Published by: CV. HAWARI

International Journal of Computer Sciences and Mathematics Engineering

Journal homepage: www.ijecom.org

ANALYSIS OF HEART FAILURE PREDICTION WITH RANDOM FOREST ALGORITHM AND LINEAR REGRESSION

Ismar Hidayat[1], Muhammad Iqbal[2], Leni Marlina[3], Andysah Putera Utama Siahaan[4], Zulham Sitorus[5].

Ism4r.h@gmail.com[1], muhammadiqbal@dosen.pancabudi.ac.id[2], lenimarlina@dosen.pancabudi.ac.id[3], andiesiahaan@gmail.com[4], zulhamsitorus@dosen.pancabudi.ac.id[5]

[1,2,3,4,5] Universitas Pembangunan Panca Budi, Indonesia

Article Info

Article history:

Received Feb 28, 2025
Revised Mar 01, 2025
Accepted Mar 03, 2025

Keywords:

Predictions
Heart Failure
Random Forest
Linear Regression
Pembelajaran mesin

ABSTRACT

Predicting the risk of heart failure is an important step in the prevention and early treatment of potentially fatal cardiovascular diseases. This study aims to compare the performance of two machine learning algorithms, namely Random Forest and Linear Regression, in predicting heart failure based on patient data that includes variables such as age, blood pressure, cholesterol levels, and other health history. The results show that the Random Forest algorithm is significantly superior in terms of prediction accuracy compared to Linear Regression, especially on data with a pattern of the number of data used. However, Linear Regression remains relevant in providing more stable results on differences in the amount of data used and has a more significant effect on the variables of heart failure. Therefore, a Random Forest-based prediction model is recommended to predict heart failure if it has a large amount of tranning data, and Linear Regression is recommended for prediction stability. The implementation of this model is expected to help medical practitioners in making more appropriate and accurate decisions to prevent the occurrence of heart failure in high-risk patients.

Corresponding Author:

Ismar Hidayat,
Magister Teknologi Informasi,
Universitas Pembangunan Panca Budi Medan, Indonesia
Alamat: Jl. Jend. Gatot Subroto Km. 4,5 Sei Sikambing 20122, Kota Medan, Propinsi Sumatera Utara, Indonesia.
Email: ism4r.h@gmail.com

1. INTRODUCTION

Heart disease is one of the number one killer diseases in the world and in Indonesia which reaches 17.8 million deaths or one in three deaths in the world every year. Heart failure can be defined as an abnormality of the structure or function that causes the failure of the heart to distribute oxygen throughout the body[11]. Heart failure prediction can be made possible by using one of the branches of artificial intelligence (AI), namely by utilizing machine learning to be able to make predictions or predictions. The classification or regression method contained in machine learning is used in data processing for early prediction of heart failure which is expected to improve the survival rate for patients. Machine learning methods can be used to make predictions or predictions using classification or regression methods in data processing, where heart failure in patients can be predicted early.

Data processing with machine learning carried out by (Ashoka, 2022) the implementation of classification using the k-NN, Decision Tree, and Random Forest algorithms has been carried out using water objects using an accuracy comparison of data carried out with the title "Automation and Analysis of Prediction Results of Clean Water Quality Research Between Classifiers Using Machine Learning." states that the k-NN, Decision Tree, and Random Forest algorithms obtained similar values 1.0 or 100% [3]. And what was done by (Nanik, Sarfiah, 2021) with the title "Random Forest Classifier for the Detection of COVID-19 Patients CT Scan Images." With the results of the algorithm, Random Forest has the highest accuracy value compared to other methods with an accuracy result of 96.9% [12] . and it can be concluded that the use of extraction in this study affects the classification results [6](Normah, B Rifai 2022).

This machine learning method has been used in research in predicting heart failure, one of which is by Polat et al. Anooj, P. K. by comparing 2 methods, namely wieghted fuzzi rules and Neural Network Ensemble where in the study the results of the Neural Network Ensemble method with an accuracy level of 89.01% are superior to Wieghted fuzzy rules [1][2]. In the research conducted by Samuel, O. W., et al using the ANN-Fuzzy_AHP method using the same dataset, the results obtained a better accuracy level of 91.10% compared to the previous study [1][2]. In the research conducted by Samuel, O. W., et al using the ANN-Fuzzy_AHP method using the same dataset, the results obtained a better accuracy level of 91.10% compared to the previous study [9].

