ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 10, No. 1, Jun 2024. pp. 127-140
doi: doi.org/10.21108/ijoict.v10i1.965

Revealing the Impact of the Combination of Parameters on SVM Performance in COVID-19 Classification

Sri Suryani Prasetiyowati [1*], Sri Harini [2], Juniardi Nur Fadila[3], Hilda Fahlena [4]

[1,4] *School of Computing, Telkom University*
Bandung, Indonesia

[2]*Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim*
Malang, Indonesia

[3]*Faculty of Computing, Universiti Teknologi Malaysia Skudai*
Johor Bahru, Malaysia

*srisuryani@telkomuniversity.ac.id,

Abstract

SVM is a method that has advantages in classification, but there are still obstacles in selecting optimal parameters. This research investigates the effect of parameter variations on SVM classification performance on the COVID-19 dataset, using linear, RBF, Sigmoid and polynomial kernels. Experiments were carried out to evaluate the combination of kernel parameter and regulations coefficient in each kernel against performance metrics such as accuracy, F1-Score, recall, and precision. In this study, the highest accuracy is 77.57% with an F1-Score of 76.67% when the parameter kernel and regulations coefficient are 0.75 and 0.75, respectively. The result is indicating an optimal balance between precision and recall. The other hand, the performance stability of polynomial kernel provides an optimal approach for analysis and prediction in classifying COVID-19 dataset, with more controlled fluctuations than other kernels. The interaction between the C and Gamma parameters shows that a Gamma value of 0.75 consistently provides good results, while adjusting the C parameter shows more controlled performance variations. This confirms that appropriate Gamma parameter settings are key in improving the accuracy and consistency of SVM model predictions in this case. This research provides insight into the potential of implementing SVM with polynomial kernels for COVID-19 classification, in analysis and prediction on complex biomedical datasets.

Keywords: Support Vector Machine, Kernel, Parameter, Performance, Covid-19

I. Introduction

CORONAVIRUS disease (COVID-19) is an infectious disease caused by a type of corona virus, namely SARS-CoV-2. The first time that there were COVID-19 cases in Indonesia were announced on March 2 2020 to December 31 2021, there were 4,262,720 confirmed cases in Indonesia with 4,292 active cases. Covid-19 transmission occurred more frequently in women than men in Bandung City, both in 2020 and 2021. In

Received on 30 June 2024. Revised on 23 July 2024. Accepted and Published on 25 July 2024.

2021, there were 17,696 men with Covid-19 (46.77%) and 20,221 (53.66%) women. Covid-19 infects many productive age groups who are high in activity and mobility. The low transmission of Covid-19 in children and adolescents can be caused by reducing activities by eliminating face-to-face processes in teaching and learning activities in schools and universities[1]. Based on this, it appears that the increase in the spread of Covid-19 cases can be influenced by several factors and the high or low number of cases can be classified to determine the status of vulnerability to its spread.

Several studies, including research [2], discuss the classification of Covid-19 using SVM, with an accuracy value of 87%, and this research uses Orange data mining software. Meanwhile, the classification of Covid-19 diagnosis using the SVM and Generalized Linear Model (GLM) methods carried out in research [3], obtained an accuracy of 95.64% for the GLM method and 98.91% for the SVM method. However, this research uses a small number of variables. Meanwhile[4] compared the Naïve Bayes and SVM methods in the classification of Diabetes Mellitus, the results showed that the accuracy of SVM with a polynomial kernel was 4.2% higher than the Naïve Bayes method. However, this study has a limited number of dataset records and only has binary class labels. Likewise, [5]–[7] are studies that use SVM for classification with a large number of attributes and the accuracy value is not high enough.

Support Vector Machine (SVM) is a supervised machine learning algorithm. SVM is used to identify hyperplanes and classify data [8]. Based on data characteristics, to find the optimal hyperplane in class separation, there are linear and non-linear SVMs[9]. The non-linear SVM functions to modify the kernel in the SVM. Each kernel function in linear and non-linear SVMs has several parameters that are used in the classification process. SVM is a method that has advantages in classification, but there are still obstacles in selecting optimal parameters. Disease classification in [3], uses a combination of parameters C (constant) 0, d (kernel degree) 3, and 4 Gamma variations, namely 0, 0.3, 0.6, 0.9, in Polynomial, RBF and Sigmoid kernels. The results obtained state that the highest accuracy is in the polynomial kernel, followed by the RBF kernel. Each kernel in SVM has advantages and disadvantages, regarding the number of features, parameter variations, data characteristics, which will affect the performance of classification results metrics[10]–[12]. In this research, the significance of parameter variations in each kernel function in SVM will be studied on the performance of classification results.

The case study used in this research is the number of Covid-19 cases with 22 features. The method used to determine the significance of the influence of selecting parameter values in the SVM kernel on classification performance is variance analysis. This research uses linear, RBF, polynomial and sigmoid kernels with scenarios of various parameter values and 10 repetitions. The difference with other research lies in the number and variety of features of the Covid-19 data set, variations in parameters and kernels used, as well as methods for studying the significance of the influence of variations in parameter values. The expected results are that there is a significant influence of parameter values on the performance of the Covid-19 classification results, with a 95% confidence interval and analysis of the impact of the combination of C and Gamma parameters on the performance of the SVM classification for each kernel.

.

II. RESEARCH METHOD

The system built in this research is a classification process using the SVM classification method with implementation scenarios of variations in several parameter values in four types of kernels, namely Linear, RBF, Polynomial and Sigmoid. Fig 1 explains the flowchart of the system design being built. There are 6 main processes in this research that are carried out sequentially, namely: data labeling, preprocessing, SVM model building, model testing and model evaluation.

