JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2394-2403

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Comparison of Parametric and Non-parametric Forecasting Methods for Daily COVID-19 Cases in Malaysia

I Made Artha Agastya [a,*], Afrig Aminuddin [b]

[a] Informatics, Universitas Amikom Yogyakarta, Depok, Sleman, 55283, Indonesia
[b] Information System, Universitas Amikom Yogyakarta, Depok, Sleman, 55283, Indonesia
Corresponding author: [*]artha.agastya@amikom.ac.id

***Abstract*—Numerous research studies are currently examining various measures to control the transmission of COVID-19. One essential task in this regard is predicting or forecasting the number of infected individuals. This predictive capability is crucial for governments to allocate resources effectively. However, the most effective approach to handling time series problems between the parametric and non-parametric methods is unclear. The parametric method utilizes a fixed number of parameters to calculate the value. On the other hand, the non-parametric method increases its parameters along with the number of observations. To address the issue, we conducted a study comparing parametric and non-parametric models for time series forecasting, specifically using Malaysia's daily confirmed COVID-19 cases from 18/3/2020 to 30/12/2020. Since there have been limited comparisons of these models in time series forecasting, we believe our study is beneficial. We considered various models, including persistence, autoregression, ARIMA, SARIMA, single, double, and triple exponential smoothing, multi-linear regression, support vector regression, artificial neural networks (ANN), K-nearest neighbor regression, decision trees regression, random forest regression, and Gaussian processes regression models. Our study revealed significant characteristics of these methods, and we found that exponential smoothing methods were the most effective in capturing the level and trend of the data compared to other methods. Additionally, ANN had the least forecasting error among the machine learning methods. In conclusion, non-parametric methods are not suitable for predicting daily cases of Covid-19 in Malaysia. Enhancing the parametric methods will be preferable in the future.**

***Keywords*—COVID-19; parametric; non-parametric; machine learning; autoregression; smoothing.**

Manuscript received 9 May. 2023; revised 13 Jun. 2023; accepted 27 Jul. 2023. Date of publication 31 Dec. 2023.
International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

I. INTRODUCTION

COVID-19 is on the Public Health Emergency of International Concern (PHEIC) list [1] due to its tendency to induce life-threatening respiratory collapse and fast transmission. The COVID-19 cases were spread from a fish market in Wuhan [1]. The number of infected people increased exponentially. Based on a genome sequencing study, the bat has been suspected of being the natural carrier of COVID-19. It is suspected that the virus was carried by bats and transmitted to humans via unknown intermediate hosts [2]. Since discovering the COVID-19 infection, scientists from numerous disciplines have investigated this new virus. Predicting the forthcoming count of COVID-19 instances holds significant importance in facilitating public health decision-making and optimizing resource utilization to mitigate the morbidity and mortality linked to a global outbreak [3].

Based on Puaschunder [4], Artificial Intelligence (AI) is a promising technology that can assist healthcare providers in alleviating their workload. By combining AI and healthcare workers' knowledge, they can produce technology that has fast execution and might be more reliable than humans in certain healthcare activities. Healthcare organizations and physicians worldwide have implemented several ML and AI technologies to support their decisions in managing COVID-19. One of the studies by Krbaş et al. [5] used nonlinear autoregression neural network (NARNN) and Long-short-term memory (LSTM) methods to forecast the number of infections each day.

Sujath et al. [6] addressed how to forecast confirmed, deceased, and recovered patients using three methods, which are Linear Regression (LR), Multilayer Perceptron (MLP), and Vector Autoregression (VAR). Their study concluded that the MLP method provides more accurate predictions than the LR and VAR methods. According to [7], the LR model is a widely utilized COVID-19 analysis and forecasting algorithm because LR models require simple calculations. Forecasting on positive has confirmed cases using MLP and an adaptive network-based fuzzy inference system (ANFIS) [8]. Their model can achieve satisfying results without asserting any assumption of the epidemiological domain.

Moreover, Suzuki et al. [9] used a hybrid XGBoost that utilizes a Multi Output Regressor to forecast the cumulative infected people. The model's accuracy is 82.4%. Gothai et al. [10] implemented LR, Support Vector Machine (SVM), and Holt's Winter to predict the number of infected individuals in the next few dates. Based on the result, Holt's Winter or triple exponential smoothing (TES) model surpassed LR and SVM algorithms in predicting COVID-19 daily verified cases with an accuracy of 87%.

Predicting or forecasting the number of infected people is mandatory as it can help the government allocate resources such as medicine, quarantine places, and hospitals. The Susceptible, Exposed, Infectious, and Removed (SEIR) model [21] was utilized to forecast the timing of the peak infection during the COVID-19 epidemic in Malaysia using data from March 17-27, 2020. Based on the SEIR model, the peak of transmission of COVID-19 will occur in April 2020 and decrease significantly in the first week of July 2020. The forecasting result of three different methods, System Dynamic, SIR, and Curve Fitting Models, was implemented by Salim et al. [22]. The models can predict the peak of a pandemic with slightly different time frames.

Similar research has been done by Zamri et al. [23]. Three SEIR models were developed to examine the transmission rate of COVID-19. They found that the COVID-19 reproducing rate was reduced by 59 % during the Movement Control Order (MCO) phase. However, their forecasting result cannot be evaluated using error rate as the result is a proportion of the population. Additionally, the disadvantage of SEIR is the dependency on complete information of the parameters. If one of the parameters is missing, the SEIR model cannot be developed. Numerical simulations [24] were conducted to evaluate the influence of diverse control strategies, including personal protection, contact tracing, testing, and medication control measures, on the spread of the disease. Using numerical simulations, the implementation of each analyzed strategy has shown a substantial potential to decrease the occurrence of COVID-19 within the population. The ordinary differential equations (ODEs) method also cannot be evaluated using an error rate.

Parametric methods such as autoregression and exponential smoothing variants are quite popular in COVID-19 studies. For example, the Double Exponential Smoothing (DES) has been compared with the linear trend model for forecasting COVID-19 cases by Konarasinghe [25]. Based on the experiment, the linear trend model yields a lower mean absolute error (MAE) than DES. Rahman et al. [26] compared the performances of Average Percent Change (APC), Single Exponential Smoothing (SES), Double Exponential Smoothing (DES) models, and ARIMA models. The ARIMA model attained RMSE = 243.59 and MAPE = 27.7787. In contrast, the most excellent exponential smoothing model achieved RMSE = 243.648 and MAPE = 27.7795. The ARIMA [27] model was utilized to make projections of the daily confirmed cases of COVID-19 between 18 April 2020 and 1 May 2020. The ARIMA (0,1,0) can achieve a 16.01 Mean Absolute Percentage Error (MAPE). The researcher [28] forecasted Malaysia's daily COVID-19 cases using the ARIMA model, comparing three different time frames based on various Movement Control Order (MCO) periods. Tan et al. [29] implemented the Seasonal Autoregressive Integrated Moving Average (SARIMA) model to predict daily COVID-19 cases in Malaysia from 22 January 2020 to 5 September 2021. The best SARIMA model, with an RMSE of 73.374, MAE of 39.716, and Bayesian information criterion (BIC) of 8.656, indicated a declining trend in COVID-19 cases. These parametric methods were successfully applied to analyze Malaysia's COVID-19 cases.

