Classification of Covid-19 Spread Patterns in the Malang City with a Radial Basis Kernel Function Approach

Sri Harini[1], Angga Dwi Mulyanto[2], Hisyam Fahmi[3]

[1,2,3]Department of Mathematics, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang, Indonesia

ABSTRACT

Corona virus (COVID-19) is a new type of virus in humans that causes upper respiratory tract infections. This virus has behaviors such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) with clinical manifestations of infection such as acute respiratory disorders (fever, cough, and shortness of breath) which can cause pneumonia, lung syndrome, kidney failure, and even cause death. To prevent the spread of this virus, WHO has recommended the provision of vaccines to actively increase a person's immunity against Covid-19 attacks, so that when exposed to the disease, they will not get sick or only experience mild illness and will not become a source of transmission. This study classifies Covid-19 data in the Malang Region before vaccination and after vaccination with the Kernel Radial Basis Function approach. The data used in this study is a series of covid-19 data in the Malang City from November 2020 to December 2021. From the results of this study, it was found that the rate of decline in Covid-19 cases had appeared in July being December 2021 with an average accuracy rate of 97%. It can be concluded that the Kernel Radial Basis Function method is suitable for classifying the pattern of the spread of COVID-19 in the Malang Region.

Keywords: Classification, Covid-19 Spread Pattern, Kernel Radial Basis Function.

INTRODUCTION

- Indonesia is a country with more than 271,5 million population on 2020 is also been affected by the spread of SARS-CoV-2 virus [1]. Where in December 2020, Indonesia have still encounters quite the big spread of this virus, so that in 12th December 2020 the total case of was going up to 611.631 people, with 18.652 death case, and 501.376 cured case, and active case by 91.602 people [2]. Also on 2020, there are a few provinces in Indonesia with the most active cases, one of them is East Java. Where it is reported to be infected 400.066 people, which contributes about 9.39% to the national data [2]. Malang is one of the cities located in East Java which also reported to have quite the increase by 29th December 2021 where the confirm positive case was 15.652 people [3][9].

- 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 [4],[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].

Plot Data COVID-19 and Prediction Death COVID-19 case in Indonesia Before vaccination

