Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 10, No. 2, June 2024, pp. 265-278
ISSN: 2338-3070, DOI: 10.26555/jiteki.v10i2.28647

Comparative Evaluation of Feature Selection Methods for Heart Disease Classification with Support Vector Machine

Winarsi J. Bidul[1], Sugiyarto Surono[1], Tri Basuki Kurniawan[2]
[1]Mathematics Study Program, UniversitasAhmad Dahlan, Yogyakarta, Indonesia.
[2]Faculties of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

ARTICLE INFO

Article history:

Received April 01, 2024
Revised May 27, 2024
Published June 29, 2024

Keywords:

Big Data;
Feature Selection;
Classification;
Heart disease

ABSTRACT

The purpose of this study is to compare the effectiveness of a variety of feature selection techniques to enhance the performance of Support Vector Machine (SVM) models for classifying heart disease data, particularly in the context of big data. The main challenge lies in managing large datasets, which necessitates the application of feature selection techniques to streamline the analysis process. Therefore, several feature selection methods, including Logistic Regression-Recursive Feature Elimination (LR-RFE), Logistic RegressionSequential Forward Selection (LR-SFS), Correlation-based Feature Selection (CFS), and Variance Threshold were explored to identify the most efficient approach. Based on existing research, these methods have shown a great impact in improving classification accuracy. In this study, it was found that combining the SVM model with LR-RFE, LR-SFS, and Variance Threshold resulted in superior evaluation, achieving the highest accuracy of 89%. Based on the comparison of other evaluation results, including precision, recall, and F1-score, the performance of these models varied depending on the feature selection method chosen and the distribution of data used for training and testing. But in general, LR-RFE-SVM and Variance Threshold-SVM tend to provide better evaluation values than LR-SFS-SVM and SVM-CFS. Based on the computation time, SVM classification with the Variance Threshold method as the feature selection method obtained the fastest time of 118.1540 seconds with the number and retention of 23 important features. Therefore, it is very important to choose a suitable feature selection technique, taking into account the number of retained features and the computation time. This research underscores the significance of feature selection in addressing big data challenges, particularly in heart disease classification. In addition, this study also highlights practical implications for healthcare practitioners and researchers by recommending methods that can be integrated into real-world healthcare settings or existing clinical decision support systems.

Corresponding Author:

Sugiyarto Surono, Mathematics Study Program, Ahmad Dahlan University, Yogyakarta, Indonesia.
Email: sugiyarto@math.uad.ac.id

1. INTRODUCTION

In today's digital age, the world is witnessing a surge of data known as "Big Data"[1]. The term "Big Data" refers to an enormous amount of data that cannot be effectively managed using traditional software or

internet-based platforms. The massive volume of data makes management and analysis difficult [2]. For example, In light of healthcare, data is often produced at a high rate. Health data must be able to be stored and processed quickly to support rapid decision-making in inpatient treatment. Delays in data processing can hurt patient diagnosis and treatment [3]. To overcome these challenges, special techniques or ways are needed that make it possible to process and analyze big data efficiently [4], [5]. One of the techniques that can be used in this context is feature selection techniques [6], [7].
By selecting the most informative features, noise, complexity, and overfitting can be minimized, improving prediction accuracy,reducing computation time, and giving the data a greater grasp [8], [9].

It is significant to remember that, particularly for classification jobs, the feature selection approach is an essential pre-processing step in data mining [10], [11]. In the healthcare field, specifically in the context of heart disease, which is one of the leading causes of death globally, it is important to select the most suitable feature selection method for the classification of this disease [12].

Much of the current literature conducts research related to the comparison of feature selection methods and classification algorithms but has not fully explored and compared feature selection methods specifically for heart disease classification. As in research [13] which compares several feature selection techniques based on Support Vector Machine (SVM) classification accuracy results and computation time for Twitter comment data, it was found that the chi-square method was superior regarding computation time, which was 0.4375 seconds, compared to the Mutual Information Feature Selection method, whose computation time was 252.75 seconds. Meanwhile, based on classification accuracy, the Mutual Information Feature Selection method is superior at 80% compared to the chi-square method, whose accuracy is 78%. In addition, research [14] evaluated various machine learning methods for human activity recognition using smartphone sensors for health research. The findings indicate that the SVM method provides superior performance, namely with an accuracy of 96.0% compared to the k-Nearest Neighbors (kNN) method with an accuracy of 90.33%, followed by the Decision Tree method, and Random Forest obtained an accuracy of 85.3% and 90.05%

In addition to the previously mentioned feature selection methods, many other feature selection methods have proven to have good performance in several studies and can be used in this study as feature selection methods that will then be compared for performance. Some of them, such as in research [15] used the Recursive Feature Elimination (RFE) method and several other methods as feature selection methods to analyze cardiovascular disease and obtained the result that the accuracy of the RFE method is higher at 88.84%.

Apart from that, in research [16], image analysis was carried out using the Variance Threshold feature selection method using a threshold of 0.05. In this study, the highest accuracy value was obtained at 92.31%, which shows that the Variance Threshold feature selection method can improve accuracy results.

In [17] the machine learning method was applied to 93 meteorological variables from the ECMWF and RDAPS short-term weather forecasts for South Korea and used the Correlation Based Feature Selection (CFS) method, which is also a feature selection method to select a discriminative subset of these variables, which proved to obtain accurate results. Rainfall predictions produced by the model were found to be 15% more accurate than ECMWF forecasts and 13% better than RDAPS forecasts.

Another feature selection method is the Sequential Forward Selection (SFS) method. In research conducted by [18], which uses the RF-SFS method as a feature selection method in classification using the Random Forest model to classify the test results of railway tunnel waterproof slab products, RF-SFS helps in identifying key factors that can affect the test results of railroad products effectively. It was found that the method successfully identified six key factors that affect the test results of railroad tunnel waterproof slab products. Using only six key features, the RF model achieved an accuracy value of 99.98%, showing very strong classification and prediction capabilities. The contribution of this research is to explore several feature selection methods that are considered to have performed quite well in some of the previously mentioned studies. This research aims to compare feature selection methods to see which method is most suitable for heart disease classification. In addition, the research will be devoted to filling the gap related to the lack of research that specifically compares feature selection methods for heart disease classification using SVM models. The hope is that this research can provide new insights and signicant contributions to the field of health and medical informatics, having great relevance in improving the understanding and application of feature selection techniques, especially in the context of health.