The Adaptive Neuro-Fuzzy Inference System (ANFIS) model was implemented in forecasting the number of infected cases by Alsayed et al. [30], and it attained 0.041 Normalized Root Mean Square Error (NRMSE), 2.45% MAPE, and 0.9964 R^2. Nyoni et al. [31] employed an artificial neural network (ANN) to forecast daily infected cases in Malaysia. The ANN model is used to predict data from 26 March to 31 July 2021. The model has shown satisfying results using error rate evaluation. Ahmad *et al.* [32] forecasted using a linear regression (LR) model. The total confirmed cases are predicted using the number of recovery patients. Purwandari et al. [33] forecasted COVID-19 cases using the MLP, Neural Network Autoregressive (NNA), and Extreme Learning Machine (ELM). All methods achieved less than 12% MAPE for the days forecast. Norwawi [34] forecasted the COVID-19 outbreak in Malaysia from early January 2020 until late April 2020. The ANN model was created to predict the cumulative recovery of COVID-19 cases. Othman et al. [35] analyzed the results from the singular spectrum analysis (SSA) analysis of the number of daily confirmed cases. The SSA struggles in forecasting non-stationary data. Shaharudin et al. [36] developed a recurrent forecasting-singular spectrum analysis (RF-SSA) to forecast the COVID-19 outbreak. They achieved 11.2549 MAE for forecasting 32 days of data. The non-parametric method was mostly developed using machine learning as an inference. Because time series data is unsuitable for training machine learning models, researchers need to utilize the sliding window method to convert single-column data into tabular or multi-column data. The multi-column can be created by stacking the n-number lags.

A few researchers tried to compare the parametric and non-parametric methods. For example, the LR and TES models were implemented by Hasri et al. [37]. They used the Ministry of Health for Malaysia dataset. The models were evaluated using MAE and MAPE. Based on the experiment, the TES was more accurate than the LR model. A more varied comparison of methods has been carried out by Kamarudin et al. [38]. The researcher implemented four machine learning models along with the traditional statistical prediction analysis method called ARIMA. It was discovered that the MLP model outperformed all other models in accurately forecasting the number of new positive cases. On the other hand, the ARIMA model showed excellent performance in predicting recovered and death cases compared to the machine learning models. Model performance was evaluated using metrics such as RMSE and MAPE.

Based on the previous discussion, the non-parametric methods have not been implemented yet in forecasting COVID-19 cases in Malaysia. These facts encourage us to explore non-parametric methods. Moreover, the performance of parametric and non-parametric models is assessed to understand the characteristics of the models. Please refer to Table I for exploring the methods, period, and performance mentioned in the previous paragraph.

Movement Control Order (MCO) significantly reduced the rapid transmission of COVID-19 infection in Malaysia [11]. Malaysians were initially oblivious to this deadly virus, as they did not restrict entry from China in early 2020. However, this measure had a more substantial effect during the first MCO phase, followed by the second and third phases in which the infection curve flattened. Despite the various techniques used to forecast the number of infected people, forecasting methods have demonstrated adequate accuracy [12]–[17]. The parametric method has finite parameters to be used for calculating the value. In contrast, the non-parametric method does not need to define the parameters as the parameter increases by the number of observations or samples. It is still debated whether the most effective method to handle time series problems [18]–[20]. To the best of the authors' knowledge, a comprehensive analysis of both parametric and non-parametric forecasting methods in COVID-19 data has not been discussed. Consequently, this study aimed to assess and compare different forecasting models to determine the most suitable model for accurately predicting daily COVID-19 cases in Malaysia.

TABLE I
LITERATURE REVIEW

Author	**Method**	**Method Type**	**Data**	**Data Period**	**RMSE**	**MAE**	**MAPE**	**Other Evaluation**
Mahmud and Lim [21]	SEIR	Parametric	Daily Positive Case	17/3/2020 - 9/7/2020	-	-	-	-
Salim *et al.* [22]	CF SD SIR	Parametric	Cumulative Positive Case	22/1/2020 - 31/5/2020	-	-	0.000128 6.76 9.9811	-
Konarasinghe [25]	LTM DES	Non-parametric	Daily Positive Case	22/1/2020 - 19/11/2020	1.0966 1.3015	1.3337 1.0608	-	-
Singh *et al* [27]	ARIMA	Parametric	Daily Positive Case	22/1/2020 - 1/5/2020	-	-	16.01	BIC: 4.17
Edre *et al* [28]	ARIMA	Parametric	Daily Positive Case	18/3/2020 - 12/5/2020	35.69	27.86	28.65	-
Alsayed et al. [30]	ANFIS	Parametric	Daily Positive Case	22/3/2020 - 5/4/2020		-	2.79	NRMSE: 46.87 R^2: 0.9998
Othman et al. [35]	SSA	Parametric	Daily Positive Case	15/2/2020 - 27/6/020	-	-	-	-
Zamri et al [23]	SEIR	Parametric	Daily Positive Case	1/10/2020 - 29/3/2021	-	-	-	-
Nyoni et al. [31]	MLP	Non-parametric	Daily Positive Case	1/1/2020 - 31/7/2021	-	149.22	-	MSE: 64093.05
Ahmad *et al.* [32]	LR	Non-parametric	Cumulative Positive Case	1/7/2020 - 1/3/2021	-	3323.1 2018.36	-	-
Rahman et al. [26]	ARIMA SES DES	Parametric	Daily Positive Case	18/3/2020 - 30/12/2020	357.98 243.65 243.59	-	47.02 357.98 27.77	-
Abidemi et al. [24]	ODEs	Parametric	Cumulative Positive Case	3/3/2020 - 28/12/2020	-	-	-	-
Hasri et al.[37]	LR TES	Non-parametric parametric	Daily Positive Case	24/1/2020 - 16/8/2020	-	-	17.3 10.36	-
Kamarudin et al. [38]	MLP ARIMA	Non-parametric parametric	Daily Positive Case	22/1/2020 - 15/4/2020	40.08 42.07	-	20.92 24.89	-
Shaharudin et al. [36]	RF-SSA	Parametric	Daily Positive Case	25/1/2020 - 31/5/2020	-	11.25	-	MSE: 0.192 R^2: 0.9619
Norwawi [34]	MLR MLP	Non-parametric	Cumulative Recovery Case	25/1/2020 - 4/5/2020	-	-	-	R^2: 0.9997 MSE: 23.41
Tan *et al* [29]	SARIMA ARIMA	Parametric	Daily Positive Case	22/1/2020 - 3/10/2020	73.374 87.301	39.716 47.836	-	BIC: 8.656 8.982
Purwandari et al. [33]	NNAR ELM MLP	Non-parametric	Daily Positive Case	22/1/2020 - 13/6/2020	60.91 40.50 21.51	48.81 28.80 34.52	2.83 3.15 3.26	-

