Comprehensive Review on Heart Disease Prediction Using Machine Learning Techniques

Veenah Ganesh1, a) and Syed Qamrun Nisa1, 2, b)

1*Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63100, Malaysia.*

2*Centre for Image and Vision Computing (CIVC), CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Malaysia.*

b) Corresponding author: qamrunnisa@mmu.edu.my

a) 1231300941@student.mmu.edu.my

**Abstract.** Heart disease is a leading cause of mortality worldwide, driving extensive research into early detection methods using data-driven techniques. This paper presents a comprehensive literature review of machine learning (ML) approaches applied to heart disease prediction. Studies from the past five years have been examined, covering traditional algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Decision Trees, as well as advanced methods including Gradient Boosting, XGBoost, AdaBoost, and Neural Networks. These models have been trained and validated on datasets from the UCI Machine Learning Repository, Kaggle, clinical records, and IoT-integrated systems. Preprocessing techniques such as data cleaning, encoding, feature selection, and Synthetic Minority Oversampling Technique (SMOTE) are commonly used to improve model performance. Evaluation metrics typically include accuracy, precision, recall, F1-score, and AUC-ROC. The review highlights the strengths and limitations of each model and discusses their applicability in real-world healthcare settings. While ensemble and deep learning models often demonstrate superior predictive capabilities, they may pose challenges in terms of interpretability and computational demands. This study provides insights into effective ML strategies for cardiovascular risk prediction and outlines areas for future research and clinical integration.

# INTRODUCTION

Heart disease is a serious worldwide problem, affecting millions of people and remaining the leading cause of death in the United States. It impacts individuals across all genders and across nearly all racial and ethnic groups, with one individual losing their life to cardiovascular disease complications every 33 seconds [1]. In 2022, it was responsible for 702,880 deaths in the U.S., accounting for nearly 20 percent of all deaths [2]. Worldwide, cardiovascular diseases caused about 17.9 million deaths in 2019, which made up nearly one-third of all global fatalities 2. In the same year, coronary heart disease alone led to over 371,000 deaths in the U.S. [3]. Cardiovascular disease (CVD) is economically very costly for health care systems. In the United States, $252.2 billion (countries in 2019–2020 was spent on health services, medication and lost productivity due to heart disease alone [3]. The problem is especially acute in low and middle-income countries where access to appropriate healthcare is not widespread, with over 75% of cardiovascular-related deaths occurring globally [3]. Preventing these major lifestyle risk factors, including smoking, poor diet, physical inactivity, and alcohol use, is crucial for preventing heart disease [3].

Machine learning can be used to make doctors' assessments more accurate and personalized by analyzing data from medical records, patient behaviour and diagnostic results in recent years. It may also have the potential to achieve the goal of accessible health in rural or resource-limited regions due to its low cost and scalability. According to IBM, the more data they receive, the better these systems become, as they adapt their predictions in light of new evidence and learn through various optimisation strategies [4]. This flexibility make machine learning become a useful tool in assisting real-time clinical decisions.

The aim is to evaluate the performance of these models across different populations and explore trade-offs between accuracy, complexity, and the extent to which they can be implemented in real-world clinical settings.

# LITERATURE REVIEW

In this review, we examine the use of machine learning in the prediction of heart disease that remains one of the most common causes of death globally. It’s centered on research from the past five years that has employed data from trustworthy sources including the UCI Machine Learning Repository, Kaggle and actual clinical records. It contains both simpler models like Logistic Regression, Random Forests and more advanced ones like Gradient Boosting Machines and Neural Networks. The goal is to determine how well these models work in varied circumstances and to better understand the trade-offs of accuracy, complexity and practical application in real-world healthcare settings. The aim is to evaluate the performance of these models in diverse settings and to gain insight into trade-offs between accuracy, complexity and utility in actual health systems.

## Dataset-Based Overview

### UCI Repository Datasets

Researchers often turn to the dataset collection hosted by UCI to experiment with and validate machine learning techniques. It includes several heart disease datasets, such as Cleveland, Hungarian, and Statlog, which contain essential clinical information often used in cardiovascular risk analysis. These datasets are popular in research because they are well-structured, easy to interpret, and suitable for comparing model performance. Ali et al. [5] combined the Cleveland and Hungarian datasets to evaluate various machine learning models, with Logistic Regression and SVM showing strong results. Chethana [6] explored different KNN configurations on the same data and consistently achieved reliable predictions. Kowsalya and Lavanya [7] applied LSTM and GRU models, demonstrating that deep learning techniques can also perform well with structured medical datasets.

