Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Comparison of Random Forest and K-Nearest Neighbors in Heart Disease Prediction

Erni[1], Ibnu Alfarobi[2]*, Wawan Kurniawan[3]

[1.2.3]*Universitas Bina Sarana Informatika*
erni.erx@bsi.ac.id[1], ibnu.iba@bsi.ac.id[2*], wawan.wwk@bsi.ac.id[3]

Abstract

Heart disease is one of the leading causes of death worldwide, with a death toll reaching 17.9 million cases annually according to the World Health Organization (WHO) and a prevalence of 1.5% in Indonesia. This high mortality rate demonstrates the importance of early detection and accurate prediction to prevent more serious complications. The development of artificial intelligence technology, particularly machine learning, offers a new approach in the medical field through the ability to analyze clinical data quickly and efficiently. This study was conducted to compare the performance of two machine learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), in predicting heart disease using a clinical dataset from Kaggle containing 20 samples and 9 attributes related to the patient's physiological condition. The parameter optimization process in both algorithms was carried out using grid search techniques with cross-validation to obtain the best model that can perform optimally on a limited dataset. Performance evaluation was carried out using accuracy, recall, and precision metrics to comprehensively measure the quality of the model predictions. The results of the study showed that the Random Forest algorithm provided superior performance with an accuracy of 0.75, a recall of 0.88, and a precision of 0.86, compared to KNN which only achieved an accuracy of 0.50, a recall of 0.67, and a precision of 0.67. These findings indicate that Random Forest is more effective in identifying the presence of heart disease, especially in terms of sensitivity to positive cases and prediction consistency. Thus, Random Forest has the potential to be a more appropriate algorithm for implementation in machine learning-based clinical decision support systems, to support the process of diagnosing heart disease more accurately and efficiently.

Keywords: *Heart Disease, Machine Learning, Random Forest, K-Nearest Neighbors (KNN)*

1. Introduction

Heart disease is one of the most serious health problems and a leading cause of death worldwide. According to the latest report from the World Health Organization (WHO), heart disease causes approximately 17.9 million deaths annually, or approximately 31% of all global deaths. This situation places heart disease as a public health threat that requires special attention. In Indonesia, the prevalence of coronary heart disease, as diagnosed by doctors, reaches 1.5%, with the highest percentage recorded in the Special Region of Yogyakarta Province at 2.0%. The high number of cases and their highly fatal consequences demonstrate the need for preventive measures, including early detection and accurate prediction to reduce the risk of dangerous complications[1][2].

As the number of heart disease sufferers increases, the need for fast, accurate, and efficient diagnostic methods becomes increasingly crucial. Conventional diagnostic methods such as physical examinations, medical records, and laboratory tests are often time-consuming and do not always provide optimal accuracy, especially under certain conditions. Therefore, the use of information technology to assist in the process of analyzing medical data is highly relevant. The development of artificial intelligence (AI) and machine learning technologies has opened up new opportunities in the medical world, particularly in supporting data-driven clinical decision-making. With the ability to process large amounts of data and recognize complex patterns in patient data, machine learning has the potential to improve the effectiveness of heart disease prediction[3].

Various machine learning algorithms have been used in medical research to predict heart disease, including Random Forest and K-Nearest Neighbors (KNN). Random Forest is an ensemble learning algorithm that combines multiple decision trees to produce more accurate, stable predictions that are less prone to overfitting. This algorithm is capable of performing well on datasets with many features and complex data variations. On the other hand, KNN is an instance-based algorithm that classifies new data based on its similarity to existing data in the training dataset. Despite its simplicity, this algorithm is often used due to its intuitive nature and its ability to produce fairly good classification results with appropriate parameters[1][4].

Although both algorithms have been widely used in various research and medical applications, there are still differences in performance results in various studies comparing the two. In some cases, Random Forest demonstrated superior performance, particularly in complex datasets with many features[5]. However, in other situations, KNN was able to provide competitive prediction results, particularly in datasets with small sample sizes or easily distinguishable data distributions. This variation in results indicates that algorithm performance is significantly influenced by the characteristics of the dataset used. Therefore, further evaluation is needed to determine which algorithm is more optimal in predicting heart disease in a given dataset[4][6] [7].

Based on this background, this study was conducted to compare the performance of the Random Forest and KNN algorithms in predicting heart disease using a clinical dataset containing various risk factors, such as age, blood pressure, cholesterol levels, and other health indicators. This study used metric-based evaluation methods such as accuracy, recall, and precision to provide a comprehensive overview of the performance of both algorithms. It is hoped that the results of this study will contribute to the development of more accurate clinical decision support systems, thereby assisting medical personnel in more effective early detection of heart disease.

