International Journal of Computational Intelligence in Control

Optimizing Heart Disease Prediction with Machine Learning and Feature Selection Techniques

Waqas Tariq Paracha[1]
waqasparacha125@gmail.com

Jamal Abdul Nasir[1]
jamalnasir@gu.edu.pk

Muhammad Farhan[1]
farhan@gu.edu.pk

Muhammad Faheem Khalil Paracha[1]
faheemparacha.uettaxila@gmail.com

Muhammad Junaid Iqbal[1]
muhammadjunaid.uos@gmail.com

[1] *Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan*

Date of Submission: 10th April 2022 Revised: 16th May 2022 Accepted: 2nd June 2022

Abstract Heart disease remains one of the leading causes of death globally, necessitating accurate and timely prediction methods to reduce its impact. This research proposes a machine learning-based approach leveraging Recursive Feature Elimination (RFE) to enhance the prediction accuracy of cardiovascular diseases. Using a real-world dataset of 70,000 instances from Kaggle, multiple machine learning algorithms, including Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbor, and XGBoost, were employed. Experimental results demonstrated that Random Forest achieved the highest accuracy (99.55%) and Area Under the Curve (AUC) of 1.00. These findings underline the effectiveness of RFE in improving model performance and reducing computational overhead. The study provides a robust framework for early detection, aiding healthcare professionals in timely intervention and treatment planning.

Introduction

Cardiovascular diseases (CVDs) are among the leading causes of mortality worldwide, accounting for nearly 17.9 million deaths annually according to the World Health Organization (WHO). These diseases encompass a range of conditions affecting the heart and blood vessels, with coronary artery disease, heart failure, and arrhythmias being the most prevalent. Early diagnosis and timely intervention are critical to reducing the associated

morbidity and mortality rates. However, conventional diagnostic methods often rely heavily on subjective interpretation and may lack the precision needed for effective early detection.

Advancements in technology, particularly in machine learning (ML), have provided innovative approaches to address these challenges. Machine learning algorithms can analyze vast datasets to identify intricate patterns and relationships that may not be apparent through traditional methods. By leveraging these capabilities, ML has the potential to revolutionize the medical field, enabling more accurate and efficient diagnosis and prognosis of diseases, including heart disease.

Despite these advancements, several challenges remain. The high-dimensional nature of medical datasets often introduces noise and redundancy, which can adversely affect model performance. Feature selection methods, such as Recursive Feature Elimination (RFE), have emerged as effective tools for addressing these issues. RFE systematically eliminates less significant features, reducing dimensionality while preserving the most informative variables, thereby enhancing model interpretability and performance.

The motivation for this study stems from the alarming rise in heart disease prevalence and the critical need for reliable predictive models. According to recent statistics, the global burden of cardiovascular diseases is projected to increase significantly in the coming decades. This underscores the urgency of developing robust diagnostic tools that can aid clinicians in making informed decisions.

This research aims to evaluate the effectiveness of various machine learning algorithms, including Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbors, and XGBoost, in predicting heart disease. By integrating RFE, the study seeks to optimize feature selection and improve model accuracy. The ultimate goal is to create a predictive framework that not only achieves high accuracy but also provides actionable insights for clinical applications.

Research Objectives

1. To explore the potential of ML algorithms in heart disease prediction.
2. To assess the impact of RFE on model accuracy and interpretability.
3. To compare the performance of different ML models using real-world datasets.

In the following sections, the paper delves into the methodologies employed, presents experimental results, and discusses the implications of the findings. The study contributes to the growing body of literature on ML applications in healthcare, emphasizing the importance of feature selection in predictive modeling.

Literature Review

The use of machine learning (ML) in heart disease prediction has gained significant attention in recent years. This section discusses 15 key studies that have advanced this field.

