Comparative Analysis of K-Nearest Neighbors and Random Forest Algorithms in Predicting Heart Failure Disease

Yoga Suwindra[1*], Muhammad Halmi Dar[2], Fitri Aini Nasution[3]

[1,2,3] Faculty of Science and Technology, Universitas Labuhanbatu**,** Sumatera Utara Indonesia.
[*]Corresponding Author:
Email: yogasuwindra48@gmail.com

Abstract.

Heart failure is one of the cardiovascular diseases with a high mortality rate worldwide. Early detection and accurate prediction of heart failure risk are critical to improving patients' quality of life and reducing mortality. With the advancement of technology and increasingly available medical data, the use of machine learning algorithms for disease prediction has become a significant area of research. The purpose of this study is to compare the performance of the K-Nearest Neighbors and Random Forest algorithms in predicting heart failure. This study follows a systematic methodology starting with the collection of relevant medical data, followed by data preprocessing to ensure good data quality. The next stage is exploratory data analysis to understand the characteristics of the data. Next, we divide the data into training and testing sets, where we train and test the K-Nearest Neighbors and Random Forest models. We perform parameter optimization for each model to achieve optimal performance. Finally, we evaluate the model performance using accuracy metrics. The results show that Random Forest outperforms K-Nearest Neighbors in terms of prediction accuracy. The training accuracy for Random Forest reaches 98.80%, while for K-Nearest Neighbors it is 93.60%. In testing, Random Forest showed an accuracy of 96.50% compared to K-Nearest Neighbors, which reached 86.00%. The smaller decrease in accuracy in Random Forest indicates better generalization ability compared to K-Nearest Neighbors. The study's results indicate that the Random Forest algorithm outperforms K-Nearest Neighbors in heart failure risk prediction. Random Forest not only has higher accuracy but also shows better stability between training and testing data. The results of this study can be a reference for medical practitioners and researchers when choosing the right algorithm for predicting heart failure.

Keywords: *Heart Failure, K-Nearest Neighbors, Machine Learning, Prediction, Random Forest.*

1. INTRODUCTION

The heart is an important organ that pumps and circulates blood throughout the body so that all organs and tissues work properly. However, there are several disorders that can cause the heart to not function normally, such as heart failure [1]. According to data collected by the World Health Organization (WHO) and the Career Development Center (CDC), heart failure has been diagnosed in 26.6 million people in several developing countries, making it one of the most common causes of death worldwide in 2020. In Indonesia, there are 82 million people with heart failure, an increase of almost 24 percent compared to 2005 [2]. One of the challenges frequently encountered by patients and laypeople is their inability to accurately identify the causes and

components of heart failure, which hinders their ability to prevent the disease and accurately determine its cause. Because of this problem, it is necessary to predict the cause of heart failure in order to minimize the risk of developing it and to prevent it from occurring.

The advancement of technology in the health sector today can make medical personnel's tasks easier, such as providing services to the community and diagnosing diseases in patients [3]. Current technology enables the diagnosis of various diseases, including heart failure, a condition where the heart fails to pump blood properly [4]. Health technology advancements will undoubtedly transform the diagnosis of diseases and the classification of heart failure causes. This progress enables faster diagnosis through machine learning [5]. Artificial intelligence, known as machine learning, enables computers to learn from created data instead of relying on direct commands [6]. The Random Forest algorithm is derived from the decision tree approach of the Classification and Regression Trees (CART) method [7]. The K-Nearest Neighbors (KNN) algorithm is a classification method that assigns categories based on the majority of categories, as well as similarities between data sets and two vectors [8].

This study drew upon several previous studies on heart failure disease as references. Firstly, the study employs the Gaussian Naive Bayes algorithm to analyze data from heart failure patients, and it explores its application in categorizing heart failure disease in 100 patients [9]. Second, the study explores the relationship between physical ability and disease duration and the quality of life of congestive heart failure patients [10]. Thirdly, the research delves into the analysis of the KNN method's performance and cross-validation on heart disease data. This study discusses the measurement of the performance of the KNN method and crossvalidation on heart disease (accuracy, precision, recall, and f-measure) [11].