2. Research Method

In this analysis, two Machine Learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), are used to predict the likelihood of heart disease based on available clinical datasets. Both algorithms were chosen because they are able to handle the characteristics of medical data that are complex, varied, and contain many determining features. Random Forest is used as an ensemble method that builds a number of decision trees to produce more stable and accurate predictions, while KNN is used as a distance-based similarity approach that classifies new data based on its proximity to existing data[8]. Through a process of training, testing, and performance evaluation using various metrics such as accuracy, precision, and recall, this study aims to determine which algorithm has better predictive ability in identifying patients who are potentially experiencing heart disease. The results of this evaluation will then serve as the basis for consideration in developing a more effective prediction model for decision support systems in the healthcare sector.

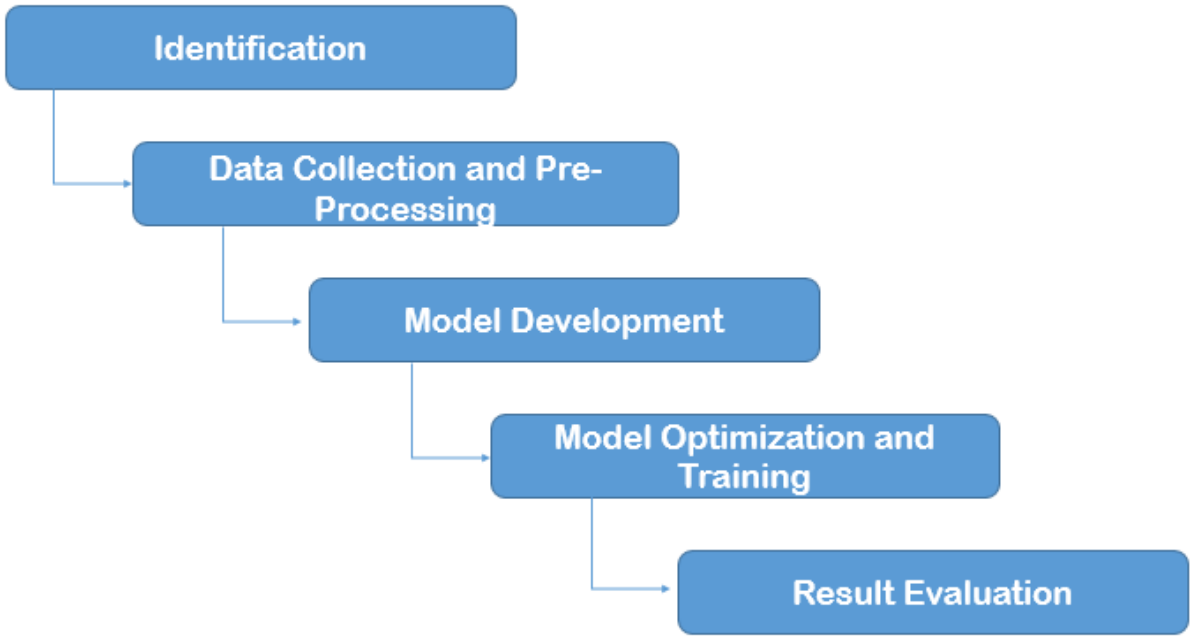

Fig. 1: Research Method

1. Problem Identification
   The initial stage of the research focused on formulating the main problem: how to predict whether a patient is at risk of heart disease based on their existing attributes or medical conditions. This problem arose due to the high mortality rate from heart disease and the need for a decision support system capable of assisting medical personnel in early detection. This problem identification served as a crucial basis for determining the research focus, data selection, and analysis methods used.
2. Data Collection and Pre-Processing
   The research data was obtained from a heart disease dataset available on the Kaggle platform, with a total of 1,200 instances and a number of relevant clinical attributes. After data collection, pre-processing was performed to improve data quality. This stage included cleaning the data from missing or inconsistent values, normalizing or standardizing features to ensure all variables were on the same scale, and dividing the dataset into training and test data. This pre-processing process is crucial to ensure the machine learning model can perform more accurately and stably.
3. Machine Learning Model Development
   This research used two main algorithms: Random Forest and K-Nearest Neighbors (KNN). Both models were built using the Python programming language and the scikit-learn library, which provides various functions for model training and testing. Random Forest was chosen for its ability to handle data with many features and reduce overfitting through an ensemble learning

approach. Meanwhile, KNN was chosen for its simplicity and ability to classify data based on the proximity between data points. At this stage, the initial model structure was established before optimization[9].