2. METHODS

This research is an experimental study with a quantitative approach based on a literature review and theoretical framework that aims to compare the performance of several feature selection techniques in the context of heart disease classification using SVM models. The ultimate goal is to provide new insights and

significant contributions to the field of health and medical informatics. The research questions include the selection of the most suitable feature selection method for heart disease classification, performance comparison between different feature selection methods, and whether or not there is a significant difference in classification performance between the feature selection methods used.

In its theoretical framework, this research will involve the use of several feature selection methods that have been proven to perform quite well in several previous studies, such as Logistic Regression-Recursive Feature Elimination (LR-RFE), Logistic Regression-Sequential Forward Selection (LR-SFS), Correlation Based Feature Selection (CFS), and Variance Threshold. This study was conducted with the system ow shown in Fig. 1.

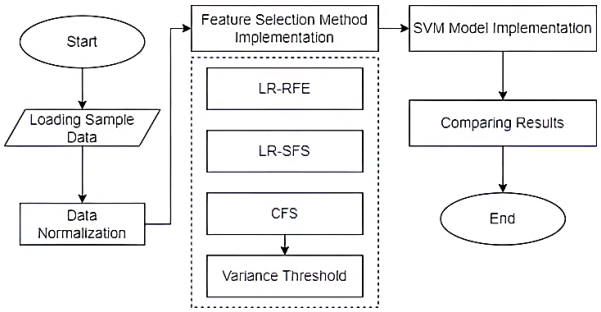

Fig. 1. Flow Chart

Based on the flowchart above, the research steps begin with loading sample data, followed by the data preprocessing stage. In this stage, the data is cleaned and normalized, and feature selection is performed to select the most relevant features from the dataset using four different feature selection methods. After that, classification is performed using the SVM model. Evaluation of the classification results of each feature selection method will be done by considering metrics such as accuracy, precision, recall, and F1-score. In addition, the computation time and number of retained features of each feature selection method will also be considered for result comparison. The final step in this research is to conduct a comparison of the results. By conducting an in-depth analysis of the results, it is expected that this research can nd the best feature selection method for heart disease classification, as well as make a significant contribution to the improvement of the understanding and application of feature selection techniques, especially in the context of healthcare.

2.1. Min-Max Normalization

This study's data was taken from the Kaggle platform, a popular data source for researchers. This data is secondary and can be accessed by users through the following link provided on Kaggle https://bit.ly/Heart_Disease_dataset. The total data used was 319,795 data samples consisting of 35 features and 1 target variable, which divided the samples into 2 classes, those suffering from heart disease and those not suffering from heart disease.

After acquiring the data and performing the processing, the next step is data normalization. In this research, the method used is Min-Max normalization. Min-Max Normalization is one of the data preprocessing techniques used in data analysis and machine learning. This method scales the data into a specific range of values, for example $[0, 1]$, without changing the data's possible bias [19]. The following formula is used to determine normalization:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(x)} \quad (1)$$

where X represents the research data, $\min(X)$ and $\max(X)$ are the minimum and maximum values in the data, respectively.

2.2. Feature Selection

Feature selection is a process in data analysis and machine learning where several features or attributes from a data set are identied and selected for use in building a more effective learning model [20], [21]. In this research, there are four feature selection techniques employed, namely:

1) Recursive Feature Elimination (RFE)

RFE is a feature selection wrapper method that makes use of internal filter-based features selection [22], [23]. The LR-RFE feature selection technique is predicated on a logistic regression algorithm and a recursive technique that iteratively removes the least significant characteristics from the dataset [24], [25]. The basic principle of LR-RFE involves the use of logistic regression to evaluate the contribution of each feature to the prediction of the target class and a recursion technique that enables an iterative process to select the best features [26]. The underlying assumption of this method is that irrelevant or unimportant features will have a low influence on the prediction of the target class, and by removing such features, model performance can be improved.

In its application, LR-RFE starts by initializing a logistic regression model with all features from the dataset and then features that are considered unimportant are iteratively removed from the dataset based on the evaluation of their contribution to the prediction of the target class. This process is repeated until the desired number of features is reached or until certain stopping criteria are met, such as a predetermined number of features or no significant improvement in model performance [27].

Logistic probability models are used to predict the probability of membership in a particular class based on given features. A logistic probability model with n independent variables and a binary dependent variable (x_i, y_i) for each i from 1 to n is as follows:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_j x_{i,k})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_j x_{i,k})} \quad (2)$$

where p_i is the probability of the i-th sample. $\beta_0, \beta_1, \ldots \beta_j$ are logistic regression coefficients, $x_{i,1}, x_{i,2}, \ldots, x_{i,k}$ are feature values of the i-th sample, and k is the number of features used in the model [28].

Then to form a logistic regression model, it needs to be transformed first to maintain the linear structure of the model [29]. The transformation performed is a logit transformation, with the following definition:

$$logit(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \quad (3)$$

If x and y are pairs of independent and dependent variables in the i-th observation and it is assumed that each pair of observations is independent of the other pair of observations, then the probability function for each pair is as follows:

$$f(x_i) = p_i^{y_i}(1 - p_i)^{i - y_i} \qquad ; y_i = 0, 1 \quad (4)$$

Parameter estimation in logistic regression uses the Maximum Likelihood Estimation (MLE) method. It involves optimizing the likelihood function, which fits the distribution of the observed data. In logistic regression, MLE provides an estimate of the coefficient β that indicates the impact of the feature on the target. The likelihood function formula for binary logistic regression is as follows:

$$l(\beta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{i - y_i} \quad (5)$$

where p_i is the probability for the i-th predictor variable and y_i is the observation on the i-th predictor variable. Optimizing the likelihood function is simpler when expressed in the form of $\log L(\beta)$, or written $L(\beta)$ [30].

