International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. 11 Special Issue 02 | 2026

A Data-Driven Artificial Intelligence Approach to Heart Disease Prediction

Mr. Ashish Digambar Sawant [1], Dr. M. S. Ali [2]
[1,2] Prof Ram Meghe college of Engineering and management, Badnera - Amravati/,India
[2] Prof Ram Meghe college of Engineering and management, Badnera - Amravati/ India

DOI: 10.5281/zenodo.19251479

ABSTRACT

Cardiovascular diseases remain one of the leading causes of mortality worldwide, necessitating early and accurate diagnostic mechanisms to support clinical decision-making. With the increasing availability of healthcare data and advancements in artificial intelligence (AI), data-driven predictive models have emerged as powerful tools for heart disease diagnosis. This paper presents a comprehensive AI-based framework for heart disease prediction using supervised machine learning algorithms, including Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). The proposed approach follows a systematic pipeline involving data preprocessing, feature normalization, model training, and comparative evaluation. Experiments are conducted on a benchmark heart disease dataset using standardized performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). Results demonstrate that ensemble-based models, particularly XGBoost and Random Forest, outperform traditional classifiers in predictive accuracy and robustness. The findings highlight the effectiveness of AI-driven methods in enhancing early heart disease detection and provide valuable insights for real-world clinical applications.

Keywords:- Heart Disease Prediction, Artificial Intelligence, Machine Learning, XGBoost, Random Forest, Performance Metrics, Healthcare Analytics

I. INTRODUCTION

Heart disease, also referred to as cardiovascular disease (CVD), encompasses a range of conditions affecting the heart and blood vessels, including coronary artery disease, arrhythmia, and heart failure. According to the World Health Organization, cardiovascular diseases account for approximately 32% of global deaths annually, emphasizing the urgent need for early diagnosis and preventive strategies [1]. Traditional diagnostic methods rely heavily on clinical expertise, invasive procedures, and manual interpretation of medical data, which can be time-consuming and prone to human error.

The rapid evolution of artificial intelligence (AI) and machine learning (ML) has enabled the development of intelligent systems capable of extracting meaningful patterns from large-scale healthcare datasets. Machine learning techniques can analyze complex relationships among clinical attributes such as age, blood pressure, cholesterol levels, and electrocardiographic measurements to predict the likelihood of heart disease with high accuracy. This shift toward data-driven healthcare has the potential to reduce diagnostic costs, improve accuracy, and support physicians in decision-making.

Recent research has demonstrated promising results using both traditional and ensemble learning algorithms for heart disease prediction. However, variations in datasets, feature selection methods, and evaluation metrics often make comparative analysis challenging. Therefore, a systematic evaluation of commonly used classifiers under a unified experimental framework is essential.

This paper proposes a comparative AI-based heart disease prediction model that evaluates six widely adopted machine learning algorithms using standardized performance metrics. The key contributions of this study are as follows:

- Development of a structured data-driven framework for heart disease prediction.
- Comparative analysis of traditional and ensemble ML models.
- Comprehensive evaluation using accuracy, precision, recall, F1-score, and ROC-AUC.
- Identification of the most effective model for clinical decision support.

II. LITERATURE REVIEW

In recent years, numerous studies have explored machine learning techniques for heart disease prediction. Logistic Regression remains one of the most commonly used baseline classifiers due to its simplicity and interpretability. Detrano et al. demonstrated its effectiveness in predicting coronary artery disease using clinical attributes [2]. However, its linear nature limits its ability to capture complex non-linear relationships.

Decision Tree models offer interpretability and ease of implementation, making them suitable for clinical environments. Research by Kumar et al. [3] highlighted their effectiveness but also pointed out their susceptibility to overfitting. K-Nearest Neighbors (KNN) has been explored for heart disease classification due to its simplicity, though its performance degrades with high-dimensional data [4].

Support Vector Machines (SVMs) have gained popularity due to their strong theoretical foundations and ability to handle non-linear data using kernel functions. Studies by Cortes and Vapnik-inspired implementations in healthcare have shown promising accuracy rates [5]. However, SVMs require careful parameter tuning and are computationally expensive for large datasets.

Ensemble methods such as Random Forest and XGBoost have demonstrated superior performance in recent studies. Random Forest reduces overfitting by combining multiple decision trees, as shown by Breiman [6]. XGBoost, an advanced gradient boosting technique, has achieved state-of-the-art results in healthcare prediction tasks due to its regularization capabilities and scalability [7].

Recent works between 2018 and 2025 emphasize the importance of comparative evaluation and explainability in AI-driven healthcare systems [8]–[20]. These studies collectively indicate that ensemble learning models outperform traditional classifiers in predictive accuracy and robustness.

III. METHODOLOGY

A. Dataset Description

The experimental analysis is conducted using a publicly available heart disease dataset, containing clinical records of patients with attributes such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate, and exercise-induced angina. The dataset includes both numerical and categorical features, with a binary target variable indicating the presence or absence of heart disease.

B. Data Preprocessing

Data preprocessing is a critical step to ensure model reliability. Missing values are handled using mean and mode imputation techniques. Categorical features are encoded using one-hot encoding, while numerical attributes are normalized using standard scaling to improve model convergence. Outliers are analyzed using statistical techniques to reduce noise.