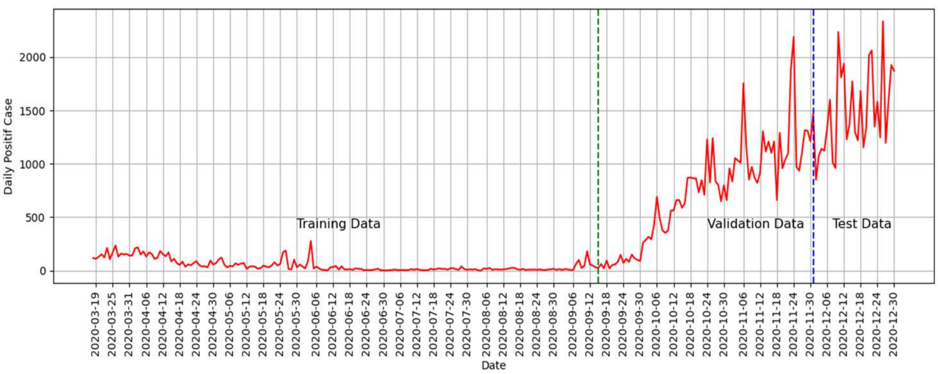


Fig. 1 Training, validation, and test data

TABLE II
LIST OF ACRONYMS AND DEFINITIONS

Acronyms	Definition
SEIR	susceptible, exposed, infectious, and removed
CF	curve fitting
SD	systems dynamic
SIR	susceptible, infectious, and removed
LTM	linear trend model
DES	double exponential smoothing
TES	triple exponential smoothing
ARIMA	autoregressive integrated moving average
ANFIS	adaptive neuro-fuzzy inference system
R-SSA	recurrent singular spectrum analysis
SSA	singular spectrum analysis
MLP	multi-layer perceptron
LR	linear regression
SES	single exponential smoothing
ODEs	ordinary differential equations
MLR	multiple linear regression
SARIMA	seasonal autoregressive integrated moving average
NNAR	neural network autoregression
ELM	extreme learning machine
MAPE	mean absolute percentage error
RMSE	root mean square error
MAE	mean absolute error
BIC	bayesian information criterion
MSE	mean square error
NRMSE	normalized root mean square error

II. MATERIALS AND METHODS

A. Dataset

We use a public database provided by the Malaysian government. From the sites, we can download information such as the flow of patients to/out of hospitals, daily recorded COVID-19 cases, daily tests by type, daily deaths due to COVID-19, and daily vaccinations at a state or country level. The link https://github.com/MoH-Malaysia/covid19-public is the dataset used in this study.

B. Experiment design

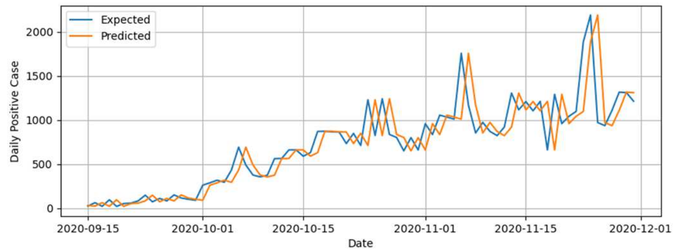
Based on the literature review, the complete experiment was conducted by Rahman et al. [26]. They split the time frame into three parts: training, validation, and testing data. We adopted their data divided as follows:

- Training date: 18/3/2020 - 14/9/2020
- Validation date: 15/9/2020 - 30/11/2020
- Testing date: 1/12/2020 - 30/12/2020

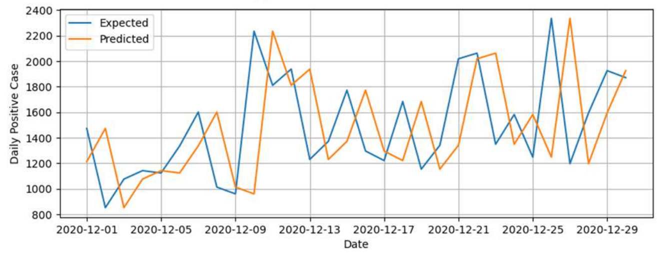
We use a walk-forward validation model that uses the most recent data for each prediction. For example, if we want to predict a seven-day confirmed case, we make one step ahead of forecasting seven times. The model is always updated by including recent data.

III. RESULT AND DISCUSSION

As shown in Fig. 1, The dataset was divided into three segments: training, validation, and test data. We tried to create a similar evaluation to that conducted by Rahman et al. [21]. The evaluation procedure is used one step ahead of prediction. The data is utilized to forecast the number of positive cases for the following day. The procedure is called walk-forward validation. Because of the characteristics of time series, we cannot use the infamous k-fold cross-validation or leave one out cross-validation (LOOCV). If we use both validation scenarios, the work becomes illogical as we predict the positive case number of a random date while future information is available. To evaluate the validation data, we use all training data. However, to assess the test data, we use both training and validation data as the traditional time series methods (i.e., ARIMA and SES) cannot handle missing values. We plot all models created by both parametric and non-parametric methods to show the prediction over expected values.



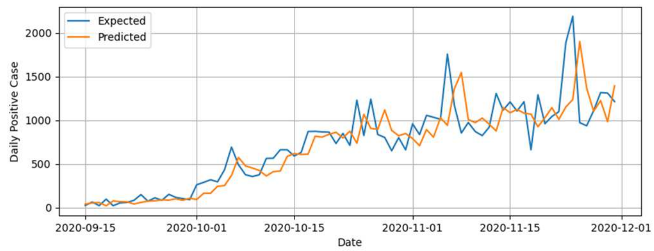
(a) Validation data



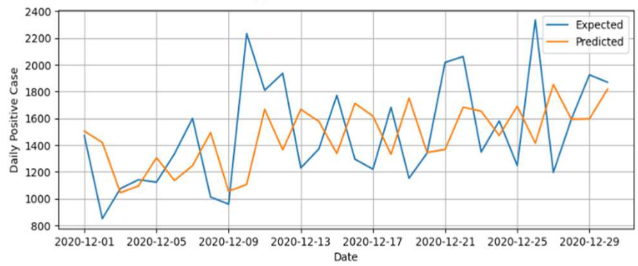
(b) Test data

Fig. 2 Persistence model's prediction on validation data and test data

In Fig. 2, the persistence model shows a shift in value from expected to predicted values. From Tables III and IV, the MAE of the persistence model is 168 and 404 for evaluation and test data, respectively. More complex methods must overcome the performance of the persistence model.

