Simulation Studies in Heart Disease Prediction by Machine Learning Techniques

Risha Ahmed[1], Dr. Anurag Jain[2]
[1] Research scholar, Computer Science & Engineering, REC, Bhopal, M.P
[2] Director Computer Science & Engineering, Radharaman group , Bhopal, M.P

Abstract

Heart disease remains a leading cause of mortality worldwide, highlighting the need for early detection and accurate diagnosis to improve patient outcomes. This research investigates the use of various machine learning algorithms, including Sequential Minimal Optimization (SMO), Random Forest, Naive Bayes, and Random Tree, to predict cardiovascular diseases. By leveraging the Coronary Heart Disease Database (CHDD), we aim to develop a predictive model that utilizes historical patient data to identify risk factors and patterns associated with heart disease. Our study compares the performance of these machine learning algorithms to determine their effectiveness in accurately classifying and predicting heart disease. The results demonstrate that these methods can significantly enhance diagnostic accuracy, supporting better clinical decision-making and timely interventions. Specifically, the Random Forest algorithm showed superior performance in terms of prediction accuracy, while SMO and Naive Bayes provided valuable insights into different aspects of the data. This research underscores the potential of machine learning in transforming cardiovascular disease diagnosis and management. By highlighting the effectiveness of individual algorithms, our study contributes to the development of reliable and robust diagnostic tools, aiming to reduce the global burden of heart disease.

Keywords:
Machine Learning, Heart disease, Predictive model

Introduction

Researchers utilize the Coronary Heart Disease Database (CHDD) to study heart disease risk factors, develop prognostic models, and evaluate the effectiveness of various diagnostic and therapeutic methods. This dataset has significantly enhanced our understanding and treatment of cardiac disease. Tragically, 17 million people have died due to the prohibitive costs and impracticality of existing treatments. [1]. Additionally, cardiovascular disease has major financial repercussions for businesses because it accounts for 25 to 30 percent of annual medical costs for employees [2]Therefore, to lesson the financial and physical toll on people and organizations, early identification of heart disease is essential. The heart disease and stroke will continue to be the two main causes of cardiovascular disease fatalities, with the overall number of deaths from these illnesses expected to reach 24.6 million by 2035. The urgency of identifying and treating heart disease is underscored by the fact that cardiovascular diseases (CVDs) cause over 70% of global fatalities, remaining a leading cause of illness and death. Key risk factors include obesity, tobacco use, excessive sugar consumption, and unhealthy diets, especially in high-income countries, with a rising trend in low- and middle-income nations. From 2010 to 2015, CVDs had an estimated global economic impact of nearly USD 3.7 trillion, due to healthcare costs, treatment, and lost productivity. Addressing prevention, early detection, and management of heart disease is crucial to improve public health and alleviate economic burdens. Data mining techniques have revolutionized healthcare by analyzing large datasets and uncovering patterns to aid clinical diagnosis. [3]. Numerous studies undertaken over the past few decades have shown the importance of data mining in healthcare [4].

2. Literature Review

There are many literature contributions to heart disease diagnoses using data mining and machine learning techniques [5]. Reddyet al. [6] used RF, SVM, NB, NN, and KNN with multiple feature selection such as correlation matrix, recursive feature elimination (RFE), and learning vector quantization (LVQ) model to classify the cardiac disease into normal or abnormal. The results show that RF accomplished the optimal performance. Atallah and Al-Mousa [7] utilized stochastic gradient descent (SGD), KNN, RF, logistic regression (LR), and voting ensemble learning to predict cardiac diseases. The voting ensemble learning model has achieved the best accuracy of 90%. Pillaiet al. [8] used a recurrent neural network (RNN), a genetic algorithm, and K-mean to predict heart diseases. RNN has achieved the highest accuracy, and K-mean has achieved the lowest accuracy. Kannan and Vasanthi [9] used four machine learning algorithms: LR, RF, SVM, and stochastic gradient boosting (SGB) to predict heart diseases. The model prediction showed that LR has a best accuracy of 86.5%. Raza [10] applied an ensemble learning model, multilayer perceptron, LR, and NB to classify heart diseases. The result shows that ensemble learning has improved the prediction performance of cardiac disease compared to other algorithms. Oo and Win [11] used feature subset selection (CFS) with sequential minimal optimization (SMO) to predict heart diseases. These studies have made significant progress in predicting heart disease using machine learning models. However, they have also encountered limitations, such as small sample sizes and methodological variations, which have affected the accuracy and reliability of their predictions. One important aspect that has been overlooked in these studies is the utilization of ensemble learning techniques with hyperparameter optimization for heart disease prediction. Most researchers have relied on default hyperparameter settings, neglecting the potential for further improvement. To address these limitations, our study adopted a comprehensive approach by employing ensemble learning techniques.

