**Accredited SINTA 2 Ranking**
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

**JURNAL RESTI**
**(Rekayasa Sistem dan Teknologi Informasi)**
Vol. 9 No. 2 (2025) 385 - 392 e-ISSN: 2580-0760

# Enhanced Heart Disease Diagnosis Using Machine Learning Algorithms: A Comparison of Feature Selection

Hirmayanti[1*], Ema Utami[2]
[1,2]Magister of Informatics Engineering, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[1]hirmayanti@students.amikom.ac.id, [2]ema.u@amikom.ac.id

***Abstract***

*Heart disease or cardiovascular disease is one of the leading causes of death in the world. Based on WHO data, in 2019, as many as 17.9 million people died from cardiovascular disease. If early prevention is not carried out immediately, of course, the victims will increase every year. Therefore, with the increasingly rapid development of technology, especially in the health sector, it is hoped that it can help medical personnel in treating patients suffering from various diseases, especially heart disease. So in this study, it will be more focused on the selection of relevant features or attributes to increase the accuracy value of the Machine Learning algorithm. The algorithms used include Random Forest and SVM. Meanwhile, for feature selection, several feature selection techniques are used, including information gain (IG), Chi-square (Chi2) and correlation feature selection (CFS). The use of these three techniques aims to obtain the main features so that they can minimize irrelevant features that can slow down the machine process. Based on the results of the experiment with a comparison of 70:30, it shows that CFS-SVM is superior by using nine features, which obtain the highest accuracy of 92.19%, while CFS-RF obtains the best value with eight features of 91.88%. By using feature selection and hyperparameter techniques, SVM obtained an increase of 10.88%, and RF obtained an increase of 9.47%. Based on the performance of the model using the selected relevant features, it shows that the proposed CFS-SVM shows good and efficient performance in diagnosing heart disease.*

*Keywords:* *heart disease; feature selection; random forest; hyperparameter*

*How to Cite:* Hirma and Ema Utami, "Enhanced Heart Disease Diagnosis Using Machine Learning Algorithms: A Comparison of Feature Selection", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 2, pp. 385 - 392, Apr. 2025.
*DOI*: https://doi.org/10.29207/resti.v9i2.6175

## 1. Introduction

The World Health Organization (WHO) estimates that 17.9 million deaths occur globally each year [1]. Although most people with heart or cardiovascular illness do not experience any symptoms, the condition can be very dangerous [2] Blood pressure, cholesterol, hyperglycemia, heart rate irregularities, lifestyle, and smoking habits are all factors associated with cardiovascular disease, which can be managed with medication and other preventative strategies [3]. However, medication cannot be used to treat characteristics, including age, race, and family history of cardiovascular disease [4].

There are several classification algorithms that are still popularly used, such as Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB) [5]. Ahamad et al. predict heart disease to determine the level of accuracy of two datasets. The UCI Kaggle Cleveland dataset includes a total of 303 records with 14 attributes, and the comprehensive UCI Kaggle dataset (Cleveland, Hungary, Switzerland, and Long Beach V) includes a total of 1024 records with 14 features. Based on the research results, the first dataset showed that SVM obtained an accuracy of 87.91%, and Random Forest had an accuracy of 84.62%. While in the second dataset, the XGB showed the best accuracy at 99,03%. However, feature selection is still needed to determine the features that affect model accuracy.

Spencer et al. [6] conducted a classification of heart disease using feature selection techniques as one of the optimizations to improve the validation of their research results. Among these models, the results of the study showed the best accuracy of 85% obtained from the Bayes Net-Chi-squared algorithm. The dataset used was heart disease, obtained from Kaggle, with a total of 720 records and 14 features. Although they have used

Received: 13-11-2025 | Accepted: 12-04-2025 | Published Online: 19-04-2025

feature selection techniques, they have not been able to explore the interactions between features that can affect heart disease predictions.

Das et al. [7] in this study compared the XGBoost, Bagging, RF, DT, KNN and NB algorithms to predict heart disease based on clinical data from patients in the United States. The experimental results obtained the best accuracy from XGBoost, which was 91.30%. Although the research of Das et al. was able to reduce irrelevant features, which initially numbered 300 features, to 18 relevant features. However, the study did not explore further the interactions between features that could be a potential for improving the model.

According to Bhatt et al. [8] heart disease prediction using various algorithms was also conducted in this study using a real-world dataset taken from Kaggle totaling 70,000 cases. Based on the results of experiments using GridSearchCV hyperparameters, GridSearchCV can increase the accuracy of Random Forests by up to 0.5%. Although there is an increase in accuracy using hyperparameters in this study, it only considers features based on health lab results and has not explored other triggering factors such as genetics and lifestyle habits.