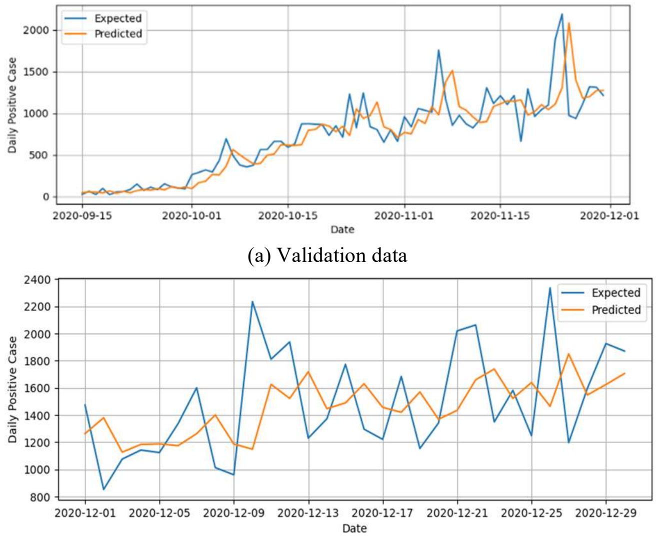


(a) Validation data



(b) Test data

Fig. 3 AR model's prediction on validation data and test data



(a) Validation data

(b) Test data

Fig. 4 ARIMA model's prediction on validation data and test data

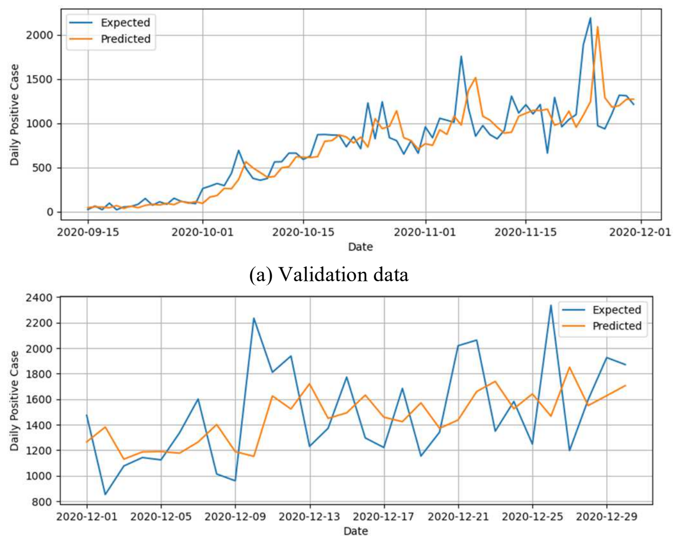
In the AR method, the predicted value can follow the validation and test data trend, as shown in Fig. 3. The predicted values are mostly below the expected values. We used grid search to find the best hyperparameter of the AR method. The training data shows that the optimal number lag is nine. So, we have ten coefficients, a linear regression model. The AR can get a four-person less incorrect prediction in the evaluation data.

In contrast, the AR can achieve 351 MAE in test data, as displayed in Table IV. The difference is more significant in the test data. The difference in the number of infected in the test period might be the reason. We can see the MAPE score for both validation and test period performance in Tables III and IV. The MAPE values are 30.61 and 28.0 for validation and test evaluations, respectively. The AR model created from training and evaluation data is superior to the model from training data only.

TABLE III
THE PERFORMANCE OF THE MODELS IN EVALUATING VALIDATION DATA

Method	hyperparameter	RMSE	MAE	MAPE
Persistence Model	-	272	168	32.71
Auto regression (AR)	(9, 0, 0)	271	164	30.61
Auto regression, Integration, and moving average (ARIMA)	(1, 1, 2)	269	162	28.4
Seasonal Auto regression, Integration, and moving average (SARIMA)	(1, 1, 2), (0, 0, 0, 0)	272	163	28.4
Single exponential smoothing (SES)	alpha: 0.3	242	**156**	**27.87**
Double exponential smoothing (DES)	alpha: 0.2, beta: 0.1, phi: 0.8	**241**	157	29.51
Triple exponential smoothing (TES)	alpha: 0.3, beta: 0.1, phi: 0.8, gamma: 0.1	242	156	27.87
Multi-linear regression (MLR)	-	274	172	30.88
Support vector regression (SVR)	kernel: linear, C:1	269	166	27.99
Artificial Neural Network (ANN)	(100,), max iter: 10,000	281	183	32.15
KNN regression (KNN-R)	n:5	256	157	31.63
Decesion Tree regression (DT-R)	criterion: squared error	361	225	39.08
Random forest regression (RF-R)	n_estimator: 100	273	164	34.21
Gaussian process regression (GP-R)	dot product and white kernel	274	172	30.62

Based on Fig. 4, the ARIMA has shown a similar pattern to the AR. As mentioned in the methods chapter, ARIMA uses differencing and residual error correction to enhance AR's capability. The ARIMA optimal model is created by performing a grid search, and the (1, 1, 2) model is selected. The ARIMA model result is better than the persistent and AR model. Based on Tables III and IV, the ARIMA can reduce the prediction error and achieve 162 and 323 MAE on validation and test data, respectively. Using one lag data can improve the autoregression model by making the time series stationary.

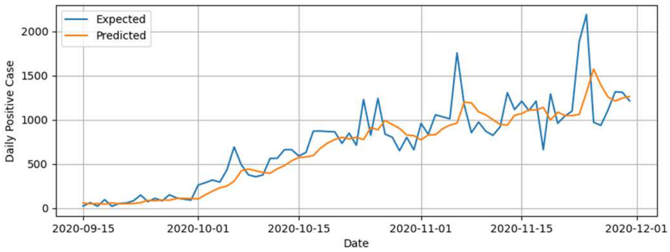


(a) Validation data

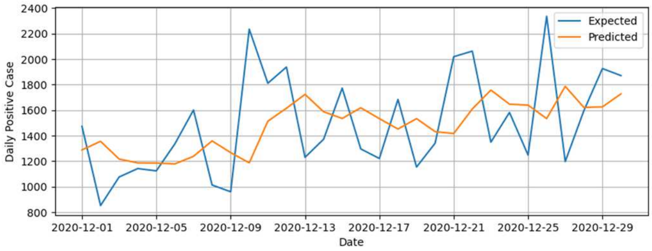
(b) Test data

Fig. 5 SARIMA model's prediction on validation data and test data

The SARIMA's best model is (1, 1, 2), (0, 0, 0, 0). The grid search mostly fails when executing the trend and seasonal model. It might be due to the time series not having a trend or seasonal component, as shown in Fig. 1. Because SARIMA uses the same model in autoregression, differencing, and residual model, it has almost the same result as ARIMA as shown in Fig 5.



