

Covid-19 forecasting model based on machine learning approaches: a review

Md Shohel Sayeed, Siti Najihah Hishamuddin, Ong Thian Song

Faculty of Information Science and Technology (FIST), Multimedia University, Melaka, Malaysia

Article Info

Article history:

Received Aug 1, 2023

Revised Apr 3, 2024

Accepted May 17, 2024

Keywords:

Analysis

Covid-19

Forecasting

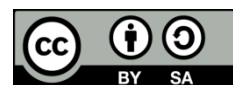
Machine learning

Predictive modeling

ABSTRACT

As coronavirus disease (Covid-19) it is a contagious disease that is spread by the SARS-CoV-2 virus, one of the most common causes of disease in humans. The disease was initially discovered in Wuhan, China, in 2019, and has now spread throughout the world, including Malaysia. A large number of people have lost their life partners and families because of this disease. Thus, in order for us to stop this epidemic spread, we have to implement social distance. The Covid-19 infection displays this type of behavior, which necessitates the development of mathematical and predictive modeling techniques capable of predicting possible disease patterns or trends, in order to assist the government and health authorities in predicting and preparing for potential outbreaks. The purpose of this paper is to provide an in-depth critique and analysis of the machine-learning approaches that have been implemented by researchers to predict Covid-19, based on existing research. As a result, future researchers will be able to use this paper as a valuable resource for their research related to the Covid-19 forecasting model.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Md Shohel Sayeed

Faculty of Information Science and Technology (FIST), Multimedia University

St. Ayer Keroh Lama, Bukit Beruang, Ayer Keroh, Melaka 75450, Malaysia

Email: shohel.sayeed@mmu.edu.my

1. INTRODUCTION

Coronavirus infection more known as Covid-19 is an illness caused by a novel coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) or 2019-nCoV [1]. Since the Covid-19 virus spread to over 223 countries around the world, more than a hundred million people have been infected and more than 6 million have been killed as of January 2023. This disease was first discovered in Wuhan, China on Dec. 31, 2019, and emerged in Wuhan, China, resulting in a formidable outbreak in many Chinese cities and spreading globally, including Thailand, Korea, the United States, and our own country, Malaysia. Covid-19 has a wide range of clinical symptoms, including fever, dry cough, and exhaustion, which are frequently associated with pulmonary involvement. Most people in the general population are susceptible to infection due to the high contagiousness of SARS-CoV-2. The disease, which is spread by respiratory droplets and direct contact, primarily originates from infected individuals and wild animal hosts [2].

Many preventive measures were taken by the whole world in order to prevent this virus outbreak. Some of the preventive measure taken to stop the Covid-19 pandemic from spreading, including enlarging quarantines, promoting social distancing, enhancing healthcare infrastructure, and permitting only necessary goods to leave the home [3]. Often washing your hands, using hand sanitizer, donning a face mask, and avoiding touching your lips or face after being in a location that may be contaminated are some of the most important preventative measures people can take. All of these preventive measures were available because no

one drug had been licensed by the food and drug administration (FDA), subjected to controlled research, or proven to have an effect on the virus that is causing the global pandemic [4].

Although most country already entering the endemic phase, Covid-19 cases still exist in many countries. Taking into account Covid-19's behavior, it is necessary to develop mathematical and predictive modeling methods capable of predicting possible disease patterns or trends, to help our government and health authorities prepare for this epidemic. There are various prediction models from machine learning that we can choose to predict or forecast Covid-19 positive cases such as autoregressive integrated moving average (ARIMA), artificial neural networks (ANN), regression models, and many more. To investigate and analyze which models or approaches would be the most appropriate to combine for this new framework, some research was conducted on existing approaches to forecasting the Covid-19 outbreak. Since each model and approach has its own limitations, and such as the models only working with large datasets.

The use of machine learning in real-time forecasting has proven to be one of the most impactful methods. To develop a real-time forecasting framework, several machine learning techniques can be used. Due to these circumstances, the objective of this review paper is to examine published journals and papers from other researchers in an effort to uncover which machine learning algorithms have been rarely or repeatedly used by researchers to predict Covid-19. Herein, we review two types of approaches that researchers use for forecasting or predicting Covid-19 positive cases which are hybrid model approaches and single model approaches.

2. LITERATURE REVIEW

2.1. Single model approaches

Artificial intelligence and machine learning (AIML) great learning researchers proposed a forecasting model based on the ARIMA model. The authors concluded that the ARIMA model outperformed other similar models such as support vector machines and wavelet neural networks. This was based on the findings of this study. The data used is from the first phase of the Indian lockdown. This data is used to gain a better understanding of how the lockdown affects the rate of increase in the number of cases. This prediction model was compared to other project models that used the modified logistic growth model. Data collected prior to the lifting of the lockdown in India were analyzed. Results of this study indicate that, except for confirmed cases, predictions are well within the range of a 95% confidence interval using ARIMA [5].

Singh *et al.* [6] proposed an ARIMA-based framework for forecasting Covid-19 cases in Malaysia. They select the best-fitting model based on the Bayesian information criteria (BIC) and mean absolute percentage error (MAPE). Since this study was conducted in 2020, the data is limited; thus, the ARIMA models are an excellent choice because they can provide accurate results though with limited data.