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[p_i] + (1 - y_i) \ln[1 - p_i]\} \quad (6)$$

In this method, features are identified for removal by considering two aspects. First, the feature with the lowest coefficient is considered to have low significance for the target. In finding the value of the estimated coefficient (β) that maximizes $L(\beta)$, the derivative of $\beta_1, \beta_2, \cdots \beta_n$ is then equalized to zero, and the following formula is obtaine

$$\beta_0 = \sum_{i=1}^{k} [y_i - p_i] \quad (7)$$

and

$$\beta_j = \sum_{i=1}^{k} x_{ij}[y_i - p_i] \quad (8)$$

In addition, the odds ratio value is also considered in feature selection. The odds ratio is the ratio between the probability of success (p_i) and the probability of failure $(1 - p_i)$, which is a measure of how much influence a feature has on the target variable [31]. Features with an odds ratio far from 1 indicate a significant influence on the target variable. If the odds ratio is > 1, then the probability of the target variable occurring increases as the feature value increases. In the context of feature selection with LR-RFE, the odds ratio is used in conjunction with a logistic probability model to evaluate the importance of each feature to the prediction of the target class. The odds ratio can also be interpreted as a comparison of odds values between two subjects or groups. The formula for the odds ratio is:

$$odds\ ratio\ (\beta) = \left[\frac{\frac{p_i(1)}{(1-p_i(1))}}{\frac{p_i(0)}{(1-p_i(0))}}\right] = \exp(\beta) \quad (9)$$

2) Sequential Forward Selection (SFS)

SFS is a feature selection technique that combines the steps of selection and model building in a sequential [32]. LR-SFS is an approach to feature selection used in logistic regression analysis. The basic principle is to iteratively add one feature at each step, selecting the feature that best improves model performance based on specified criteria [33]. This method assumes that the relationship between the independent and dependent variables can be described linearly and uses a logistic function to model the probability of membership in the target class [34]. The process starts with the initialization of a logistic regression model, which is initially built with one feature or no features at all. Then, each feature that has not been incorporated into the model is evaluated at each iteration. The features that improve the model's performance the most are added to the model, and after that, the model is re-evaluated to see if the specified evaluation criteria are met. The process stops when certain stopping criteria are met, such as no significant improvement in model performance after the addition of certain features or when a predetermined number of features have been incorporated into the model [35], [36]. This research uses accuracy to evaluate model performance and decide which features should be included. The stopping criteria use the maximum number of features as a limit to avoid overfitting or complexifying the model more than necessary.

The LR-SFS method provides a systematic approach to selecting the features that are most informative in predicting the target variable in logistic regression, which helps reduce the dimensionality of the features, improves model interpretation, and allows for building simpler and more efficient models [37].

3) Variance Threshold

Variance Threshold feature selection is a technique used to select features in a dataset by evaluating their variance values [38]. The basic principle behind this approach is that features with very low variance, which tend to have almost constant values across data samples, provide little additional information for distinguishing the target class. Therefore, these features are considered uninformative or redundant in the context of analysis. This technique is suitable for use on data that has features measured on a similar scale [39]. The process starts by calculating the variance of each feature in the dataset. To calculate the variance of each feature, the following formula is used:

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \mu_j\right)^2 \quad (10)$$

where σ_j^2 is the variance of feature j, x_{ij} is the value of feature j in the j-th sample, μ_j is the average of feature j and n is the number of elements in the data set.

Next, a threshold value is set in this study. 0.05 is taken as the threshold, and features with variation below the threshold are removed from the dataset. This step helps filter out features that have too little variation to make a significant contribution to the model. Once the low-variation features are removed, the purified dataset can be used for further analysis, such as classification model building [16].

4) Correlation Based Feature Selection (CFS)

CFS is a technique based on the hypothesis that a good feature subset contains features that are highly correlated with the class label and mostly uncorrelated with each other [17]. This is because features that have a high correlation with the class label have the potential to provide significant information in prediction, while

features that have a high correlation with each other are considered redundant and tend to provide similar information.

The application process starts by calculating the correlation between each feature and the class label, as well as the correlation between features. The Pearson correlation formula is used to calculate the correlation in the CFS method, with the following equation:

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\ \sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}) \quad (11)$$

where r is the correlation coefficient, x_i is the i-th x value, y_i is the i-th y value, and n is the number of data samples [40].

Next, a merit value is calculated, which describes the quality of each feature combination or feature subset based on a heuristic value. This merit value helps evaluate how well a combination of features can distinguish or predict the target variable [41], [42]. The CFS method uses this merit value as a criterion for selecting the optimal feature subset to be included in the model. The following is the formula for calculating the merit value:

$$M = \frac{kr_cf}{\sqrt{k + k(k-1)r_{ff}}} \quad (12)$$

where M is the merit score, k is the number of features, kr_cf is the average correlation between features and class labels, and r_{ff} is the average intercorrelation between features [43].

2.3. Support Vector Machine (SVM)

After performing the feature selection process with four different methods, the next step is the classification stage. The new dataset generated from feature selection is then used to perform classification. Classification is a process that categorizes or groups objects into predetermined classes or categories based on a number of specific attributes or features [44]. The main purpose of classification is to develop a model or algorithm that can predict the class or category that corresponds to an unknown data object [45], [46]. In this research, the SVM model is used as a model for classification.

