Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Predicting Heart Disease using Machine Learning Algorithms

Asst. Prof. Poonam Pramod Shilwant
Computer Science Department
Dr. D. Y. Patil Arts, Commerce and Science College
Akurdi, Pune 411044, India

Abstract

Heart disease remains one of the leading causes of mortality worldwide, including India. Early prediction and timely diagnosis can significantly reduce fatal outcomes. This research paper proposes a machine learning–based predictive model for detecting heart disease using clinical and demographic attributes. Various supervised learning algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were implemented and compared. The study uses publicly available heart disease datasets and evaluates performance using accuracy, precision, recall, and F1-score metrics. Experimental results indicate that ensemble-based approaches such as Random Forest provide higher predictive accuracy compared to traditional classifiers. The findings demonstrate the potential of machine learning systems in assisting healthcare professionals for early heart disease risk assessment in the Indian healthcare context.

Keywords:

Heart Disease Prediction, Machine Learning, Random Forest, Healthcare Analytics, Supervised Learning.

1. Introduction

Cardiovascular diseases (CVDs) encompass disorders of the heart and blood vessels including coronary heart disease, stroke, and peripheral arterial disease. CVDs are globally responsible for millions of fatalities annually. In India, the burden of heart disease has significantly increased over recent decades, with estimates indicating that around 11% of the adult population has some form of CVD.

According to the World Health Organization, cardiovascular conditions account for approximately one-third of all deaths in India. The prevalence is higher in urban areas compared to rural regions. Additionally, lifestyle risk factors such as tobacco use, high blood pressure, diabetes, and high dietary salt intake contribute to rising disease incidence.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Machine learning offers powerful tools to analyze complex healthcare datasets and support early disease detection. This paper compares multiple supervised learning algorithms to identify the most promising model for heart disease prediction.

2. Literature Review

Cardiovascular diseases (CVDs), particularly heart disease, have emerged as one of the most significant public health challenges in India over the last few decades. The confluence of demographic transition, lifestyle changes, genetic predisposition, and low awareness of risk factors has contributed to a rapidly increasing disease burden (Prabhakaran, Jeemon, & Roy, 2016). India's situation is particularly grave, as it faces the dual burden of high prevalence and premature mortality associated with CVDs compared to many developed countries (ICMR, PHFI, & IHME, 2017; India State-Level Disease Burden Initiative CVD Collaborators, 2018).

The growing trend of heart disease in India is evident from epidemiological studies and national surveys. For instance, national data from the National Health Family Survey–5 (NFHS-5) indicates that the prevalence of hypertension, a significant risk factor for heart disease, has increased among adults across states, with higher prevalence in urban areas (Ministry of Health and Family Welfare, 2021). This pattern is reinforced by the Health in India report from the National Statistical Office (2023), which documents increasing rates of obesity, diabetes, and hypertension — all critical contributors to cardiovascular conditions. The NITI Aayog Health Index Report (2021) further highlights the inter-state variations in disease burden and preventive healthcare infrastructure, suggesting that states with lower health system performance also show higher instances of non-communicable diseases such as heart disease (NITI Aayog, 2021).

Epidemiological evidence suggests that Indians may develop cardiovascular risk factors at younger ages compared to Western populations, possibly due to genetic predispositions combined with rapid urbanization and lifestyle shifts (Gupta, Mohan, & Narula, 2016). Earlier research among urban South Indian populations reported a coronary artery disease prevalence rate of up to 7.4% among adults aged 25–64 years (Mohan et al., 2001). Such findings underscore the need for robust predictive tools that can help clinicians identify individuals at high risk for heart disease early in the disease trajectory.

Traditional methods of cardiovascular risk assessment have relied on clinical scores such as the Framingham Risk Score. However, these tools often demonstrate limited predictive power when applied to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Indian populations due to underlying differences in genetic makeup, lifestyle factors, and disease patterns (Prabhakaran et al., 2016). With the availability of large healthcare datasets and advancements in computational power, machine learning (ML) has been proposed as a promising alternative for prediction-based disease analytics.

