© OCT 2026 | IRE Journals | Volume 8 Issue 4 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I4-1719791

Feature-Optimised Heart Disease Prediction Using Machine Learning Techniques

SUNITA SHARMA[1], TANISHKA[2], RAHUL[3], AVIRAL JAIN[4]
[1, 2, 3, 4]*Department of CSE-Artificial Intelligence and Machine Learning, Panipat Institute of Engineering and Technology, Samalkha, Panipat*

Abstract-* *Heart disease is a major cause of mortality worldwide, making early and accurate prediction essential for effective clinical intervention. This paper presents a feature-optimized heart disease prediction framework using machine learning techniques to improve diagnostic performance. The proposed approach incorporates data preprocessing, feature optimization, and classification to identify the most relevant clinical attributes while reducing redundancy. Several machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Experimental results demonstrate that feature optimization enhances prediction accuracy and model efficiency, with ensemble-based methods achieving superior performance. The proposed framework offers an effective decision-support tool for the early detection of heart disease and has the potential to assist healthcare professionals in improving diagnostic accuracy and patient care.

Keywords— Heart Disease Prediction, Machine Learning, Feature Optimization, Feature Selection, Classification, Random Forest, Support Vector Machine (SVM), Xgboost, Clinical Decision Support.

I. INTRODUCTION

Heart disease is one of the leading causes of mortality worldwide and continues to pose a major public health challenge. According to the World Health Organization (WHO), cardiovascular diseases account for nearly 18 million deaths each year, emphasizing the need for early diagnosis and effective preventive strategies [1]. Early detection of heart disease enables timely clinical intervention, reduces healthcare costs, and improves patient survival rates. However, conventional diagnostic approaches often rely on extensive clinical examinations, laboratory investigations, and physician expertise, making the diagnostic process time-consuming and susceptible to variability [2].

The rapid advancement of machine learning (ML) has significantly improved the development of intelligent clinical decision-support systems. By learning complex patterns from patient data, ML algorithms can accurately classify individuals at risk of developing heart disease. Several supervised learning techniques, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and Extreme Gradient Boosting (XGBoost), have demonstrated promising performance in disease prediction tasks [3], [4]. These models assist clinicians by providing accurate and consistent predictions based on historical clinical records.

Despite these advancements, the predictive performance of ML models is often affected by the presence of redundant, irrelevant, or highly correlated features within medical datasets. Such features increase computational complexity, reduce model interpretability, and may lead to overfitting [5]. Feature optimization has therefore become an essential preprocessing step in medical data analysis. By selecting the most informative clinical attributes, feature optimization enhances classification accuracy, reduces training time, and improves the overall efficiency of prediction models [6].

Several researchers have investigated feature selection and optimization techniques for heart disease prediction. Statistical methods, wrapper-based approaches, embedded techniques, and metaheuristic optimization algorithms have been employed to identify significant risk factors from clinical datasets [7], [8]. These optimized feature subsets have consistently demonstrated improved predictive performance compared with models trained using the complete feature set. Moreover, ensemble learning methods such as Random Forest and XGBoost have shown superior robustness and generalization capabilities in handling complex healthcare datasets [9].

In this paper, a feature-optimized heart disease prediction framework using machine learning techniques is proposed. The framework integrates data preprocessing, feature optimization, and comparative evaluation of multiple supervised classification algorithms to improve prediction accuracy while minimizing computational complexity. The proposed models are evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to identify the most effective classifier for heart disease prediction [10]. The proposed framework aims to provide an efficient and reliable decision-support tool that assists healthcare professionals in the early detection of heart disease and supports timely clinical decision-making.

- The major contributions of this study are summarized as follows:
- Development of a feature-optimized framework for heart disease prediction using machine learning techniques.
- Comparative analysis of multiple supervised classification algorithms using optimized clinical features.
- Performance evaluation based on standard classification metrics to determine the most effective prediction model.
- A scalable and efficient decision-support approach that can assist healthcare professionals in improving the accuracy of early heart disease diagnosis.

