RESEARCHERID
THOMSON REUTERS
[Vineet et al., 11(7), July 2026]
ISSN : 2455-9679
SJIF Impact Factor : 6.190

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"HEART DISEASE PREDICTION USING MACHINE LEARNING CLASSIFICATION TECHNIQUES: A SURVEY"

Vineet Richhariya [1], Sumbul [2]

[1] Professor, Department of Computer Science and Engineering, LNCT Bhopal, Madhya Pradesh, India
[2] M. Tech Scholar, Department of Computer Science and Engineering, LNCT Bhopal, Madhya Pradesh, India
vineetr@lnct.ac.in
sumbulkhan283@gmail.com

ABSTRACT

Artificial intelligence (AI) is utilized to mimic human intelligence through machines and software simulators, enhancing the ability to make prompt and data-driven decisions. In the medical sector, where vast amounts of data often overlap, AI can assist in identifying and mitigating risk factors through advanced classification techniques and data analysis. This research paper study examines the efficacy of a machine learning model designed to predict heart disease. The machine learning model performance assessed through various metrics, including recall, accuracy, precision F1 score, and the area under the receiver operating characteristic (ROC) curve. This research work also states the comparative literature study using machine learning classification for heart disease prediction.

Key Words: *Artificial intelligence, Machine learning, Supervised learning, Disease.*

I. INTRODUCTION

Heart disease is the leading cause of death globally, accounting for 17.9 million fatalities in 2019, according to the World Health Organization. Given its status as a critical organ, it is essential to predict heart disease effectively [1]. The complexity of medical diagnoses, coupled with the fact that 20-40% of heart attacks happen in undiagnosed patients due to instrument inaccuracies, highlights the need for improved disease prediction algorithms. Consequently, the development of an optimal machine-learning model for heart disease prediction is vital for enhancing diagnostic accuracy [2], with various researchers striving to achieve this in their work.

Heart disease, including conditions like coronary artery disease, heart failure, and arrhythmias, represents a leading global health challenge, resulting in millions of deaths annually. Early detection and accurate diagnosis are essential for reducing mortality rates and enhancing patient outcomes [3]. Traditional diagnostic approaches depend heavily on clinical expertise, including physical examinations and various medical tests (e.g., ECG, blood pressure, cholesterol analysis). However, these methods can be time-consuming, costly, and susceptible to human error, often leading to misdiagnosis or delayed treatment, especially in early stages where symptoms may not be easily identifiable [4, 5].

The coronary arteries supply blood to the heart muscle, and their narrowing leads to coronary artery disease (CAD), primarily caused by the accumulation of atherosclerotic material. This buildup results in stenosis, reducing blood supply, particularly during increased heart activity, and causing myocardial ischemia. Atherosclerotic material consists of fatty deposits that interact with blood components, eventually forming hardened plaques due to calcium deposits [6,7]. Angina pectoris is a common symptom of CAD, presenting as chest discomfort that may radiate to the arms,

ISSN : 2455-9679
SJIF Impact Factor : 6.190

neck, jaw, shoulders, or upper abdomen, and can be accompanied by sweating, nausea, or vomiting during intense episodes. Notably, atherosclerosis can progress asymptomatically [8], with individuals not exhibiting symptoms until significant coronary artery stenosis occurs.

Early detection and accurate diagnosis of cardiac disease significantly enhance patient outcomes and decrease mortality rates by enabling timely intervention and tailored treatment plans [9]. This proactive approach addresses risk factors and improves long-term disease management, yielding cost savings and alleviating the public health burden of cardiovascular issues. Traditional diagnostic methods, which typically depend on clinical assessments, medical histories, and invasive procedures, have limitations such as subjectivity, time consumption, and the requirement for specialized expertise [10, 11]. The intricate and varied nature of heart disease further complicates accurate diagnosis for healthcare professionals. However, recent advancements in machine learning (ML) and extensive healthcare datasets have revolutionized disease detection and diagnosis. ML algorithms can autonomously extract pertinent information and make predictions based on data patterns and correlations, effectively identifying heart disease trends across large sets of medical, imaging, and genetic data [12, 13, 14]. This paper aims to provide an in-depth review of existing literature concerning heart disease detection through machine learning techniques, encompassing methodology, datasets, performance indicators, and the challenges involved in developing reliable prediction models.