In general, SVM is defined as a method that classifies linear and non-linear data [47], [48]. In the process, SVM classifies data points into two different groups in a d-dimensional space. Each data point x in the data set has a label y, which can be either $+1$ or -1. Under the assumption that the data can be linearly separated by a hyperplane. The general hyperplane equation is as follows:

$$h(x) = \boldsymbol{w}^T\boldsymbol{x} + b \quad (13)$$

So that the general hyperplane equation for two classes on the margin applies:

$$\boldsymbol{w}^T\boldsymbol{x} + b = +1 \quad (14)$$

$$\boldsymbol{w}^T\boldsymbol{x} + b = -1 \quad (15)$$

where $\boldsymbol{w}$ is the weight vector and b is a scalar or bias. The distance between the hyperplane and the closest points of the two separated classes is called the margin which is defined as $\frac{2}{\|w\|}$, where $\|w\|$ is the Euclidean norm of the weight vector. To find the optimal hyperplane, a formulation is required that involves finding w that minimizes

$$\frac{1}{2}\|w\|^2 \text{ with conditions } y_i(w^T x_i + b) \geq 1, \forall_i = 1, 2, \ldots, n \quad (16)$$

In the SVM formulation, the classification function is obtained by calculating the result of the dot product between the feature vector x and the weight vector w. But in the dual SVM formulation, the weight vector is expressed as:

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (17)$$

where α_i is the Lagrange coefficient obtained from the solution of the SVM optimization problem. By substituting equation (17) into the decision function, the form of the SVM classification function is obtained:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x_i^T x_i} + b \tag{18}$$

when the data is not linearly separable, a slack variable ξ_i is used, and the parameter C controls the misclassification penalty. The optimization problem in this case is to find the minimum value so that equation (16) changes to:

$$\min \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{19}$$

with constraints

$$y_i(\boldsymbol{w}^T x_i + b) \geq 1 - \xi_i \tag{20}$$

In the process, the SVM model tries to find the best line or hyperplane that can separate two classes of data with optimal distance [49]. For data that cannot be linearly separated, the SVM model is modified by transforming it into a higher-dimensional vector space [50]. This creates a separating plane called a separating hyperplane that separates the data according to its class. This transformation process, known as the kernel trick, in its application to non-linear SVM models requires a kernel function:

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \varphi(x_i)^T \varphi(x_j) \tag{21}$$

In this study, the dataset used is non-linear, which requires the use of a kernel function. The kernel function used is the polynomial kernel because it is suitable for data that has complex non-linear patterns and can handle a high degree of interaction between its features. The formula for calculating the polynomial kernel is as follows:

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\boldsymbol{x_i^T x_j} + C)^d \tag{22}$$

where C is a bias constant that allows adjustment to the margin distance and d indicates the degree of the polynomial used in the kernel mapping.

The values of parameters C and d in this study were adjusted based on an understanding of the data and empirical experiments to achieve optimal performance. To determine the optimal value of C, experiments were conducted with several different values, and the one that gave the best performance on validation or test data was selected.

2.4. Confusion Matrix

In the operation of classification algorithms to assess two classification groups, a confusion matrix is used to obtain information about the data used in the comparison between the actual classification results and the classification results obtained through the SVM method [51]. The shape of the confusion matrix is presented in Table 1.

Table 1. Confusion Matrix

PREDICTION	ACTUAL	
	Positive	Negative
Positive	**TP**	**FP**
Negative	**FN**	**TN**

Based on Table 1, True Positive (TP) is a correct positive prediction when the true class is also positive, True Negative (TN) is a correct negative prediction when the true class is also negative, False Positive (FP) is an incorrect positive prediction when the true class is negative, and False Negative (FN) is an incorrect negative prediction when the true class is positive.

2.5. Performance Evaluation

This study employed a number of metrics, including accuracy, precision, recall, and F1-score, to assess the effectiveness of SVM models based on the confusion matrix.

1) Accuracy

Accuracy is an evaluation metric that measures the overall ability of a model to identify both True Positive and True Negative compared to the total number of tests or classifications. The formula for calculating accuracy is as follows [44]:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FN} \quad (23)$$

2) Precision

Precision is a measure used to evaluate the extent to which a model or system is able to identify truly relevant examples among the set of examples considered positive by the model. Precision is calculated using the following formula [52]:

$$Precision = \frac{TP}{TP + FP} \quad (24)$$

3) Recall

Recall is known as the success rate in retrieving information, which is a way to evaluate the extent to which the model can find relevant information. The formula for calculating recall can be expressed in the following equation [53]:

$$Recall = \frac{TP}{TP + FN} \quad (24)$$

4) F1-Score

F1-Score is a metric used to provide an overview of the average comparison between precision and recall. The formula for calculating F1-Score can be expressed by the following equation [54]:

$$F1 - Score = \frac{2(Recall \times Precision)}{Recall + Precision} = \frac{TP}{TP + FN} \quad (25)$$

3. RESULTS AND DISCUSSION

The feature selection methods used in this research include LR-RFE, LR-SFS, CFS, and Variance Threshold. After obtaining the best features from each feature selection method, the next step is to apply the SVM model to perform classification. In this context, classification is used to predict whether someone has heart disease or not based on the features in the dataset.

Simulations were conducted by dividing the data into training data and testing data that varied from a sample ratio of 50:50 to 80:20. After that, to gauge the accuracy of the model's performance, a confusion matrix is used. The confusion matrix provides an overview of how well the model can predict data classes.

One indicator of good model performance is when the accuracy value is high, close to 100%. By calculating the accuracy using equation (23), the accuracy value of the SVM model using four feature selection methods is obtained, as shown in Table 2.