Plot Data COVID-19 Case in Indonesia in March 2 is April 7, 2020

Prediction Death COVID-19 Case in Indonesia
Double Exponential Method
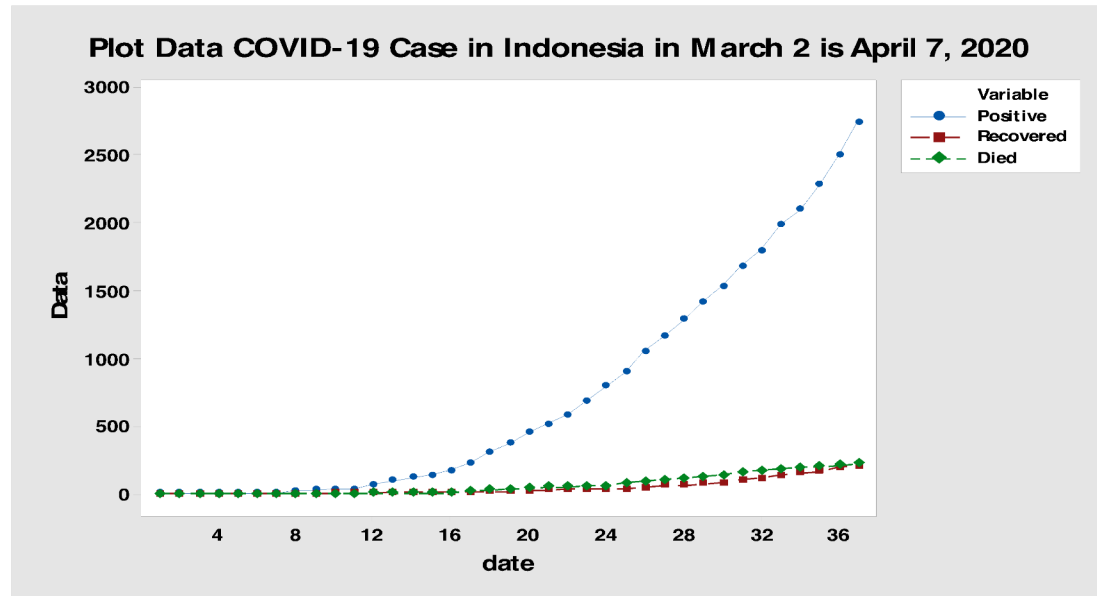
Using the time-series model approach, the pattern of COVID-19 data distribution behavior in Indonesia shows an exponential distribution pattern, where the addition of positive cases of COVID-19 increases significantly from day today.
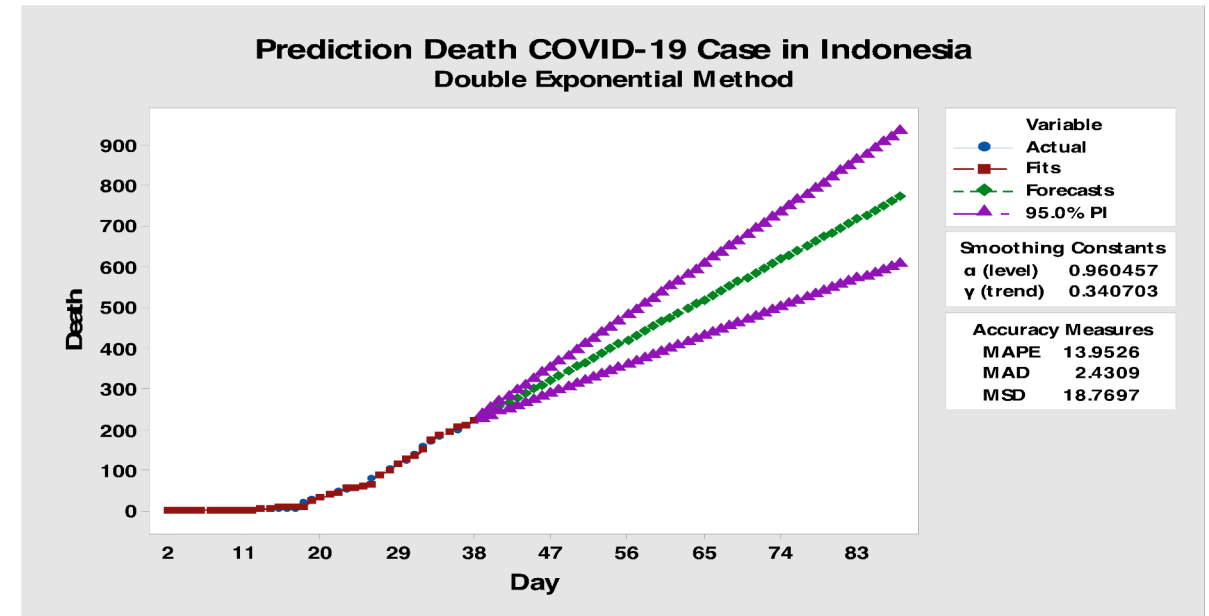
Using a 10 percent error rate, the estimated value of the parameter death patients the MAPE of 13.95 percent. MAPE value is greater than the error value set at 10%. the high death rate of COVID-19 patients due to the still limited number of health workers, supporting health facilities to prevent and Before vaccination

- In 2020, there is a research about the performance estimation the analysis of factor in Covid-19 dataset. The purpose of this research is to predict and classify anything related to Covid-19 uses logistic regression, ARIMA, Spasial Analysis, Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and atc. [8][10][11].

- Based on this research, in this study, Radial Basis Kernel Function Analysis will be applied to classify Covid-19 cases in Malang City. Testing the goodness of the model is based on the values of accuracy, precision and recall.

Research Method

1. Dataset

The dataset used in this research are a data time series of Covid–19 of every single ward in Malang 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 and woman, 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 Malang. 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.

