OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 8, No. 2, Dec 2022. pp. 40-52
doi: 10.21108/ijoict.v8i2.671

The Analysis of Support Vector Machine (SVM) on Monthly Covid-19 Case Classification

Rifaldo Sitepu [1], Aniq A. Rohmawati [2], Sri Suryani Prasetyowati [3]

[1,2,3]*School of Computing, Telkom University*
Jl. Telekomunikasi 1 Terusan Buah Batu, Bandung 40257, Indonesia

[*] rifaldositepu@student.telkomuniversity.ac.id

Abstract

Covid-19 is disease caused by the new corona virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The effect of this virus usually causes infection on respiratory system. Covid-19 was rapidly spread globally. Experts said that the factor that caused this to spread rapidly is human mobility. Therefore, several countries create new rules so that it can suppress the spreading of this disease, by prohibiting a large scale gathering, keeping away distance with each other, mandatory rule of using mask, and the prohibition for the entry of their country. This research proposes a performance analysis of Support Vector Machine (SVM) to classify the monthly data of covid-19. The dataset used was the covid-19 data of towns in Bandung from November 2020 to December 2021 which was acquired from The Public Health Office of Bandung. The data that was collected includes the vaccination data of covid-19, the obligatory to use mask in public and keeping up distance with each other. This research also uses the weather data from BMKG Bandung, which includes temperature, rainfall pattern, and sunshine pattern. From the evaluation it is found that the best accuracy comes from December 2021 with 100% accuracy followed by August 2021 with 97% accuracy and October 2021 with 90%. From the result it is also found that the average that produces the best accuracy, precision, f1-score, and recall came from RBF kernel. Which can be concluded that support vector machine (SVM) is good to classify the amount of active cases of covid-19.

Keywords: Classification, Covid-19, Support Vector Machine.

I. INTRODUCTION

Atypical pneumonia plague was first reported by the public officers of Wuhan, Hubei, China, on December 2019 to be exact [1]. That disease was caused by the new corona virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1],[4],[6]. This disease is caused by SARS-COV-2 which is called Covid-19 [2],[3],[4],[5],[6]. The spread of this disease is rapidly fast considering it was spread throughout China, and it can be found all around the world. By 11 March 2020 World Health Organization (WHO) declared that this disease which was caused by SARS-CoV-2 or Covid-19 as a pandemic [2],[4],[6].

Indonesia is a country with more than 270,20 million population on 2020 is also been affected by the spread of SARS-CoV-2 virus [17]. In 2021, Indonesia also shows the trend of increasing active cases, where in 7 January 2021 the active cases found were 9321 cases, with the total of 23520 death cases, 659437 cured cases and 114766 active cases, which contributes 14.4% to the overall world cases [22]. At the same year, that in a

Received on 05 Sep 2022. Revised on 17 Sep 2022. Accepted and Published on 30 Des 2022.

few provinces in Indonesia the cured cases found to be increasing from the last seven days or which also can be called seven day moving average in 31st October 2021, where one of the provinces is West Java. Where the cases that has been reported is 741 new cases from 1424 total active cases [23].

The most known factor for this disease was caused by human mobility. A few countries established their restriction to enter their country to decrease human mobility in hope to suppress the spread of Covid-19, by prohibiting a large scales human grouping activity, using a mask to go to places, and to keeping distance with others [2],[5],[6],[7]. The new variant of Covid-19 virus was keep being discovered until now, the new variant supposedly easier to be transmitted and there a few countries that still have worse cases of this new variant of Covid-19 [6].

In 2020, there is research about the performance estimation algorithm of machine learning with the analysis of factor in covid-19 dataset. The purpose of this research is to predict and classify anything related to covid-19. Where this research uses machine learning algorithm like linear and logistic regression, Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and SVM with grid search. The data used on that research was a dataset of covid-19 collected from The Center for Systems Science and Engineering (CSSE) of Jhon Hopkins University (JHU). From that research, they found that the use of SVM method using grid shows that the highest accuracy approximately is 95%, following with the decision tree method by an approximately 94%. Even though with this great result, there are still a few boundaries found in that research. One of them was the availability of covid-19 patient data [11].

There is also other research about prediction of patient covid-19 which uses machine learning algorithm. This research uses five classification method which are Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor (KNN), and Decision Tree [19]. From this research it is stated that the accuracy of Decision Tree is yet to best, with the highest accuracy at 94,5% followed by KNN, SVM, LR, and GNB. Where the method SVM was found to have an accuracy of 85% and a precision value approximately around 81,1%. With that stated this research has not used time series data.

Based on the research above, we can conclude that Support Vector Machine is a good method to predict and classify a time series data of covid-19. Which can be seen from the advantage of SVM which has a theoretical base that can be learned and implemented easily. In a restricted sample, SVM also one of the methods that can solve problem of higher dimensions and a small amount of sample [24],[26],[28]. With that being said, there has not been a research regarding the issue, therefore this research propose an analysis of SVM performance to calculate the monthly cases of covid-19. The purpose of this research is to give a larger picture about the performance of SVM in classification which can help give more insights to predict diseases [20], in hope to help people or the government.

