Performance Comparison of Naive Bayes, Decision Tree, and Random Forest for Heart Disease Prediction

Eko Rian Sri Raharjo[1], Akhsanul Khuluq[2]
Communication and Information Service Klaten Regency, Klaten, Indonesia[1]
Directorate of Academics and Admissions, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia[2]
ekoriansr26@gmail.com

***Abstract*—Medical datasets can support early clinical screening when they are analyzed with appropriate and reproducible machine learning procedures. In this paper, Naive Bayes, Decision Tree, and Random Forest are compared for heart disease prediction on a publicly available dataset with 14 attributes and 270 patient records. The experimental workflow consisted of variable identification, data quality checking, target label transformation, train-test split at 80:20 ratio, model training in Python, and evaluation using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), confusion matrix, and computation time. The results demonstrate that Naive Bayes achieved the best overall performance with 90.74% accuracy, 94.44% precision, 80.95% recall, 87.18% F1-score, and 0.925 AUC. The second best model was Random Forest with 79.63% accuracy and 0.885 AUC. The Decision Tree model achieved an accuracy of 68.52% and an AUC of 0.690. Naive Bayes also has the lowest training and prediction times which means it is the most efficient model for this dataset. However, due to the relatively limited size of the dataset, this conclusion should be interpreted as a preliminary comparison rather than a final clinical recommendation. Future work should include larger datasets, cross-validation, class balancing, feature selection, hyperparameter optimization and external validation.**

Keywords—heart disease prediction, machine learning, Naive Bayes, decision tree, random forest, model evaluation.

I. INTRODUCTION

The rapid development of information technology has increased the use of data-driven approaches in healthcare. Machine learning is particularly useful because it can learn patterns from historical data and transform those patterns into predictions that may support clinical screening, risk stratification, and decision support [1], [2]. In cardiology, artificial intelligence is increasingly discussed because cardiovascular data often contain interacting risk factors that are difficult to evaluate manually with a single rule-based approach [3].

Heart disease prediction remains an important research problem because early identification may help both clinicians and patients prioritize further examination and preventive action. The classical heart disease datasets used in machine learning research originate from clinical variables such as age, chest pain type, blood pressure, cholesterol, electrocardiographic results, maximum heart rate, exercise-induced angina, ST segment depression, slope, number of vessels, thallium test result, and disease status [4]. Although such datasets are compact compared with modern electronic health records, they remain useful for benchmarking classification algorithms and for learning how different models behave under controlled experimental conditions.

A critical issue in medical prediction research is that model accuracy alone is not enough. Prediction models require transparency in reporting, reproducible preprocessing, proper validation, and careful interpretation of performance metrics [5], [6]. A model that performs well on a small test split may still generalize poorly to different populations or data collection settings. Therefore, the comparative experiments in this study report accuracy, precision, recall, F1-score, AUC, confusion matrix, and computation time.

Previous work suggests that the relative performance of machine learning algorithms varies across datasets. Also, promising results for heart disease prediction have been shown by hybrid and ensemble models [7]. Supervised classifiers such as Naive Bayes, Decision Tree, Random Forest, k-nearest neighbors, support vector machines, and logistic regression are commonly compared on structured cardiovascular datasets [8]–[10]. Hossain et al. [11] showed that Random Forest worked well in the study of cardiovascular disease. Pal et al. [12] and Mehrabani-Zeinabad et al. [13] pointed out the data-, feature- and validation-design-dependence of the performance of the model. Studies on feature selection and dimensionality reduction also suggest that preprocessing choices can significantly affect prediction outcomes [14].

These results suggest that no single classifier can be assumed superior for all heart disease datasets. For that reason, this study compares Naive Bayes, Decision Tree, and Random Forest on the same public dataset and under the same train-test split. The objective is to identify which model provides the best balance of predictive performance and computational efficiency in an initial heart disease prediction experiment.

II. METHODS

A. Research Design

The study was designed as a quantitative experiment. The experiment is based on the training of three supervised classification algorithms with one data set and using the same testing records and performance metrics to evaluate the results. This design was selected to ensure that the comparison reflects model behavior rather than differences in the composition of preprocessing or test data.

The process was divided into five main steps: dataset collection, variable analysis, data preprocessing, model training, and performance evaluation. The experimental process was designed to improve reproducibility and to reduce ambiguity in model comparison. This follows the general principle that prediction studies should describe data, modeling procedures and evaluation criteria clearly [5], [6].

B. Dataset and Research Variables

The dataset used in this study is the publicly available Heart Disease Prediction dataset from Kaggle, comprising 270 patient records and 14 attributes. Thirteen attributes were used as independent variables: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar > 120 mg/dl, electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression, ST-segment slope, number of major vessels, and thallium test result. The dependent variable was heart disease, a binary target indicating the presence or absence of heart disease.