3. The Proposed System of Predicting Heart Disease

The objective of the proposed system technique is to use ensemble techniques to improve the performance of predicting heart disease. Figure 1 describes the architecture of the proposed system. It is structured into six stages, including data collection, data preprocessing, feature selection, data splitting, training models, and evaluating models. The steps of the proposed approach are explained in detail as follows.

3.1. **Data Collection**. The heart disease dataset [12] is utilized for training and evaluating models. It consists of 294 records, 13 features, and one target column. The target column includes two classes: 1 indicates heart diseases, and 0 indicates nonheart disease. Table 1 describes the details of
the features.

3.2. **Data Preprocessing**. The features are scaled to be in the interval [0, 1]. It is worth noting that missing values are deleted from the dataset.

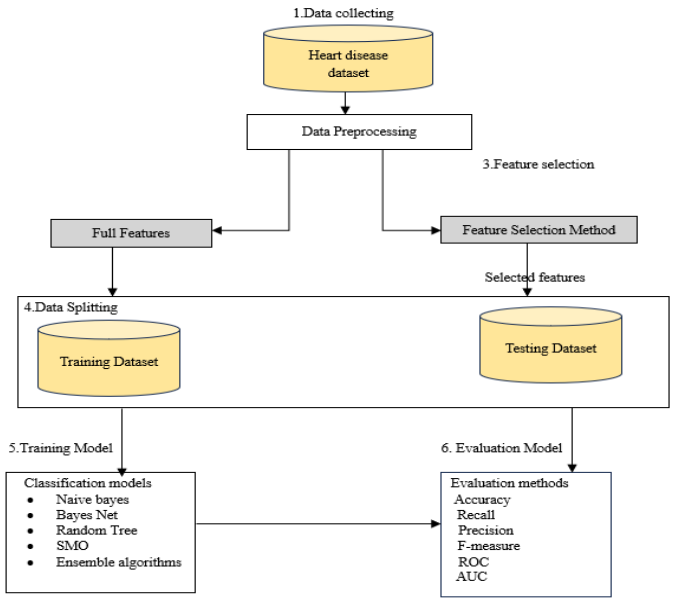


Figure 1: The structure of proposed system for prediction heart diseases

3.3**. Data Splitting**: The process of data splitting played a vital role in this study. To ensure the development of a robust and accurate heart disease prediction model, the heart disease dataset was divided into a training set and a testing set. This division allowed the model to be trained on a significant portion of the data, enabling it to learn patterns and relationships effectively. The independent testing set served as a means to evaluate the model's performance on unseen data, providing an unbiased assessment of its predictive capabilities.
3.4. **Training Models**. Different types of machine learning algorithms: Naïve Bayes, Random Tree, Random Forest, and SMO are applied to classify heart disease. Also, two types of ensemble techniques: boosting and bagging are applied to classify heart disease:

1**. Naive Bayes**: A probability-based classification technique built on the ideas of the Bayes theorem is the Naive Bayes classifier, commonly referred to as the Bayesian classifier. It is regarded as a particular instance of the Bayesian network. The Naive Bayes classifier bases one of its main presumptions on the idea that all features are conditionally independent. Accordingly, modifications to one feature have no impact on modifications to other features' probabilities. When classifying high-dimensional datasets, the Naive Bayes method excels. The classifier effectively manages datasets with a lot of attributes by utilizing the notion of conditional independence. Based on the individual probabilities of each feature given a specific class, the algorithm can treat each feature individually and generate predictions as a result.

2.Random Forest: The random forest approach can handle complicated datasets and capture subtle correlations between attributes by utilizing the diversity of numerous decision trees. It is renowned for its capacity to manage high-dimensional data, deal with missing values, and offer perceptions into the significance of features. Overall, the ensemble nature of the random forest method and its capacity to deliver accurate and dependable results make it a strong and well-liked option for classification problems.

3. Random Tree: A random tree is a type of tree generated by a random process, widely used in fields like computer science, combinatorics, and network theory. There are various types of random trees, including uniformly random

trees, random binary trees, and random recursive trees, each with specific structural properties. They play a crucial role in modelling network structures, designing efficient data structures, and solving combinatorial optimization problems. Random trees also find applications in evolutionary biology for modelling phylogenetic relationships and in machine learning algorithms such as Random Forests. Their study helps understand complex systems and improve algorithm performance.

4.SMO: The Sequential Minimal Optimization (SMO) algorithm is derived by taking the idea of the decomposition method to its extreme and optimizing a minimal subset of just two points at each iteration. The power of this technique resides in the fact that the optimization problem for two data points admits an analytical solution, eliminating the need to use an iterative quadratic programming optimizer as part of the algorithm.