Table 2. Accuracy Results of SVM Models with Different Distributions of Training and Testing Data from Four Different Feature Selection Methods

Training:Testing	Method			
	LR-RFE-SVM	LR-SFS-SVM	CFS-SVM	Variance threshold-SVM
50:50	88%	88%	87%	88%
60:40	89%	88%	87%	88%
70:30	89%	89%	87%	89%
80:20	89%	89%	87%	89%

Table 2 and Fig. 2 above provide the results of the SVM model accuracy comparison using four different feature selection methods, which have been evaluated with various training and testing data. The analysis shows that the LR-RFE-SVM, LR-SFS-SVM, and Variance Threshold-SVM methods achieved the highest accuracy, reaching 89%. This happens because the three methods combine effective approaches to addressing the data classification problem. First, LR-RFE-SVM combines the Logistic Regression technique for recursive feature selection with SVM for classification. By performing recursive feature selection, the model can identify and use the most important features to improve classification accuracy. Furthermore, LR-SFS-SVM also utilizes Logistic Regression as the first step, but then performs forward feature selection to select the best subset

of features. By incrementally selecting the most relevant features for classification, the model can improve accuracy by reducing the dimensionality of unimportant or noisy features. Finally, Variance Threshold-SVM, this method selects features that have a variance above a threshold. In this research, 0.05 is taken as the threshold. By discarding features with low variance, the model can focus on more useful features to improve classification accuracy.

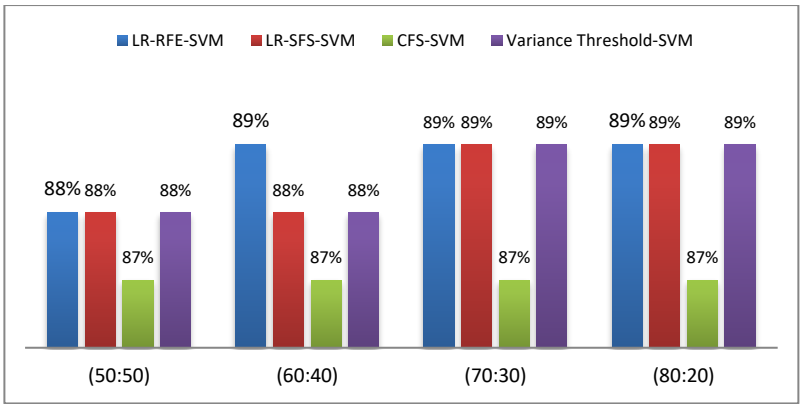

Fig. 2. Accuracy Results Comparison Chart

Meanwhile, the CFS-SVM method has lower accuracy compared to the other three methods because this method is less effective in selecting the most relevant features for the classification of heart disease data. CFS-SVM uses correlation to select features, which may not be as optimal as the recursive or forward selection approaches used by the other methods.

The importance of high accuracy in heart disease classification cannot be overlooked, given the huge impact of diagnostic errors in the clinical world. The ability to accurately predict the presence of heart disease can enable timely intervention and more effective management of patients, which in turn can reduce the risk of serious complications and improve clinical outcomes.

Additionally, there exist supplementary evaluation metrics for SVM models, namely precision, recall, and f1-score, which are computed by the application of equations (24), (25) and (26). The outcomes of these computations are displayed in Table 3.

Table 3. Model Evaluation Results of Each Feature Selection Method

Training:Testing	Model Evaluation	Method			
		LR-RFE-SVM	LR-SFS-SVM	CFS-SVM	Variance threshold-SVM
50:50	Precision	89%	89%	88%	89%
	Recall	88%	88%	87%	88%
	F1-Score	88%	88%	87%	88%
60:40	Precision	89%	89%	88%	89%
	Recall	89%	88%	87%	88%
	F1-Score	88%	88%	87%	88%
70:30	Precision	89%	89%	88%	89%
	Recall	89%	89%	87%	89%
	F1-Score	88%	89%	87%	89%
80:20	Precision	90%	89%	88%	90%
	Recall	89%	89%	87%	89%
	F1-Score	89%	89%	87%	89%

Table 3 and Fig. 3 above show a comparison of evaluation results including precision, recall, and F1-score for SVM models using four different feature selection methods across different training and testing data. Based on Table 3 and Fig. 3, it can be seen that there are variations in the performance of SVM models depending on the feature selection method used and the proportion of training-testing data selected. However, in general, LR-RFE-SVM and Variance threshold-SVM tend to provide better evaluation scores than LR-SFS-SVM and CFS-SVM, especially at higher proportions of training data. While choosing a feature selection technique, consideration of the computation time and the number of retained features of each method is crucial.

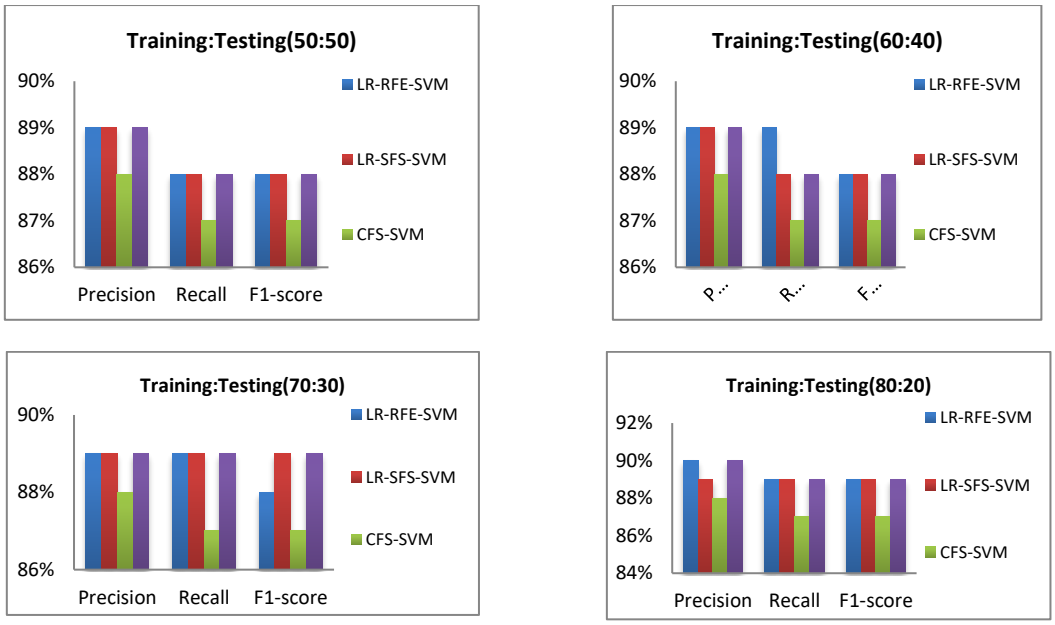

Fig. 3. Model Evaluation Result Chart

In the process of selecting a feature selection method, consideration of the computation time and the number of retained features of each method is crucial. In this study, LR-RFE-SVM takes a longer computation time, 329,7388 seconds, to retain 18 features. On the other hand, LR-SFS-SVM and CFS-SVM took 247,7429 seconds and 221,6335 seconds, respectively, to retain 20 and 17 important features. Variance Threshold-SVM, on the other hand, tends to be faster with only 118,1540 seconds but manages to retain a larger number of features, namely 23 features.

