PLOS ONE

RESEARCH ARTICLE

Developing forecasting model for future pandemic applications based on COVID-19 data 2020–2022

Wan Imanul Aisyah Wan Mohamad Nawi[1], Abdul Aziz K. Abdul Hamid[1,2], Muhamad Safiih Lola[1,3]*, Syerrina Zakaria[1,3], Elayaraja Aruchunan[4], R. U. Gobithaasan[1,3], Nurul Hila Zainuddin[5], Wan Azani Mustafa[6,7], Mohd Lazim Abdullah[1,3], Nor Aieni Mokhtar[8], Mohd Tajuddin Abdullah[9,10]

1 Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia, **2** Special Interest Group on Applied Informatics and Intelligent Applications (AINIA) Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia, **3** Special Interest Group on Modeling and Data Analytics (SIGMDA), Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia, **4** Faculty of Science, Institute of Mathematical Sciences, Universiti Malaya, Kuala Lumpur, Kuala Lumpur Federal Territory, Malaysia, **5** Mathematics Department, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak Darul Ridzuan, Malaysia, **6** Faculty of Electrical Engineering & Technology, Universiti Malaysia Perlis, UniCITI Alam Campus, Sungai Chuchuh, Padang Besar, Perlis, Malaysia, **7** Advanced Computing (AdvCOMP), Centre of Excellence, Universiti Malaysia Perlis (UniMAP), Arau, Perlis, Malaysia, **8** Institute of Oceanography and Environment, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia, **9** Faculty of Fisheries and Food Science, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia, **10** Fellow Academy of Sciences Malaysia, Kuala Lumpur, Kuala Lumpur Federal Territory, Malaysia

* safiihmd@umt.edu.my

Citation: Wan Mohamad Nawi WIA, K. Abdul Hamid AA, Lola MS, Zakaria S, Aruchunan E, Gobithaasan RU, et al. (2023) Developing forecasting model for future pandemic applications based on COVID-19 data 2020–2022. PLoS ONE 18(5): e0285407. https://doi.org/10.1371/journal.pone.0285407

Editor: Ahmed Hamza Osman, King Abdulaziz University, SAUDI ARABIA

Received: August 28, 2022

Accepted: April 13, 2023

Published: May 12, 2023

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The publication is partially sponsored by the Research Management Office, Universiti Malaysia Terengganu (UMT). No additional external funding was received for this study.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Improving forecasting particularly time series forecasting accuracy, efficiency and precisely become crucial for the authorities to forecast, monitor, and prevent the COVID-19 cases so that its spread can be controlled more effectively. However, the results obtained from prediction models are inaccurate, imprecise as well as inefficient due to linear and non-linear patterns exist in the data set, respectively. Therefore, to produce more accurate and efficient COVID-19 prediction value that is closer to the true COVID-19 value, a hybrid approach has been implemented. Thus, aims of this study is (1) to propose a hybrid ARIMA-SVM model to produce better forecasting results. (2) to investigate in terms of the performance of the proposed models and percentage improvement against ARIMA and SVM models. statistical measurements such as MSE, RMSE, MAE, and MAPE then conducted to verify that the proposed models are better than ARIMA and SVM models. Empirical results with three real datasets of well-known cases of COVID-19 in Malaysia show that, compared to the ARIMA and SVM models, the proposed model generates the smallest MSE, RMSE, MAE and MAPE values for the training and testing datasets, means that the predicted value from the proposed model is closer to the actual value. These results prove that the proposed model can generate estimated values more accurately and efficiently. As compared to ARIMA and SVM, our proposed models perform much better in terms of error reduction percentages for all datasets. This is demonstrated by the maximum scores of 73.12%, 74.6%, 90.38%, and 68.99% in the MAE, MAPE, MSE, and RMSE, respectively. Therefore, the proposed model

can be the best and effective way to improve prediction performance with a higher level of accuracy and efficiency in predicting cases of COVID-19.

Introduction

The city of Wuhan in the province of Hubei, China is etched in the folds of history for being the first place of the spread of the Coronavirus disease (COVID-19), due to severe acute respiratory syndrome. The World Health Organisation (WHO) on 31st January was firstly declared that COVID-19 as a "Public Health Emergency of International Concern" [1]. Originally, it was thought that the virus has been derived from a seafood market in Wuhan. However, on 11 January 2020 the genetic sequence of which was overtly shared by China through human-to-human contacts have driven its rapid spread with a total of 9,129,146 confirmed cases, including 473,797 deaths across the globe until June 24, 2020 [2]. Nonetheless, the COVID-19 pandemic has infected more than 151 million of the humans all over the world and caused 3 million deaths as of May 1, 2021. The countries like USA, Brazil, Russia, Spain, UK, Italy, France, Germany, China, India, Iran, and Pakistan become the most affected from COVID-19. The first few COVID-19 cases were reported in Malaysia on 24th January 2020 were detected from Chinese tourists entering the country from Singapore [3]. In the early stage, only in single digit of daily cases were reported, however it had increased to 235 by 26th March [4]. The number of daily cases in Malaysia were continued to rise exponentially hitting around 20,000 by August 2021. The Malaysian government was declared the implementation of the Movement Control Order (MCO), Conditional MCO (CMCO) and Recovery MCO (RMCO) from 18th March to 12th May 2020, 13th May to 9th June, and 9th June to 31st December, respectively. All travelling and socio-economic activities (gatherings for religious and cultural occasions were not allowed) were restricted nationwide to keep new infections at bay and avoid overloading the country's healthcare system during this period. All government and private offices, and education institutions including transport hubs were closed and instructing citizens to stay at home and interstate travelling was banned with fines of up to RM10,000 for violators.

Since WHO declared as the outbreak of COVID-19 as a pandemic, a lot of effort have been attempts not only from government worldwide but effort also from medical institution are committed to finding vaccines and treatments to control the spread of the virus, statistical modelling particularly forecasting on the COVID-19 cases also have been extensively carried out by statisticians and health scientists to support the health system to inhibit the disaster of infection as well. In this scenario, the capability to pinpoint the growth rate more effectively at which the epidemic is spreading is very crucial to fight back and assist the governments mindfulness concerning society planning and policymaking to accurately deal with the consequences of the infection. Thus, the motivation behind this research compared to the existing research work, namely, (i) to develop the forecasting model that more accurate and efficient regarding the spread of COVID-19 in Malaysia, and (ii) to compare the performance of this novel model with ARIMAS and SVM. This model can assist the public health authorities for pre-emptive and preventive planning to curtail the impact of future pandemics.

During pandemic many studies have been carried out through different mathematical and statistical models to predict the spread of the COVID-19 pandemic. One of the most popular and widely time series forecasting models used to analyse and predict the spread of the disease is the ARIMA (p,d, q) model [5–7]. Forecasting daily new cases of COVID-19 was a difficult

undertaking because the cases were growing daily. In the first wave, the cases of COVID-19 pattern has been continuously increasing for some period then decline. However, for the second wave it seen to be increased again and some of the COVID-19 cases are difficult to predict. In this scenario, a few researchers predict COVID-19 pattern using ARIMA [8–15]. However, ARIMA model have a limitation where it's normally only can handle a linear time series data structure [16]. However, approximations by ARIMA models are inadequate in representing a barrier in time series forecasting for researchers particularly for nonlinear pattern [17]. Despite its superior performance, Support Vector Machines (SVM's) classification performance and classifier's generalisation ability are frequently impacted by the dimension or quantity of feature variables as mentioned by Lee [18] is used. As a sequence of the development of Vector Machines model, this process will be able to provide the accurate and efficient result in any case of prediction. The SVMs, which were first introduced by Vladimir Vapnik in 1995 [19] in the domain of statistical learning theory and structural risk minimization, have been shown to operate well on a variety of forecasting and classification issues. The SVMs could also cope with or address difficulties like nonlinearity, local minimum, and high dimension in which ARIMA model [16, 20–22]. SVMs models have recently been used to handle issues such as nonlinear, local minimum, and high dimension. SVMs can ensure higher accuracy for a long-term prediction compared to other computational approaches even in many practical applications. However, single SVM model as single ARIMA model also have some limitation where SVM model only can handle nonlinear data, instead of linear data. With the constrains of a single ARIMA and SVM models as well, in-dept analysis of time series forecasting, hybrid approaches become the best approach to overcome both limitations and it's a very significant impact in numerous fields due to their dynamic nature and capability to predict at a higher level of accuracy, efficiency, and precision. This approach is crucial due to issues that arise in time series forecasting where almost all real-world time series contain both linear and nonlinear correlation patterns between the data. Recently, the hybridization of forecasting methods has been used with great achievement to reach enhanced forecasting accuracy [16, 17, 20–26].

In terms the spread of COVID-19, the hybrid time series model approach is crucial in predicting the impact of COVID-19 outbreak and it has been shown to be successful in predicting COVID-19 [27–30]. Thus. this study aims (a) to propose the hybrid ARIMA -SVM models approach for produce better forecasting results where its capability to produce the best estimator, i.e., generating small error terms; (b) to investigate the performance of the proposed models by comparing with the ARIMA and SVM models using three daily cases of COVID-19 data in Malaysia which are daily new positive cases, daily new fatalities cases, and daily new recovered cases. In spite of recent advances in time series and in particular in COVID-19, the model building process does not include cases of COVID-19 specifically in Malaysia to assist the authorities in dealing with the spread of this outbreak by producing more efficient, accurate and precise forecast results in the future. Therefore, in this study rather than rely on conventional approaches to deal with the COVID-19 data, this study relies on intelligent-based prediction methods to better predict the future pandemic. According to Moore [31], the scenario for the next likely new pandemic of strain of bird influenza H7N9 virus, or a novel coronavirus. Despite the fact that future outbreaks are inevitable, however, this intelligent-based prediction methods can produce more efficient, accurate and precise forecasts for pre-emptive prevention medicinal procedures by the local health care authorities [32, 33]. The model can also be used to predict Coronavirus or bird flu in the future, especially in tropical rainforest countries like Malaysia. Additionally, the intelligent-based prediction methods will produce prediction models that are more accurate, precise, and efficient in predicting the dynamic spread of the virus in the future. Although, the vaccine is currently available and the number

of deaths worldwide is low, this model will be useful for making very accurate predictions if similar outbreaks occur in the future. As a result, the spread of COVID-19 can be predicted earlier so that better health facilities can be built, legislative measures can be taken, and economic losses, especially human losses, can be avoided.

The rest of this paper is organized as follows. Details of the method we used to develop our proposed model are discussed in materials and methods. Followed by a brief formulation of the hybrid ARIMA-SVM model used in this study. The performance of our proposed model based on three well-known COVID-19 case datasets is presented in the results and discussion. Finally, we conclude the paper and provide recommendations for further work.

Materials and methods

The ARIMA modelling. The Autoregressive Integrated Moving Average, The ARIMA (p, d,q) model is one of the families in time series forecasting that is commonly used for time series forecasting because of its flexibility with various categories of time series datasets [17]. It also expressly caters to a set of standard patterns in time series analysis, enabling an easy-to-use yet powerful way for creating accurate time series predictions However, limitations may occur with pre-assumptions due to the existence of a linear form that is a linear relationship between the future value of the time series with the current value, past and white noise in the model [16–18, 22, 34]. In the ARIMA model, let p and q be the numbers of autoregressive and moving average terms and they are always mentioned in the order of the model while, d be the integer representative of the differential order. The type of ARIMA model with mean, μ is represented mathematically as follows.

$$y_t = \mu + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

where, y_t and ε_t are the actual value and the random error at time t, respectively. Both are assumed to be independently and identically distributed (iid) with mean 0 and constant variance of σ^2, $\emptyset_i (i = 1,2,\ldots,q)$ and $\theta_j (j = 0,1,2,\ldots,q)$ are the model parameters that need to be predicted.

Support vector machines model

The support vector machine (SVM) introduced by Vladimir Vapnik [19] which involves statistical learning theory can better handle larger dimensional data, even with a small number of training examples, and has excellent generalization. Because the models choose limit support vectors from input data, they process data quickly. The SVM regression function is written as follows.

For linear and regressive data set $\{x_i, y_i\}$ the function is formulated as follows

$$f(x) = w^T x + b \quad (2)$$

The coefficient w and b are estimated by minimizing

$$\frac{1}{2} w^T w + C \frac{1}{n} \sum_{i-1}^{n} L_\varepsilon (y_i, f(x_i)) \quad (3)$$

where L_ε is called the ε-intensive loss function and is formulated as follows:

$$L_\varepsilon (y, f(x)) = \begin{cases} 0 \text{ if } |y - f(x)| \leq \varepsilon \\ |y - f(x)| \text{ others} \end{cases} \quad (4)$$

By introducing positive slack variable ξ and ξ_i^*, Eq (3) can be transformed to the following constrained formulation:

$$\begin{aligned} &min\frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &wx_i + b_i - y_i \leq \xi + \xi_i^* \\ &-wx_i - b_i + y_i \leq \xi + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0 \\ &i = 1, 2, \ldots, N \end{aligned} \tag{5}$$

When solving the above formula, we always utilize dual theory to convert it into a convex quadratic programming problem. Introducing the Lagrange Eq(5) change into the following term:

$$min\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\alpha_i^T\alpha_j - \sum_{i=1}^{n}\alpha_i^*(y_i - \varepsilon) - \alpha_i(y_i + \varepsilon) \tag{6}$$

subject to

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0,\ \alpha_i, \alpha_i^* \in [0, C]$$

When the data set cannot be regressed linearly, we also map them to a high dimension feature space and make linear regress. Then the formulation is as follows:

$$min\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\varphi(x_i)^T\varphi(x_j) - \sum_{i=1}^{n}\alpha_i^*(y_i - \varepsilon) - \alpha_i(y_i + \varepsilon) \tag{7}$$

subject to

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \alpha_i, \alpha_i^* \in [0, C]$$

Let $K(X_i, X_j) = \{\varphi(X_i) \cdot \varphi(X_j)\} = \varphi^T(X_j)\varphi(X_i);\ K(x, x)$ is the inner product of feature space and is called kernel function. Any symmetric function that satisfies Mercer condition can be used as Kernel Function [19]. The Gaussian kernel function is specified in this study.

$$K(x_i, x_j) = exp(-||x_i - x_i||^2/(2\sigma^2)) \tag{8}$$

The SVMs were employed to estimate the nonlinear behaviour of the forecasting data set as Gaussian kernels tend to give good performance under general smoothness assumptions [23].

