Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

Implementation of Feature Selection Information Gain in Support Vector Machine Method for Stroke Disease Classification

Anisa Fitri, Iis Afrianty*, Elvia Budianita, Siska Kurnia Gusti

Faculty of Science and Technology, Informatics Engineering, Sultan Syarif Kasim Riau State Islamic University, Pekanbaru, Indonesia
Email: 12150120006@students.uin-suska.ac.id, *iis.afrianty@uin-suska.ac.id, elvia.budianita@uin-suska.ac.id, siskakurniagusti@uin-suska.ac.id
Correspondence Author Email: iis.afrianty@uin-suska.ac.id

Abstract
Stroke is a disease with a high mortality and disability rate that requires early detection. However, the main challenge in the classification process of this disease is data imbalance and the large number of irrelevant features in the dataset. This study proposes a combination of Support Vector Machine (SVM) method with Information Gain feature selection technique and data balancing using Synthetic Minority Over-sampling Technique (SMOTE) to improve classification accuracy. The dataset used consists of 5,110 data with 10 variables and 1 label. Feature selection was performed with three threshold values (0.04; 0.01; and 0.0005), while SVM classification was tested on three different kernels: Linear, RBF, and Polynomial. Model evaluation was performed using Confusion Matrix and training and test data sharing using k-fold cross validation with k=10. The best results were obtained on the RBF kernel with Cost=100 and Gamma=5 parameters at an Information Gain threshold of 0.0005, with accuracy reaching 90.51%. These results show that the combination of techniques used aims to determine the variables that most affect SVM classification in detecting stroke disease.

Keywords: Information Gain; Stroke Classification; Machine Learning; SMOTE; Support Vector Machine

1. INTRODUCTION

Stroke is a sudden brain disorder due to circulatory disorders that can permanently damage brain cells [1] [2] [3]. The cause of stroke occurs due to rupture of blood vessels (hemorrhagic stroke) or blockage of blood vessels in the brain (ischemic stroke) which blocks the flow of oxygen and nutrients, if not treated immediately this condition can cause brain cell death [4] [5]. If left untreated, stroke can lead to permanent disability or death [6]. Especially in hemorrhagic stroke, rupture of cerebral blood vessels can cause extensive bleeding and significantly damage brain tissue [7].

Some of the risk factors that contribute to stroke incidence include age, gender, history of hypertension, cholesterol levels, obesity, coronary heart disease, smoking, consumption of high-salt foods, and lack of physical activity [8]. Among these factors, hypertension is the most dominant, with the highest influence value of a mean of 0.994 [9]. Hypertension is also known to be a major cause of intracerebral hemorrhage, with a prevalence of more than 60% among stroke patients [7]. Early detection of stroke is very important because the initial symptoms are often not recognized, and low public knowledge contributes to delays in treatment [10].

Advances in technology now allow the use of artificial intelligence, particularly Machine Learning, in stroke disease detection more effectively [11]. By using Machine Learning, the system can learn patterns from medical data without the need for explicit programming, which can speed up and improve the accuracy of medical analysis [12]. One method in Machine Learning that is often used for medical data classification is Support Vector Machine (SVM). SVM works by finding a hyperplane to separate data classes with the largest margin so that it can produce more accurate classifications [13] [14].

Previous studies have shown that SVM can be applied for stroke disease classification with varying results. For example, research by [15]. Using linear, polynomial, sigmoid, and RBF kernels on a dataset containing 5,110 data with 11 variables. The results show that the polynomial kernel produces the highest accuracy of 78.86%, with 73.98% precision and 56.75% recall on 80:20 training and test data. However, this accuracy is still very low. Another study by [16] implemented the SVM algorithm in predicting stroke disease using polynomial, RBF, and sigmoid kernels with 80:20 training and test data, the dataset used was 5,110 data with 11 attributes and produced the highest accuracy of 95%. Research by [17] used the SMOTE method to balance data that had 5110 data and 11 attributes, achieving 92% accuracy with RBF parameters, Cost 100, and Gamma 1. In addition, research by[18] shows that the application of SVM with Gaussian RBF kernel is also able to provide high accuracy, which is 93.90% on elementary school accreditation data, with parameters $\sigma = 3$ and $C = 1$. This shows that the selection of the right kernel and parameters greatly affects the performance of SVM classification. Although SVM has shown good potential in a wide range of parameters, the accuracy obtained in some studies can still be improved with parameter adjustments and optimal data processing approaches.

One way to improve SVM accuracy is to apply Feature Selection techniques before classification [19]. Feature Selection is the process of selecting the most relevant features to reduce data dimensions, improve accuracy, prevent overfitting and make the model easier to understand [20]. In this study using Information Gain feature selection. Information gain (IG) is a method widely used for feature selection and filtering [21]. Related research on Information Gain is one of them by [22] applying Information Gain in feature selection for Naïve Bayes Classifier classification with 200 data and obtaining an increase in accuracy with a difference of 4% with a threshold between 0.01 to 0.10. Without using the Information Gain method (threshold 0) of 70%, while using the Information Gain method (threshold 0.01) of

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

74%. In addition, research by [23] shows that the Information Gain feature selection on sentiment analysis with SVM using a tweet dataset from the Twitter application as much as 496 data produces 92% accuracy, 90% precision, and 92% recall, with 80:20 testing using a linear kernel. further research by [24] shows that the SVM model with the application of feature selection using Information Gain (SVM-IG) produces the best accuracy of 72.45%, an increase of 3.08% from the initial accuracy of 69.36%. The average increase in accuracy after optimization is 2.51%. SVM-IG shows better performance, so the proposed model is proven to be able to improve classification accuracy in SVM..

In optimizing the performance and accuracy of the model on stroke disease classification, this research will use Information Gain to select the best features from the 10 variables in the dataset. This approach aims to select the most relevant features and reduce less important data, so as to speed up processing time and improve model accuracy. Thus, the application of SMOTE to handle data imbalance and feature selection. It is expected to produce a faster, more accurate, and more effective SVM model in early stage stroke classification, providing a more efficient and beneficial solution for stroke prevention.

2. RESEARCH METHODOLOGY

Research methodology is the stages that will be carried out during the implementation of research and is well organized and systematic. The research methodology is used as a reference or guideline during research in order to achieve the expected goals. This research uses the Support Vectore Machine method for stroke disease classification.