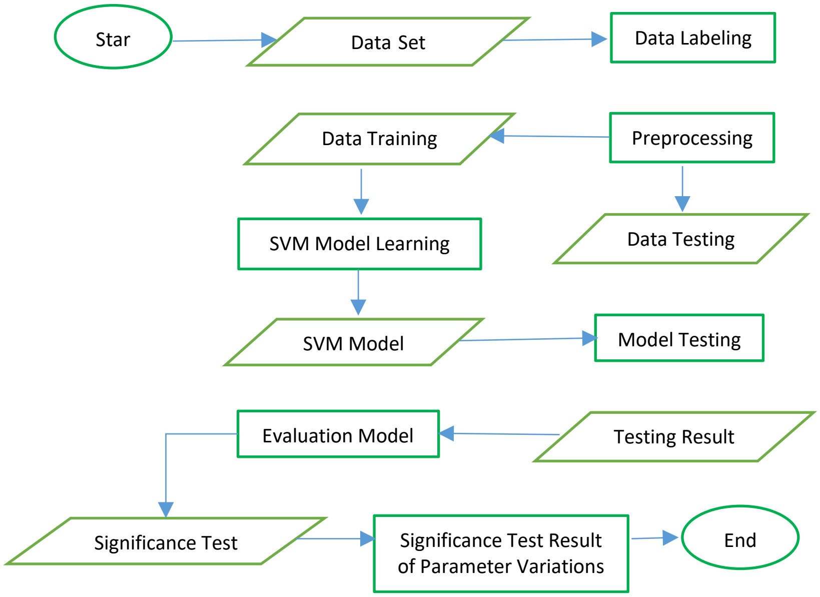


Fig. 1. Flowchart for SVM Classification

A. Data Set and Labelling

This research uses the Covid-19 dataset in the city of Bandung for the period November 2020 – April 2022. The Covid-19 dataset collected has 23 attributes and 2718 records. The data set is presented in Table I. Next, the dataset is split, with a proportion of 80:20, where 80% is training data and 20% is testing data.

TABLE I
DATA SET FEATURE DESCRIPTION

Data Set	Features
Number of Covid-19 cases	Number of male population, number of female population, rainfall (%), sunlight (^{0}C), average temperature (^{0}C), maximum temperature (^{0}C), minimum temperature (^{0}C), number of people who are not in school, kindergarten graduates, elementary school graduates, middle school graduates, high school graduates, D1 graduates, D2 graduates, D3 graduates, S1 graduates, Masters graduates, S3 graduates, proportion of first dose of vaccine (%), proportion of second dose of vaccine (%), proportion of third dose of vaccine (%) , proportion of compliance with wearing masks (%), proportion of compliance with maintaining distance (%)

The target class labeling for the number of Covid-19 cases is classified into 3 levels, with descriptions explained in Table II.

TABLE II
THE TARGET CLASS LABLING FOR THE NUMBER OF COVID-19 CASES

Class	Range	Label
Low	Cases < 218	0
Medium	218 ≤ Cases < 419	1
High	Cases ≥ 419	2

B. Preprocessing

Preprocessing is the stage of preparing the dataset before the dataset is used in the classification process. There are two data preprocessing activities, namely

1) Handling Missing Values means checking the dataset if there are missing values in the data rows. The activity carried out is deleting data in the dataset row.

2) Normalization functions to overcome problems in datasets that have high data range values ranging from 0 to 1. The activity carried out is adjusting the values in the dataset columns using the Standard Scaler in the Sklearn library. This method normalizes the data based on the average value of the data and the standard deviation value of the data [11].

C. Support Vectoc Machine (SVM) Model Learning

Support Vector Machines (SVM) is a supervised learning method used for classification, regression and ranking. This method is a very principled and robust machine learning method in maximizing prediction accuracy for detection and classification [12]. The SVM method is built on the idea of decision hyperplanes and constructs a linear function from a set of labeled training datasets [13]. Hyperplanes can define decision boundaries in input space or high-dimensional feature space and can separate positive samples from negative samples. This linear separator is created by the closest maximum distance from the hyperplane to the negative and positive samples. In linear classification, SVM can be divided into two, namely linear separable and linear non-separable.

Given a binary classification with the number of training samples N and each sample is denoted (x_i, y_i), where $i = 1, 2, 3 \dots, N$, $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})$ corresponds to the ith set of attributes, and $y_i = \{-1, 1\}$. The decision boundary of a linear classifier (separator) can be written as follows:

$$w^T x + b = 0 \tag{1}$$

where w is the weight vector and b is the bias factor. A linear classifier is defined as:

$$f(x) = sign(w^T x + b) \tag{2}$$

For a given data set and decision hyperplane, the margin function of the i^{th} sample x_i with respect to a hyperplane (w, b) is defined as in (3).

$$(w^T x + b) \geq 1 \rightarrow \; y_i = +1, \qquad (w^T x + b) \leq -1 \rightarrow \; y_i = -1 \tag{3}$$

Fig 2 shows that the margin function on the data set from the boundary decision is twice the margin function of one of the samples in the data set with the minimum margin function.

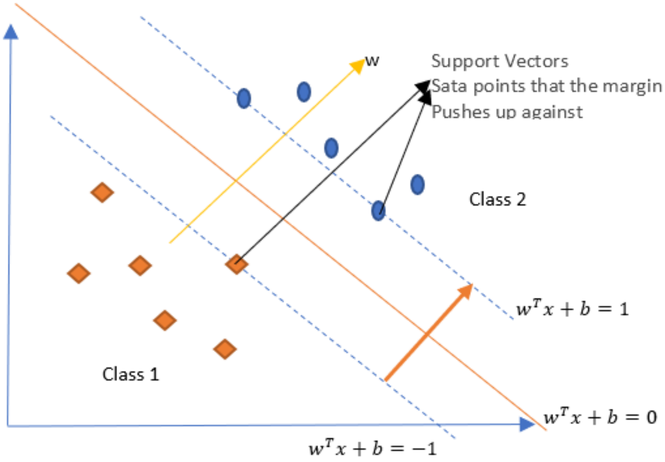


Fig. 2. Decision boundary and margin of SVM

The linear nonseparable case classifies linear data that cannot be separated, so the constraints in equation (3) must be changed linearly by adding a slack variable ξ_i which indicates a penalty to the separation accuracy that allows a point to be within the margin of error $(0 \leq \xi_i \leq 1, \forall i)$, or called misclassification $\xi > 1$, so that x_i is classified into:

$$(w^T x + b) \geq 1 - \xi_i \rightarrow \; y_i = +1, \qquad (w^T x + b) \geq -(1 - \xi_i) \rightarrow \; y_i = -1 \tag{4}$$

Search for the best dividing field by adding variable ξ_i often also called the soft margin hyperplane. The best dividing plane search formula or objective function changes to

$$\min_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \;\; , \xi_i \geq 0, C > 0 \tag{5}$$