Variable	Attribute
X_1	Village
X_2	Male Population
X_3	Female Population
X_4	Rainfall (mm)
X_5	Sunlight (%)
X_6	Average Temperature (^oC)
X_7	Maximum Temperature (^oC)
X_8	Minimum Temperature (^oC)
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

TABLE I

DATA ATTRIBUTE

TABLE II. CLASS LABELLING

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

2. ANALYSIS

Processing 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 consists of a few process 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 is :

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

3) *Kernel Radial Basis Function (RBF)*

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

3. 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)
Positif(1)	TP	FP
Negatif(0)	FN	TN

From table 3. 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]. Which can be obtained by the formula of:

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

2. *Precision*

Precision is a ratio of a correct positive prediction compared to the entire positive prediction data[7], which can be obtained by the formula of:

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

3. *Recall*

Recall is a comparison of True Positive (TP) with the amount of the actual positive prediction data[7], which can be obtained by the formula of :

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

RESULTS AND DISCUSSION

In this research the dataset that was used is a dataset of Covid-19 in Malang 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. The model that was built 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 *Kernel Radial Basis Function method* . 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 [8]. The result of the classification of the monthly case covid-19 with accuracy, precision and recall will be shown in the below table.

1. Accuracy

TABLE IV. EVALUATION OF MODEL ACCURACY

|	Gamma	C	Nov 20	Dec 20	Jan 21	Feb 21	Mar 21	Apr 21	May21	Jun 21	Jul 21	Aug 21	Sep 21	Oct 21	Nov 21	Dec 21
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%

Table IV shows that the accuracy of each month from the parameter that was experimented gives the Radial Basis Kernel Function model. 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 resulting the accuracy with approximately 97% - 100%.
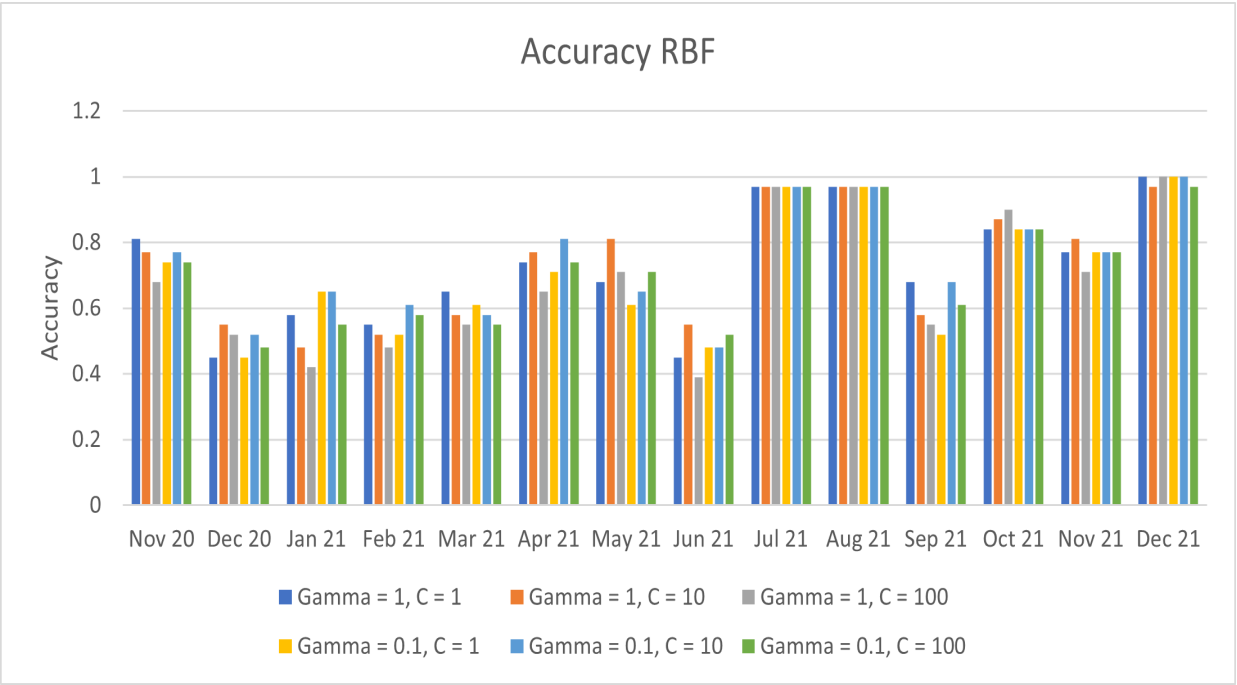
Accuracy RBF