2.1 Research Stages

Research stages will perform several stages as shown below. First, starting from data collection, then selecting data on the dataset, and then preprocessing the data where there are several processes carried out in preprocessing the data, including data cleaning, data transformation, normalization and smote. After preprocessing, continue to perform feature selection using Information Gain, after that classify the Support Vectore Machine (SVM) method, and continue to evaluate using Confusion Matrix. So that conclusions can be drawn in this study. The following stages of research can be seen in Figure 1.

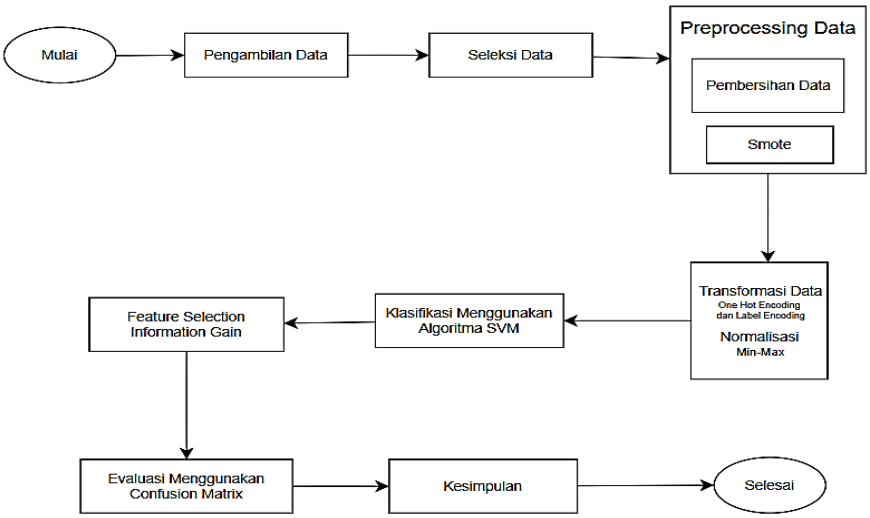

Figure 1. Research Stages

2.2 Data Collection

The dataset used in the study is a stroke prediction dataset obtained from Kaggle (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset). This dataset consists of 5,110 data with 10 variables and 1 label. The available variables include id, stroke (label), gender, age, hypertension, heart disease, ever married, work type, residence type, avg glucose level, BMI, and smoking status. Can be seen in Table 1.

Table 1. Initial Research Data

Id	Gender	Age	Hypertension	Heart_diease	Ever_married	Work_type	Residence_type	Avg_glucose	BMI	Smoking-status	Stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1

Id	Gender	Age	Hypertension	Heart_diease	Ever_married	Work_type	Residence_type	Avg_glucose	BMI	Smoking-status	Stroke
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	Never smoked	1
56669	Male	81	0	0	Yes	Private	Urbal	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
44679	Female	44	0	0	Yes	Govt_job	Urban	85.28.00	26.02.00	Unknown	0

2.3 Data Selection

Data selection is the process of selecting relevant features in the initial dataset. Data that is not important or not needed in the data will be deleted. The initial dataset consists of 12 variables, but after going through the feature selection process, only 11 variables are used, namely: stroke, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, BMI, and smoking_status. Of all these variables, stroke is used as the target variable or label, while the other 10 variables function as variables or features in the classification process. The id variable contained in the initial dataset was removed because it had no relevance to the classification process and did not contain information that could help in predicting stroke risk.

2.4 Preprocessing

Data preprocessing is the initial stage in data mining that aims to transform raw data into a cleaner format ready for use in analysis or modeling. This process includes data cleaning, transformation, normalization, and handling of missing or inconsistent data to make the analysis results more accurate and efficient.

2.4.1 Data Cleaner

Data cleaning aims to handle missing values so that only valid and relevant variables are used in the analysis. In the BMI variable, the missing value (NaN) is replaced with the average value (mean) of the BMI variable. This step is done to avoid bias and ensure the data is ready to be used in the model training process.

2.4.2 Data Transformation

Data transformation is the process of changing the coding of categorical variables into numerical form so that they can be understood and processed by machine learning models, one of which is by using One-Hot Encoding and Label Encoding techniques. One Hot Encoding is converting categorical variables into a format that the model can use by creating a binary column for each category. While Label Encoding is converting categories into numbers that represent these categories [25]. In the dataset there are 5 categorical variables such as Gender, Ever_married, Work_type, Residence_type, and Smoking_status that need to be transformed to a numerical format in order to be processed by the Support Vector Machine (SVM) algorithm. Such as Gender, Work_type, and Smoking_status variables are transformed using One-Hot Encoding, which results in new variables such as Gender_male, Work_type_private and Smoking_status_never. Features that have been One-Hot Encoded produce categories with values of 1 and 0. Meanwhile, Ever_married and Residence_type variables are transformed using Label Encoding. Besides categorical transformation, there are also other types of transformation such as normalization.

Data normalization aims to rescale numerical values into a smaller range, from a scale of 0 to 1 or from -1 to 1[26]. One commonly used method is min-max normalization. Min-max normalization is the process of performing a linear transformation on the original data [27]. The min-max normalization formula is shown in Equation 1.

$$'\mathrm{X_i} = \frac{\mathrm{X_i} - \min_{(x)}}{\max_{(x)} - \min_{(x)}} (\max_{(baru)} - \min_{(baru)}) + \min_{(baru)} \quad 1$$

a. X_i = State the original value
b. $'X_i$ = Values after normalization
c. $max_{(x)} - min_{(x)}$ = Minimum and maximum values of the change in X
d. $max_{(baru)} - min_{(baru)}$ = New desired range

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

2.5 Synthetic Minority Over-Sampling Technique (SMOTE)

Synthetic Minority Over sampling Technique (SMOTE) is an oversampling used to overcome the problem of lack of data related to minority groups, the aim is to multiply minority data by creating synthetic data similar to existing minority data [28]. In this study, the stroke case uses the smote method to overcome the problem of lack of data. In Figure 2 is the data before SMOTE, there are 4,861 total data in class 0 (no stroke) and 249 data in class 1 (stroke) from a of 5110 data.

