

## The Comparison of ReliefF and C.45 for Feature Selection on Heart Disease Classification Using Backpropagation

Anita Desiani<sup>\*1</sup>, Yuli Andriani<sup>2</sup>, Irmeilyana<sup>3</sup>, Rifkie Primartha<sup>4</sup>, Muhammad Arhami<sup>5</sup>,  
Dwi Fitrianti<sup>6</sup>, Henny Nur Syafitri<sup>7</sup>

<sup>1,6,7</sup>Mathematics, Sriwijaya University, Indonesia

<sup>2,4</sup>Technical Information, Sriwijaya University, Indonesia

<sup>3,5</sup>Technical Information, Politeknik Negeri Lhokseumawe, Indonesia

e-mail: <sup>\*1</sup>anita\_desiani@unsri.ac.id, <sup>2</sup>yuliandriani@unsri.ac.id, <sup>3</sup>irmeilyana@unsri.ac.id,

<sup>4</sup>rifkie@ilkom.unsri.ac.id, <sup>5</sup>muhammad.arhami@pnl.ac.id, <sup>6</sup>dwifitrianti28@gmail.com,

<sup>7</sup>henny.nur.sya@gmail.com

### Abstrak

Salah satu dataset yang digunakan untuk klasifikasi penyakit jantung adalah dataset UCI. Sayangnya, dataset ini memiliki data hilang. Backpropagation merupakan metode yang mudah dan cepat, namun sangat bergantung pada input data sehingga jika terdapat data hilang dapat menurunkan kinerja backpropagation. Salah satu teknik yang digunakan untuk menangani data hilang adalah seleksi fitur. Penelitian ini membandingkan algoritma ReliefF dan C4.5 pada seleksi fitur untuk menangani data hilang. Tujuan dari perbandingan adalah untuk menemukan teknik terbaik dalam mengatasi data hilang dengan melakukan seleksi fitur untuk meningkatkan kinerja backpropagation pada klasifikasi penyakit jantung menggunakan Backpropagation. Hasil yang diperoleh diukur berdasarkan akurasi, presisi, dan recall. Hasil kinerja dari ReliefF dan Backpropagation diatas 82%. Hasil kinerja C4.5 dan backpropagation rata-rata 80,54%. Berdasarkan hasil tersebut disimpulkan bahwa ReliefF memberikan kinerja yang lebih baik pada backpropagation dibandingkan C4.5. ReliefF juga mampu menangani data hilang dengan melakukan pemilihan fitur untuk meningkatkan kinerja backpropagation pada klasifikasi penyakit jantung dibandingkan dengan C4.5. Meskipun algoritma C4.5 mampu memberikan peningkatan performa pada backpropagation, namun C4.5 kurang tepat digunakan sebagai metode pemilihan fitur untuk menangani data yang hilang.

**Kata Kunci**— Backpropagation, C4.5, Feature Selection; Missing Data, ReliefF

### Abstract

One of the datasets used to classify heart disease is UCI dataset. Unfortunately, the dataset contains missing data. Backpropagation is an easy and fast method, but it is very dependent on input data so if there is missing data, it can reduce the performance of the backpropagation. One of the techniques used to handle missing data is feature selection. This study compares ReliefF and C4.5 algorithm in feature selection. The purpose of the study is to find way in overcoming missing data by feature selection to improve backpropagation performance in the heart disease classification. The results of these algorithms are applied to the classification by Backpropagation. The results will be measured based on accuracy, precision, and recall. The performance results of the ReliefF and Backpropagation are above 82%. The performance results of of C4.5 and backpropagation are 80.54% on average for accuracy, recall and precision. Based on the results it can be concluded the ReliefF gives better

*performance on backpropagation than C4.5. ReliefF is also able to handle missing data by performing feature selection to improve the performance of the backpropagation method for heart disease classification compared to C4.5. Although the C4.5 algorithm is able to provide increased performance on backpropagation, C4.5 is not appropriate to be used as a feature selection method for handling missing data.*

**Keywords**— *Backpropagation, C4.5, Feature Selection; Missing Data, ReliefF*

## 1. INTRODUCTION

Classification is the process of classifying an object based on a specific model [1]. Classification is often used in various fields, such as the fields of health, environment, and data investigations [2]. The application of classification has been widely used in the health sector, one of which is to predict heart disease. Heart disease is currently one of the biggest causes of death in the world [2]. Heart disease is caused by narrowing of the arteries which function to deliver nutrients and oxygen to the heart [3]. Some of the factors that influence heart disease are high cholesterol levels, high blood pressure, obesity, heredity, and others [4]. Heart disease can be predicted by carrying out an electrocardiogram test, in order to predict early symptoms of heart disease by measuring and recording heart activity [5]. The importance of classification in data mining and machine learning is that it can be used to concluding obscure classes using overall sample learning [6].