II. LITERATURE REVIEW

Covid 19 is a disease caused by the new virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). This virus usually causes an infection to the respiratory system. The symptoms that it caused can be varied, such as: fever, coughing, headache, low energy, and stomach pain [4]. On 11th March 2020 WHO stated that covid-19 is considered to be pandemic, this is because it is quite a concern because it affected people globally. All of the searching topic about covid-19 was increased and dominate all of the information over social media. This causes hoax concerning covid-19 also surfaces. For example, the information about how to avoid the disease which are not effective and dangerous, the use of drugs out of context, and a test without any reasonable knowledge [6]. In example on a few cases, the use of ibuprofen or any anti-inflammation drugs which was stated to increase the probability of getting covid-19, which is not reasonable and leading people to the wrong way. There is also a few cases of misusing vitamin D which causes poisoning because of the use of methanol [8]. Besides of that, the statement from experts which causes panic buying, so that there was a time where mask and personal protective equipment become so extremely rare which causes the price to be out of this world [8].

Referring to the data from The Center for Systems Science and Engineering (CSSE) in Jhon Hopkins University, there are at least 1.400.000 case that have been reported from around the world by 7th April 2020 [4]. Then on 23rd October 2021, the spread of covid-19 has yet to be infected to more than 243,547,503 people globally. With the total of 4,949,697 death cases, and over 220,661,965 people cured cases [9]. From the first case that has been announced, a researcher Jiang et al stated that the percentage of death cases caused by covid-19 was 4,5%, where 8,0% was a patient with a range 70-79 of age, and 14,8% in patient with the age of more than 80 [7]. Jiang et al also stated that patient in the range >50 with a chronic disease has also a high risk if not being treated carefully [7].

In 2020, there has been research that conduct the prediction towards the spread of covid-19 using SVM method. In this research they use a time series data which was acquired from 22nd January 2020 until 25th April 2020 globally, where the data consists of location, death case, and cured case in attribute. These data were collected from Public Repository Center for System Science and Engineering (CSSE) of Jhon Hopkins University (JHU). In this research the SVM method was used to explore the impact of identifying death and cure. The research stated that the prediction from patient of covid-19 is depends on the score of the attribute when using the calculator model of SVM. Besides of that, the researcher also found that the ideal hyperplane using BRF and C, can help to make a comparison between hyperplane and to find better vector. In that research they also found that by conducting a statistic analysis using a bar chart to differentiate the subject. They use kernel function SVM to make an optimal performance value in predicting the case of covid-19. The researcher also stated that the SVM model is one of the techniques in machine learning that can give great result and it is easy to use [10].

Research on the performance estimation algorithm of machine learning with the analysis of factor in covid-19 dataset was conducted by [11]. The purpose of this research is to predict and classify anything related to covid-19. Where this research uses machine learning algorithm like linear and logistic regression, Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and SVM with grid search. The data used on that research was a dataset of covid-19 collected from The Center for Systems Science and Engineering (CSSE) of Jhon Hopkins University (JHU). From that research, they found that the use of SVM method using grid shows that the highest accuracy approximately is 95%, following with the decision tree method by an approximately 94%. Even though with this great result, there are still a few boundaries found in that research. One of them was the availability of covid-19 patient data.

Similar research that explain the analysis of the pandemic covid-19 using SVM classifier was also conducted. Which also uses other machine learning methods like Random Forest (RF), and Artificial Neural Network (ANN). The purpose of this research is to help to decrease the time of waiting the medical result and to predict covid-19 which to help predict the detection of the virus faster, in hope to decrease the mortal rate. The research stated that the prediction to count the case of covid-19 depends on the feature rank that was chosen by SVM calculator. The design of SVM is used to find the ideal hyperplane to divide cluster, which on one side consists of variable target and the other side with other categories. Support vector is a vector that can possibly be a hyperplane. In this research they conclude that one of the best qualities of SVM model is their ease of use in AI procedure to predict [12].

In other studies, there are research about machine learning and statistic modelling to predict covid-19 in Jordan that is conducted in 2020. This study uses a few prediction models which based on statistic model called Logistic Regression (LR), and a machine learning model Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). The purpose of this study is to predict the potential patient for covid-19 by identifying the signs and symptoms. From this study it was stated that the SVM model is a model that can give a precision value up to 91,67%. This is because the high correlation which can be seen in the data collection. While MLP shows the accuracy up to 91,62% compared to other models. On the other hands the usage of Logistic Regression model was stated to be poor to predict patient of covid-19 because it has a relative failure value and a precision value up to 66,67% [13].