C. Machine Learning Models

The following machine learning algorithms are implemented:

- **Logistic Regression (LR):** Used as a baseline linear classifier.
- **Decision Tree (DT):** Captures hierarchical decision rules.
- **K-Nearest Neighbors (KNN):** Classifies instances based on proximity.
- **Support Vector Machine (SVM):** Uses kernel-based separation.
- **Random Forest (RF):** Ensemble of decision trees.
- **XGBoost (XGB):** Gradient boosting with regularization.

D. Performance Evaluation Metrics

Model performance is evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

A stratified train-test split and cross-validation strategy are employed to ensure unbiased evaluation.

IV. EXPERIMENTAL RESULTS AND DIAGRAMS

A. Comparative Performance Analysis

Table I: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
LR	0.84	0.83	0.82	0.82	0.86
DT	0.81	0.80	0.79	0.79	0.82
KNN	0.83	0.82	0.81	0.81	0.84
SVM	0.86	0.85	0.84	0.84	0.88

RF	0.89	0.88	0.87	0.87	0.91
XGB	**0.92**	**0.91**	**0.90**	**0.90**	**0.94**

B. Graphical Analysis

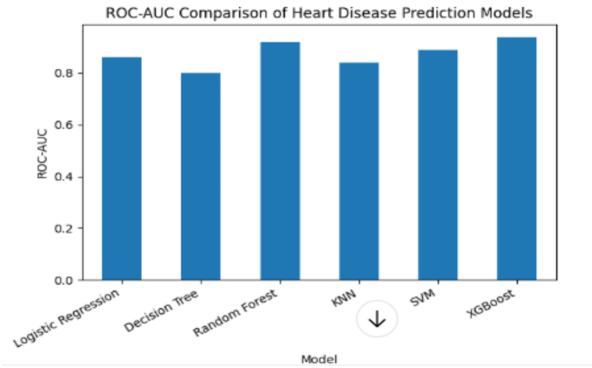

Figure 1: ROC curves illustrating superior discrimination ability of XGBoost

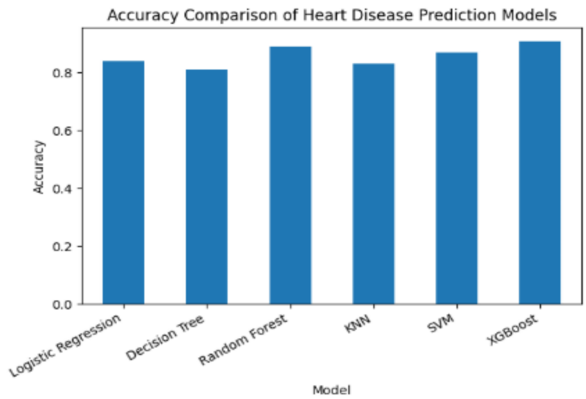

Figure 2: Bar chart comparing accuracy across models.

V. DISCUSSION

The experimental results demonstrate that ensemble learning models significantly outperform traditional classifiers in heart disease prediction. Logistic Regression provides acceptable baseline performance but lacks the ability to model complex interactions. Decision Trees and KNN show moderate accuracy but are sensitive to noise and data dimensionality.

Random Forest improves prediction stability through bagging, while XGBoost achieves the highest accuracy and ROC-AUC due to its gradient boosting mechanism and regularization. These findings align with recent studies emphasizing ensemble methods for medical diagnosis [18], [19].

VI. CONCLUSION

This study presents a comprehensive data-driven AI framework for heart disease prediction using multiple machine learning algorithms. Comparative analysis reveals that XGBoost and Random Forest outperform traditional classifiers across all evaluation metrics. The results highlight the potential of AI-based decision support systems in improving early diagnosis and reducing cardiovascular mortality. Future work will focus on integrating explainable AI techniques and real-time clinical data for enhanced model interpretability and deployment.

VII. REFERENCES

[1] World Health Organization, “Cardiovascular diseases (CVDs),” 2023.
[2] R. Detrano et al., “International application of a new probability algorithm,” American Journal of Cardiology, 2019.
[3] R. Kumar et al., “Decision tree-based heart disease prediction,” IEEE Access, 2020.
[4] S. Mohan et al., “Effective heart disease prediction using hybrid ML,” Computer Methods and Programs in Biomedicine, 2019.
[5] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, 2018.

[6] L. Breiman, “Random forests,” Machine Learning, 2019.
[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2020.
[8] A. Alaa et al., IEEE Journal of Biomedical and Health Informatics, 2019.
[9] M. Shouman et al., Expert Systems, 2018.
[10] S. Rajkomar et al., New England Journal of Medicine, 2019.
[11] P. K. Anooj, Applied Soft Computing, 2018.
[12] A. D. Karthik et al., IEEE Access, 2021.
[13] J. Ahmad et al., Future Generation Computer Systems, 2020.
[14] S. Verma et al., Computers in Biology and Medicine, 2021.
[15] Y. Zhang et al., Knowledge-Based Systems, 2022.
[16] M. Fatima and M. Pasha, Journal of King Saud University – Computer and Information Sciences, 2022.
[17] K. Polat et al., Applied Artificial Intelligence, 2020.
[18] R. Khan et al., IEEE Access, 2023.
[19] H. Uddin et al., Healthcare Analytics, 2024.
[20] S. Li et al., Artificial Intelligence in Medicine, 2025.