One of the dataset that is often used in research to classify heart disease prediction is dataset from the University of California Irvine (UCI) which can be accessed at <https://archive.ics.uci.edu/ml/datasets/heart+disease>. In this dataset there are 76 features for diagnosing heart disease, of the 76 features only published by UCI are 14 features namely age, gender, type of chest pain, blood pressure, cholesterol, fasting blood sugar, electrocardiography, maximum heart rate, induced angina, oldpeak, ST slope, fluoroscopy, heart rate type and label features consisting of healthy and sick. Features in dataset UCI have missing data, those are fluoroscopy (ca) as many as 291 data, heart rate type (thal) as many as 266 data, ST slope as many as 166 data, cholesterol (chol) as many as 23 data, fasting blood sugar (fbs) as many as 8 data, and electrocardiographic, blood pressure, maximum heart rate, angina exercise has 1 data. Missing data causes a lack of information and affects classification performance [7] [8]. Handling of missing data can be solved by several methods, namely feature selection, missing data imputation, deletion and etc. feature selection is a technique that reduces the number of features, removes irrelevant, redundant, or noisy data, speeds up the classification algorithm, and improves the classification performance [9]. Missing data can be handled using feature selection by removing features that have the least amount of information or have the most missing data.

Feature selection works to eliminate features that are not important [10]. One of method that can be used to select the most important features is the ReliefF algorithm. ReliefF was ranked each features by using the weight from each feature. The greater weight of a feature, the more important feature in a dataset [11]. ReliefF algorithm is a filter-based feature selection model that is widely applied and has great classification efficiency [12]. The advantages of this algorithm are it does not limit the data types used and can effectively deal with multi-class problems, missing data, and noise tolerance [13]. Several studies use the ReliefF algorithm to handle missing data in their studies. Baliarsingh et al., [14] applied ReliefF and used Support Vector Machine (SVM) to classify several diseases, one of which is Colon tumor disease with an accuracy of 84.26%. Unfortunately, this research only calculates the result of accuracy. Q. Liu et al., [15] combining ReliefF and Random Forest (RF) algorithms to diagnoses dermatology data in the health sector. The study resulted the accuracy of 91.82%, but they did not measure precision and recall. Yahdin et al., [16] used ReliefF algorithm to find the most

important features of prediction of educational background relevance with jobs after graduating from school. This study used Naive Bayes and KNN methods to measure the success rate of ReliefF algorithm. The results of the accuracy of the data before the feature selection process for Naive Bayes method were 73.43% and KNN method were 66.24%. After the feature selection process the accuracy obtained in both methods increased to 74.38% for the Naive Bayes method and 72.22% for the KNN method. However, this study did not measure the RMSE value and the results of the accuracy were still in the pretty good category.

Another method that can provide information on the importance of a feature is the C4.5 algorithm. Algorithm C4.5 is a classification algorithm based on a tree algorithm, where the results of each root and sub-branch show the features that most influence the classification of the dataset. The features that have no effect will not appear in the tree of C4.5. C4.5 algorithm is one of the most frequently used classification algorithms for making decisions [17]. Algorithm C4.5 has several advantages, namely it can work with a smaller training dataset than is required for perfect accuracy, and it can make decisions that are more accurate as a result. [18]. Suyatno, Nhita & Rohmawati [19]) applied the C4.5 algorithm to rainfall forecasts in Bandung district. In this study, before using the C4.5 algorithm it resulted the accuracy of 60%. After features selection, the accuracy increase to 93.33%. Pujianto et al., [20] The prediction study of inpatients with diabetes in the hospital used the C4.5 algorithm to obtain an accuracy of 82.74%, a precision of 87.1% and a sensitivity of 82.7%. Prasetyo & Prasetyo [21] conducted research on heart disease diagnoses using the C4.5 algorithm and obtained an accuracy of 72.98% and then combined with Bagging and obtained an accuracy of 81.84%. Datasets with complete data are better at classifying things than datasets with any missing data.

Backpropagation is a simple, fast, easy to analyze and flexible classification method. Backpropagation is an algorithm that use to classify a new weight for each feature and minimize the number of errors between the actual value and the predicted value [22]. The advantage of backpropagation is that it has good computational properties, especially when processing data on a large and complex scale [23]. Unfortunately, backpropagation is a very sensitive algorithm and depends on the conditions of the input data. If there are features that contain missing data, it can seriously affect the performance degradation of the backpropagation method. If the problem of missing data is not resolved properly it can result in overfitting. Backpropagation uses random numbers to help learn or classify patterns from available input data. If data is not available, backpropagation will assume that the input value is zero, it can affect the classification results. [8]. Zhang et al., [24] used the Backpropagation Neural Network (BNN) to study heart disease and colorectal cancer. It had an accuracy of 75%, a precision of 70% and a recall of 80%. Mhatre & Varma [25] conducted a study to predict heart disease. The study produced an accuracy of 78.76%. However, the study does not calculate other evaluation performance measures. Al-Barzinji et al., [26] used the backpropagation method to predict heart disease and the results were very good. It had an accuracy result of 82.17%, a precision of 81%, and a recall of 79%.

This study combines feature selection algorithms to overcome missing data and classification method for heart disease classification. In the UCI dataset there are several attributes that have missing data with different percentages. For feature selections, this study applies the ReliefF and the C4.5 algorithms. The results of feature selection will be classified using the backpropagation method. This comparison was carried out to find which algorithm can help handle missing data problems and improve the performance of the backpropagation method in the classification of heart disease. This study also compares the performance results between backpropagation classification with feature selection and the performance results of backpropagation classification without using feature selection in the UCI dataset. The UCI dataset is a heart disease dataset which has some features that contain missing data. In this study, performance comparisons were made based on the results of accuracy, precision and recall obtained in each classification either with feature selection or without feature selection. The results of this study are expected to be able to provide recommendations for the best technique that can be used to overcome missing data in classification using the

backpropagation method. The results of this study can be applied to various datasets that have missing data problems.