AI applications are pivotal in promoting a decentralized rehabilitation model that integrates intelligent connected tools to enhance clinical decision-making and health outcome monitoring. Recent years have seen various AI-based methods developed to address the future demands of assisted physical therapy and patient assessments in a minimally supervised, decentralized context, particularly within home settings [15]. Nevertheless, to date, no comprehensive reviews exist on machine learning methods and applications specifically for remote rehabilitation monitoring and assistance. While some literature has explored the integration of machine learning algorithms with specific rehabilitation technologies, such as wearable sensors and vision-based motion capture systems, a thorough examination of the landscape is lacking [16].

In response to these limitations, the integration of machine learning (ML), deep learning (DL), and advanced optimization techniques has become a transformative force in predicting heart disease through the analysis of clinical and healthcare datasets [17]. The shift from manual diagnosis to data-driven methodologies has highlighted the efficacy of ML algorithms such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Deep Learning (DL). Furthermore, employing advanced techniques like ensemble learning, feature selection, and optimization algorithms [18] (including Particle Swarm Optimization and Genetic Algorithm) enhances the robustness and accuracy of predictions.

The recent availability of large-scale datasets, such as those from UCI and Kaggle, combined with enhancements in computational techniques like quantum-enhanced machine learning and hybrid deep learning have significantly boosted prediction performance [19]. Nevertheless, challenges remain in the field, including issues related to data imbalance, overfitting, high computational costs, and a lack of interpretability in models. Consequently, there is an urgent need for the development of an efficient, accurate, and scalable system to support healthcare professionals in early diagnosis and clinical decision-making, ultimately striving towards better management of heart disease [20, 21].

II. MOTIVATION

The research paper highlights the critical need for enhanced heart disease detection and diagnosis through advanced technologies like machine learning, given the significant morbidity and mortality associated with heart disease. Early identification of heart disease can improve patient outcomes and alleviate pressure on healthcare systems. However, the complexity and variability in heart disease presentations complicate accurate diagnosis. Machine learning techniques, proven effective in various domains including healthcare, show great potential in recognizing complex patterns within large datasets and making accurate predictions to improve detection. By utilizing automated decision-support systems, clinicians could efficiently process diverse patient data, yielding valuable insights for diagnosis and treatment planning.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

III. Literature Review

Table 1: This table presents that comparative study of heart disease prediction using supervised machine learning classification.

Author(s)	Pub. Year	Technique / Technology Used	Dataset Used	Performance Metrics	Parameters Value in Percentage	Limitations & Future Scope
Amrit Singh et al. [1]	2024	DT, RF, SVM, PCA	Healthcare dataset	Accuracy	91 %	Models suffer from overfitting and limited generalization on unseen datasets.
Jingyuan Yi et al. [2]	2025	Transformer + PSO	Medical dataset	Accuracy	96 %	Limited dataset diversity.
Hammoud et al. [3]	2024	LR, SVM, KNN, RF, GB	Cleveland, Statlog, Hungary	Accuracy, F1-score	94 %	Performance depends heavily on feature selection and dataset quality.
Al-Alshaikh et al. [4]	2024	GA + RFE + CNN	Clinical dataset	Accuracy, Precision, Recall	95 %	Deep CNN models are computationally expensive and difficult to interpret clinically.
El-Sofany et a. [6]	2024	XGBoost, SVM, RF, SMOTE	Public + private datasets	Accuracy, AUC	95 %	Requires balanced datasets and high preprocessing efforts.
Anjum et al. [7]	2024	LR, SVM, LightGBM, XGBoost	Clinical dataset	Accuracy, AUC	94 %	Limited evaluation on multiclass disease prediction.
Omkari & Shaik et al. [11]	2025	TLV Ensemble Model	UCI + Kaggle	Accuracy	93 %	Integrate big data analytics and cloud computing for scalable disease prediction systems.
Rao et al. [12]	2024	AttGRU + HMSI	Big healthcare dataset	Accuracy	95 %	Apply federated learning and edge computing
Gaikwad et al. [13]	2023	SVM, RF, DT, LR	UCI dataset	Accuracy	88 %	Limited dataset size.
Rahman et al. [15]	2024	Transformer (Self-attention)	UCI Cleveland	Accuracy	92 %	Limited dataset size, need for ensemble learning.

