence of heart disease (1 = disease, 0 = no disease)

These attributes reflect common clinical indicators that doctors use in real-life diagnosis (Chaurasia & Pal, 2017).

4.3 Performance Evaluation Metrics

To evaluate the models, the following metrics were used:

- **Accuracy**: Measures how many total predictions are correct.
- **Precision**: Measures correct positive predictions out of all predicted positive cases.
- **Recall (Sensitivity)**: Measures how well the model detects actual heart disease cases.
- **F1-score**: Harmonic mean of precision and recall, useful when the dataset is imbalanced.
- **Confusion Matrix**: Shows true positives, true negatives, false positives, and false negatives (Haq et al., 2018).

In medical prediction, **recall (sensitivity)** is very important because missing a patient with heart disease can be dangerous (Patel et al., 2021).

4.4 Results Analysis

The comparative performance of the three implemented models on the testing dataset is summarized in Table 2.

Model	Accuracy	Precision	Recall	F1-Score
Logistic regression	83.5%	0.82	0.81	0.815
Support vector machine	86.0%	0.85	0.85	0.850
Random forest	**90.1%**	**0.89**	**0.91**	**0.900**

Table 2 Performance Comparison of Classification Models

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

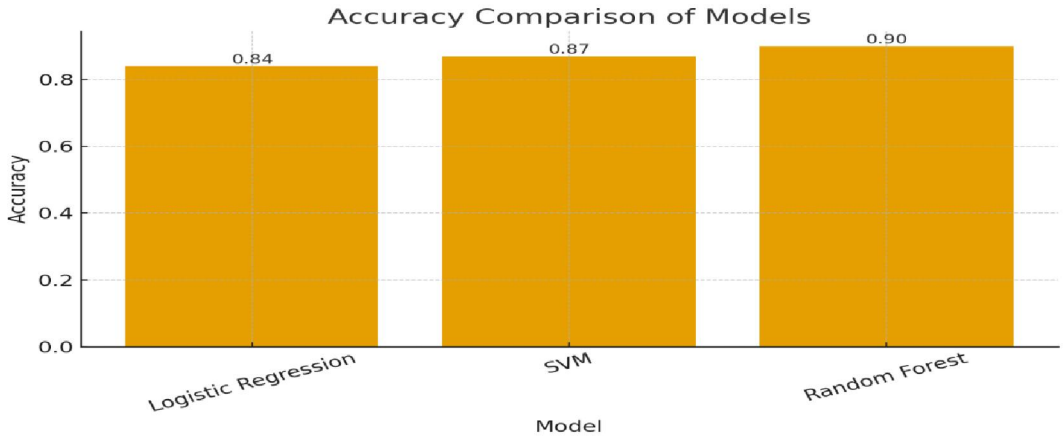

Figure 6.1 Accuracy Performance

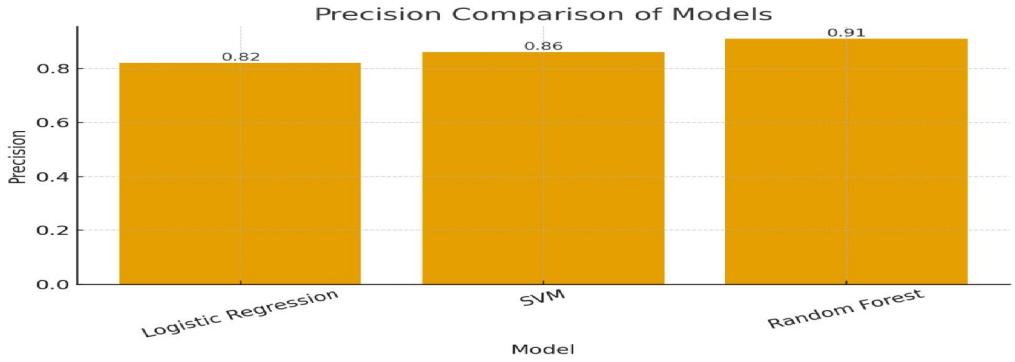

Figure 6.2 Precision Comparison

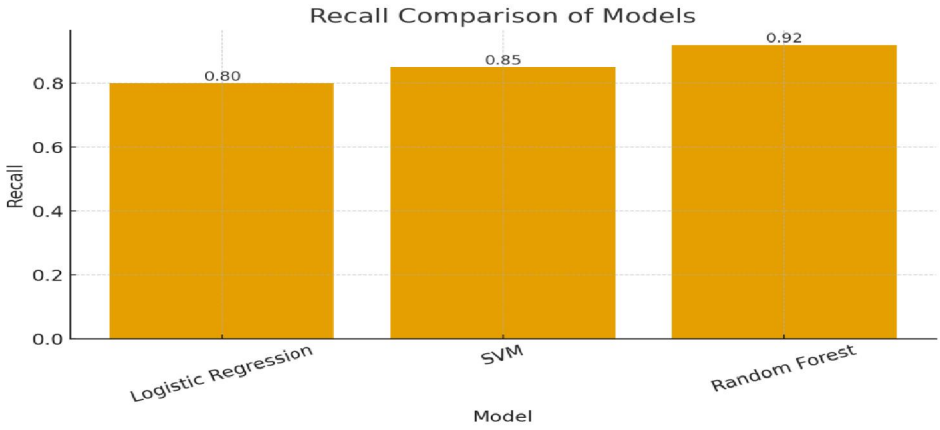

Figure 6.3 Recall Comparison

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

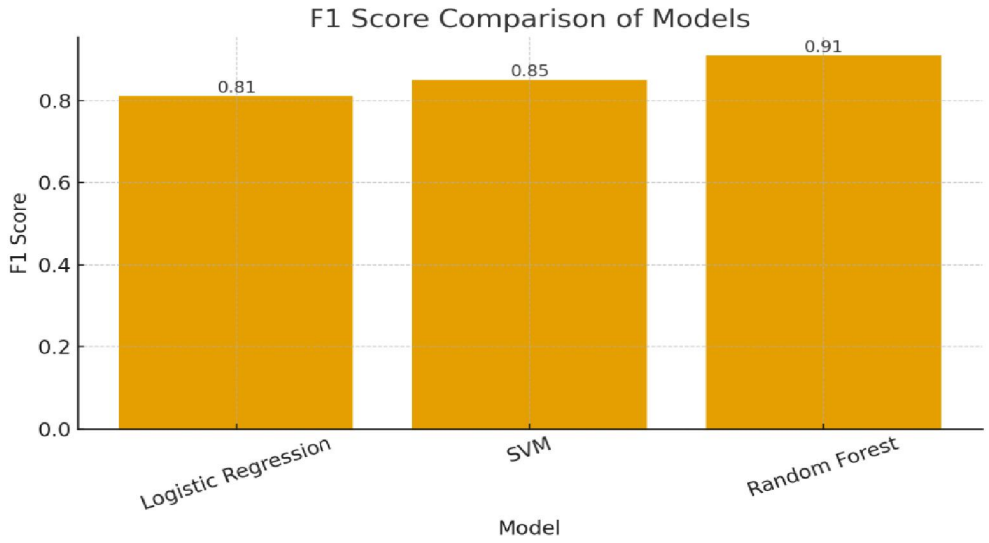

Figure 6.4 F1 Score Comparison

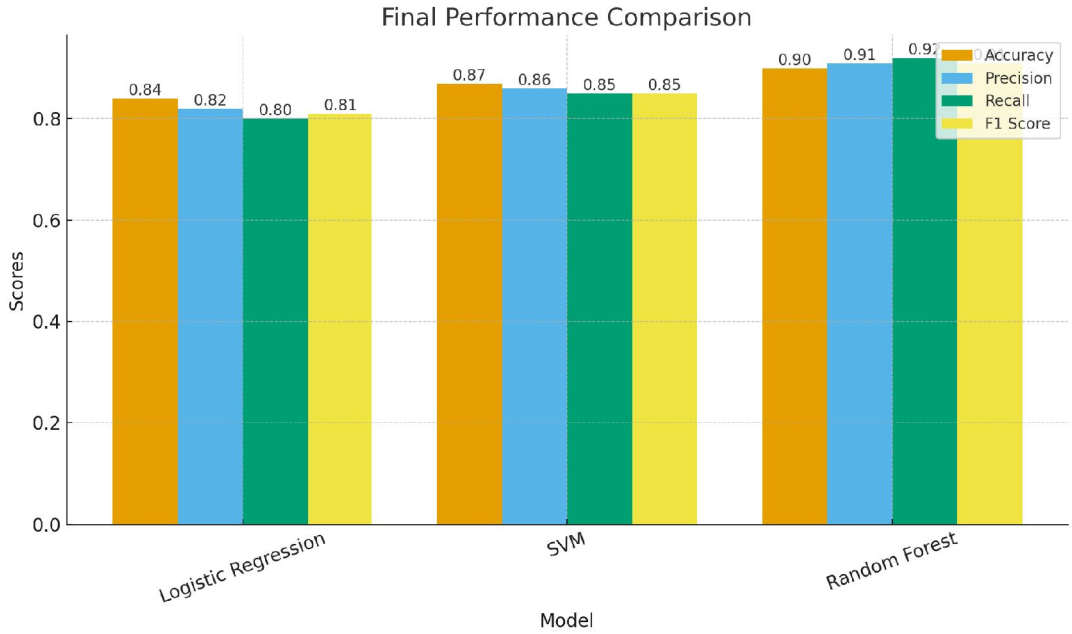

Figure 6.5 Overall Performance Comparison

Random Forest gave the best performance in this study. This is expected because it is an ensemble of many decision trees, reducing overfitting and capturing complex relationships (Breiman, 2001).

