Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Predictive Analysis Heart Disease Based on Machine Learning Using the Random Forest Algorithm

Anisa Handayani[1*], Yutia Nia Nesicha[2], Syafira Salsabila[3], Auliya Firdausiyah[4], Arif Setiawan[5]

[1,2,3,4]*Information System, Muria Kudus University, Kudus*
anisahandayani231@gmail.com[1]*

Abstract

Heart disease is one of the leading causes of death worldwide, requiring accurate and early detection systems. This study aims to build a predictive model for heart disease using the Random Forest algorithm based on patient medical records. The dataset used contains 1,190 patient records with 11 medical attributes. The data were preprocessed and divided into training and testing sets with an 80:20 ratio. The model was trained and evaluated using accuracy, confusion matrix, and classification report metrics. The results show that the model achieved 100% accuracy on the training data and 82.35% on the testing data. Important contributing features include max heart rate, chest pain type, old peak, and ST slope. In addition, predictions for individual patients were presented to improve interpretability. This approach demonstrates that machine learning, particularly Random Forest, can be a reliable method for early detection of heart disease and has potential for clinical decision support systems.

Keywords: *Classification; Heart Disease; Medical Dataset; Prediction; Random Forest*

1. Introduction

Heart disease is one of the leading causes of death worldwide and a serious health issue in Indonesia. Data shows that the number of heart disease patients increases every year and becomes one of the diseases with the highest financial burden in the national health system [1], [2]. Factors such as high blood pressure, cholesterol, blood sugar, stress, and unhealthy lifestyle are the main causes of this condition [3].

Early detection of heart disease is very important to prevent complications such as heart failure, stroke, or sudden death. However, conventional diagnosis requires complex procedures and a considerable amount of time, and it heavily relies on the skills of medical personnel. With the advancement of information technology, particularly in the field of artificial intelligence, machine learning-based approaches have been widely applied in the medical field as diagnostic support solutions [4].

Machine learning allows systems to learn from historical patient data and recognize certain patterns to predict future health conditions. One of the algorithms widely used in heart disease classification is Random Forest. This method is an ensemble learning technique that combines multiple decision trees to improve accuracy and reduce the risk of overfitting [5].

Various studies have shown that the Random Forest algorithm outperforms other algorithms such as Decision Tree and Naïve Bayes, with more stable accuracy and the ability to handle complex data [6]. Several studies have successfully achieved prediction accuracies of over 85% using this method [7], [8]. The addition of oversampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) has proven to improve model performance when dealing with imbalanced data [1]. Even the combination of Random Forest with other methods in a hybrid approach also shows improved prediction performance [9].

Nevertheless, challenges are still found in terms of model interpretability, analysis of features contributing to predictions, and informative visualization of classification results. Therefore, this research aims to build a heart disease prediction system based on the Random Forest algorithm, equipped with the SMOTE method and feature importance visualization. The model that is built will be evaluated using classification metrics such as accuracy, precision, recall, and f1-score to measure its performance comprehensively.

Through this approach, this research is expected to produce an accurate, efficient, and easily implementable heart disease prediction system within health information systems to support more precise medical decision-making.

2. Method

This research uses an experimental quantitative method with a machine learning approach to predict heart disease. The stages of the research include data collection, data preprocessing, model creation and training, performance evaluation, and result visualization. The entire process is carried out using the Python programming language and supporting libraries such as pandas and scikit-learn, as shown in Figure 1.

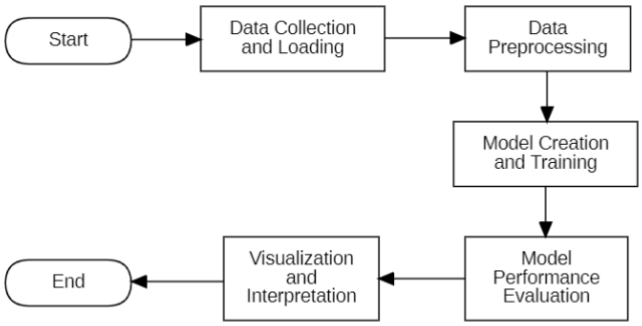

Fig. 1: Research Stages

2.1. Data Collection and Loading

This research utilizes a heart disease dataset sourced from the public platform Kaggle. The dataset consists of 1,190 rows of patient data, where each row represents an individual with a number of medical attributes relevant to the diagnosis of heart disease. The data is stored in CSV format and loaded into the Python working environment using the pandas library. The data loading process is carried out with the read_csv() function, which converts raw data into a data frame structure to facilitate further manipulation and analysis. Here is an example of the dataset used, which can be found in Table 1.