According to the presented references, it's crucial to initiate early treatment and categorize the causes of heart failure to prevent heart failure. Based on the background information, this study focuses on comparing the effectiveness of the KNN and Random Forest algorithms in predicting heart failure. This study uses the help of artificial intelligence and machine learning to classify the factors that cause heart failure with the KNN and Random Forest algorithms.

II. METHODS

We created a work procedure to ensure the smooth operation and timely completion of this research. Figure 1 displays the work procedure for this research.

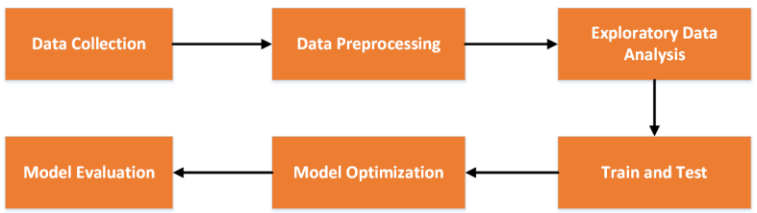

Fig. 1. Research methodology

Researchers collect data. In this study, they used a dataset of heart failure patients, consisting of 299 rows and 13 columns [12]. Data preprocessing involves transforming raw data into a more efficient format, with the goal of improving the created model and achieving accurate results [13]. Exploratory Data Analysis (EDA) is a data exploration process that aims to understand the data's contents and components. In this study, we performed EDA on the dataset to examine its contents, correlations, and distributions. The process of train and test involves processing the dataset with the KNN and Random Forest algorithms for data training, with the goal of assessing the accuracy of the generated model [14]. Model optimization involves enhancing the precision of a trained model using the KNN and Random Forest algorithms to achieve a more optimal model [15]. Model evaluation is the process of identifying a model that aims to optimize the trained and tested model to obtain a good accuracy value and to find the best model that represents our data [16].

III. RESULT AND DISCUSSION

In this study, the dataset consists of the following features: 'age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets', 'serum_Creatinine', 'serum_Sodium', ‘sex’, 'smoking', 'time', and 'death_event'. Figure 2 displays the heart failure patient dataset used in this study.

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets
0	75.0	0	582	0	20	1	265000.00
1	55.0	0	7861	0	38	0	263358.03
2	65.0	0	146	0	20	0	162000.00
3	50.0	1	111	0	20	0	210000.00
4	65.0	1	160	1	20	0	327000.00
...	...	...	...	...	...	...	...
294	62.0	0	61	1	38	1	155000.00
295	55.0	0	1820	0	38	0	270000.00
296	45.0	0	2060	1	60	0	742000.00
297	45.0	0	2413	0	38	0	140000.00
298	50.0	0	196	0	45	0	395000.00

299 rows × 13 columns

Fig. 2. Heart Failure Patient Dataset

Figure 2 shows a portion of the heart failure patient dataset used in this study. The dataset comprises several features relevant to the patient's heart failure medical condition. Each row represents one patient, with the following variables: The patient's age is measured in years. Anaemia is an indicator of whether the patient is anemic (0 = no, 1 = yes). Creatinine phosphokinase (mcg/L) is the level of the enzyme creatinine phosphokinase in the blood. Diabetes is an indicator of whether the patient has diabetes (0 = no, 1 = yes). Ejection fraction is the percentage of blood pumped out of the left ventricle with each heartbeat (%). High blood pressure is an indicator of whether the patient has high blood pressure (0 = no, 1 = yes). Platelets (kiloplatelets/mL) is the number of platelets in the blood. Serum_creatinine is the blood creatinine level

