Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 22, No. 3, July 2023, pp. 539∼552
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v22i3.xxx

Comparison of Support Vector Machine Performance with Oversampling and Outlier Handling in Diabetic Disease Detection Classification

Firda Yunita Sari , Maharani Sukma Kuntari , Winda Ari Yati , Hani Khaulasari
Universitas Islam Negeri Sunan Ampel, Surabaya, Indonesia

Article Info

Article history:

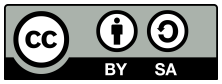
Received May 19, 2023
Revised June 08, 2023
Accepted July 02, 2023

Keywords:

Accuracy
Diabetes Mellitus
Support Vector Machine
Synthetic Minority Over-Sampling Technique.

ABSTRACT

Diabetes mellitus is a disease that attacks chronic metabolism, characterized by the body's inability to process carbohydrates, fats so that glucose levels are high. Diabetes mellitus is the sixth cause of death in the world. Classifying data about diabetes mellitus makes it easier to predict the disease. As technology develops, diabetes mellitus can be detected using machine learning methods. The method that can be done is the support vector machine. The advantage of SVM is that it is very effective in completing classification, so it can quickly separate each positive and negative point. This study aimed to obtain the best SVM classification model based on accuracy, sensitivity, and precision values in detecting diabetes by adding Synthetic Minority Over-Sampling Technique (SMOTE) and handling outliers. The SMOTE method was applied to handle class imbalance. The Support Vector Machine (SVM) method aimed to produce a function as a dividing line or what can be called a hyperplane that matches all input data with the smallest possible error. The data studied were indications of diabetes, consisting of 8-factor variables and 1 class variable. The test results show that the SVM-SMOTE scenario produces the best accuracy. The SVM SMOTE scenario produced an accuracy value of the RBF kernel of 88% with an error of 12%, and this is obtained from the division of test data and training data of 90:10. This SVM-SMOTE scenario produced a precision value of 0.880 and a sensitivity value of 0.880. The research results showed that factor classification was more accurate if it is carried out using the support vector machine (SVM) method with imbalance data handling (SMOTE), and it can be concluded that the distribution of test data and training data influences a test scenario.



Corresponding Author:

Hani Khaulasari,
Department of Science and Technology,
Universitas Islam Negeri Sunan Ampel, Surabaya, Indonesia,
Email: hani.khaulasari@uinsby.ac.id

How to Cite:
F. Yunita Sari, M. Kuntari, H. Khaulasari, and W. Ari Yati, "Comparison of Support Vector Machine Performance with Oversampling and Outlier Handling in Diabetic Disease Detection Classification", MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 3, pp. 539-552, Jul. 2023.

1. INTRODUCTION

Diabetes Mellitus (DM) is the highest cause of death in Indonesia. In other words, this disease is a health problem that is quite risky. Sourced from a global study providing a fact in 2011, 366 million people became patients with Diabetes Mellitus. Diabetes can occur when the body's condition is lacking or ineffective insulin is produced [1]. Diabetes mellitus (DM) is a group of metabolic diseases accompanied by several signs, including hyperglycemia, due to insulin action and secretion effects. Meanwhile, the pancreas produces a hormone that transports glucose from food into the body's cells, commonly known as insulin. Then the glucose is converted into energy to drive muscle and organ functions. In people with diabetes who are unable to absorb glucose properly, glucose will increase in the blood (hyperglycemia) and cause damage to organs and tissues over time. According to the World Health Organization (WHO), cases of diabetes mellitus in the world have increased every year by 171 million diabetes mellitus patients in 2000, and there will be an increase of 114% or 336 million people in the coming 2030. The number of diabetes mellitus is increasing every year, especially in pregnant women. Diabetes Mellitus is one of three problems that cause complications during pregnancy [2]. The latest data released by the International Diabetes Federation (IDF) in 2021 records 537 million adults aged 20-79 years worldwide as people with diabetes. An indicated number of 3 out of 4 adults with diabetes are in low to middle-income countries. In 2021 as many as 6.7 million people will die as diabetics [3]. Diabetes Mellitus (DM) is a disease that attacks chronic metabolism with a sign that the body cannot process carbohydrates, proteins, and fats, so glucose levels are high [4]. Diabetes Mellitus occurs not only due to genetic factors but is also caused by a person's lifestyle and living habits [5]. Unfavorable lifestyles, such as lack of physical activity and excessive consumption of sweet foods, result in obesity or overweight. One of the causes of diabetes is an inconsistent diet so that the blood sugar levels that enter cannot be absorbed by the body [6]. Diabetes mellitus, which has attacked the human body, is very dangerous because it can cause complications in the human body that can lead to death [7]. Diabetes can be diagnosed by checking blood sugar, insulin, BMI, and blood pressure levels in the patient's body [8].

Along with the development of the times, diabetes mellitus can be detected by taking blood samples from patients and then classifying the data they already have with the help of machine learning methods. One of the machine learning methods that can be used to detect diabetes is a support vector machine [9]. SVM is a method that separates two classes using a linear function so that a dividing line (hyperplane) is found between classes [10]. The advantage of SVM is that it is very effective in completing classification, so it can quickly separate each positive and negative point [11]. Classification label data is often found in imbalanced data conditions or unbalanced data in each class, causing the classification accuracy results to be much higher for the majority class than for the minority class [12]. One solution for overcoming data imbalance cases is the Synthetic Minority Over-sampling Technique (SMOTE) [13]. Handling between classes is done to avoid misclassification results which can lead to errors in patient handling because they depend on the majority class [14]. The process of detecting diabetes is still using invasive techniques (injuries) using a medical device called a glucometer. Examination of diabetes using a glucometer has drawbacks, namely the limited measurement of interval analysis. Temperature greatly affects the accuracy of results, and the price is more expensive than other methods. So from this deficiency, the authors developed the SVM method to classify patients with confirmed diabetes.