(a) Validation data

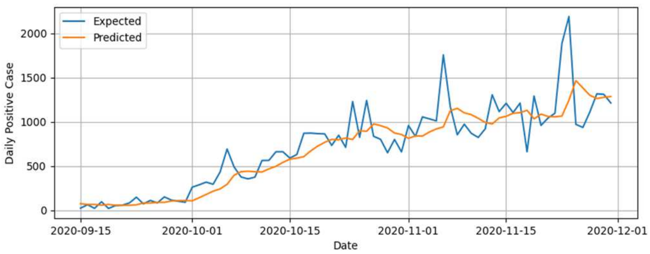


(b) Test data

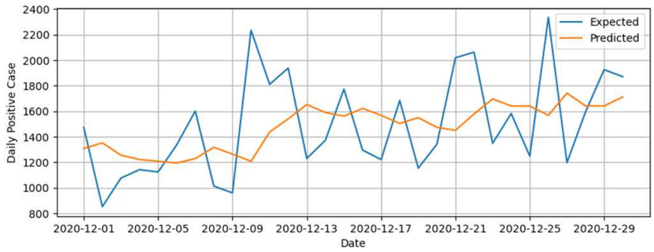
Fig. 6 SES model's prediction on validation data and test data

Based on Fig. 6 and Fig. 7, the SES and DES predicted values could mimic the evaluation and test data trend. Both models had similar prediction values. The DES has shown smoother values compared to the SES. The SES has a lower error compared to DES in evaluation data. The hyperparameter of SES is only alpha: 0.3. In contrast, the DES has alpha: 0.2, beta: 0.1, and phi: 0.8. The DES cannot give better results because it was trained on stationary data (i.e., no trend component is available). At the same time, the evaluation was conducted on data with a trend component.

In contrast, the test data evaluation can be predicted better as the combination of training and validation data is used as the historical data in the modeling process. The result is that DES can be slightly better than SES with one person error difference in MAE. Even though the TES is the most advanced method in the exponential smoothing variant, the error rate is like the SES, as TES has the same alpha as SES. As shown in Fig. 8, the TES has an identical value to SES.

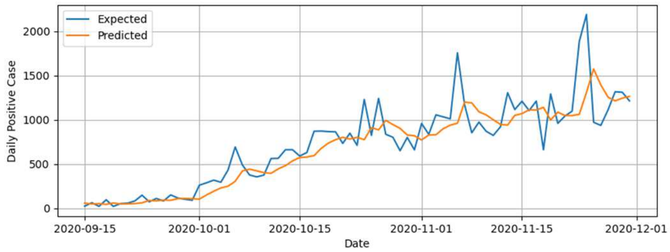


(a) Validation data

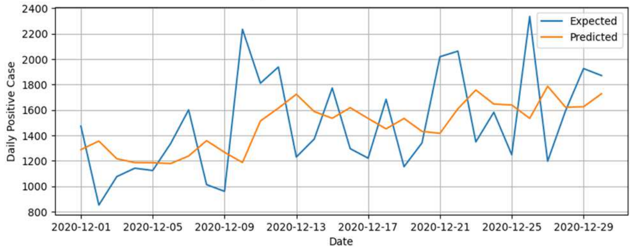


(b) Test data

Fig. 7 DES model's prediction on validation data and test data



(a) Validation data

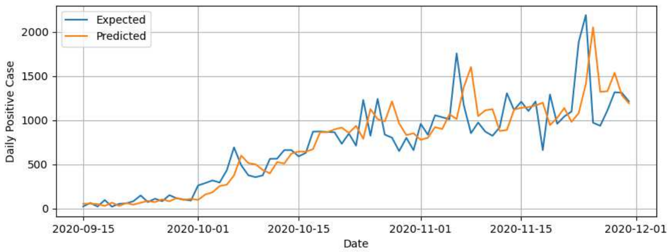


(b) Test data

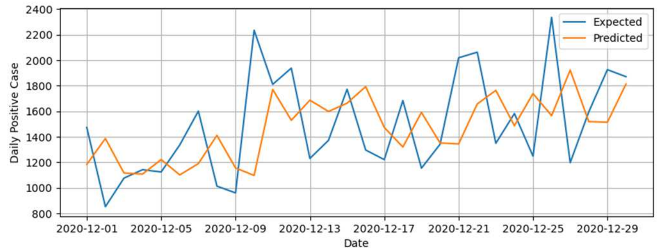
Fig. 8 TES model's prediction on validation data and test data

The MLR is the simplest machine learning-based model. The MLR uses five lags or previous data points to create supervised data. While the MLR model can capture the trend observed in the validation data, it also tends to capture the noise present in the data, resulting in higher errors compared to a persistence model, as shown in Fig. 9. It is a bad sign as a complex model suppose has better forecast capability than simply shifting the data to predict the next day's value. In contrast, the MLR model can outperform the persistence model on test data by achieving 343 MAE as shown in Table IV. The MLR can better predict the test data because the model is trained using a combination of training data and validation data. The validation data contains the trend of the dataset.

The SVR cannot predict better than MLR in predicting evaluation data. In contrast, the SVR has higher error rates than MLR in forecasting test data. The SVR is optimized based on the fitting line between the hyperplane and boundary line. The SVR cannot handle noise well like the original method named SVM. This is because the optimization does not directly calculate the error. Tables III and IV show that the SVR can achieve 269 and 437 RMSE on forecasting validation and test data, respectively.



(a) Validation data

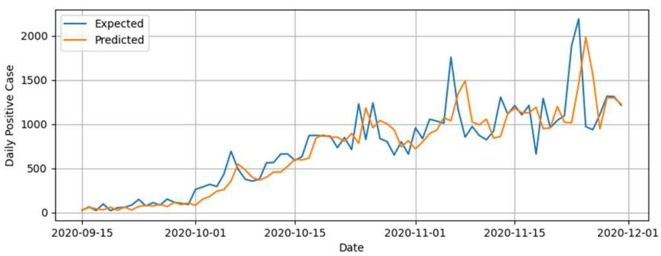


(b) Test data

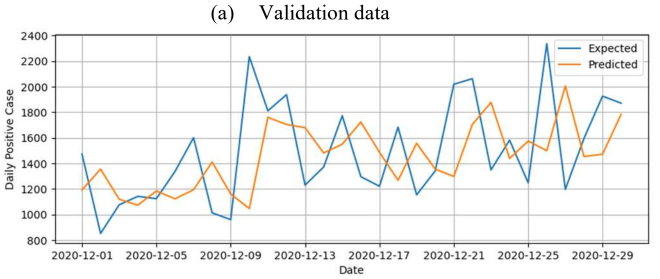
Fig. 9 MLR model's prediction on validation data and test data