Based on the description that has been explained above, machine learning is able to predict well [3], However, there has been no research that uses the Random Forest and Linear Regression methods as the techniques used in predicting heart failure data and looking at some of the comparison of accuracy results from the use of classification methods and expected regression of heart failure prediction, this research was carried out to be able to analyze heart failure prediction by using the Random Forest and Regression Liner algorithms to measure the accuracy value with the parameters that are in the dataset.

2. RESEARCH METHOD

2.1. *Dataset*

The variables used in this study are derived from 13 Independent Variables that can be seen in table 3.2 that determine the prediction results, and 1 Dependent Variable (Target) which has a value of 1 (There is Heart Disease) / 0 (No Heart Disease).

Tabel 1. Independent Variables (Input Features)[9]

No	Atribut	Kode Atribut	Alternative	Range
1	Age (years)	AGE	Young, Medium, Old, Very old	<33, 34-40, 41-52, >52
2	Sex	SEX	Male, Female	1, 0
3	Chest pain type	CP	Typical angina, Atypical angina, Non-angina pain, Asymptomatic	1, 2, 3, 4

4	Resting blood Pressure	TRESTBPS	Low, Medium, High, Very high	<128, 128-142, 143-154, >154
5	Serum cholesterol	CHOL	Low, Medium, Very High	<188, 189-217, 218-281, >182
6	Fasting blood sugar	FBS	True (>120mg/dl), False (<120mg/dl)	1, 0
7	Resting electrocardiographic results	RESTECG	Normal, St-T abnormal, Hypertrophy	0, 1, 2
8	Maximum heart rate achieved	THALAC	Low, Medium, High	<112, 112-152, <152
9	Exercise induced angina	EXANG	True, False	1, 0
10	Old peak	OPK	Low, Risk, Terrible	<1.5, 1.5-2.55, >2.55
11	Peak exercise slope	SLOPE	Upsloping, Flat, Downsloping	1, 2, 3
12	Bunber of majir vessels colored by fluoroscopy	CA	Fluoroscopy-0, Fluoroscopy-1, Fluoroscopy-2, Fluoroscopy-3, Fluoroscopy-4	0, 1, 2, 3 4
13	Thallium scan	THAL	Normal, Fixed Defect, Reversible defect	1, 2, 3

Dengan keterangan sebagai berikut :

1. Age = Age
2. Sex = Gender
3. CP = Type of chest pain = Range 0 - 3 how much chest pain the patient has
4. trestbps = Blood pressure at rest (Unit in Hg)
5. chol = Serum Cholesterol Levels in mg/dl
6. fbs = Fasting blood sugar> 120 mg/dl ? (1 = Yes, 0 = No)
7. restecg = Electrocardiography results after rest (value 0,1,2)
8. thalach = Detak jantung maksimum yang dicapai
9. exang = Experiencing angina (chest pain) after exercise? (1= Yes, 0 = No)
10. oldpeak = ST segment depression caused by relative exercise
11. slope = Gradient of ST segment

12. ca = Number of major blood vessels (0-3) stained with fluoropyethal
13. thal = Declining Disease / Genetic Thalassemia : 1 = normal; 2 = deformity ; 3 = Correctable defect (carrier)

2.2. *Split Dataset*

The dataset is divided into 5 groups with 2 subsets, namely training data and test data, which are as follows,

1. 70% (index 0 - 716) totaled 717 train data, and 30% (index 717 - 1024) totaled 308 test data (X_train_1, X_test_1, y_train_1, y_test_1)
2. 30% (index 0 - 306) totaled 308 train data, 70% (index 307 - 1024) totaled 717 test data (X_train_2, X_test_2, y_train_2, y_test_2)
3. 50% (index 0 – 511) totaling 513 train data, 50% (index 512 – 1024) totaling 512 test data (X_train_3, X_test_3, y_train_3, y_test_3)
4. 90% (index 0 - 921) totaling 922 train data, 10% (index 922 - 1024) totaling 103 test data (X_train_4, X_test_4, y_train_4, y_test_4)
5. 10% (index 0 – 101) totals 102 train data, and 90% (index 102 – 1024) totals 923 test data (X_train_5, X_test_5, y_train_5, y_test_5)