Research has been done on predictions of diabetes based on Logistic Regression and SVM. The results obtained are SVM accuracy of 78.6% which is higher than the Logistic Regression method of 78.1% [15]. Previous research on Diabetes Prediction using the Support Vector Machine (SVM) method has an accuracy of 98% [16]. The previous research using the SMOTE method aims to balance classes in the diabetes classification using C4.5, random forest, and SVM with SMOTE oversampling, which achieves good performance for handling unbalanced data. Research on modeling SVM imbalance data for classifying study success of IPB master students has concluded that handling unbalanced data using the Synthetic Minority Oversampling Technique (SMOTE) method successfully increases SVM performance in classifying students who fail [17]. Previous research on detecting COVID-19 used the outlier handling method with SMOTE SVM as a filter to increase the prediction of someone being exposed to the Covid-19 virus. This study explained that unbalanced data sets greatly affected classification performance and resulted in the conclusion that outlier handling and SMOTE worked better [18]. There are several differences between previous studies and our research, namely predictor variables and response variables. In the previous study, 39 predictor variables were used to detect COVID-19, while this study used eight predictor variables to detect diabetes in a patient [18]. The novelty of this research aims to compare how much accuracy is obtained in the diabetes dataset using the SVM method with or without outlier treatment and with or without the SMOTE method because SVM is one of the methods that has strong and optimal accuracy results. This study aims to obtain the best SVM classification model based on accuracy, sensitivity, and precision values in detecting diabetes by adding Synthetic Minority Over-Sampling Technique (SMOTE) and handling outliers.

2. RESEARCH METHOD

This study uses the Data Mining method, which results from observational analysis using a data set that aims to determine a correlation (relationship) and narrow down the data using different methods depending on the characteristics possessed by the data [19]. This study uses a Support Vector Machine or what can be called Support Vector Classification [20]. SVM aims to produce a function as a dividing line or what can be called a hyperplane which corresponds to all input data with the smallest possible error [21]. This study uses data on the classification of diabetes from Kaggle. The data consisted of response variables in the form of diabetes status labeled (+ and -) and predictor variables consisting of pregnancy, glucose, blood pressure, skin thickness, insulin, pedigree diabetes, and age. This data has 768 rows with eight columns in which each column contains factors causing indications of diabetes which will be examined to find out the results of classification using the SVM method [22].

2.1. Flowchart

This research was conducted to show the best classification model based on accuracy, sensitivity, and precision values. Results from this study were processed using the support vector machine method, which was optimized using the Synthetic Minority Over-Sampling Technique. Figure 1 shows the research stages.

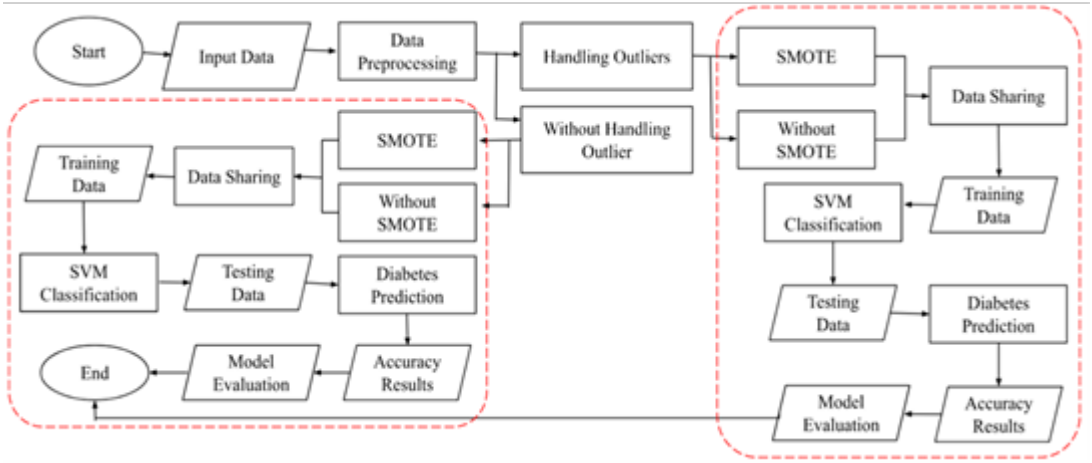


Figure 1. Stages of research using a Support Vector Machine with the SMOTE and Outlier Handling

This study's steps are as follows:

1. Input data for classification of factors indicating diabetes
2. Data description
3. Preprocessing Data
   a. Missing Value detection
   b. Outlier detection
   c. Multikolinieritas detection
4. Then the data is divided, training data with data testing with a ratio of 90:10, 80:20, 70:30
5. Klasifikasi SVM dengan 4 skenario :
   (i) SVM original data The SVM algorithm is as follows [23]
   Input: Input data (X), target data (Y), kernel parameters (σ), penalty (C)
   Output: a,b accuracy
   Begin:
   a. Dividing the class into binary groups $\frac{v(v-1)}{2}$
      For $k = 1 : v, 1 = k + 1 : v$
   b. Defines the RBF kernel function parameters (K)
   c. Determining the minimum point with the Equation $min_{w,\xi}\frac{1}{2}||w||^2 + \frac{1}{2}C\Sigma_{i=1}^{n}\xi_i^2$

d. Determine the parameter penalty (C)
e. Minimizes the primal langrange function
f. Changing the form to a linear system instead of Quadratic Programming
g. Calculate the value (a,b) with Karush Kuhn Tucker (KTT) Optimization
h. Creating the hyperplane equation for the construct : $\widehat{f^{kl}}(x) = sign(\widehat{w^{klT}}x) + b^{kl} = 0$

(ii) SVM SMOTE

The SMOTE algorithm is as follows [23]

Input: Number of class minority data (T); The number of data for the majority of classes (P); Number of SMOTE replications (N); Number of nearest neighbors (KNN)

Output: Synthetic data x_{syn}

Begin:

a. Determine the amount of data from the minor class (T), and it is said to be a minor class if the percentage of the total class data is less than 50%
b. Determines the amount of data from the major class; there is only one major data class
c. Calculates the k-nearest neighbor or Euclidean distance using formula 1.
for x =1:n
for z = 1:n

$$d(x,z) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \ldots + (x_p - z_p)^2} \quad (1)$$

end :

d. Determine replication for minority data $N = \dfrac{Total\ major\ class\ data(P)}{Total\ minor\ class\ data(T)}$
e. Specifies the data to be replicated in the minor class x_i
f. Determines the data with the shortest distance from the data to be replicated in the same minor class (x_{knn})
g. Determine random value γ (γ a random number between values [0,1])
h. Calculating the synthesis using the formula 2.