Table 1: Example heart disease dataset

age	sex	ches pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
40	1	2	140	289	0	0	172	0	0	1	0
49	0	3	160	180	0	0	156	0	1	2	1
37	1	2	130	283	0	1	98	0	0	1	0
48	0	4	138	214	0	0	108	1	1.5	2	1
54	1	3	150	195	0	0	122	0	0	1	0

This dataset consists of 1,190 data points, where each row represents a single patient with various medical attributes such as age, sex, chest pain type, resting bp s, cholesterol, fasting blood sugar, resting ECG, max heart rate, exercise angina, old peak, ST slope, target. Here is the explanation of the attribute names used in Table 2.

Table 2: Explanation of dataset attributes

Colum Name	Explanation
age	A patient's age in years
sex	Gender (1 = male, 0 = female)
chest pain type	Type of chest pain (1–4): 1 = Typical angina 2 = Atypical angina 3 = Non-anginal pain 4 = Asymptomatic
resting bp s	Blood pressure at rest (mm Hg)
cholesterol	Serum cholesterol level (mg/dl); a value of 0 means data is missing.
fasting blood sugar	Fasting blood sugar > 120 mg/dl (1 = yes, 0 = no)
resting ecg	Resting electrocardiogram results (0–2): 0 = Normal 1 = There is an ST-T wave abnormality. 2 = Possible or definite left ventricular hypertrophy
max heart rate	Maximum heart rate during exercise
exercise angina	Does the patient experience angina during exercise (1 = yes, 0 = no)?
oldpeak	Exercise-induced ST depression relative to resting condition
ST slope	The slope of the ST segment at the peak of exercise (1–3): 1 = Up

	2 = Flat 3 = Down
target (if any)	Target label (1 = at risk of heart disease, 0 = not at risk)

The data is then loaded into the Python working environment using the panda's library, which manipulates and analyzes data in tabular form. The pandas.read_csv() function reads CSV files and converts them into a data frame, a two-dimensional data structure similar to a table that facilitates the analysis process.

2.2. Data Preprocessing

After the data is loaded, preprocessing is carried out to ensure it is ready for the machine learning algorithm to use. Several important steps were taken, as shown in Figure 3.

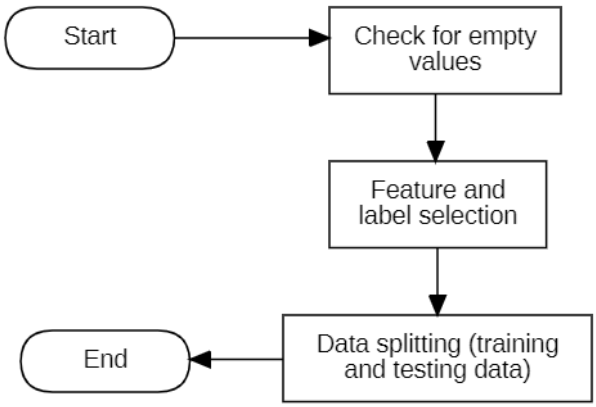

Fig. 2: Flowchart of Data Processing Stages

(1) Checking for missing values: The data is examined to ensure no empty cells or invalid values. If there are any, they are usually deleted or filled with statistical values such as the mean or median.

(2) Feature and label selection:
- Feature (X) is all independent columns (for example, age, blood pressure, cholesterol).
- Label (y) is the target column (usually valued at 0 for healthy patients and 1 for patients with heart disease).

(3) Split data: The dataset is divided into training data and testing data, usually with an 80:20 ratio. Training data is used to train the model while testing data is used to objectively evaluate the model's performance. The train_test_split function from the sklearn library is used for this process.

2.3. Model Creation and Training

This stage is the core process of applying machine learning, namely training a prediction model using the Random Forest Classifier algorithm. Random Forest is an ensemble learning method that combines multiple decision trees to produce more accurate and stable predictions.