In 2021, there is a research where the purpose is to analyze and compare the better machine learning algorithm that can be used to build the covid-19 prediction using dataset of symptoms and the signs of covid-19 from Kaggle. This study uses dataset that consists of 20 attributes and 1 class attribute. Where they use a few Supervised Machine Learning algorithm which are, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors and Naïve Bayes which is being used in a machine learning software called WEKA. From this study the researchers use dataset of Covid-19 to optimize the hyperparameter to select the value in which algorithm will work best. This research conducts a crossing validation for approximately 10 times and comparing all of them based on the amount of accuracy, kappa, MAE (Mean Absolute Error), and the amount of time to build the model. This study stated that Support Vector Machine (SVM) using kernel universal pearson VII is giving the accuracy of 98,81% compared than the other algorithm and the Mean Absolute Error value of 0,012 [14].

In other study, which proposed the analysis of risk factor of covid-19 and identifying the region with high risk in Fars, Iran uses an algorithm of machine learning which was based on geographic information system (GIS), and the use of SVM model. This research uses a validation technique of crossing the curve of ROC-AUC to validate the map of covid-19 risks from the SVM model. After the validation is done, they use the last 30% testing dataset which were gathered on 20th March 2020, which then resulting that the SVM model reaches the AUC value of 0,786 and the fault standard of 0,031 where it shows a good prediction. Which then retested on an active case location on 29th March 2020, which give the AUC value increased to 0,799 and it also proves more stable prediction on the data of 10th April 2020 and it shows that the accuracy of the model was increased to 86,6% with the AUC value of 0,868 [15].

The study of predict covid-19 using SVM approach. This research built a model in predicting people that has been exposed by covid-19 or not using the model. In this research the SVM model was tested to a different kind of frequencies and many data. The dataset that was used was 200 records with 8 attributes. The prediction classification is divided into 3 classes, which are low infection, high infection, and no infection. The reason of choosing this SVM model is that because they use a kernel trick to divide the input of N-low dimension and N-high dimension which is inseparable to be separable. SVM can use hyperplane to divide the data linearly. Where the hyperplane divides every class and makes it to be in distance as far as possible [16]. The researchers also took the value parameter of C = 10, to find the hyperplane with the least boundary to classify the infection with least inaccuracy and with higher accuracy. From this research it is found that the SVM model has an accuracy to 87% on class with low infection rate, and it is stated that the method was found to be efficient in predicting covid-19. On the other hand, the two other classes, doesn't have a good accuracy because the behavior of the covid-19 variable which is difficult to shows signs so that the uncertainty of the data collection caused the accuracy to be low [16].

Research on dynamic monitoring of covid-19 and the prediction of Space-Temporal real time with the approach of machine learning was conducted by [6]. This research uses the prediction of spatio-temporal for the period of 25th – 27th of May 2020 for the state of Pernambuco and Brazil. This research is also using the system of COVID-SGIS to combine the Public Health data which then to be distributed to the Public Health System by considering the geographic feature and the amount of time spend to make the prediction. This study, uses 4 methods of Regression, which are Linier Regression, Support Vector Machines (where kernel polynomial and RBF were used), Multilayer Perceptrons, and Random Forests. This study stated that the Linier Regression has the best prediction among the others with the Correlation Coefficient > 0,99 and the RMSE < 4% for Pernambuco and 5% for Brazil [6].

III. RESEARCH METHOD

The system that will be built is a classification process of Covid – 19 spreads using the method of Support Vector Machine (SVM). The flow of the system that will be built in this research is shown in Fig. 1.

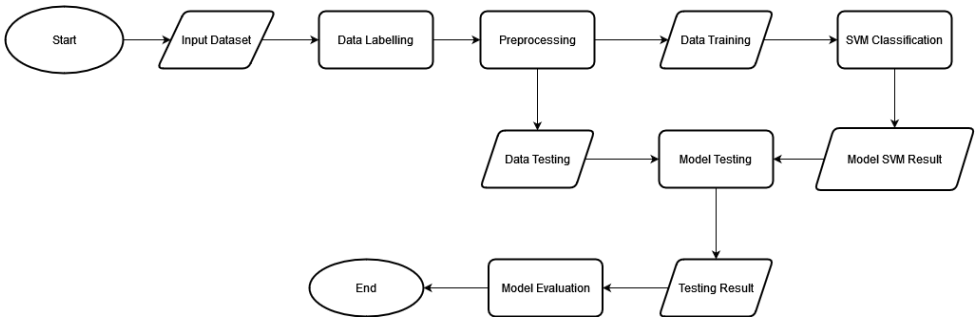

Fig. 1. System Design Flow

A. Dataset

The dataset used in this research are a data time series of Covid – 19 of every single ward in Bandung with the time series from November 2020 – December 2021. The data time series of Covid – 19 that were gathered have 23 attributes which are, name of the ward, total population of men, total population of women, rainfall pattern, sunshine pattern, the temperature average, the temperature minimum, and maximum, population based on their school degree, vaccination dosage, the obedience of mask usage, and the obedience of keeping distance. The datasets that were gathered from the Public Health Office of Bandung. These datasets were going to be divided into two, with the proportion of 80:20 which is 80% for training and 20% for testing. The training data will be used to build the model, while the testing data will be used to measure the performance of the model created. The attribute that will be used in this research for the dataset will be shown in Table 1.