From the surveyed literature, XGBoost (Extreme Gradient Boosting) is the most suitable classifier for future heart disease prediction because it combines high predictive accuracy with strong generalization capability. Unlike conventional algorithms such as Decision Tree, Random Forest, and Support Vector Machine, XGBoost incorporates regularization techniques that reduce overfitting while improving prediction performance. It efficiently captures complex nonlinear relationships among clinical attributes and automatically handles missing values, making it well suited to healthcare datasets. Additionally, XGBoost provides feature importance scores, enabling clinicians and researchers to identify the most influential risk factors contributing to heart disease.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Several researchers have explored machine learning and deep learning techniques for heart disease prediction. Singh et al. (2024) proposed a supervised learning-based diagnostic system using Decision Tree, Random Forest, SVM, and PCA, addressing overfitting issues and achieving improved performance on both training and testing datasets. Yi et al. (2025) introduced a Transformer model optimized with Particle Swarm Optimization (PSO)**,** achieving an accuracy of 96.5%, outperforming traditional models such as Random Forest (92.2%). This demonstrates the effectiveness of optimization techniques in improving prediction models. Hammoud et al. (2024) conducted a comparative study of seven ML algorithms, including Logistic Regression, SVM, KNN, Random Forest, and Gradient Boosting. Their results showed that Random Forest achieved the highest accuracy (~94.96%) using cross-validation and feature selection methods. Al-Alshaikh et al. (2024) proposed a hybrid ML model combining Genetic Algorithm (GA), Recursive Feature Elimination (RFE), and deep CNN, achieving 95.5% accuracy and improved sensitivity and specificity, highlighting the importance of feature selection and optimization. El-Sofany (2024) evaluated multiple classifiers with feature selection techniques (Chi-square, ANOVA, MI) and SMOTE for data balancing. The study reported XGBoost as the best-performing model with 97.57% accuracy and strong evaluation metrics. Anjum et al. (2024) compared ML models such as Logistic Regression, SVM, LightGBM, and XGBoost, concluding that XGBoost achieved the highest accuracy (94.8%) and AUC, proving its effectiveness in clinical prediction tasks. Venkatesh Babu et al. (2024) explored quantum-enhanced machine learning (QuEML)**,** demonstrating improved accuracy and reduced training time compared to traditional ML methods, indicating future potential in healthcare analytics. Elmi et al. (2024) emphasized advanced feature selection techniques, improving model interpretability and reducing dimensionality, which enhances prediction performance in cardiovascular disease datasets. Omkari & Shaik (2024) proposed a Two-Layer Voting (TLV) ensemble model**,** achieving up to 99.03% accuracy using large datasets, showing the effectiveness of ensemble learning techniques. Rao et al. (2024) developed a hybrid deep learning model (AttGRU-HMSI) integrated with big data analytics and SMOTE, achieving 95.42% accuracy, demonstrating the role of deep learning in handling large healthcare datasets.

IV. Problem Statement

Heart disease is a leading cause of death globally, highlighting the need for accurate early prediction systems to enhance patient outcomes and lower healthcare expenses. Traditional diagnostic methods, which heavily depend on clinical judgement and manual evaluation, can result in delays, misdiagnoses, and heightened complication risks. While various machine learning and deep learning models for heart disease prediction have been introduced, several challenges remain unresolved. These include issues like overfitting, data imbalance, limited dataset sizes, lack of interpretability, and the high computational demands of advanced models such as deep learning and transformer-based approaches. Furthermore, many existing studies concentrate on single algorithms without effectively leveraging optimization techniques or ensemble methods to boost performance. Thus, there is a pressing need for the development of a robust heart disease prediction model that integrates machine learning, feature selection, and optimization strategies to enhance accuracy, decrease computational costs, and facilitate real-time clinical decision-making.