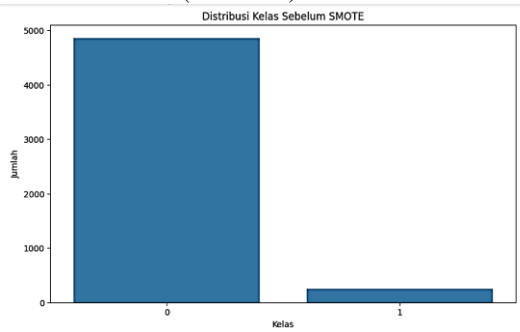

Figure 2. Data before Smote

Figure 3 is the data after applying Smote, the data becomes balanced with a total of 9,722 samples, consisting of 4,861 for class 0 and 4,861 for class 1. With this balance, the model can detect patterns in both classes more effectively.

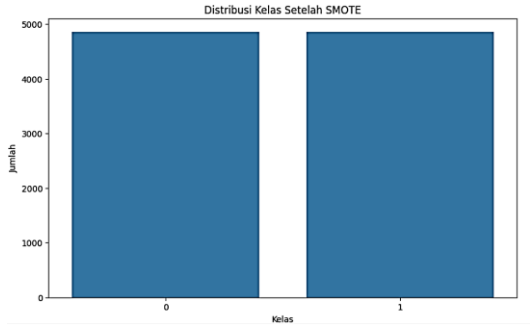

Figure 3. Data after smote

2.6 Information Gain Feature Selection

Feature selection is to determine the most influential variables in improving model accuracy. This research uses Information Gain, which measures the contribution of each variable based on entropy reduction. The steps of Information Gain are as follows: The first stage of feature selection separates 10 variables according to class, then calculates the total entropy. Next, calculate the Information Gain then the variables are arranged descendingly based on the Information Gain value. The final stage is to select the top ranking variable. This research uses three thresholds: 0.01; 0.04; and 0.0005 to determine the most influential features. Information gain formulas include [21]. The Information Gain formula is shown in Equations 2 and 3

$$Entropy\ (S) = \sum_{i}^{c} -p_i log_2 p_j \quad (2)$$

$$Information\ Gain\ (S, A) = Entropy\ (S) - \sum_{values(A)} \frac{S_v}{S_v} Entropy\ (s_v) \quad (3)$$

Information Gain (S, A) is a measure of the reduction in entropy (uncertainty) that occurs after the data S is divided based on variable A. It is calculated from the difference between the entropy (S) before division and the weighted average of the entropy (Sv), i.e. the entropy on each subset Sv for v values of variable A. The higher the Gain (S, A) value, the greater the contribution of variable A in differentiating the data towards the classification target.

2.7 K-Fold Cross Validation

K-Fold Cross Validation is testing data that is divided into folds and the model is trained and tested on each different fold in turn [29]. This research test uses k-fold cross validation with k = 10, which divides the data into 10 folds. Each iteration, the data is divided into training data to train the model and test data to evaluate its performance. This technique ensures a more accurate evaluation of the model.

2.8 Support Vector Machine (SVM) Method

The Support Vector Machine method uses kernels to map low-dimensional nonlinear data into a higher dimensional space, thus allowing linear separation of data with a hyperplane. There are several types of kernels in SVM, namely linear kernels

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

that are suitable for linearly separable data, polynomials that use polynomials of a certain degree to handle more complex data, RBF is often used because it is able to handle complex and irregular data patterns. The choice of kernel depends on the characteristics of the dataset and the classification needs.

Classification testing is done using the k-fold cross-validation technique with a value of k=10. Three types of kernels used in the SVM method are Linear, Radial Basis Function (RBF), and Polynomial. For each kernel, testing is done with various parameters. The Linear kernel uses the C parameter with values of 1, 10, and 100. The second kernel is RBF using various values of C = 1, 10 and 100, and the gamma used (γ) = 1, 4, and 5. Finally, the Polynomial kernel, with degree values = 1 and 2. This test aims to determine the best configuration of parameters in detecting the risk of stroke disease and evaluate the effect of parameter combinations on accuracy. The following kernel functions are used in Table 2 [21].

Table 2. Kernel Function

Kernel Name	Kernel Function	
Linear	$K(x_i, x_j) = x_j., x_j$	4
RBF	$K(x_i, x_j) = (x_j., x_j + c)^d$	5
Polynomial	$K(x_i, x_j) = exp(-y\|\|x_j - x_j\|\|^d$	6

2.9 Confusion Matrix Evaluation

Confusion matrix is a part of machine learning by studying existing data and grouping it as new data by giving out results with categorical Nominal or Ordinal variables. In Confusion Matrix there are 4 (four) terms used to represent the process of classification results TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). The confusion matrix formula is shown in Equations 7 through 10.

$$\text{Accuracy} \quad \frac{(\text{TP} + \text{TN})}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{X100\%} \quad (7)$$

$$\text{Precision} \quad \frac{(\text{TP} + \text{TN})}{\text{TP} + \text{FP}} \text{X100\%} \quad (8)$$

$$\text{Recall} \quad \frac{\text{TP}}{\text{TP} + \text{FN}} \text{X100\%} \quad (9)$$

$$\text{F1 Score} \quad \frac{2(\text{Precision x Recall})}{(\text{Precision} + \text{Recall})} \quad (10)$$

a. Accuracy: is a measure that shows how close the prediction results are to the actual data, which is the percentage of correctly classified test data out of all test data.
b. Precision: is a measure of the accuracy of positive predictions, which is the ratio of correct positive predictions to all positive predictions made by the model.
c. Recall: is a measure of the model's ability to detect positive data correctly, which is the ratio of positive data that is successfully recognized to all actual positive data.
d. F1 Score: is a harmonic mean between precision and recall that reflects the balance of the two, especially useful when the data is not balanced.

3. RESULTS AND DISCUSSION

This study uses 5,110 stroke data with 10 variables that have previously been carried out in the data cleaning stage in the preprocessing process. The implementation was done using Python programming using Google Colab. Experiments were conducted by applying feature selection using Information Gain. Using the Support Vector Machine (SVM) classification method, using three types of kernels namely linear, RBF, and polynomial. The performance of the model is evaluated using Confusion Matrix to measure accuracy results and testing using K-Fold 10. The accuracy results of each kernel are presented in tabular form to be compared and analyzed further.