II. RELATED WORK

Machine learning has become an essential tool in healthcare for predicting cardiovascular diseases by analyzing patient records and identifying complex relationships among clinical parameters. Over the past decade, numerous studies have investigated the effectiveness of supervised learning algorithms for heart disease prediction, demonstrating significant improvements in diagnostic accuracy and decision support.

Sultana *et al.* [11] proposed a machine learning-based framework for heart disease prediction and compared multiple classification algorithms. Their study showed that ensemble learning methods, particularly Random Forest, achieved higher prediction accuracy due to their robustness in handling complex clinical datasets. Khan *et al.* [12] investigated the impact of feature selection techniques on heart disease prediction using supervised machine learning algorithms. Their findings indicated that selecting the most relevant clinical attributes significantly improved classification accuracy while reducing computational complexity.

Singh and Kumar [13] evaluated several machine learning classifiers, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN), using the UCI heart disease dataset. The study concluded that Support Vector Machine and Random Forest consistently outperformed traditional classifiers in terms of prediction performance.

Ahmed *et al.* [14] developed a hybrid heart disease prediction model by integrating feature optimization with ensemble learning techniques. Experimental results demonstrated that optimized feature subsets enhanced model generalization and reduced overfitting, resulting in improved diagnostic accuracy. Patel and Shah [15] presented a comparative analysis of feature selection algorithms for cardiovascular disease prediction. Their research highlighted that eliminating redundant and irrelevant features improved both model efficiency and interpretability while maintaining high predictive performance.

Recent advances have also focused on explainable artificial intelligence (XAI) to improve clinician confidence in machine learning-based diagnosis. Lee *et al.* [16] incorporated explainable learning techniques into heart disease prediction models, enabling clinicians to interpret feature importance and understand the reasoning behind prediction outcomes. Such approaches enhance the practical applicability of machine learning in healthcare environments.

Although previous studies have reported encouraging prediction performance, many of them either utilize the complete feature set or evaluate only a limited number of machine learning algorithms. Furthermore, comprehensive comparisons between feature-optimized models and conventional approaches remain relatively limited. These gaps motivate the development of a robust framework that combines efficient feature optimization with multiple supervised learning algorithms for accurate and reliable heart disease prediction.

The proposed study addresses these challenges by integrating data preprocessing, feature optimization, and comparative evaluation of multiple machine learning classifiers. The framework aims to improve prediction accuracy, reduce computational complexity, and provide an effective clinical decision-support system for the early detection of heart disease.

III. PROPOSED METHODOLOGY

The proposed framework aims to improve the accuracy and efficiency of heart disease prediction by integrating data preprocessing, feature optimization, and machine learning classification. The overall workflow consists of six major stages: dataset acquisition, data preprocessing, feature optimization, model training, performance evaluation, and prediction. Figure 1 illustrates the architecture of the proposed heart disease prediction framework.

A. Dataset Description

The proposed model utilizes the Cleveland Heart Disease dataset obtained from the UCI Machine Learning Repository, one of the most widely used benchmark datasets for cardiovascular disease prediction [17]. The dataset contains 303 patient records with 14 clinical attributes, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression, slope of the ST segment, number of major vessels, thalassemia, and the target class indicating the presence or absence of heart disease. The diversity of clinical features makes the dataset suitable for evaluating machine learning algorithms in medical diagnosis.

B. Data Preprocessing

Data preprocessing is performed to improve data quality and ensure reliable model training. Missing values, duplicate records, and inconsistencies are identified and appropriately handled. Numerical attributes are normalized using Min-Max Scaling to ensure uniform feature ranges, while categorical variables are encoded into numerical representations suitable for machine learning algorithms [18]. Data preprocessing reduces noise and improves model convergence, leading to better classification performance.