The optimal hyperplane is determined to maximize generalization. However, if the training data is not linearly separable, then the designed classifier may not have high generalization ability, even if the hyperplanes are determined optimally. In nonlinear SVM, the scalar product $x_i^T x_j$ is transformed to $\varphi\,(x_i)^T \varphi(x_j)$. The transformation function in SVM uses the "Kernel Trick" (Scholkopf & Smola, 2002), which functions to calculate the scalar product through a kernel function. Function projection $\varphi : \, R^P \; \rightarrow H$ ensures that the inner product $\varphi\,(x_i)^T \varphi(x_j)$ is represented by the kernel function.

$$K(x_i, x_j) = \varphi\,(x_i)^T \varphi(x_j) \tag{6}$$

The classes in the COVID-19 data set are multiclass, so the output can be more than one class and must be divided into k mutually exclusive classes. There are several multi-class classification methods for SVM including Directed Acyclic Graph (DAG), Binary Tree (BT), One-Against One (OAO) and One Against All (OAA) classifiers, Decision Directed Acyclic Graph, Adaptive Directed Acyclic Graph. The DDAG method works like the OAO method in the training stage by building $\frac{k\,(k-1)}{2}$ binary classifier.

D. Scenario

The scenario carried out in this research is implementing the SVM method to classify the class of the number of Covid-19 cases by varying the parameters in various kernel functions. The parameter values used are $r = 1, d = 2, C = 0.5, 0.75, 1, 10, 100\;, Gamma(\gamma) = 0.005, 0.05, 0.1, 0.5, 0.75$. Meanwhile, the kernel functions used are Linear, Radial Basis Function (RBF), Sigmoid, Polynomial and are expressed in equations (7), (8), (9), and (10) respectively.

Linear Kernel
$$K(x_i, x_j) = x_i^T \, x_j \tag{7}$$

RBF Kernel
$$K(x_i, x_j) = EXP\left(-\,\gamma\, \left\|x_i - x_j\right\|^2\right) \tag{8}$$

Sigmoid Kernel
$$K(x_i, x_j) = tanh(\gamma \, x_i^T \, x_j + r) \tag{9}$$

Polynomial Kernel
$$K(x_i, x_j) = (\gamma \, x_i^T \, x_j + r)^d \tag{10}$$

E. Evaluation

Evaluation of SVM classification performance on the COVID-19 data set using performance metrics on a multi-class function matrix, because the number of target classes is more than two. The multiclass confusion matrix is described in Table 3 which has dimensions NxN, where N s the number of different class labels $C_1, C_2, \dots, C_N$. The analysis only focuses on certain classes and can be carried out based on the characterization described in Table III. Equations (7) to (10) are formulas for calculating classification performance based on a multiclass confusion matrix [14].

TABLE III
MULTI CLASS CONFUSION MATRIX

		Predicted Class			
		$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	...	$\boldsymbol{C_N}$
	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
True Class	C_2	FN	TP	...	FN
	...	...	...	...	
	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

$$Accuracy = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{i,j}} x100\% \tag{7}$$

$$Recall = \frac{TP(C_i)}{TP(C_i)+FN(C_i)} x100\% \tag{8}$$

$$Precision = \frac{TP(C_i)}{TP(C_i)+FP(C_i)} x100\% \tag{9}$$

$$F1 - Score = 2\frac{TPR(C_i)\ PPV(C_i)}{TPR(C_i)\ PPV(C_i)} x100\% \tag{10}$$

F. Significance Testing of the Effect of Parameter Variations

Testing the effect of parameter variations on SVM used analysis of variance (ANOVA). Analysis of variance is a statistical method that can detect differences between several experimental groups, with one or more independent variables. In this research, ANOVA can detect the significance of the effect of implementing varying C and Gamma parameters on SVM classification performance on COVID-19 data. The advantage of this method is that it is able to analyze experimental designs with several independent variables. The hypothesis and table IV explain the process of testing the significance of the influence of parameter variations:

H_0 : All scenarios provide the same response
H_1 : There are at least pairs of scenarios that provide different responses

TABLE IV
VARIANCE ANALYSIS TABLE

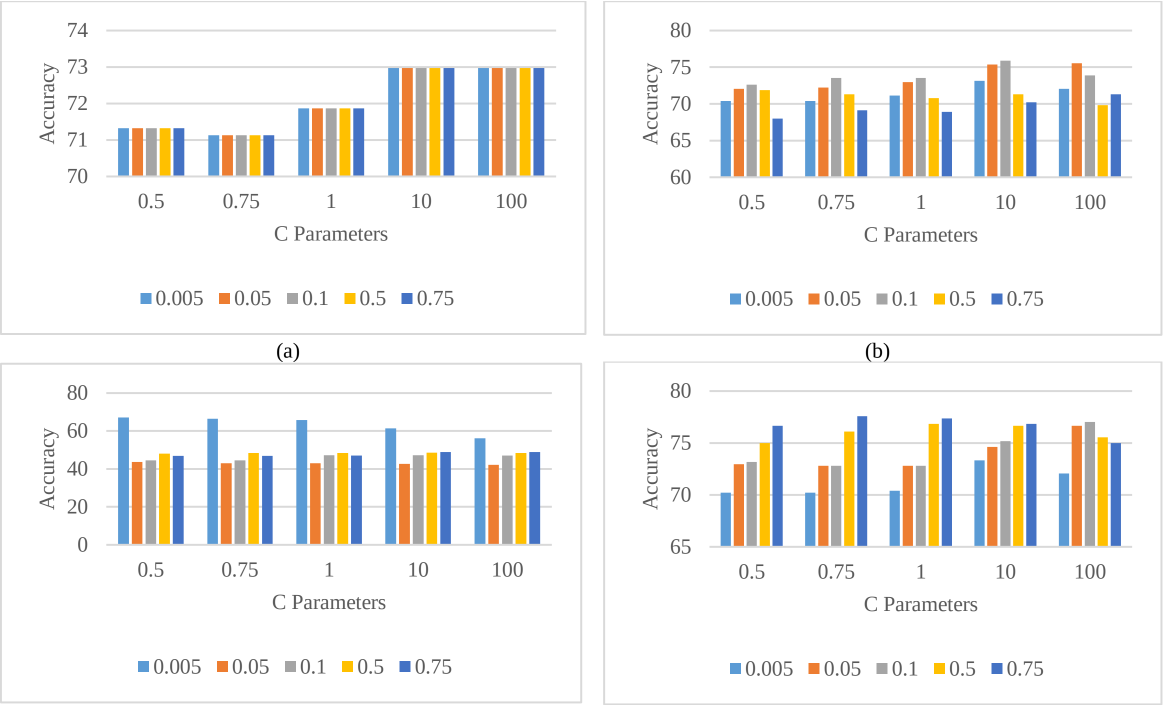
Source of variation	**Sum Square (SS)**	**Degree of free (Df)**	**Mean Square (MS)**	**F**
Alpha (between groups.BG)	SSBG	k - 1	$MSBG = \frac{SSBG}{k-1}$	$\frac{MSBG}{MSWG}$
Within groups (WG)	SSWG	k (r – 1)	$MSWG = \frac{SSWG}{k\ (r\ -1)}$	
Total	SST	kr - 1		