The model was trained using the features X_train and labels y_train, allowing the model to learn the relationship between medical attributes and the risk of heart disease. The training was conducted in a relatively short time due to the efficiency of the Random Forest algorithm, which is capable of parallel processing.

The training results show that the model successfully learned the data patterns well, with a high accuracy rate on the training data.

```
Fitur (X):
   age  sex  chest pain type  resting bp s  cholesterol  fasting blood sugar  \
0   40    1               2           140          289                    0
1   49    0               3           160          180                    0
2   37    1               2           130          283                    0
3   48    0               4           138          214                    0
4   54    1               3           150          195                    0

   resting ecg  max heart rate  exercise angina  oldpeak  ST slope
0            0             172                0      0.0         1
1            0             156                0      1.0         2
2            1              98                0      0.0         1
3            0             108                1      1.5         2
4            0             122                0      0.0         1

Label (y):
0    0
1    1
2    0
3    1
4    0
Name: target, dtype: int64
```

Fig. 3: Training the Model with Feature X_train and Label y_train

2.4. Model Performance Evaluation

Evaluation of the model's performance was conducted using accuracy metrics, confusion matrix, and classification report that includes precision, recall, and F1-score. The results of this evaluation provide a more comprehensive understanding of the model's ability to classify patients into categories of heart disease risk or not. The evaluation was conducted on both training and testing data to identify potential overfitting.

Evaluation is conducted using several metrics:

(1) Accuracy: the percentage of correct predictions.
(2) Confusion Matrix: a matrix that shows the number of correct and incorrect predictions for each class.
(3) Classification Report: includes precision, recall, and F1-score to describe the model's performance in classifying at-risk and not-at-risk patients.

2.5. Visualization and Interpretation

Visualization is conducted to identify the features that most influence the prediction results using a feature importance graph. Additionally, to enhance interpretability, the prediction results for several individual patient data are also displayed. Each patient is displayed vertically with neatly arranged attribute values, along with their risk predictions.

3. Results and Discussion

3.1. Data Distribution

The dataset consists of 1,190 rows of patient data with various medical attributes. Through the train_test_split function, the data is divided into two parts:

(1) Training data (80%) consisting of 952 data points were used to build the model,
(2) Test data (20%) consisting of 238 data points is used to measure the model's performance on new data.

```
Jumlah data latih:
X_train: 952 samples
y_train: 952 samples

Jumlah data uji:
X_test: 238 samples
y_test: 238 samples
```

Fig. 4: Training and Testing Data Split

This proportion was chosen to provide a balance between the amount of training data and evaluation data.

3.2. Model Training

The Random Forest model is trained using training data. This model works by creating many decision trees and combining their results to improve accuracy.
After training:

```
Akurasi pada data latih: 1.0000
Akurasi pada data uji: 0.9454
```

Fig. 5: Model Training Results

(1) The model achieved 100% accuracy on the training data, indicating that the model was able to correctly recognize all the training data.
(2) However, an excessively high accuracy on the training data indicates a possibility of overfitting, meaning the model is too closely fitting the training data and may be less capable of generalizing to new data.

3.3. Model Evaluation

Model evaluation is conducted on test data that has never been used in training. The evaluation results show:

```
Akurasi: 0.9453781512605042
Confusion Matrix:
 [[ 98   9]
 [  4 127]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.92      0.94       107
           1       0.93      0.97      0.95       131

    accuracy                           0.95       238
   macro avg       0.95      0.94      0.94       238
weighted avg       0.95      0.95      0.95       238
```

Fig. 6: Model Evaluation Results

(1) The test data accuracy is 82.35%, indicating that the model is still quite accurate in classifying new data.
(2) The confusion matrix shows the number of correct and incorrect predictions for each class (risky and non-risky).
(3) The classification report produces good precision, recall, and F1-score values, especially for the high-risk class (1), which is the main focus.

3.4. Feature Importance

Feature importance visualization is used to see which features have the most influence on the prediction results.