## 2. METHODS

There are four stages in the research method, namely:

### 2.1 Definition of dataset

The data used in this study is in the form of secondary data which is data from patients with heart disease obtained from the UCI Machine Learning Repository which can be downloaded at <https://archive.ics.uci.edu/ml/datasets/heart+disease>. The features used in the prediction of heart disease patients are 14 features, there are, age, sex, cp, chol, flaurosopy, fbs, induced angina, electrocardiography, tresrbps, thalac, oldpeak, ST slope or slope, and heart rate. An explanation of each feature and the amount of missing data for each feature is described in Table 1. In Table 1, it can be seen that the features that have the most missing data are the ca or flaurosopy features of 97.97%, the thal or heart rate type features of 90.48%, and the ST slope feature of 56.46%. Some other features such as cholesterol and fasting blood sugar (fbs) only have some missing data, cholesterol features of 7.82% and fbs features is only 2.72%. The rest of the features in the dataset only experience a small amount of missing data, with an average of 0.15%.

### 2.2 Feature Selection

Feature selection consists of two steps. The steps are:

#### 2.2.1 The ReliefF Algorithm

The steps of the ReliefF algorithm are as follows:

- i. Initialize all weights with a value of 0.
- ii. Determine the probability value of each class appearing in the data. The probability value is used to calculate the weight of the features that are calculated using the ReliefF algorithm.
- iii. Calculate the k-near hit and k-near miss values with Equation 1 [27].

$$dist(p, q) = \sum_{i=1}^n |p_i - q_i| \quad (1)$$

- iv. Calculate the diff value using Equation 2 which is then used in calculating the weights [28].

$$diff(A, R_i, R_j) = \frac{|\text{nilai}(A, R_i) - \text{nilai}(A, R_j)|}{\max(A) - \min(A)} \quad (2)$$

- v. Calculate the weights for each feature, using Equation 3 [29].

$$w[A] = w[A] - \sum_{j=1}^k \frac{diff(A, R_i, H_j)}{m \cdot k} + \sum_{C \neq \text{class}(R_i)} \left[ \frac{P(C)}{1 - P(\text{class}(R_i))} \sum_{j=1}^k \frac{diff(A, R_i, M_j(C))}{m \cdot k} \right] \quad (3)$$

- vi. Sort the results of the weight value (weight) of each feature from the largest to the smallest (rank).

Compare the accuracy values in each experiment by removing features from the smallest weight. Issue features that have been reduced based on rank and the accuracy value of the influence of these features on labels.

Table 1. Definition Features

No.	Feature name	Definitions	Data type	Missing data
1.	age	year	continuous	-
2.	sex	0 = female 1 = male	discrete	-
3.	chestpain (cp)	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic	discrete	-
4.	resting blood pressure (restbtps)	mm Hg (taken at the time of admission to the hospital)	continuous	1
5.	cholesterol (chol)	in mg/dl	continuous	23
6.	fasting blood sugar(fbs)	Blood sugar > 120 mg/dl 1 = true 0 = false	discrete	8
7.	resting electrocardiographicresults (restecg)	0 = normal 1 = have a wave disorder ST-T 2 = showed left ventricular hypertrophy	discrete	1
8.	maximum heart rate (thalac)		continuous	1
9.	angina exercise (exang)	0 = no 1 = yes	discrete	1
10.	Old peak or depression induced by exercise relative to rest		continuous	-
11.	slope (ST)	1 = tilt up 2 = flat 3 = tilt down	discrete	114
12.	ca or flaurosopy	0-3 value	discrete	291
13.	heart rate type (thal)	3= normal 6= fixed 7= reversible defect	discrete	266
14.	Num (heart disease diagnosis label)	0 = <50% diameter narrowing (healthy) 1 = >50% diameter narrowing (sick)	discrete	-

### 2.2.2 C4.5 Algorithm

The application of Algorithm C4.5 can be applied as follows:

- i. Count the number of cases and the number of occurrences.
- ii. Calculate the total entropy value based on using Equation 4.

$$Entropy(S) = \sum_{i=1}^n -p_i \times \log_2 p_i \quad (4)$$

- iii. Calculate the gain value with Equation 5.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{S_i}{S} \times Entropy(S_i) \quad (5)$$

- iv. Find the highest gain value for each feature in the heart disease dataset.
- v. Obtain a conclusion which features have the most influence in predicting heart disease.

### 2. 3 Classification with Backpropagation Neural Network

The features that have been selected using the ReliefF and C4.5 algorithms will be tested using the Backpropagation Neural Network method. Backpropagation works by passing information from the input layer to the hidden layer, and then from the hidden layer to the output layer. The architecture of backpropagation can be seen in Figure 1 [30].