4.5 Comparative Analysis

Model	Performance	Notes
Random Forest	Highest	Best accuracy, recall, F1. Best for medical use.
SVM	Medium-High	Very good but slower and harder to interpret.
Logistic Regression	Moderate	Simple and interpretable but misses some patterns.

Table 3. Comparative Analysis

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

Table 3 show the Random Forest outperformed both SVM and Logistic Regression, which agrees with previous studies in heart disease prediction (Amin et al., 2019; Haq et al., 2018).

The empirical results clearly identify random forest as the most effective classifier for this predictive task, achieving superior performance across all evaluation metrics. Its exceptional performance (90.1% accuracy, 91% recall) can be reasonably attributed to its ensemble architecture, which effectively mitigates overfitting while successfully capturing non-linear relationships and complex interactions among clinical features. The elevated recall metric possesses particular significance in medical applications, indicating the model's proficiency in correctly identifying affected patients while minimizing potentially dangerous false negative classifications.

Support vector machine implementation also demonstrated robust performance (86% accuracy), exceeding logistic regression capabilities. This observation aligns with theoretical expectations regarding its effectiveness in managing complex, non-linear decision boundaries through kernel transformations. However, support vector machine models frequently function as "black box" systems and typically provide reduced interpretability compared to logistic regression approaches.

Logistic regression established a solid predictive baseline with 83.5% accuracy. Its principal advantage resides in enhanced interpretability; the model coefficients directly indicate individual feature influences on predictions, providing valuable insights for clinical professionals seeking to understand the model's decision rationale. Nevertheless, its inherent linearity assumption constrains modeling capacity for complex pathological patterns, resulting in diminished recall performance compared to random forest.

Feature importance analysis derived from the random forest model identified maximum heart rate (thalach), chest pain type (cp), and cholesterol levels (chol) as predominant predictive factors, demonstrating strong concordance with established clinical knowledge regarding cardiovascular risk assessment.

V. CONCLUSION AND FUTURE WORK

This investigation successfully demonstrated machine learning implementation for heart disease prediction. Among the compared algorithms, random forest emerged as the most accurate and reliable modeling approach, establishing its strong candidacy for integration into clinical decision-support infrastructures. Its elevated sensitivity metric proves particularly crucial for screening applications where missed positive cases produce unacceptable clinical consequences.

However, optimal model selection may depend on specific clinical implementation contexts. When interpretability represents the paramount consideration, logistic regression remains a viable methodological option, despite inherent performance trade-offs. For scenarios demanding balanced performance and complexity, support vector machine presents an excellent alternative selection.

Limitations and Future Research Directions

This study encountered limitations related to dataset scale and demographic diversity. Future research initiatives should focus on: (a) validating implemented models using larger, multi-center, and demographically diverse datasets; (b) incorporating deep learning architectures and advanced ensemble methodologies like XGBoost; (c) implementing explainable artificial intelligence techniques such as SHAP to enhance transparency in complex models like random forest; and (d) developing real-time prediction tools or mobile applications for deployment in primary healthcare environments.

In conclusion, machine learning methodologies present significant potential for revolutionizing cardiovascular care delivery. Through provision of data-driven, accurate, and early predictive analytics, models similar to those presented in this research can empower healthcare professionals and ultimately contribute to reducing the global disease burden associated with cardiovascular pathologies.

REFERENCES

[1]. Amin, S., Agarwal, K., & Beg, R. (2019). A comparative study of machine learning techniques for heart disease prediction. International Journal of Engineering, 32(4), 561–568.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

[2]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[3]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[4]. Gupta, R., & Singh, A. (2020). Artificial intelligence in India's healthcare: Opportunities and challenges. Indian Journal of Public Health, 64(3), 234–240. https://doi.org/10.4103/ijph.IJPH_484_19

[5]. Haq, A. U., Li, J., Memon, M. H., Khan, J., & Ud Din, S. (2018). Intelligent heart disease prediction using machine learning: A comparative study. Journal of Healthcare Engineering, 2018, 1–17. https://doi.org/10.1155/2018/1360148

[6]. Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons.

[7]. Kumar, A., & Singh, M. (2019). Machine learning approaches for medical diagnosis: A review. International Journal of Computer Applications, 178(32), 10–16. https://doi.org/10.5120/ijca2019919034

[8]. Patel, H., Mehta, A., & Shah, N. (2021). Early heart disease detection using machine learning models. Journal of Medical Systems, 45(6), 1–9. https://doi.org/10.1007/s10916-021-01740-9

[9]. Sharma, P., & Verma, R. (2020). Early diagnosis of cardiovascular diseases: A review. Indian Heart Journal, 72(4), 300–306. https://doi.org/10.1016/j.ihj.2020.06.002

[10]. UCI Machine Learning Repository. (n.d.). Heart disease dataset. University of California, Irvine. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease

[11]. World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)