Proposed hybrid models

Despite various time series models presented, the accuracy, effectively as well as precisely of time series forecasting at this time become the fundamental to many decision-making processes. However, those factors do not occur in the ARIMA and SVM models. This also become the most reason why time series forecasting model is crucial, most challenging, and dynamic as well as active research in many fields of studies. ARIMA and SVM models also have achieved success in their linear or nonlinear areas [16, 25, 26]. However, none of these are generic principles that can be generalized to all situations. Hence, a hybrid strategy that employs both linear and nonlinear modelling skills is recommended. This approach is suggested mainly for improving overall prediction effectiveness. Therefore, there is no research

on how to improve the effectiveness of forecasting models conducted especially in the case of COVID-19 in Malaysia.

In this study two motivation for hybrid models. First, a single model of ARIMA and SVM may not be sufficient to identify all the characteristics of the time series. Second, the assumption that either one or both cannot recognize the actual data generating process. Building the hybrid models of this study involved of two parts. Part I about linear autocorrelation composition and follow with nonlinear component in part II. Thus,

$$y_t = L_t + N_t \tag{9}$$

Where L_t and N_t signifies the linear composition and the nonlinear component, respectively. These two parts must be approximated based on the data. In the part I, linear modelling become the focus using ARIMA model to model the linear composition. The model from the first model involved the residuals which is the nonlinear interactions, and it cannot be model by linear model, and maybe linear relationship as well. Thus,

$$L_t = [\sum_{i=1}^{p} \emptyset_i z_{t-i} - \sum_{i=1}^{p} \theta_j \varepsilon_{t-j}] + e_t = \hat{L}_t + e_t \tag{10}$$

Let e_i signify the residual from the linear model at time t, then

$$e_t = y_t - \hat{L}_t$$

where $\hat{L}_t$ is the predicted value for time t from the estimated relationship in (1) with e_t is the residual at time t from the linear model. According to Aisyah, et al., [16] the residual data set after ARIMA fitting will only contain non-linear relationships and can be properly represented by a linear model. Results of first stage which contains the forecast values and residuals of linear modelling then used in Part II.

In Part II, the focus is for nonlinear modelling which SVM used to model the nonlinear (maybe linear) relationship occurring in residuals of linear modelling and original data as well. Then, the residual can be calculated using SVM by modelling various configurations as follows:

$$e_t = f(e_{t-1}, e_{t-2}, ..e_{t-n}) + \varepsilon_t \tag{11}$$

$$e_t = f(e_{t-1}, e_{t-12}) + \varepsilon_t \tag{12}$$

$$y_t = f(y_{t-1}, y_{t-12}, \hat{L}_t) + \varepsilon_t \tag{13}$$

$$y_t = f(y_{t-1}, y_{t-12}) + \varepsilon_t \tag{14}$$

where f is a nonlinear function determined by the SVMs model and ε_t is the random errors.

Thus, the combined forecast is

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{15}$$

Eqs (11) and (12) can be identified as $\hat{N}_t$, therefore the forecasted values can be achieved by summation of linear and nonlinear components Fig 1 shows the functional flowchart of hybrid models

In short, the proposed methodology of the hybrid process consists of two parts. In the part I, the ARIMA model is employed to analyse the problem of linear composition. In the part II, a SVM model is developed to model the residuals from part I. Since the ARIMA model in part I cannot handle the nonlinear component of the data, the residuals of linear model will include

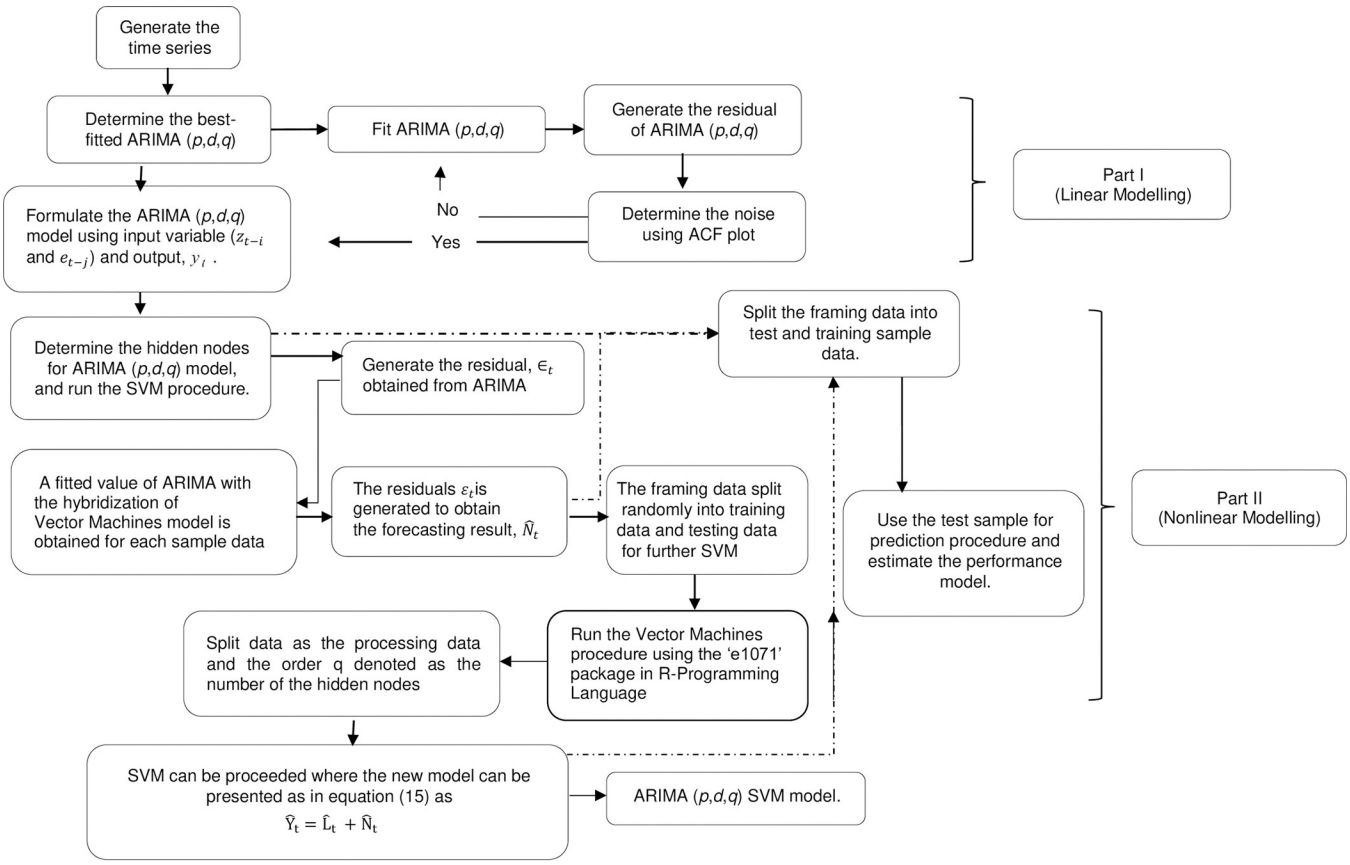

Fig 1. Flowchart process for hybrid ARIMA SVM models.

https://doi.org/10.1371/journal.pone.0285407.g001

information about the nonlinearity. The results from the SVM can be treated as forecasts of the error terms for the ARIMA model. The hybrid model utilizes the distinctive feature and strength of ARIMA and SVM model as well in defining various patterns. Therefore, it is more effective to model linear and non-linear patterns separately by using two different models and re-hybridize the forecast results obtained to improve overall modelling and forecasting performance.

Proposed algorithm

Step 1: Three selected time series of COVID-19 cases datasets (1st of October 2020-4th of November 2022), namely daily new positive cases, daily new deaths cases and daily new recovered cases are generated in R programming Language

Step 2: Every of the generated datasets is defined as $\{X_{1i} = x_{11,}x_{12}, x_{13,} \ldots, x_{n1}\}$, $\{X_{2i} = x_{21,}x_{22}, x_{23,} \ldots, x_{2n}\}$ and $\{X_{3i} = x_{31,}x_{32}, x_{33,} \ldots, x_{3n}\}$ for daily new positive cases, daily new deaths cases and, daily new recovered cases, respectively. Then, selected the best ARIMA (*p*,*d*,*q*) after checking the autocorrelation function (ACF) plot of ARIMA (*p*,*d*,*q*) residuals. The best fitted value for daily new positive cases is ARIMA (2,1,2), while ARIMA (1,1,2) and ARIMA (0,1,1) for daily new fatalities cases, and daily new recovered cases of COVID-19, respectively.

Step 3: The fitted value, $y_{t-i} = (y_{t-1}, y_{t-2}, \ldots, y_{t-m})$ and the residuals $e_{t-i} = (e_{t-1}, e_{t-2}, \ldots., e_{t-n})$

Step 4: Combine the values in step 3 as a set of input variables to get the output y_t

Step 5: The ARIMA (p,d,q) is defined by the order of q. According to the information in step 4, Vector Machines is carried out to examine the residuals to get the output L_t using R-programming Language.

Step 6: A fitted value of ARIMA with the hybridization of Vector Machines model is obtained for each sample data. Then, the residuals ε_t is generated to obtain the forecasting result, $\hat{N}_t$

Step 7: The framing data split randomly into training data and testing data for further Vector Machines model. Run the Vector Machines procedure using the 'e1071' package in R-Programming Language

Step 8: Assume the split data as the processing data and the order q as in Step 5. Therefore, the combine forecast as in Eq (15): $\hat{Y}_t = \hat{L}_t + \hat{N}_t$

Step 9: Estimate the model performance using the statistical measurement which are MSE, RMSE, MAE and MAPE.

Forecasting evaluation criteria

In order to evaluate the performance of the proposed hybrid models, the different statistical measurements criteria which followed by [16, 17, 32], such as MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error) are used.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |\hat{y}_t - y_t|$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} (\hat{y}_t - y_t)^2$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\hat{y}_t - y_t)^2} = \sqrt{\mathrm{MSE}}$$

For ARIMA model, normally, the measurement tools such as Akaike's information Criterion (AIC) and the Bayesian information criterion (BIC) have been widely used in time series analysis to determine the appropriate length for distributed lag [16, 17]. Therefore, model selection is made based on the model with the smallest value of AIC and BIC to provide measures of model performance which gives the selection of the best ARIMA model. Meanwhile, for the SVMs models, three parameters such as γ, C and ε are used as the measurement tools to determine the best fitted model. Inappropriate selection of SVM model parameters can result in either over or under fitting the training data. As with the ARIMA model, the parameter sets of the SVMs model with the lowest MSE value will be selected for use in the best fitting model. Thus, for the hybrid models, first the ARIMA worked as a pre-processor to filter the linear pattern of data sets. Then, the error term generated from the ARIMA model will be fed into the SVM in the hybrid models. The SVMs were performed to reduce the error function from the ARIMA.

Results and discussion

Application of the hybrid model to daily cases of COVID-19 in Malaysia

This section analysed the performance of the proposed model in respect to two aspects: (1) the performance of the proposed models against ARIMA and SVM models, and (2) the percentage

improvement of the proposed models against ARIMA and SVM models. Since the World Health Organisation (WHO) was declared that COVID-19 is pandemic worldwide, the COVID-19 time series data sets have been widely studied. Next, the predictive capability of the developed novel models was compared using three well-known data sets of daily cases of COVID-19 in Malaysia- daily new positive cases data, daily new fatalities cases data and daily new recovered cases data–used to demonstrate the performance of the proposed model in terms of accuracy, effectively and accurately. All these data are reported from the 1st of October 2020 to 4th of November 2022 and retrieved from the COVIDNOW website at https://covidnow.moh.gov.my/

In the Table 1, the minimum value of the new death, new cases and new recovered are zero, 2600 and 1.8, respectively, while the maximum value of new cases, death and recovered cases are 33872.0, 592 and 33406 respectively. Similarly, the mean and median for the number of new cases, death and recovered cases are 6322.7, 47.51, 6415.5, where the parenthesis indicates the median in (3471, 11, 3447.0). While the first quartile value of daily new cases, death and recover cases are 1922, 4 and 1843 respectively. The third quartile value of number of daily new cases, death and recover cases are 6824, 58 and 6775 respectively. Moreover, the standard deviation of new cases, death and recover cases are 7097.8, 81.12 and 7058.3 respectively.