(mg/dL). The level of sodium in the blood (mEq/L) is denoted by serum_sodium. Sex is the patient's gender (1 = male, 0 = female). Smoking is an indicator of whether the patient smokes (0 = no, 1 = yes). Time measures the number of days from the patient's enrollment until the study's final outcome (death or end). Death_event is an indicator of whether the patient died during the study period (0 = no, 1 = yes). Next, we change the names of the features in the dataset to Indonesian.

```
df=df.drop_duplicates()

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Usia                  299 non-null    int64
 1   Anemia                299 non-null    int64
 2   Enzim                 299 non-null    int64
 3   Diabetes              299 non-null    int64
 4   Ef(%)                 299 non-null    int64
 5   Tekanan_Darah_Tinggi  299 non-null    int64
 6   Trombosit             299 non-null    int64
 7   serum_creatinine      299 non-null    int64
 8   serum_sodium          299 non-null    int64
 9   Jenis_Kelamin         299 non-null    int64
 10  Perokok               299 non-null    int64
 11  Waktu                 299 non-null    int64
 12  Kematian              299 non-null    int64
dtypes: int64(13)
memory usage: 32.7 KB
```

Fig. 3. Deduplication of Data

Figure 3 shows the process of removing duplicate data using the Python programming language on Google Colab. The process of removing duplicate data using the syntax "df=df.drop_duplicates()" aims to clean up the same data so that it does not affect the model during training and testing.

```
#Melihat jumlah kelas 0 dan 1
print(df['Kematian'].value_counts())
cls_0=df[df['Kematian']==0]
cls_1=df[df['Kematian']==1]
```

```
0    203
1     96
```

Fig. 4. View the number of classes

Figure 4 illustrates the program code, which in turn displays the number of classes. The process of viewing this class aims to see the number of 'death' classes that have a value of 0 and 1, where 0 is the number of patients who died while 1 is the number of patients who did not die. In this study, the number of deaths was 203, while those who did not die were 96.

```
cls_0=cls_0.sample(500,replace=True)
cls_1=cls_1.sample(500,replace=True)
df=pd.concat([cls_0,cls_1],axis=0)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000 entries, 174 to 67
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Usia                    1000 non-null   int64
 1   Anemia                  1000 non-null   int64
 2   Enzim                   1000 non-null   int64
 3   Diabetes                1000 non-null   int64
 4   Ef(%)                   1000 non-null   int64
 5   Tekanan_Darah_Tinggi    1000 non-null   int64
 6   Trombosit               1000 non-null   int64
 7   serum_creatinine        1000 non-null   int64
 8   serum_sodium            1000 non-null   int64
 9   Jenis_Kelamin           1000 non-null   int64
 10  Perokok                 1000 non-null   int64
 11  Waktu                   1000 non-null   int64
 12  Kematian                1000 non-null   int64
dtypes: int64(13)
memory usage: 109.4 KB
```

Fig. 5. Sampling Data

Figure 5 shows the data sampling process. In this study, we carried out data sampling on the 'death' class, using a sampling amount of 500 data points as a replacement statistical strategy to find patterns and trends in a larger data set.