In a separate study, Singh *et al.* [7] proposed the use of an advanced ARIMA model to forecast the spread of the Covid-19 disease over the next two months. This model is intended to forecast the spread of the disease among the world's top 15 countries. They used the Akaike information criterion (AIC) for data that was available at the time to validate the ARIMA model during that time period [7].

Another study that utilized the ARIMA model was conducted by Januri *et al.* [8]. This study was conducted in Malaysia, and five models were developed using the Box-Jenkins approach. They compare the value of AIC, BIC, root mean square error (RMSE), and Ljung-Box test statistics to determine the best model. They examined the p-value of the Ljung-Box test and discovered that four out of five models had probability values greater than 0.05, indicating that four models had no serial correlation. As a result, they chose ARIMA (1, 1, 3) as their final forecasting model due to its low AIC, BIC, and RMSE values [8].

A prediction model using ARIMA has been proposed by Ratu *et al.* [9] for the prediction of Covid-19 positive cases in Bangladesh. Through the analysis of the AIC values of the models, they were able to compare and select the best model. Considering ARIMA is the most commonly used model to forecast future data for time series data, they have chosen it as their proposed model. It was also discovered that the ARIMA model closely matched the distribution in Bangladesh [9].

A group of Malaysian researchers proposed developing the seasonal autoregressive integrated moving average model, abbreviated as the SARIMA model. In contrast to ARIMA, this model allows for the explicit inclusion of seasonal components in univariate time series data. They create 3 models with SARIMA and use the Ljung Box test to select the best one. They choose the SARIMA model to improve model accuracy, and SARIMA can provide a 28-day forecast for Covid-19 [10].

A study by Mofthakhar *et al.* [11] proposed to use of the ARIMA model and ANN in predicting the positive cases in Iran. This study was conducted in order to forecast the number of daily positive cases over the next 30 days. They compared the values of mean square error (MSE) and mean absolute error (MAE) for both models to choose a model with great accuracy. It should be noted, however, that although this study is

well conducted, they assumed that their models might not be well trained due to the limited number of observations provided by this type of prediction algorithm.

Another research study from a group of researchers at JSS Science and Technology University, India proposed two supervised learning models to predict the Covid-19 pandemic. They choose linear regression (LR) and support vector regression (SVR) and make a comparison between both models to study the performance of the prediction that they made. As a result of this study, they concluded that LR algorithms perform better as a linear dataset was used, whereas SVR algorithms cannot handle large linear datasets very well [12].

A genetic programming-based (GP) prediction model proposed by Salgotra *et al.* [13] to predict Covid-19 cases in India. This paper presents explicit formulas for the proposed prediction models, as well as assesses the ineffectiveness of the prediction variables. Input variables and metrics were used to evaluate and validate the models. In accordance with the findings of the study, the proposed GEP-based models utilize simple linkage functions and are highly capable of predicting the time series of Covid-19 cases in India.

The study was conducted in Kuwait by a group of researchers who studied mathematical modeling in order to provide a real-time real-time monitoring and predicting tool for this Covid-19. For infectious disease transmission, they used a deterministic compartmental model, and the model will simulate the SEIR model, testing, and hospitalization dynamics. In this study, it was found that early gradual and aggressive control measures delayed and reduced the severity of the pandemic by protecting a significant percentage of the population [14].

A group of researchers from Turkey proposed long short-term memory (LSTM) networks model to forecast Covid-19 cases in Istanbul, Turkey. They compared their model with the other three existing models. The LSTM's neural network showed a better performance than the ARIMA model (6, 1, 0) and Prophet model. Despite this, Holt-Winters' additive method with a damped trend outperformed LSTM networks in predicting Covid-19 cases [15].

Yuan *et al.* [16] proposed an internet search-interest-based model to predict the daily new cases and deaths of Covid-19 in the United States. Google Trends search data was used to estimate US Covid-19 cases and deaths. The results suggested that internet search data could predict Covid-19 infections and death rates. The model's presented patterns for new cases and deaths were close to the actual trends, indicating good predictive capability.

Zou [17] used a machine learning model to predict the US Covid-19 epidemic. The author developed a machine learning model that predicted US confirmed cases, deaths, and hospitalisations using real-world data. The model predicted Covid-19 case trends with an average error of 5%, according to the findings of the study. The author also stated the model could demonstrate how public health interventions like locking individuals up and keeping them apart from the effect of the disease transmission [17].

Mustafa and Fareed [18] suggested a Box-Jenkins ARIMA model to predict Covid-19 in Iraq. The Iraqi Ministry of Health and World Health Organization (WHO) provided Covid-19 case data from February 24 to May 20, 2020. The research found that the Box-Jenkins ARIMA model could accurately predict Covid-19 cases in Iraq.

Podder and Mondal [19] suggested a machine learning model to forecast Covid-19 cases and intensive care unit (ICU) requirements. Decision trees, random forests, and SVM were utilized to predict Covid-19 cases and ICU needs. The SVM method predicted Covid-19 cases and ICU needs. The SVM method may predict Covid-19 cases and requirements for ICU care. A stacking ensemble used with random forests also been tested and concluded with the great accuracy in order to forecast Covid-19.