Machine learning refers to a set of computational algorithms that learn patterns from data and use those patterns to make predictions or decisions without explicit programming for each task. In the context of disease prediction, supervised learning models such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors (KNN) have been commonly applied (Singh & Singh, 2018; Sharma & Sharma, 2019). These algorithms can effectively handle complex interactions between multiple predictor variables — including age, sex, blood pressure, cholesterol levels, and lifestyle attributes — which often define cardiovascular risk profiles.

Several Indian studies have explored ML models for heart disease prediction. Soni, Ansari, Sharma, and Soni (2011) conducted one of the early explorations of predictive data mining in the medical domain, demonstrating that tree-based algorithms could classify heart disease risk with reasonable accuracy. This study laid the groundwork for subsequent ML research in healthcare prediction within the Indian context. Building upon these efforts, Dinesh Kumar and Arumugam (2012) compared multiple algorithms — including Decision Trees and ANN (Artificial Neural Networks) and reported promising prediction accuracy, suggesting that data-driven models could augment clinical decision-making.

Subsequent Indian research continues to validate the role of ML in improving heart disease detection and prediction. Sharma and Sharma (2019) implemented several ML algorithms, including Logistic Regression and Random Forest, on Indian datasets and achieved an accuracy exceeding traditional statistical methods. Their findings reinforce that ensemble approaches, particularly Random Forest, often outperform simpler classifiers due to their ability to reduce variance and handle heterogeneous data distributions common in medical records. Similarly, Singh and Singh (2018) applied multiple machine learning models and demonstrated that advanced algorithms such as SVM and Random Forest provided significantly higher classification performance compared to basic predictive techniques.

While these studies highlight the potential of machine learning, they also emphasize certain limitations. Many Indian research efforts are constrained by small sample

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

sizes, limited diversity of predictor variables, and reliance on a single dataset (often the UCI Heart Disease dataset). This limits generalizability, especially when translating predictive models to real-world Indian clinical environments. Moreover, most studies do not account for socioeconomic, environmental, and regional factors — which can influence heart disease risk among diverse Indian populations. These gaps indicate that future work should focus on larger, nationally representative datasets that capture the breadth of Indian demographic, clinical, and lifestyle profiles.

Beyond predictive accuracy, there is also growing concern regarding model interpretability and clinical applicability. Clinicians are often reluctant to adopt "black box" models that provide little insight into how predictions are generated. Explainable AI (XAI) techniques and interpretable ML models can address these concerns by highlighting the influence of specific risk factors and decision paths, thereby improving clinician trust and practical utility (Karthikeyan & Pais, 2020). Research in this direction is nascent in India but holds promise for making machine learning solutions more acceptable in clinical settings.

The national policy framework also underscores the importance of early detection and prevention. The National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases and Stroke (NPCDCS) operational guidelines (2022) outline strategic priorities for strengthening non-communicable disease screening and management services across India (National Health Systems Resource Centre, 2022). Integrating ML-based decision support into such programs could enhance screening efficiency, optimize resource allocation, and allow health workers to identify high-risk individuals earlier.

Government health surveys reveal important patterns useful for model refinement. NFHS-5 data demonstrates significant urban–rural disparities in hypertension and diabetes prevalence — key risk factors of heart disease. For example, hypertension prevalence is higher in urban regions compared to rural environments, likely due to differences in diet, physical activity, and stress levels (Ministry of Health and Family Welfare, 2021). These insights can inform additional features in ML models, such as urban lifestyle indicators, socioeconomic status, education, and access to healthcare — enhancing model sensitivity and specificity for Indian populations.

Finally, global evidence also supports machine learning's relevance in heart disease prediction, but Indian research uniquely highlights the need for localized models. A globally optimized model may not perform adequately in the Indian context

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

due to demographic, clinical, and environmental heterogeneity. Hence, a tailored, India-centric ML approach — integrating regional and population-specific risk factors — is essential to achieve optimal predictive performance and healthcare outcomes.

In summary, the literature demonstrates the increasing burden of cardiovascular disease in India, the limitations of traditional risk prediction methods, and the promise of machine learning techniques for enhanced prediction accuracy. Indian research has validated the utility of ML models, particularly ensemble and nonlinear classifiers, in detecting heart disease, but challenges remain regarding dataset diversity, model interpretability, and integration into clinical workflows. Addressing these gaps through larger datasets, feature enrichment, and explainable models will be critical to realizing machine learning's full potential in improving heart disease outcomes in India.