The Random Forest and SVM algorithms, on average, provide the best performance in classifying heart disease. However, it is not yet possible to find out the features that are relevant to the accuracy value of each classification model based on feature selection, and descriptive analysis is still needed to solve this problem. One way to select features in a classification model is feature selection. Feature selection is a technique used to reduce the dimensionality of patterns for classification by selecting the most informative features, not irrelevant and/or redundant features [9].

In the study of Noroozi et al. [10] They predicted heart disease using feature selection in the machine learning algorithm [10]. The dataset used was UCI Cleveland, with 303 cases with 14 attributes and the best accuracy at 85.5% from the Random Forest-CFS (Correlation Feature Selection) algorithm. Using the same dataset [11]Khurana et al. in their study predicted heart disease with the best performance of 83.41% using the SVM algorithm with Chi-square and Information Gain as feature selection. Research [5], [12] and [13] also concluded that chi-square works better than other feature selection techniques. While [14] obtains a superior CFS.

Feature selection with traditional extraction often results in features that affect the classification model being lost. However, not only that, the removal of attributes can sometimes cause errors in decision making, due to the many considerations involved [15]. In addition, manual feature selection requires people who are experts or experts in their fields. Because if you choose the wrong feature, it can result in the loss of relevant features [5]. Not only that, manual feature selection also takes a long time and requires high concentration to maintain consistency in feature selection [1].

Therefore, to overcome this problem, this study will use a feature selection technique because by focusing on the main characteristics, the feature selection approach often produces better model performance compared to using raw data directly [16]. By using feature selection also to prevent overfitting, irrelevant features will be removed because they can slow down the machine's performance. To get maximum results, researchers will compare several feature selection techniques such as Chi-square, Information Gain and CFS (Correlation Feature Selection). To get accurate results, the dataset and classification model used are popular and superior models based on previous research.

In order not to widen the research conducted, there are several limitations, such as the dataset used was collected from UCI Cleveland, the model used was Random Forest (RF) and Support Vector Machine (SVM) with the feature selection compared, namely CFS, chi-square and information gain. The model development was implemented with Google Colab using the Python language, with evaluation models using a confusion matrix. Finally, optimisation was carried out on the model using GridSearchCV hyperparameters to obtain accurate results, so that researchers can conduct descriptive analysis to determine the best feature selection technique in selecting relevant features in diagnosing heart disease.

## 2. Research Methods

This study compares the performance of feature selection in determining relevant features for machine learning models. To obtain balanced comparison results, popular and well-performing algorithms such as Random Forest and SVM are used. The flow of this study can be seen in Figure 1. Based on Figure 1, it can be seen that the first thing to do is select a dataset. The dataset used in this study is the same as in previous studies, namely the UCI Cleveland heart disease dataset, consisting of 303 cases with 14 attributes. The dataset can be accessed at the UCI Repository Machine Learning or by accessing this address: https://archive.ics.uci.edu/dataset/45/heart+disease .

The next step is preprocessing, where the preprocessing results show that there are no missing values. Therefore, to prepare the dataset, researchers only perform preprocessing, including outliers, correlation, encoding, and data transformation. Here, encoding functions are used to change the value of the dataset from category to numeric, because there are two attributes whose values are found in the object, namely, ca and thal. While transformation data is used to change the target multi-class into a binary class, this is done to improve machine performance. For more details, please see the research flow diagram in Figure 1.

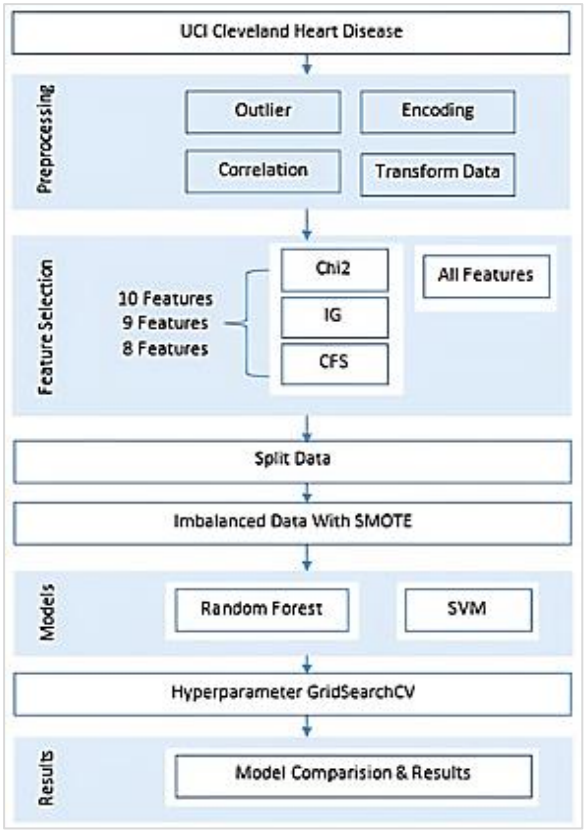


Figure 1. Research Flow Diagram

There are three feature selection techniques used, including Chi-square (Chi2), Information Gain (IG) and Correlation Feature Selection (CFS). Of all the features, researchers conducted three trials of the three feature selection techniques, namely starting from 10 features, 9 features and 8 features, the results of which will later be compared with all features. The data split is done with a ratio of 70:30, 70% training data and 30% testing data. To handle unbalanced data, the SMOTE technique is used to obtain balanced data.