### Kaggle Datasets

Kaggle offers a broad selection of open datasets, many of which are frequently used in both research and real-world machine learning projects. Several of its heart disease datasets are based on the Cleveland dataset, though they often come with added features or have been preprocessed in different ways. Because they’re easy to access and ready for experimentation, these datasets are commonly used to evaluate various classification techniques. Jindal et al. [8] experimented with models like KNN, decision trees, and logistic-based approaches, reaching accuracy levels close to 88%. In their study, Kumar et al. [9] applied ensemble methods alongside neural networks to similar datasets and observed that the models handled different prediction tasks well without major drops in accuracy.

### Clinical/Hospital Datasets

Data sourced from hospitals and clinics offers valuable insights drawn from real patient interactions. Since this data comes directly from hospitals and clinics, it reflects a wide variety of health conditions and patient backgrounds, which helps researchers test how well machine learning models perform in real-world situations. However, using clinical data isn’t always straightforward, as privacy regulations and the need to clean and prepare the data properly can make the process more complex. For example, Maini et al. [10] applied Random Forest and Gradient Boosting on over 1,600 hospital records, achieving 93.8% accuracy in predicting heart disease. Similarly, Sekhar et al.  [11] leveraged clinical data to train deep learning models like CNN and LSTM, highlighting their usefulness in medical diagnostics.

### IoT-Integrated and Custom Systems

People are now using smartwatches, health apps, and in-home sensors along with machine learning to better monitor their heart health. These tools track things like heart rate and activity levels throughout the day, making them especially useful for those who can’t easily access medical care or need to keep an eye on existing conditions from home. Still, these systems aren’t perfect cause problems like faulty devices or unreliable data can affect their performance. For example, Senthil et al. [12] created a mobile system that gathers health data to assist in predicting heart disease. In another study, researchers integrated supervised learning models with relational databases to enhance early risk detection in personalized healthcare settings [13].

## Model-Based Overview

### K-Nearest Neighbors (KNN)

The K-Nearest Neighbors method predicts a result by comparing it to nearby points. It finds the ‘k’ closest examples and chooses the most frequent label for classification or the average for regression. KNN is non-parametric and does not require model training, but it is sensitive to feature scaling and can be computationally expensive for large datasets. Harshit Jindal, Sarthak Agrawal, Rishabh Khera1, Rachna Jain and Preeti [8] applied the K-Nearest Neighbors (KNN) algorithm to Kaggle’s cardiac dataset and achieved 88.52% accuracy. Chethana C [6], explored different KNN configurations using the UCI dataset and reported stable performance across models. Ekta Maini et al. [10] also included KNN among other models in their clinical dataset study, validating its usability despite sensitivity to feature scaling. Abhishek Gupta, Vansh Misra, Ketan Chauhan, Kumar Manoj [14] included KNN as one of several models in their performance evaluation for heart disease detection, confirming its feasibility for small-to-mid-sized datasets. Narendra Mohan, Vinod Jain, Gauranshi Agrawal 13 also evaluated KNN as part of their comparative study of supervised models for heart disease.

### Random Forest

Random Forest is built by training multiple decision trees and then blending their predictions. This technique helps improve performance and makes the model more robust, especially when working with large datasets. Ekta Maini et al. [10] used Random Forest on 1670 hospital records and achieved a diagnostic accuracy of 93.8%. Harshit Jindal et al.8 also applied Random Forest with high accuracy on Kaggle data. Sunitha Guruprasad et al. [15] have surpassed Decision Tree in performance and generalizability. Mamta et al. [16] implemented a hybrid Random Forest-XGBoost framework that improved the classification of heart disease symptoms. Aman Solanki, Anand Vardhan, Aman Jharwal, Prof. Narender Kumar [17] also adopted Random Forest in their multi-model approach, demonstrating consistent accuracy across categorical and numerical features. Narendra Mohan, Vinod Jain, Gauranshi Agrawal [13] reported favorable performance using Random Forest in combination with other supervised learning methods.

### Logistic Regression

Logistic Regression is a statistical model used for binary classification problems. It estimates the probability of a binary response based on one or more independent variables using the logistic function. Although simple, Logistic Regression is effective and interpretable, especially in linearly separable datasets. Farman Ali et al. [5] achieved 92.2% accuracy using Logistic Regression on a combined Cleveland-Hungarian dataset. Harshit Jindal1, Sarthak Agrawal1, Rishabh Khera1, Rachna Jain2 and Preeti 8 applied Logistic Regression on Kaggle data, demonstrating its robustness. Ekta Maini et al. [10] also used it as a benchmark model in their clinical data analysis.