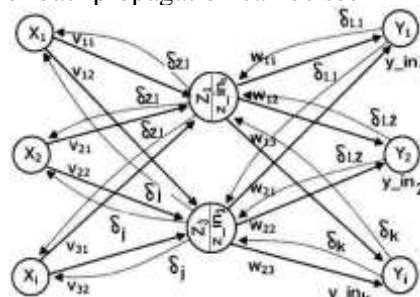


Figure 1. Backpropagation Neural Network Architecture

## 2.4 Evaluation

The results obtained from these stages will be compared using several classification performances. Classification performance can be measured using a confusion matrix, namely accuracy, precision, and recall [31].

### a. Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (6)$$

### b. Precision

$$\text{Precision} = \frac{TP}{FP + TP} \times 100\% \quad (7)$$

### c. Recall

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (8)$$

The entire stages carried out at the study can be seen in Figure 2.

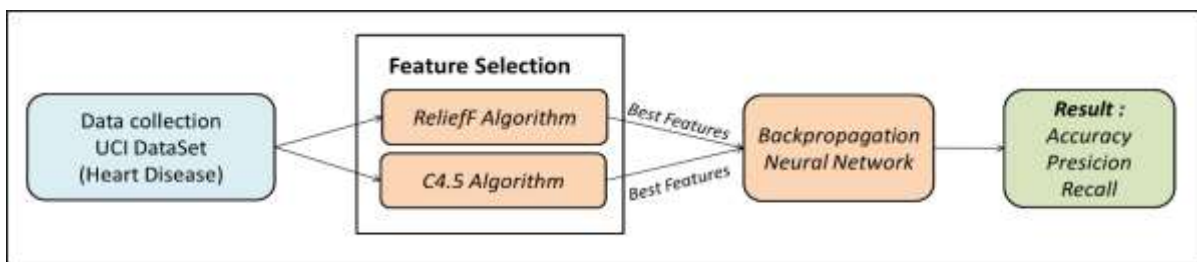


Figure 2 Research Stages in The Study

In addition to evaluating the results of accuracy, precision, and recall, this study will also use RMSE (Root Mean Squared Error) to calculate how accurate the test results are. RMSE is a measure of how accurate a prediction model is. It is calculated by adding up the squared errors from all of the predictions made by the model. This can help to show how big a error was generated by the model when making predictions. Equation 9 is the formula used to calculate RMSE [32].

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2}; \quad 0 \leq \text{RMSE} \leq \infty \quad (9)$$

## 3. RESULT AND DISCUSSION

The UCI dataset consists of 14 features, 13 independent features and 1 label feature for classifying heart disease. In the dataset there are several features that experience missing data, including ca (291 data), thal (266 data), slope (144 data), chol (23 data), fbs (8 data), tresbps (1 data), restecg (1 data), thalac (1 data), and exang (1 data). In this dataset there are 9 features that contain missing data. It can affect the performance of the classification. The classification method used in this study is backpropagation. Before the study applies backpropagation for heart disease classification, the first step is feature selection using the ReliefF and C4.5 algorithms.

### 4.1 ReliefF Algorithm

In the feature selection process, ReliefF algorithm calculates the weight of each feature used in heart disease patient data. The resulting weights for each feature are sorted based on the largest to the smallest feature weight values. The feature selection process reduces the dimensions of the data on heart disease patients based on the weight values for each feature. ReliefF algorithm calculates the weights of the features using Equation 3, the largest weight

is shown first and the smallest weight is shown last. The results of the calculation are shown in Table 2. From Table 2 it can be seen that the features that have the most missing data are slope features, and ca has the lowest weight. The highest weight is obtained by the Cp feature. Cp feature is a feature that does not contain missing data. The greater the weight, the more important the effect of the feature on the dataset. Conversely, the smaller the weight obtained, the less the influence of the feature in the dataset. To determine the number of features to be used in classification, classification is carried out using the backpropagation method by eliminating a single feature unit based on the lowest weight. The results of applying ReliefF and backpropagation can be seen in table 3.

From table 3 it can be seen that the lowest accuracy, precision and recall are obtained when classification uses all features and when eliminating 5 features so that only 8 features are used in backpropagation classification. The results of the highest accuracy, precision and recall were obtained by backpropagation classification with 9 features and eliminating the 4 lowest attributes namely slope, thal, ca and chol. This shows that there are 9 features that most influence the performance of backpropagation classification in heart disease namely "cp", "sex", "oldpeak", "exang", "restecg", "age", "thalach", and "trestbps" although those features still contain missing data.

Table 2 The Weight Value of Each Feature with ReliefF Algorithm

No.	Weight	Feature Description
1.	$1,4025 \times 10^{-1}$	<i>cp</i>
2.	$1,1871 \times 10^{-1}$	<i>sex</i>
3.	$8,121 \times 10^{-2}$	<i>oldpeak</i>
4.	$3,878 \times 10^{-2}$	<i>exang</i>
5.	$3,061 \times 10^{-2}$	<i>restecg</i>
6.	$2,083 \times 10^{-2}$	<i>age</i>
7.	$2,032 \times 10^{-2}$	<i>thalach</i>
8.	$1,758 \times 10^{-2}$	<i>trestbps</i>
9.	$1,088 \times 10^{-2}$	<i>fbs</i>
10.	$6,09 \times 10^{-3}$	<i>chol</i>
11.	$-0,1 \times 10^{-18}$	<i>ca</i>
12.	$-1,811 \times 10^{-2}$	<i>thal</i>
13.	$-3,605 \times 10^{-2}$	<i>slope</i>