After the dataset is ready to use, it is then applied to the Random Forest and SVM classification algorithms. Here, testing will be carried out to find relevant features based on the three feature selection techniques. The first thing to do is to apply all the features, and the second test will use each feature selected based on the three feature selection techniques, be it 10 features, 9 features and 8 features. And finally, to optimize the models, researchers use the GridSearchCV hyperparameter with k-fold 10.

Random Forest is a machine learning method introduced in 2001 by Leo Breiman, this method uses a large series of Decision Tree structures with low mutual correlation and randomly selected features using the bagging method (Bootstrap AGGregatING) [17]. The selection of this model is based on the fact that Random Forest is widely considered one of the most successful and widely used Machine Learning methods to date [18]. Random Forest has been widely used and shows the best results [19]-[21]. Random Forest is widely used because it is simple and flexible [22]Random Forest is also capable of processing high-dimensional data, even though it uses a lot of data, its performance remains high when compared to other Machine Learning models [17].

SVM, first proposed in the mid-1990s [23]. While SVM is used to analyze data and recognize patterns, SVM was especially developed to perform classification, regression and novelty detection [24]. In simple terms, the concept of SVM is an effort to find the best hyperplane that functions as a separator of two classes in the input space [25]. In previous studies [26], [10], [11] SVM showed better performance than other machine learning models. Feature selection is widely used to eliminate irrelevant and redundant features. The feature selection method has its characteristics in selecting features, such as data type, data size and the presence of noise in the data [27]. The function of feature selection in this study is to improve model performance, reduce irrelevant features, so that it can reduce computational time and costs.

The use of the Random Forest algorithm in this study aims to minimize the confusion of Random Forest in making decisions [28]. Therefore, this study uses a feature selection technique that aims to overcome these problems. Meanwhile, the SVM algorithm is an algorithm that has been widely compared and is superior to other popular algorithms, such as Logit Boost, MLM (Multivariate Linear Model) [10], ANN [26], and XGB [5].

CFS (Correlation-based Feature Selection) is a multivariate filter feature selection that works by selecting features based on correlation by measuring the relationship between two variables [29]. Irrelevant features will be ignored because they do not have a high correlation value [10]. The CFS used in this study is based on the Pearson correlation coefficient with Equation 1.

$$r = \frac{\sum(X-\dot{X})(Y-\dot{Y})}{\sqrt{\sum(X-\dot{X})^2 \sum(Y-\dot{Y})^2}} \quad (1)$$

r is Pearson correlation, X is the first variable value, Y is the second variable value, $\dot{X}$ is the average of variable X, and $\dot{Y}$ is the average of variable Y.

Information Gain is a univariate filter feature selection that works by selecting features based on the reciprocal dependency between the feature and the target variable [11]. Information gain is also defined as the reduction of information uncertainty (entropy) based on certain features in the dataset. This feature selection is calculated using Equation 2.

$$IG(D,A) = H(D) - H(D|A) \quad (2)$$

*IG(D,A)* is information gain on dataset D with feature A, *H(D)* is entropy of dataset D, and *H(D|A)* is conditional entropy of dataset D given feature A.

Chi-square is a filter feature selection that works by selecting features that are related to each other based on

the target class [6]. The chi-square statistic is calculated by finding the difference between each observed frequency and the expected frequency for each possible outcome, squaring it, dividing it by the expected frequency, and summing the results [30]. So, the higher the chi-square value, the higher the relationship between the attribute and the target class. The formula for calculating chi-square is in Equation 3.

$$chi = \sum \frac{(Oi-Ei)^2}{Ei} \quad (3)$$

chi2 is univariate statistics, Oi is the observed frequency is the actual frequency obtained from the dataset, and Ei is the expected frequency is the frequency that will occur if both variables are independent.

SMOTE (Synthetic Minority Over-sampling Technique) is a method used to handle data imbalance. This method works by creating artificial data (synthesis) from k-nearest neighbors with the aim of improving class representation [31]. In the study [32], the SMOTE method performed better when compared to ADASYN. Meanwhile, optimising machine learning performance is often known as hyperparameters. Hyperparameters are one method to optimize the performance of each model by identifying the best parameters in the machine learning process [5]. This study uses the GridSearchCV approach as a hyperparameter optimization to improve model accuracy.