3.1 Feature Selection Result

The figure below shows the feature selection process using Information Gain with a threshold of 0.04. This means that only features that have an Information Gain value greater than or equal to 0.04 will be selected for the next classification stage. The use of this threshold is quite high, so only features that are highly influential on the target class will be used.

```
# Threshold for feature selection
threshold = 0.04
```

Figure 4. Threshold 0,04

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

Figure 5 shows the same process, but with a threshold of 0.01. Compared to the previous threshold, this threshold is lower, so more features are used. This allows the model to consider features with moderate influence on the target class. This threshold is usually used to maintain a balance between efficiency and diversity of information from the features.

```
# Threshold for feature selection
threshold = 0.01
```

Figure 5. Threshold 0,01

The last figure uses a very low threshold of 0.0005. Almost all features that have even a small Information Gain value will still be selected. The aim is that no potentially informative features are missed, although this may increase the risk of including less relevant features.

```
# Threshold for feature selection
threshold = 0.0005
```

Figure 6. Threshold 0,0005

Each of these thresholds was used to explore their impact on feature selection and model performance. A higher threshold tends to result in a simpler model, while a lower threshold allows the model to consider more features.

The figure below shows the results of feature selection using the Information Gain algorithm with a threshold of 0.0005. Through this process, 18 features were selected that are considered to have the greatest contribution to the classification process. This selection aims to simplify the model without sacrificing important information while maintaining features that are relevant to the prediction target. The selected features include: avg_glucose_level, bmi, age, work_type_children, ever_married, work_type_Self-employed, smoking_status_formerly smoked, smoking_status_Unknown, work_type_Private, work_type_Govt_job, smoking_status_smokes, hypertension, heart_disease, gender_Male, gender_Female, smoking_status_never smoked, work_type_Never_worked, and Residence_type. Of all these features, avg_glucose_level has the highest Information Gain value of 0.981916, which indicates that this feature has the most significant influence on class prediction. Meanwhile, Residence_type is the feature with the lowest Information Gain value, which is 0.001453, but still exceeds the threshold value and is considered to still contribute enough to be included in the modeling. The results of the feature selection can be seen in Figure 7.

```
=== Information Gain Calculation Results (All Features) ===
Fitur: avg_glucose_level, Information Gain: 0.981916
Fitur: bmi, Information Gain: 0.858017
Fitur: age, Information Gain: 0.851187
Fitur: work_type_children, Information Gain: 0.072480
Fitur: ever_married, Information Gain: 0.049179
Fitur: smoking_status_Unknown, Information Gain: 0.027713
Fitur: smoking_status_formerly smoked, Information Gain: 0.025432
Fitur: work_type_Self-employed, Information Gain: 0.024124
Fitur: work_type_Private, Information Gain: 0.023035
Fitur: work_type_Govt_job, Information Gain: 0.020457
Fitur: smoking_status_smokes, Information Gain: 0.011132
Fitur: hypertension, Information Gain: 0.008315
Fitur: heart_disease, Information Gain: 0.007504
Fitur: smoking_status_never smoked, Information Gain: 0.006512
Fitur: gender_Male, Information Gain: 0.006398
Fitur: gender_Female, Information Gain: 0.006390
Fitur: work_type_Never_worked, Information Gain: 0.002267
Fitur: Residence_type, Information Gain: 0.001632
Fitur: gender_Other, Information Gain: 0.000103

=== Selected Features (IG > threshold) ===
- avg_glucose_level
- bmi
- age
- work_type_children
- ever_married
- smoking_status_Unknown
- smoking_status_formerly smoked
- work_type_Self-employed
- work_type_Private
```

Figure 7. Feature Selection Result

3.2 Classification Results

This study shows that without the application of SMOTE and Information Gain, the SVM model performance is still not optimal due to data imbalance. The RBF kernel gives the best results with 82.54% accuracy and 86.49% F1-Score, while the Linear and Polynomial kernels are quite low in accuracy. When compared to previous research [15] which only achieved 78.86% accuracy, 73.98% precision, and 56.75% recall on 80:20 training and test data, these results show a significant improvement in performance. This improvement shows that the selection of the right kernel such as RBF can produce better performance even before the feature selection and data balancing stages are carried out, but further optimization is still needed so that recall and accuracy increase. This can be seen in Table 3.

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

Table 3. Overall Average Results without Smote and Information Gain

Kernel	Accuracy	Precision	Recall	F1-Score
Linear	72.64%	94.51%	72.64%	80.42%
RBF	82.54%	91.68%	82.54%	86.49%
Polynomial	72.81%%	94.31%	72.81%	80.53%

The Overall average results table without SMOTE using Information Gain presents the performance of the model after applying feature selection using three threshold values (0.04; 0.01; and 0.0005), but without the use of SMOTE. The results displayed from the average accuracy across all parameter combinations for each kernel and threshold value show that the RBF kernel remains superior across all thresholds, with the highest accuracy reaching 82.78% and F1-Score up to 86.50%. The Polynomial kernel shows stable performance with high precision, but the recall accuracy values remain low. The Linear kernel also showed similar results, with high precision but low recall and accuracy. This shows that although feature selection can help simplify the model, without data balancing such as SMOTE, the model still struggles to optimally recognize minority classes. This can be seen in Table 4.

Table 4. Overall Average Results without Smote using Information Gain

Kernel	Threshold	Accuracy	Precision	Recall	F1-Score
Linear	0,04	72.68%	94.52%	72.68%	80.46%
RBF		82.78%	91.71%	82.78%	86.16%
Polynomial		72.77%	94.32%	72.77%	80.53%
Linear	0,01	72.57%	94.51%	72.57%	80.38%
RBF		82.69%	91.75%	82.69%	86.60%
Polynomial		72.72%	94.29%	72.72%	80.48%
Linear	0,0005	72.68%	94.51%	72.68%	80.46%
RBF		82.63%	91.71%	82.63%	86.56%
Polynomial		72.74%	94.37%	72.74%	80.50%

The test results show the performance of the Support Vector Machine (SVM) model using three types of kernels, namely Linear, RBF, and Polynomial without the application of feature selection. Before the model training process, the data is balanced using the SMOTE method to overcome class imbalance by generating synthetic data in the minority class. The test results shown are the average of all tested parameters for each kernel. The RBF kernel shows the best performance with the highest accuracy reaching 88.51%, precision 89.06%, recall 88.51%, and F1-Score 88.47%. Furthermore, the polynomial kernel obtained an accuracy of 80.82%, while the linear kernel showed the lowest result with an accuracy of 79.60%. Can be seen in Table 5.