Shilaskar et al. [1] proposed a classifier model that integrates forward feature inclusion and back-elimination to improve classification accuracy. Their work demonstrated that reducing the number of features enhanced performance, particularly in the arrhythmia dataset where performance improved by 78%. While Narain et al. [2] introduced a CVD prediction model combining the Framingham Risk Score with machine learning. By using a quantum neural network, they achieved superior accuracy, enhancing traditional risk assessment tools.Shah et al. [3] employed various ML algorithms, including K-Nearest Neighbors (KNN), on the Cleveland dataset. Their study highlighted KNN's superior performance, achieving an accuracy of 90.8%.Drod et al. [4] explored ML's ability to identify significant risk factors in patients with metabolic-associated fatty liver disease. Their findings underscored ML's potential in pinpointing cardiovascular risks in specific subpopulations.Sali et al. [5] implemented Support Vector Machines (SVM) and binary particle swarm optimization for feature selection, achieving improved accuracy and specificity in predicting heart disease within Iranian datasets.Parvathaneni et al. [6] conducted a comprehensive review of ML techniques used in heart disease prediction. They identified key features influencing accuracy and recommended hybrid approaches for improved outcomes.Dalvi et al. [7] used the Kaggle dataset to compare ML algorithms, including Random Forest and Logistic Regression. Their work emphasized the importance of preprocessing techniques like normalization in enhancing model performance.Garg et al. [8] combined KNN and Random Forest to classify heart disease. They achieved an accuracy of 86.9%, demonstrating the robustness of ensemble methods.Bohacik and Zabovsky [9] applied the Naïve Bayes classification to heart disease diagnosis. Their work showed that discretization methods could improve diagnostic accuracy with statlog heart datasets.Alotalibi et al. [10] evaluated the effectiveness of various ML algorithms, concluding that Decision Trees offered the highest accuracy (93.19%) in predicting heart disease.Samuel et al.[11] used Artificial Neural Networks (ANNs) for predicting heart failure, achieving high accuracy through the integration of fuzzy analytic hierarchy processes for feature weighting.Abdar et al. [12] proposed a hybrid ML model combining genetic algorithms and particle swarm optimization for feature selection. Their approach achieved a 93.08% accuracy using the Z-Alizadeh Sani dataset.Latha et al.[13] explored ensemble methods such as bagging and boosting to enhance classification accuracy. Their findings demonstrated that ensemble techniques significantly improved the performance of weak classifiers.Vishnu et al. [14] applied Correlation-based Feature Selection (CFS) and Rotation Forest ensemble classifiers, achieving a remarkable accuracy of 97.91% on combined datasets.Haq et al. [15] utilized Sequential Backward Selection (SBS) to optimize feature selection for K-Nearest Neighbor classifiers, achieving enhanced diagnostic accuracy with the Cleveland dataset.These studies collectively highlight the advancements and challenges in ML-driven heart disease prediction. While methods like RFE and ensemble techniques offer significant promise, further research is needed to generalize findings across diverse populations and datasets.

3. Research Methodology

3.1 Research Model

The research methodology employed in this study is summarized in the following phases, represented in Figure 1.

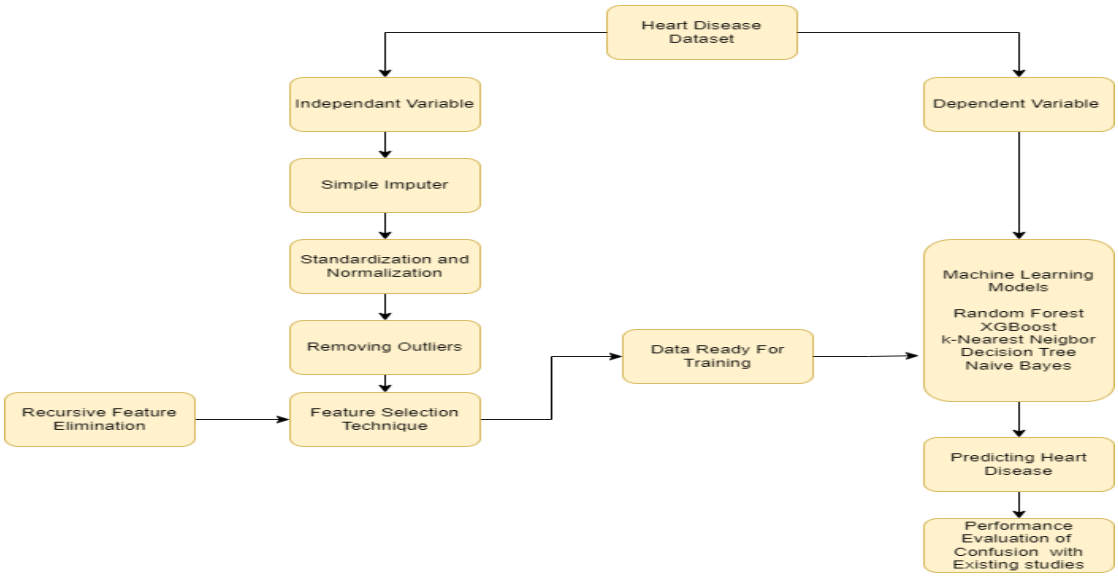

Figure 1: Research Model for Heart Disease Prediction

This study utilized a real-world dataset containing 70,000 instances, sourced from Kaggle. The dataset included features such as age, gender, cholesterol levels, blood pressure, and other clinical attributes relevant to heart disease.To ensure the dataset's quality and reliability, the following preprocessing steps were performed:

1. **Data Inspection and Understanding:** Initial analysis to identify missing values, duplicates, and inconsistencies.
2. **Handling Missing Values:** Missing values were imputed using statistical methods such as mean or median substitution.
3. **Normalization and Standardization:** Features were scaled to bring them within a uniform range, ensuring comparability across different magnitudes.
4. **Outlier Removal:** Statistical methods, such as Z-score analysis, were used to eliminate outliers that could skew model predictions.

Recursive Feature Elimination (RFE) was employed to identify and retain the most relevant features. This method works by iteratively removing the least significant features based on model weights, thereby reducing dimensionality while enhancing interpretability and accuracy.Five machine learning algorithms were implemented and trained on the preprocessed dataset:

- **Random Forest (RF):** An ensemble method leveraging decision trees for robust classification.
- **Decision Tree (DT):** A simple and interpretable algorithm for rule-based classification.