C. Feature Optimization

Feature optimization is employed to identify the most informative clinical attributes while eliminating redundant and irrelevant features. In this work, Recursive Feature Elimination (RFE) is adopted because it iteratively ranks features based on their contribution to the prediction model and removes less significant attributes until an optimal subset is obtained [19]. Optimizing the feature set reduces computational complexity, minimizes overfitting, and enhances the generalization capability of machine learning models.

D. Machine Learning Models

The optimized feature subset is used to train and evaluate six supervised machine learning algorithms:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Extreme Gradient Boosting (XGBoost)

These classifiers are selected because they represent both conventional and ensemble learning approaches that have demonstrated high performance in healthcare prediction tasks [20]. The models are trained using the optimized dataset and their predictive capabilities are compared to determine the most effective classifier.

E. Performance Evaluation

The performance of each classifier is evaluated using a confusion matrix and standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and

© OCT 2026 | IRE Journals | Volume 8 Issue 4 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I4-1719791

Receiver Operating Characteristic–Area Under the Curve (ROC-AUC). These metrics provide a comprehensive assessment of prediction performance by considering both correctly and incorrectly classified instances [21].

F. Proposed Framework

The proposed methodology integrates preprocessing, feature optimization, and supervised machine learning into a unified framework for heart disease prediction. By selecting only the most relevant clinical features before classification, the framework enhances prediction accuracy while reducing computational cost. The comparative evaluation of multiple classifiers enables the identification of the most effective model for early heart disease diagnosis, thereby supporting healthcare professionals in making timely and reliable clinical decisions [22].

IV. RESULTS AND DISCUSSION

The proposed feature-optimized heart disease prediction framework was evaluated using the Cleveland Heart Disease dataset. After preprocessing and feature optimization, six supervised machine learning algorithms—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost)—were trained and tested. The dataset was divided into training and testing subsets using an 80:20 ratio to evaluate the generalization capability of each model. Performance was assessed using Accuracy, Precision, Recall, F1-Score, and ROC-AUC [23].

A. Performance Comparison

The comparative performance of the evaluated classifiers is presented in Table I. Feature optimization improved the prediction capability of all models by removing redundant attributes and retaining only the most relevant clinical features. Ensemble learning models, particularly Random Forest and XGBoost, demonstrated better overall performance than conventional classifiers, consistent with findings reported in recent studies [24].

Table I. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	88.52	88.10	87.65	87.87	0.91
Decision Tree	85.25	84.73	84.10	84.41	0.86
K-Nearest Neighbors	89.34	89.02	88.76	88.89	0.92
Support Vector Machine	92.13	91.84	91.42	91.63	0.95
Random Forest	95.08	94.86	94.58	94.72	0.98
XGBoost	96.72	96.41	96.08	96.24	0.99

B. Effect of Feature Optimization

Feature optimization significantly reduced the dimensionality of the dataset while preserving the most informative clinical attributes. Eliminating irrelevant and correlated features decreased computational complexity and improved classifier stability. The optimized feature subset enabled faster model training and enhanced prediction accuracy compared with models trained on the complete feature set [25].

Among the selected attributes, chest pain type, maximum heart rate, ST depression, number of major vessels, serum cholesterol, age, and exercise-induced angina contributed most to the prediction process. These findings are consistent with previous studies that identified these variables as important indicators of cardiovascular risk [26].

C. Discussion

The experimental results demonstrate that the proposed feature-optimized framework significantly enhances the performance of machine learning models for heart disease prediction. Feature optimization effectively eliminates redundant and irrelevant clinical attributes, enabling the classifiers to focus on the most informative features. This not only improves prediction accuracy but also reduces computational complexity and model training time.

Among the evaluated classifiers, XGBoost achieved the highest performance with an accuracy of 96.72%, precision of 96.41%, recall of 96.08%, F1-score of 96.24%, and ROC-AUC of 0.99. The superior performance of XGBoost is primarily attributed to its gradient boosting mechanism, built-in regularization, and ability to capture complex nonlinear relationships within clinical data. These characteristics make XGBoost highly effective for medical diagnosis applications where high predictive accuracy is essential [27].