Table 5. Overall Average Results using Smote without Information Gain

Kernel	Accuracy	Precision	Recall	F1-Score
Linear	79.60%	80.32%	79.60%	79.49%
RBF	88.51%	89.06%	88.51%	88.47%
Polynomial	80.82%%	81.46%	80.88%	80.72%

Test results after three trials using Information Gain feature selection and various types of SVM kernels, namely Linear, RBF, and Polynomial. The value shown is the average accuracy of all parameter combinations for each kernel and threshold value. Based on these results, the RBF kernel provides the best performance with the highest accuracy of 88.42% at a threshold of 0.0005. In contrast, the Linear kernel at a threshold of 0.04 only achieved the highest accuracy of 78.84%, which is the lowest accuracy in this test. This shows that the selection of the kernel and the number of features based on the threshold greatly affects the performance of the model. It can be seen in Table 6.

Table 6. Overall Average Results using Smote and Information Gain

Kernel	Threshold	Accuracy	Precision	Recall	F1-Score
Linear	0,04	78.84%	79.38%	78.84%	78.75%
RBF		80.38%	81.51%	80.38%	80.21%
Polynomial		78.89%	79.71%	78.89%	78.74%
Linear	0,01	79.16%	79.79%	79.16%	79.06%
RBF		84.12%	84.75%	84.12%	84.05%
Polynomial		80.18%	80.84%	80.18%	84.07%
Linear	0,0005	79.40%	80.05%	79.40%	79.29%
RBF		88.42%	88.93%	88.42%	88.38%
Polynomial		80.91%	81.54%	80.91%	80.82%

Support Vector Machine (SVM) linear kernels very often used when the data is high-dimensional and linearly separable. Kernels do not require mapping to higher dimensions as they do not provide significant performance

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

improvement. In this study, Information Gain feature selection is used to select the most robust features before applying the model, in order to improve classification accuracy and efficiency. The C parameter with values of 1, 4, and 5 is used to control the balance between the decision limit and the classification error rate. Accuracy results after three tests of feature selection and model application.

Table 7. Average Results of Linear Kernel using Smote with Information Gain

Threshold	Cost	Accuracy	Precision	Recall	F1-Score
0,04	1	79.84%	79.38%	78.84%	78.74%
	10	78.85%	79.38%	78.85%	78.72%
	100	78.84%	79.38%	78.84%	78.75%
0,01	1	79.15%	79.79%	79.15%	79.04%
	10	79.17%	79.79%	79.17%	79.06%
	100	79.17%	79.79%	79.17%	79.06%
0,0005	1	79.39%	80.06%	79.39%	79.28%
	10	79.39%	79.05%	79.39%	79.28%
	100	79.40%	80.06%	79.40%	79.30%

The average results of the linear kernel with various Cost parameter values for three tests. Shows the test results on the feature selection process with a threshold of 0.04 which produces the highest accuracy of 79.84%, precision of 79.38%, recall of 78.84%, and F1-Score of 78.74% at a Cost value of 1. Furthermore, testing with a threshold of 0.01 shows the highest accuracy of 79.17%, precision of 79.79%, recall of 79.17%, and F1-Score of 79.06% at Cost values of 10 and 100. Then, with a threshold of 0.0005, the best results were obtained at a Cost value of 100 with an accuracy of 79.40%, precision of 80.06%, recall of 79.40%, and F1-Score of 79.30%. Overall, although there is an increase in performance at each threshold. The linear kernel shows relatively low performance and is less responsive to changes in the Cost parameter value. It can be seen in Table 7.

The RBF kernel in Support Vector Machine (SVM) is a popular function because it is effective on data that is not linearly separable. This kernel has two main parameters, namely Cost (C) and Gamma (γ), where Gamma regulates the range of influence of the training data. Gamma values that are too small make the model less able to capture the complexity of the data, while values that are too large can cause overfitting. Before testing, feature selection was applied using Information Gain with thresholds of 0.04; 0.01; and 0.0005. In this test, various parameters C = 1, 10, and 100 and γ = 1, 4, and 5 were used. The accuracy results of the parameter combinations are shown in Table 8.

Table 8. Average Results of RBF Kernel using Smote and Information Gain

Threshold	Cost	Gamma	Accuracy	Precision	Recall	F1-Score
0,04	1	1	79.04%	81.92%	79.04%	78.72%
		4	80.10%	81.14%	80.10%	79.93%
		5	80.29%	81.24%	80.29%	80.14%
	10	1	79.76%	81.27%	79.76%	79.52%
		4	80.68%	81.57%	80.68%	80.54%
		5	80.74%	81.62%	80.74%	80.61%
	100	1	80.63%	81.67%	80.63%	80.47%
		4	81.07%	82.04%	81.07%	80.93%
		5	81.16%	82.12%	81.16%	81.02%
0,01	1	1	81.67%	82.44%	81.67%	81.56%
		4	82.56%	83.13%	82.56%	82.48%
		5	82.90%	83.45%	82.90%	82.48%
	10	1	82.86%	82.52%	82.86%	82.77%
		4	84.59%	85.21%	84.59%	83.06%
		5	85.11%	85.80%	85.11%	85.04%
	100	1	84.54%	85.05%	84.54%	84.48%
		4	86.23%	86.89%	86.23%	86.17%
		5	86.60%	87.23%	86.60%	86.54%
0,0005	1	1	85.34%	85.99%	85.34%	85.28%
		4	86.79%	87.42%	86.79%	86.74%
		5	87.16%	87.74%	87.16%	87.11%
	10	1	87.62%	88.21%	87.62%	87.58%
		4	89.18%	89.67%	89.18%	89.15%
		5	89.39%	89.84%	89.39%	89.36%
	100	1	89.36%	89.87%	89.36%	89.32%
		4	90.39%	90.79%	90.39%	90.37%
		5	90.51%	90.89%	90.51%	90.49%

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

Based on Table 8, the test results in three tests on the RBF kernel with feature selection using Information Gain show that the highest accuracy is obtained at a threshold of 0.0005 resulting in the highest accuracy of 90.51% with Precision 90.89%, Recall 90.51%, and F1-Score 90.49% (C = 100, Gamma = 5); followed by a threshold of 0.01 with an accuracy of 86.60%, Precision 87.23%, Recall 86.60%, and F1-Score 86.54% (C = 100, Gamma = 5); and a threshold of 0.04 with an accuracy of 81.28%, Precision 82.25%, Recall 81.28%, and F1-Score 81.14% (C = 100, Gamma = 5), which concludes that the smaller the threshold value in feature selection, the higher the performance of the resulting model.