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. There are a few months that gives the best accuracy, which are July 2021, August 2021, and December 2021 are the ones with the accuracy of 97%, 97% and 100%. While the accuracy on the other month like June 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 etc
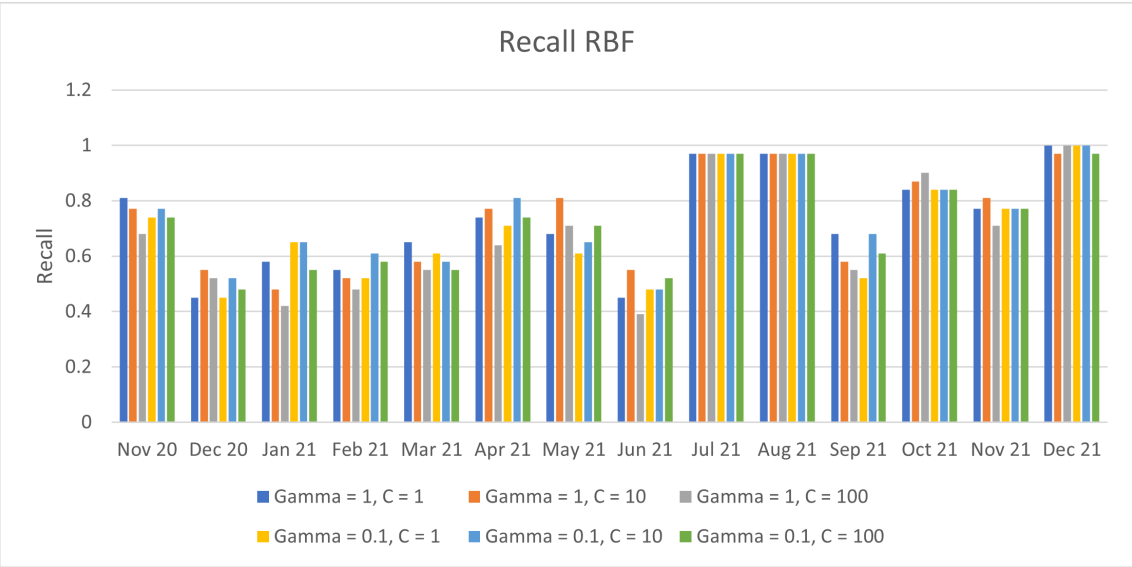
2. Precision