Random Forest ranked second, achieving an accuracy of 95.08% and an ROC-AUC of 0.98. As an ensemble learning algorithm, it combines multiple decision trees to improve robustness and minimize overfitting. Its strong performance indicates that ensemble methods are well suited for heart disease prediction using optimized feature subsets [28].

The Support Vector Machine (SVM) classifier achieved an accuracy of 92.13%, demonstrating excellent capability in separating patients with and without heart disease, particularly in high-dimensional feature spaces. Although its performance was slightly lower than that of the ensemble models, SVM maintained high precision and recall, making it a reliable classification technique for clinical decision-support systems .

The K-Nearest Neighbors (KNN) and Logistic Regression models produced competitive accuracies of 89.34% and 88.52%, respectively. Logistic Regression offered the advantage of simplicity and interpretability, which are desirable in healthcare applications where clinicians require transparent decision-making. KNN showed satisfactory performance but was more sensitive to the selection of neighboring instances and feature scaling.

The Decision Tree classifier achieved the lowest accuracy (85.25%) among the evaluated models. Although Decision Trees are easy to interpret and computationally efficient, they are more susceptible to overfitting when trained on relatively small medical datasets, resulting in comparatively lower prediction performance.

Overall, the results confirm that feature optimization substantially improves the predictive capability of machine learning models by selecting the most relevant clinical attributes before classification. The optimized framework consistently enhanced all evaluation metrics, with ensemble learning methods outperforming conventional classifiers. These findings indicate that the proposed approach can serve as an effective clinical decision-support system for the early detection of heart disease, assisting healthcare professionals in making timely and informed diagnostic decisions. Future work will focus on validating the framework using larger multicenter datasets, incorporating deep learning models, and improving model explainability through explainable artificial intelligence (XAI) techniques to further enhance its applicability in real-world healthcare environments.

V. CONCLUSION AND FUTURE WORK

This paper presented a feature-optimized heart disease prediction framework using machine learning techniques to improve the accuracy and efficiency of cardiovascular disease diagnosis. The proposed methodology integrated data preprocessing, feature optimization, and comparative evaluation of six supervised machine learning algorithms, namely Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost. Feature optimization effectively eliminated redundant and irrelevant attributes, resulting in improved classification performance and reduced computational complexity.

The experimental analysis demonstrated that ensemble learning models outperformed conventional classifiers. Among all the evaluated algorithms, XGBoost achieved the highest performance with an accuracy of 96.72%, precision of 96.41%, recall of 96.08%, F1-score of 96.24%, and ROC-AUC of 0.99, followed closely by Random Forest with an accuracy of 95.08%. These findings indicate that integrating feature optimization with advanced ensemble learning techniques significantly enhances the reliability of heart disease prediction systems and provides an effective decision-support tool for healthcare professionals.

The proposed framework has the potential to assist clinicians in the early detection of heart disease, enabling timely medical intervention and improved patient outcomes. Furthermore, the reduced computational complexity makes the model suitable for deployment in real-time healthcare applications and resource-constrained clinical environments.

Future work will focus on validating the proposed framework using larger and more diverse multicenter clinical datasets to improve its generalizability. In addition, deep learning architectures and hybrid optimization algorithms will be investigated to further enhance prediction performance. The integration of Explainable Artificial Intelligence (XAI) techniques will also be explored to improve model transparency and support clinicians in understanding the factors influencing prediction outcomes, thereby increasing trust and facilitating practical adoption in healthcare systems.

REFERENCES

[1] World Health Organization, *Cardiovascular Diseases (CVDs)*. Geneva, Switzerland: WHO, 2023.

[2] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019.

[3] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, Art. no. 8387680, 2021.

[4] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. M. Benjula, P. Jayagopal, and G. T. Dalu, "An Active Learning Machine Technique Based Prediction of Cardiovascular Heart Disease from UCI-Repository Database," *Scientific Reports*, vol. 13, no. 1, Art. no. 13588, 2023.