The polynomial kernel in Support Vector Machine (SVM) is an effective kernel function for large, normalized datasets, as it maps data to feature space while maintaining relationships between samples. This kernel is similar to the linear kernel in the way it works, but considers a combination of features in measuring similarity. The polynomial kernel has three main parameters, namely Cost (C), Gamma (γ), and Degree (d), which affect the complexity of data separation. Before testing, feature selection is done using Information Gain with thresholds of 0.04; 0.01; and 0.0005 to get the most relevant features.

Table 9. Average Results of Polynomial Kernel using Smote and Information Gain

Threshold	Cost	degree	coef0	Accuracy	Precision	Recall	F1-Score
0,04	1	1	0	78.36%	78.86%	78.36%	78.27%
			1	78.36%	78.86%	78.36%	78.26%
		2	0	79.70%	80.10%	79.70%	79.64%
			1	78.77%	80.48%	78.77%	78.47%
	10	1	0	78.35%	78.84%	78.35%	78.26%
			1	78.35%	78.85%	78.35%	78.26%
		2	0	80.04%	80.57%	80.04%	79.96%
			1	78.86%	80.81%	78.86%	78.52%
	100	1	0	78.35%	78.84%	78.35%	78.26%
			1	78.36%	78.85%	78.36%	78.27%
		2	0	79.96%	80.53%	79.96%	79.88%
			1	79.18%	80.97%	78.18%	79.87%
0,01	1	1	0	79.19%	79.83%	79.19%	79.08%
			1	79.19%	79.83%	79.19%	79.08%
		2	0	81.00%	81.45%	81.00%	80.94%
			1	80.94%	81.55%	80.94%	80.85%
	10	1	0	79.22%	79.82%	79.22%	79.12%
			1	79.23%	79.83%	79.23%	79.13%
		2	0	81.20%	81.82%	81.20%	81.11%
			1	81.07%	81.89%	81.07%	80.95%
	100	1	0	79.20%	79.80%	79.20%	79.10%
			1	79.22%	79.81%	79.22%	79.12%
		2	0	81.40%	82.30%	81.40%	81.28%
			1	81.24%	82.20%	81.24%	81.10%
0,0005	1	1	0	79.50%	80.22%	79.50%	79.38%
			1	79.50%	80.22%	79.50%	79.38%
		2	0	81.50%	81.90%	81.50%	81.44%
			1	81.47%	81.93%	81.47%	81.40%
	10	1	0	79.51%	80.19%	79.51%	79.40%
			1	79.52%	80.19%	79.52%	79.40%
		2	0	82.41%	82.89%	82.41%	82.35%
			1	82.42%	83.01%	82.42%	82.34%
	100	1	0	79.53%	80.20%	79.53%	79.42%
			1	79.53%	80.19%	79.53%	79.41%
		2	0	83.04%	83.75%	83.04%	82.95%
			1	83.04%	83.80%	83.04%	82.94%

Evaluation results of Support Vector Machine (SVM) model with Polynomial kernel at three different threshold values, namely 0.04; 0.01; and 0.0005. Each threshold displays various model parameters such as Cost, Gamma, degree, and coef0 values, as well as evaluation results based on accuracy, precision, recall, and F1-score. At threshold 0.04, the best performance is obtained when Cost=10, degree=2, and coef0=0, with an accuracy of 80.04% and F1-Score of 79.96%. Then at threshold 0.01 the performance increases at Cost=100, degree=2, and coef0=0, resulting in an accuracy of 81.40% and F1-Score of 81.28%. A more significant improvement is seen at threshold 0.0005, where the Cost=100, degree=2, and coef0=0 configuration produces the highest accuracy of 83.04% and F1- Score of 82.95%. Overall, these results show that decreasing the threshold value (which means more features are used) has a positive impact on model performance, especially in improving accuracy and F1-Score. This can be seen in Table 9.

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 22−33
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.116
https://ejurnal.pdsi.or.id/index.php/bids/index

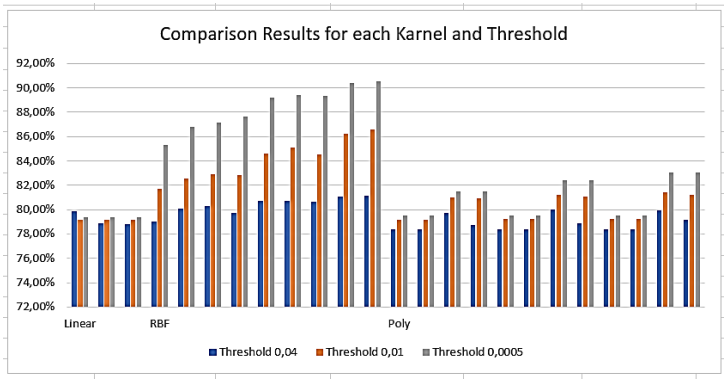

Figure 8. Comparison Results for each Karnel and Threshold

The graph above shows the results of the SVM model accuracy comparison with three types of kernels, namely Linear, RBF, and Polynomial, against three different threshold values, namely 0.04, 0.01, and 0.0005. Based on the graphs, it can be seen that the Linear kernel provides the lowest accuracy and is less affected by changes in the threshold value, with accuracy ranging from 78% to 80%. In contrast, the RBF and Polynomial kernels show a significant increase in accuracy as the threshold decreases. The RBF kernel with a threshold of 0.0005 yields the highest accuracy of around 91%, which indicates that selecting more features (smaller threshold) can improve the overall performance of the model. Thus, it can be concluded that the RBF kernel is superior to Polynomial and Linear, and the use of a lower threshold can improve classification performance.

The study showed that the combination of Information Gain technique in feature selection and SMOTE for data balancing successfully improved the performance of SVM model in predicting stroke disease. Without this technique, the model has difficulty recognizing stroke patients because the amount of data is much smaller. After applying SMOTE, the data distribution becomes balanced, so the model can learn better. Feature selection using Information Gain (0.04, 0.01, and 0.0005) helps the model focus on relevant features. The SVM model with RBF kernel gave the best results, especially at threshold 0.0005 with Cost 100 and Gamma 5, resulting in the highest accuracy of 90.51%. The Polynomial kernel is also quite good (83.04%), while the linear kernel is the most stable but has lower accuracy (79.84%). Thus, proper feature selection, data balancing, and optimal parameter tuning are essential for building an accurate and effective stroke detection model.