Table 3. The Results of Comparison of the Accuracy Values for Each Feature Selection Using the ReliefF Algorithm

No.	Features	Features used	Feature Selection	Accuracy	Precision	Recall
				Backpropagation		
1.	13	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs, chol, ca, thal, slope</i>	-	77,211%	77,1%	77,2%
2.	12	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs, chol, ca, thal</i>	<i>slope</i>	81,296%	81,1%	81,3%
3.	11	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs, chol, ca</i>	<i>slope, thal</i>	80,952%	80,7%	81,0%
4.	10	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs, chol</i>	<i>slope, thal, ca</i>	80,952%	80,7%	81,0%
5.	9	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs</i>	<i>slope, thal, ca, chol</i>	<b>82,653%</b>	<b>82,7%</b>	<b>82,7%</b>
6.	8	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps</i>	<i>slope, thal, ca, chol, fbs</i>	77,211%	77,1%	77,2%

### 3. 2. C4.5 Algorithm

In this study, the C4.5 algorithm was used to calculate the gain and entropy values by using equations 4 and 5. The result of C4.5 algorithm decision tree can be seen in Figure 3.

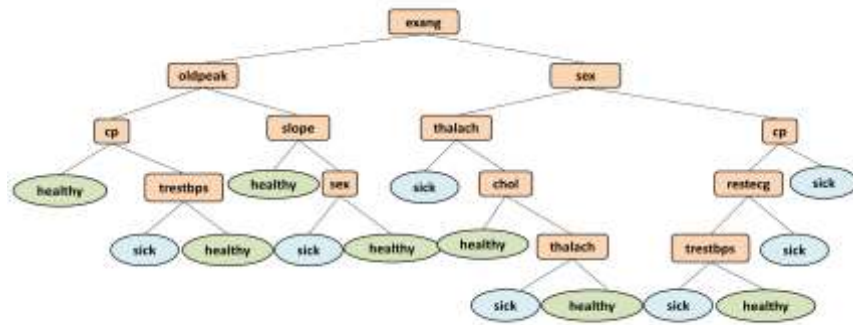


Figure 3. The Result of C4.5 Algorithm Decision Tree on Feature Selection

The tree results in Figure 3 give different results from the results from ReliefF where the results of C4.5 show that there are 10 very important features in the classification of heart disease. This is different from the results obtained by ReliefF. the age feature in ReliefF is a feature that includes an effect on the UCI dataset but in C4.5, the age feature has absolutely no effect. the chol feature contains 23 missing data but in Figure 3 it can be seen that the chol feature has an influence on the dataset. It can be concluded that C4.5 is not suitable for dealing with missing data problems. ReliefF algorithm is more suitable for dealing with missing data.

3. 3 Implementation of Backpropagation

In this study, the Backpropagation architecture uses 3 learning layers, namely the input layer, the hidden layer and the output layer. The features that have been selected by the ReliefF and C4.5 algorithms will be applied to the Backpropagation classification. Implementation of Backpropagation on the UCI dataset will produce a confusion matrix. Confusion matrix is a matrix used to measure the success of classification. This study will also compare the results of backpropagation classification before feature selection, and after feature selection. The results of the confusion matrix obtained can be seen in table 4.

Table 4. Confusion Matrix of Backpropagation Method on Heart Disease Classification

Method	Confusion Matrix		Features	Features used
	Predicted Values	Actual Values		
Before feature selection	healthy	healthy	13	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs, chol, ca, thal, slope.</i>
	sick	sick		
	healthy	171		
After feature selection uses the ReliefF Algorithm	healthy	healthy	9	<i>cp, sex, oldpeak, exang, restecg, age, thalach, trestbps, fbs.</i>
	sick	sick		
	healthy	173		
After feature selection uses the C4.5 Algorithm	healthy	healthy	10	<i>exang, oldpeak, sex, cp, slope, thalach, trestbps, chol, restecg.</i>
	sick	sick		
	healthy	163		
	sick	32	74	

Based on Table 4, the following information are obtained:

1. Before feature selection, it was found that 171 instants of True Positive (TP), 50 instants of False Negative (FN), 17 instants of False Positive (FP), and 56 instants of True Negative (TN).
2. After feature selection uses the ReliefF Algorithm, it iss found that 173 instants of True Positive (TP), 37 instants of False Negative (FN), 15 instants of False Positive (FP), and 69 instants of True Negative (TN).



3. After feature selection uses the C4.5 Algorithm, it was found that 163 instants of True Positive (TP), 32 instants of False Negative (FN)s, 25 instants of False Positive (FP), and 74 instant of True Negative (TN).

From Table 4, it can be calculated the results of accuracy, precision, and recall using Equations (6), (7), and (8). The results of accuracy, precision, and recall can be seen in Figure 4. For RMSE (Root Mean Squared Error) is calculated by Equation (9). The results of RMSE can be seen in Figure 5.