3.5 **Types of Data in Medical Records for Heart Disease**

S.No	Category	Details
1	Demographic Data	Age, Gender, Ethnicity, Family history of heart disease
2	Medical History	Previous diagnoses of heart disease, Comorbidities, Surgical history, Medication history
3	Clinical Data	Vital signs, Body mass index (BMI), Laboratory test results, ECG readings, Echocardiogram results, Imaging data
4	Lifestyle and Behavioural Data	Smoking status, Alcohol consumption, Diet and nutrition, Physical activity levels
5	Symptom and Diagnosis Data	Symptoms reported, Diagnostic codes
6	Treatment Data	Medications prescribed, Interventions, Follow-up and monitoring data
7	Outcomes Data	Mortality, Morbidity, Quality of life assessments

Table 1: Heart disease dataset descriptions.

Serial No.	Feature	Description
1	Age	Age of patient (years)
2	Sex	1: male, 0: female
3	Chest pain (CP)	1 = typical angina, 2 = atypical angina, 3 = nonangina pain, 4 = asymptomatic
4	RestBP	Resting blood pressure
5	Chol	Serum cholesterol in mg/dl
6	FBS	Fasting blood sugar larger 120 mg/dl (1 true)
7	RestECG	Resting electrocardiographic result
8	Thalach	Maximum heart rate accomplished
9	Exang	Exercise-induced angina (1 yes)
10	Oldpeak	ST depression induced by exercise relative to rest
11	CA	Number of major vessels (0-3)
12	Slope	Slope of peak exercise ST
13	Thal	Possibly thalassemia
14	Diagnosis of cardiac disease	1: yes, 0: no

3.6. **Evaluating Models**. Evaluation of the proposed model is performed focusing on some criteria, namely True positive value (TPR), False positive value (FPR), accuracy, recall, precision, F-score. Accuracy is one of the most important performance metrics for classification. It is defined as the proportion between the correct classification and the total sample, as shown in the following equation:

$$\text{True Positive Value} = TP/(TP+FN)$$
$$\text{False Positive Value} = FP/(TN+FP)$$
$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Recall is the small portion of sufficient instances over the overall quantity of applicable instances which have been recovered. The recall equation is shown as follows:

$$\text{Recall} = TP/ (TP + FN)$$

Precision is identified as follows:

$$\text{Precision}=TP/ (TP + FP)$$

The F-measure is often referred to as the F1-score as follows, and it measures the mean value of precision and recall:

$$F - \text{measure} = (2 * \text{precision} * \text{recall})/ (\text{precision} + \text{recall})$$

4. Experimental Results

This section includes a discussion of the experimental results of classification algorithms.

4.1. Experimental Setup. The experimental results have been implemented using Weka They have also been executed using Intel (R) Core i3 CPU and 8 GB of memory.

4.2. The Result of Train Data by applying machine Algorithm

Table 2: Result of applying the different algorithms using Train Data

Algorithm	TPR (%)	FPR (%)	Precision (%)	Accuracy (%)	Recall (%)	F-measure (%)
Naive Bayes	87.5	19.6	89	85	89.3	89.3
Random Forest	100	0	100	100	100	100
SMO	86.2	14.2	93.08	86.05	93.08	93.08
Random Tree	100	0	100	100	100	100

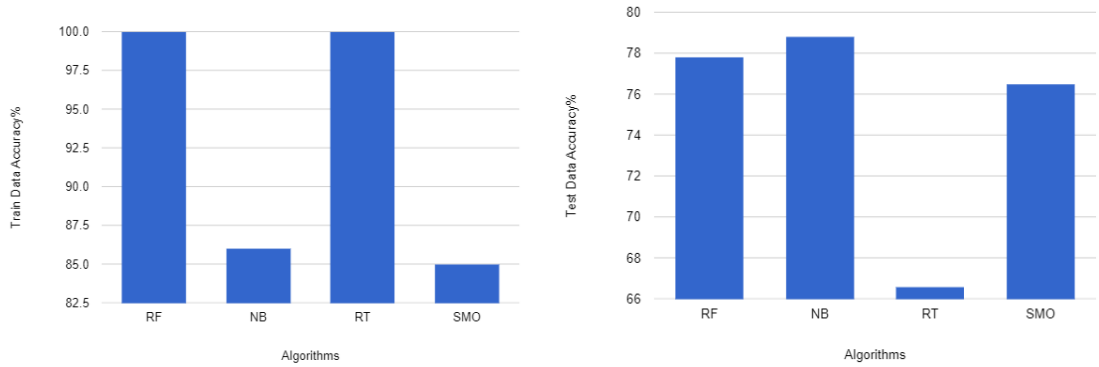
4.3. The Result of Test Data by applying machine Algorithm