TABLE IV
THE PERFORMANCE OF THE MODELS IN EVALUATING TEST DATA

Method	**hyperparameter**	**RMSE**	**MAE**	**MAPE**
Persistence Model	-	521	404	28
Auto regression (AR)	(9, 0, 0)	443	351	23.93
Auto regression, Integration, and moving average (ARIMA)	(1, 1, 2)	405	323	21.9
Seasonal Auto regression, Integration, and moving average (SARIMA)	(1, 1, 2), (0, 0, 0, 0)	404	323	21.95
Single exponential smoothing (SES)	alpha: 0.3	397	328	22.47
Double exponential smoothing (DES)	alpha: 0.2, beta: 0.1, phi: 0.8	**390**	**327**	**22.4**
Triple exponential smoothing (TES)	alpha: 0.3, beta: 0.1, phi: 0.8, gamma: 0.1	397	328	22.47
Multi-linear regression (MLR)	-	429	343	23.55
Support vector regression (SVR)	kernel: linear, C:1	437	345	23.47
Artificial Neural Network (ANN)	(100,), max iter: 10,000	426	349	24.62
KNN regression (KNN-R)	n:5	457	372	23
Decesion Tree regression (DT-R)	criterion: squared error	573	488	32.29
Random forest regression (RF-R)	n_estimator: 100	432	327	20.26
Gaussian process regression (GP-R)	dot product and white kernel	429	343	23.61



(a) Validation data

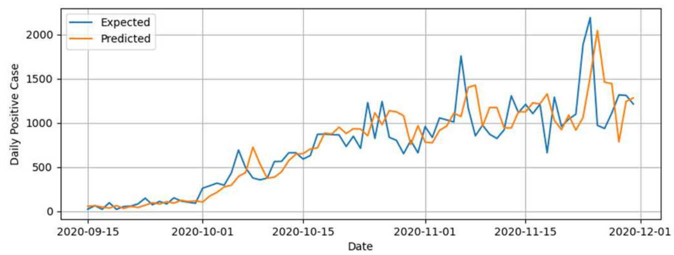


(b) Test data

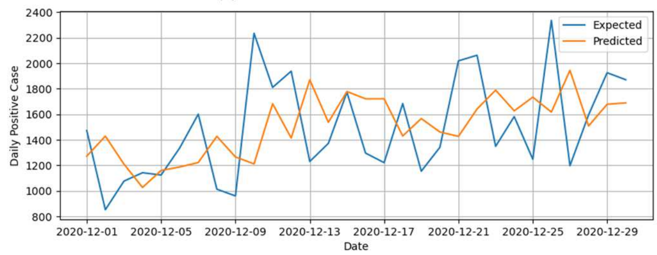
Fig. 10 SVR model's prediction on validation data and test data

The artificial neural network (ANN) can be used in both classification and regression tasks. The ANN can perform best when the historical or training data has a pattern. Based on Fig. 1, the training data does not show similar tendencies with the validation and test data. So, we got 23.55 accuracy, 1.15 higher than the best method as shown in Table IV.

The KNN regressor (KNN-R) is a simple non-parametric because the algorithm does not build the model by learning from historical data but instead compares the distance of the historical data against the new datum. The neighbor number influences the performance of the method significantly. We use grid search to overcome the hyperparameter hassle. The number of neighbors is five. The KNN regressor is also not suitable for predicting time series. As a result, it is not as significant as shown in Fig. 12, and based on Tables III and IV, the KNN regressor achieves 257 and 457 RMSE for predicting evaluation and test data, respectively.

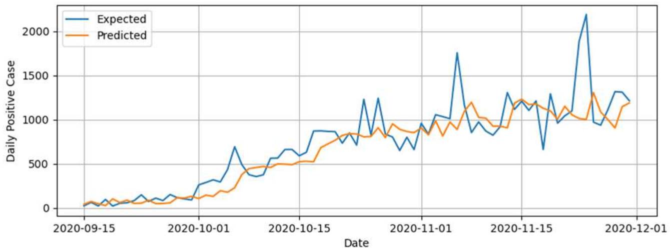


(a) Validation data

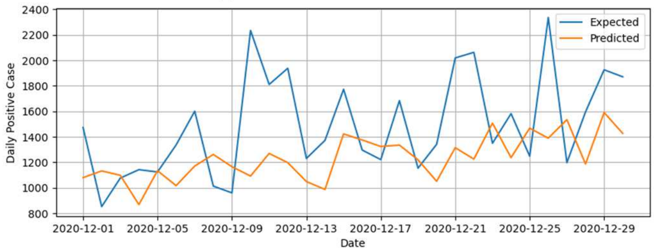


(b) Test data

Fig. 11 ANN model's prediction on validation data and test data



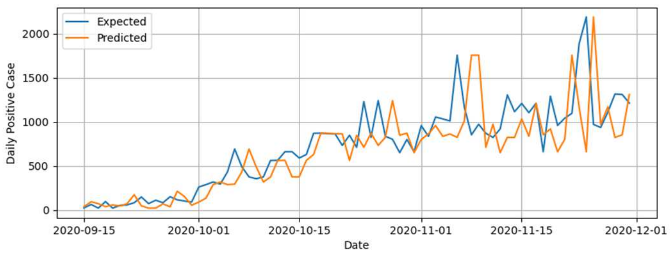
(a) Validation data



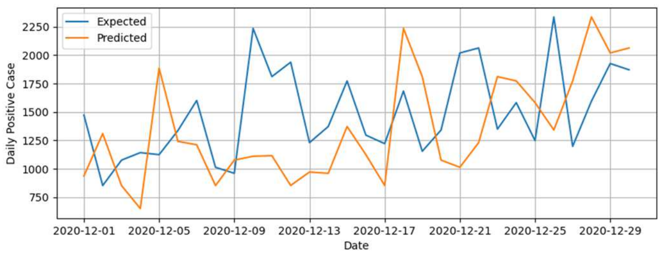
(b) Test data

Fig. 12 KNN-R model's prediction on validation data and test data

The decision tree regression (DT-R) is the worst model of all implemented methods. The decision tree regressor model cannot follow the trend of the DT-R. Instead, it mimics the data spike, as shown in Fig. 13. From Table III and Table IV, the DT-R achieved 361 and 573 RMSE on predicting evaluation and test data. The model cannot outperform the persistence model, which means it is not worth further exploration.



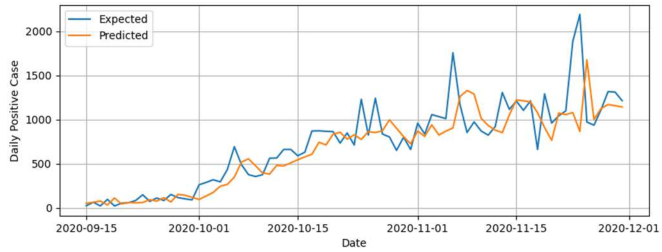
(a) Validation data



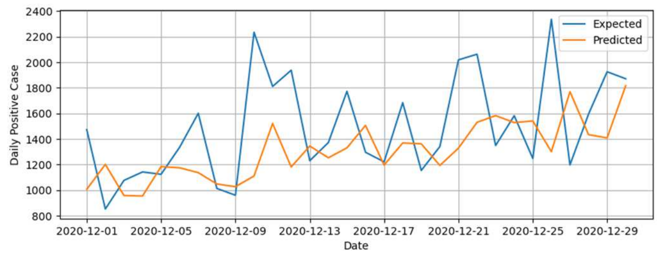
(b) Test data

Fig. 13 DT-R model's prediction on validation data and test data

Random forest regression (RF-R) has utilized several decision tree regressions that are trained on several samples and features in random selection. The RF-R can solve the high variance problem of DT-R. Using grid search, the RF-R has the best result using 100 decision trees. The RF-R achieved 164 and 327 MAE in predicting evaluation and test data. Based on Fig. 13 and Fig. 14, the RF-R has shown a significant improvement compared to DT-R. The RF-R prediction can follow the trend of the evaluation and test data.