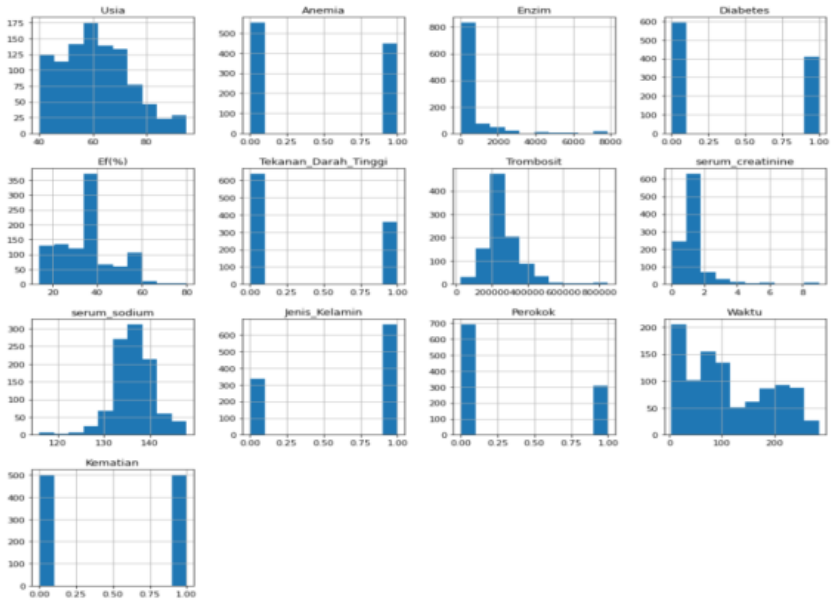

Fig. 6. Dataset Distribution

Figure 6 illustrates the distribution of data in the dataset using a histogram, which aims to simplify the interpretation of data on heart failure patients. The data is presented in the form of diagrams representing age, enzymes, ef(%), platelets, and

time. The histogram shows that there are 125 patients aged 40 years, 175 patients aged 60 years, and 49 patients aged 80 years.

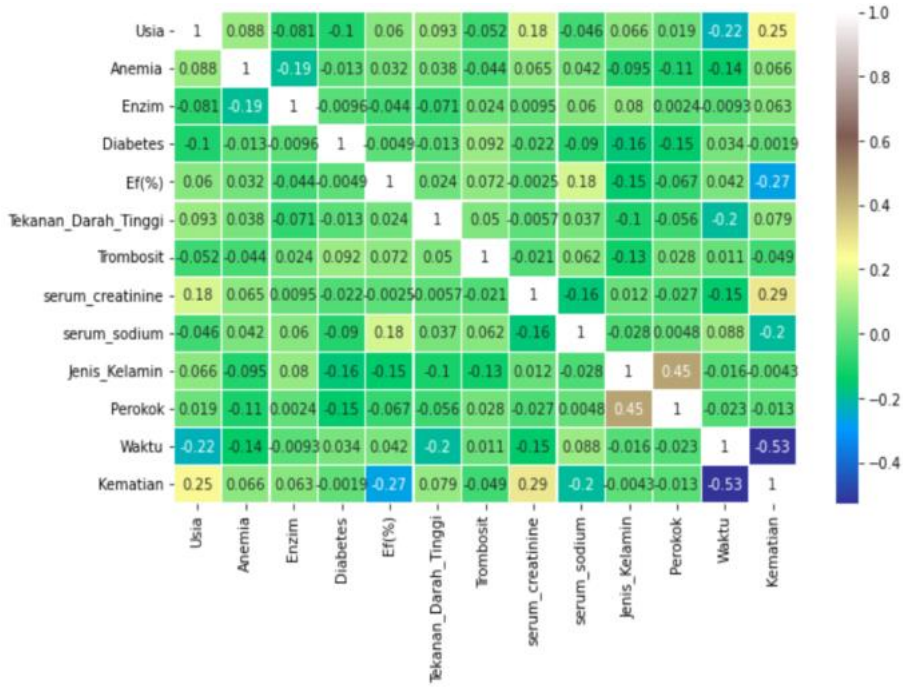

Fig. 7. Fitur Correlation

Figure 7 shows the correlation between the features in the dataset. Creating a correlation table aims to show the relationship between variables that are interrelated, with the provision that the closer to the number 1, the more related the variables are. As shown in Figure 13, the variables that are interrelated or related are the smoker variable and the gender variable.

```
#import library data partitioning/splitting data dan konfigurasi
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test= train_test_split
(X,Y, test_size=0.2, random_state=101,shuffle=True)
```

Fig. 8. Spliting Dataset

Figure 8 shows the process of separating the dataset into a training dataset and a testing dataset. To perform the train and test, we separate variables X and Y into two parts: variable X includes age, anemia, enzymes, diabetes, EF (%), high blood pressure, platelets, serum creatinine, serum sodium, gender, and smokers, while variable Y includes death and time. We use the Python scikit-learn library to create the train and test dataset. The results of the 80:20 division of the train and test with random state 101 are presented.