4. Model Optimization and Training
   To ensure optimal model performance, a parameter optimization process (hyperparameter tuning) was performed using grid search with cross-validation. The optimization included adjusting key parameters such as the number of trees (n_estimators) and tree depth in Random Forest, and the number of neighbors (n_neighbors) and distance metrics in KNN. This technique ensured that the model was tested on various parameter combinations to find the configuration that yielded the highest prediction accuracy and stability. After optimization was complete, the model was trained using training data to learn patterns from clinical data.
5. Result Evaluation and Analysis
   The final stage of the study was to evaluate the performance of each model using accuracy, recall, and precision metrics. Evaluation was conducted using test data to assess the model's ability to predict previously unseen data. Based on the evaluation results, Random Forest demonstrated superior performance with higher accuracy, recall, and precision compared to KNN. This analysis concluded that Random Forest is more effective in predicting heart disease and can be implemented as part of a machine learning-based clinical decision support system.

3. Result and Discussion

3.1. Data Collection

The dataset used in this study is the Kaggle heart disease dataset, which contains patient clinical data with diagnostic labels to identify the presence or absence of heart disease. This dataset was chosen because it has comprehensive features, is relevant, and has been widely used in previous research, thus facilitating the analysis process and comparison of results. The attributes are Age, Sex, Chest Pain Type (CP), Trestbps, Chol, Thalach, Exang, Oldpeak, and Target as variables that determine the presence of heart disease. Each attribute plays a crucial role in the analysis process, starting from basic patient characteristics, clinical indicators, to risk factors that contribute to the diagnosis of heart disease. This dataset is then used as the basis for the data processing process and the development of research models.

Table 1: Data Collection

No.	Age	Sex	Cp	Trestbps	Chol	Thalach	Exang	Oldpeak	Target
1	52	1	0	125	212	168	0	1.0	0
2	53	1	0	140	203	155	1	3.1	0
3	70	1	0	145	174	125	1	2.6	0
4	61	1	0	148	203	161	0	0.0	0
...	…	…	…	…	…	…	...	...	...
1199	60	1	2	140	185	155	0	3.0	0
1200	67	0	0	106	223	142	0	0.3	1

3.2. Pre-processing

Before being used to build a model, the dataset undergoes the following preprocessing steps:

1. Missing Value Handling: Checking for and handling missing or incomplete values using appropriate imputation techniques (mean for numeric features and mode for categorical features).
2. Data Normalization: Scaling numeric features to the [0,1] range using min-max scaling or standardization with Z-score to avoid the dominance of large-scale features.
3. Categorical Feature Encoding: Transforming categorical features into numeric form using one-hot encoding or label encoding techniques.
4. Feature Selection: Identifying and selecting the most relevant and influential features for heart disease diagnosis using methods such as correlation-based feature selection or feature importance from a Random Forest model.
5. Train-Test Split: Dividing the dataset into training data (80%) and testing data (20%), stratified by target class to maintain a balanced class distribution.

3.3. Modeling

1. Random Forest
   Random Forest is an ensemble learning algorithm used in this study due to its ability to produce stable and accurate predictions by combining multiple decision trees. This algorithm works based on the principle of bagging, which builds multiple decision trees from randomly selected data samples with replacement (bootstrap sampling), then combines the prediction results of each tree using majority voting. In the tree formation process, only a random subset of features is used at each node, thereby reducing correlation between trees and improving model performance. Important parameters such as n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features were optimized using a grid search method with cross-validation to obtain the best configuration. This approach ensures that the resulting Random Forest model is capable of providing optimal performance in the process of classifying heart disease.

2. K-Nearest Neighbors (KNN)
   K-Nearest Neighbors is an instance-based learning algorithm that classifies new data based on its similarity to existing data. This algorithm works on the principle that similar data points tend to be in the same class.
   KNN Working Principle:

a. Distance Calculation: Calculates the distance between a new data point and all data points in the training dataset using a specific distance metric (Euclidean, Manhattan, Minkowski, etc.).
b. Determining K-Nearest Neighbors: Selects the K closest data points based on the calculated distances.
c. Majority Voting: Classifies new data points based on the majority class of the K nearest neighbors.

Optimized KNN Parameters:
a. n_neighbors (K): Number of nearest neighbors considered (typically tested in the range 1-30)
b. weights: Point weighting (uniform or distance-weighted)
c. metric: Distance metric used (Euclidean, Manhattan, Minkowski)
d. p: Minkowski distance parameter (p=1 for Manhattan, p=2 for Euclidean)

KNN parameters were optimized using grid search with cross-validation to find the parameter combination that yielded the best performance.