Part I (Linear Modelling)–the best ARIMA model for the daily new positive case dataset is derived from ARIMA (2,1,2). The best fitting ARIMA model for the daily new death case data set is ARIMA (1,1,2). Meanwhile, in the case of the daily new recovered cases dataset, the best ARIMA model is reported as ARIMA (0,1,1). The results of this ARIMA (*p*,*d*,*q*) model are summarized in Table 2. The estimates of all parameters are shown in Table 3. From this table, it can be observed that the *p*-values of all parameters are small. Therefore, the models were statistically significant for confirmed, recovered, and death cases, and could be used to forecast the future [33, 35].

Part II (Nonlinear Modelling)–In order to obtain an optimal machine learning algorithm, based on the concepts of support vector machine design and using pruning algorithms in R-programming software. For the daily new positive COVID-19 cases datasets, parameters $\gamma = 2$, $C = 256$, $\varepsilon = 0.2$ shows the smallest values of MSE i.e., 10321275 (see Table 4). Therefore, this parameters value was selected for use in the best-fitting model for the datasets of daily new positive COVID-19 cases. Whereas the smallest value of MSE is 1431.732 and 9885746 (Table 4), with parameters $\gamma = 2$, $C = 256$, $\varepsilon = 0.2$ are selected as the best-fitting model for daily new death cases of COVID-19 and daily new recovered cases of COVID-19, respectively.

Table 1. Descriptive statistics of COVID-19 daily new cases, death and recovered cases of Malaysia.

	New Case	New Death	New Recovered
Min	2600	0	1.8
1st Qu	1922	4	1843.0
Median	3471	11	3447.0
Mean	6415.5	47.5098	6322.7
3rd Qu	6824	58	6775.0
Max	33406	592	33872.0
SD	7097.8	81.1215	7058.3

In addition, this section also discusses the process of proposed models at once for both part i.e., Part I (Linear modelling) and Part II (Nonlinear Modelling) using three well-known data sets of COVID-19 i.e., daily new positive cases, daily new deaths cases and daily new recovered cases are discussed in order to demonstrate the effectiveness of the proposal models. Both linear and nonlinear modelling as well as well as the data used in this study are executed through programming using the R-language.

https://doi.org/10.1371/journal.pone.0285407.t001

Table 2. The best ARIMA(p,d,q) model selection.

COVID-19 daily cases	ARIMA(p,d,q)	AIC	BIC
Daily New Positive Cases	(2,1,2)	12564.54	12587.73
Daily New Deaths Cases	(1,1,2)	6930.12	6948.63
Daily New Recovered Cases	(0,1,1)	13044.74	13054.01

https://doi.org/10.1371/journal.pone.0285407.t002

New positive cases data forecasts

The daily new positive cases datasets series is recoded from the 1st of October 2020 to 4th of November 2022 (see Fig 2) contains 765 data points. The number of daily new positive cases of COVID-19 in Malaysia continued to show a significant increase starting in July 2021 dropped below 5,000 new cases. However, it's continued an increased again around March-April 2022 to the maximum of 33,406.00. But this number showed a drastic decrease until November 4, 2022. The daily new positive cases of COVID-19 datasets, which is consider in this investigation and the COVID-19 datasets also have been extensively used with a vast variety of linear and nonlinear time series models including ARIMA, ANN and machine learning methods [8–10, 12, 14, 17, 20–26, 34]. The study of the daily new positive cases of COVID-19 has crucial as an indication of the effectiveness of preventive measures that have been, are being and will be taken by the authorities in controlling the spread of this epidemic more effectively.

Therefore, to investigate the performance of the proposal models on daily new positive cases of COVID-19 datasets, which is similar approach by Aisyah et al., [16] is used where the dataset is divided into two samples, known as training sample and testing sample. According to Aisyah et al., [16] and Nurul Hila et al., [17], the datasets should be divided into two (2) which are 70–80% the data for training and the remaining 20–30% for testing yields the greatest outcomes [36, 37]. The training data are used to assemble the models while testing data is used to evaluate based on the statistical measurement the forecasting performances of the models. Thus, in this study the daily new positive cases of COVID-19 data set are divided into two samples which the training data set and test data set. For training data sets consists of 612 observations from day 1 to day 612, which is 80% of the data sets from October 1st, 2020, to June 4th, 2022, exclusively used to formulate. The test sample data sets used about 153 observations from days 613–765 (20%) for the period of 5th June 2022- 4th November 2022 in order to evaluate the forecasting performance of proposed models.

The performance of the proposed model of the daily new positive COVID-19 cases datasets are shown in Table 5. The results were obtained from the proposed models in terms of

Table 3. Parameter estimates of ARIMA (p,d,q) models and their p-values.

Model parameters	Estimate	z-stat	p-value
New Case *ARIMA(2,1,2)*			
θ_1	1.2408	120.085	< 0.0001
θ_2	-0.9715	-98.320	< 0.0001
φ_1	-1.2628	-42.225	< 0.0001
φ_2	0.8738	48.102	< 0.0001
Recovered Case *ARIMA(0,1,1)*			
φ_1	-0.3473	-9.953	< 0.0001
Death Case *ARIMA(1,1,2)*			
θ_1	0.8595	19.852	< 0.0001
φ_1	-1.6196	-35.651	< 0.0001
φ_2	0.7039	20.432	< 0.0001

https://doi.org/10.1371/journal.pone.0285407.t003

Table 4. SVMs model parameters for the daily new COVID-19 cases datasets.

COVID-19 daily cases	SVM Parameter	MSE
Daily New Positive Cases	γ = 2, C = 128, ε = 0.1	11260319
	γ = 2, C = 8, ε = 0.2	17519221
	γ = 2, C = 256, ε = 0.2	**10321275**
	γ = 2, C = 4, ε = 0.3	19499587
	γ = 2, C = 128, ε = 0.3	11058873
Daily New Deaths Cases	γ = 2, C = 16, ε = 0.2	1599.659
	γ = 2, C = 128, ε = 0.2	1418.033
	γ = 2, C = 256, ε = 0.2	1378.962
	γ = 2, C = 8, ε = 0.3	1711.465
	γ = 2, C = 256, ε = 0.3	**1431.732**
Daily New Recovered Cases	γ = 0.5, C = 256, ε = 0.2	38847557
	γ = 1, C = 4, ε = 0.2	36617325
	γ = 1, C = 8, ε = 0.2	34666712
	γ = 2, C = 128, ε = 0.2	10463694
	γ = 2, C = 256, ε = 0.2	**9885746**

https://doi.org/10.1371/journal.pone.0285407.t004

measurement error terms, namely MSE and MAE have the smaller values of 42552.7137 and 90.34845 Similar results were also obtained from the testing datasets with values of 61223.474, 0.05633, 247.4337 and 146.9841 for MSE, MAPE, RMSE and MAE, respectively. Based on these numerical results, the findings are examined in more detail using figures as illustrated in Fig 3. This figure, illustrates the estimated values for the proposed model (test sample) of daily new positive COVID-19 cases. As can be seen from this figure, the proposed model line closely matches the actual data. As a further example, Figs 4–6 provide estimated values of our model for test data and ARIMA, SVM, and SVM models for COVID-19 cases. A comparison of the proposed model's lines for the test sample (Fig 6) with ARIMA and SVM models clearly shows

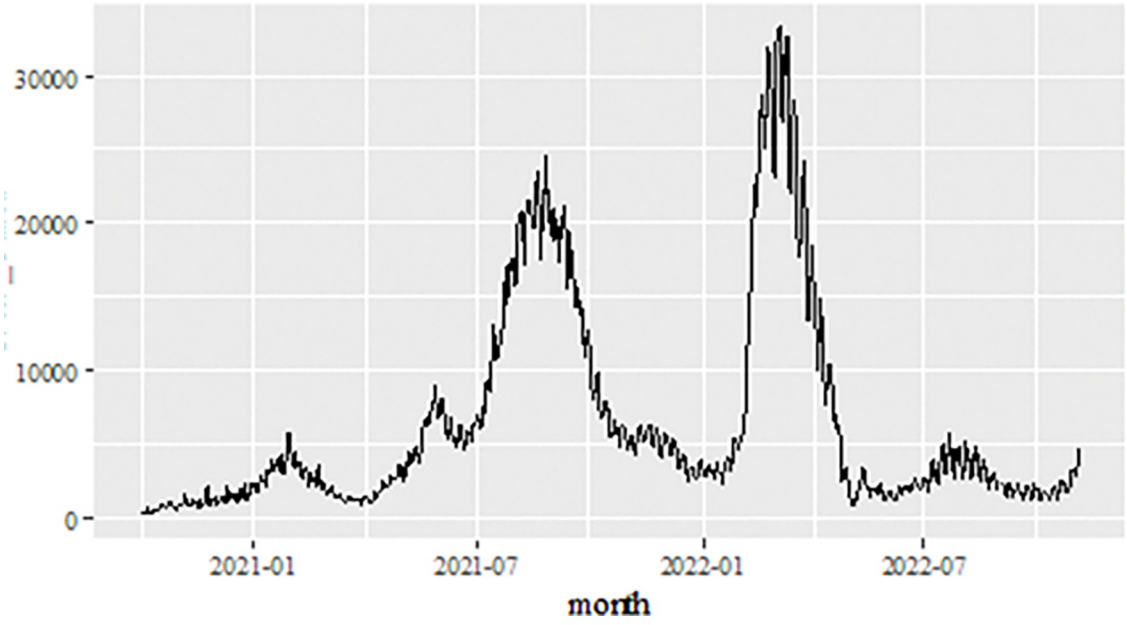

Fig 2. Malaysian daily new positive COVID-19 cases (1st of October 2020 to 4th of November 2022).

https://doi.org/10.1371/journal.pone.0285407.g002

Table 5. Performance measures of the proposed model for daily new positive COVID-19 cases datasets.

Models	Train		Test			
	MSE	MAE	MSE	MAPE	RMSE	MAE
ARIMA	929843.169	611.0274	298988.28	0.15167	546.7982	397.57
SVM	8355184.483	2001.644	274588.16	0.15421	524.0116	390.3848
ARIMA-SVM	42552.7137	90.34845	61223.474	0.05633	247.4337	146.9841

https://doi.org/10.1371/journal.pone.0285407.t005

that the proposed model's lines are somewhat similar to actual data. Comparing the performance of our proposal models with that of ARIMA and SVM models, this indicated that our proposal models are efficient, accurate, and precise. In addition, as in Fig 7, the number of daily new positive COVID-19 cases is plotted. From this figure, the daily new positive cases of COVID-19 for Malaysia are forecasted for the forthcoming three weeks.

Based on Table 6, we further analysed the performance of the proposed models for the daily newly positive COVID-19 cases dataset by comparing at the percentage of MSE, MAPE, RMSE and MAE. The study hypothesis investigates assumptions of the proposed hybrid model (ARIMA-SVM) approach to single ARIMA and SVM models. The proposed model achieved a higher percentage of improvement in MAE, MAPE, MSE and RMSE compared to the ARIMA model with improvements of 63.03%, 62.86%, 79.52%, 54.74%, where the parenthesis indicates the SVM model that results in (62.34%, 63.47%, 77.70%, 52.78%). Therefore, based on these results (Tables 4–6 and Figs 3–7), it can be concluded that the proposed model that has been developed has produced higher accuracy as well as efficiency compared to results achieved by ARIMA and SVM

New deaths cases data forecasts

Besides the Malaysian daily new positive COVID-19 cases datasets, the Malaysian daily new deaths cases datasets are also considered and used to analyse the performance of the proposed

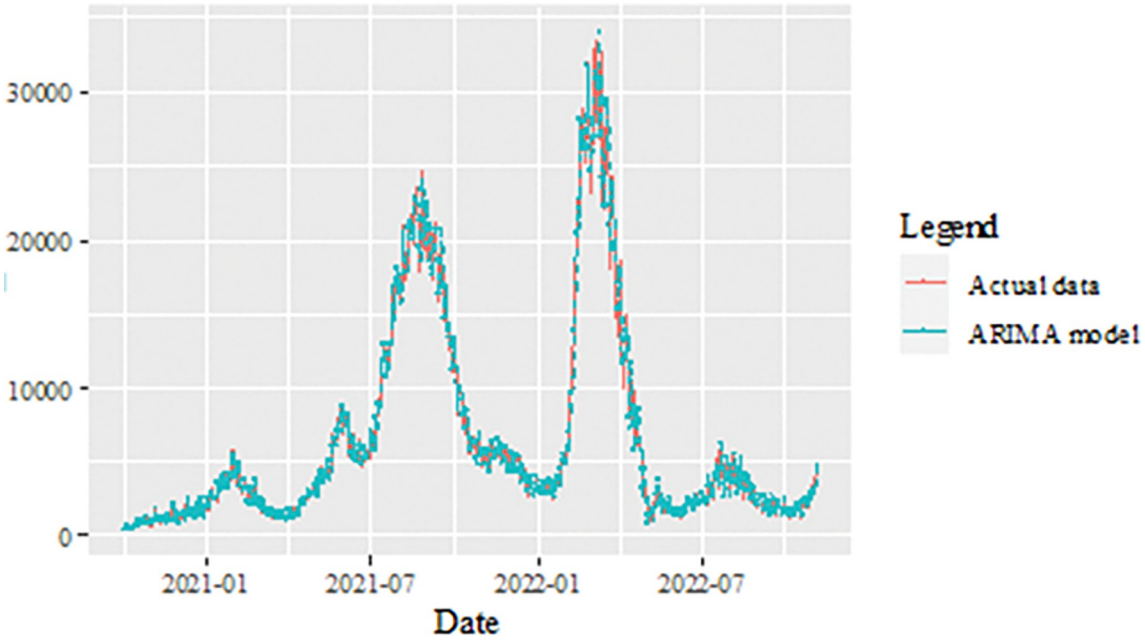

Fig 3. Results obtained from the proposed model for daily new positive COVID-19 cases dataset.