### 3.4 Comparison of the Results Between the ReliefF Algorithm and C4.5 Algorithm with Backpropagation

The backpropagation method was used to compare the performance of three different methods: without handling missing data, with handling missing data by ReliefF algorithms, and with handling missing data by C4.5 algorithms. The results are shown in Figure 4.

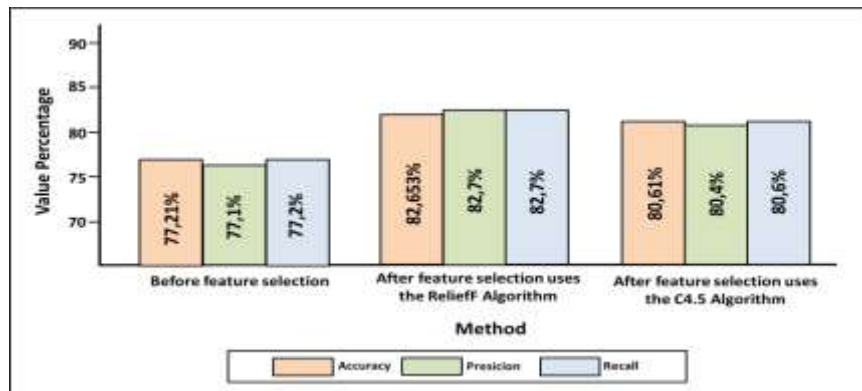


Figure 4 Graph of Comparison of the Application ReliefF Algorithm and the C4.5 Algorithm with Test Results Using the Backpropagation Neural Network (BNN) Method

It can be seen that the ReliefF algorithm on backpropagation has better accuracy, precision, and recall results than the C4.5 algorithm on backpropagation. The improvement obtained by C4.5 algorithm are accuracy of 3.4%, precision of 3.3%, and recall of 3.4%. On the other hand, ReliefF algorithm improves the results of accuracy by 5.44%, precision by 5.6%, and recall by 5.5%. Based on Figure 4, it can be conclude that the results of ReliefF algorithm on backpropagation is better than C4.5. The study looked at how well different algorithms of feature selection work when dealing with missing data. It is tested by backpropagation method, which is compared with other studies. The results of this comparison are shown in Table 5.

Table 5. Comparison of Research Results with Other Studies

Research	Dataset	Accuracy	Precision	Recall
Zhang <i>et al.</i> (2016)	Heart Disease & Colorectal Cancer	75%	70%	80%
Mhatre & Varma (2019)	Heart Disease	78,76%	-	-
Al-Barzinji <i>et al.</i> (2020)	Heart Disease	82,17%	81%	79%
Without handling missing data	UCI Dataset Heart Disease	77,21%	72,31%	76,60%
Handling missing data with ReliefF		<b>82,65%</b>	<b>82,7%</b>	<b>82,7%</b>
Handling missing data with C4.5		80,61%	80,4%	80,6%

Note: the value in bold is the highest value

The studies in Table 5 used the same classification system, namely the backpropagation method. The datasets used in each study were all from the same type, the heart disease dataset. Based on Table 7, the study that uses ReliefF algorithm produces the highest accuracy value, followed by research by Al-Al-Barzinji *et al.* (2020) with a difference of 0.48%. The results with the greatest precision were also obtained by ReliefF algorithm, with a good category, followed by research by Al-Barzinji *et al.* (2020) and research of C4.5 algorithm. However, the study by Al-Barzinji *et al.* (2020), the result of recall is not better than the results are achieved

by the C4.5 algorithm. In table 7 it can be seen that handling missing data can improve the performance of the backpropagation method in the classification of heart disease. The application of each algorithm also produces RMSE (Root Mean Squared Error) that calculated using Equation 6. The results of RMSE can be seen in Figure 5.

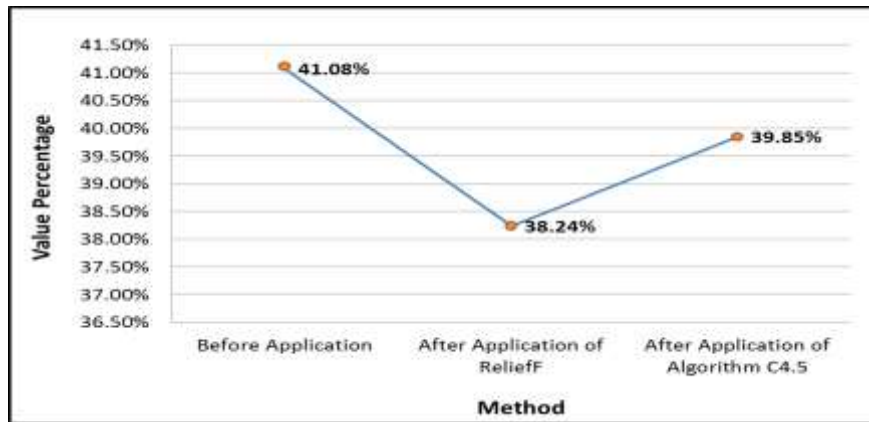


Figure 5. The Comparison Root Mean Squared Error (RMSE) Results of ReliefF-Backpropagation and C4.5-Backpropagation