### Decision Tree

Supervised learning approaches like Decision Trees classify data by applying a series of conditional rules structured in a hierarchical tree format. Internal nodes represent conditions or tests, while the leaves show the final outcomes. Although Decision Trees are quick to build and simple to understand, they can be prone to overfitting and may not handle small variations in the data well. Guruprasad et al. [15] applied a Decision Tree classifier on the UCI dataset and reported 87.3% accuracy. Jindal et al. [8] also used Decision Tree but found it less robust compared to ensemble methods like Random Forest. Nagaraj M. Lutimath [6] applied Decision Tree in their comparative study and noted its ease of use but limitations in scalability. Gupta et al. [14] also implemented Decision Tree in their comparison and observed satisfactory interpretability outcomes. Sridharan Kannan [18] implemented Decision Tree and Support Vector Machine within an optimized clinical decision support framework, demonstrating effective classification outcomes.

### Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes’ Theorem with a strong assumption of feature independence. It is simple, fast, and works well with high-dimensional data, though its performance may decline when the independence assumption is violated Ekta Maini et al. [10] included Naive Bayes in a set of classifiers applied to clinical data, finding it efficient but less accurate. Sunitha et al. [15] reported improved results when combining Naive Bayes with SVM. Chethana C [6] evaluated Naive Bayes as part of a comparative study and noted its simplicity and quick training speed. Narendra Mohan, Vinod Jain, Gauranshi Agrawal [13] included Naive Bayes in a supervised ML comparison and observed solid baseline performance, though it trailed ensemble models in accuracy.

### Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classifier that constructs a hyperplane to separate data points in high-dimensional space. It works when the data is linearly and non-linearly separable using kernel function to be mapped to higher dimensions. Farman Ali et al. [5] applied Support Vector Machine (SVM) on a merged Cleveland dataset and achieved 84.4% accuracy. Sunitha et al. [15] experimented with an SVM-Naive Bayes hybrid and reported improved accuracy on the UCI dataset. Nagaraj M. Lutimath [19] also included SVM and emphasized its performance on high-dimensional datasets. Abhishek Gupta, Vansh Misra, Ketan Chauhan, Kumar Manoj [14] employed Support Vector Machine alongside Decision Tree and Logistic Regression to evaluate comparative efficiency on hospital-based datasets. Sridharan Kannan [18] also incorporated SVM in their optimization framework for heart disease prediction.

### Gradient Boosting

Gradient Boosting involves building a series of models where each one improves upon the errors of the last, using gradient-based optimization. Although effective in boosting accuracy, it can be slower due to its iterative nature. Ekta Maini et al. [10] implemented Gradient Boosting and observed competitive accuracy. S. Sasipriya, K Bala Nimisha, R Anusha, G Arun Pandiyan [20] applied Gradient Boosting in a hybrid ensemble method on UCI data and reported superior performance. Senthil G. A [12] integrated Gradient Boosting in a real-time prediction system using patient inputs. Mamta Gagoriya. Mukesh Kumar Khandelwal [16] noted that combining Gradient Boosting with Random Forest resulted in higher diagnostic reliability in their study. In the paper "Heart Disease Prediction Using Ensemble Model" [17] the authors employed a voting-based ensemble approach combining Gradient Boosting with other classifiers, enhancing prediction accuracy. Paul T Sheeba, Deepjyoti Roy, Mohammad Haider Syed [21] used metaheuristic-enabled ensemble learning with Gradient Boosting, demonstrating enhanced classification effectiveness.

### Neural Network

Neural networks represent a branch of deep learning that mimics certain brain-like processes by using layers of artificial neurons to interpret data patterns. Advanced versions such as LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are particularly suited for analyzing time-dependent information. These models can capture complex, sequential trends but often demand high computational power and large datasets to perform optimally. In one project, LSTM and GRU were used with the UCI heart disease dataset and achieved close to 90% accuracy [7]. Clinical datasets have also shown that while neural networks can deliver strong predictions, training them often requires more time, data, and computing power [11]. In some studies, these models were used as part of a larger ensemble approach, helping improve the accuracy of heart disease diagnosis [9].