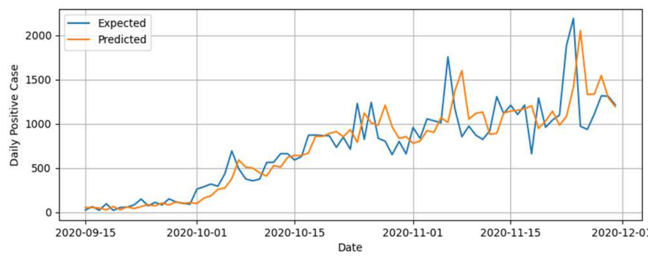
(a) Validation data



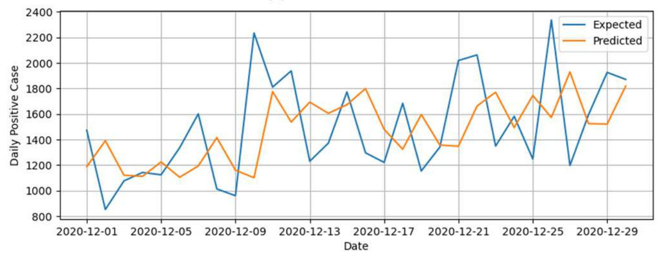
(b) Test data

Fig. 14 RF-R model's prediction on validation data and test data

The Gaussian process regression (GP-R) uses a dot product and white kernel. The GP-R can almost outperform the non-parametric methods. The GP-R can predict positive cases even though the trend is available in the data. As shown in Fig. 15, the evaluation and test performances are 164 and 343 MAE. The Gaussian process prediction can follow the trend of validation and test data.



(a) Validation data



(a) Test data

Fig. 15 GP-R model's prediction on validation data and test data

Based on those results, the parametric method, primarily Holt's, performs very well compared to non-parametric methods. The best non-parametric method is KNN-R, with a 23% error rate. In comparison with the best method, i.e., DES, the difference is 0.6%.

IV. Conclusion

The study aims to analyze parametric and non-parametric methods in forecasting daily COVID-19 cases in Malaysia. We found that Holt's methods can produce the finest model. Moreover, the parametric models are better than non-parametric models. Machine learning methods mostly make non-parametric models. It has shown that machine learning is not the answer to all problems. Therefore, it is suggested that parametric forecasting approaches such as ARIMA and exponential smoothing methods be explored before implementing machine learning or deep learning methods.

Future researchers are encouraged to explore the implementation of deep learning forecasting models to assess their performance in various contexts. Additionally, multivariate forecasting can be explored to know the performance of multivariate forecasting methods compared to univariate forecasting.

ACKNOWLEDGMENT

This research has received support from Universitas Amikom Yogyakarta in the form of a fundamental research grant. (grant number: RP-1660631307000).

REFERENCES

[1] A. A. Al-Qahtani, "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): Emergence, history, basic and clinical aspects," Saudi Journal of Biological Sciences, vol. 27, no. 10, pp. 2531–2538, Oct. 2020, doi: 10.1016/j.sjbs.2020.04.033.

[2] Y.-R. Guo et al., "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status," Military Medical Research, vol. 7, no. 1, Mar. 2020, doi: 10.1186/s40779-020-00240-0.

[3] V. J. Jayaraj, S. Rampal, C.-W. Ng, and D. W. Q. Chong, "The Epidemiology of COVID-19 in Malaysia," The Lancet Regional Health - Western Pacific, vol. 17, p. 100295, Dec. 2021, doi: 10.1016/j.lanwpc.2021.100295.

[4] J. M. Puaschunder, "The Potential for Artificial Intelligence in Healthcare," SSRN Electronic Journal, 2020, doi: 10.2139/ssrn.3525037.

[5] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches," Chaos, Solitons & Fractals, vol. 138, p. 110015, Sep. 2020, doi: 10.1016/j.chaos.2020.110015.

[6] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India," Stochastic Environmental Research and Risk Assessment, vol. 34, no. 7, pp. 959–972, May 2020, doi: 10.1007/s00477-020-01827-8.

[7] N. Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review," Informatics in Medicine Unlocked, vol. 24, p. 100564, 2021, doi: 10.1016/j.imu.2021.100564.

[8] H. Zakiyyah and S. Suyanto, "Prediction of Covid-19 Infection in Indonesia Using Machine Learning Methods," Journal of Physics: Conference Series, vol. 1844, no. 1, p. 012002, Mar. 2021, doi: 10.1088/1742-6596/1844/1/012002.

[9] Y. Suzuki, A. Suzuki, S. Nakamura, T. Ishikawa, and A. Kinoshita, "Machine learning model estimating number of COVID-19 infection cases over coming 24 days in every province of South Korea (XGBoost and MultiOutputRegressor)," *medRxiv*, 2020.

[10] E. Gothai, R. Thamilselvan, R. R. Rajalaxmi, R. M. Sadana, A. Ragavi, and R. Sakthivel, "Prediction of COVID-19 growth and trend using machine learning approach," *Mater. Today Proc.*, 2021.

[11] M. Amiruzzaman, M. Abdullah-Al-Wadud, R. M. Nor, and N. A. Aziz, "Evaluation of the Effectiveness of Movement Control Order to Limit the Spread of COVID-19," Annals of Emerging Technologies in Computing, vol. 4, no. 4, pp. 1–9, Oct. 2020, doi: 10.33166/aetic.2020.04.001.

[12] A. H. Elsheikh, A. I. Saba, H. Panchal, S. Shanmugan, N. A. Alsaleh, and M. Ahmadein, "Artificial Intelligence for Forecasting the Prevalence of COVID-19 Pandemic: An Overview," Healthcare, vol. 9, no. 12, p. 1614, Nov. 2021, doi: 10.3390/healthcare9121614.

[13] I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," Neural Computing and Applications, vol. 35, no. 33, pp. 23671–23681, Feb. 2021, doi: 10.1007/s00521-020-05626-8.

[14] W. S. Putra, A. Aminuddin, I. H. Purwanto, R. S. Kurnia, and I. A. Astuti, "Estimation of Transmission Rate and Recovery Rate of SIR Pandemic Model Using Kalman Filter," International Journal of Advanced Computer Science and Applications, vol. 13, no. 12, 2022, doi: 10.14569/ijacsa.2022.0131239.