## 3. Results and Discussions

### *3.1 Results*

This study presents evaluation results based on (1) diagnosis using all features (without feature selection) in the classification algorithm, (2) identifying relevant features with accuracy values using feature selection techniques and (3) model performance using feature selection and GridSearchCV hyperparameters. This experiment was implemented using Google Colab with the Python language. The dataset used was 303 cases with 14 attributes, as in Tables 1 and 2.

Table 1. Attribute for The Heart Disease Dataset

| Attribute | Description | Type Data |
|---|---|---|
| Age | Age | Float |
| Sex | Gender | Float |
| Cp | Chest pain type | Float |
| Trestbps | Resting blood pressure | Float |
| Chol | Cholesterol | Float |
| Fbs | Fasting blood sugar > 120 mg/dl | Float |
| Restecg | Rest ECG test | Float |
| Thalach | Maximum heart rate achieved | Float |
| Exang | Exercise-induced angina | Float |
| Oldpeak | ST depression induced by exercise relative to rest | Float |
| Slope | Slope (ST depression) | Float |
| Ca | Number of major vessels (0-3) colored by fluoroscopy | Object |
| Thal | Thalassemia (hemolytic disease) | Object |
| Num | Diagnosis of heart disease | Int |

Table 2. Attributes Description

| Attributes | Description |
|---|---|
| Age | Patient age |
| Sex | 1: male, 0: female |
| Cp | 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic |
| Trestbps | Resting blood pressure upon hospital admission, measured in mm/Hg. |
| Chol | Blood cholesterol level measured in mg/dL. |
| Fbs | fasting blood sugar > 120 mg/dl, (1: true; 0: false) |
| Restecg | 0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy by Estes' criteria |
| Thalach | Max heart rate during exercise. |
| Exang | 1: yes, 0: no |
| Oldpeak | ST depression induced by exercise relative to rest |
| Slope | 1: upsloping, 2: flat, 3: downsloping |
| Ca | 0-4 |
| Thal | 3: normal, 6: fixed defect; 7: reversable defect |
| Num | 0: normal, 1: mild, 2: moderate, 3: several, 4: very severe |

After preprocessing, the next step is to transform the data. Based on Table 2, num (target class) has 5 heart disease diagnosis labels, including 0: normal, 1: mild, 2: moderate, 3: severe, and 4: very severe. Therefore, because the dataset contains multi-classes, the transformation function is performed to change the multi-class into binary in the dataset, so that the machine can easily understand the data used. For more details, see Table 3 for the differences in the dataset before and after transformation.

Table 3. Data Transformation

| | Num (target class) | | | | |
|---|---|---|---|---|---|
| Multiclass | 0 | 1 | 2 | 3 | 4 |
| Total | 164 | 55 | 36 | 35 | 13 |
| Binary | 0 | | | 1 | |
| Total | 219 | | | 84 | |

SMOTE is used to handle imbalance in data in order to obtain balanced data for processing. Based on how SMOTE works, which has been explained previously, the synthetic data that is formed will be new data in the dataset. This synthetic data is formed on target class data 1 (true has heart disease), for more details, see Figure 2.

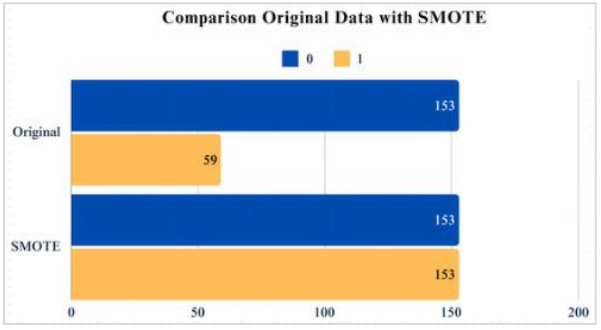


Figure 2. Comparison Original Data with SMOTE

By comparing 70% training data and 30% testing, 212 training data were obtained with 153 class 0 and 59 class 1. To handle this imbalance, SMOTE was used to obtain 153 class 0 and class 1. Next, feature selection is

carried out based on information gain, chi-square and correlation-based feature selection (CFS).

*3.1.1 Models with All Attributes*

In this study, the dataset testing used Random Forest and SVM models with a ratio of 70% training set and 30% testing set, so that the accuracy results were obtained as in Figure 3.