$$x_{syn} = x_i + (x_{knn} - x_i)\gamma \quad (2)$$

(iii) Outlier-SVM
(iv) Outlier- SMOTE-SVM

6. Model evaluation with the criteria of accuracy, precision, sensitivity
7. Conclusion

Figure 1 shows the flow of research stages from start to finish. The first stage is to enter data on the classification of diabetes indication factors. The data preprocessing is done to detect missing values, outliers, and multicollinearity. Dividing data by comparison of training and testing 90:10, 80:20,70:30. SVM classification is carried out with four scenarios, namely initial data pure SVM, SMOTE and SVM, SVM and outlier handling, SMOTE SVM, and outlier handling. Finally, evaluate the model with accuracy, precision, and sensitivity criteria.

2.2. Data

The data used in this research is secondary data, which comes from Kaggle. The data used in this study are things that can cause someone to be confirmed to have diabetes. There are eight predictor variables: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The predictor variable is labeled 'x,' while the response variable is denoted by the letter 'y'. The data is presented in tabular form in Table 1.

Table 1. Diabetes indication data

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
x_2	x_2	x_2	x_4	x_4	x_8	x_7	x_2	y
6	148	72	35	0	33,6	0,627	50	Positive
1	85	66	29	0	26,6	0,351	31	Negative
8	183	64	0	0	23,3	0,672	32	Positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	93	70	31	0	30,4	0,315	23	Negative

2.3. Data Preprocessing

Preprocessing is an important sequence in any processing system. Data analysis will be hindered, and work is not easy if the data is very large, so it can be solved with data preprocessing. Data preprocessing is done by preparing the data by cleaning the data from noise or changing the data format. This is necessary because raw data is often found to be incomplete and has a format that changes frequently. Preprocessing itself is divided into data validation and imputation. Validation was carried out to assess the level of completeness and accuracy of the filtered data. On the other hand, imputation aims to minimize the error rate and manually enter missing values or automatically through a business process automation (BPA) program [24]. The missing value is the void of some existing data in the data. Listwise Deletion is the most suitable method for dealing with missing values. If empty data is found, it will be removed from the analysis [25]. An outlier is a data object with an abnormal value, either too low or too high, so the difference is large with other objects [26]. Multicollinearity describes a perfect or definite linear relationship between some or all of the independent variables [27]. The formula for determining the Pearson correlation value using formula (3). Where $r_{(XY)}$ represents the correlation coefficient, $X_{(i)}$ represents the variable x to i, $Y_{(i)}$ represents the variable y to i.

$$r_{XY} = \frac{n\Sigma_{i=1}^{n} X_i Y_i - \Sigma_{i=1}^{n} X_i \Sigma_{i=1}^{n} Y_i}{\sqrt{n\Sigma i = 1^{n} X_I^2 - (\Sigma_{i=1}^{n} X_i)^2}\sqrt{n\Sigma_{i=1}^{n} Y_I^2 - (\Sigma_{i=1}^{n} Y_i)^2}} \tag{3}$$

2.4. Support Vector Machine

SVM stands for Support Vector Machine. SVM uses the process of finding the optimal dividing line (hyperplane) to find the maximum margin size between inputs, using a linear function in a high-dimensional feature space that works by separating two groups of data classes using space and feature space using kernel rules [28]. SVM is a derived model of statistical learning theory with better results than other methods. In SVM, each training data is known as (x_i, y_i), where $i = 1, 2, \ldots, N$ dengan $x_i = x_{i1}, x_i 2, \ldots, x_{iq}$ T is an attribute for training data I, $y_i - 1, +1$ is the label class [29]. Figure 2 shows that various alternative separators can separate all datasets according to their class, but the best separators can separate data and have the largest margins. The SVM model equation is as in formula 4 [30]. where w represents the weight of SVM and b is a scalar. Several types of kernels will form a hyperplane to turn and produce the best accuracy, represented in Figure 3. The formula for solving the problem linearly and nonlinearly in SVM is shown in Table 2.

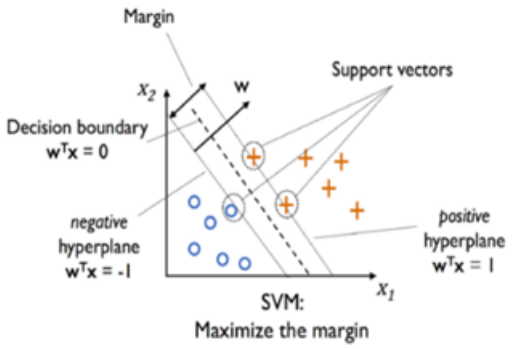


Figure 2. Support vector machine

$$x.w + b = 0 \tag{4}$$

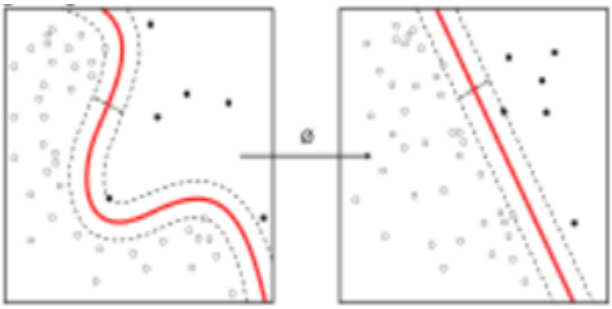


Figure 3. Hyperplane on linear and nonlinear problems

Table 2. Kernel Formulas

Kernel Type	Formula
Linear	$K(x.x') = (x.x')$
Polynomial	$K(x.x') = (x.x')$
RBF Gaussian	$K(x.x') = exp(-\gamma x - x'2)$
Sigmoid	$K(x.x') = tanh(x.x' + \beta)$

2.5. Synthetic Minority Over-Sampling Technique

The Synthetic Minority Over-Sampling Technique (SMOTE) method is generally implemented to overcome the class imbalance. This technique balances the dataset by resampling minority class samples by adding new data from minority classes.