2.3. *Machine learning*

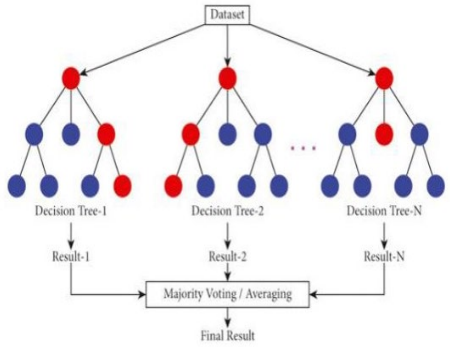
Machine Learning is a study to make computers can make decisions automatically using various algorithms that can be improved in accuracy through the learning process with historical data [7]. Machine learning is also concerned with how a computer can build the performance of its program with multiple tasks through experience. Supervised learning is learning to predict targets based on a training model using labeled data [4]. Then it can be classified into the main categories, namely classification and regression. Classification is a data approach used to estimate data instances. Regression is a technique in machine learning and is commonly used for two theories. First, regression analysis is used to predict. Second, regression analysis is used to determine the causal relationship between independent and dependent variables[5].

2.4. *Random Forest*

Random Forest is one of the types of classification algorithms that consists of more than one decision tree where each decision tree will be formed depending on the random vector values of the sample independently and identically distributed equally for all trees[10]. Random forests are also usually trained using the 'bagging' method which is a collection of several decision trees that have a combination of learning models to be able to improve the overall data results (DQLab 2022). Suppose a classifier ensemble $h1(x)$, $h2(x)$, … , $hK(x)$, and with the training data randomly selected from the random distribution Y and X, the margin function ($mg(X, Y)$) of Random Forest is defined in the equation below.

$$mg(X,Y) = \frac{\sum_{K=1}^{K} I(h_K(X) = Y)}{K} - \max_{j \neq Y} \left[\frac{\sum_{K=1}^{K} I\big(h_K(X)\big) = j}{K}\right]$$

where I is the indication function and K is the number of trees. The structure of the Random Forest is depicted in the image below [6].



Gambar 1. Structure of Random Forest

2.5. Linear Regression

Regression analysis is an analysis method that can be used to analyze data and draw meaningful conclusions about the relationship of variable dependence on other variables. If the regression analysis equation involves two or more independent variables, then this regression is called multiple linear regression analysis[9]. The general form of multiple linear regression can be expressed statistically as follows:

$$\mathrm{y} = \mathrm{a} + \mathrm{b}_1\mathrm{x}_1 + \mathrm{b}_2\mathrm{x}_2 + \ldots. + \mathrm{b}_\mathrm{n}\mathrm{x}_\mathrm{n}$$

- y: Dependent variables.
- x1,x2,...,xn: Independent variables.
- a: Intercept.
- b1,b2,...,bn: Regression coefficients for each independent variable.

2.6. Evaluation Criteria

The performance of the random forest and linear regression models was evaluated with an accuracy level based on training data and testing data, then in the linear regression was added the evaluation of Mean Square Error (MSE), and R2 Score. For the results of the prediction analyzed through True Positive (TP) is a prediction for heart disease patients who are found to have heart disease, and True Negative (TN) is a prediction for patients without heart disease who are found not to have heart disease, False Negative (FN) is a prediction for patients without heart disease who are found to have heart disease, and False Positive (FP) is a prediction for heart disease patients who are found not to have heart disease.

$$\mathrm{MSE} = \frac{1}{n}\sum_{n-1}^{n}(y_i - \hat{y}_i)^2$$

Keterangan:

- y_i: Actual value (ground truth).
- $\hat{y}_i$: Prediction value.
- n: Amount of data

$$R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Keterangan:

- yi: Actual value.
- $\hat{\mathrm{y}}_\mathrm{i}$: Prediction value.