### Linear Regression

Linear Regression is among the earliest techniques developed in statistical modeling and remains widely used for identifying trends between variables. In machine learning, although it is primarily applied to estimate continuous outcomes, it is occasionally repurposed as a baseline classifier to compare against more complex algorithms. The approach relies on the assumption that there is a straight-line relationship between input variables and the predicted target. In a comparative study by Olawade, et al. [22], Linear Regression was employed to provide a reference point for evaluating the performance of more sophisticated heart disease prediction models. While its simplicity makes it easy to interpret, the method showed limitations when applied to datasets with intricate or non-linear patterns.

### XGBoost

XGBoost is an algorithm created as part of machine learning that combines many small decision trees to create better predictions. It’s faster and more accurate than plain boosting because it makes use of techniques like regularization and parallel processing. These characteristics allow for the management of very large and fine datasets without overfitting. Gagoriya et al. [16] have been able to predict heart disease more accurately using XGBoost along with Random Forest. The combination of the two methods allowed the model to locate complicated patterns in the data. Another team included XGBoost into a battery of models they tested on public data-set such as those from UCI. They had better results which proves the XGBoost can improve other models accuracy.

### AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble learning method that uses a series of simple models (mostly small decision trees in this case) to make better predictions. It puts more emphasis on the data points we classified incorrectly in previous rounds so that subsequent models can learn from those mistakes. This is because it helps the system as a whole to become more stable and accurate in the long run. In a study by Sheeba et al. [21], AdaBoost was incorporated into a larger model that drew upon optimization methods, and contributed significantly to promoting predictability of heart disease prediction. Together with other elementary models, mainly decision trees, they significantly advanced the algorithm work well on other forms of data. Another study that utilized an ensemble for the prediction of heart disease highlighted AdaBoost as a key part of the approach to improve classification performance for multiple clinical markers.

## COMPARATIVE ANALYSIS OF REVIEWED STUDIES

To make sense of the diverse results reported in the literature, we organized a comparative summary that captures how different ML models performed across various datasets. This helps clarify which techniques show strong results and under what conditions they are most effective. It also uncovers recurring limitations and areas that have yet to be adequately addressed, providing a roadmap for future research directions. Table 1 shows a comparison of machine learning methods for heart disease prediction.

# CONCLUSION

This research examined the application of multiple machine learning models for predicting heart disease, using data from sources including the UCI repository, Kaggle datasets, hospital records, and IoT-enabled systems. Techniques such as Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and Neural Networks were assessed based on their accuracy and real-world feasibility. Advanced ensemble and deep learning methods achieved high performance, though they often demand significant computing power and can be difficult to interpret. Preprocessing steps like SMOTE and effective feature selection contributed to improved predictive results. Some of these models were also incorporated into web platforms for real-time use. However, challenges remain in adapting these tools across different healthcare environments. Moving forward, greater focus should be placed on improving model generalization, ease of use, and seamless integration into clinical workflows to aid early diagnosis and lower heart disease-related deaths.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TABLE 1.** Comparison of machine learning methods for heart disease prediction | | | | | |
| **Authors** | **Model Used** | **Accuracy** | **Dataset Used** | **Weakness/Improvements** | **Study Gaps** |
| [20] | Artificial Neural Network (ANN) | 90% | Publicly available datasets | ANN requires significant computational resources; lacks model generalization analysis. | Need for real-time application testing and wider dataset inclusion. |
| [18] | AdaBoost, SMOTE | 95% | Statlog dataset | Requires fine-tuning for highly imbalanced data. | Further testing required on real-world noisy datasets. |
| [19] | Random Forest | 98% | UCI ML repository | Limited exploration of advanced hyperparameter tuning; lacks interpretability for clinical use. | Needs broader validation with real-world clinical data. |
| [11] | KNN, Random Forest | 96% | Public datasets | Limited real-world testing of the application; lacks longitudinal evaluation. | More studies on user experience and healthcare practitioner usability. |
| [14] | Decision Tree, Random Forest, Naïve Bayes | 95.08% | UCI ML repository | Limited focus on expanding features beyond standard attributes. | Need to test across varied real-world settings for robustness. |
| [7] | Hybrid DPA-RNN+LSTM | 99.20% | UCI and clinical datasets | High computational demands; limited testing in live healthcare environments. | Scalability and real-time applicability not fully explored. |
| [9] | ANN, SVM, Logistic Regression | 97% | Indian clinical datasets | Lacks thorough analysis of interpretability and explainability of deep learning outputs. | More work needed on bias reduction in predictions. |