The dataset combines numerical variables, including age, blood pressure, cholesterol, maximum heart rate, and ST depression, with categorical or ordinal variables, including sex, chest pain type, electrocardiographic results, exercise-induced angina, ST-segment slope, number of vessels, and thallium. The measurement scale of each variable is relevant to data preparation and to the interpretation of model behavior.

TABLE I. HEART DISEASE DATASET ATTRIBUTES

No.	Attribute	Description
1	Age	Patient age
2	Sex	Patient sex
3	Chest Pain Type	Chest pain category
4	BP	Resting blood pressure
5	Cholesterol	Serum cholesterol
6	FBS Over 120	Fasting blood sugar > 120 mg/dl
7	EKG Results	Electrocardiogram result
8	Max HR	Maximum heart rate
9	Exercise Angina	Exercise-induced angina
10	ST Depression	ST-segment depression
11	Slope of ST	Slope of the ST segment
12	Number of Vessels	Number of major vessels
13	Thallium	Thallium test result
14	Heart Disease	Target class

Based on Table 1, the dataset contains 13 predictor variables and one binary target variable. Predictors included demographic, clinical, electrocardiographic, exercise, and diagnostic data. Quantitative measures like age, blood pressure, cholesterol, maximal heart rate, and ST depression are supplemented with categorical or ordinal indicators such as sex, chest pain type, exercise angina, ST-segment slope, number of vessels, and thallium. This mixture provides various information to train the model. And the Heart Disease attribute defines the class (present or absent) to be predicted.

C. Data Preprocessing

Preprocessing was done to make the dataset uniformly processable by the selected classification algorithms. An initial check showed that there were no missing values in the dataset, so there was no need for imputation. The Heart Disease target was converted into numerical labels using LabelEncoder. The data were then separated into a feature matrix (X) and target vector (y), followed by an 80:20 training-testing split.

D. Classification Models

Three classification models were evaluated. Naive Bayes was trained using GaussianNB, which estimates the class probabilities from the distributions of the features and chooses the class with the highest posterior probability. The Decision Tree was implemented with DecisionTreeClassifier (random_state=42) that recursively partitions the feature space to informative split criteria. For Random Forest, RandomForestClassifier (n_estimators=100, random_state=42) was used and the final prediction was based on the aggregation of predictions of multiple decision tree models.

E. Model Evaluation

The random state parameter was set to 42 to make the experiment reproducible. All three models were trained and evaluated using the same split, which reduces the risk that performance differences arise from variations in the test samples rather than from the models themselves. Because the dataset contains only 270 records, the results should be interpreted cautiously; previous evidence shows that small medical datasets can produce unstable performance estimates when validation is limited [17].

The models were evaluated on 54 testing records. Accuracy was used to measure the proportion of correct predictions, precision to assess the reliability of positive predictions, recall to measure the ability to identify positive cases, and F1-score to balance precision and recall. AUC was also reported because ROC-based measures are widely used to evaluate classifier separability and can provide more information than accuracy alone under varying decision thresholds [19], [20].

The distribution of true positive, false positive, true negative and false negative predictions were described using the confusion matrix. Model efficiency was assessed by recording the computation time for training and prediction. The use of multiple metrics is important because the performance of binary classification can appear different depending on whether the focus is on overall correctness, positive-case detection, or class separability [21], [22].

III. RESULTS AND DISCUSSION

A. Model Evaluation Results

Table 2 summarizes the performance and computation time of the three models. Accuracy, precision, recall, and F1-score are percentages; AUC is reported on a 0-1 scale; and computation times are reported in seconds.

Table 2 indicates that Naive Bayes exhibited the most formidable predictive performance, achieving the highest accuracy (90.74%), precision (94.44%), F1-score (87.18%), and AUC (0.925). The Decision Tree has a recall of 71.43% which was higher than the Random Forest recall of 66.67% but had the lowest precision of 57.69% suggesting the Decision Tree was more likely to make false-positive predictions. Random Forest exhibited moderate performance, outperforming Decision Tree in accuracy, precision, F1-score

and AUC but taking significantly more time to train and predict. This table shows that Naive Bayes achieved the best compromise in terms of predictive accuracy and computational efficiency for this dataset.

TABLE II. MODEL PERFORMANCE COMPARISON

Metric	Naive Bayes	Decision Tree	Random Forest
Accuracy (%)	90.74	68.52	79.63
Precision (%)	94.44	57.69	77.78
Recall (%)	80.95	71.43	66.67
F1-score (%)	87.18	63.83	71.79
AUC	0.925	0.690	0.885
Train Time (s)	0.0058	0.0087	0.2841
Predict Time (s)	0.0017	0.0021	0.0187

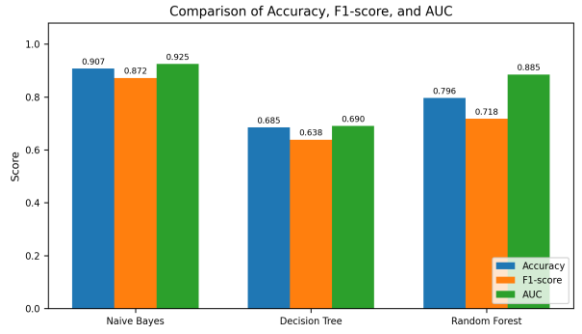

Fig. 1. Comparison of accuracy, F1-score, and AUC across the three models.