- $\bar{y}$: Average actual value ($\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$)
- n: Amount of data.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Keterangan:

- True Positive (TP).
- True Negative (TN).
- False Negative (FN)
- False Positive (FP)

3. RESULTS AND DISCUSSIONS

3.1 Random Forest & Linear Regression Prediction Results

In this section, the results and evaluations will be presented. The results of the heart failure prediction study using the random forest method obtained the highest level of accuracy shown in the data (90:10) (index 922 – 1024) with an accuracy level of 100% with the number of test data 103 and showing higher when compared to other test data, and the lowest level in data sharing (10:90) with an accuracy of 82.77%. With a total of 923 test data. In table 2, it can be seen that the division of training data and data sets greatly affects the accuracy level of the random forest method.

Tabel 2 Random Forest Model Prediction Results

Kategori	Akurasi	TP	TN	FP	FN	Total
data (10:90)	82,77%	377	387	100	59	923
data (30:70)	90,25%	325	323	37	33	718
data (50:50)	92,98%	231	246	22	14	513
data (70:30)	96,10%	149	147	5	7	308
data (90:10)	100,00%	56	47	0	0	103

The results of the linear regression model prediction showed the highest accuracy in the data group (50:50) (index 512 – 1024) showing an accuracy level of 82.46% with a total of 513 test data, the lowest accuracy level in the data group (70:30) with a total of 308 test data. In table 3, the Linear Regression model shows that the level of accuracy is not determined by the amount of training data used.

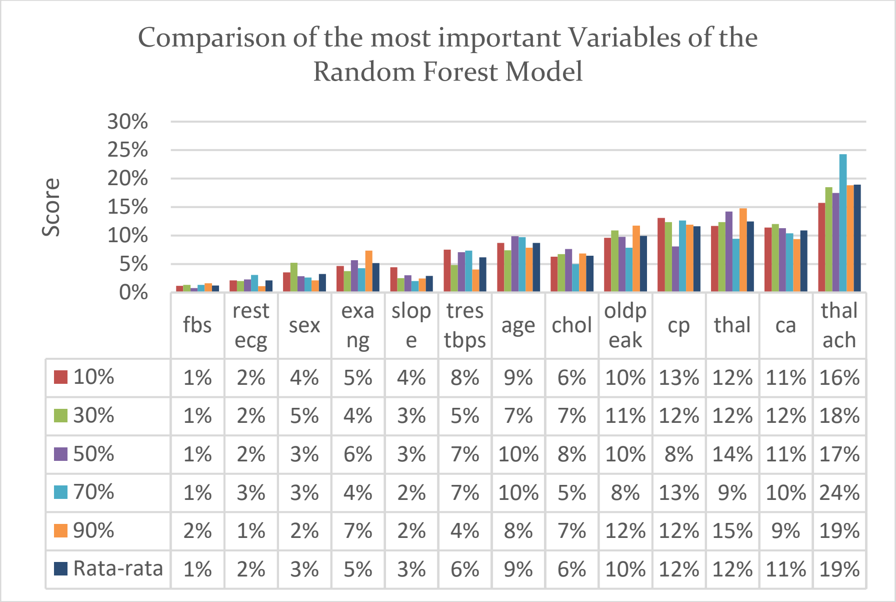
Tabel 3 Linear Regression Model Prediction Results

Kategori	R2 Score	MSE	Akurasi	TP	TN	FP	FN	Total
data (10:90)	46%	0,132	82,02%	757	166	403	354	923
data (30:70)	44%	0,139	81,48%	585	133	323	262	718
data (50:50)	45%	0,138	82,46%	225	198	28	62	513
data (70:30)	40%	0,148	79,55%	132	113	22	41	308
data (90:10)	43%	0,141	79,61%	47	35	9	12	103