## REFERENCES

1. “Heart Disease Facts | Heart Disease | CDC,” (n.d.).
2. “Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report - World Heart Federation,” (n.d.).
3. “Cardiovascular diseases (CVDs),” (n.d.).
4. “What Is Machine Learning (ML)? | IBM,” (n.d.).
5. F. Ali, S. El-Sappagh, S.M.R. Islam, D. Kwak, A. Ali, M. Imran, and K.S. Kwak, “A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion,” Information Fusion **63**, 208–222 (2020).
6. C. Chethana, “Prediction of heart disease using different KNN classifier,” in *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, (Institute of Electrical and Electronics Engineers Inc., 2021), pp. 1186–1194.
7. V. Baviskar, M. Verma, P. Chatterjee, and G. Singal, “Efficient Heart Disease Prediction Using Hybrid Deep Learning Classification Models,” IRBM **44**(5), (2023).
8. H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” in *IOP Conf Ser Mater Sci Eng*, (IOP Publishing Ltd, 2021).
9. R. Kumar, K. Divya, and S. Verma, “Heart Disease Diagnosis using Deep Learning,” in *2024 IEEE 1st Karachi Section Humanitarian Technology Conference, Khi-HTC 2024*, (Institute of Electrical and Electronics Engineers Inc., 2024).
10. E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, “Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India,” Med J Armed Forces India **77**(3), 302–311 (2021).
11. A. Sekhar, A. Babu, V. K. Jayalekshmi, and A. Udayan, “Machine Learning Based Heart Disease Prediction,” in *Proceedings - 2nd International Conference on Next Generation Intelligent Systems, ICNGIS 2022*, (Institute of Electrical and Electronics Engineers Inc., 2022).
12. G. A. Senthil, R. Prabha, M. Razmah, T. Veeramakali, S. Sridevi, and R. Yashini, “Machine Learning Heart Disease Prediction Using KNN and RTC Algorithm,” in *3rd International Conference on Power, Energy, Control and Transmission Systems, ICPECTS 2022 - Proceedings*, (Institute of Electrical and Electronics Engineers Inc., 2022).
13. N. Mohan, V. Jain, and G. Agrawal, “Heart Disease Prediction Using Supervised Machine Learning Algorithms,” in *2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021*, (Institute of Electrical and Electronics Engineers Inc., 2021).
14. A. Gupta, V. Misra, K. Chauhan, and K. Manoj, “Heart Disease Prediction Using Machine Learning,” in *Proceedings - IEEE 2023 5th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2023*, (Institute of Electrical and Electronics Engineers Inc., 2023), pp. 108–112.
15. S. Guruprasad, V. L. Mathias, and W. Dcunha, “Heart Disease Prediction Using Machine Learning Techniques,” in *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, ICEECCOT 2021 - Proceedings*, (Institute of Electrical and Electronics Engineers Inc., 2021), pp. 762–766.
16. M. Gagoriya, and M. K. Khandelwal, “Heart Disease Prediction Analysis Using Hybrid Machine Learning Approach,” in *Proceedings of the International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2023*, (Institute of Electrical and Electronics Engineers Inc., 2023), pp. 896–899.\
17. A. Solanki, A. Vardhan, A. Jharwal, and I. N. Kumar, “Heart Diseases Prediction Using Machine Learning,” in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, (Institute of Electrical and Electronics Engineers Inc., 2023).
18. S. Kannan, “Modelling an Efficient Clinical Decision Support System for Heart Disease Prediction Using Learning and Optimization Approaches,” CMES - Computer Modeling in Engineering and Sciences **131**(2), 677–694 (2022).
19. N. M. Lutimath, N. Sharma, and B. K. Byregowda, “Prediction of Heart Disease using Random Forest,” in *2021 IEEE International Conference on Emerging Trends in Industry 4.0, ETI 4.0 2021*, (Institute of Electrical and Electronics Engineers Inc., 2021).
20. S. Sasipriya, K. Bala Nimisha, R. Anusha, and G. Arun Pandiyan, “Heart Disease Prediction Using Machine Learning Algorithm,” in *Proceedings of the 5th International Conference on Inventive Research in Computing Applications, ICIRCA 2023*, (Institute of Electrical and Electronics Engineers Inc., 2023), pp. 66–69.
21. P. T. Sheeba, D. Roy, and M. H. Syed, “A metaheuristic-enabled training system for ensemble classification technique for heart disease prediction,” Advances in Engineering Software **174**, (2022).
22. Olawade, David B., et al. "Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection." *International Journal of Cardiology* (2025): 133443.