TABLE I
DATA ATTRIBUTE

Variable	Attribute
X_1	Village
X_2	Male Population
X_3	Female Population
X_4	Rainfall (mm)
X_5	Sunshine (%)
X_6	Average Temperature (°C)
X_7	Maximum Temperature (°C)
X_8	Minimum Temperature (°C)
X_9	Not yet in school
X_{10}	Not yet graduated from elementary school
X_{11}	Elementary school graduate
X_{12}	Junior high school
X_{13}	High school
X_{14}	Diploma I and II
X_{15}	Diploma III
X_{16}	Bachelor's Degree / Diploma IV
X_{17}	Master's Degree
X_{18}	Doctoral degree
X_{19}	Vaccination Dose1
X_{20}	Vaccination Dose2
X_{21}	Vaccination Dose3
X_{22}	The obedience in mask wear
X_{23}	The obedience in keeping distance
Y	Class DA

B. Data Labelling

Binning method is used to refine the value in a data, with the concept of sorting the value with the value around it [21]. In this research the Binning method that was used is equal depth (or frequency) binning, where the value will be divided into N interval which consists with approximately the amount of the sample. This will be resulting the data class to be balanced so that the data labelling will be ready to be used. The data labelling will be shown in Table II.

TABLE II
CLASS LABELLING

Interval	Class	Label	Total
confirmed cases < 52	Low	0	704
52 ≤ confirmed cases < 196	Medium	1	705
confirmed cases ≥ 196	High	2	705

C. Preprocessing

Preprocessing stage is one of the important steps to prepare the dataset so that it can be ready to be used. In general, data will consist of flaws, which can be missing value, data redundant, outliers, or the format that is not suitable for the system. Because of that reason it is important to do the data preprocessing step. The preprocessing step will be consisting of a few process [18], which:

1) *Handling Missing Value*: Missing value is a condition where the data in an attribute is gone or not available, which has to be handled.
2) *Normalization:* Normalization is a process of making a variable to be a numeric type data which has a range or scale from 0 – 1[18]. The formula of the minmax equality used in this research can be seen in equation (1) [25]

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

D. Support Vector Machine Classification

Support Vector Machine (SVM) is one of the approaches of algorithm classification (Support Vector Classification) and regression (Support Vector Regression) which are a supervised machine learning. Support Vector Machine is used to identify hyperplane in a feature quantity or an N – dimension [11],[12],[19], which main purpose is to classify the data dots as clear as possible. In this research the classification process will be using the model of Support Vector Machine Classification. With that reason, we need the input x to maximize the width w which can be acquired with the formula of equation (2) [25]:

$$f(x) = w.x + \mathrm{b} \quad (2)$$

1) SVM Hyperplane:

The first approach that SVM do is finding the best hyperplane by maximizing the distance between class [25]. To determine the optimal hyperplane results, the right value of parameter C is needed, which cannot be too small and too large. where the equation can be seen in equation (3)[25].

$$min\frac{1}{2} \parallel \omega \parallel^2 + C\sum_{i=1}^{n} \varsigma_i \quad (3)$$

Hyperplane is a function that can be used as a divisible line for two data from two classes [24], where the equation can be seen in equation (4)[12]:

$$y = m \times x + c \quad (4)$$

2) SVM Classifier:

In the process of predicting hyperplane, we have to use the hypothesis H_0 function [12], where the equation can be seen in equation (5)[12]:

$$h_0(x_i) = \begin{cases} +1, & \text{if } w \times x + C \geq 0 \\ -1, & \text{otherwise} \end{cases} \quad (5)$$

From both formulas above, the obtained hyperplane can be classified into class +1 and class -1.

SVM also has a few kernel functions that is often used which are Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid [24]. The kernel function is used to change the data into a higher dimension to help the dimension problem [24]. Where linear kernel chooses hyperplane and is one of the simple kernels. On the other hand, RBF kernel is used to map the sample in linear number to a higher dimension and a more appropriate multiclass. While polynomial kernel is used to train denormalization [24]. These are the functions of Linear, Polynomial, Sigmoid, and RBF kernel can be seen in the equation (6), (7), (8), and(9) [26], [27].

a) Kernel linear

$$K(x_i, x_j) = x_i^T \, x_j \quad (6)$$

Where x_i is the training data, x_j is the test data [27].

b) Kernel Polynomial

$$K(x_i, x_j) = x_i^T \, x_j \quad (7)$$

Where x_i is the training data x_j is the test data, d is degree of polynomial [27].

c) Sigmoid

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \quad (8)$$

Where x_i is the training data, x_j is the test data, r is coefficient [27].

d) Kernel Radial Basis Function (RBF)

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i^T - x_j\|^2\right), \gamma > 0 \quad (9)$$

E. Model Evaluation

Model evaluation is done to measure the performance and to determine the best model by evaluating the performance of the model that has been created. Model evaluation that has been in this research used confusion matrix which can also be called error matrix. Confusion matrix itself is a table matrix which can be picture the model classification performance in a sequence of testing data [7]. The confusion matrix can be shown in Table III.