III. RESULTS AND DISCUSSION

A. Result

1) Results of Implementation of Parameter Variations in Each Kernel

Measurement of the performance of implementation scenarios for variations in the gamma and C parameters in each SVM kernel was carried out by varying the performance metric values, namely accuracy, F1-Score, recall and precision. Fig 3, Fig 4, Fig 5, and Fig 6 explain the performance pattern of the combination of parameter variations for each kernel in COVID-19 classification.



(a)

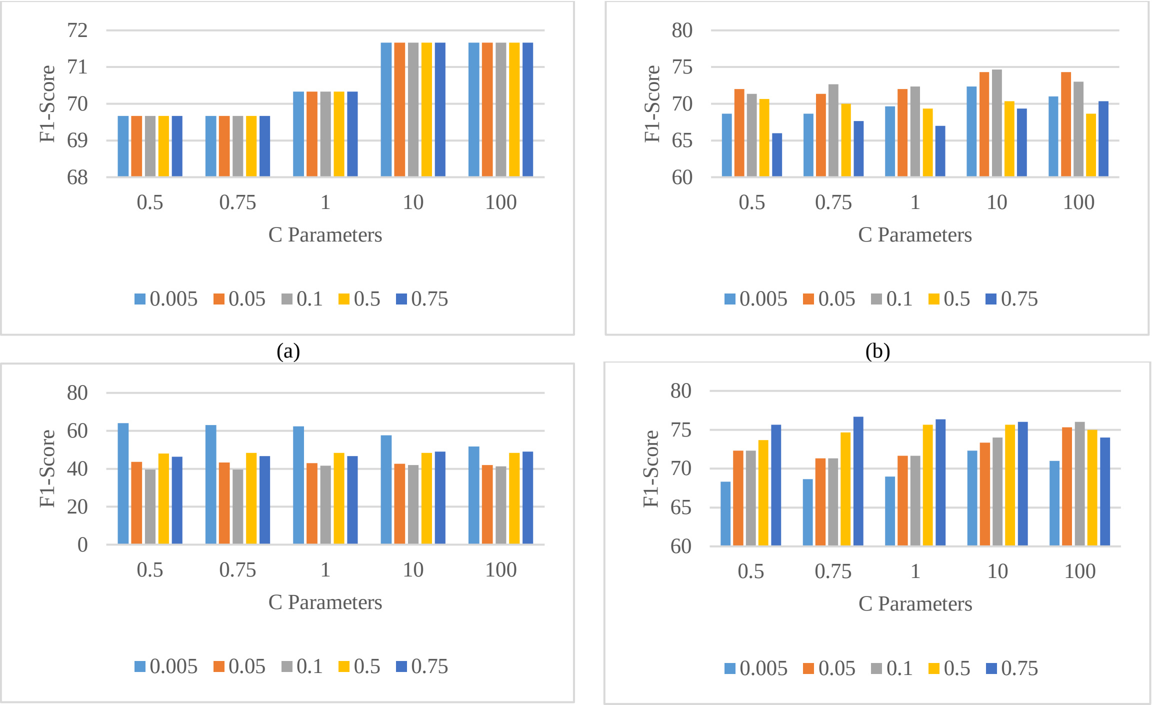
(b)

(c)

(d)

Fig 3. The accuracy pattern for each kernel for all variations of Gamma and Parameter C values, (a) Linear, (b) RBF, (c) Sigmoid, (d) Polynomial

The accuracy value of a model is a measure that can reflect the model's performance in predicting the class as a whole. In Figure 3(a) it can be seen that the combination of parameters c=100 and gamma 0.75 provides the highest accuracy for the linear kernel compared to other combinations. In this figure it can also be seen that the gamma parameter has no effect when working alone, because all the accuracy values are the same. However, in Figures 3(b), 3(c), and 3(d), the combination of C and Gamma parameters can be seen to have an influence on achieving accuracy performance, although it is not very significant for the sigmoid kernel, as explained in Figure 3c. In the sigmoid kernel, the highest accuracy value occurs for each C parameter when combined with Gamma 0.005 and reaches the lowest accuracy value when each C parameter is combined with Gamma 0.05. Meanwhile, in the RBF kernel, variations in the combination of parameters C and Gamma greatly influence the fluctuating accuracy patterns of COVID-19 classification using SVM. The highest accuracy can be achieved at any C value when combined with a Gamma value of 0.1 and reaches the lowest value when combined with a Gamma of 0.75. However, for a C value of 100, it will achieve the highest accuracy when combined with Gamma 0.05 and the lowest value when combined with Gamma 0.5. While the SVM classification performance for polynomial kernels, the highest accuracy was achieved at all C values when combined with Gamma 0.75 and the lowest value when combined with Gamma 0.005. However the highest accuracy is achieved for a C value of 100 when combined with Gamma 0.1.