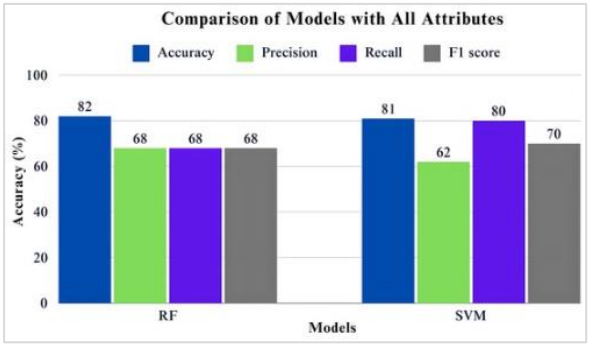


Figure 3. Comparison of Models with All Attributes

Based on Figure 3, the evaluation results of each model are illustrated in Figure 3. In this study, researchers compared three feature selection techniques, including Chi2, IG, and CFS, applied to the Random Forest and SVM algorithms. By using 14 features (without feature selection) with "diagnosis" as the target class, Random Forest obtained the best performance with accuracy 82%, precision 68%, recall 68%, and F1 score 68%. While SVM obtained results that were not much different, with accuracy 81%, precision 62%, recall 80%, and F1 score 70%.

*3.1.2 Models with Feature Selection*

In testing the model using feature selection, researchers conducted experiments by trying several of the highest attributes (such as 10 attributes, 9 attributes and 8 attributes) from each feature selection technique used. Each feature is measured based on the way each feature selection technique works, so that the results of the features obtained are also diverse.

By using feature selection techniques to select features rather than manually selecting them, it can minimize errors in selecting relevant features. It can be seen in Figure xxx, from the three feature selection techniques that were compared, there were several similarities in features such as "oldpeak", "ca", "thal", "cp", "thalach", "slope", and "exang". After that, there were also different features such as "chol", if in the IG technique "chol" was ranked 1 and in the Chi2 technique "chol" was ranked 10, but in the CFS technique "chol" was ranked 13 (last). The results of the feature selection calculation are sorted from the highest to the lowest value for each feature selection technique, as in Table 4.

Each feature selection is tested three times with the number of attributes 10, 9 and 8. This is to determine the effect of the number of attributes on the accuracy of the models. By using the features selected based on the three feature choices, Table 5 shows the test results based on chi-square (Chi2), information gain (IG) and correlation feature selection (CFS).

Table 4. Feature Selection Ranking Based on Feature Selection Techniques

| Ranking | Feature Selection Techniques | | |
|---|---|---|---|
| | Chi2 | IG | CFS |
| 1 | Oldpeak | Chol | Oldpeak |
| 2 | Thalach | Thalach | Thal |
| 3 | Ca | Oldpeak | Ca |
| 4 | Exang | Ca | Thalach |
| 5 | Cp | Thal | Cp |
| 6 | Slope | Cp | Exang |
| 7 | Thal | Trestbps | Slope |
| 8 | Sex | Slope | Age |
| 9 | Trestbps | Age | Sex |
| 10 | Chol | Exang | Trestbps |
| 11 | Restecg | Sex | Restecgs |
| 12 | Fbs | Restecg | Fbs |
| 13 | Age | Fbs | Chol |

After applying the three feature selection techniques to the model, the evaluation results were obtained as shown in Table 5.

Table 5. Models' Performance Based on Feature Selection Techniques

| Models | Selected Feature | Number of Feature (%) | | |
|---|---|---|---|---|
| | | 10 | 9 | 8 |
| Random Forest | IG | 82,41 | 81,31 | 80,21 |
| | Chi2 | 82,41 | 83,51 | **85,71** |
| | CFS | 82,41 | 83,51 | 82,41 |
| SVM | IG | 84,61 | 81,31 | 81,31 |
| | Chi2 | 83,51 | 84,61 | 84,61 |
| | CFS | 82,41 | 85,71 | 84,61 |

Based on Table 5, it can be concluded that the use of relevant features can produce higher accuracy than using all attributes (13 attributes, except num). The highest accuracy using all attributes was obtained from the Random Forest model of 82%, while using only eight features could achieve an accuracy of 85.71% in Random Forest. This shows that the use of feature selection can make the machine work efficiently and effectively. When compared with the results of accuracy with all features in random forest, a difference of 3.71% was found.

*3.1.3 Models with Feature Selection and Hyperparameter*

To improve the performance of the model in diagnosing heart disease, researchers also use GridSearchCV as a hyperparameter. As for the code on the Google Colab used in Figure 4.

```
from sklearn.model_selection import GridSearchCV

grid_models = [(RandomForestClassifier(),
                [{'n_estimators':[100,150,200],
                'criterion':['gini','entropy'],
                'random_state':[0]}]),
                (SVC(),[{'C':[0.25,0.5,0.75,1],
                'kernel':['linear', 'rbf'],
                'random_state':[0]}])]
```

Figure 4. Code for Hyperparameter GridSearchCV Models

The best parameters obtained from RF include: {'criterion': 'gini', 'n_estimators': 100, 'random_state': 0}. The best parameters of SVM include: {'C': 0.5, 'kernel': 'rbf', 'random_state': 0}.