TABLE III
CONFUSION MATRIX

Predicted Values	**Actual Values**	
	Positive (1)	**Negative (0)**
Positive (1)	TP	FP
Negative (0)	FN	TN

From Table III. There are 4 combinations for Prediction Value and Actual Value, which are:

- *True Positive* (TP) where the actual value is 1 and the prediction value is 1
- *False Positive* (FP) where the actual value is 0 and the prediction value is 1
- *False Negative* (FN) where the actual value is 1 and the prediction value is 0
- *True Negative* (TN) where the actual value is 0 and the prediction value is 0

These are a few performance metrics that will used, which are:

1) Accuracy

Accuracy is a ratio of a correct prediction from the entire data collection [7]. Where the accuracy equation can be seen in the equation (10) [14]:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

2) Precision

Precision is a ratio of a correct positive prediction compared to the entire positive prediction data [7], Where the precision equation can be seen in the equation (11) [14]:

$$precision = \frac{TP}{TP + FP} \tag{11}$$

3) Recall

Recall is a comparison of True Positive (TP) with the amount of the actual positive prediction data [7], Where the recall equation can be seen in the equation (12) [14]:

$$recall = \frac{TP}{TP + FN} \tag{12}$$

4) F1-Score

F1 Score is a harmonic average of precision and recall [7], Where the f1-score equation can be seen in the equation (13) [14]:

$$F1\ score = 2\left(\frac{precision \times recall}{(precision + recall)}\right) \tag{13}$$

IV. RESULTS AND DISCUSSION

In this research the dataset that was used is a dataset of covid-19 of towns in Bandung with the periodic time of November 2020 until December 2021. Where this dataset was divided into 2 parts, which 80% was used to be the training data, and 20% for the test data which will be used to test and calculate the performance the model algorithm that was built. From the model that was built, it is found that the comparison between kernels, using 2 parameters which are Gamma, and C, resulting with Gamma value of 1 and 0,1 and C with 1, 10, 100 value. The testing uses two combination of every parameter value which gives 6 combinations in every kernel. The test on parameter C and Gamma then will be used to find the optimal value of C and the value of Gamma from the classification [11], [24], [26], [27]. The result of the classification of the monthly case covid-19 with accuracy, precision, recall, and f1-score will be shown in the below table.

Table IV shows that the accuracy of each month from the parameter that was experimented gives the comparison between each kernel in SVM such as, Kernel Linier, RBF, Polynomial, and Sigmoid. From the table we can see the optimal value from the combination of C and Gamma parameter. From the table, we can see that there are a few months data that gives the highest accuracy with are July, August, and December 2021. Which in each month the value of Gamma and C parameter for each kernel resulting the accuracy with approximately 97% - 100% on those 3 months. On the other hand, the combination of C and Gamma parameter didn't give their best accuracy on other months. For example, on November 2020, the accuracy was 81% with the kernel RBF value of gamma = 1, and C = 1.