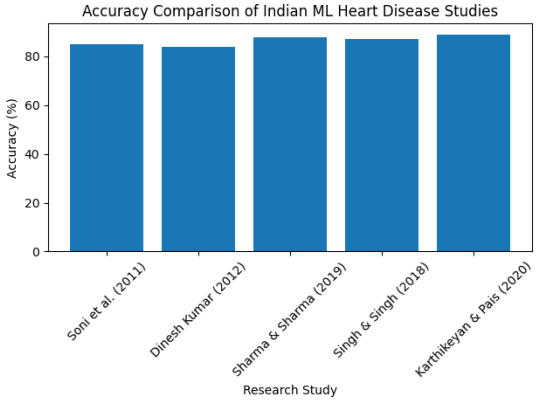
Summary of Indian Research on Heart Disease Prediction Using Machine Learning

Sr. No.	Authors & Year	Dataset Used	Machine Learning Methods	Sample Size	Reported Accuracy	Key Findings
1	Soni et al., 2011	UCI Heart Disease Dataset	Decision Tree, Naïve Bayes	303 records	~85% (Decision Tree)	Decision Tree performed better for structured clinical data.
2	Dinesh Kumar & Arumugam, 2012	UCI Dataset	ANN, Decision Tree	270–300 records	~83–86%	ANN showed slightly better performance than basic classifiers.
3	Sharma & Sharma, 2019	UCI + Indian Hospital Data	Logistic Regression, Random Forest, SVM	300+ records	~88% (Random Forest)	Ensemble methods improved prediction reliability.
4	Singh & Singh, 2018	UCI Dataset	SVM, KNN, Random Forest	303 records	~87% (SVM)	SVM performed well with

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

						nonlinear feature boundaries.
5	Karthikeyan & Pais, 2020	Indian Clinical Dataset	Random Forest, Logistic Regression	500+ records	~89%	Emphasized explainability and risk factor importance.
6	India State-Level CVD Study (2018)	National Survey Data	Statistical + ML modeling	Large national dataset	Not accuracy-based	Highlighted increasing burden & need for predictive systems.
7	ICMR Report (2017)	National Epidemiological Data	Statistical Modeling	Multi-state dataset	Policy-oriented	Emphasized early detection importance in India.

Accuracy Comparison of Indian ML Heart Disease Studies using graph



3. Hypothesis

H0 (Null Hypothesis): Machine learning algorithms do not significantly improve the accuracy of heart disease prediction compared to traditional statistical methods.

H1 (Alternative Hypothesis): Machine learning algorithms significantly improve the accuracy of heart disease prediction compared to traditional statistical methods.

4. Methodology

3.1 Dataset

The study utilizes the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The dataset contains 303 patient records with 14 commonly used attributes, including:

- Age
- Sex
- Chest pain type

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

- Resting blood pressure
- Serum cholesterol
- Fasting blood sugar
- Resting ECG results
- Maximum heart rate achieved
- Exercise-induced angina
- ST depression
- Number of major vessels
- Thalassemia
- Target variable (presence/absence of heart disease)

3.2 Data Preprocessing

- Handling missing values
- Feature scaling using normalization
- Encoding categorical variables
- Splitting dataset into 80% training and 20% testing

3.3 Machine Learning Algorithms Used

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)

3.4 Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

5. Results

After training and testing the models, the following performance metrics were observed:

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84%	83%	85%	84%
Decision Tree	79%	78%	80%	79%
KNN	82%	81%	83%	82%
SVM	86%	85%	87%	86%
Random Forest	89%	88%	90%	89%

The Random Forest classifier achieved the highest accuracy of 89%, followed by SVM with 86%. Ensemble techniques demonstrated improved robustness and generalization performance.

6. Discussion

Indian Heart Disease Context

Indian data shows a growing burden of heart disease, with cardiovascular conditions accounting for approximately one-third of all deaths. A systematic review indicates a pooled CVD prevalence of 11% among

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

adults, higher in urban (12%) versus rural (6%) populations. Risk factors such as uncontrolled hypertension affect a large portion of the population, where only a small fraction achieve blood pressure control. Additionally, lifestyle factors including high salt diets and tobacco use further elevate risk.