https://doi.org/10.1371/journal.pone.0285407.g003

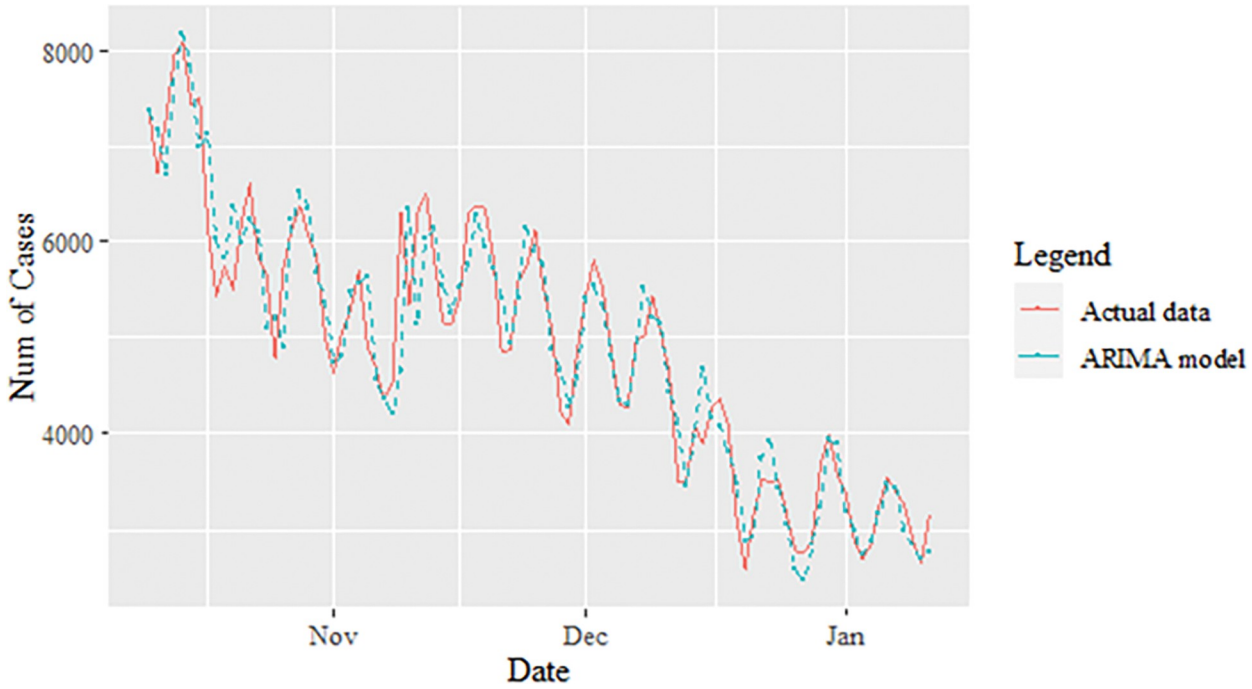

Fig 4. ARIMA model prediction of daily new positive COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g004

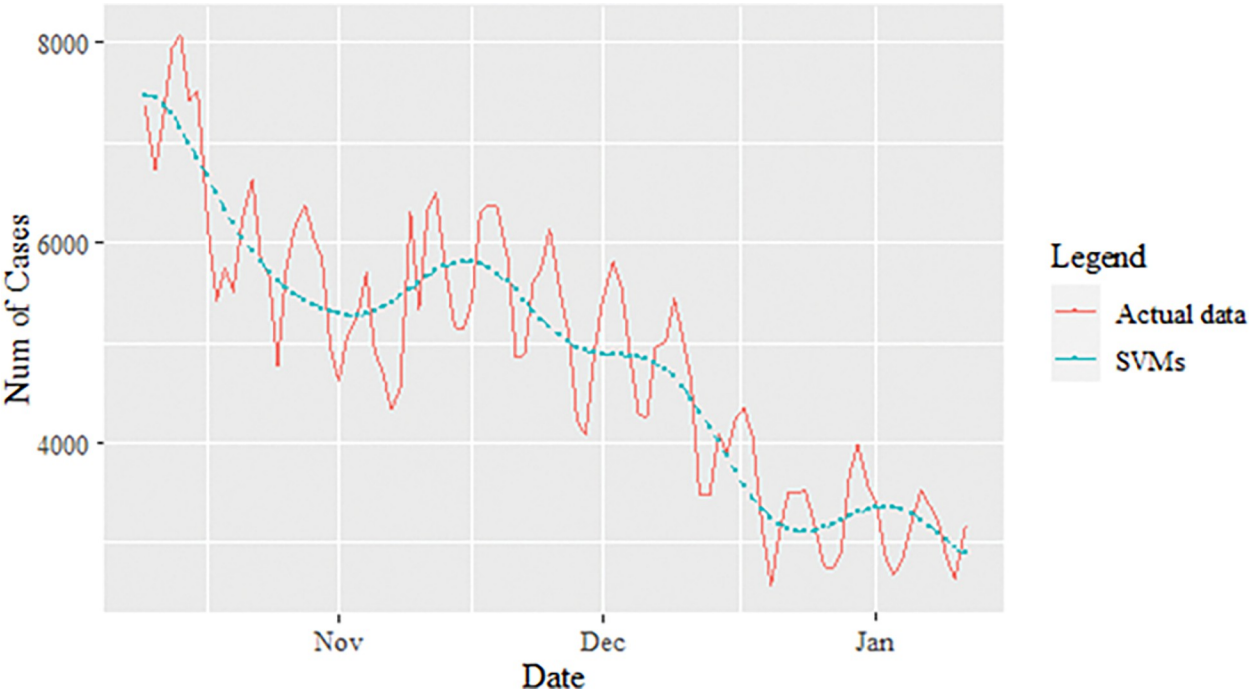

Fig 5. SVM model prediction of daily new positive COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g005

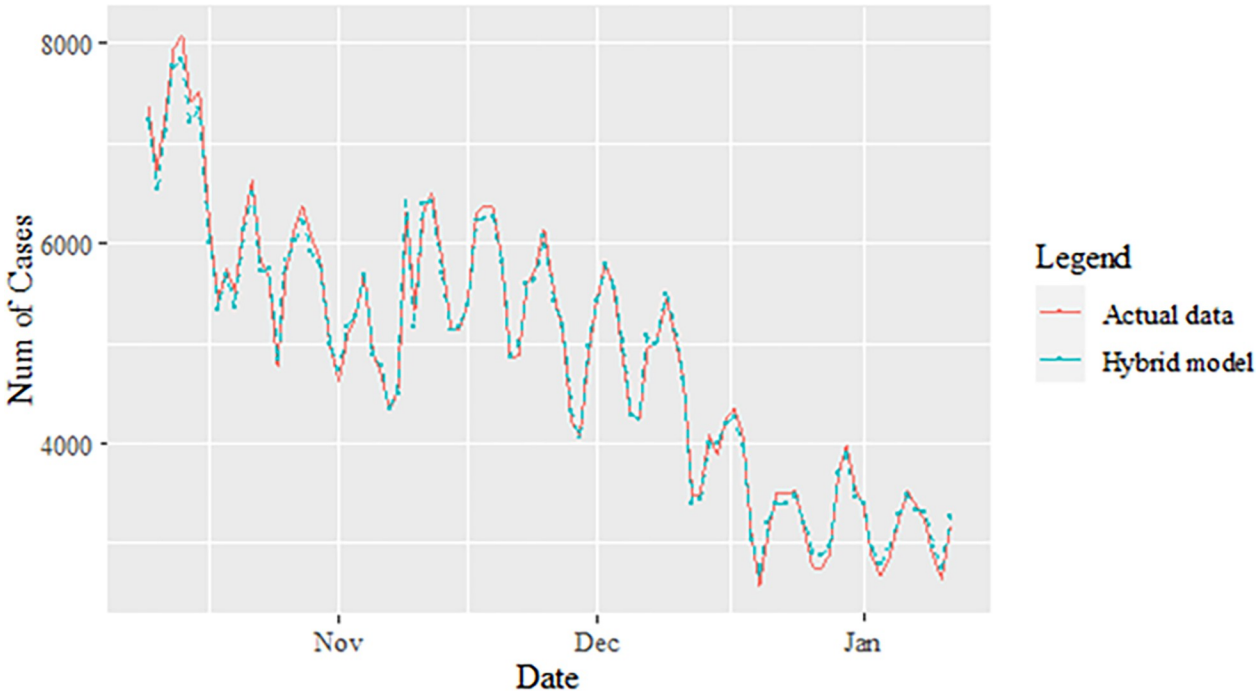

Fig 6. Proposed models prediction of daily new positive COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g006

models. Similar to the daily new positive data set as well as the daily new death case data set, the recording period of this data set from 1st of October 2020 to 4th of November 2022 (see Fig 8) contains 765 data points and is divided into two samples. As a result of the increase in the number of daily positive cases of COVID-19 reported, this also shows that there is a significant increase in the number of deaths around 600. In order to formulate the model, the training

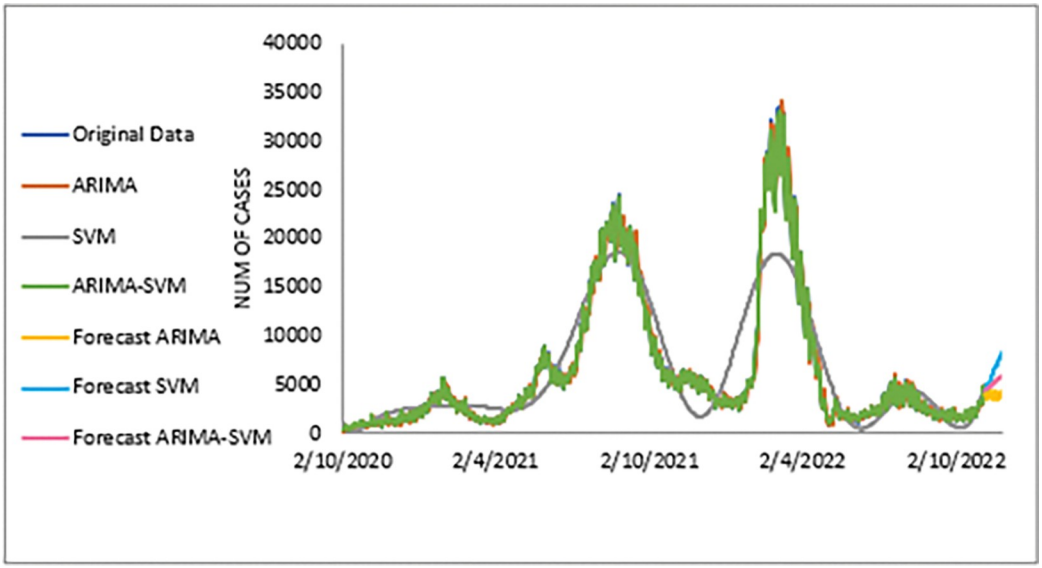

Fig 7. Actual and three weeks ahead forecasted values of ARIMA, SVM and ARIMA SVM models for new cases of COVID-19 of the 80% training and 20% testing set.

https://doi.org/10.1371/journal.pone.0285407.g007

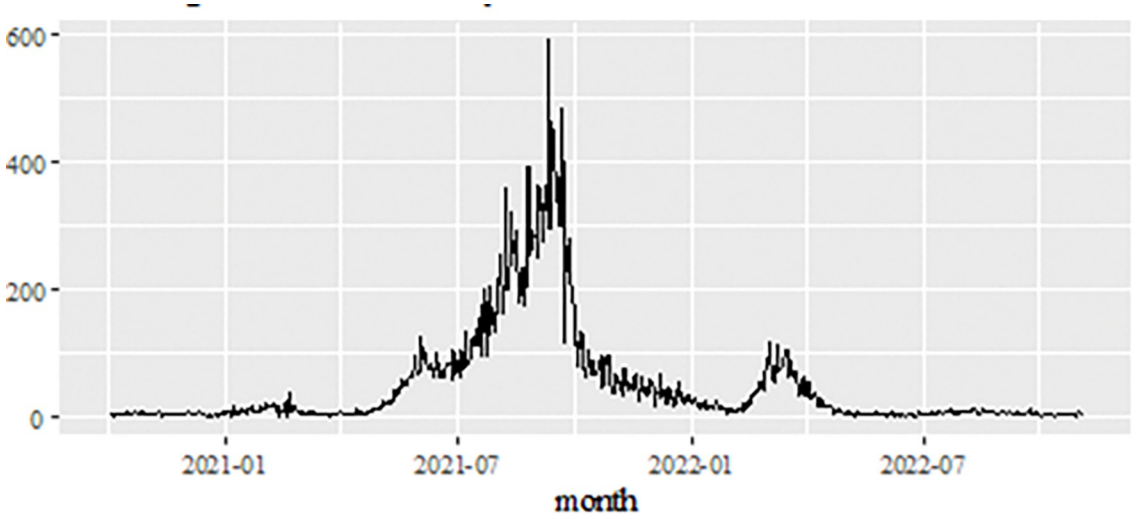

Fig 8. Malaysian daily new deaths COVID-19 cases (1st of October 2020 to 4th of November 2022).

https://doi.org/10.1371/journal.pone.0285407.g008

data set involves 612 observations (80%) from October 1, 2020- June 4, 2022, the test sample uses approximately 153 observations (20%) for the period June 5, 2022- November 4, 2022, to evaluate the prediction performance of the proposed model.