E-ISSN 2962-4274

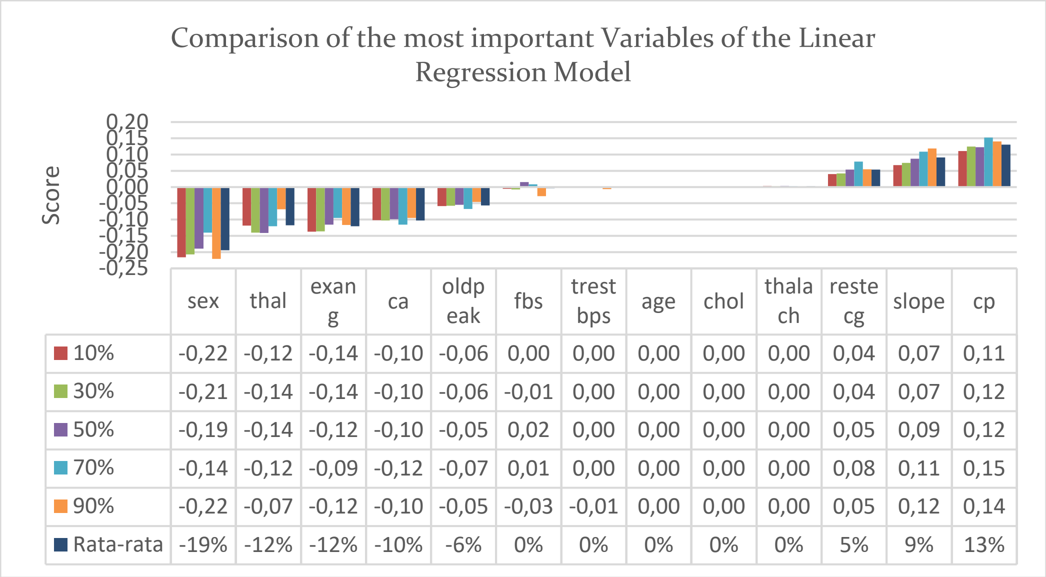
3.2 Variable Analysis

The Random Forest model algorithm in the data division of random split 1 and sequential indices gave quite good results with an accuracy value of 96.10%, from 13 data variables. The variable that had the greatest influence on the Random Forest algorithm was thalach with an average of 18% and in the Linear Regression algorithm it was cp with an average of 13%. Furthermore, there are variables with small values in the Random Forest algorithm which are fbs, restecg, and sex which can be seen in figure 2, while in the linear regression algorithm are sex, thal, and exang which can be seen in figure 3.



	fbs	rest ecg	sex	exa ng	slop e	tres tbps	age	chol	oldp eak	cp	thal	ca	thal ach
🟥 10%	1%	2%	4%	5%	4%	8%	9%	6%	10%	13%	12%	11%	16%
🟩 30%	1%	2%	5%	4%	3%	5%	7%	7%	11%	12%	12%	12%	18%
🟪 50%	1%	2%	3%	6%	3%	7%	10%	8%	10%	8%	14%	11%	17%
🟦 70%	1%	3%	3%	4%	2%	7%	10%	5%	8%	13%	9%	10%	24%
🟧 90%	2%	1%	2%	7%	2%	4%	8%	7%	12%	12%	15%	9%	19%
⬛ Rata-rata	1%	2%	3%	5%	3%	6%	9%	6%	10%	12%	12%	11%	19%

Gambar 2 Comparison of variable values in random forest algorithm



	sex	thal	exan g	ca	oldp eak	fbs	trest bps	age	chol	thala ch	reste cg	slope	cp
🟥 10%	-0,22	-0,12	-0,14	-0,10	-0,06	0,00	0,00	0,00	0,00	0,00	0,04	0,07	0,11
🟩 30%	-0,21	-0,14	-0,14	-0,10	-0,06	-0,01	0,00	0,00	0,00	0,00	0,04	0,07	0,12
🟪 50%	-0,19	-0,14	-0,12	-0,10	-0,05	0,02	0,00	0,00	0,00	0,00	0,05	0,09	0,12
🟦 70%	-0,14	-0,12	-0,09	-0,12	-0,07	0,01	0,00	0,00	0,00	0,00	0,08	0,11	0,15
🟧 90%	-0,22	-0,07	-0,12	-0,10	-0,05	-0,03	-0,01	0,00	0,00	0,00	0,05	0,12	0,14
⬛ Rata-rata	-19%	-12%	-12%	-10%	-6%	0%	0%	0%	0%	0%	5%	9%	13%

Gambar 3 Comparison of variable values in Linear Regression algorithm

3.3 Method Comparison

The Random Forest model showed the highest performance with 100% accuracy in the group with data sharing (90:10) and showed an average accuracy of 92.42%, where this result was better when compared to the Linear Regression model with the highest accuracy of 82.46% on and with data sharing (50:50) and showed an average accuracy of 81.02%. This shows that Random Forest is better able to handle non-linear relationships between features and targets in heart failure datasets.