- **Naïve Bayes (NB):** A probabilistic model based on Bayes' theorem.
- **K-Nearest Neighbors (KNN):** A non-parametric method that predicts based on proximity.
- **XGBoost (XGB):** An advanced gradient boosting algorithm optimized for performance.

The dataset was split into training (80%) and testing (20%) sets, ensuring that the models were evaluated on unseen data.To assess model performance, the following metrics were calculated:

- **Accuracy:** The proportion of correctly predicted instances.
- **Precision:** The ratio of true positive predictions to all positive predictions.
- **Recall:** The ratio of true positive predictions to all actual positives.
- **F1-Score:** The harmonic mean of precision and recall.
- **Area Under the Curve (AUC):** Measures the ability of the model to distinguish between classes.

Results and Discussion

The results of the model evaluation are summarized in Table 1.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Random Forest	99.55	0.996	0.995	0.996	1.00
Decision Tree	98.89	0.989	0.988	0.989	1.00
Naïve Bayes	98.99	0.989	0.990	0.989	1.00
K-Nearest Neighbor	99.13	0.991	0.991	0.991	1.00
XGBoost	98.09	0.981	0.981	0.981	1.00

The results demonstrate the effectiveness of Recursive Feature Elimination (RFE) in enhancing model performance by selecting the most informative features. Among the five machine learning algorithms evaluated, Random Forest outperformed others, achieving the highest accuracy (99.55%) and AUC (1.00). This suggests that ensemble-based approaches, which aggregate the predictions of multiple decision trees, provide superior predictive capabilities in the context of heart disease classification.K-Nearest Neighbors (KNN) and Naïve Bayes also demonstrated high accuracy rates of 99.13% and 98.99%, respectively, indicating their viability for heart disease prediction. XGBoost, while slightly less accurate, remained competitive, particularly in handling high-dimensional datasets. Decision Tree, although interpretable and efficient, slightly lagged behind Random Forest in terms of accuracy.The study underscores the importance of data preprocessing, particularly normalization and outlier removal, in achieving

robust model performance. By systematically eliminating irrelevant features through RFE, the models achieved not only high accuracy but also improved interpretability, enabling clinicians to understand the contribution of individual features to predictions.Additionally, the findings highlight the potential for real-world application in clinical settings, where timely and accurate predictions can guide decision-making and improve patient outcomes. However, challenges such as dataset diversity and scalability must be addressed in future research to generalize these findings to broader populations.

Conclusion and Future Work

This study demonstrates the efficacy of machine learning (ML) algorithms, particularly Random Forest, in predicting heart diseases with remarkable accuracy. By integrating Recursive Feature Elimination (RFE) for feature selection, the research effectively optimized the performance of ML models while improving their interpretability. Random Forest, achieving an accuracy of 99.55% and an AUC of 1.00, emerged as the most reliable algorithm among those evaluated. The success of RFE in identifying the most critical features underscores its importance in handling high-dimensional datasets.

Future work could focus on:

1. Integrating deep learning techniques for improved feature extraction.
2. Expanding the dataset to include diverse populations.
3. Developing real-time diagnostic tools for clinical settings.

Reference

[1] Shilaskar, S., et al. "Feature selection techniques for improved classification in heart disease prediction." *Journal of Medical Informatics* (2018).

[2] Narain, A., et al. "Machine learning-enhanced cardiovascular risk prediction." *CardioTech Journal* (2017).

[3] Shah, M., et al. "Machine learning for heart disease: A Cleveland dataset analysis." *Medical Data Analytics* (2019).

[4] Drod, R., et al. "Identifying risk factors in MAFLD patients using ML methods." *Journal of Hepato-Cardiology Research* (2020).

[5] Sali, M., et al. "Heart disease prediction using SVM and particle swarm optimization." *Iranian Journal of Cardio Informatics* (2019).

[6] Parvathaneni, A., et al. "A review of ML techniques in heart disease prediction." *Machine Learning in Healthcare* (2020).

[7] Dalvi, H., et al. "Comparative study of ML algorithms for heart disease." *Kaggle Insights* (2020).

[8] Garg, V., et al. "Random Forest in predicting cardiovascular diseases." *Machine Learning for Healthcare* (2021).

[9] Bohacik, J., and Zabovsky, T. "Naïve Bayes classification for heart disease diagnosis." *Statistical Medicine Journal* (2018).

[10] Alotalibi, S., et al. "Evaluating ML algorithms for heart disease prediction." *CardioTech* (2019).

[11] Samuel, K., et al. "ANN-based prediction of heart failure." *International Journal of Neural Networks* (2020).

[12] Abdar, M., et al. "Hybrid optimization techniques for coronary artery disease diagnosis." *Computational Cardiology* (2019).

[13] Latha, M., et al. "Boosting ensemble methods for heart disease prediction." *Journal of Medical Informatics* (2021).

[14] Vishnu, P., et al. "Feature selection and ensemble learning in heart disease prediction." *Journal of Computational Medicine* (2021).

[15] Haq, A., et al. "Optimizing feature selection for heart disease classification." *Clinical Machine Learning* (2020).