In Figure 5, the RMSE in ReliefF-Backpropagation produces a lower RMSE than C4.5 algorithm. The RMSE is the magnitude of the prediction error rate, where the smaller the RMSE, the more accurate the prediction results will be. Based on the picture above, it can be said that ReliefF algorithm can improve the performance of backpropagation on heart disease classification. The results show that handling missing data can improve classification performance. The results of feature selection with C4.5 show that C4.5 is able to show the most influential features in a dataset but is not sensitive if there is missing or damaged data. The results of implementing feature selection C4.5 and backpropagation classification are quite good but do not provide a significant increase in backpropagation performance on heart disease classification. From these results it can be interpreted that ReliefF is the best algorithm compared to C4.5 in selecting features to overcome the problem of missing data. Algorithm C4.5 is a classification method that works by opening nodes based on the importance of each feature, easy to implement, but unfortunately C4.5 has not been able to handle the existence of missing data. It can reduce the performance of backpropagation in the classification of heart disease.

#### 4. CONCLUSIONS

The ReliefF algorithm and the C4.5 algorithm can help improve the accuracy, precision and recall of the backpropagation method in the classification of heart disease. The ReliefF algorithm and the C4.5 algorithm can improve backpropagation performance with feature selection. ReliefF can improve backpropagation performance. This increase in performance can be seen from an increase in accuracy of 5.44%, an increase in precision of 5.6%, and an increase in recall of 5.5%. the implementation of ReliefF provides improved performance on backpropagation which is better than C4.5. the feature selection results using C4.5 are still able to improve the performance of Backpropagation but are said to be still not significant. In addition, the results of selecting features using C4.5 were not able to cope with missing data features. It can be seen from the decision tree produced by C4.5 which still contains features that have quite a lot of missing data. From the results obtained that the handling missing data can improve the performance of the backpropagation method in classification problems, especially in heart disease dataset. Feature selection is one technique that can be used to handle missing data. The ReliefF algorithm is the best choice for performing feature selection to improve backpropagation classification performance.

## REFERENCES

- [1] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An Improved Method to Construct basic Probability Assignment based on The Confusion Matrix for Classification Problem," *Inf. Sci. (Ny)*, vol. 340–341, pp. 250–261, 2016, doi: 10.1016/j.ins.2016.01.033.
- [2] 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.
- [3] T. Johnson, L. Zhao, G. Manuel, H. Taylor, and D. Liu, "Approaches to therapeutic angiogenesis for ischemic heart disease," *J. Mol. Med.*, vol. 97, no. 2, pp. 141–151, 2019, doi: 10.1007/s00109-018-1729-3.
- [4] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *J. Biomed. Inform.*, vol. 58, pp. S171–S182, 2015, doi: 10.1016/j.jbi.2015.09.006.
- [5] C. M. Otto *et al.*, "2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *J. Am. Coll. Cardiol.*, vol. 77, no. 4, pp. e25–e197, 2021, doi: 10.1016/j.jacc.2020.11.018.
- [6] Z. F. Hussain *et al.*, "A new model for iris data set classification based on linear support vector machine parameter's optimization," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, pp. 1079–1084, 2020, doi: 10.11591/ijece.v10i1.pp1079-1084.
- [7] A. F. Costa, M. S. Santos, J. P. Soares, and P. H. Abreu, *Missing data imputation via denoising autoencoders: The untold story*, vol. 11191 LNCS. Springer International Publishing, 2018.
- [8] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, "LSTM-based traffic flow prediction with missing data," *Neurocomputing*, vol. 318, pp. 297–305, 2018, doi: 10.1016/j.neucom.2018.08.067.
- [9] J. Zhang, Y. Xiong, and S. Min, "A new hybrid filter/wrapper algorithm for feature selection in classification," *Anal. Chim. Acta*, vol. 1080, no. 2, pp. 43–54, 2019, doi: 10.1016/j.aca.2019.06.054.
- [10] K. M. Lang and T. D. Little, "Principled missing data treatments," *Prev. Sci.*, vol. 19, no. 3, pp. 284–294, 2018, doi: 10.1007/s11121-016-0644-5.
- [11] V. Vakharia, V. K. Gupta, and P. K. Kankar, "Efficient fault diagnosis of ball bearing using ReliefF and Random Forest classifier," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 39, no. 8, pp. 2969–2982, 2017, doi: 10.1007/s40430-017-0717-9.
- [12] L. Sun, X. Kong, J. Xu, Z. Xue, R. Zhai, and S. Zhang, "A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, 2019, doi: 10.1038/s41598-019-45223-
- [13] O. Osanaiye, O. Ogundile, F. Aina, and A. Periola, "Feature selection for intrusion detection system in a cluster-based heterogeneous wireless sensor network," *Facta Univ. - Ser. Electron. Energ.*, vol. 32, no. 2, pp. 315–330, 2019, doi: 10.2298/fuee1902315o.
- [14] S. K. Baliarsingh, W. Ding, S. Vipsita, and S. Bakshi, "A memetic algorithm using emperor penguin and social engineering optimization for medical data classification," *Appl. Soft Comput. J.*, vol. 85, p. 105773, 2019, doi: 10.1016/j.asoc.2019.105773.
- [15] Q. Liu, X. Xu, Y. Tao, and X. Wang, "An Improved Decision Tree Method Base on RELIEFF for Medical Diagnosis," *Proc. - 2016 Int. Conf. Digit. Home, ICDH 2016*, pp. 133–138, 2017, doi: 10.1109/ICDH.2016.037.
- [16] S. Yahdin, A. Desiani, N. Gofar, K. Agustin, and D. Rodiah, "Application of the Relief-f Algorithm for Feature Selection in the Prediction of the Relevance Education Background with the Graduate Employment of the Universitas Sriwijaya," *Comput. Eng. Appl.*, vol. 10, no. 2, pp. 71–80, 2021.
- [17] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetyo, and S. Alimah, "Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis," *J.*