The results of models using feature selection and hyperparameters are as in Table 6.

Table 6. Performance of Feature Selection and Hyperparameters on Models

| Models | Selected Feature | Accuracy (%) 10 | 9 | 8 |
|---|---|---|---|---|
| Random Forest | IG | 90,89 | 90,25 | 89,91 |
| | Chi2 | 89,28 | 89,29 | 90,59 |
| | CFS | 90,26 | 90,91 | 91,88 |
| SVM | IG | 90,25 | 89,92 | 89,59 |
| | Chi2 | 91,23 | 91,87 | 91,56 |
| | CFS | 91,55 | 92,19 | 91,55 |

Table 6 shows that SVM using nine selected features based on CFS obtained the highest accuracy of 92.19%. When compared to Figure 3, the difference in improvement is 10.88% with k-10 fold cross-validation. Although SVM is superior here, Random Forest also provides the best performance with an accuracy of 91.88%. When compared to Figure 3, which shows Random Forest obtaining the best score using eight features, Random Forest has an improvement by a difference of 9.47%.

Based on the results of the experiment, considering the number of datasets and attributes in them, it can provide different results, because one attribute is definitely different from the other attributes. If you choose the wrong attribute, it can affect the accuracy of the model. However, choosing and using the right features can make the results more accurate. And by only focusing on certain features, it can speed up the performance of the model. This study still uses a dataset that includes little data, which affects the accuracy results. By using a large number of datasets and applying the feature selection technique along with the hyperparameters that we use, of course will get much better results.

*3.2 Discussion*

The accuracy of Machine Learning algorithms can be influenced by many factors, one of which is the use of features. After the researcher evaluated the results, it was found that not always using all the features would provide the best accuracy results. This study shows that using only a few relevant features can produce a higher level of accuracy than using all the features.

Figure 5 shows the comparison results when the feature selection technique is optimised on the Random Forest and SVM algorithms. The use of nine relevant features obtained from CFS-SVM obtained the highest results, and Random Forest showed the best results with eight features from CFS. Thus, the performance of CFS-SVM is superior to other feature selection techniques.

The increase in accuracy shows that the model works well, with the use of feature selection being able to help the machine process the dataset more efficiently in diagnosing heart disease. This study not only shows an increase in accuracy results, but also, by comparing feature selection techniques, can provide knowledge about which attributes or features influence the increase in accuracy values, and provide knowledge that, from the various feature selection techniques compared, CFS is known to be superior.

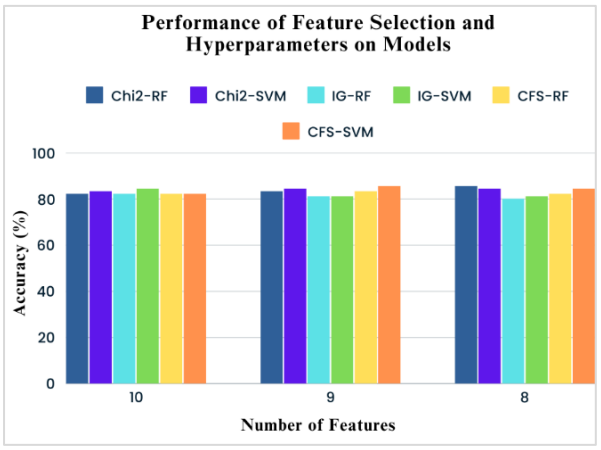


Figure 5. Performance of Feature Selection and Hyperparameters on Models

The Chi2 feature selection technique obtains the three lowest features with the same value, including restecg, fbs and age. However, in the CFS and IG feature selection techniques, all three features are taken into account. Therefore, it can be concluded that there is diversity in evaluating relevant features in diagnosing heart disease. The results of the model evaluation will later be influenced based on the selected features according to the diversity of the results of the feature selection calculation.

Based on Table 7, by comparing our research results with previous studies, we found that Reddy et al. [13] predict heart disease using the Cleveland UCI dataset with the SMO (Sequential Minimal Optimization) model, their study compared several feature selections to obtain a better Chi-square with an accuracy of 86.46%. Khurana et al. [11] predict heart disease using the NB, DT, LR, RF, SVM and KNN models. The dataset used was the Cleveland UCI, which also compared feature selection. The results showed the best performance with an accuracy of 83.41% from SVM-Chi2 and IG. While the study [10] analyzed the effect of feature selection on machine learning performance. The study Noroozi et al. compared models (Bayes Net, NB, MLP, SVM, Logis Boost, J48 and RF) and feature selection techniques including CFS, IG, Gain Ratio, Relief and SU (Symmetrical Uncertainly), and obtained the best accuracy value of 85.5% from SVM-CFS, IG and SU.