3.3 Evaluation

The confusion matrix in the figure below shows two confusion matrices of the Support Vector Machine (SVM) model with RBF kernel (C=100, gamma=5, Fold=6), comparing the performance of the model without feature selection and with feature selection using Information Gain. In the first confusion matrix without Information Gain, the model successfully classified 437 negative data and 465 positive data correctly, and produced 51 false positives and 19 false negatives, while in the second confusion matrix with Information Gain, the model slightly decreased in the classification of negative data with 419 true negatives and increased in the classification of positive data with 473 true positives, and produced 59 false positives and 21 false negatives, which overall shows that the use of Information Gain helps the model in recognizing positive cases (stroke) better although it slightly reduces the accuracy in the negative class. The Confusion matrix results can be seen in Figure 9.

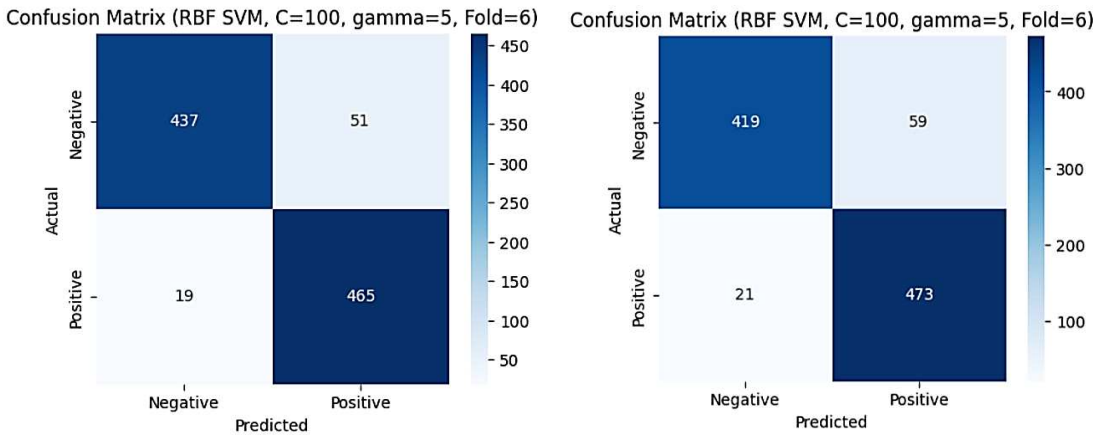

Figure 9. Comparison of confusion matrix results with the highest accuracy in SVM testing (a) without Information Gain and (b) using Information Gain

3.4 Discussion

This study shows that the application of Feature Selection using Information Gain and data balancing with SMOTE successfully improves the performance of the Support Vector Machine (SVM) model in stroke disease classification. In contrast to previous research by [15] which only evaluated various SVM kernels with parameters or data imbalance handling. And was able to achieve a maximum accuracy of 78.86% with the Polynomial kernel. this study explores the model parameters in more depth. The experimental results show that at a threshold of 0.0005 the RBF kernel with parameters Cost = 100 and Gamma = 5 gives the highest accuracy of 90.51%. While the Polynomial kernel produces 83.04% accuracy with parameters Cost=100, Degree=2, and Coef0=0 which has very stable results. Linear kernel has lower results than other kernels with the highest accuracy of 79.84%. Thus, it can be concluded that the use of Feature Selection techniques, SMOTE, and appropriate parameter settings significantly improve the performance of stroke classification models. This approach contributes better than previous studies and shows more optimal results.

4. CONCLUSION

This study aims to evaluate the performance of Information Gain feature selection in the classification process using the Support Vector Machine (SVM) algorithm to detect stroke disease. The method used combines feature selection with Information Gain and data balancing using (SMOTE), which is proven effective in overcoming class imbalance problems and improving model accuracy. Feature selection is performed by applying a threshold of 0.0005, resulting in 18 features that are considered most relevant to the classification target. These features include: avg_glucose_level, bmi, age, work_type_children, ever_married, work_type_Self-employed, smoking_status_formerly smoked, smoking_status_Unknown, work_type_Private, work_type_Govt_job, smoking_status_smokes, hypertension, heart_disease, gender_Male, gender_Female, smoking_status_never smoked, work_type_Never_worked, and Residence_type. On the other hand, the gender_Other feature was not used because it did not pass the selection threshold. The results showed that the RBF kernel with a combination of Cost = 100 and Gamma = 5 parameters gave the best performance with the highest accuracy of 90.51%. The Polynomial kernel with parameters Cost = 100, Degree = 2, and Coef0 = 0 produced an accuracy of 83.04%, while the Linear kernel showed the lowest accuracy of 79.84%, although stable. Overall, the application of feature selection, data balancing, and appropriate parameter settings were proven to determine the variables that most influence SVM classification, thus improving the model's ability to detect stroke disease. As a direction for further development, it is recommended to use other feature selection methods such as Gain Ratio, Recursive Feature Elimination (RFE), and Lasso Regression to compare performance and find a more optimal approach. In addition, the use of larger and more diverse datasets is expected to improve the generalizability and reliability of the model. Further research could also include comparing SVM with other algorithms such as Random Forest, XGBoost, or deep learning approaches to broaden insights in the development of stroke classification systems. These findings have great potential to be applied in technology-based early detection systems to support the stroke diagnosis process quickly, accurately, and efficiently.

REFERENCES

[1] K. Wirastuti, N. S. Riasari, D. Djannah, dan M. Silviana, “Upaya Pencegahan Stroke melalui Skrining Skor Risiko Stroke dengan Intervensi Penyuluhan dan Pemeriksaan Faktor Risiko Stroke di Kelurahan Bojong Salaman Kecamatan Pusponjolo Selatan Semarang Barat,” *J. ABDIMAS-KU J. Pengabdi. Masy. Kedokt.*, vol. 2, no. 1, hal. 23–29, 2023, doi: 10.30659/abdimasku.2.1.23-29.

[2] D. Kuriakose dan Z. Xiao, “Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives,” *Int. J. Mol. Sci.*, vol. 21, no. 20, hal. 1–24, 2020, doi: 10.3390/ijms21207609.