A similar approach to the daily new positive cases of the COVID-19 dataset was used to study the performance of the proposed model on the daily new death cases of the COVID-19 dataset, the dataset was divided into two samples, namely, training sample and testing sample. For the training sample, it represents approximately 80% of the daily new death cases for the COVID-19 dataset (involving 612 observations with the period October 1, 2020, until June 4, 2022). The remaining 20% is for the test sample, involving approximately 153 observations starting from June 5, 2022- November 4, 2022.

The performance of the proposed models using the daily new deaths COVID-19 cases datasets is first characterized by statistical measurement such as the MSE, MAPE, RMSE and MAE as shown in Table 7. The results for the training data from this table show that the proposed model gives the smallest values of 49.4459 and 3.53812 for MSE and MAE values, respectively, compared to ARIMA and SVM for MSE and MAE values, respectively, compared to ARIMA and SVM. The same trend also occurs on the test data where all the values of the statistical measures used show the smallest values compared to the ARIMA and SVM models.

The study continues by investigating the estimated value of the proposed model for the daily new death COVID-19 case data set as illustrated in Fig 8. This figure clearly indicates that the proposed model line is almost no difference with the actual data. In addition, the estimated values of ARIMA, SVM and proposed models for test sample are plotted in Figs 9–11, respectively. Again, it clearly shows that our proposed model's lines (Fig 12) for test sample are relatively closed to actual data compared to ARIMA and SVM models. This shows that the results of our proposed model are consistent with previous findings, which are efficient, accurate and precise compared to ARIMA and SVM models. In addition, as in Fig 12, the number

Table 6. Percentage improvement of the proposed models with other forecasting models (The COVID-19 cases of daily new positive cases).

Model	MAE	MAPE	MSE	RMSE
ARIMA	63.0294	62.8602	79.5231	54.7486
SVM	62.3489	63.4719	77.7035	52.7809

https://doi.org/10.1371/journal.pone.0285407.t006

Table 7. Performance measures of the proposed model for daily new deaths COVID-19 cases datasets.

Models	Train		Test			
	MSE	MAE	MSE	MAPE	RMSE	MAE
ARIMA	697.999	11.8083	6.06741	0.56838	2.46321	1.92791
SVM	1409.19	21.8006	5.38920	0.53687	2.32146	1.85605
ARIMA-SVM	49.4459	3.53812	0.92630	0.19088	0.96303	0.76230

https://doi.org/10.1371/journal.pone.0285407.t007

of daily COVID-19 death cases is plotted. As a result of this figure, the daily new death cases of COVID-19 in Malaysia for the next three weeks are forecast to decrease, showing a downward trend in the next few weeks.

Here, a similar approach as in the daily new positive COVID-19 case dataset is used to investigate the performance of the proposed model for the daily new death COVID-19 case dataset through percentage MSE, MAPE, RMSE and MAE, as reported in the Table 8. Again, the percentage of improvement reveals that our proposed model produces better improvement for all statistical measures than the ARIMA and SVM models with results of 60.46%, 66.42%, 84.73%, 60.90%; improvement (58.93%, 64.45%, 82.81%, 58.52%) for MAE, MAPE, MSE and RMSE, respectively. The SVM model results reported in the parenthesis. The presented results (see Tables 7,8 and Figs 9–11, 13) clearly conclude that our proposed model has produced efficiently and accurately as well compared to ARIMA and ASV models.

New recovered cases data forecasts

The last dataset considered in this investigation to study the performance of the proposed model, is the dataset of new daily recovered cases of COVID-19 in Malaysia. Predicting

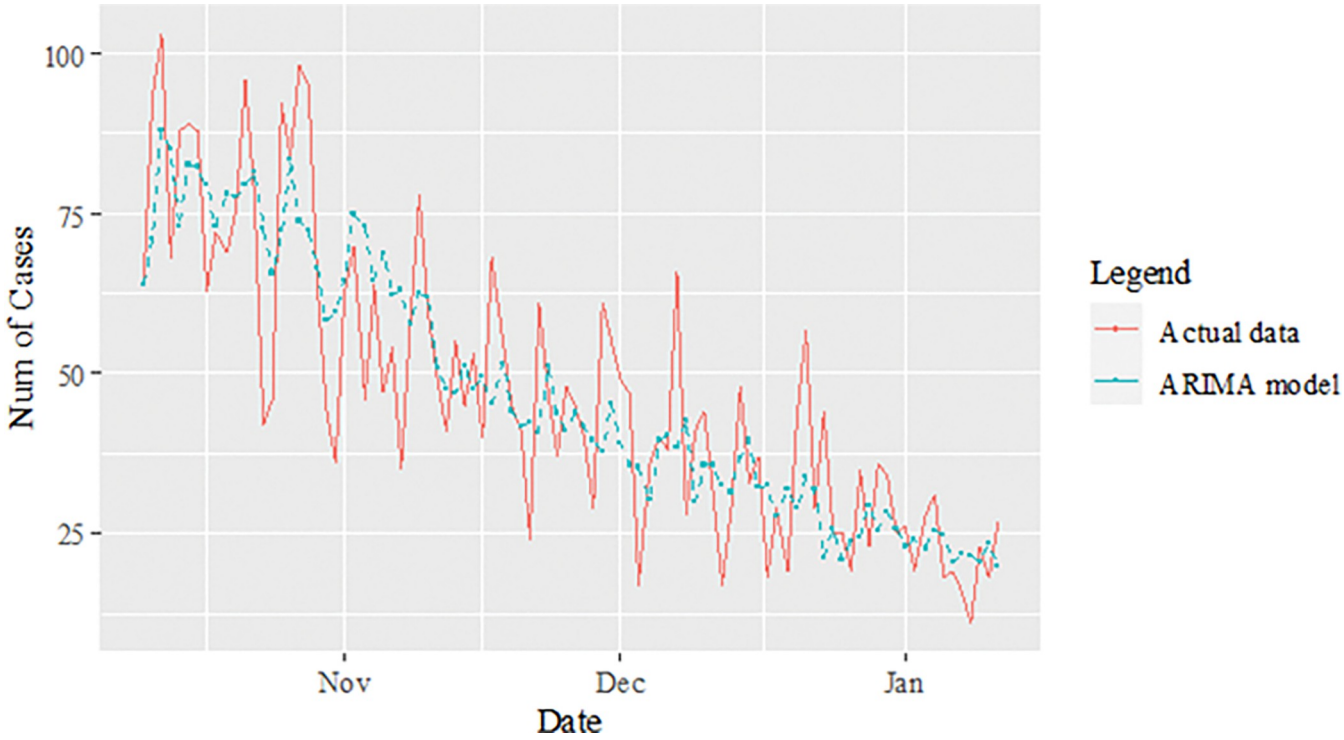

Fig 9. ARIMA model prediction of daily new deaths COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g009

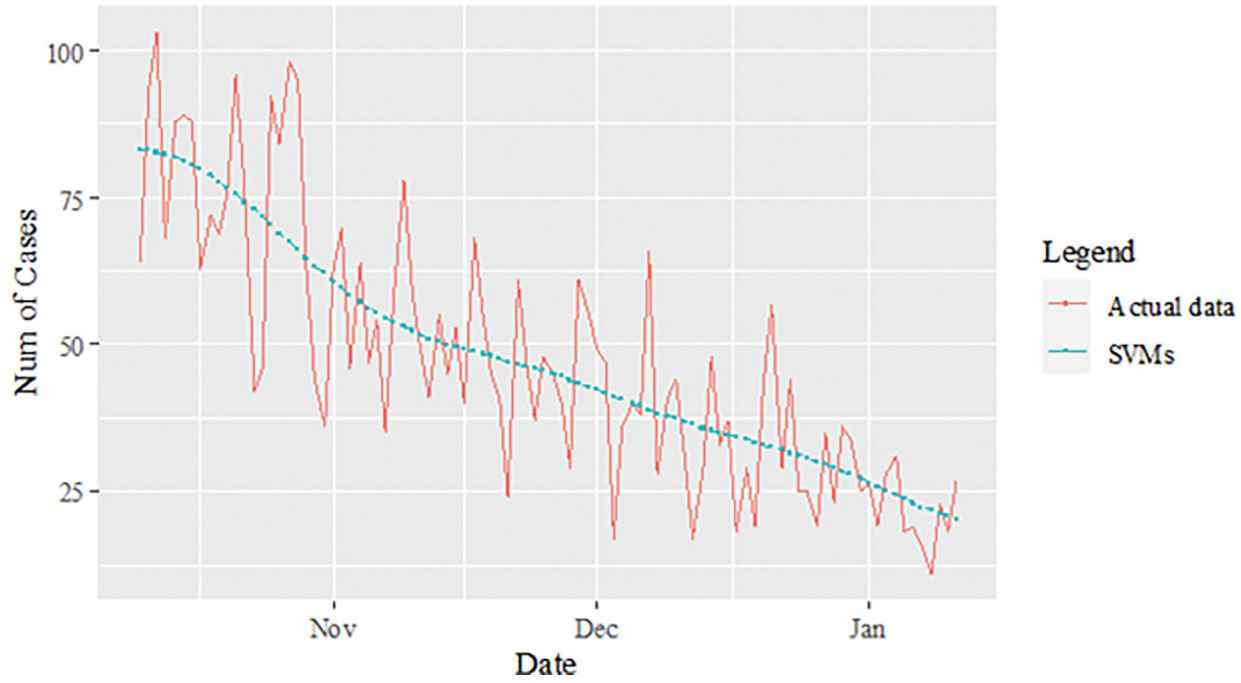

Fig 10. SVMs model prediction of daily new deaths COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g010

Malaysia's daily new recovered COVID-19 cases is equally important as the two datasets discussed earlier. The data used in this paper contain daily observation from the 1st of October 2020 to 4th of November 2022, giving 765 data points in the time series. The same trend is also

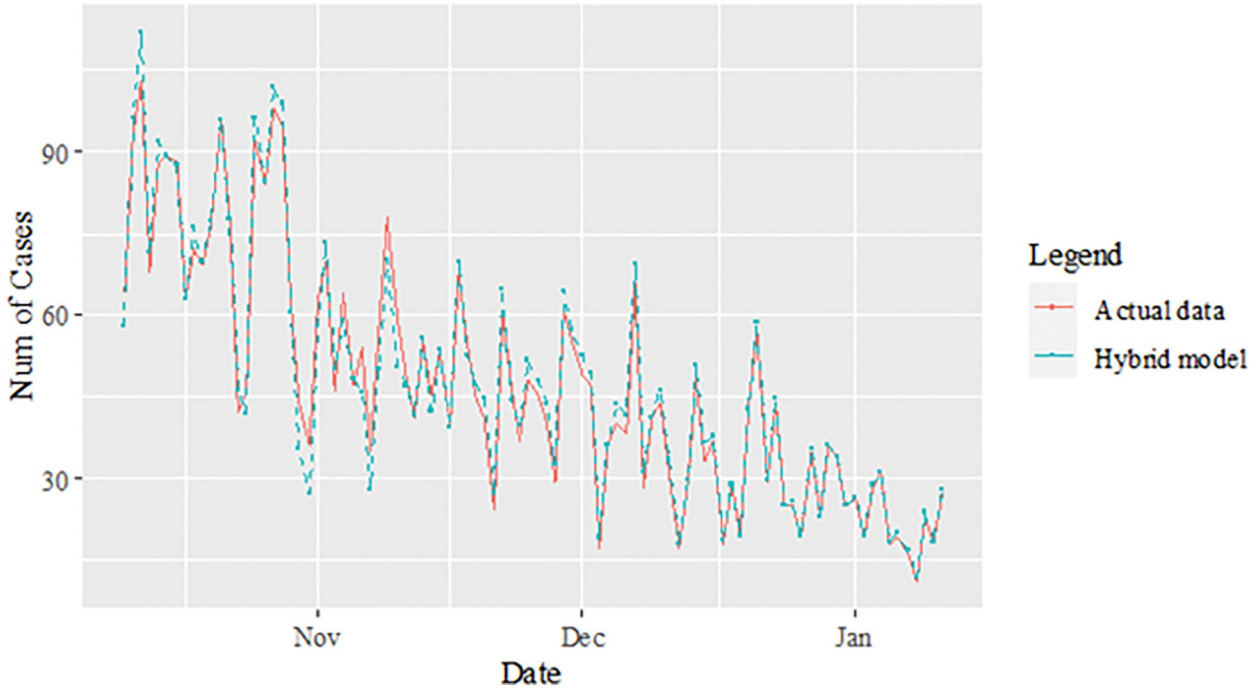

Fig 11. Proposed models prediction of daily new deaths COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g011