Computational efficiency and the number of retained features can have an impact on model performance. In this study, although LR-RFE-SVM requires longer computation time to retain 18 features, it provides good evaluation results. Meanwhile, other methods, such as LR-SFS-SVM and SVM-CFS, require less time but maintain almost the same number of features as or fewer than LR-RFE-SVM. Whereas Variance Threshold-SVM, despite requiring faster computation time, managed to maintain a larger number of features.

Thus, these results show that sometimes there is a compromise between computation time and the number of features retained, but in this case, there is not always a clear trade-off between the two.

The practical implications of factors such as computation time and number of features become important in the context of real-world applications or decision-making processes. For example, in a clinical scenario for heart disease classification, efficient computation time is crucial as it can affect the speed of diagnosing and treating patients. In this case, choosing a feature selection method that requires shorter computation time, such as Variance Threshold-SVM, could be a better choice if there is no significant degradation in model performance.

In addition, the number of features retained also has significant implications. In the development of heart disease classification models, retaining important and relevant features is crucial. However, too many retained features may lead to overfitting and worsen the generalization of the model. Therefore, finding a balance between retaining important features and avoiding overfitting is a top priority. In this regard, the selection of an appropriate feature selection method plays an important role in ensuring that the resulting model performs well and is reliable in clinical practice.

Taking these factors into account, this study provides deeper insights into how the choice of feature selection method can affect the practical application of heart disease classification models. As such, the findings are not only relevant in an academic context but also have significant implications for improving the diagnosis and treatment of heart disease in daily clinical practice.

4. CONCLUSION

Based on the analysis, it is concluded that feature selection techniques play an important role in improving the performance of SVM models for heart disease classification. The LR-RFE-SVM, LR-SFS-SVM, and Variance Threshold- SVM methods show the highest accuracy, reaching 89%. The variation in model

performance depends on the feature selection method and the proportion of training-testing data used. In general, LR-RFE-SVM and Variance Threshold-SVM tend to provide better evaluation results. However, consideration of computation time and the number of retained features are equally important in the selection of feature selection methods. The Variance Threshold-SVM method emerged as an efficient choice by retaining a larger number of 23 features in a shorter time compared to the other methods.

These results reflect the importance of selecting a suitable feature selection method for large datasets, such as heart disease. This research, therefore offers a better understanding of the effectiveness of various feature selection methods in improving the performance of SVM models in classification. Future development prospects could involve the exploration of other feature selection methods, the exploration of new feature selection techniques, and the exploration of ensemble approaches that combine several classification models. In addition, the application to more diverse datasets can be considered for future research.

Thus, it is expected that this research not only contributes substantially to the understanding of the efficacy and efficiency of feature selection methods in the the context of heart disease classification but can also inspire further research in the development of better solutions for big data analysis.

REFERENCES

[1] M. M. Alani, "Big data in cybersecurity: a survey of applications and future trends," *J. Reliab. Intell. Environ.*, vol. 7, pp. 85–114, Jun. 2021, https://doi.org/10.1007/s40860-020-00120-3.

[2] Z. Allam and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," *Cities*, vol. 89, pp. 80–91, 2019, https://doi.org/10.1016/j.cities.2019.01.032.

[3] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *J. Big Data*, vol. 6, p. 54, Dec. 2019, https://doi.org/10.1186/s40537-019-0217-0.

[4] J. Sheng, J. Amankwah-amoah, Z. Khan, and X. Wang, "COVID-19 Pandemic in the New Era of Big Data Analytics : Methodological Innovations and Future Research Directions," *Br. J. Manag.*, vol. 32, pp. 1164–1183, 2021, https://doi.org/10.1111/1467-8551.12441.

[5] L. Yang, K. Zhang, Z. Chen, and Y. Liang, "Fault diagnosis of WOA-SVM high voltage circuit breaker based on PCA Principal Component Analysis," *Energy Reports*, vol. 9, pp. 628–634, 2023, https://doi.org/10.1016/j.egyr.2023.04.341.

[6] G. T. Reddy *et al.*, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020, http://dx.doi.org/10.1109/ACCESS.2020.2980942.

[7] S. Singh and K. Malik, "Feature selection and classification improvement of Kinnow using SVM classifier," *Meas. Sensors*, vol. 24, p. 100518, Dec. 2022, https://doi.org/10.1016/j.measen.2022.100518.

[8] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019, https://doi.org/10.1016/j.eswa.2019.05.028.

[9] M. Alishahi, V. Moghtadaiee, and H. Navidan, "Add noise to remove noise: Local differential privacy for feature selection," *Comput. Secur.*, vol. 123, p. 102934, 2022, https://doi.org/10.1016/j.cose.2022.102934.

[10] S. Ihlenfeldt, S. Wrobel, D. Weichert, P. Link, A. Stoll, and R. Stefan, "A review of machine learning for the optimization of production processes," *Orig. Artic.*, vol. 104, pp. 1882–1902, 2019, https://doi.org/10.1007/s00170-019-03988-5.

[11] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction : a review," *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2663–2693, 2022, https://doi.org/10.1007/s40747-021-00637-x.

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

[13] S. K. Nasib, F. I. Pammus, Nurwan, and L. O. Nashar, "comparison of feature selection based on computation time and classification accuracy using support vector machine," *Indones. J. Appl. Res.*, vol. 4, pp. 63–74, Apr. 2023, https://doi.org/10.30997/ijar.v4i1.252.