2.6. Evaluation of Classification Results

Confusion Matrix (Confusion Matrix) is a matrix used to indicate the level of accuracy of classified images relative to the reference data. A classification model can be said to be good if it gets a relatively small error rate. The confusion matrix table contains True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values, as in Table 3. The calculation of the confusion matrix will produce a sensitivity value, specificity value, and accuracy value. The specificity value is the value resulting from the classification of the class of diabetes that really belongs to that class of diabetes. The specificity value is the value resulting from the classification of a class of diabetes and is a class of diabetes. At the same time, the success value of a system that runs the classification is the accuracy value [31].

Table 3. Confusion Matrix

Diabetes Classification Results	Actual Diabetes Data	
	Yes	No
Yes	TP	FP
No	FN	TN

Where True Positive (TP), namely positive information successfully identified correctly, enters the positive class. True Negative (TN) is a negative statement correctly identified and entered into the negative class. False Positive (FP), namely negative information, but identified as wrong by the system and entered into the positive class. False Negative (FN), namely a positive statement identified as wrong by the system and entered into the negative class [32].

Accuracy is a value that describes the accuracy of the classification results on positive or negative data. The higher the accuracy value obtained, it means that the system has succeeded in classifying properly. The formula can know the accuracy value in the binary class (two classes) as in Equation (5). The precision value represents the number of correctly identified positive data points divided by the total number of positive data points. Precision can be known by equation (6). Recall is a value that describes how much data

with positive information can be correctly identified as being in a positive class. A higher sensitivity value means the classification system is better at detecting objects. The sensitivity value can be found using Equation (7). F1-score is an evaluation metric used in classification to measure the balance between precision and recall of a model. F1-score provides an overall picture of the quality of the model's performance in predicting the positive class. The formula of the F1-score is as in Equation (8). Error or error is a problem that identifies errors in a number of data so you can see the error level in the system used. The percentage of error can be found using Equation (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - score = 2 \times \frac{Precicion \times Recall}{Precicion + Recall} \quad (8)$$

$$Error = FP \times TP \times 100\% \quad (9)$$

3. RESULT AND ANALYSIS

All variables tested have a total of 768 rows. Then it detects missing values in nine variables, and there are no missing values in Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree Diabetes, Age, or group variables. Figure 5 shows the boxplot results with dots at the top and bottom, indicating that this variable has outlier data. The data has 218 outlier data. After handling the outlier data by deleting data containing outliers, there are 550 diabetes data without outlier data.

3.1. Descriptive Statistics

Based on Table 4, all the variables tested totaled 768 rows. The pregnancy variable describes the number of pregnancies in women; the maximum value is 17. According to medical science, a pregnant person has a condition where the body is unable to produce insulin during pregnancy, so a mother who has had multiple pregnancies will be at a higher risk of developing diabetes. The glucose variable referred to in the table is the result of measurements using an oral glucose tolerance test within 2 hours, with a maximum value of 199. According to medical science, it is classified as high because normal blood sugar is 70-130 mg/dl. Blood pressure in the data has a maximum value of 122. From a scientific point of view, the normal human blood pressure is 90/60 mmhg to 120/80 mmhg. Skin Thickness is the thickness of the skin multiples of the triceps and has an average value of 20.5. The insulin variable is the number of serum insulin in 2 hours, with a maximum value of 846. Variable BMI has an average of 31.99. If reviewed according to medical science, the BMI figure of 31.99 is classified as obese and can be at risk of diabetes. The Pedigree variable is the number of hereditary history of diabetes, and the average is 0.4. Then the age variable with an average value of 33.

Table 4. Statistical Descriptive

	Pregnancy	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Pedigree Diabetes	Age
Count	768	768	768	768	768	768	768	768
Average	3,845	120,894	69,105	20,536	79,799	31,992	0,471	33,240
Std. deviation	3,369	31,972	19,355	15,952	115,244	7,884	0,331	11,760
Minimum	0,000	0,000	0,000	0,000	0,000	0,000	0,078	21,000
25%	1,000	99,000	62,000	0,000	0,000	27,300	0,243	24,000
50%	3,000	117,000	72,000	23,000	30,500	32,000	0,372	29,000
75%	6,000	140,250	80,000	32,000	127,250	36,600	0,626	41,000
Maximum	17,000	199,000	122,000	99,000	846,000	67,100	2,420	81,000

Based on Figure 4 illustrates the histogram of 8 predictor variables causing diabetes. Histogram of pregnancy with positive skewness and leptokurtosis, glucose histogram with negative skewness and leptokurtosis, blood pressure histogram with zero skewness and leptokurtosis, inulin histogram with positive skewness and leptokurtosis, BMI histogram with normal skewness and leptokurtosis, pedigree histogram of diabetes with positive skewness and leptokurtosis, age histogram with positive skewness and

leptokurtosis. From several histograms that have been visualized, there is a skewness worth 0 and kurtosis worth 3, so it can be concluded that the eight variables that cause diabetes above have uneven data.