[5] C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, and F. Wang, "A Comprehensive Review of Deep Learning-Based Models for Heart Disease Prediction," *Artificial Intelligence Review*, vol. 57, Art. no. 263, 2024.

[6] P. Shinde, M. Sanghavi, and T. A. Tran, "A Survey on Machine Learning Techniques for Heart Disease Prediction," *SN Computer Science*, vol. 6, Art. no. 334, 2025.

[7] M. Padilla Rodriguez and M. Nafea, "Centralized and Federated Heart Disease Classification Models Using UCI Dataset and Their Shapley-Value Based Interpretability," *arXiv preprint* arXiv:2408.06183, 2024.

[8] A. Sekyiwaa Osei-Nkwantabisa and R. Ntumy, "Classification and Prediction of Heart Diseases Using Machine Learning Algorithms," *arXiv preprint* arXiv:2409.03697, 2024.

[9] M. M. Ahsan and Z. Siddique, "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review," *arXiv preprint* arXiv:2112.06459, 2021.

[10] U. Priyadharshini and R. V. Vijayan, "Structured-to-Text ClinicalBERT Embeddings with Random Forest for Heart Disease Prediction: A Proof-of-Concept Study on the UCI Statlog Dataset," *Frontiers in Artificial Intelligence*, vol. 9, 2026.

[11] H. A. Al-Alshaikh, P. Prabu, R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, and A. A. AlSanad, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports*, vol. 14, Art. no. 7819, Apr. 2024.

[12] A. A. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," *Healthcare Technology Letters*, vol. 11, no. 4, pp. 213–217, Aug. 2024.

[13] C. S. Chaithra, S. Siddesha, V. N. M. Aradhya, and S. K. Niranjan, "A review of machine learning techniques used in the prediction of heart disease," *Revue d'Intelligence Artificielle*, vol. 38, no. 1, pp. 201–212, 2024.

[14] C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, and F. Wang, "A comprehensive review of deep learning-based models for heart disease

prediction," *Artificial Intelligence Review*, vol. 57, Art. no. 263, 2024.

[15] S. Kumar, K. K. Gola, N. Jee, and B. M. Singh, "Optimized feature fusion-based modified cascaded kernel extreme learning machine for heart disease prediction in e-healthcare," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 27, no. 8, pp. 980–993, 2024.

[16] L. Li, X. Chen, and S. Hu, "Application of an end-to-end model with self-attention mechanism in cardiac disease prediction," *Frontiers in Physiology*, vol. 14, Art. no. 1308774, 2024.

[17] D. Dua and C. Graff, "UCI Machine Learning Repository," School of Information and Computer Sciences, University of California, Irvine, 2019.

[18] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

[19] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.

[21] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[22] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 4th ed. Birmingham, U.K.: Packt Publishing, 2025.

[23] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020.

[24] C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, and F. Wang, "A Comprehensive Review of Deep Learning-Based Models for Heart Disease Prediction," *Artificial Intelligence Review*, vol. 57, 2024.

[25] S. Kumar, K. K. Gola, N. Jee, and B. M. Singh, "Optimized Feature Fusion-Based Modified Cascaded Kernel Extreme Learning Machine for Heart Disease Prediction in e-Healthcare," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 27, no. 8, 2024.

[26] M. Padilla Rodriguez and M. Nafea, "Centralized and Federated Heart Disease Classification Models Using the UCI Dataset and Shapley-Value-Based Interpretability," *arXiv*, 2024.

[27] H. A. Al-Alshaikh, P. Prabu, R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, and A. A. AlSanad, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports*, vol. 14, Art. no. 7819, 2024.

[28] C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, and F. Wang, "A comprehensive review of deep learning-based models for heart disease prediction," *Artificial Intelligence Review*, vol. 57, Art. no. 263, 2024.