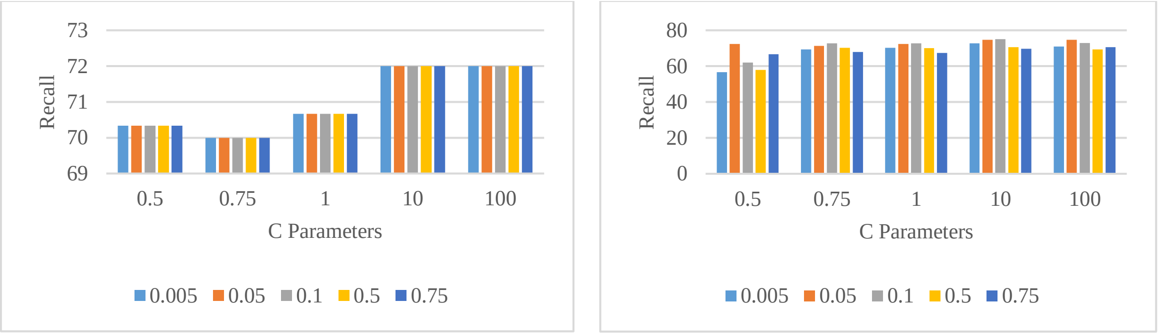


(a) (b)

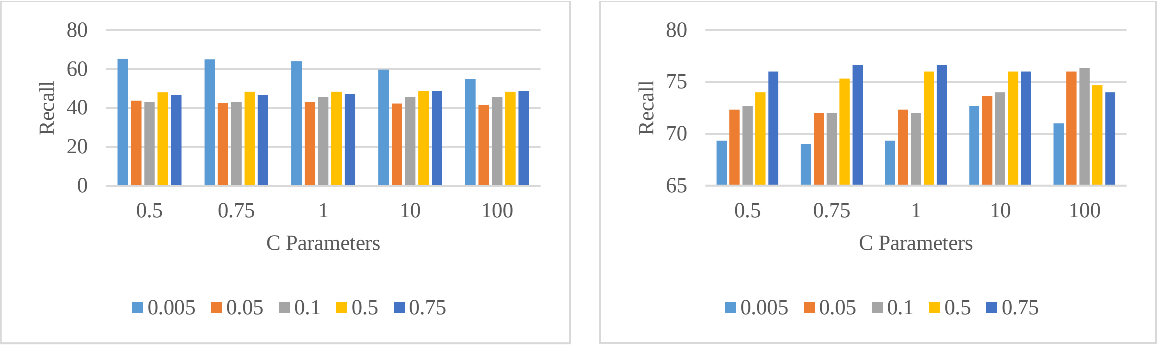
(c) (d)

Fig 4. The F1-Score pattern for each kernel for all variations of Gamma and Parameter C values, (a) Linear, (b) RBF, (c) Sigmoid, (d) Polynomial

F1-Score is calculated by considering the recall and precision values for each class. A high F1 Score value indicates that the model has a good balance between precision and recall for all classes. The F1-Score pattern for linear kernels is the same as the accuracy pattern, namely the value is only influenced by the value of parameter C. Meanwhile, the F1-Score pattern for RBF kernels is very volatile, where the highest F1-Score is achieved when C 0.5 and C 100 combined with Gamma 0.05, whereas for C 0.75, 1 and 10 when combined with Gamma 0.1. Meanwhile, the lowest F1-Score was achieved when Gamma 0.75 was combined with all C, except C 100 with Gamma 0.5. The pattern in the Sigmoid shows that the highest F1-Score value is achieved when all C parameter values are combined with Gamma 0.005 and the lowest when all C parameter values are combined with Gamma 0.1. Meanwhile for polynomial kernels, the combination of parameters C with Gamma reaches the highest F1-Score value when Gamma is 0.75, except for C 100 with Gamma 0.1. Meanwhile, the lowest F1-Score value was achieved by all C parameter values when combined with Gamma 0.005. The highest F1-Score value for COVID-19 classification using SVM was 76.67 achieved in a polynomial kernel with a combination of C 0.75 and Gamma 0.75.



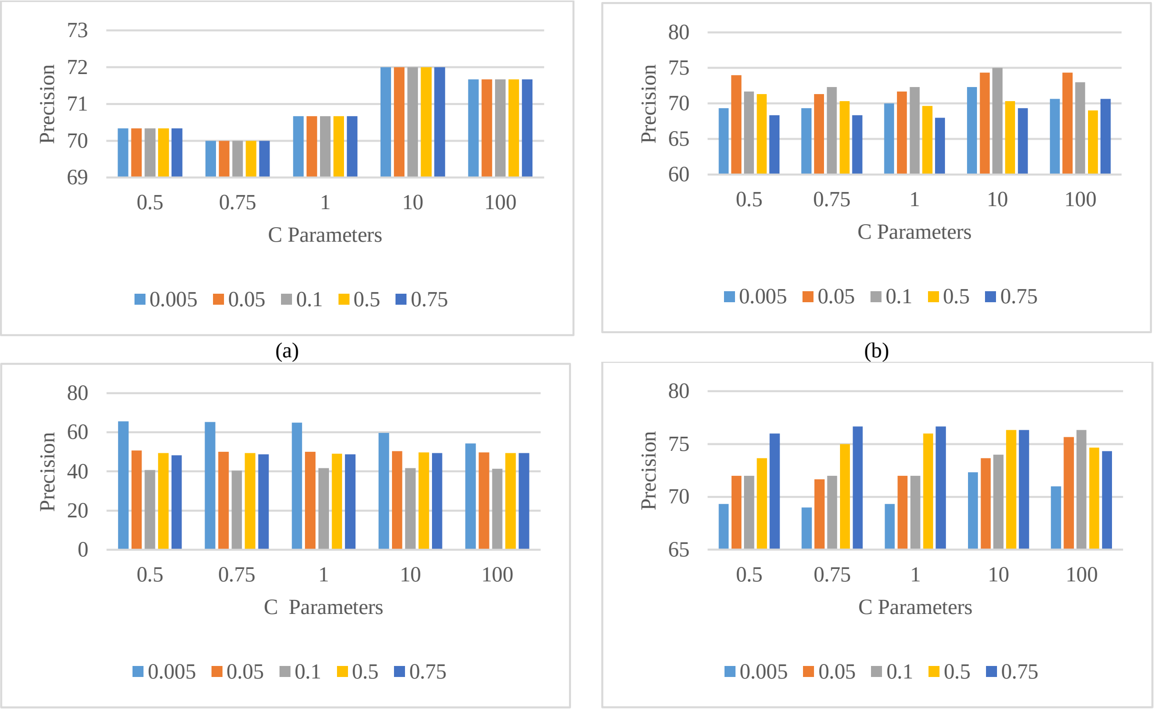
(a) (b)



(c) (d)

Fig 5. The recall pattern for each kernel for all variations of Gamma and Parameter C values, (a) Linear, (b) RBF, (c) Sigmoid, (d) Polynomial

Recall is a measure of model performance in identifying all instances of a class that are true positives. The pattern of recall values for the linear kernel in Fig 5 is the same as for accuracy and F1-Score, the Gamma parameter value has no effect. Meanwhile, the recall pattern in the RBF and sigmoid kernels does not fluctuate, where the highest value is achieved for all C values with Gamma 0.005, the lowest with Gamma 0.05. While the recall pattern in the polynomial kernel fluctuates, with the highest value achieved for all C with Gamma 0.75, but for C 100 the highest recall is with Gamma 0.1. Meanwhile, the lowest recall value was achieved for all C parameters with Gamma 0.005. The highest Recall value in the SVM classification results of 76.67 was achieved in the polynomial kernel with a combination of C 0.75 and Gamma 0.75. The recall value of 76.67 indicates that the model is able to recall around 76.67% of all true positive cases for all classes.