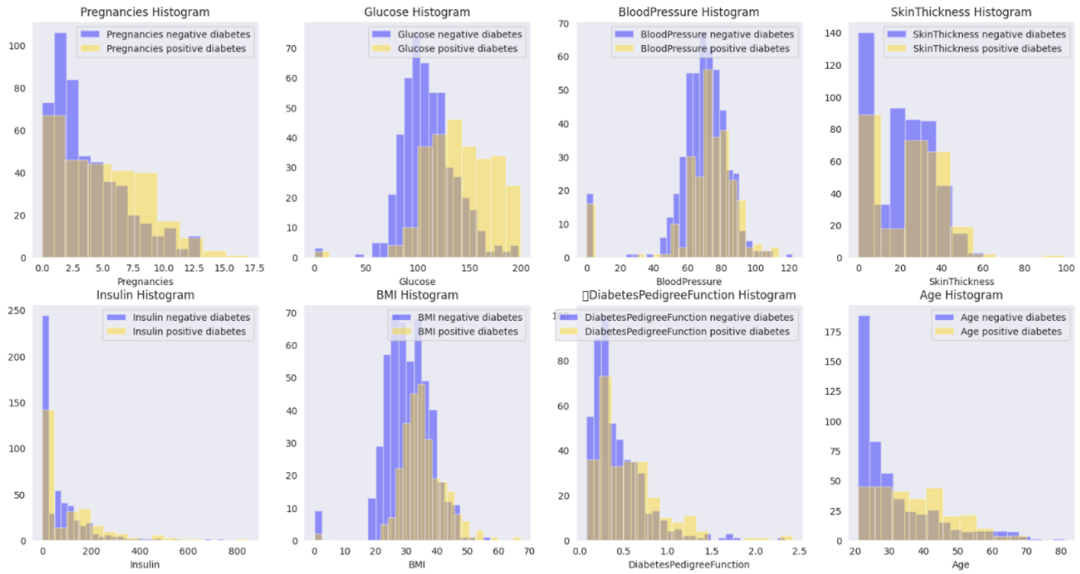


Figure 4. Visualize histogram data

3.2. Preprocessing Data

Missing Value or empty data is the loss of some data that has been obtained. After detecting empty data on nine variables, it can be seen that there is no empty data on the variables of Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree Diabetes, Age, or class variables. So it is true that the data consists of 768 rows with no empty data. Based on Figure 5, it can be seen that in the boxplot, there are dots at the top and bottom, which indicate that the variable has outlier data. The data has as many as 218 outlier data, and this number follows previous research with the same data source. After handling outlier data by deleting data containing outliers, the diabetes data totals 550 without outlier data. Table 5 shows that several variables have the highest correlation of 0.544, namely the age variable with pregnancy, or there is no correlation between variables that are 0.650. So it can be concluded that all variables in the diabetes indication data have a relationship or correlation in the absence of multicollinearity data.

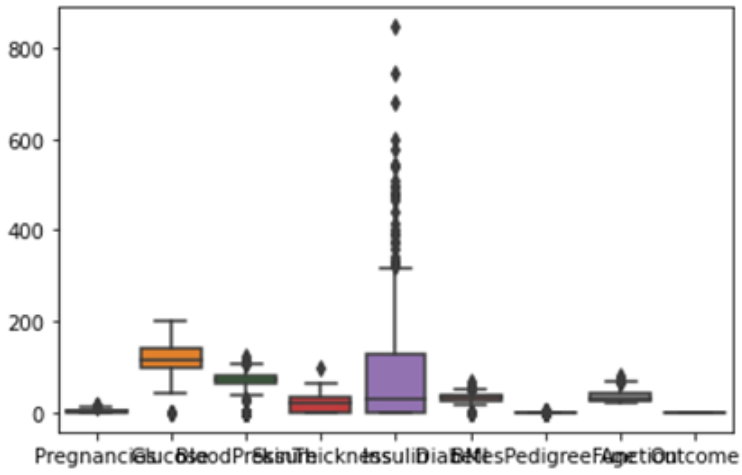


Figure 5. Boxplot outlier data

Table 5. Correlation Value Between Variables

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Pedigree Diabetes	Age	Outcome
Pregnancies	1,000	0,129	0,141	-0,082	-0,074	0,018	-0,034	0,544	0,222
Glucose	0,129	1,000	0,153	0,057	0,331	0,221	0,137	0,264	0,467
Blood Pressure	0,141	0,153	1,000	0,207	0,089	0,282	0,041	0,240	0,065
Skin Thickness	-0,082	0,057	0,207	1,000	0,437	0,393	0,184	-0,114	0,075
Insulin	-0,074	0,331	0,089	0,437	1,000	0,198	0,185	-0,042	0,131
BMI	0,018	0,221	0,282	0,393	0,198	1,000	0,141	0,036	0,293
Pedigree Diabetes	-0,034	0,137	0,041	0,184	0,185	0,141	1,000	0,034	0,174
Age	0,544	0,264	0,240	-0,114	-0,042	0,036	0,034	1,000	0,238
Outcome	0,222	0,467	0,065	0,075	0,131	0,293	0,174	0,238	1,000

3.3. SVM classification

Based on Figure 6, it can be seen that in Figure (i), variable 0 or negative has a percentage of 65.1% or amounts to 500 data. In comparison, variable 1 or positive has a percentage of 34.9% or amounts to 268 data in the diabetes indication dataset, so it can be concluded that the data predominantly indicates a negative case of diabetes. In Figure (ii), after balancing data on variable 0 or negative and variable 1 or positive has a percentage of 50% or amounts to 500 data, the data has 1000 data after SMOTE. In Figure (iii), after handling the outlier data, variable 0 or negative has a percentage of 64.7% or a total of 356 data. In comparison, variable 1 or positive has a percentage of 35.3% or a total of 194 data on the dataset of indications of diabetes. In Figure (iv), after handling the outlier data and balancing with SMOTE, variable 0 or negative has a percentage of 50% or a total of 356 data. In comparison, variable 1 or positive has a percentage of 50% or a total of 356 data on the dataset of indications of diabetes. Our research results follow several previous studies using the same data.

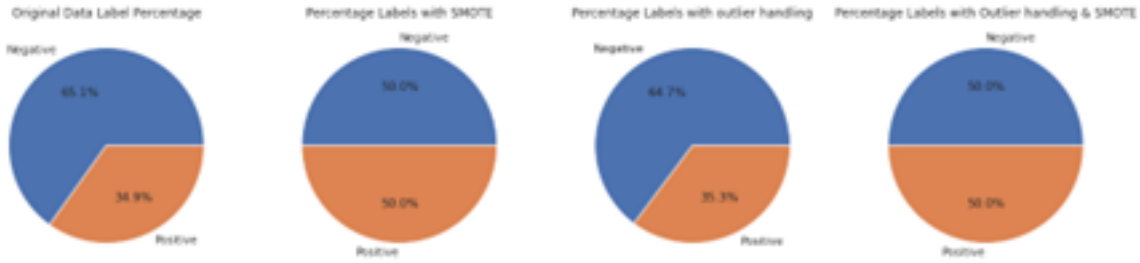


Figure 6. Pie Chart Percentage Labels

3.4. Accuracy Value of SVM

This research uses a separation strategy by dividing the dataset into training and testing data with a ratio of 90:10, 80:20, and 70:30. This is because the three divisions are optimal data-sharing strategies in implementing the support vector machine method and using four scenarios, namely SVM without outlier handling and SMOTE, SVM no outlier handling and with SMOTE, SVM outlier handling, and no SMOTE, and SVM outlier handling and no SMOTE. Based on Table 6, the test results show the accuracy or the best model for diabetes data in scenario B, namely SVM without outliers with SMOTE in the 90:10 data distribution (Kernel RBF), which has an accuracy value of 0.880 or 88% with an evaluation of results. With the error of 0.120 or 12%, it can be concluded that the accuracy that has been obtained is accurate.