In the study [13] shows the best performance of SVM using eleven relevant features from Chi-Square measurement results with an accuracy of 86.46%. The features include thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg, and fbs. In this research by

 

Reddy et al., to handle the missing dataset on the attributes "ca" and "thal" are replaced with the majority of these values, namely 0 and 3. So that the dataset used remains intact (303 records, 14 attributes) without reducing the data at all.

Table 7. Compare Our Work Results with the Previous Results

| Feature Selection Methods | ML Algorithms | Dataset | Best Methods | Best Acc. | Year | Reference |
|---|---|---|---|---|---|---|
| CFS, Chi-square and ReliefF | NB, LR, SMO, IBk/KNN, ABM1+DS, ADM1+LR, Bagging+REPTree, Bagging+LR, JRip and RF | Cleveland | SMO with Chi-square | 86,46 (%) | 2021 | [13] |
| Chi-square, Gain Ratio, Information Gain, One-R and Relief | NB, DT, LR, RF, SVM, KNN | Cleveland | SVM with Chi-square and Information Gain | 83,41 (%) | 2021 | [11] |
| Filter methods (CSF, Information Gain, Gain Ratio, Relief, Symmetrical uncertainty), Wrapper (Forward and backward selection, Naïve Bayes, Decision tree, KNN, NN, SVM, Logistic regression), and evolutionary (PSO, ABC, and genetic algorithms) | Bayes Net, Naïve Bayes (BN), multivariate linear model (MLM), (SVM), logit boost, j48 and RF | Cleveland | SVM with CFS, Information Gain and Symmetrical Uncertainty | 85,5 (%) | 2023 | [10] |
| Chi-square, Information Gain and CFS | Random Forest and SVM | Cleveland | SVM with CFS | 92,19 (%) | 2024 | **Our study** |

Meanwhile, in the study, Khurana et al. used nine features to obtain the best accuracy results of 83.41% [11]. These results were obtained from the SVM algorithm with Chi-square and Information Gain, which were superior to other feature selections. The nine features include thal, cp, ca, oldpeak, exang, thalach, slope, age and sex.

In the study [10] obtained the best accuracy result of 85.5% using SVM-Filter methods (CFS-IG-SU). Noroozi et al.'s study obtained ten features based on filter methods, including age, sex, cp, restecg, thalach, exang, oldpeak, ca, thal, slope. In the Cleveland dataset used, 6 missing values were detected, which were then dropped, so that a clean dataset of 297 was obtained.

In our study, when compared to previous studies, it can be said to be superior based on the accuracy results and based on the number of features used, namely with nine features including oldpeak, thal, ca, thalach, cp, exang, slope, age and sex. In our study, no missing values were found in the dataset, but there were category values which we then replaced by converting them to numeric form using the encoding function. We hope that there will be more future studies discussing feature selection techniques, because feature selection is very helpful in determining relevant features and affecting the accuracy results of Machine Learning.

## 4. Conclusions

The conclusion of this study, after testing and obtaining evaluation results, it can be concluded that the use of feature selection techniques and hyperparameters can increase the accuracy value of the Machine Learning model used, especially in the SVM algorithm. The feature selection techniques compared include Chi-square (Chi2), Information Gain (IG) and Correlation Feature Selection (CFS) which are implemented in the Random Forest and SVM classification algorithms. Based on the evaluation results, CFS is superior to the SVM algorithm with an accuracy of 92.19%, while Random Forest obtains the best results from CFS as much as 91.88%. In this study, the CFS feature selection technique and hyperparameters were able to increase the accuracy of SVM by 10.88% and increase Random Forest by 9.47%. The nine relevant features selected based on CFS-SVM include oldpeak, thal, ca, thalach, cp, exang, slope, age, and sex. We can conclude that using the nine features from CFS can improve the performance of SVM and make CFS superior to other feature selection techniques. For further research, we suggest exploring more datasets with larger numbers so that we can find more diverse features in diagnosing heart disease. We hope that the use of feature selection will continue to be developed by trying new algorithms to further explore which features can affect Machine Learning performance.

## References

[1] A. Al Ahdal *et al.*, "Monitoring Cardiovascular Problems in Heart Patients Using Machine Learning," *J. Healthc. Eng.*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9738123.

[2] S. H. Rampengan, *Buku praktis kardiolaogi*. Jakarta: Badan Penerbit Fakultas Kedokteran Universitas Indonesia, 2014.

[3] R. Vijaya Saraswathi, K. Gajavelly, A. Kousar Nikath, R. Vasavi, and R. Reddy Anumasula, "Heart Disease Prediction Using Decision Tree and SVM," *Springer*, no. March, pp. 69–78, 2022, doi: 10.1007/978-981-16-7389-4_7.