3.4. Modeling Evalution

Table 2: Accuracy Comparison

Method	Accuracy
Random Forest	0.75
K-Nearest Neighbors	0.50

Based on the accuracy test results, the Random Forest method performed better than K-Nearest Neighbors. Random Forest achieved an accuracy score of 0.75, indicating that 75% of the data was correctly classified. Meanwhile, KNN only achieved an accuracy of 0.50, indicating that this model's ability to recognize data patterns is still low on the dataset used. This difference indicates that Random Forest is more effective in handling feature variations and is more stable in the process of classifying heart disease.

Table 3: Recall Comparison

Method	Recall
Random Forest	0.88
K-Nearest Neighbors	0.67

The recall evaluation results show that Random Forest has a higher sensitivity than KNN. Random Forest produced a recall value of 0.88, indicating that this model was able to correctly detect 88% of positive cases of heart disease. In contrast, KNN only achieved a recall of 0.67, indicating that some positive cases were still not identified by the model. The high recall valuein Random Forest indicates that this algorithm is better able to minimize false negative errors, making it more reliable in detecting the presence of heart disease.

Table 4: Precision Comparison

Method	Precision
Random Forest	0.86
K-Nearest Neighbors	0.67

In precision testing, Random Forest also demonstrated superior performance with a value of 0.86, indicating that the majority of positive predictions generated by the model were correct. On the other hand, KNN only achieved a precision of 0.67, indicating that this model has a higher false positive error rate. This difference in precision values confirms that Random Forest is more accurate in ensuring that each positive prediction is truly a case of heart disease, making it more appropriate for use in diagnostic contexts that require high accuracy.

4. Conclusion

Based on the analysis and testing of two machine learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), it can be concluded that Random Forest provides significantly better performance in predicting heart disease using the clinical dataset used in this study. Random Forest achieved an accuracy of 0.75, a recall of 0.88, and a precision of 0.86, demonstrating strong ability to identify positive cases and minimize prediction errors. Meanwhile, KNN achieved an accuracy of 0.50, a recall of 0.67, and a precision of 0.67, indicating that this algorithm is less effective on datasets with characteristics such as those in this study. These results confirm that Random Forest is superior in stability, generalization ability, and classification accuracy. Therefore, Random Forest is recommended as a more suitable algorithm for implementation in clinical decision support systems to detect heart disease more accurately.

References

[1] A. Agrawal, S., Gupta, R., & Kumar, "Comparative analysis of machine learning algorithms for heart disease prediction: A comprehensive study of Random Forest and KNN approaches.," *J. Biomed. Inform.*, pp. 104–118, 2023.
[2] B. Kumar, S., Patel, A., & Sharma, "Deep comparative analysis of machine learning algorithms for cardiovascular disease prediction: Focus on ensemble and lazy learning methods.," *Biomed. Signal Process. Control*, pp. 105–119, 2024.
[3] S. Hassan, A., Ibrahim, K., & Ahmed, "Comparative analysis of supervised learning algorithms for heart disease classification using clinical datasets," *J. Healthc. Eng.*, 2022.
[4] V. Gupta, M., Singh, R., & Arora, "Hybrid feature selection techniques for improving Random Forest and KNN performance in cardiac risk assessment," *IEEE Trans. Biomed. Eng.*, pp. 2245–2256, 2024.

[5] N. Das, P., Sharma, K., & Patel, "Performance evaluation of ensemble methods vs. instance-based learning for cardiovascular disease prediction. Expert Systems with Applications," *Expert Syst. Appl.*, pp. 117–132, 2022.
[6] P. Johnson, R., Thompson, D., & Wilson, "Evaluating the effectiveness of Random Forest versus K-Nearest Neighbors in predicting coronary artery disease," *Int. J. Med. Inform.*, pp. 105–118, 2023.
[7] S. Lee, H., Park, J., & Kim, "No Optimization of hyperparameters in Random Forest and KNN algorithms for heart disease prediction using grid search and cross-validation.," *Appl. Sci.*, 2023.
[8] M. Nakamura, T., Suzuki, Y., & Tanaka, "Performance comparison of machine learning classifiers for heart disease prediction: Random Forest vs. K-Nearest Neighbors analysis," *omputers Electr. Eng.*, pp. 107--120, 2022.
[9] M. Chen, L., Wang, H., & Liu, "Enhanced heart disease prediction using optimized Random Forest and weighted K-Nearest Neighbors algorithms. Computers in Biology and Medicine," *Comput. Biol. Med.*, pp. 107–121, 2024.