Table 6. Accuracy and Error Details of SVM and SVM Scenarios with SMOTE

Ratio	Kernel Type	Scenarios	Accuracy	Error
70:30	Linear	SVM without outlier handling and SMOTE	0,784	0,216
		SVM has no outlier handling and with SMOTE	0,787	0,213
		SVM outlier handling and no SMOTE	0,788	0,212
		SVM outlier handling and no SMOTE	0,724	0,276
	RBF	SVM without outlier handling and SMOTE	0,758	0,242
		SVM has no outlier handling and with SMOTE	0,833	0,167
		SVM outlier handling and no SMOTE	0,764	0,236
		SVM outlier handling and no SMOTE	0,762	0,238
	Polynomial	SVM without outlier handling and SMOTE	0,745	0,255
		SVM has no outlier handling and with SMOTE	0,793	0,207
		SVM outlier handling and no SMOTE	0,776	0,224
		SVM outlier handling and no SMOTE	0,762	0,238
	Sigmoid	SVM without outlier handling and SMOTE	0,719	0,281
		SVM has no outlier handling and with SMOTE	0,700	0,300
		SVM outlier handling and no SMOTE	0,745	0,255
		SVM outlier handling and no SMOTE	0,636	0,364
80:20	Linear	SVM without outlier handling and SMOTE	0,825	0,175
		SVM has no outlier handling and with SMOTE	0,780	0,220
		SVM outlier handling and no SMOTE	0,791	0,209
		SVM outlier handling and no SMOTE	0,692	0,308
	RBF	SVM without outlier handling and SMOTE	0,792	0,208
		SVM has no outlier handling and with SMOTE	0,850	0,150
		SVM outlier handling and no SMOTE	0,764	0,236
		SVM outlier handling and no SMOTE	0,734	0,266
	Polynomial	SVM without outlier handling and SMOTE	0,753	0,247
		SVM has no outlier handling and with SMOTE	0,800	0,200
		SVM outlier handling and no SMOTE	0,782	0,218
		SVM outlier handling and no SMOTE	0,713	0,287
	Sigmoid	SVM without outlier handling and SMOTE	0,766	0,234
		SVM has no outlier handling and with SMOTE	0,766	0,234
		SVM outlier handling and no SMOTE	0,745	0,255
		SVM outlier handling and no SMOTE	0,629	0,371
90:10	Linear	SVM without outlier handling and SMOTE	0,870	0,130
		SVM has no outlier handling and with SMOTE	0,760	0,240
		SVM outlier handling and no SMOTE	0,800	0,200
		SVM outlier handling and no SMOTE	0,667	0,333
	RBF	SVM without outlier handling and SMOTE	0,831	0,169
		SVM has no outlier handling and with SMOTE	0,880*	0,120*
		SVM outlier handling and no SMOTE	0,800	0,200
		SVM outlier handling and no SMOTE	0,694	0,306
	Polynomial	SVM without outlier handling and SMOTE	0,766	0,234
		SVM has no outlier handling and with SMOTE	0,810	0,190
		SVM outlier handling and no SMOTE	0,782	0,218
		SVM outlier handling and no SMOTE	0,611	0,389
	Sigmoid	SVM without outlier handling and SMOTE	0,740	0.260
		SVM has no outlier handling and with SMOTE	0,720	0,280
		SVM outlier handling and no SMOTE	0,545	0,455
		SVM outlier handling and no SMOTE	0,528	0,472

Figure 7 is the best model confusion matrix, namely Kernel RBF with a data division of 90:10. From the SVM scenario without any outliers to SMOTE, and this study shows a significant advantage in the context of diabetes classification. From the given confusion matrix, it appears that the model has a high level of sensitivity (recall) in classifying patients who are actually positive with only 7 cases of wrong positive predictions (false negatives). This shows that the model can accurately identify patients with potential diabetes, which is very important in preventing and controlling this disease. In addition, with only 5 cases of false positive predictions, the model also has a good level of precision in avoiding a wrong diagnosis in patients who do not actually have diabetes. With this combination of high sensitivity and precision, this study demonstrates the superiority of using a classification model to detect diabetes, which can positively impact medical practice and clinical decision-making accurately.

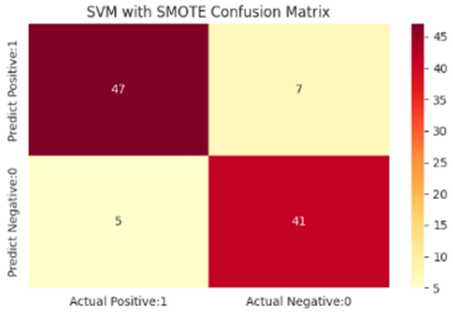


Figure 7. Confusion matrix

Table 7 shows the value of the best evaluation model, namely Kernel RBF, with a data division of 90:10. From the SVM scenario without outliers with SMOTE, the evaluated model shows good performance with high precision, recall, f1-score, and accuracy. The precision is 0,900 for class 0 and 0,850 for class 1, which shows the model's ability to give a slight error in positive predictions. Recall has a value of 0,870 for class 0 and 0,890 for class 1, which shows the model's ability to identify the most positive data. The F1 scores, the harmonic averages of precision and gain, are 0,890 for class 0 and 0,870 for class 1, indicating a good balance between precision and gain. With an accuracy of 0,880, the overall model can classify data accurately. The weighted average and macro average score is 0,880 for precision, recall, and f1 scores, indicating consistent and balanced performance across classes.