[14] V. Ghate and S. Hemalatha C, "A comprehensive comparison of machine learning approaches with hyper-parameter tuning for smartphone sensor-based human activity recognition," *Meas. Sensors*, vol. 30, p. 100925, Dec. 2023, https://doi.org/10.1016/j.measen.2023.100925.

[15] P. Theerthagiri, "Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique," *Intell. Syst. with Appl.*, vol. 16, no. 2667–3053, p. 200121, Nov. 2022, https://doi.org/10.1016/j.iswa.2022.200121.

[16] Y. Siti Ambarwati and S. Uyun, "Feature Selection on Magelang Duck Egg Candling Image Using Variance Threshold Method," in *3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 694–699, Dec. 2020, https://doi.org/10.1109/ISRITI51436.2020.9315486.

[17] S. Moon and Y. Kim, "An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression," *Atmos. Res.*, vol. 240, p. 104928, Aug. 2020, https://doi.org/10.1016/j.atmosres.2020.104928.

[18] X. Hou, S. Li, X. Zhang, and H. Jiang, "Identification of key influencing factors of railway tunnel water-proof

slab test data based on RF-SFS algorithm," *Meas. Sensors*, vol. 18, p. 100144, Dec. 2021, https://doi.org/10.1016/j.measen.2021.100144.

[19] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, https://doi.org/10.1016/j.asoc.2019.105524.

[20] E. Tuba, I. Strumberger, T. Bezdan, N. Bacanin, and M. Tuba, "Classification and Feature Selection Method for Medical Datasets by Brain Storm Optimization Algorithm and Support Vector Machine," *Procedia Comput. Sci.*, vol. 162, pp. 307–315, 2019, https://doi.org/10.1016/j.procs.2019.11.289.

[21] N. V Sharma and N. Singh, "An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers," *Microprocess. Microsyst.*, vol. 85, pp. 0141–9331, 2021, https://doi.org/10.1016/j.micpro.2021.104293.

[22] M. Artur, "Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features," *Procedia Comput. Sci.*, vol. 190, pp. 564–570, 2021, https://doi.org/10.1016/j.procs.2021.06.066.

[23] D. Albashish, A. I. Hammouri, M. Braik, J. Atwan, and S. Sahran, "Binary biogeography-based optimization based SVM-RFE for feature selection," *Appl. Soft Comput.*, vol. 101, p. 107026, Mar. 2021, https://doi.org/10.1016/j.asoc.2020.107026.

[24] B. Richhariya, M. Tanveer, and A. H. Rashid, "Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE)," *Biomed. Signal Process. Control*, vol. 59, p. 101903, May 2020, https://doi.org/10.1016/j.bspc.2020.101903.

[25] H. Wu, "A Deep Learning-Based Hybrid Feature Selection Approach for Cancer Diagnosis," *J. Phys. Conf. Ser.*, vol. 1848, no. 1, p. 012019, Apr. 2021, http://dx.doi.org/10.1088/1742-6596/1848/1/012019.

[26] J. Sheng, M. Shao, Q. Zhang, R. Zhou, L. Wang, and Y. Xin, "Alzheimer's disease, mild cognitive impairment, and normal aging distinguished by multi-modal parcellation and machine learning," *Sci. Rep.*, vol. 10, no. 1, p. 5475, Mar. 2020, https://doi.org/10.1038/s41598-020-62378-0.

[27] Karthikeyan T., K. Sekaran, Ranjith D., Vinoth Kumar V., and Balajee J M, "Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques," *Int. J. Web Portals*, vol. 11, no. 2, pp. 41–52, Jul. 2019, http://dx.doi.org/10.4018/IJWP.2019070103.

[28] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *J. Data Anal. Inf. Process.*, vol. 07, pp. 190–207, 2019, https://doi.org/10.4236/jdaip.2019.74012.

[29] A. Balboa, A. Cuesta, J. González-Villa, G. Ortiz, and D. Alvear, "Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations," *Saf. Sci.*, vol. 174, no. March, p. 106485, Jun. 2024, https://doi.org/10.1016/j.ssci.2024.106485.

[30] S. D. Id, M. Avalos, E. Lagarde, and M. Schumacher, "Penalized logistic regression with low prevalence exposures beyond high dimensional settings," pp. 1–14, 2019, https://dx.doi.org/10.1371/journal.pone.0217057.

[31] T. J. VanderWeele, "Optimal approximate conversions of odds ratios and hazard ratios to risk ratios," *Biometrics*, vol. 76, no. 3, pp. 746–752, Sep. 2020, https://doi.org/10.1111/biom.13197.

[32] S. O. Aregbesola, J. Won, S. Kim, and Y. Byun, "Sequential backward feature selection for optimizing permanent strain model of unbound aggregates," *Case Stud. Constr. Mater.*, vol. 19, no. 8, p. e02554, Dec. 2023, https://doi.org/10.1016/j.cscm.2023.e02554.

[33] Y. Yulianti and A. Saifudin, "Sequential Feature Selection in Customer Churn Prediction Based on Naive Bayes," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, p. 012090, Jul. 2020, http://dx.doi.org/10.1088/1757-899X/879/1/012090.

[34] K. Chotchantarakun and O. Sornil, "An Adaptive Multi-levels Sequential Feature Selection," vol. 13, no. 9, pp. 10–19, 2021, http://dx.doi.org/10.5614/itbj.ict.res.appl.2021.15.1.1.

[35] 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, https://doi.org/10.1109/ACCESS.2022.3153047.

[36] S. Shafiee, L. M. Lied, I. Burud, J. A. Dieseth, M. Alsheikh, and M. Lillemo, "Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery," *Comput. Electron. Agric.*, vol. 183, no. 1432, p. 106036, Apr. 2021, https://doi.org/10.1016/j.compag.2021.106036.