(a) (b)

(c) (d)

Fig 6. The Precision pattern for each kernel for all variations of Gamma and Parameter C values, (a) Linear, (b) RBF, (c) Sigmoid, (d) Polynomial

The number of correct positive predictions compared to the total positive predictions made by the model is measured by the precision value. For the precision pattern in the linear kernel, the performance is also the same as the other three patterns, and in the sigmoid kernel, the highest precision value is achieved for all C parameters when combined with Gamma 0.005 and the lowest when combined with Gamma 0.1. The precision pattern is quite fluctuating in the RBF and polynomial kernels. In the RBF kernel, the highest precision values were achieved by the combination of C 0.5 and 100 with Gamma 0.05, while for C 0.75, 1, and 10 with Gamma 0.1. The lowest precision value was achieved at all C values with Gamma 0.75. Meanwhile, for the polynomial kernel, the highest precision values were achieved at C 0.5 and 0.75 with Gamma 0.75, and the lowest value was achieved at C 100 with Gamma 0.1. The precision performance value of COVID-19 classification using SVM, the highest precision value of 76.67 was achieved in the polynomial kernel with a combination of parameters C 0.75 and Gamma 0.75.

2) Testing the effect of parameter variations on performance metrics for each kernel

In this study, testing the effect of variations in the C and Gamma parameters on each classification performance for each kernel was measured using the analysis of variance significance test. Testing uses a 95% confidence interval. To find out the significance of variations in the C and Gamma parameters, you can see the significance of the F table value or P-value. Table V, Table VI, Table VII, and Table VIII illustrate the response of classification performance to the implementation of parameter variations for each kernel.

TABLE V
Significance Test of Gamma and C Parameters on Accuracy

Kernel	Source of Variance	Degrees of Freedom	F Table	P -Value	F Critical
RBF	Gamma	4	18.39	$7.83 e^{-6}$	3.01
	C Parameters	4	4.64	0.01	3.01
Sigmoid	Gamma	4	56.32	$3.17 e^{-9}$	3.01
	C Parameters	4	0.42	0.79	3.01
Polynomial	Gamma	4	13.56	$5.2 e^{-5}$	3.01
	C Parameters	4	1.96	0.15	3.01

TABLE VI
Significance Test of Gamma and C Parameters on F1-Score

Kernel	Source of Variance	Degrees of Freedom	F Table	P- Value	F Critical
RBF	Gamma	4	21.01	$3.31 e^{-6}$	3.01
	C Parameters	4	5.66	0.004	3.01
Sigmoid	Gamma	4	40.42	$3.58 e^{-8}$	3.01
	C Parameter	4	0.49	0.75	3.01
Polynomial	Gamma	4	14.97	$2.86 e^{-5}$	3.01
	C Parameter	4	2.42	0.09	3.01

TABLE VII
Significance Test of Gamma and C Parameters on Recall

Kernel	Source of Variance	Degrees of Freedom	F Table	P -Value	F Critical
RBF	Gamma	4	3.37	0.04	3.01
	C Parameter	4	8.73	0.001	3.01
Sigmoid	Gamma	4	53.26	$4.79 e^{-9}$	3.01
	C Parameter	4	0.42	0.79	3.01
Polynomial	Gamma	4	12.69	$7.69 e^{-5}$	3.01
	C Parameter	4	1.61	0.22	3.01

TABLE VIII
SIGNIFICANCE TEST OF GAMMA AND C PARAMETERS ON PRECISION

Kernel	Source of Variance	Degrees of Freedom	F Table	P= Value	F Critical
RBF	Gamma	4	16.99	1.3 e^{-5}	3.01
	C Parameter	4	3.47	0.03	3.01
Sigmoid	Gamma	4	52.75	5.14 e^{-9}	3.01
	C Parameter	4	0.75	0.57	3.01
Polynomial	Gamma	4	15.07	2.75 e^{-5}	3.01
	C Parameter	4	2.38	0.09	3.01

The results obtained in Fig 2, Fig 3, Fig 4, and Fig 5 are in accordance with the test results using analysis of variance, which show that the Gamma parameter has a very significant influence on the performance values of all SVM classification metrics on the RBF, Sigmoid and polynomial kernels. Meanwhile, the C parameter only has a significant effect on classification performance when implemented in the RBF kernel.

3) Classification performance for each combination of C and Gamma parameters

SVM classification performance is measured based on accuracy, F1-Score, recall and precision values. Fig 7, Fig 8 and Fig 9 explain the classification performance patterns for each parameter combination in each kernel.

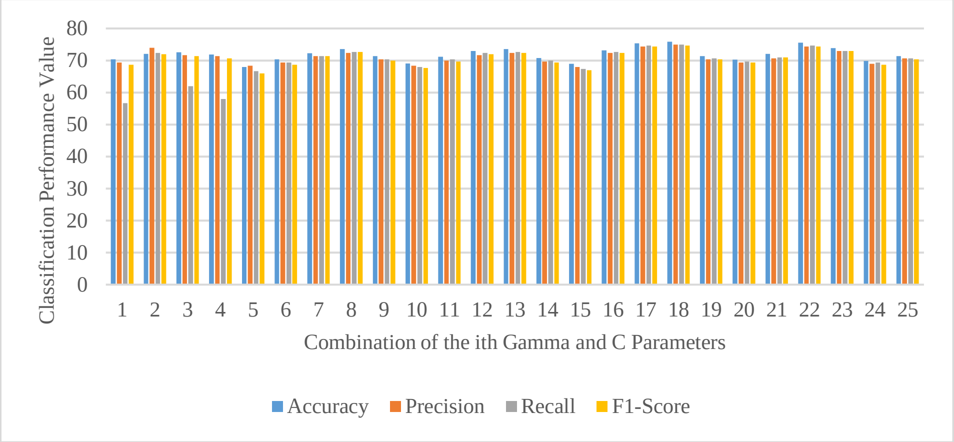


Fig 7. The Effect of combined parameter variations on classification performance patterns for RBF kernels