V. Machine Learning Techniques

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that enables computer systems to learn patterns from data and make predictions or decisions without being explicitly programmed. In the healthcare sector, machine learning techniques are widely used for disease prediction, diagnosis, medical image analysis, and decision support systems. These techniques analyze large amounts of patient data to identify hidden patterns and improve the accuracy of disease prediction models [4]. Different machine learning algorithms are used depending on the type of dataset, prediction requirements, and computational complexity.

Decision Tree (DT)

Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It works by dividing the dataset into smaller subsets based on feature values, forming a tree-like structure consisting of nodes and branches [1, 3]. Each internal node represents a decision based on an attribute, while leaf nodes represent the final output or class label. Decision Trees are easy to understand and interpret, making them highly useful in medical

ISSN : 2455-9679
SJIF Impact Factor : 6.190

diagnosis systems. They can handle both numerical and categorical data efficiently. However, Decision Trees are prone to overfitting when trained on complex datasets.

Random Forest (RF)

Random Forest is an ensemble learning technique that combines multiple Decision Trees to improve prediction accuracy and reduce overfitting. Each tree is trained on a random subset of the dataset, and the final prediction is obtained through majority voting or averaging [4, 6]. Random Forest is widely used in healthcare applications because of its robustness, high accuracy, and ability to handle large datasets. It can identify important features contributing to disease prediction.

Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm primarily used for classification tasks. SVM works by finding an optimal hyperplane that separates data points into different classes with the maximum margin.

In heart disease prediction, SVM is effective in handling high-dimensional medical datasets and provides good classification performance [7]. Kernel functions such as linear, polynomial, and radial basis function (RBF) can be used to improve classification accuracy.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a supervised learning algorithm that classifies a data point based on the majority class among its nearest neighbors. The similarity between data points is usually measured using distance metrics such as Euclidean distance [6, 7]. KNN is simple and effective for small datasets but becomes computationally expensive for large-scale medical datasets.

XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is one of the most powerful supervised machine learning algorithms for classification and regression tasks. It is an optimized implementation of the Gradient Boosting algorithm that builds an ensemble of decision trees sequentially, where each new tree corrects the errors made by the previous trees. XGBoost is widely used in healthcare applications, including heart disease prediction, because of its high accuracy, robustness, and ability to handle complex relationships in medical data.The proposed model employs the XGBoost (Extreme Gradient Boosting) classifier as the core prediction algorithm for heart disease diagnosis. Initially, the clinical dataset undergoes preprocessing, including handling missing values, encoding categorical attributes, normalization, and removal of noisy records. To improve model performance, feature selection techniques such as Recursive Feature Elimination (RFE), correlation analysis, or mutual information can be applied to retain only the most informative clinical features. If class imbalance exists, techniques such as SMOTE may be used to generate balanced training samples.

The processed dataset is then divided into training and testing subsets, and the XGBoost classifier is trained using optimized hyperparameters obtained through Grid Search or Randomized Search with cross-validation. Finally, the trained model is evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. Because of its gradient boosting framework, built-in regularization, and efficient tree-pruning strategy, the proposed XGBoost model is expected to provide superior prediction accuracy, improved generalization, and enhanced robustness compared with traditional classifiers such as Decision Tree, Random Forest, SVM, and conventional Gradient Boosting.

VI. CONCLUSION

The human body comprises various organs, with the heart being essential for pumping blood throughout the body, indicating that human life relies on its proper function. A malfunctioning heart can impact other organs, including the mind and kidneys. Meanwhile, technological advancements, particularly in the healthcare sector, have introduced artificial intelligence (AI) techniques, including machine learning (ML) and deep learning (DL). These AI approaches enhance disease diagnosis systems for patients, yielding results that are often more accurate and efficient compared to traditional methods. This research work presents comparative study for heart disease using machine learning and deep learning techniques, in future work implement an efficient system for heart disease prediction using machine learning algorithm and enhance the performance of an existing system techniques.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

REFERENCES

[1] A. Singh, H. Mahapatra, A. K. Biswal, M. Mahapatra, D. Singh, and M. Samantaray, "Heart Disease Detection Using Machine Learning Models," Procedia Computer Science, vol. 235, pp. 937–947, 2024.