TABLE IV
EVALUATION OF MODEL ACCURACY

	Gamma	C	Nov 20	Dec 20	Jan 21	Feb 21	Mar 21	Apr 21	May 21	Jun 21	Jul 21	Aug 21	Sep 21	Oct 21	Nov 21	Dec 21
Kernel Linear	1	1	77%	55%	58%	52%	61%	71%	65%	48%	97%	97%	68%	84%	77%	100%
	1	10	71%	55%	62%	52%	65%	77%	61%	52%	97%	97%	61%	84%	77%	100%
	1	100	68%	48%	45%	55%	61%	68%	71%	48%	94%	97%	55%	84%	77%	100%
	0.1	1	77%	55%	58%	52%	61%	71%	65%	48%	97%	97%	68%	84%	77%	100%
	0.1	10	71%	55%	61%	52%	65%	81%	61%	52%	97%	97%	61%	84%	77%	100%
	0.1	100	68%	48%	58%	55%	61%	68%	71%	48%	94%	97%	55%	84%	77%	100%
Kernel RBF	1	1	81%	45%	58%	55%	65%	74%	68%	45%	97%	97%	68%	84%	77%	100%
	1	10	77%	55%	48%	52%	58%	77%	81%	55%	97%	97%	58%	87%	81%	97%
	1	100	68%	52%	42%	48%	55%	65%	71%	39%	97%	97%	55%	90%	71%	100%
	0.1	1	74%	45%	65%	52%	61%	71%	61%	48%	97%	97%	52%	84%	77%	100%
	0.1	10	77%	52%	65%	61%	58%	81%	65%	48%	97%	97%	68%	84%	77%	100%
	0.1	100	74%	48%	55%	58%	55%	74%	71%	52%	97%	97%	61%	84%	77%	97%
Kernel Polynomial	1	1	74%	55%	58%	61%	68%	61%	65%	55%	97%	97%	45%	84%	77%	100%
	1	10	74%	52%	42%	55%	65%	65%	58%	42%	97%	97%	58%	84%	81%	100%
	1	100	77%	52%	35%	58%	58%	48%	61%	39%	97%	97%	48%	81%	84%	97%
	0.1	1	58%	45%	52%	45%	58%	68%	61%	55%	97%	97%	48%	84%	77%	100%
	0.1	10	68%	55%	55%	42%	65%	71%	65%	55%	97%	97%	48%	84%	81%	100%
	0.1	100	71%	55%	58%	48%	65%	68%	65%	55%	97%	97%	48%	84%	77%	100%
Sigmoid	1	1	23%	19%	26%	39%	35%	16%	39%	35%	97%	97%	39%	84%	61%	100%
	1	10	35%	10%	26%	29%	29%	19%	32%	32%	97%	97%	39%	68%	45%	100%
	1	100	26%	10%	26%	26%	29%	19%	26%	32%	94%	94%	42%	61%	74%	97%
	0.1	1	71%	45%	58%	45%	61%	61%	61%	52%	97%	97%	52%	84%	77%	100%
	0.1	10	77%	55%	58%	52%	65%	68%	65%	45%	97%	97%	74%	87%	77%	100%
	0.1	100	58%	61%	58%	45%	45%	71%	65%	42%	97%	97%	55%	81%	74%	100%

Fig. 2. Histograms of Accuracy Result

Fig. 2 is a histogram that represents the accuracy results of each parameter that has been tested. From the histogram, we can see that the accuracy pattern on each month is rather up and down. There are a few months that gives the best accuracy, which are July 2021, August 2021, November 2021, and December 2021 are the ones with the accuracy of 97%, 97%, 90%, 84%, and 100%. While the accuracy on the other month like July comes up with the value of 55%. The high and low accuracy of the data can be caused by a few things, like the attribute of high positive confirm cases, low vaccination distribution, or the low obedience of civilians to do the protocols.

TABLE V
THE BEST RESULT OF EACH ACCURACY, F1-SCORE, PRECISION, AND RECALL ON EACH MONTH

	Accuracy	F1-Score	Precision	Recall
Nov 20	0.81	0.78	0.75	0.81
Dec 20	0.61	0.61	0.67	0.61
Jan 21	0.65	0.61	0.68	0.65
Feb 21	0.61	0.57	0.63	0.61
Mar 21	0.65	0.64	0.74	0.68
Apr 21	0.81	0.81	0.82	0.81
May 21	0.81	0.81	0.83	0.81
Jun 21	0.55	0.51	0.58	0.55
Jul 21	0.97	0.95	0.94	0.97
Aug 21	0.97	0.95	0.94	0.97
Sept 21	0.74	0.75	0.77	0.74
Oct 21	0.90	0.90	0.90	0.90
Nov 21	0.84	0.83	0.82	0.84
Dec 21	1	1	1	1

Table V shows the best result of each accuracy, f1-score, precision, and recall on each month. Where on December 2021 it is found that the accuracy value is 100%, with the precision and f1-score also recall with the best value of 100%, followed by July and August with the accuracy of 97%, and October with the value of 90% While on June 2021 was found to be the lowest, with the accuracy of 55%, precision 58%, f1-score 51%, and recall 55%. Based on Table IV, we can see that by implementing the method using 6 combination values from 2 parameters Gamma and C, the approach that produces the average value of accuracy, precision, and F-1 score for each month is RBF kernel.