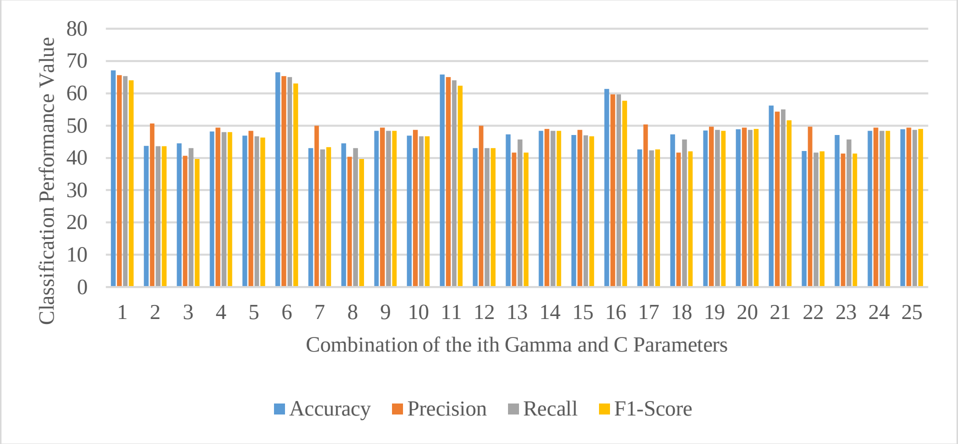


Fig 8. The Effect of combined parameter variations on classification performance patterns for Sigmoid kernels

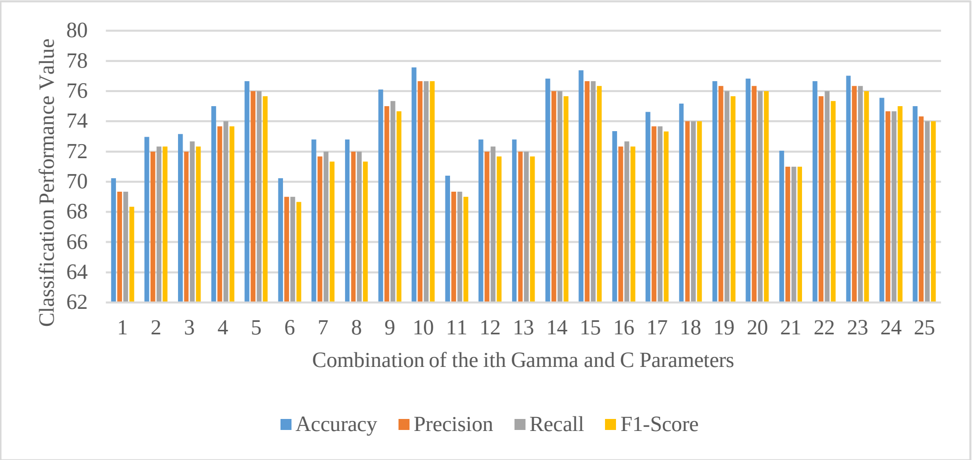


Fig 9. The Effect of combined parameter variations on classification performance patterns for Polynomial kernels

The performance of classifying the number of COVID-19 cases using SVM is shown by the best metric performance for the RBF kernel, namely accuracy of 75.91, F1-Score of 74.67 with recall and precision of 75 achieved for a combination of parameter values C 10 and Gamma 0.1. Meanwhile, for the sigmoid kernel, the best performance was respectively accuracy of 67, F1-Score of 64 with recall of 65.33 and precision of 65.67 which was achieved for a combination of C values of 0.5 and Gamma of 0.005. Meanwhile, in the polynomial kernel, the best SVM classification performance value is shown by an accuracy value of 77.57, F1-Score 76.67 with recall 76.67 and precision 76.67 achieved for a combination of C values 0.75 and Gamma 0.75.

B. Discussion

The problem discussed in this research is to determine the effect of implementing parameter variations on SVM classification performance for the COVID-19 data set on linear, RBF, Sigmoid and polynomial kernels. The discussion of implementation results focuses on three things, namely the performance pattern of combinations of parameter variations for each kernel, testing the influence of parameter variations on performance metrics for each kernel, and classification performance for each combination of C and Gamma parameters. Based on the results presented in the form of a bar chart, the performance pattern of each combination of C and Gamma parameter variations for each kernel can be seen: In the linear kernel, accuracy and F1-Score are influenced by the C parameter value, while gamma does not have a significant influence if it works alone. In this kernel, the main influence is seen in the C parameter value on accuracy and F1-Score. Gamma, although present, did not significantly affect performance when tested individually. This shows that in the case of COVID-19 data, the complexity of the SVM model with a linear kernel may not require complicated adjustments to the Gamma parameter. The combination of the C and Gamma parameters in the RBF kernel greatly influences the fluctuation of the accuracy pattern, with the highest values at C 10 and Gamma values 0.1. large fluctuations in accuracy depending on the combination of C and Gamma parameters. The highest accuracy was found at C 10 with Gamma 0.1. This indicates that for the COVID-19 dataset, proper setting of the C and Gamma parameters is very important to obtain optimal results. The RBF kernel is naturally capable of handling non-linear relationships in data, and parameter variations can provide flexibility in adapting SVM models to complex data characteristics. Meanwhile, in the sigmoid kernel, the influence of the Gamma parameter is more significant than C on accuracy, with the best value when C is 0.5 and Gamma is 0.005. Based on these results, it appears that the Gamma parameter has a more significant influence than the C parameter on accuracy. The best performance was found at Gamma=0.005 for all values of C. The sigmoid kernel is used when the relationship between input and output variables is predicted using the sigmoid function. Using Gamma values requires less input data. Meanwhile, different Gamma values in the polynomial kernel have an effect on the F1-Score value, with the best results at C 0.75 and Gamma 0.75. This shows that in this case, the C and Gamma parameters support each other to improve the accuracy of SVM classification.

Testing the influence of parameter variations on the performance metrics for each kernel shows that the Gamma parameter has a significant influence on the SVM classification accuracy value for the RBF, Sigmoid and polynomial kernels. Meanwhile, the C parameter also has a significant effect on the RBF kernel, although not as strong as Gamma for Sigmoid and polynomial kernels.