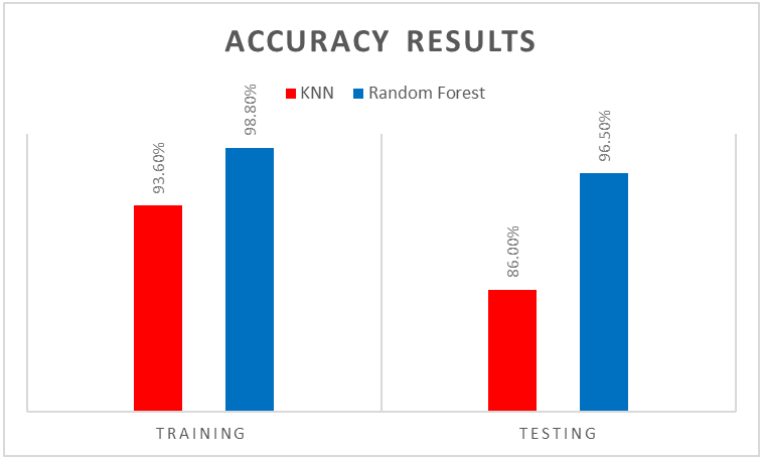

Fig. 9. Comparison of Accuracy Results of KNN vs Random Forest

Figure 9 displays a contrast in the precision of two algorithms, K-Nearest Neighbors (KNN) and Random Forest, when they were trained and tested to predict heart failure disease. KNN shows quite high accuracy results during the training phase, which is 93.60%. This shows that KNN is able to recognize patterns well in the training data. Random Forest has a higher training accuracy compared to KNN, which is 98.80%. This shows that Random Forest is more effective in capturing the complexity of the training data. KNN experienced a decrease in accuracy from training to testing by 7.6% (93.60% to 86.00%). This could suggest a slight overfitting or difficulty in generalizing to previously unseen data in KNN. Random Forest experienced a decrease in accuracy from training to testing by 2.3% (98.80% to 96.50%). Random Forest showed more stable and consistent performance on both training and testing data, indicating better generalization capabilities. Random Forest is superior with 98.80% accuracy compared to KNN, which is only 93.60% in the training process. This indicates that Random Forest is more effective in capturing patterns from training data. Random Forest remains superior with 96.50% accuracy compared to KNN with 86.00% in the testing process. This shows that Random Forest is better at training and testing. Although KNN has excellent accuracy, this algorithm experiences a significant decrease in accuracy from training to testing, which could indicate problems with generalization. The Random Forest algorithm shows superior performance in both training and testing, with a minimal decrease in accuracy, indicating better generalization ability to previously unseen data.

IV. CONCLUSION

With the advancement of technology in the health sector, it is anticipated that machine learning based on classification algorithms will assist patients in understanding the factors that cause heart failure, enabling early prediction of heart failure parameters. The processes carried out include renaming columns, deleting duplicate data, viewing the number of classes 0 and 1, and data sampling. We carry out

exploratory data analysis by examining the data distribution, creating correlation tables, gender distributions, and box plot diagrams. The analysis reveals that there are 125 patients aged 40 years, 175 patients aged 60 years, and 49 patients aged 80 years. There is a relationship between the smoker variable and the gender variable. We have created a machine learning model that can identify the factors causing heart failure in patients. Based on the analysis of the two K-Nearest Neighbors (KNN) algorithms and the Random Forest algorithm, the Random Forest algorithm had the best accuracy results at 96.5%.

Future research and development of technology in the health sector should continue to facilitate medical personnel in providing services and diagnosing diseases, including heart failure. The issue of public and patient ignorance about the causes and factors causing heart failure should be addressed through classification and education. We concentrate on the early management of heart failure by assessing and contrasting the effectiveness of K-Nearest Neighbors (KNN) and Random Forest algorithms in addressing these issues.

REFERENCES

[1] A. P. Lumi, V. F. F. Joseph, and N. C. I. Polii, "Rehabilitasi Jantung pada Pasien Gagal Jantung Kronik," *J. BiomedikJBM*, vol. 13, no. 3, p. 309, 2021, doi: 10.35790/jbm.v13i3.33448.