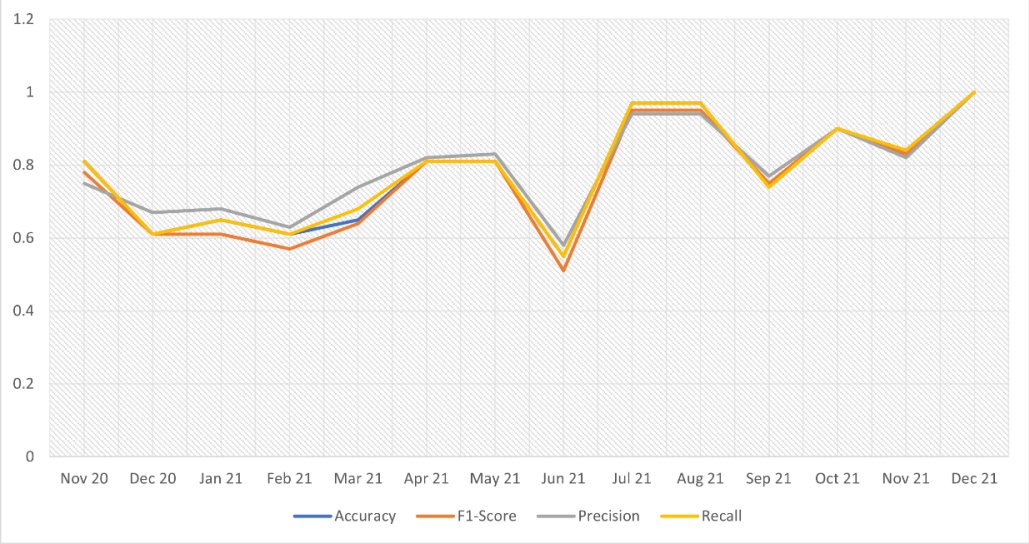

Fig. 3. Line Chart the Best Result of Each Accuracy, F1-Score, Precision, And Recall on Each Month

Fig 3. Shows the line chart of the best in accuracy, f1-score, precision, and recall on each month. From the line chart, we can see that in November 2020 the accuracy was found to be 81%, the f1-score 78%, precision 75%, and recall 81%, and then it started to go down in December 2020 until March 2021 with the average accuracy from 60% -65%. Then it started to start back up in April – May 2021 with the accuracy of 81%. In June 2021, it is found to have quite a significant decrease, which is going down to 55% and then it goes back up in August 2021 where the accuracy was found to be 97%, and lastly on December 2021 it is found to have 100% accuracy, f1 score 100%, precision 100%, and recall 100%. From the analysis obtained, the decrease and increase in accuracy, precision, f1-score and recall patterns every month is strongly influenced by attribute data. we can see that in June 2021 is the one month that gives the lowest accuracy which indicates that the mandatory use of mask, and to keep distance to each other in month June is rather low, also the amount of confirm active cases that is high.

Based on Table IV we can see that in July to December 2021, the use of 4 kernel functions and the same gamma value will obtaine the same accuracy value. This is because the attribute of those dataset was complete, for example the data of vaccination until 3 doses which is distributed since July 2021. Also the obligation to use mask and keeping distance with each other are rather increasing, so that in December 2021 the accuracy was found to be 100%.

V. CONCLUSION

In this study, researchers tried to analyze the performance of SVM to the classification of monthly cases of covid-19. Where the data used in this research was the data from Bandung in November 2020 until December 2021, which was acquired from the public health office of Bandung. That data includes vaccination data of covid-19, the obligatory to use mask in public and keeping up distance with each other. This research also uses the weather data from BMKG Bandung, which includes temperature, rainfall pattern, and sunshine pattern. In this research, 2 parameters were used, Gamma and C. Where the value of gamma used was 1 and 0.1, while C was 1, 10, and 100 which produces 6 combinations to be tested on every SVM kernels. It is found that in December 2021 the accuracy was 100%, followed by July and August 2021 with 97% and October with 90%. From the results it is also found that the kernel that produces the best accuracy, precision, f1-score, was from RBF. Which can be concluded, that SVM is a good method to classify the monthly data of covid-19 cases. We hope that in the future there will be more research regarding the issue with the addition of coef() parameter and degree.

ACKNOWLEDGMENT

The authors would like to thank Telkom University, and the Public Health Office of Bandung for the support of facilities and infrastructure in providing information and data, and I would also like to show my gratitude and thanks to the references I received that made this whole research done properly.

REFERENCES

[1] World Health Organization, “Covid-19 Situation Report,” *World Heal. Organ.*, vol. 31, no. 2, pp. 61–66, 2020.

[2] S. Kim and M. C. Castro, “Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020),” *Int. J. Infect. Dis.*, vol. 98, pp. 328–333, 2020, doi: 10.1016/j.ijid.2020.07.004.

[3] R. K. Singh *et al.*, “Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model,” *JMIR Public Heal. Surveill.*, vol. 6, no. 2, pp. 1–10, 2020, doi: 10.2196/19115.

[4] K. Yuki, M. Fujiogi, and S. Koutsogiannaki, “Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ’ s public news and