[3] D. Majumder, “Ischemic Stroke: Pathophysiology and Evolving Treatment Approaches,” *Neurosci. Insights*, vol. 19, hal. 1–8, 2024, doi: 10.1177/26331055241292600.

[4] M. N. Aziz dan A. Supriyadi, “Pengaruh Proprioceptive Neuromuscular Facilitation Techniques Terhadap Penurunan Spastisitas Otot Pasien Stroke: a Critical Review,” *Universitas Muhammadiyah Surakarta*, 2021, [Daring]. Tersedia pada: http://eprints.ums.ac.id/id/eprint/91145.

[5] L. Wang *et al.*, “Remote ischemic conditioning enhances oxygen supply to ischemic brain tissue in a mouse model of stroke: Role of elevated 2,3-biphosphoglycerate in erythrocytes,” *J. Cereb. Blood Flow Metab.*, vol. 41, no. 6, hal. 1277–1290, 2021, doi: 10.1177/0271678X20952264.

[6] Setiawan *et al*, “Diagnosis Dan Tatalaksana Stroke Hemoragik,” *J. Med. Hutama*, vol. 03, no. 01, hal. 402–406, 2021.

[7] T. G. Rahayu, “The Analysis of Stroke Risk Factors and Stroke Types,” *Faletehan Heal. J.*, vol. 10, no. 01, hal. 48–53, 2023, doi: 10.33746/fhj.v10i01.410.

[8] Y. A. Utama dan S. S. Nainggolan, “Faktor Resiko yang Mempengaruhi Kejadian Stroke: Sebuah Tinjauan Sistematis,” *J. Ilm. Univ. Batanghari Jambi*, vol. 22, no. 1, hal. 549–553, 2022, doi: 10.33087/jiubj.v22i1.1950.

[9] Astannudinsyah, Rusmegawati, dan C. K. Negara, “Hubungan Kadar Kolesterol Darah dan Hipertensi dengan Kejadian Stroke di RSUD Ulin Banjarmansin Tahun 2020,” *J. Medika Karya Ilmiah Kesehatan*, vol. 5, no. 2, 2020, doi: 10.35728/jmkik.v5i2.129.

[10] Harmawati, Etriyanti, dan S. Hardini, “Deteksi Dini Gejala Awal Stroke,” *J. Abdimas Saintika*, vol. 3, no. 2, hal. 186–189, 2021, doi: 10.30633/jas.v3i2.1253.

[11] N. Chafid *et al.*, "Kecerdasan Buatan," Batam: *Yayasan Cendikia Mulia Mandiri*, Juli 2024.

[12] M. M. Ahsan, S. A. Luna, dan Z. Siddique, “Machine-Learning-Based Disease Diagnosis : A Comprehensive Review,” *Healthcare*, vol. 10, no. 3, hal. 1–30, 2022, doi: 10.3390/healthcare10030541.

[13] M. W. Sanjaya, "*Fisika Komputasi Berbasis Machine Learning Dengan Pemrograman Python,"*. Bandung: BolaBot, 2024.
[14] R. Guido, S. Ferrisi, D. Lofaro, dan D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information*, vol. 15, no. 4, hal. 1–36, 2024, doi: 10.3390/info15040235.
[15] S. Rahayu dan Y. Yamasari, "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)," *J. Informatics Comput. Sci.*, vol. 5, no. 03, hal. 440–446, 2024, doi: 10.26740/jinacs.v5n03.p440-446.
[16] A. Nazri dan R. A. Panbudi, "Implementasi Algoritma SVM dalam Memprediksi Penyakit Stroke," *Journal Zetroem,* vol. 06, no. 02, hal. 4–8, 2024, doi: 10.36526/ztr.v6i2.3676.
[17] L. Pasiolo *et al.*, "Penerapan Teknik SMOTE pada Klasifikasi Penyakit Stroke dengan Algoritma Support Vector Machine," *ZONAsi: Jurnal Sistem Informasi,* vol. 7, no. 1, hal. 61–74, 2025.
[18] D. Ispriyanti, P. A. Octaviani, dan Y. Wilandari, "Penerapan Metode SVM Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang," *J. Gaussian*, vol. 3, no. 4, hal. 811–820, 2014, doi: 10.14710/j.gauss.3.4.811-820.
[19] I. Santoso, Windu Gata, dan Atik Budi Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, hal. 364–370, 2019, doi: 10.29207/resti.v3i3.1084.
[20] N. A. Amri *et al.*, *Image Processing*. Yogyakarta: PT. Green Pustaka Indonesia, 2025.
[21] A. Angela Sitinjak *et al.*, *Matematika Pada Kecerdasan Buatan*. Makasar: CV. Tohar Media, 2024.
[22] A. W. Attabi, Lailil Muflikhah, dan Mochammad Ali Fauzi, "Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, hal. 4548–4554, 2018.
[23] A. Kulsumarwati, I. Purnamasari, dan B. A. Darmawan, "Penerapan SVM dan Information Gain Pada Analisis Sentimen Pelaksanaan Pilkada Saat Pandemi," *J. Teknol. Inform. dan Komput.*, vol. 7, no. 2, hal. 101–109, 2021, doi: 10.37012/jtik.v7i2.641.
[24] D. Apriliani dan O. Somantri, "Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, hal. 537, 2018, doi: 10.25126/jtiik.201855867.
[25] B. Aribowo dan S. Fairuz, *Panduan Praktis Machine Learning Klasifikasi Menggunakan Python*. Yogyakarta: Dandra, 2024.
[26] D. W. Lestari, D. A. Lusia, M. Y. Rochayani, dan U. Sa'adah, *Kupas Tuntas Algoritma Data Mining Dan Implementasinya Menggunakan R*. Malang: UB Press, 2021.
[27] A. Apriyanto *et al.*, *DATA MINING (Teori dan Penerapannya dalam Berbagai Bidang)*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.
[28] M. S. Amrullah dan S. F. Pane, *Analisis Sentiment Masyarakat Terhadap Kebijakan Polisi Tilang Manual Di Indonesia Dengan Svm (Support Vector Machine)*. Bandung Barat: Buku Pedia, 2023.
[29] S. Agustin *et al.*, *No Title*. Batam: Yayasan Cendikia Mulia Mandiri, 2025.