As we can see in Figure 1, Naive Bayes has always given the best scores in terms of accuracy (0.907), F1-score (0.872), and AUC (0.925). The values are tightly clustered, which indicates balanced overall performance and not the strength of performance in just one measure of evaluation. The second best performance was achieved by Random Forest with an AUC of 0.885, higher than its accuracy (0.796) and F1-score (0.718), and showing good class separation ability, but a lower threshold-based classification performance.

Figure 1 also reveals that the Decision Tree had the lowest accuracy (0.685), F1-score (0.638), and AUC (0.690). Its AUC was only slightly higher than its accuracy, whereas Naive Bayes and Random Forest showed clearer AUC advantages. This pattern suggests that the single-tree model was less stable on the unseen test records than the probabilistic and ensemble alternatives.

Taken together, Figure 1 and Table 2 show that Naive Bayes was the best-performing model under the selected train-test split, Random Forest was the second-best predictive model but the most computationally demanding, and Decision Tree produced the weakest overall results. These comparisons are specific to the present dataset and validation design and should not be interpreted as universal rankings of the algorithms.

B. *Confusion Matrix Analysis*

This research also analyzes the model's performance using a confusion matrix, which is shown in Figure 2. Figure 2 presents the Naïve Bayes confusion matrix for the 54 test records. The model correctly classified 32 absence cases (true negatives) and 17 presence cases (true positives), with one false positive and 4 false negatives. Thus, 49 of the 54 test records were classified correctly. The single false positive is consistent with the high precision of 94.44%, whereas the four false negatives explain why recall (80.95%) was lower than precision.

Based on Figure 2, false negatives were the dominant error type. In a clinical screening context, these errors may be more concerning than false positives because they represent patients with heart disease who were predicted as absence cases. Consequently, although Naive Bayes achieved the best overall performance in this experiment, future work should evaluate decision-threshold adjustment, cross-validation, and external validation to determine whether false-negative errors can be reduced without producing an excessive increase in false positives.

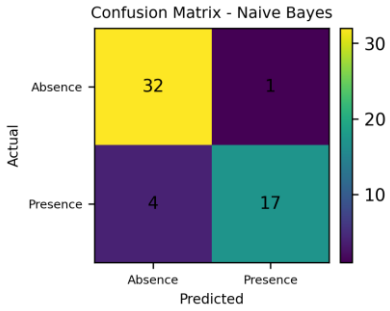

Fig. 2. Confusion matrix of the Naive Bayes model.

C. *Computational Efficiency*

Based on Table 2, Naive Bayes required 0.0058 seconds for training and 0.0017 seconds for prediction. Decision Tree was slightly slower, requiring 0.0087 seconds for training and 0.0021 seconds for prediction. The Random Forest's computation time was the highest (0.2841 seconds for training and 0.0187 seconds for prediction). Such a difference is expected, since Random Forest trains and combines many decision tree models, while Naive Bayes estimates a smaller number of probabilistic parameters.

Considering both performance and efficiency, Naive Bayes was the most suitable model for this particular dataset. However, the efficiency advantage should not be interpreted separately from the dataset size. Because the dataset contains only 270 records, all models are computationally light in absolute terms. The observed difference is more meaningful as an indication of algorithmic complexity than as a practical runtime limitation.

D. *Critical Discussion*

The good performance of Naive Bayes suggests that a simple probabilistic classifier is sufficient to model the relationship between the available attributes and the target class in this dataset. Although the independence assumption of Naive Bayes is often considered restrictive, the result indicates that this assumption did not prevent useful classification performance on this compact and structured dataset.

The weaker result of Decision Tree highlights an important methodological point: interpretability does not automatically imply higher predictive accuracy. A single tree can be attractive because it produces a clear decision structure, but it may also overfit training patterns. Random Forest reduced this weakness through an ensemble mechanism, consistent with previous heart disease studies that reported robust performance from ensemble-based classifiers [7], [11], [16], [23].

At the same time, several studies have reported different best-performing models depending on data source, preprocessing, and validation design [8], [9], [10], [11], [12], [13], [14], [15], [24], [25]. This means the present result should not be generalized as a universal claim that Naive Bayes is always superior for heart disease prediction. Rather, the result shows that Naive Bayes was the best among the three tested algorithms under the specific experimental conditions used in this study.