information ," no. January, 2020.
[5] D. Haritha, N. Swaroop, and M. Mounika, "Prediction of COVID-19 Cases Using CNN with X-rays," *Proc. 2020 Int. Conf. Comput. Commun. Secur. ICCCS 2020*, 2020, doi: 10.1109/ICCCS49678.2020.9276753.
[6] C. C. da Silva *et al.*, "Covid-19 Dynamic Monitoring and Real-Time Spatio-Temporal Forecasting," *Front. Public Heal.*, vol. 9, no. April, pp. 1–17, 2021, doi: 10.3389/fpubh.2021.641253.
[7] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari, "COVID-19 prediction and detection using deep learning," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 12, no. April, pp. 168–181, 2020.
[8] F. Tagliabue, L. Galassi, and P. Mariani, "The 'Pandemic' of Disinformation in COVID-19," *SN Compr. Clin. Med.*, vol. 2, no. 9, pp. 1287–1289, 2020, doi: 10.1007/s42399-020-00439-1.
[9] "COVID Live - Coronavirus Statistics - Worldometer." https://www.worldometers.info/coronavirus/ (accessed Oct. 23, 2021).
[10] V. Singh *et al.*, "Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine," *J. Discret. Math. Sci. Cryptogr.*, vol. 23, no. 8, pp. 1583–1597, 2020, doi: 10.1080/09720529.2020.1784535.
[11] A. K. Dubey, S. Narang, A. Kumar, S. M. Sasubilli, and V. García-Díaz, "Performance estimation of machine learning algorithms in the factor analysis of COVID-19 dataset," *Comput. Mater. Contin.*, vol. 66, no. 2, pp. 1921–1936, 2020, doi: 10.32604/cmc.2020.012151.
[12] S. A. Kumar, H. Kumar, V. Dutt, and ..., "COVID-19 Pandemic analysis using SVM Classifier: Machine Learning in Health Domain," *Glob. J. ...*, vol. 4, no. 1, pp. 30–38, 2020, [Online]. Available: http://technology.eurekajournals.com/index.php/GJADSIT/article/view/637
[13] E. Fayyoumi, S. Idwan, and H. Aboshindi, "Machine learning and statistical modelling for prediction of Novel COVID-19 patients case study: Jordan," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 122–126, 2020, doi: 10.14569/IJACSA.2020.0110518.
[14] C. N. Villavicencio, J. J. E. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using weka," *Algorithms*, vol. 14, no. 7, 2021, doi: 10.3390/a14070201.
[15] H. R. Pourghasemi *et al.*, "Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020)," *Int. J. Infect. Dis.*, vol. 98, pp. 90–108, 2020, doi: 10.1016/j.ijid.2020.06.058.
[16] S. Guhathakurata, S. Kundu, A. Chakraborty, and J. S. Banerjee, *A novel approach to predict COVID-19 using support vector machine*. Elsevier Inc., 2021. doi: 10.1016/B978-0-12-824536-1.00014-9.
[17] [BPS] Badan Pusat Statistik, "Berita resmi statistik," *Bps.Go.Id*, no. 27, pp. 1–52, 2019, [Online]. Available: https://papua.bps.go.id/pressrelease/2018/05/07/336/indeks-pembangunan-manusia-provinsi-papua-tahun-2017.html
[18] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Leaning," *Int. J. Comput. Sci.*, vol. 1, 2006, doi: doi.org/10.5281/zenodo.1082415.
[19] M. Buvana and K. Muthumayil, "Prediction of covid-19 patient using supervised machine learning algorithm," *Sains Malaysiana*, vol. 50, no. 8, pp. 2479–2497, 2021, doi: 10.17576/jsm-2021-5008-28.
[20] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.
[21] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.
[22] Tim Komunikasi Komite Penanganan Corona Virus Disease 2019 (Covid-19) dan Pemulihan Ekonomi Nasional, "Lonjakan Kasus Dampak Abaikan Protokol Kesehatan - Berita Terkini | Covid19.go.id." https://covid19.go.id/p/berita/lonjakan-kasus-dampak-abaikan-protokol-kesehatan (accessed Sep. 16, 2022).
[23] Tim Komunikasi Komite Penanganan Corona Virus Disease 2019 (Covid-19) dan Pemulihan Ekonomi Nasional, "Pemda Harus Mengidentifikasi Tren Kenaikan Kasus Jelang Nataru Untuk Mencegah Lonjakan | Covid19.go.id." https://covid19.go.id/artikel/2021/11/04/pemda-harus-mengidentifikasi-tren-kenaikan-kasus-jelang-nataru-untuk-mencegah-lonjakan (accessed Sep. 16, 2022).
[24] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," *Int. J.*

Adv. Res. Comput. Eng. Technol., vol. 1, no. 10, pp. 2278–1323, 2012.

[25] A. Property and F. Extension, "Buzzer Detection on Indonesian Twitter using SVM and Account Property," vol. 5, no. 158, pp. 663–669, 2022.

[26] A. Z. Praghakusma and N. Charibaldi, "Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi)," *JSTIE (Jurnal Sarj. Tek. Inform.*, vol. 9, no. 2, p. 88, 2021, doi: 10.12928/jstie.v9i2.20181.

[27] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," *J. Theor. Appl. Inf. Technol.*, vol. 12, no. 1, pp. 1–7, 2010.

[28] D. A. Pisner and D. M. Schnyer, "Support vector machine," *Mach. Learn. Methods Appl. to Brain Disord.*, pp. 101–121, 2019, doi: 10.1016/B978-0-12-815739-8.00006-7.