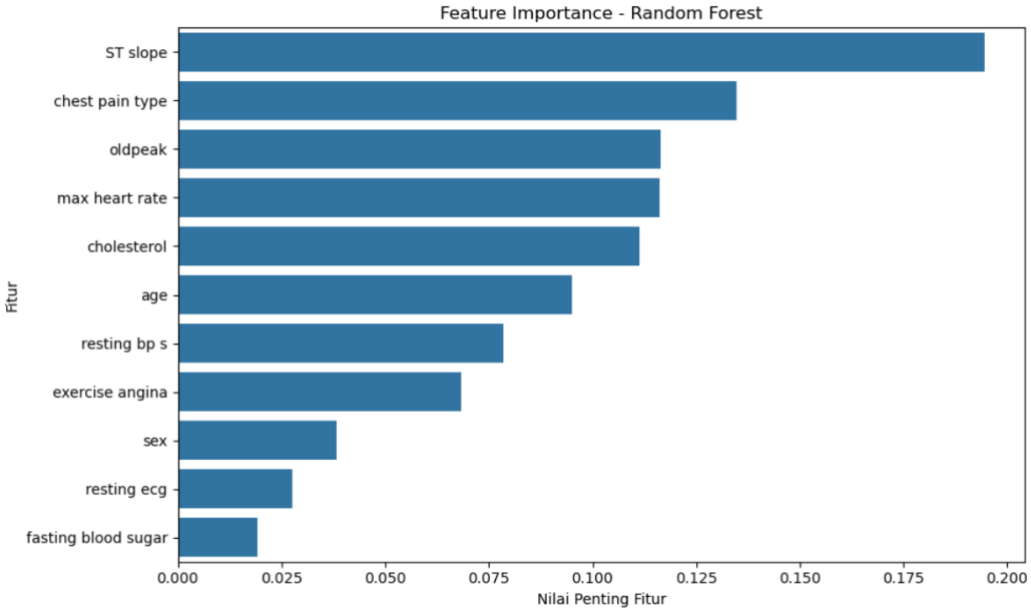

Fig. 7: Visualization of Prediction Results

Features with the highest weight include:
(1) max heart rate,
(2) chest pain type,
(3) oldpeak, and
(4) ST slope

These features align with the medical indicators commonly used in the diagnosis of heart disease, thereby enhancing the model's reliability.

To test the model's capability in real cases, predictions were made for 5 patients individually. Each patient's data is displayed vertically, showing the value of each attribute and the prediction result.

From the prediction results:

(1) There are patients predicted to be at risk because they have high old peak values and abnormal chest pain types.
(2) There are also patients predicted to be at low risk, with high heart rates and normal ST slopes.

This line-by-line interpretation facilitates the analysis of results and can be used as an initial reference in data-driven medical decision-making.