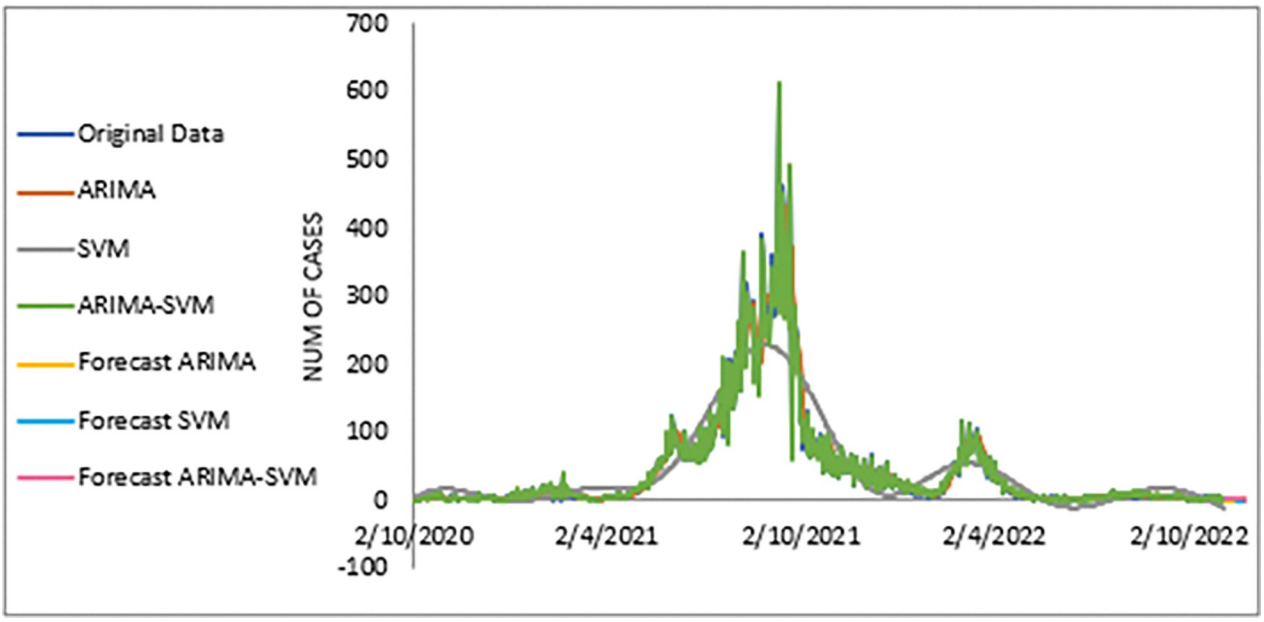

Fig 12. Actual and three weeks ahead forecasted values of ARIMA, SVM and ARIMA SVM models for daily new deaths COVID-19 cases of the 80% training and 20% testing set.

https://doi.org/10.1371/journal.pone.0285407.g012

Table 8. Percentage improvement of the proposed models with other forecasting models (The COVID-19 cases of daily new deaths cases).

Model	MAE	MAPE	MSE	RMSE
ARIMA	60.4598	66.4168	84.7332	60.9035
SVM	58.9289	64.4458	82.8119	58.5162

https://doi.org/10.1371/journal.pone.0285407.t008

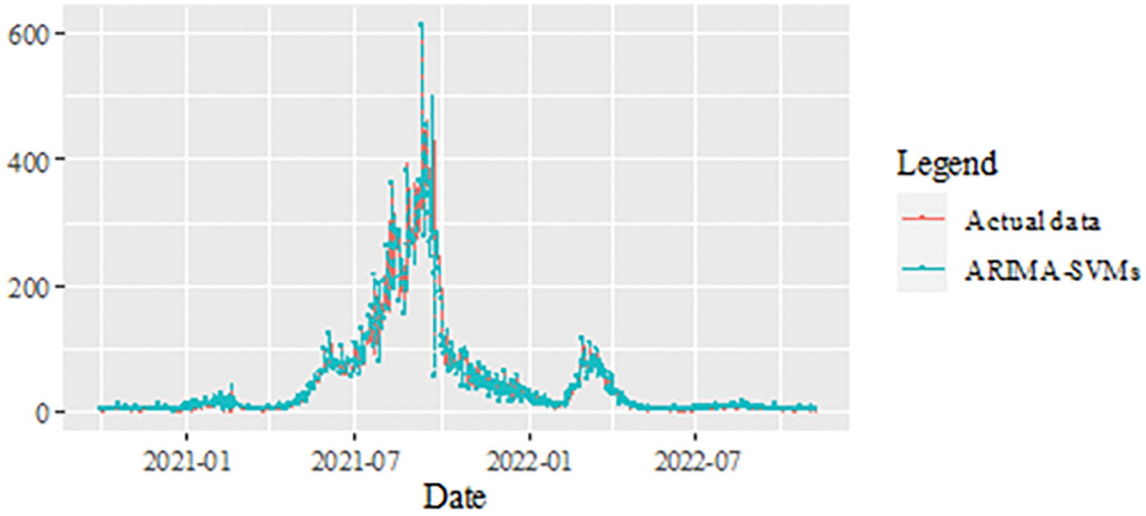

Fig 13. Proposed models prediction of daily new deaths COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g013

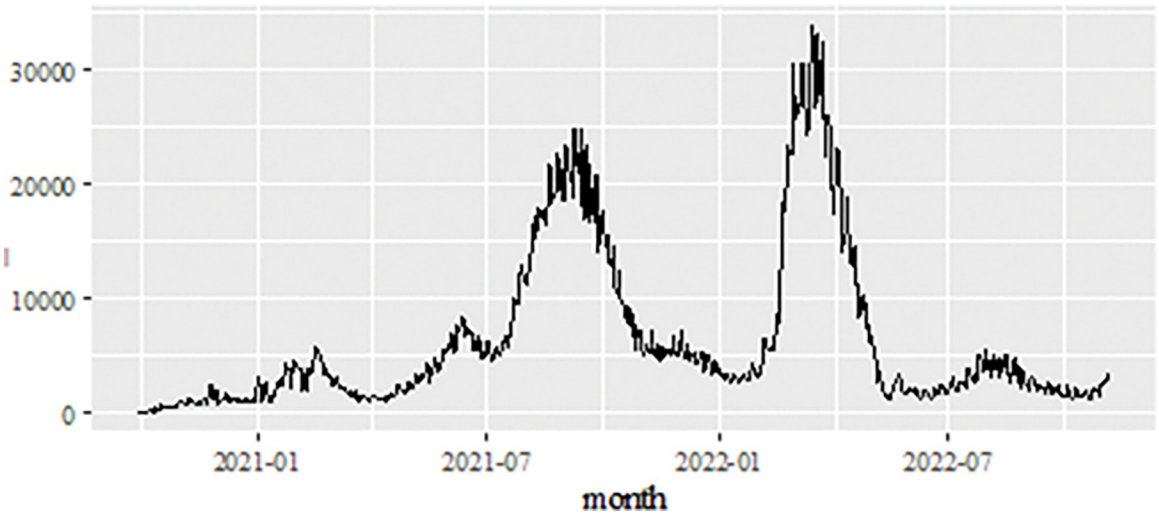

Fig 14. Malaysian daily new recovered COVID-19 cases (1st of October 2020 to 5th of November 2022).

https://doi.org/10.1371/journal.pone.0285407.g014

shown by the number of patients recovered from COVID-19 where there is a significant increase twice. Starting in July 2021, the number of recovered patients also shows an exponential increase until it reaches over 22,500.00 in August 2021 (the time series plot is given in Fig 14) and drop. However, around March—April 2022, the number of recovered COVID-19 cases increased again until a maximum of 33,872.00 and then decreased and it showed a relatively stable movement after that. This dataset also divided into two samples, i.e., the training data set and test data set. Like the previous datasets, training data set is implemented in order to formulate the model, which involved 612 observations (80%) from 1st October 2020-4th October 2022. Whereas, to evaluate the forecasting performance of the proposed model, the test sample uses approximately 153 observations (20%) for the period 5 June 2022- November 2022.

Table 9 presented the performance of the proposed model of the daily new recovered COVID-19 cases datasets based on training sample and test sample. The results in Table 9 clearly show that the proposed training sample model produces the smallest MSE and MAE values with 99205.699 and 136.8519, respectively compared to the MSE and MAE models of the ARIMA model and the SVM model. For the test sample also revealed that the same scenario as the training sample ie, produced the smallest MSE, MAPE, RMSE and MAE with values of 26108.02, 0.0396, 161.5797 and 104.1002, respectively compared to ARIMA and SVM as well.

Meanwhile, the estimated value for the test sample of the proposed model for the dataset of daily new COVID-19 cases is depicted in Fig 15. Again, this figure clearly shows that the predicted value from the proposed models appear to be close to the actual values. A further

Table 9. Performance measures of the proposed model for daily new recovered COVID-19 cases datasets.

Models	Train		Test			
	MSE	MAE	MSE	MAPE	RMSE	MAE
ARIMA	1802678.36	804.4378	271462.22	0.1560	521.0203	387.2768
SVM	7636804.13	1890.917	239672.00	0.1504	489.5630	371.6573
ARIMA-SVM	99205.699	136.8519	26108.02	0.0396	161.5797	104.1002

https://doi.org/10.1371/journal.pone.0285407.t009

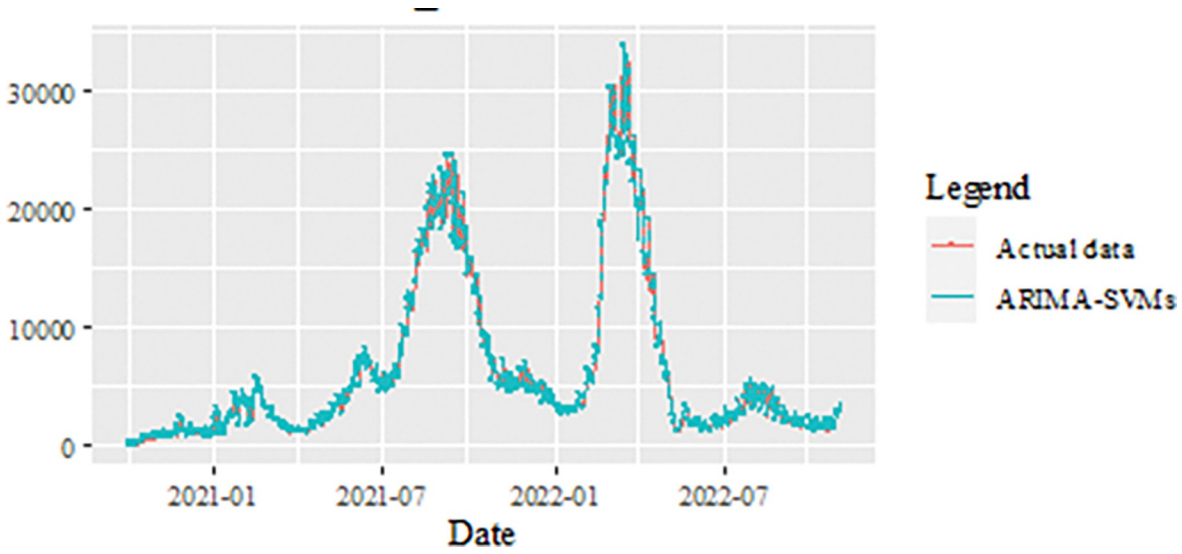

Fig 15. Results obtained from the proposed model for daily new recovered COVID-19 cases dataset.

https://doi.org/10.1371/journal.pone.0285407.g015

investigation of the proposed model's results is displayed in Figs 16–18. These three figures (Figs 16–18) reveal that the predicted values extracted from ARIMA, SVM, and the proposed model for the test samples seem to be close to the actual values. However, as we will see in Fig 8, these models are dominated by the proposed model i.e., they are closed to the true value. The number of daily new recovered COVID-19 cases is plotted as in Fig 19. In this figure, it's clearly shown that the proposed model follows the original sharpness of the data. From this figure, the daily new recovered cases of COVID-19 for Malaysia are forecasted for the

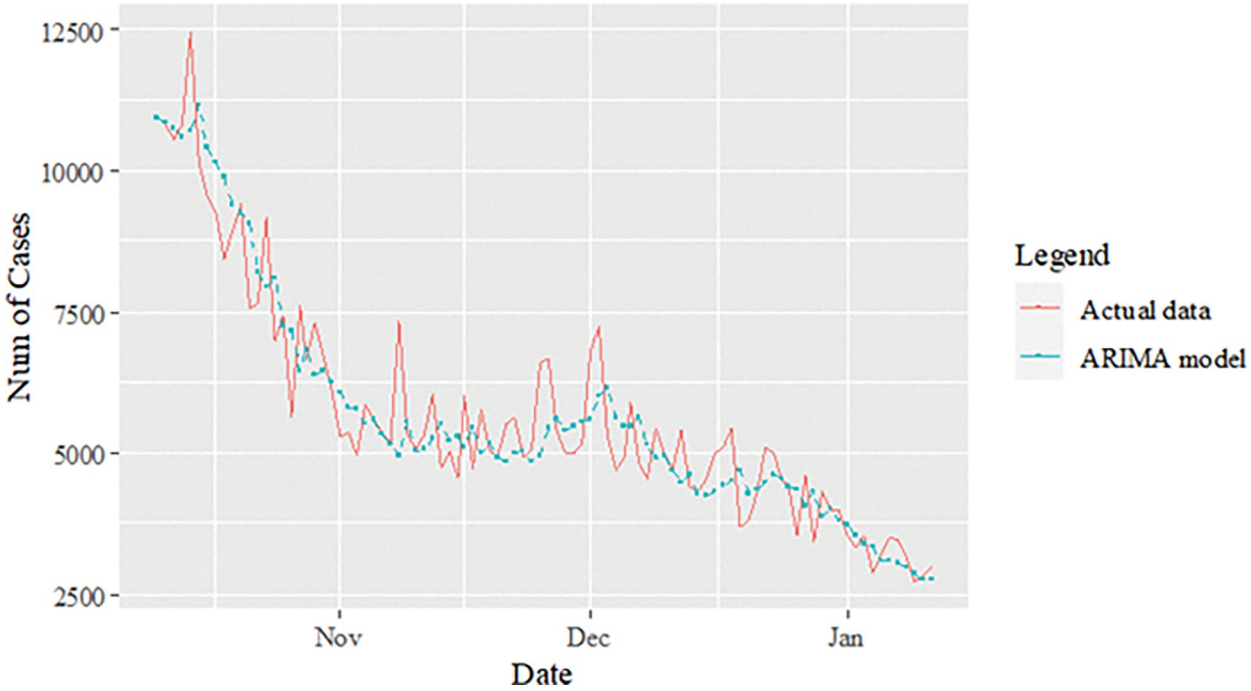

Fig 16. ARIMA model prediction of daily new recovered COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g016