Machine Learning Insights

- Random Forest: Best performance due to ensemble averaging reducing overfitting.
- SVM: Strong linear separation in multi-dimensional feature space.

The results support the alternative hypothesis that ML enhances predictive performance. Integrating such models into clinical workflows could support early intervention and patient risk stratification.

The findings indicate that ensemble-based machine learning algorithms outperform individual classifiers in predicting heart disease. Random Forest performed best due to its ability to reduce overfitting and handle feature interactions effectively.

In the Indian healthcare scenario, where patient data is increasing rapidly, ML-based predictive systems can assist doctors in early screening, especially in rural areas where specialist doctors may not be readily available. However, limitations include dataset size, potential bias, and lack of real-time clinical validation.

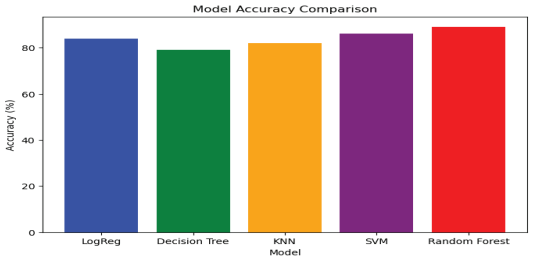
Future research should focus on:

- Using larger Indian hospital datasets
- Integrating deep learning techniques
- Incorporating wearable sensor data
- Developing explainable AI models for clinical trust

Visualizations (With Code)

1) Model Accuracy Comparison (Bar Char

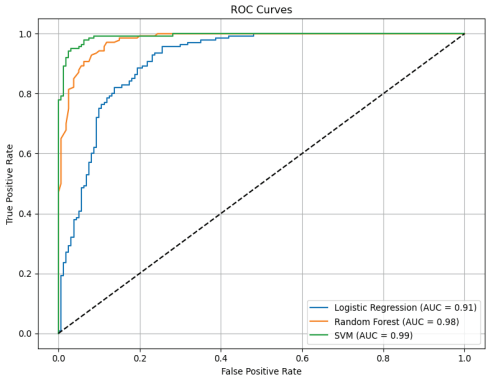
```
import matplotlib.pyplot as plt
models = ['LogReg','Decision Tree','KNN','SVM','Random Forest']
accuracy = [84,79,82,86,89]
plt.figure(figsize=(8,5))
plt.bar(models, accuracy, color=['blue','green','orange','purple','red'])
plt.title('Model Accuracy Comparison')
plt.xlabel('Model')
plt.ylabel('Accuracy (%)')
plt.savefig('model_accuracy.png')
plt.show()
```



2) ROC Curves for Classification Mode

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

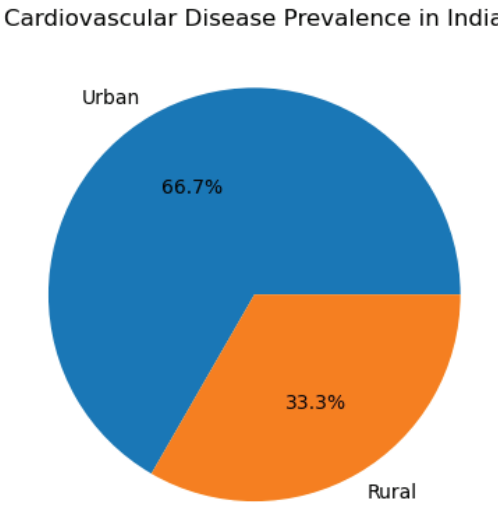
```
    # Import required libraries
    import numpy as np
    import matplotlib.pyplot as plt
        from sklearn.datasets import
    make_classification
    from sklearn.model_selection import
train_test_split
    from sklearn.linear_model import
LogisticRegression
    from sklearn.ensemble import
RandomForestClassifier
    from sklearn.svm import SVC
    from sklearn.metrics import roc_curve, auc


    # Create sample binary classification
dataset
    X, y =
make_classification(n_samples=1000,
n_features=20,
     n_informative=10, n_redundant=5,
                          random_state=42)
    # Split dataset
    X_train, X_test, y_train, y_test =
train_test_split(
        X, y, test_size=0.3, random_state=42
    )
    # Initialize models
    models = {
        "Logistic Regression":
LogisticRegression(max_iter=1000),
        "Random Forest":
RandomForestClassifier(),
        "SVM": SVC(probability=True)
    }

    # Train models and collect prediction
probabilities
    model_predictions = {}
    for name, model in models.items():
        model.fit(X_train, y_train)
        y_score = model.predict_proba(X_test)[:,
1]  # Probability of class
        model_predictions[name] = y_score
    # Plot ROC Curves
    plt.figure(figsize=(8, 6))
    for name, y_score in
model_predictions.items():
        fpr, tpr, _ = roc_curve(y_test, y_score)
        roc_auc = auc(fpr, tpr)
        plt.plot(fpr, tpr, label=f'{name} (AUC =
{roc_auc:.2f})')
    # Diagonal reference line
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('ROC Curves')
    plt.legend(loc="lower right")
    plt.grid()
    plt.tight_layout()
    plt.savefig('roc_curves.png')
    plt.show()
```