[4] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," *MDPI*, vol. 12, no. 1, pp. 1–31, 2023, doi: 10.3390/biology12010117.

[5] G. N. Ahamad *et al.*, "Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease," *MDPI*, vol. 11, no. 3, 2023, doi: 10.3390/pr11030734.

 

[6] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Heal.*, vol. 6, pp. 1–10, 2020, doi: 10.1177/2055207620914777.

[7] R. C. Das, M. C. Das, M. A. Hossain, M. A. Rahman, M. H. Hossen, and R. Hasan, "Heart Disease Detection Using ML," *IEEE*, pp. 983–987, 2023, doi: 10.1109/CCWC57344.2023.10099294.

[8] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart-Disease Prediction by Using Hybrid Machine Learning Technique," *MDPI*, pp. 1670–1675, 2023, doi: 10.1109/ICCPCT58313.2023.10245785.

[9] E. Chitsaz, M. Taheri, S. D. Katebi, and M. Z. Jahromi, "An improved fuzzy feature clustering and selection based on chi-squared-test," *Proc. Int. multiconference Eng. Comput. Sci.*, vol. 1, no. June 2015, pp. 18–20, 2009.

[10] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-49962-w.

[11] P. Khurana, S. Sharma, and A. Goyal, "Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques," *Proc. 8th Int. Conf. Signal Process. Integr. Networks, SPIN 2021*, no. August, pp. 510–515, 2021, doi: 10.1109/SPIN52536.2021.9565963.

[12] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Elsevier*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

[13] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," *MDPI*, vol. 11, no. 18, 2021, doi: 10.3390/app11188352.

[14] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Prediction of Heart Disease Risk Using Machine Learning with Correlation-based Feature Selection and Optimization Techniques," *2021 7th Int. Conf. Signal Process. Commun. ICSC 2021*, no. December 2022, pp. 228–233, 2021, doi: 10.1109/ICSC53193.2021.9673490.

[15] B. C. L. Adiatma, E. Utami, and A. D. Hartanto, "PENGENALAN EKSPRESI WAJAH MENGGUNAKAN DEEP CONVOLUTIONAL NEURAL NETWORK," *EXPLORE*, vol. 11, no. 2, p. 75, Jul. 2021, doi: 10.35200/explore.v11i2.478.

[16] Padathala Visweswara Rao, "Extraction and Feature Selection for Precise Cardiovascular Disease Classification," *Int. J. Multidimens. Res. Perspect.*, vol. 2, no. 7, pp. 79–87, 2024, doi: 10.61877/ijmrp.v2i7.172.

[17] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *Elsevier*, vol. 34, no. 6, pp. 2733–2742, 2022, doi: 10.1016/j.jksuci.2022.03.012.

[18] M. Brendel, C. Su, Z. Bai, H. Zhang, O. Elemento, and F. Wang, "Application of Deep Learning on Single-cell RNA Sequencing Data Analysis: A Review," *Elsevier*, vol. 20, no. 5, pp. 814–835, 2022, doi: 10.1016/j.gpb.2022.11.011.

[19] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *MDPI*, vol. 11, no. 4, 2023, doi: 10.3390/pr11041210.

[20] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," *Health Soc. Care Community*, vol. 2023, no. Cvd, pp. 1–10, 2023, doi: 10.1155/2023/1406060.

[21] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection," *IEEE Access*, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.

[22] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[23] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999, doi: 10.1109/72.788640.

[24] M. Awad and R. Khanaa, *Efficient Learning Machine: Theories, Concepts and Application for Engineers and System Designers*, no. 112. Apress, 2015.

[25] P. Eko, *Data Mining: Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: ANDI Yogyakarta, 2012.

[26] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, no. Ml, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[27] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, 1997, doi: 10.3233/IDA-1997-1302.

[28] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning," *IEEE Access*, vol. 11, no. June, pp. 56214–56224, 2023, doi: 10.1109/ACCESS.2023.3281484.

[29] A. G. Karegowda, A. S. Manjunath, G. Ratio, and C. F. Evaluation, "Comparative study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[30] R. L. Plackett, "Karl Pearson and the Chi-squared Test," *Int. Stat. Inst.*, vol. 64, no. 1, pp. 50–53, 1984, doi: 10.47316/cajmhe.2024.5.1.05.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *J. Artif. Intell. Res.*, vol. 4, no. 16, pp. 321–357, 2002, doi: 10.46880/jmika.vol4no1.pp67-72.

[32] S. P. R. Yulianto, A. Z. Fanani, A. Affandy, and M. I. Aziz, "Analisis Metode Smoote pada Klasifikasi Penyakit Jantung Berbasis Random Forest Tree," *J. Media Inform. Budidarma*, vol. 8, no. 3, p. 1460, 2024, doi: 10.30865/mib.v8i3.7712.

 