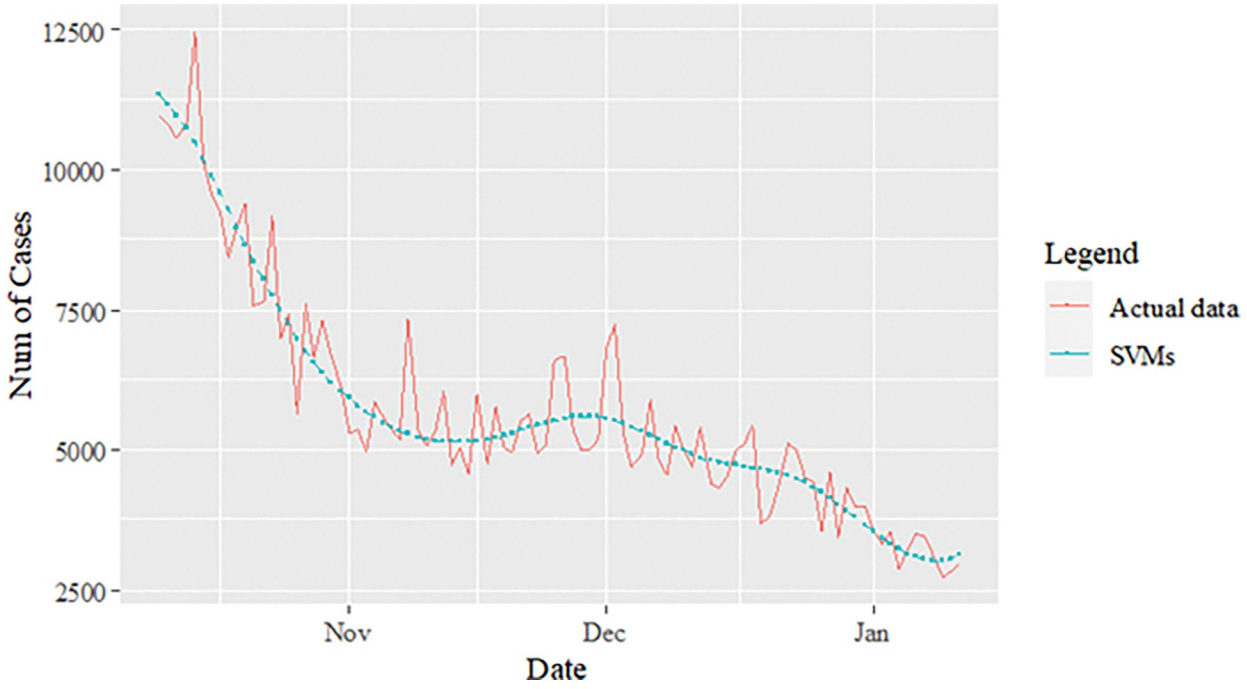

Fig 17. SVM model prediction of daily new recovered COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g017

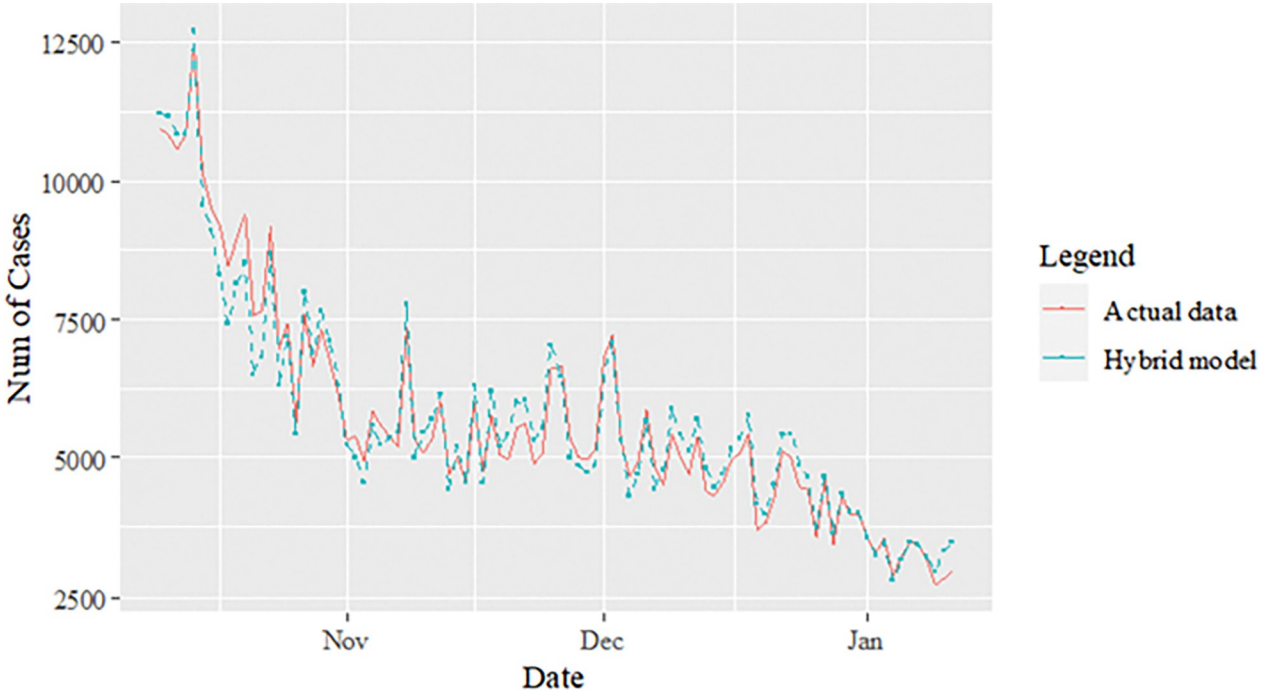

Fig 18. Proposed model prediction of daily new recovered COVID-19 cases dataset (test sample).

https://doi.org/10.1371/journal.pone.0285407.g018

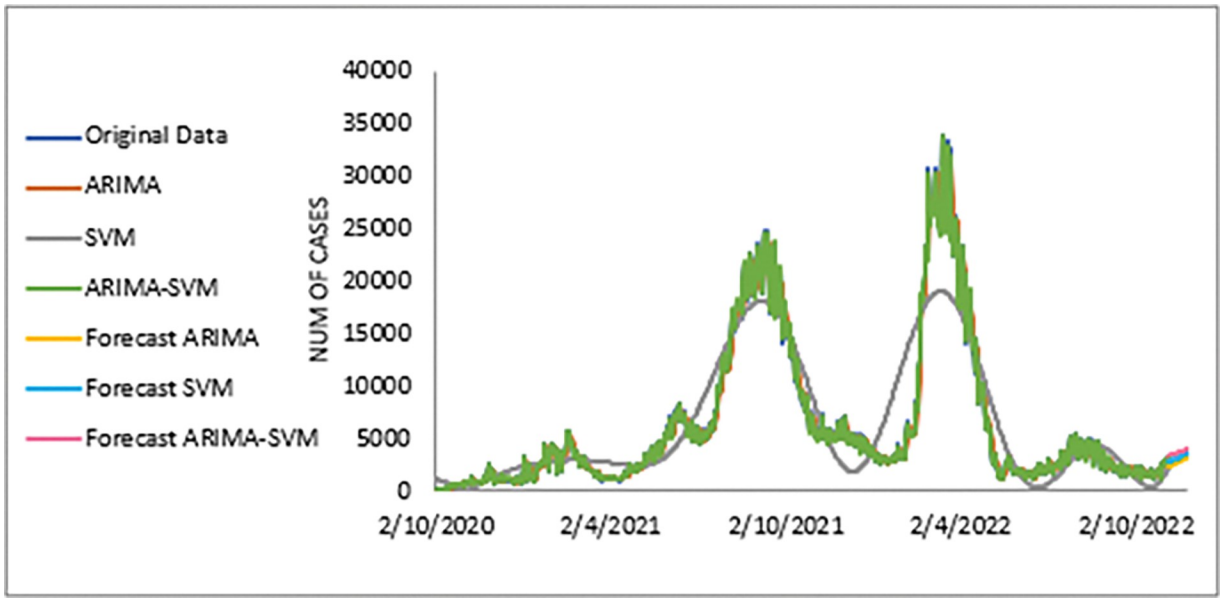

Fig 19. Actual and three weeks ahead forecasted values of ARIMA, SVM and ARIMA SVM models for daily new recovered COVID-19 cases of the 80% training and 20% testing set.

https://doi.org/10.1371/journal.pone.0285407.g019

Table 10. Percentage improvement of the proposed models with other forecasting models (The COVID-19 cases of daily new recovered cases).

Model	MAE	MAPE	MSE	RMSE
ARIMA	73.1199	74.6153	90.3824	68.9878
SVM	71.9902	73.6702	89.1067	66.9951

https://doi.org/10.1371/journal.pone.0285407.t010

forthcoming three weeks and indicates that daily new recovered COVID-19 cases would increase in the upcoming days in Malaysia.

The performance of the proposed models for the daily new recovered COVID-19 cases datasets was further investigated for MSE, MAPE, RMSE and MAE in terms of the percentage, as reported in Table 10. By looking at the percentage of improvement for statistical measurements such as MSE, MAPE, RMSE and MAE, the results observed for the proposed model show a better improvement compared to ARIMA and SVM, respectively, with results of 73.12%, 74.62%, 90.38% and 68.99% improvement (71.99%, 73.67%, 89.11% and 66.99%) (where the results reported in the parenthesis are the SVM model). Therefore, based on the results, it can be concluded that the proposed model that has been developed has produced higher accuracy and efficiency compared to the results achieved by ARIMA and SVM models.

Conclusion

Accuracy and efficiency in predicting the spread of COVID-19 is crucial but often difficult for decision makers, especially the frontline and authorities. Although the spread of COVID-19 seems to be endless, but many efforts in the development of time series models, research to improve the effectiveness of forecasting models has never stopped. Among them is the hybrid approach and one of the most popular categories of hybrid models that decompose time series into linear and non-linear forms. in this study, a hybrid model as a combination of predictions produced by linear and some non-linear is proposed. The proposed model was investigated using three well-known COVID-19 data sets, namely, daily new positive cases, daily new death

cases and daily new recovered cases based on (1) performance of the proposed model and (2) percentage improvement compared to ARIMA and SVM models. The proposed model with cross-validation check based on MSE, RMSE, MAE and MAPE is the most accurate prediction compared to ARIMA and SVM models. The performance of the proposed models produces the smallest values of MSE, RMSE, MAE and MAPE for both training and testing datasets. This means, the predicted value from the proposed model is closer to the actual value. In other words, the proposed model can generate estimated values more accurately and efficiently. In addition, percentage improvement of the proposed models against the ARIMA and SVM models (where the results reported in the parenthesis is SVM model) are 63.03%, 62.86%, 79.52%, 54.74% improvement, (62.34%, 63.47%, 77.70%, 52.78%); 60.46%, 66.42%, 84.73%, 60.90% improvement (58.93%, 64.45%, 82.81%, 58.52%) and 73.12%, 74.62%, 90.38% and 68.99% improvement (71.99%, 73.67%, 89.11% and 66.99%) for daily new positive cases, daily new deaths cases and daily new recovered cases, respectively. Therefore, our proposed models showed higher degree of precision and could be recommended for forecasting COVID-19. It can be concluded that the proposed model can be the best and effective way to improve the prediction accuracy performance, especially to predict and prevent the infection of COVID-19 cases is a priority.

Limitations and future recommendation

An effort was made in this research study to forecast the total number of confirmed cases, fatalities, and recoveries of COVID-19 in Malaysia. Nowadays, the change in daily numbers of COVID-19 is affected by a very large number of factors, such as the population's adherence to prevention measures, vaccination, social isolation, and new variants of the virus. As such, in order to improve future predictions and forecasts, it is imperative that the study of COVID-19 be taken into consideration in terms of (i) the clinical and behavioural aspects, and (ii) the possibility of underreporting cases, deaths, or delays in notifying as part of the study of COVID-19 in the future. Besides that, to improve the accuracy of the forecast in future work, investigation in SVM performance with different kernel functions and optimal hyper parameters of SVM forecasting model can be developed. Next, multi-step forecasts can be centralized in the future work since only one-step- ahead forecasting is considered in this paper. It is proven that multi-step forecasts can make the trading system much more realistic [38]. Finally, another approach, such as bootstrapping, can also be added as a hybridization of ARIMA and SVM [39]. Bootstrap is a reliable method given the lack of researchers adding this method in daily cases of COVID-19 forecasting. Many studies have shown that the bootstrap resampling technique provides a more accurate estimation [17, 40–42].