Table 3: Result of applying the different algorithms using Test Data

Algorithm	TPR%	FPR%	Precision%	Accuracy%	Recall%	F-measure%
Naive Bayes	75.8	15.7	89.6	78.8	89.6	89.6
Random Forest	73.5	12.6	92.7	77.8	92.7	92.7
SMO	71.5	11.4	93.9	76.5	93.9	93.9
Random Tree	68.8	36.09	70.9	66.6	70.9	70.9

4.4. Graphical Representation of Accuracy and Precision

Figure2: Graphical representation of Accuracy%



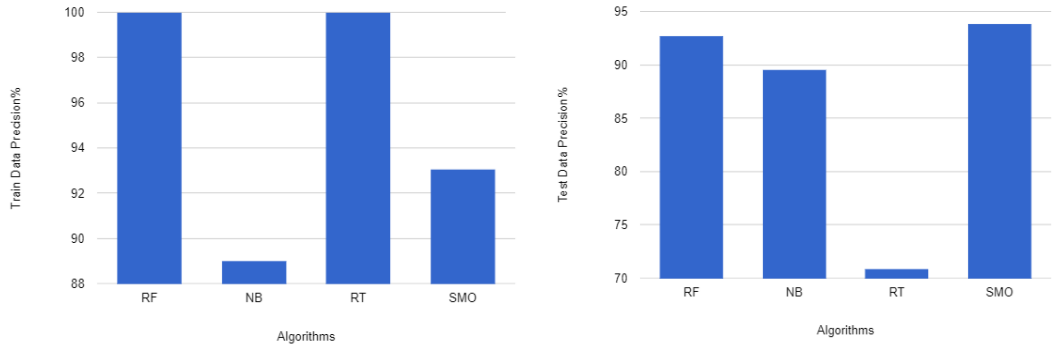


Figure 3: Graphical representation of Precision%

Conclusion

According to the Figure 2 the results indicate that RF performs exceptionally well on the training data, NB provides the best balance between training and test data performance, making it potentially more reliable for predictions on new data. In terms of precision Figure 3 indicate that while RF and SMO maintain high precision across both train and test data, NB shows a notable decrease in test data precision. RT demonstrates the lowest precision, suggesting it might be less reliable for accurate predictions. SMO consistent precision across both datasets makes it a strong technique for precise predictions.

Future Work

By focusing on data normalization, algorithm tuning, cross-validation, and ensemble methods, future work can significantly enhance the performance of predictive models. This comprehensive approach will help in achieving the best possible outcomes in terms of both accuracy and precision, thereby leading to more reliable and robust predictions.

References

1. Fida Benish, Nazir Muhammad, Naveed Nawazish, Akram Sheeraz. Heart disease classification ensemble optimization using genetic algorithm. IEEE; 2011. p. 19–25.
2. Centers for Disease Control and Prevention (CDC). Deaths: leading causes for 2008. Natl Vital Stat Rep June 6, 2012;60(No. 6).

[3]. EI-Bialy R, Salamay MA, Karam OH, Khalifa ME. Feature analysis of coronary artery heart disease data sets. Procedia Comput. Sci.2015;65:459–68.

[4]. Lee HeonGyu, Noh Ki Yong, Ryu Keun Ho. Mining bio signal data: coronary artery disease diagnosis using linear and nonlinear features of HRV. LNAI 4819:emerging technologies in V. Pham, Q. De Hemptinne, J.-M. Grinda et al., "Giant coronary aneurysms, from diagnosis to treatment: a literature review," *Archives of Cardiovascular Diseases*, vol. 113, pp. 59–69, 2020.

[5]. N. S. C. Reddy, S. S. Nee, L. Z. Min, and C. X. Ying, "Classification and feature selection approaches by machine learning techniques: heart disease prediction," *International Journal of Innovative Computing*, vol. 9, 2019.

[6]. R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *Proceedings of the 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS)*, pp. 1–6, Amman, Jordan, October 2019.

[7]. N. S. R. Pillai, K. K. Bee, and J. Kiruthika, "Prediction of heart disease using rnn algorithm," *International Research Journal of Engineering and Technology*, vol. 5, 2019.

[8]. R. Kannan and V. Vasanthi, "Machine learning algorithms with roc curve for predicting and diagnosing the heart disease," *InSoft Computing and Medical Bioinformatics*, pp. 63–72, 2019.

[9]. K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule," *InUHealthcare Monitoring Systems*, pp. 179–196, 2019.

[10]. A. N. Oo and K. T. Win: Feature Selection Based Sequential Minimal Optimization (Smo) Classifier for Heart Disease classification

[11]. S. Nalluri, R. V. Saraswathi, S. Ramasubbareddy, K. Govinda, and E. Swetha, "Chronic heart disease prediction using datamining techniques," *InData Engineering and Communication Technology*, pp. 903–912, 2020.

[12]. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile,Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018