Then the test by eliminating variables that lack a predictive impact uses the Random Forest and Linear Regression algorithms, namely the sex variable. The follow-up test used 12 independent variables (age, fbs, restecg, exang, cp, testfbs, chol, thalac, oldpeak, slope, ca, and thal), which gave results in the random forest model had a slightly increased average accuracy of 92.43% but in the linear regression model the accuracy level was higher in the test using 13 variables resulting in an average value of 81.02% while in the 12-variable test it decreased to 79.64%.

Tabel 4 Variable Model 13 and 12 Performance Accuracy Levels

Kategori	Jumlah Variabel	Data (10:90)	Data (30:70)	Data (50:50)	Data (70:30)	Data (90:10)	Rata-rata
Akurasi RF	**13**	82,77%	90,25%	92,98%	96,10%	100,00%	92,42%
Akurasi RL	**13**	82,02%	81,48%	82,46%	79,55%	79,61%	81,02%
Akurasi RF	**12**	86,13%	86,35%	93,57%	96,10%	100,00%	92,43%
Akurasi RL	**12**	81,15%	79,25%	79,92%	78,25%	79,61%	79,64%

4. CONCLUSION

The Random Forest algorithm tends to provide higher accuracy in predicting heart failure with an average accuracy of 92.42% for 12 variables and 92.43% compared to Linear Regression which gets an accuracy rate of 81.02% for 13 variables and 79.64% for 12 variables. This is because Random Forest can handle the complexity of data, such as blood pressure, cholesterol levels, age, and other medical histories that are non-linear and feature that has a complicated relationship with the target variable. The results of the analysis showed that the difference in the number of training data affected the results in the Random Forest algorithm, while in the linear regression algorithm, the difference in the number of training data did not affect the accuracy of the prediction results.

REFERENCES (10 PT)

[1] Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), 27-40.

[2] Anooj, P. K. (2013, December). Implementing decision tree fuzzy rules in clinical decision support system after comparing with fuzzy based and neural network based systems. In 2013 International Conference on IT Convergence and Security (ICITCS) (pp. 1-6). IEEE.

[3] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680

[4] Erlangga, A. W. Otomasi dan analisis hasil prediksi penentuan kualitas air bersih antar classifier menggunakan machine learning (Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta).

[5] Hariyadi, yaya and Wahyuno, Teguh (2020). Machine Learning : Konsep dan Implementasi. Penerbit Gava Media Yogyakarta.

[6] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018, May). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE symposium on security and privacy (SP) (pp. 19-35). IEEE.

[7] Kamila, S. A., Sulistijowati, R. S., & Susanto, I. (2023, January). Classification of Heart Disease Using Decision Tree and Random Forest. In Seminar Nasional Teknologi & Sains (Vol. 2, No. 1, pp. 7-12).

[8] Normah, N., Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. Jurnal Teknik Komputer, 8(2), 174-180.

[9] Nursyafitri, Gifa Delyani (2022). Machine learning. https://dqlab.id/machine-learning-model-untuk-prediksi-data-2022 (accessed Oct. 11, 2024).

[10] Rahmadeni, R., & Anggreni, D. (2014). Analisis jumlah tenaga kerja terhadap jumlah pasien RSUD Arifin Achmad Pekanbaru menggunakan metode Regresi Gulud. SITEKIN: Jurnal Sains, Teknologi dan Industri, 12(1), 48-57.

[11] Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert systems with applications, 68, 163-172

[12] Suryanegara, G. A. B., & Purbolaksono, M. D. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(1), 114-122.

[13] Thesiana, Yumi, (2024). Gagal Jantung. Kementrian Kesehatan Kementerian Kesehatan Republik Indonesia https://ayosehat.kemkes.go.id/gagal-jantung (accessed Oct. 11, 2024).

[14] Utomo, D. P., Sirait, P., & Yunis, R. (2020). Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5. 0. Jurnal Media Informatika Budidarma, 4(4), 994-1006.

[15] Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science, 120, 588-593.

[16] Wuryani, N., & Agustiani, S. (2021). Random Forest Classifier untuk Deteksi Penderita COVID-19 berbasisCitra CT Scan. Jurnal Khatulistiwa Informatika, 7(2), 187-193.