Table 7. Model Evaluation Value

	Precision	Recall	f1-score
0	0,900	0,870	0,890
1	0,850	0,890	0,870
accuracy			0,880
Weight Avg	0,880	0,880	0,880
Macro Avg	0,880	0,880	0,880

3.5. Visualization of Prediction Results

Based on Figure 8 shows the predicted results of SVM without Outlier Handling with SMOTE with the RBF 90:10 kernel data division; it is known that the yellow center has information 1, namely positive diabetes, and the red center has information 0, namely negative diabetes, according to the color of the stem. At the same time, the blue line is called a hyperplane or the distinction between the two groups, and the dotted line is called the margin or estimated class difference. A center on the margin is close to the hyperplane because the accuracy is 0.880 with an error value of 0.120. These results experienced a 1% difference from the research [33]. This difference was due to an update in this study by adding the SMOTE feature selection. The best accuracy results were obtained in the second scenario, namely SVM with SMOTE. So it can be concluded that the SMOTE feature selection can improve accuracy so that the accuracy obtained is 88%.

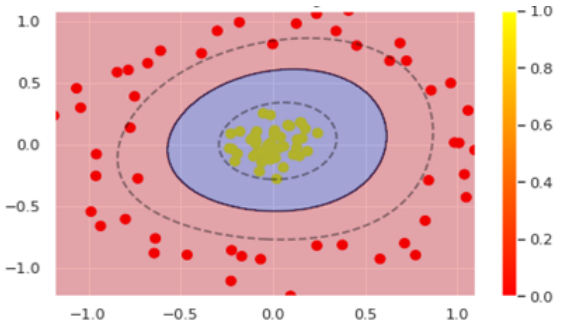


Figure 8. Best Model SVM Prediction Visualization

4. CONCLUSION

This study uses Diabetes data with several influencing variables. In variable 0 or negative, there are 500 data, while in variable 1 or positive, there are 268 data, so it can be concluded that negative cases of diabetes dominate this data. In this study, four tests were carried out to compare the final results of these tests to find the model with the highest accuracy. Of the four scenarios that have been tested, the best accuracy is produced by the SVM-SMOTE scenario with the RBF kernel, which produces an accuracy value of 88% with an error of 12%, and this is obtained from the division of testing data. Training data of 90:10. The results showed that the classification of diabetes using the 8-factor variable was more accurate if it was carried out using the Support Vector Machine (SVM) method with unbalanced data handling (SMOTE), and the distribution of data affected the test scenario. The main advantage of the Machine Learning method, especially the Support Vector Machine (SVM), in classifying diabetes compared to diagnosing diabetes in medicine is SVM's ability to process and analyze complex medical data with high accuracy. SVM can find hidden patterns and nonlinear relationships that are difficult for humans to detect in medical data involving many features. SVM can also address class imbalance problems in medical datasets and provide consistent and objective predictions. Although the Support Vector Machine (SVM) has advantages in classifying diabetes, some drawbacks need to be considered in its comparison with diagnosing diabetes in medicine; namely, SVM requires large amounts of data to train the model properly. In diabetes cases where medical data is often limited, using SVM can be challenging as it can lead to overfitting or underfitting issues and requires proper parameter tuning and selection of the appropriate kernel for optimal performance, which can be time-consuming and requires considerable technical expertise.

Suggestions for future research are to focus on developing more sophisticated feature selection strategies, such as genetic algorithms or more in-depth data mining, to select the most informative and relevant feature subsets in diabetes data. In addition, research can extend the application of SMOTE with various other oversampling techniques, such as ADASYN (Adaptive Synthetic Sampling), to address more complex class imbalances in diabetes datasets. Thus, this research is expected to make a significant contribution to increasing the accuracy and reliability of the SVM model in diagnosing diabetes, as well as providing new insights for the development of better classification techniques in the medical field.

5. ACKNOWLEDGEMENTS

Thank you to Ms. Hani Khaulasari, M.Si as the lecturer in the mathematical statistics course, who has patiently guided us in completing this journal, and also thanks to the team members who have worked well together to prepare this journal.

6. DECLARATIONS

AUTHOR CONTIBUTION

Firda Yunita Sari: Original Draft, Investigation. Maharani Sukma Kuntari: Data and Editing. Winda Ari Yati: Editing. Hani Khaulasari: Conceptualization, Methodology, Writing Review & Editing, Supervision.

FUNDING STATEMENT

This research received no specific grant from any funding agency.

COMPETING INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] R. Amelia, "Hubungan Perilaku Perawatan Kaki dengan Terjadinya Komplikasi Luka Kaki Diabetes pada Pasien Diabetes Melitus Tipe 2 di Puskesmas Tuntungan Kota Medan," *Talenta Conference Series: Tropical Medicine (TM)*, vol. 1, no. 1, pp. 124–131, 2018.

[2] B. Delvika, S. Nurhidayarnis, and P. D. Rinada, "Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil," vol. 2, no. 2 october 2022, pp. 68–75, 2022.

[3] M. D. M. Tito Putri, P. Wahjudi, and I. Prasetyowati, "Gambaran Kondisi Ibu Hamil dengan Diabetes Mellitus di RSD dr. Soebandi Jember Tahun 2013-2017," *Pustaka Kesehatan*, vol. 6, no. 1, p. 46, 2018.

[4] I. Diabetes Atlas, "International Diabetes Federation," *Diabetes Research and Clinical Practice*, vol. 10, no. 2, pp. 1–133, 2021.

[5] I. Maria, *Asuhan Keperawatan Diabetes Mellitus Dan Asuhan Keperawatan Stroke*. Deepublish, 2021.

[6] D. P. Paramita and A. W. Lestari, "Pengaruh Riwayat Keluarga Terhadap Kadar Glukosa Darah Pada Dewasa Muda Keturunan Pertama Dari Penderita Diabetes Mellitus Tipe 2 Di Denpasar Selatan," *Jurnal Medika*, vol. 8, no. 1, pp. 61–66, 2019.

[7] M. K. Murtiningsih, K. Pandelaki, and B. P. Sedli, "Gaya Hidup sebagai Faktor Risiko Diabetes Melitus Tipe 2," *Jurnal Ilmiah Kedokteran Klinik*, vol. 9, no. 2, p. 328, mar 2021.