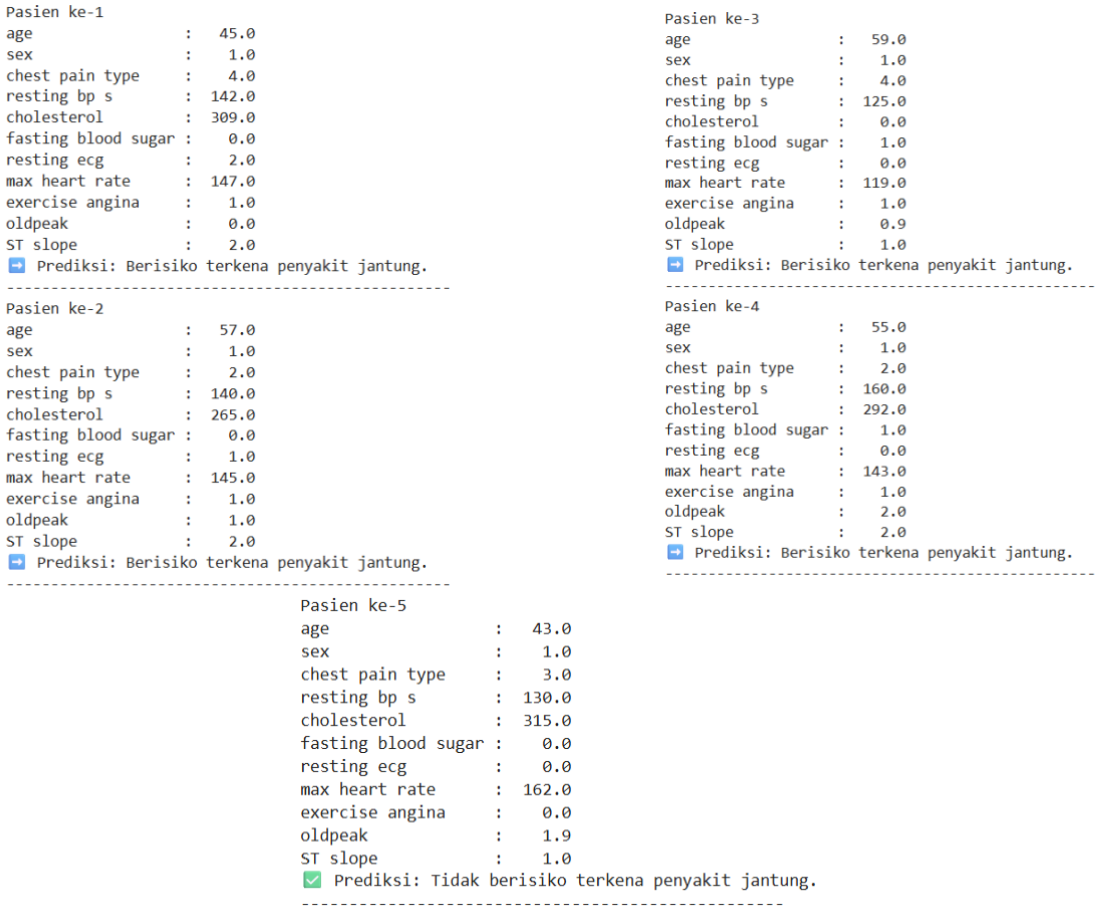

Fig. 8: Prediction Results

4. Conclusion

This research shows that the Random Forest algorithm is capable of building a heart disease prediction model with high accuracy. The training results achieved an accuracy of 100% on the training data and 82.35% on the test data. This indicates that the model has a fairly good generalization capability towards new data. The features that contribute the most to the predictions are max heart rate, chest pain type, old peak, and ST slope, which are also clinically relevant in the diagnosis of heart disease. Visualization of prediction results for individual patients enhances interpretability and supports the use of the model in medical practice.
In the future, model development can be focused on handling imbalanced data and integrating with hospital information systems so that this prediction system can be directly used by medical staff as a decision-support tool.

References

[1] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier," *Indonesian Journal of Computer Science*, vol. 12, no. 5, pp. 2995–3011, 2023, doi: 10.33022/ijcs.v12i5.3413.

[2] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *Informatik : Jurnal Ilmu Komputer*, vol. 18, no. 3, p. 239, 2022, doi: 10.52958/iftk.v18i3.4694.

[3] N. S. S. Silmi Ath Thahirah Al Azhima, D. Darmawan, N. Fahmi Arief Hakim, I. Kustiawan, M. Al Qibtiya, "Hybrid Machine Learning Model Untuk Memprediksi Penyakit," *Jurnal Teknologi Terpadu*, vol. 8, no. 1, pp. 40–46, 2022.

[4] A. S. Prabowo and F. I. Kurniadi, "Analisis Perbandingan Kinerja Algoritma Klasifikasi dalam Mendeteksi Penyakit Jantung," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 7, no. 1, pp. 56–61, 2023, doi: 10.47970/siskom-kb.v7i1.468.

[5] M. S. Ummah, *BUKU SAKU FARMAKOTERAPI JANTUNG*, vol. 11, no. 1. 2019. [Online]. Available: http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

[6] H. Hidayat, A. Sunyoto, and H. Al Fatta, "Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 7, no. 1, pp. 31–40, 2023, doi: 10.47970/siskom-kb.v7i1.464.

[7] S. P. Tamba and E. -, "Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random Forest," *Jurnal Sistem Informasi dan Ilmu Komputer Prima(JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 176–181, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445.

[8] J. Elektronik *et al.*, "Analisis Algoritma Random Forest Dalam Memprediksi Penyakit Jantung Koroner," vol. 11, no. 4, pp. 2654–5101, 2023, [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[9] N. H. Alfajr and S. Defiyanti, "METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA)," vol. 12, no. 3, 2024.