The main limitations of this study are the small sample size, the use of secondary data, and the absence of cross-validation. The experiment did not consider feature selection, probability calibration, class balancing or hyperparameter optimization. These limitations are important because medical prediction models have to be robustly validated and transparently reported before they can be trusted for practical use [5,6,17]. Therefore, this study should be understood as a preliminary algorithm comparison and not as a final clinical decision system.

IV. CONCLUSIONS

This study compared Naive Bayes, Decision Tree, and Random Forest for heart disease prediction using a public dataset containing 270 patient records and 14 attributes. Naive Bayes performed better than the other models with the maximum accuracy (90.74%), F1-score (87.18%), and AUC (0.925), and the minimum training and prediction times. Random Forest came in second and performed better than Decision Tree on most of the predictive metrics but took more computation time. The Decision Tree performed the worst overall in this dataset.

The findings indicate that Naive Bayes was the most effective and computationally efficient model under the experimental conditions used in this study. However, the result should be treated as preliminary because it was obtained from a relatively small dataset and a single train-test split. Future work should determine whether the observed performance is preserved with larger datasets, k-fold cross-validation, external validation, class balancing, feature selection, hyperparameter optimization, and probability calibration.

V. DATA AVAILABILITY

The dataset used in this study is publicly available in the Kaggle Heart Disease Prediction repository at https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction. The analyzed file contains 270 patient records, 13 predictor attributes, and one binary target attribute.

REFERENCES

[1] R. C. Deo, "Machine Learning in Medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[2] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[3] K. W. Johnson et al., "Artificial Intelligence in Cardiology," Journal of the American College of Cardiology, vol. 71, no. 23, pp. 2668–2679, Jun. 2018, doi: 10.1016/j.jacc.2018.03.521.

[4] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[5] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a multivariable prediction model for individual prognosis or Diagnosis (TRIPOD): The TRIPOD statement," Annals of Internal Medicine, vol. 162, no. 1, pp. 55–63, Jan. 2015, doi: 10.7326/M14-0697.

[6] K. G. M. Moons et al., "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and elaboration," Annals of Internal Medicine, vol. 162, no. 1, pp. W1–W73, Jan. 2015, doi: 10.7326/M14-0698.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[8] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, vol. 1, no. 345, 2020, Art no. 345, doi: 10.1007/s42979-020-00365-y.

[9] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, Art. no. 88, Feb. 2023, doi: 10.3390/a16020088.

[10] C. A. ul Hassan et al., "Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers Effectively," Sensors, vol. 22, no. 19, article 7227, Sep. 2022, doi: 10.3390/s22197227.

[11] S. Hossain et al., "Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023," BMC Cardiovascular Disorders, vol. 24, no. 214, Apr. 2024, doi: 10.1186/s12872-024-03883-2.

[12] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," Open Medicine, vol. 17, no. 1, pp. 1100–1113, 2022, doi: 10.1515/med-2022-0508.

[13] K. Mehrabani-Zeinabad, A. Feizi, M. Sadeghi, H. Roohafza, M. Talaei, and N. Sarrafzadegan, "Cardiovascular disease incidence prediction by machine learning and statistical techniques: a 16-year cohort study from eastern Mediterranean region," BMC Medical Informatics and Decision Making, vol. 23, article 77, 2023, doi: 10.1186/s12911-023-02169-5.

[14] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, p. 100330, 2020, doi: 10.1016/j.imu.2020.100330.

[15] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," Biology, vol. 12, no. 1, p. 117, Jan. 2023, doi: 10.3390/biology12010117.

[16] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Improving the Prediction of Heart Diseases Using Ensemble Learning and Hyperparameter Optimization," Algorithms, vol. 16, no. 6, article 308, Jun. 2023, doi: 10.3390/a16060308.

[17] A. Althnian et al., "Effect of Dataset Size on Classification Performance: An Empirical Evaluation on Medical Datasets," Applied Sciences, vol. 11, no. 2, article 796, Jan. 2021, doi: 10.3390/app11020796.

[18] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[19] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[20] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.

[21] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, pp. 1–6, Jan. 2020.

[22] D. G. Altman and J. M. Bland, "Diagnostic tests 1: Sensitivity and specificity," BMJ, vol. 308, no. 6943, p. 1552, Jun. 1994, doi: 10.1136/bmj.308.6943.1552.

[23] N. A. Baghdadi, S. M. F. Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," Journal of Big Data, vol. 10, p. 144, 2023, doi: 10.1186/s40537-023-00817-1.

[24] 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, article 8387680, 2021, doi: 10.1155/2021/8387680.

[25] F. Ali et al., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," Information Fusion, vol. 63, pp. 208–222, Nov. 2020, doi: 10.1016/j.inffus.2020.06.008.