SVM classification performance for the COVID-19 dataset can be optimized by choosing the right combination of parameters for each kernel. The polynomial kernel shows the best performance with high F1-Score values, indicating a good balance between precision and recall for all classes in this dataset. In addition, Gamma parameter settings also play a crucial role in improving classification accuracy and consistency in RBF, Sigmoid, and polynomial kernels. Linear Kernel: Stable classification performance with the highest value at C=100. Kernel RBF: Significant fluctuations in accuracy depending on the combination of C and Gamma, where the best accuracy was 75.91%, with an F1-Score of 74.67, recall of 75, and precision of 75 for C 10 and Gamma 0.1. The best performance on the sigmoid kernel is achieved with Gamma 0.005 for all C, where the best accuracy is 67%, with F1-Score 64, recall 65.33, and precision 65.67 for C 0.5 and Gamma 0.005. While on the polynomial kernel, the best balance between precision and recall at C 0.75 and Gamma 0.75, with the highest accuracy of 77.57%, F1-Score 76.67, recall 76.67, and precision 76.67. The polynomial kernel shows stability in classification performance, with relatively lower fluctuations compared to the RBF and Sigmoid kernels in the tests carried out. This stability can be an advantage in the context of COVID-19 datasets, where consistent predictions can help in making better decisions regarding case classification. The high F1-Score (76.67%) indicates that the polynomial kernel is able to achieve a good balance between precision (accuracy of positive predictions) and recall (ability of the model to find all instances of the class that are true positives). This is important because in the context of COVID-19, it is important to have a model that is not only accurate in predicting, but also able to identify positive cases correctly. The Gamma 0.75 parameter consistently provides good results when combined with various C values, such as 0.5, 0.75, and 100. This indicates that Gamma 0.75 has a stable positive impact on the performance of the polynomial kernel. Adjustment of the C parameter showed more controlled variations in performance, with the highest values achieved at C 0.75 and Gamma 0.75, as well as more limited changes in performance at different C values.

IV. CONCLUSION

Based on the results and discussion, it can be concluded that the polynomial kernel shows excellent performance in classifying the COVID-19 dataset using SVM. The optimal combination of C and Gamma parameters, especially C 0.75 and Gamma 0.75, provides the highest accuracy and a balanced F1-Score between precision and recall. The performance stability and ability of polynomial kernels to handle non-linear relationships in data makes them a good choice. A Gamma value of 0.75 makes a significant contribution in improving the accuracy and consistency of model predictions. The highest accuracy value in COVID-19 classification of 77.57% was achieved when a C value of 0.75 and Gamma of 0.75 was implemented in the polynomial kernel, this shows that around 77.57% of all predictions made by the model were correct. Meanwhile, the F1-Score value for the polynomial kernel shows that the model has a good balance between precision and recall for all classes. For further research, another multiclass SVM method is used and implemented in other case studies with different parameter values.

ACKNOWLEDGMENT

The authors wish to thank Telkom University for the support of funds and infrastructure, so that this research can be completed.

REFERENCES

[1] Dinkes Jawa Barat, "Profil Kesehatan Jawa Barat," *Dinas Kesehat. Provinsi Jawa Barat*, vol. 5, no. 3, pp. 248–253, 2020.

[2] S. Guhathakurata, S. Kundu, A. Chakraborty, and J. S. Banerjee, "A novel approach to predict COVID-19 using support vector machine," *Data Sci. COVID-19 Vol. 1 Comput. Perspect.*, no. May, pp. 351–364, 2021, doi: 10.1016/B978-0-12-824536-1.00014-9.

[3] Y. Rismawati, I. Made Tirta, and Y. Setia Dewi, "Klasifikasi Data Diagnosis Covid-19 Menggunakan Metode Support Vector Machine (SVM) dan Generalized Linear Model (GLM) (Classification of Covid-19 Diagnosis Data Using Support Vector Machine (SVM) and Generalized Linear Model (GLM) Methods)," [Online]. Available: https://magestic.unej.ac.id/.

[4] H. Apriyani, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," vol. 1, no. 3, pp. 133–143, 2020.

[5] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman, "Poverty classification using machine learning: The case of Jordan," *Sustain.*, vol. 13, no. 3, pp. 1–16, 2021, doi: 10.3390/su13031412.

[6] N. Nofriani, "Machine Learning Application for Classification Prediction of Household's Welfare Status," *JITCE (Journal Inf. Technol. Comput. Eng.*, vol. 4, no. 02, pp. 72–82, 2020, doi: 10.25077/jitce.4.02.72-82.2020.

[7] M. Yanto, Y. Wiyandra, and S. Defit, "Deep learning approach analysis model prediction and classification poverty status," *IAES Int. J. Artif. Intell.*, vol. 12, no. 1, pp. 459–468, 2023, doi: 10.11591/ijai.v12.i1.pp459-468.

[8] V. Kadam, S. Kumar, A. Bongale, S. Wazarkar, P. Kamat, and S. Patil, "Enhancing surface fault detection using machine learning for 3d printed products," *Appl. Syst. Innov.*, vol. 4, no. 2, 2021, doi: 10.3390/asi4020034.

[9] Z. P. D. Marston *et al.*, "Linear Support Vector Machine Classification of Plant Stress From Soybean Aphid (Hemiptera: Aphididae) Using Hyperspectral Reflectance," *J. Econ. Entomol.*, vol. 115, no. 5, pp. 1557–1563, 2022, doi: 10.1093/jee/toac077.

[10] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, pp. 5497–5506, 2020, doi: 10.11591/IJECE.V10I5.PP5497-5506.

[11] N. S. J. Jailani, Z. Muhammad, N. S. Damanhuri, M. H. F. Rahiman, and M. N. Taib, "Kaffir Lime Oil Quality Grading Using Non-Linear Support Vector Machine With Different Kernels," *Malaysian J. Anal. Sci.*, vol. 26, no. 2, pp. 295–302, 2022.

[12] S. Wu, "Information System Establishment and Prediction Analysis for SVM Model of RBF Kernel Function," *J. Phys. Conf. Ser.*, vol. 1982, no. 1, 2021, doi: 10.1088/1742-6596/1982/1/012155.

[13] K. Poddar, G. B. D. Amali, and K. S. Umadevi, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," *2019 Innov. Power Adv. Comput. Technol. i-PACT 2019*, pp. 1–5, 2019, doi: 10.1109/i-PACT44901.2019.8960044.

[14] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," 2021.