[8] L. Hansur, D. Ugi, and A. Febriza, "Pencegahan Penyakit Diabetes Melitus Di Kelurahan Tamarunang Kec Sombaopu Kabupaten Gowa Sulawesi Selatan," *SELAPARANG Jurnal Pengabdian Masyarakat Berkemajuan*, vol. 4, no. 1, p. 417, 2020.

[9] F. Andaresta, S. Sudarsih, and M. Achwandi, "Asuhan Keperawatan Dengan Ketidakstbilan Kadar Gula Darah Pada Klien Diabetes Mellitus," Ph.D. dissertation, 2022.

[10] V. K. Putri and F. I. Kurniadi, "Klasifikasi Diabetes Menggunakan Model Pembelajaran Ensemble Blending," *Jurnal ULTIMATICS*, vol. 10, no. 1, pp. 11–15, 2018.

[11] A. Rahman Isnain, A. Indra Sakti, D. Alita, and N. Satya Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm," *Jdmsi*, vol. 2, no. 1, pp. 31–37, 2021.

[12] A. Muqiit WS and R. Nooraeni, "Penerapan Metode Resampling Dalam Mengatasi Imbalanced Data Pada Determinan Kasus Diare Pada Balita Di Indonesia (Analisis Data Sdki 2017)," *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 8, no. 1, p. 19, 2020.

[13] R. D. Fitriani, H. Yasin, and Tarno, "Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes (Studi Kasus: Status Peserta Kb Iud Di Kabupaten Kendal," *Jurnal Gaussian*, vol. 10, no. 1, pp. 11–20, 2021.

[14] S. Mutmainah, "Penanganan Imbalance Data Pada Klasifikasi," in *SNATi*, vol. 1, 2021, pp. 10–16.

[15] P. M. Joshi, T. N., & Chawan, "Logistic Regression and Svm Based Diabetes," *International Journal For Technological Research In Engineering*, vol. 5, no. July, pp. 4347–4350., 2018.

[16] V. C. Bavkar and A. A. Shinde, "Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement," *Indian Journal of Science and Technology*, vol. 14, no. 10, pp. 869–880, 2021.

[17] O. D. Amelia, A. M. Soleh, and S. Rahardiantoro, "Pemodelan Support Vector Machine Data Tidak Seimbang Keberhasilan Studi Mahasiswa Magister IPB," *Xplore: Journal of Statistics*, vol. 2, no. 1, pp. 33–40, 2018.

[18] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intelligence-Based Medicine*, vol. 3-4, no. November, p. 100023, 2020.

[19] D. Sepri, A. Fauzi, R. Wandira, O. S. Riza, Y. F. Wahyuni, and H. Hutagaol, "Prediksi Harga Cabai Merah Menggunakan Support Vector Regression," *Computer Based Information System Journal*, vol. 02, pp. 1–5, 2020.

[20] D. I. Ramadhan and B. Santosa, "Analisis Kinerja Peramalan dan Klasifikasi Permintaan Auto Parts Berbasis Data Mining," *Jurnal Teknik ITS*, vol. 9, no. 2, pp. 162–169, jan 2021.

[21] R. M. Mashita, S. Basuki, and N. Hayatin, "Prediksi Pemakaian Kwh Listrik Menggunakan Metode Support Vector Regression (SVR) (Studi Kasus: PT. PLN (Persero) Rayon Seririt)," *Jurnal Repositor*, vol. 2, no. 4, pp. 525–540, 2020.

[22] D. A. Agatsa, R. Rismala, and U. N. Wisesty, "Klasifikasi Pasien Pengidap Diabetes Metode Support Vector Machine," *e-Proceeding of Enginering*, vol. 7, no. 1, pp. 2517–2525, 2020.

[23] H. Khaulasari, "Combine Sampling Least Square Support Vector Machine Untuk Klasifikasi Multi Class Imbalanced Data," *Jurnal Widyaloka IKIP Widya Darma*, vol. 5, no. 3, pp. 261–278, 2018.

[24] L. Luo, S. Bao, and X. Peng, "Robust monitoring of industrial processes using process data with outliers and missing values," *Chemometrics and Intelligent Laboratory Systems*, vol. 192, p. 103827, sep 2019.

[25] E. A. Sembiring, "Pengaruh metode pencatatan persediaan dengan sistem periodik dan perpetual berbasis SIA terhadap stock opname pada perusahaan dagang di PT Jasum Jaya," *Accumulated Journal (Accounting and Management Research Edition)*, vol. 1, no. 1, pp. 69–77, 2019.

[26] P. R. Fitrayana and D. R. S. Saputro, "Algoritme Clustering Large Application (CLARA) untuk Menangani Data Outlier," in *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 5, 2022, pp. 721–725.

[27] R. Andhykha, H. R. Handayani, and N. Woyanti, "Analisis Pengaruh PDRB, Tingkat Pengangguran, dan IPM Terhadap Tingkat Kemiskinan di Provinsi Jawa Tengah," *Media Ekonomi dan Manajemen*, vol. 33, no. 2, pp. 113–123, 2018.

[28] D. Alita, Y. Fernando, and H. Sulistiani, "Implementasi Algoritma Multiclass Svm Pada Opini Publik Berbahasa Indonesia Di Twitter," *Jurnal Tekno Kompak*, vol. 14, no. 2, p. 86, 2020.

[29] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma SVM untuk Analisis Sentimen pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *Edutic - Scientific Journal of Informatics Education*, vol. 7, no. 1, pp. 1–11, 2020.

[30] D. Wahyuni, "Optimasi parameter support vector machine (svm) classifier menggunakan firefly algorithm (ffa) optimization untuk klasifikasi mri tumor otak," Ph.D. dissertation, 2019.

[31] N. Nafiah, "Klasifikasi Kematangan Buah Mangga Berdasarkan Citra HSV dengan KNN," *Jurnal Elektronika Listrik dan Teknologi Informasi Terapan*, vol. 1, no. 2, pp. 1–4, 2019.

[32] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance analysis and comparison of machine and deep learning algorithms for IoT data classification," *arXiv preprint arXiv:2001.09636*, 2020.

[33] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 1–22, 2020.