Supporting information

S1 Dataset.
(XLSX)

Acknowledgments

The authors would like to express his gratitude to the Research Management Centre, Universiti Malaysia Terengganu (UMT) for partially grant of the journal publication fee as well as to the editors and the referees for careful reading and for comments which greatly improved the paper.

Author Contributions

Conceptualization: Wan Imanul Aisyah Wan Mohamad Nawi, Muhamad Safiih Lola, R. U. Gobithaasan, Nurul Hila Zainuddin, Mohd Lazim Abdullah, Nor Aieni Mokhtar, Mohd Tajuddin Abdullah.

Data curation: Wan Imanul Aisyah Wan Mohamad Nawi, Nor Aieni Mokhtar.

Formal analysis: Wan Imanul Aisyah Wan Mohamad Nawi, Muhamad Safiih Lola, Syerrina Zakaria, Elayaraja Aruchunan, Nurul Hila Zainuddin, Mohd Lazim Abdullah.

Funding acquisition: Abdul Aziz K. Abdul Hamid, Syerrina Zakaria, Elayaraja Aruchunan, R. U. Gobithaasan, Wan Azani Mustafa, Mohd Lazim Abdullah, Nor Aieni Mokhtar, Mohd Tajuddin Abdullah.

Investigation: Muhamad Safiih Lola, Nurul Hila Zainuddin, Mohd Tajuddin Abdullah.

Methodology: Wan Imanul Aisyah Wan Mohamad Nawi, Muhamad Safiih Lola, Syerrina Zakaria, Elayaraja Aruchunan, R. U. Gobithaasan, Nurul Hila Zainuddin, Mohd Lazim Abdullah.

Project administration: Muhamad Safiih Lola.

Resources: Elayaraja Aruchunan.

Software: Abdul Aziz K. Abdul Hamid, R. U. Gobithaasan, Wan Azani Mustafa.

Supervision: Muhamad Safiih Lola.

Validation: Wan Imanul Aisyah Wan Mohamad Nawi, Muhamad Safiih Lola, Elayaraja Aruchunan, Nurul Hila Zainuddin, Mohd Tajuddin Abdullah.

Visualization: Wan Imanul Aisyah Wan Mohamad Nawi, Abdul Aziz K. Abdul Hamid, Muhamad Safiih Lola, Wan Azani Mustafa.

Writing – original draft: Wan Imanul Aisyah Wan Mohamad Nawi, Muhamad Safiih Lola.

Writing – review & editing: Muhamad Safiih Lola, Mohd Lazim Abdullah, Nor Aieni Mokhtar, Mohd Tajuddin Abdullah.

References

1. Mohd Tajuddin A., Muhamad Safiih L., Hisham AE, Sabreena S, Nor Fazila CM, Idham K, et al. Framework of Measures for COVID-19 Pandemic in Malaysia: Threats, Initiatives and Opportunities. Journal of Sustainability Science and Management. 2022; 17(3):8–18.3
2. Ali M, Khan DM, Aamir M, Khalil U, Khan Z. Forecasting COVID-19 in Pakistan. PLoS One. 2020; 15 (11): e0242762. https://doi.org/10.1371/journal.pone.0242762 PMID: 33253248
3. WHO. (2020). Coronavirus disease (COVID-19) in Malaysia. Accessed on 23 May 2020, from https://www.who.int/malaysia/emergencies/coronavirus-disease-(covid-19)-in-Malaysia.
4. KKM. (2020b). COVID-19 Malaysia: Situasi Terkini 25 Oktober 2020. Accessed on 25 June 2022, from covid-19.moh. gov.my/archive:June_2022.
5. Gecili E, Ziady A, Szczesniak RD Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. PLoS ONE, 2021; 16(1): e0244173. https://doi.org/10.1371/journal.pone.0244173
6. Awwad FA, Mohamoud MA, Abonazel MR Estimating COVID-19 cases in Makkah region of Saudi Arabia: Space-time ARIMA modeling. PLoS ONE, 2021; 16(4): e0250149. https://doi.org/10.1371/journal.pone.0250149
7. Sahai AK., Rath N., Sood V., Singh MP. ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2020; 14(5), 1419–1427. https://doi.org/10.1016/j.dsx.2020.07.042 PMID: 32755845

8. Alzahrani SI., Aljamaan IA., Al-Fakih EA. Forecasting the Spread of The COVID-19 Pandemic In Saudi Arabia Using ARIMA Prediction Model Under Current Public Health Interventions. J Infect Public Health. 2020; 13: 914–919. https://doi.org/10.1016/j.jiph.2020.06.001 PMID: 32546438

9. Benvenuto D., Giovanetti M., Vassallo L., Angeletti S., Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in Brief. 2020; 105340. https://doi.org/10.1016/j.dib.2020.105340 PMID: 32181302

10. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of The Total Environment. 2020; 138817. https://doi.org/10.1016/j.scitotenv.2020.138817 PMID: 32360907

11. Hernandez-Matamoros A., Fujita H., Hayashi T., Perez-Meana H. Forecasting of COVID19 per regions using ARIMA models and polynomial functions. Applied Soft Computing. 2020; 106610. https://doi.org/10.1016/j.asoc.2020.106610 PMID: 32834798

12. Khan FM., Gupta R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. Journal of Safety Science and Resilience. 2020; 1, 12–18. https://doi.org/10.1016/j.jnlssr.2020.06.007

13. Kayode O., Fahimah A., Mustapha R., Jacques D. Data Analysis and Forecasting of COVID-19 Pandemic in Kuwait Based on Daily Observation and Basic Reproduction Number Dynamics. Kuwait J. Sci. Special Issue. 2021; 1–28. https://doi.org/10.48129/kjs.splcov.14501

14. Rahman MS, Chowdhury AH., Amrin M. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. PLOS Glob Public Health. 2022; 2(5): e0000495. https://doi.org/10.1371/journal.pgph.0000495 PMID: 36962227

15. Singh S, Murali Sundram B, Rajendran K, Boon Law K, Aris T, Ibrahim H, et al. Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. J Infect Dev Ctries. 2020 Sep 30; 14(9):971–976. https://doi.org/10.3855/jidc.13116 PMID: 33031083.

16. Aisyah WI WMN, Muhamad Safiih L, Razak Z, Nurul Hila Z, Abd. Aziz KAH, Elayaraja A, et al. Improved of Forecasting Sea Surface Temperature based on Hybrid ARIMA and Vector Machines Model. Malaysian Journal of Fundamental and Applied Sciences. 2021; 17:609–620. https://doi.org/10.11113/mjfas.v17n5.2356

17. Nurul Hila Z., Muhamad Safiih L., Maman Abdurachman D., Fadhilah Y., Mohd Noor Afiq R., Aziz D., et al. Improvement of Time Forecasting Models using A Novel Hybridization of Bootstrap and Double Bootstrap Artificial Neural Networks. Applied Soft Computing Journal. 2019; 105676. https://doi.org/10.1016/j.asoc.2019.105676

18. Lee MC. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Journal of Expert Systems with Applications. 2009; 36(8): 10896–10904. https://doi.org/10.1016/j.eswa.2009.02.038

19. Vapnik VN. The Nature of Statistical Learning Theory. 1st Edn., Springer-Verlag, New York, USA; 1995.

20. Sudheer C., Maheswaran R., Panigrahi BK Mathur S. A hybrid SVM-PSO model for forecasting monthly streamflow. Neural Computing and Applications. 2013; 24(6), 1381–1389. https://doi.org/10.1007/s00521-013-1341-y

21. Chakraborty T., Chakraborty AK., Biswas M., Banerjee S. & Bhattacharya S. Unemployment Rate Forecasting: A Hybrid Approach. Computational Economics. 2020; 57:183–201 https://doi.org/10.1007/s10614-020-10040-2

22. Zhang GP. Time series forecasting using a hybrid ARIMA and Neural Network Model.Neurocomputing. 2003; 50: 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

23. Terui N., Van Dijk H. Combined forecasts from linear and nonlinear time series models. International Journal of Forecasting. 2002; 18(3): 421–438. https://doi.org/10.1016/s0169-2070(01)00120-0

24. Wang X., Meng M. A Hybrid Neural Network and ARIMA Model for Energy Consumption Forecasting. Journal Of Computers. 2012; 7(5): 1184–1190. https://doi.org/10.4304/jcp.7.5.1184–1190

25. Pai PF. Lin C.-S. A hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting. International Journal of Management Science. 2005; 3(3): 497–505. https://doi.org/10.1016/j.omega.2004.07.024

26. Lee N-U., Shim J-S., Ju Y-W. Park S-C. Design and Implementation of the SARIMA–SVM time series analysis algorithm for the improvement of atmospheric environment forecast accuracy. Soft Computing. 2017; 22(13): 4275–4281. https://doi.org/10.1007/s00500-017-2825-y

27. Hao Y, Xu T, Hu H, Wang P, Bai Y Prediction and analysis of Corona Virus Disease 2019. PLoS ONE. 2020; 15(10): e0239960. https://doi.org/10.1371/journal.pone.0239960

28. Roy S, Ghosh P Factors affecting COVID-19 infected and death rates inform lockdown- related policymaking. PLoS ONE. 2020; 15(10): e0241165. https://doi.org/10.1371/journal.pone.0241165

29. Mahdavi M, Choubdar H, Zabeh E, Rieder M, Safavi-Naeini S, Jobbagy Z, et al. A machine learning based exploration of COVID-19 mortality risk. PLoS ONE. 2021; 16(7): e0252384. https://doi.org/10.1371/journal.pone.0252384 PMID: 34214101

30. Singhal T. A Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr. 2020; 87, 281–286. https://doi.org/10.1007/s12098-020-03263-6 PMID: 32166607

31. Moore Sarah. (2022, January 17). The Future of Pandemics. News-Medical. Retrieved on November 05, 2022 from https://www.news-medical.net/health/The-Future-of-Pandemics.aspx.

32. Naeem M, Yu J, Aamir M, Khan SA, Adeleye O, Khan Z. Comparative analysis of machine learning approaches to analyse and predict the COVID-19 outbreak. Peer J Comput. Sci. 2021; 17: e746 https://doi.org/10.7717/peerj-cs.746 PMID: 35036527

33. Qiang X, Aamir M, Naeem M, Ali S, Aslam A, Shao Z., Analysis and Forecasting COVID-19 Outbreak in Pakistan Using Decomposition and Ensemble Model. Computers, Materials & Continua. 2021; 68(1): 842–856. https://doi.org/10.32604/cmc.2021.012540

34. Muhamad Safiih L., Nurul Hila Z., Mohd Tajuddin A., Vigneswary P., Mohd Noor Afiq R., Razak Z., et al. Improving the Performance of ANN-ARIMA Models for Predicting Water Quality in The Offshore Area of Kuala Terengganu, Terengganu, Malaysia. Journal of Sustainability Science and Management. 2018; 13(1): 27–37

35. Adhikari SP., Meng., Wu Y-U., Mao Y-P., Ye R-X., Wang Q-Z., et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Infectious Diseases of Poverty. 2020; 9(1): 29 https://doi.org/10.1186/s40249-020-00646-x PMID: 32183901

36. Ahmadini AAH, Naeem M, Aamir M, Dewan R, Alshqaq SSA and Mashwani WK Analysis and Forecast of the Number of Deaths, Recovered Cases, and Confirmed Cases From COVID-19 for the Top Four Affected Countries Using Kalman Filter. Front. Phys. 2021; 9:629320. https://doi.org/10.3389/fphy.2021.629320

37. Alessa AA., Alotaibie TM., Elmoez Z., Alhamad HE. Impact of COVID-19 on Entrepreneurship and Consumer Behaviour: A Case Study in Saudi Arabia. The Journal of Asian Finance, Economics and Business. 2021; 8(5), 201–210. https://doi.org/10.13106/JAFEB.2021.VOL8.NO5.0201

38. Huck N. Pairs trading and outranking: The multi-step-ahead forecasting case. European Journal of Operational Research. 2010; 207(3): 1702–1716. https://doi.org/10.1016/j.ejor.2010.06.043

39. Nurul Hila Z Muhamad Safiih. The Performance of BB-MCEWMA Model: Case Study on Sukuk Rantau Abang Capital Berhad, Malaysia. International Journal of Applied, Business and Economic Research. 2016; 14(2): 63–77

40. Nurul Hila Z., Muhamad Safiih L. Nur Shazrahanim K. Modelling Moving Centreline Exponentially Weighted Moving Average (MCEWMA) with bootstrap approach: Case study on sukuk musyarakah of Rantau Abang Capital Berhad, Malaysia. International Journal of Applied, Business and Economic Research. 2016; 14(2): 621–638.

41. Muhamad Safiih L., Nurul Hila Z., Mohd Noor Afiq R., Hizir S. Double Bootstrap Control Chart for Monitoring SUKUK Volatility at Bursa Malaysia. Jurnal Teknologi. 2017; 79 (6):149–157. https://doi.org/10.11113/jt.v79.10410

42. Nisbet R., Elder J. Miner G. Chapter 11—Model Evaluation and Enhancement. In: Handbook of Statistical Analysis and Data Mining Applications. 2018; pp. 215–233. https://doi.org/10.1016/B978-0-12-374765-5.00013–9