[2] J. Yi, P. Yu, T. Huang, and Z. Xu, "Optimization of Transformer Heart Disease Prediction Model Based on Particle Swarm Optimization Algorithm," arXiv preprint arXiv:2412.02801, 2025.

[3] A. Hammoud, A. Karaki, R. Tafreshi, S. Abdulla, and M. Wahid, "Coronary Heart Disease Prediction: A Comparative Study of Machine Learning Algorithms," 2024.

[4] H. A. Al-Alshaikh et al., "Comprehensive Evaluation and Performance Analysis of Machine Learning in Heart Disease Prediction," Scientific Reports, vol. 14, 2024.

[5] H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques," IEEE Access, vol. 12, 2024.

[6] N. Anjum et al., "Improving Cardiovascular Disease Prediction through Comparative Analysis of Machine Learning Models," Journal of Computer Science and Technology Studies, vol. 6, no. 2, pp. 62–70, 2024.

[7] S. Venkatesh Babu, P. Ramya, and J. Gracewell, "Revolutionizing Heart Disease Prediction with Quantum-Enhanced Machine Learning," Scientific Reports, vol. 14, 2024.

[8] A. H. Elmi, A. Abdullahi, and M. A. Barre, "A Machine Learning Approach to Cardiovascular Disease Prediction with Advanced Feature Selection," Indonesian Journal of Electrical Engineering and Computer Science, vol. 33, no. 2, 2024.

[9] N. Seth, H. Kaushik, K. Sharma, and H. Sharma, "A Comparative Analysis for Genetic Disease Detection Accuracy Through Machine Learning Models on Datasets," International Journal of Enhanced Research in Management & Computer Applications, vol. 13, no. 8, 2024.

[10] S. Roobini, M. S. Kavitha, and S. Karthik, "A Systematic Review on Machine Learning and Neural Network Based Models for Disease Prediction," Journal of Integrated Science and Technology, vol. 12, no. 4, 2024.

[11] B. Duraisamy, R. Sunku, K. Selvaraja, V. V. R. Pilla, and M. Sanikala, "Heart Disease Prediction Using Support Vector Machine," Multidisciplinary Science Journal, 2024.

[12] A. U. Rahman et al., "Enhancing Heart Disease Prediction Using a Self-Attention-Based Transformer Model," Scientific Reports, vol. 14, 2024.

[13] D. Y. Omkari and K. Shaik, "An Integrated Two-Layered Voting (TLV) Framework for Coronary Artery Disease Prediction Using Machine Learning Classifiers," IEEE Access, 2024.

[14] G. M. Rao et al., "AttGRU-HMSI: Enhancing Heart Disease Diagnosis Using Hybrid Deep Learning Approach," Scientific Reports, vol. 14, 2024.

[15] Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators", Appl. Sci. 2021, pp. 1-22.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

[16] Ashir Javeed, Muhammad Asim Saleem, "Decision Support System for Predicting Mortality in Cardiac Patients Based on Machine Learning", Appl. Sci. 2023, pp. 1-14.

[17] Vigneshwaran Namasivayam, Nithyashree Senguttuvan, "Artificial Intelligence and Its Application in Cardiovascular Disease Management", 2022, pp. 189-207.

[18] Rustem Yilmaz, Fatma Hilal Yagın, "Early Detection of Coronary Heart Disease Based on Machine Learning Methods", MEDICAL RECORDS-International Medical Journal, 2022, pp. 1-6.

[19] Peng Wang, Zihuai Lin, Xucun Yan, Zijiao Chen, Ming Ding, Yang Song, Lu Meng, "A Wearable ECG Monitor for Deep Learning Based Real-Time Cardiovascular Disease Detection", IEEE 2022, pp. 1-9.

[20] Sandeep Chandra Bollepalli, Ashish Kumar Sahani, "An Optimized Machine Learning Model Accurately Predicts In-Hospital Outcomes at Admission to a Cardiac Unit", Diagnostics 2022, pp. 1-14.

[21] Jasjit S. Suri, Mrinalini Bhagawati, "Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review", Computers in Biology and Medicine, 2022, pp. 1-24.