- Phys. Conf. Ser.*, vol. 983, no. 1, 2018, doi: 10.1088/1742-6596/983/1/012063.
- [18] A. Cherfi, K. Nourira, and A. Ferchichi, "Very fast C4.5 decision tree algorithm Cherfi, A., Nourira, K., & Ferchichi, A. (2018). Very fast C4.5 decision tree algorithm. *Applied Artificial Intelligence*, 32(2), 119–137. <https://doi.org/10.1080/08839514.2018.1447479>," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 119–137, 2018, doi: 10.1080/08839514.2018.1447479.
- [19] J. A. Suyatno, F. Nhita, and A. A. Rohmawati, "Rainfall forecasting in Bandung regency using C4.5 algorithm," *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, vol. 0, no. c, pp. 324–328, 2018, doi: 10.1109/ICoICT.2018.8528725.
- [20] U. Pujiyanto, A. L. Setiawan, H. A. Rosyid, and A. M. M. Salah, "Comparison of naïve bayes algorithm and decision tree C4.5 for hospital readmission diabetes patients using HbA1c Measurement," *Knowl. Eng. Data Sci.*, vol. 2, no. 2, p. 58, 2019, doi: 10.17977/um018v2i22019p58-71.
- [21] E. Prasetyo and B. Prasetyo, "Increased Classification Accuracy C4 . 5 Algorithm Using Bagging Techniques in Diagnosing Heart Disease," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 5, pp. 1035–1040, 2020, doi: 10.25126/jtiik.202072379.
- [22] N. Mohd Nawi, E. T. Tosida, H. Hasbi, and N. Abdul Hamid, "An Implementation of First and Second Order Neural Network Classification on Potential Drug Addict Repetition," *Emerg. Adv. Integr. Technol.*, vol. 02, no. 01, pp. 18–29, 2021
- [23] S. Setti and A. Wanto, "Analysis of Backpropagation Algorithm in Predicting the Most Number of Internet Users in the World," *J. Online Inform.*, vol. 3, no. 2, p. 110, 2019, doi: 10.15575/join.v3i2.205.
- [24] B. Zhang, X. L. Liang, H. Y. Gao, L. S. Ye, and Y. G. Wang, "Models of logistic regression analysis, support vector machine, and back-propagation neural network based on serum tumor markers in colorectal cancer diagnosis," *Genet. Mol. Res.*, vol. 15, no. 2, 2016, doi: 10.4238/gmr.15028643.
- [25] T. Mhatre and S. Varma, "IJERT-Heart Disease Prediction using Evolutionary based Artificial Neural Network Heart Disease Prediction using Evolutionary based Artificial Neural Network," *Int. J. Eng. Res. Technol.*, vol. 8, no. 08, 2019.
- [26] Y. M. S. Al-barzinji, M. A. Ahmad, and B. K. Saeed, "International Transaction Journal of Engineering , Management , & Applied Sciences & Technologies GENETIC DISTANCES AMONG EIGHT ORNAMENTAL," vol. 11, no. 2, pp. 1–10, 2020, doi: 10.14456/ITJEMAST.2020.294.
- [27] S. Furmanek *et al.*, "University of Louisville Journal of Respiratory Infections The City of Louisville Encapsulates the United States Demographics," pp. 1–6, 2020, doi: 10.18297/jri/vol4/iss2/4.Abstract.
- [28] H. Shan, H. Xu, S. Zhu, and B. He, "A novel channel selection method for optimal classification in different motor imagery BCI paradigms," *Biomed. Eng. Online*, vol. 14, no. 1, p. 1, 2015, doi: 10.1186/s12938-015-0087-4.
- [29] K. Celikmih, O. Inan, and H. Uguz, "Failure Prediction of Aircraft Equipment Using Machine Learning with a Hybrid Data Preparation Method," *Sci. Program.*, vol. 2020, 2020, doi: 10.1155/2020/8616039.
- [30] L. Zajmi, F. Y. H. Ahmed, and A. A. Jaharadak, "Concepts, Methods, and Performances of Particle Swarm Optimization, Backpropagation, and Neural Networks," *Appl. Comput. Intell. Soft Comput.*, vol. 2018, 2018, doi: 10.1155/2018/9547212.
- [31] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.
- [32] A. Seifi and F. Soroush, "Pan evaporation estimation and derivation of explicit optimized equations by novel hybrid meta-heuristic ANN based methods in different climates of Iran," *Comput. Electron. Agric.*, vol. 173, no. February, 2020, doi: 10.1016/j.compag.2020.105418.