[15] V. Geetha Mohan, A.-F. Mubarak Ali, M. Ariff Ameedeen, B. Lathakumary Vijayan, A. Aminuddin, and W. Widayani, "Predictive Models Using Supervised Neural Network for Pollutant Removal Efficiency in Petrochemical Wastewater Treatment," 2022 5th International Conference on Information and Communications Technology (ICOIACT), Aug. 2022, doi: 10.1109/icoiact55506.2022.9971929.

[16] Q. Waseem, W. Isni Sofiah Wan Din, A. Aminuddin, M. Hussain Mohammed, and R. F. Alfa Aziza, "Software-Defined Networking (SDN): A Review," 2022 5th International Conference on Information and Communications Technology (ICOIACT), Aug. 2022, doi: 10.1109/icoiact55506.2022.9972067.

[17] A. Cucus, L. B. Aji, A.-F. Bin Mubarak Ali, A. Aminuddin, and L. D. Farida, "Selection of Prospective Workers Using Profile Matching Algorithm on Crowdsourcing Platform," 2022 5th International Conference on Information and Communications Technology (ICOIACT), Aug. 2022, doi: 10.1109/icoiact55506.2022.9972155.

[18] A. Gautam and V. Singh, "Parametric Versus Non-Parametric Time Series Forecasting Methods: A Review," Journal of Engineering Science and Technology Review, vol. 13, no. 3, pp. 165–171, 2020, doi: 10.25103/jestr.133.18.

[19] R. Khaldi, A. El Afia, and R. Chiheb, "Forecasting of BTC volatility: comparative study between parametric and nonparametric models," Progress in Artificial Intelligence, vol. 8, no. 4, pp. 511–523, Jul. 2019, doi: 10.1007/s13748-019-00196-w.

[20] T. Kocsis, I. Kovács-Székely, and A. Anda, "Comparison of parametric and non-parametric time-series analysis methods on a long-term meteorological data set," Central European Geology, vol. 60, no. 3, pp. 316–332, Dec. 2017, doi: 10.1556/24.60.2017.011.

[21] A. Mahmud and P. Y. Lim, "Applying the SEIR Model in Forecasting The COVID-19 Trend in Malaysia: A Preliminary Study," *medRxiv*, p. 2020.04.14.20065607, 2020.

[22] N. Salim *et al.*, "COVID-19 epidemic in Malaysia: Impact of lockdown on infection dynamics," *medRxiv*, no. May, pp. 1–27, 2020.

[23] A. S. S. M. Zamri et al., "Effectiveness of the movement control measures during the third wave of COVID-19 in Malaysia," Epidemiology and Health, vol. 43, p. e2021073, Sep. 2021, doi: 10.4178/epih.e2021073.

[24] A. Abidemi, Z. M. Zainuddin, and N. A. B. Aziz, *Impact of control interventions on COVID-19 population dynamics in Malaysia: a mathematical study*, vol. 136, no. 2. Springer Berlin Heidelberg, 2021.

[25] K. M. U. B. Konarasinghe, "Forecasting COVID-19 Spread in Malaysia, Thailand, and Singapore," *J. New Front. Healthc. Biol. Sci.*, vol. 1, no. 2, pp. 1–13, 2020.

[26] H. A. Abd Rahman, M. A. Rahman, A. N. Rozaimi, and I. B. Zulnahar, "Forecasting of COVID-19 in Malaysia: Comparison of Models," 2021 IEEE International Conference on Computing (ICOCO), Nov. 2021, doi: 10.1109/icoco53166.2021.9673498.

[27] S. Singh et al., "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models," The Journal of Infection in Developing Countries, vol. 14, no. 09, pp. 971–976, Sep. 2020, doi: 10.3855/jidc.13116.

[28] M. A. Edre, M. A. ZA, and A. R. Jamalludin, "Forecasting Malaysia COVID-19 incidence based on movement control order using ARIMA and expert modeler," *IIUM Med. J. Malaysia*, vol. 19, no. 2, 2020.

[29] C. V. Tan et al., "Forecasting COVID-19 Case Trends Using SARIMA Models during the Third Wave of COVID-19 in Malaysia," International Journal of Environmental Research and Public Health, vol. 19, no. 3, p. 1504, Jan. 2022, doi: 10.3390/ijerph19031504.

[30] A. Alsayed, H. Sadir, R. Kamil, and H. Sari, "Prediction of Epidemic Peak and Infected Cases for COVID-19 Disease in Malaysia, 2020," International Journal of Environmental Research and Public Health, vol. 17, no. 11, p. 4076, Jun. 2020, doi: 10.3390/ijerph17114076.

[31] S. P. Nyoni, T. Nyoni, and T. A. Chihoho, "Forecasting Covid-19 New Cases in Malaysia," *Int. Res. J. Innov. Eng. Technol.*, vol. 5, no. 6, pp. 316–321, 2021, doi: https://doi.org/10.47001/IRJIET/2021.506056 Forecasting.

[32] W. M. A. W Ahmad et al., "Forecasting cumulative COVID-19 cases in Malaysia and rising to unprecedented levels," Bangladesh Journal of Medical Science, vol. 20, no. 3, pp. 504–510, Apr. 2021, doi: 10.3329/bjms.v20i3.52791.

[33] T. Purwandari, S. Zahroh, Y. Hidayat, S. Sukonob, M. Mamat, and J. Saputra, "Forecasting model of COVID-19 pandemic in Malaysia: An application of time series approach using neural network," Decision Science Letters, vol. 11, no. 1, pp. 35–42, 2022, doi: 10.5267/j.dsl.2021.10.001.

[34] N. M. Norwawi, *Sliding window time series forecasting with multilayer perceptron and multiregression of COVID-19 outbreak in Malaysia*. Elsevier Inc., 2021.

[35] A. A. A. Othman, Nurul ashikin Binti Aziz, N. A. Ahmad, M. H. Mohd, and S. I. M. Adam, "Analysing Trends and Forecasting of COVID-19 Pandemic in Malaysia using Singular Spectrum Analysis," *Mat. Malaysian J. Ind. Appl. Math.*, pp. 121–134, 2021.

[36] S. M. Shaharudin, S. Ismail, N. A. Hassan, M. L. Tan, and N. A. F. Sulaiman, "Short-Term Forecasting of Daily Confirmed COVID-19 Cases in Malaysia Using RF-SSA Model," Frontiers in Public Health, vol. 9, Jun. 2021, doi: 10.3389/fpubh.2021.604093.

[37] H. Hasri, S. A. M. Aris, and R. Ahmad, "Linear Regression and Holt's Winter Algorithm in Forecasting Daily Coronavirus Disease 2019 Cases in Malaysia: Preliminary Study," 2021 IEEE National Biomedical Engineering Conference (NBEC), Nov. 2021, doi: 10.1109/nbec53282.2021.9618763.

[38] A. N. Akramin Kamarudin, Z. Zainol, N. F. Abu Kassim, and R. Sharif, "Prediction of COVID-19 Cases in Malaysia by Using Machine Learning: A Preliminary Testing," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Mar. 2021, doi: 10.1109/widstaif52235.2021.9430222.