[37] K. Chotchantarakun, "Optimizing Sequential Forward Selection on Classification Using Genetic Algorithm Related work Feature selection process," *Informatica,* vol. 47, pp. 81–90, 2023, http://dx.doi.org/10.31449/inf.v46i9.4964.

[38] M. Al Fatih Abil Fida, T. Ahmad, and M. Ntahobari, "Variance Threshold as Early Screening to Boruta Feature Selection for Intrusion Detection System," in *13th International Conference on Information & Communication Technology and System (ICTS)*, pp. 46–50, Oct. 2021, https://doi.org/10.1109/ICTS52701.2021.9608852.

[39] T. Avraham, M. Dror, S. Ariel, and D. Amit, "Word embedding dimensionality reduction using dynamic variance thresholding (DyVaT)," *Expert Syst. Appl.*, vol. 208, no. 0957–4174, 2022, https://doi.org/10.1016/j.eswa.2022.118157.

[40] D. Risqiwati, "Feature Selection for EEG-Based Fatigue Analysis Using Pearson Correlation," *Humanification Reliab. Intell. Syst. ISITIA*, pp. 164–169, 2020, https://dx.doi.org/10.1109/ISITIA49792.2020.9163760.

[41] A. Ranjan, V. P. Singh, R. B. Mishra, A. K. Thakur, and A. K. Singh, "Sentence polarity detection using stepwise

greedy correlation based feature selection and random forests: An fMRI study," *J. Neurolinguistics*, vol. 59, no. 0911–6044, p. 100985, Aug. 2021, doi: https://doi.org/10.1016/j.jneuroling.2021.100985.
[42] T. Ridwan, B. Kushal, and M. S. Illindala, "Electrical Power and Energy Systems Correlation-based feature selection for resilience analysis of MVDC shipboard power system ☆," *Electr. Power Energy Syst.*, vol. 117, no. November 2019, p. 105742, 2020, https://doi.org/10.1016/j.ijepes.2019.105742.
[43] M. Mohamad, A. Selamat, O. Krejcar, R. G. Crespo, E. Herrera-viedma, and H. Fujita, "Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection," *Electronics*, pp. 1–24, 2021, https://dx.doi.org/10.3390/electronics10232984.
[44] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties," *Commun. Stat. - Theory Methods*, vol. 49, no. 9, pp. 2080–2093, May 2020, https://doi.org/10.1080/03610926.2019.1568485.
[45] V. Blanco, A. Japón, and J. Puerto, "A mathematical programming approach to SVM-based classification with label noise," *Comput. Ind. Eng.*, vol. 172, p. 108611, Oct. 2022, https://doi.org/10.1016/j.cie.2022.108611.
[46] J. Cai and N. Xi, "Site classification methodology using support vector machine: A study," *Earthq. Res. Adv.*, p. 100294, Mar. 2024, https://doi.org/10.1016/j.eqrea.2024.100294.
[47] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazanci, and O. Oral, "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture," *J. King Saud Univ. - Sci.*, vol. 35, no. 6, p. 102754, Aug. 2023, https://doi.org/10.1016/j.jksus.2023.102754.
[48] K. C. Onyelowe, C. B. Mahesh, B. Srikanth, C. Nwa-David, J. Obimba-Wogu, and J. Shakeri, "Support vector machine (SVM) prediction of coefficients of curvature and uniformity of hybrid cement modified unsaturated soil with NQF inclusion," *Clean. Eng. Technol.*, vol. 5, p. 100290, Dec. 2021, https://doi.org/10.1016/j.clet.2021.100290.
[49] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)," *Informatics Med. Unlocked*, vol. 16, no. May, p. 100200, 2019, https://doi.org/10.1016/j.imu.2019.100200.
[50] N. A. Utami, W. Maharani, and I. Atastina, "Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method," *Procedia Comput. Sci.*, vol. 179, pp. 177–184, 2021, https://doi.org/10.1016/j.procs.2020.12.023.
[51] G. Phillips *et al.*, "Setting nutrient boundaries to protect aquatic communities: The importance of comparing observed and predicted classifications using measures derived from a confusion matrix," *Sci. Total Environ.*, vol. 912, p. 168872, Feb. 2024, https://doi.org/10.1016/j.scitotenv.2023.168872.
[52] A. Y. Mahmoud, D. Neagu, D. Scrimieri, A. Rashad, and A. Abdullatif, "Early diagnosis and personalised treatment focusing on synthetic data modelling : Novel visual learning approach in healthcare," *Comput. Biol. Med.*, vol. 164, 2023, https://dx.doi.org/10.1016/j.compbiomed.2023.107295.
[53] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, p. 012076, Jul. 2020, https://doi.org/10.1088/1757-899X/879/1/012076.
[54] Z. DeVries *et al.*, "Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability," *Spine J.*, vol. 21, no. 7, pp. 1135–1142, Jul. 2021, https://doi.org/10.1016/j.spinee.2021.02.007.

BIOGRAPHY OF AUTHORS

Winarsi J. Bidul, is a mathematics student at Ahmad Dahlan University who is interested in data science. She has attended basic to advanced classes and improved her skills to become a tutor at the Data Science Study Center of the Mathematics Study Program. Email: winarsi2000015011@webmail.uad.ac.id.

Sugiyarto Surono, is a senior lecturer in Mathematics at Ahmad Dahlan University. His deep interest focuses on the implementation of Big Data, Machine Learning, Artificial Intelligence, and Deep Learning. As a professor, he actively teaches and mentors Mathematics students. Sugi combines several scientific fields through the application of Data Science to make his students aware of the importance of this course in technological development. Email: sugiyarto@math.uad.ac.id.

Tri Basuki Kurniawan, is an experienced lecturer, researcher, and senior Python programmer. He has expertise in Oracle Database, PhoneGap, PHP, Computer Science, and Windows Server. He is also a data scientist who holds a PhD in Computer Science from Universiti Kebangsaan Malaysia. Email: tribasukikurniawan@ukm.edu.my.