	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 RBF	1	1	75%	34%	59%	50%	62%	76%	60%	42%	94%	94%	73%	70%	60%	100%
	1	10	75%	46%	54%	55%	69%	78%	83%	48%	94%	94%	62%	89%	76%	100%
	1	100	69%	53%	46%	52%	64%	72%	73%	42%	94%	94%	56%	90%	77%	100%
	0.1	1	69%	20%	66%	45%	58%	58%	38%	46%	94%	94%	27%	70%	60%	100%
	0.1	10	72%	44%	63%	54%	55%	82%	52%	44%	94%	94%	73%	70%	60%	100%
	0.1	100	69%	37%	58%	52%	62%	74%	72%	47%	94%	94%	66%	70%	73%	100%
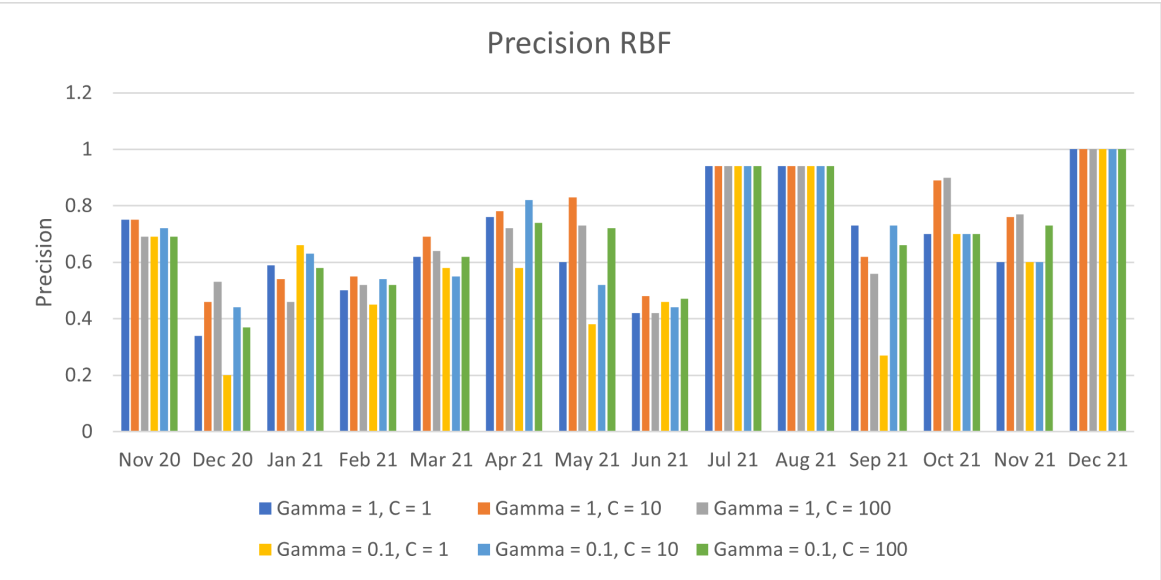
The result of precision from the experiment. It shows there are a few months that gives the highest precision which are on July, August, and December 2021. Where on those 3 months July 2021, August 2021, and December 2021 the value on each Kernel RBF gives an accuracy of 94% - 100%.

3. Recall

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%	64%	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%

Recall RBF

From the table we can see that there are a few months that gives the highest precisions which are July, August, and December 2021. Where on those 3 months gives the result of recall with a 100% accuracy on RBF kernel with the value of gamma = 1, and C = 100

CONCLUSION

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

The best result of each accuracy, precision, and recall on each month. Where on December 2021 it is found that the accuracy value is 100%, with the precision and recall with the best value of 100%, followed by July and August with the accuracy of 97%. While on June 2021 was found to be the lowest, with the accuracy of 55%, precision 58%, and recall 55%.

REFERENCES

[1]   "Jumlah Penduduk Indonesia 271,4 Juta Jiwa pada 2020, Mayoritas di Jawa." https://databoks.katadata.co.id/datapublish/2021/06/22/2714-juta-penduduk-ri-pada-2020-mayoritas-di-jawa (accessed Sep. 04, 2022).

[2]   "Kasus Corona Indonesia 611.631, Ini 5 Provinsi dengan Kasus Tertinggi Halaman all - Kompas.com." https://www.kompas.com/tren/read/2020/12/13/124500065/kasus-corona-indonesia-611.631-ini-5-provinsi-dengan-kasus-tertinggi?page=all (accessed Sep. 04, 2022).

[3]   "Kasus Covid-19 Meningkat, Bandung 2 Pekan Masuk Zona Merah." https://www.cnnindonesia.com/nasional/20201218224535-20-584094/kasus-covid-19-meningkat-bandung-2-pekan-masuk-zona-merah (accessed Sep. 04, 2022).

[4]   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.

[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]   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.

[9] 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

[10] M. N. Ikechukwu, E. Ebinne, U. Idorenyin, and N. I. Raphael, "Accuracy Assessment and Comparative Analysis of IDW, Spline and Kriging in Spatial Interpolation of Landform (Topography): An Experimental Study," *J. Geogr. Inf. Syst.*, vol. 09, no. 03, pp. 354–371, 2017, doi: 10.4236/jgis.2017.93022

[11] 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