[2] S. Suraji, A. C. Fauzan, and H. Harliana, "Penerapan Algoritma Decision Tree C5.0 untuk Memprediksi Tingkat Kematian Pasien Penyakit Gagal Jantung," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 02, pp. 216–222, 2022, doi: 10.46772/intech.v4i02.682.

[3] S. Sutrisno, "Sistem Pakar Mendiagnosa Penyakit Gagal Jantung pada Manusia dengan Menggunakan Metode Certainty Factor Berbasis Web," *J. Ilmu Komput. dan Sist. Inf.*, vol. 5, no. 1, pp. 20–27, 2022, doi: 10.55338/jikomsi.v5i1.207.

[4] A. Desiani, M. Akbar, I. Irmeilyana, and A. Amran, "Implementasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Pada Klasifikasi Penyakit Kardiovaskular," *J. Tek. Elektro dan Komputasi*, vol. 4, no. 2, Aug. 2022, doi: 10.32528/elkom.v4i2.7691.

[5] D. Prihatiningsih and T. Sudyasih, "Perawatan Diri pada Pasien Gagal Jantung," *J. Pendidik. Keperawatan Indones.*, vol. 4, no. 2, 2018, doi: 10.17509/jpki.v4i2.13443.

[6] T. Wahyono, *Fundamental of Python for Machine Learning: Dasar-Dasar Pemrograman Python untuk Machine Learning dan Kecerdasan Buatan*. Gava Media Yogyakarta, 2018.

[7] M. R. Adrian, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi Random Forest dan SVM pada Analisis Sentimen PSBB," *J. Inform. UPGRIS*, vol. 7, no. 1, pp. 36–40, 2021, doi: 10.26877/jiu.v7i1.7099.

[8] M. R. Noviansyah, T. Rismawan, and D. M. Midyanti, "Penerapan Data Mining menggunakan Metode K-Nearest Neighbor untuk Klasifikasi Indeks Cuaca Kebakaran berdasarkan Data AWS (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya)," *J. CODING*, vol. 6, no. 2, pp. 48–56, 2018, doi: 10.26418/coding.v6i2.26672.

[9] Q. Hasanah, H. Oktavianto, and Y. D. Rahayu, "Analisis Algoritma Gaussian Naive Bayes terhadap Klasifikasi Data Pasien Penderita Gagal Jantung," *Smart Teknol.*, vol. 3, no. 4, 2022.

[10] Haryati, Saida, and L. Rangki, "Kualitas Hidup Penderita Gagal Jantung Kongestif berdasarkan Derajat Kemampuan Fisik dan Durasi Penyakit," *Faletehan Heal. J.*, vol. 7, no. 02, pp. 70–76, 2020, doi: 10.33746/fhj.v7i02.134.

[11] I. P. Putri, "Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 21–28, 2021,

doi: 10.33096/ijodas.v2i1.25.
[12] Z. T. Anggara and M. F. Dzulqarnain, "Visualisasi Data Citra untuk Klasifikasi Kalimantan's Batik Production Menggunakan Neural Network," Politeknik 'Aisyiyah Pontianak, 2022.
[13] H. Said, N. H. Matondang, and H. N. Irmanda, "Penerapan Algoritma K-Nearest Neighbor untuk Memprediksi Kualitas Air yang Dapat Dikonsumsi," *Techno.com*, vol. 21, no. 2, 2022, doi: 10.33633/tc.v21i2.5901.
[14] S. Amos, "When Training and Test Sets Are Different: Characterizing Learning Transfer," *Dataset Shift Mach. Learn.*, pp. 2–28, 2013, doi: 10.7551/mitpress/9780262170055.003.0001.
[15] Amalia, M. Radhi, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, "Prediksi Harga Mobil Menggunakan Algoritma Regressi dengan Hyper-Parameter Tuning," *JUSIKOM PRIMA*, vol. 4, no. 2, pp. 28–32, 2021, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2479.
[16] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations," *Am. Soc. Agric. Bilogical Eng.*, vol. 50, no. 3, pp. 885–900, 2007.