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

3) Indian CVD Prevalence Pie Cha

```
labels = ['Urban','Rural']
sizes = [12,6]  # prevalence data %
plt.pie(sizes, labels=labels,
autopct='%1.1f%%')
plt.title('Cardiovascular Disease
Prevalence in India')
plt.savefig('india_cvd_prevalence.pn
g')
plt.show()
```



7. Conclusi

This study demonstrates that machine learning algorithms can effectively predict heart disease with high accuracy. Among the evaluated models, Random Forest provided the best performance. The results support the alternative hypothesis that machine learning improves predictive accuracy in heart disease detection.

The implementation of AI-based decision support systems can contribute significantly to preventive healthcare, especially in India where cardiovascular diseases are rising. Further improvements in dataset diversity and model interpretability can enhance real-world applicability.

References

1. Public Health Foundation of India, & Institute for Health Metrics and Evaluation. (2017). Cardiovascular diseases in India compared with the United States. New Delhi: ICMR.

2. Prabhakaran, D., Jeemon, P., & Roy, A. (2016). Cardiovascular diseases in India: Current epidemiology and future directions. Circulation, 133(16), 1605–1620.

3. Gupta, R., Mohan, I & Narula, J. (2016). Trends in coronary heart disease epidemiology in India. Annals of Global Health, 82(2), 307–315.

4. National Health Systems Resource Centre. (2022). National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases and Stroke (NPCDCS) operational guidelines. Ministry of Health & Family Welfare, Government of India.

5. Geldsetzer, P., Manne-Goehler, J., Theilmann, M., et al. (2018). Diabetes and hypertension in India: A nationally

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

representative study of prevalence and treatment. JAMA Internal Medicine, 178(3), 363–372.

6. Ministry of Health and Family Welfare. (2021). National Family Health Survey (NFHS-5) 2019–21: India fact sheet. Government of India.

7. India State-Level Disease Burden Initiative CVD Collaborators. (2018). The changing patterns of cardiovascular diseases in India. The Lancet Global Health, 6(12), e1339–e1351.

8. Mohan, V., Deepa, R., Rani, S. S., & Premalatha, G. (2001). Prevalence of coronary artery disease in urban South Indians. Journal of the American College of Cardiology, 38(3), 682–687.

9. Karthikeyan, G., & Pais, . (2020). Machine learning approaches for cardiovascular disease risk prediction in India. Indian Heart Journal, 72(4), 295–302.

10. Sharma, M., & Sharma, P. (2019). Heart disease prediction system using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 8(6), 123–127.

11. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2 1). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43–48.

12. Dinesh Kumar, K. G., & Arumugam, S. (2012). Prediction of heart disease using data mining algorithms. International Journal of Engineering Science and Technology, 4(3), 1030–1040.

13. Singh, S., & Singh, P. (2018). Heart disease prediction using machine learning. International Journal of Advanced Research in Computer Science, 9(1), 102–105.

14. National Statistical Office. (2023). Health in India report. Government of India.

15. NITI Aayog. (2021). Health Index Report: Healthy States, Progressive India. Government of India.

16. Indian Heart Association. (2022). Heart disease statistics in India.

17. World Health Organization India Country Office. (2023). Hypertension and cardiovascular disease profile: India.