A model proposed by Saadah and Permana [20] is used for predicting Covid-19 cases in Aceh, Indonesia using a fuzzy time series technique. This model was used to create fuzzy time series estimated of daily Covid-19 cases in Aceh from March 10 to November 29, 2020. The fuzzy time series technique utilising Chen and Lee's model which cab able to predict Aceh Covid-19 cases with high accurately.

2.2. Hybrid model approaches

A study conducted by Safi and Sanusi [21] proposed a hybrid model that combines three models, namely ETS, ARIMA, and ANN. A comparison was conducted between their hybrid model and a single model that included ETS, ARIMA, and ANN in order to determine which model works best for their prediction model. It was found that ARIMA and ETS were more effective than ANNs and hybrid models based on the results of this study [21].

The SutteARIMA method was proposed by Ahmar and del Val [22] to forecast short-term Covid-19 cases. SutteARIMA obtains average forecasting results by combining the Sutte indicator with ARIMA. The comparison of the two models revealed that SutteARIMA has a lower MAPE value (0.036) than the ARIMA model, making the SutteARIMA model more suitable for predicting Covid-19.

Poleneni *et al.* [23] proposed a study to forecast Covid 19 by using a combination of two methods that are ARIMA model and the Facebook Prophet. They choose to combine these two models because they

believe that this combination will give the highest accuracy for the prediction. This study forecasts the overall number of active cases in India for the next 15 days.

Kumar and Kaur [24] developed a hybrid technique for future forecasting of Covid-19 cases that utilizes self-organized maps and fuzzy time series (SOMFTS). In their study, they chose these approaches since previous research had not taken them into account. As a result of the experiment, the proposed SOMFTS technique is shown to be most effective for forecasting Covid-19 cases in the future. Their forecasting model ranks at the top when more than one conflicting performance measure is present for both confirmed cases and cured cases in Delhi.

The study by Prasanth *et al.* [25] developed a hybrid model of LSTM and grey wolf optimizer (GWOLSTM) to predict the future accumulated cases of Covid-19, new cases, and deaths related to Covid-19. MAPE was reduced by 74% overall by using related Google Trends for particular keywords related to the disease outbreak. The accuracy of their forecasting of pandemic cases was confirmed by this study [25].

An ensemble machine-learning approach proposed by Maaliw *et al.* [26] for Covid-19 time series forecasting. In order to predict the number of Covid-19 cases in the Philippines, the authors used real-world data from their country and trained multiple machine learning models, including SVR, decision trees, random forests, and ANN. It was found that a combination of multiple machine learning models resulted in a higher level of accuracy and robustness than a single model. Accordingly, Maaliw *et al.* [26] ensemble machine learning method provides a promising method for forecasting Covid-19 spread and might prove beneficial for policymakers and public health officials in developing effective mitigation strategies [26].

A hybrid artificial intelligence (AI) model was developed by Zheng *et al.* [27] to predict Covid-19 cases in China. This model comprises three components: the LSTM neural network, the gradient boosting decision tree (GBDT) algorithm, and the extreme gradient boosting (XGBoost) algorithm. LSTM was also observed to play a key role in capturing the temporal dynamics of the pandemic, whereas the GBDT and XGBoost components provided additional information to enhance the accuracy of the predictions.

3. METHOD, DATASET, AND EVALUATION METRICS

Researchers have used a variety of approaches and techniques in their current research to forecast and predict Covid-19 cases. Some researchers stated that their research was limited in scope due to the timeline of this pandemic disease and lacked adequate datasets and variables [28]. Choosing a good and excellent model to forecast Covid-19 cases or any pandemic disease requires taking into consideration several factors. These factors ensure that our accuracy and results will be accurate and reliable.

Table 1 summarises a wide range of machine-learning methodologies and techniques used in Covid-19 forecasting. Researchers' methodology varies, with some using hybrid models that incorporate numerous techniques in order to capitalize on the benefits of each approach. Other researchers, on the other hand, prefer to use a single model for all of their forecasting efforts. The decision between these methodologies is frequently determined by the specific properties of the data, the required level of model interpretability, and the forecasting study's broader objectives. An exploration of the datasets utilized in these studies is presented in detail, illuminating the context that influences the predictive models. This table also summarizes the different models used, providing researchers with a holistic understanding of the varied techniques used in Covid-19 prediction, and providing an invaluable resource.

3.1. Hybrid model approaches for Covid-19 forecasting

A hybrid model can give a lot of different results than using a single model. Hybrid models are typically a combination of multiple single methods, where each method contributes to improved precision and effectiveness in forecasting [29]. The advantage of choosing a hybrid model for this forecasting model is that each model can support the limitations of the other. These hybrid forecasting methods could lead to more accurate predictions as well as augment and improve visual analytics tools for making judgments or assisting decision-making processes. For example, hybrid methods can provide a solution to the issue of linearity assumed in linear methods [30]-[32]. Some examples of hybrid approaches selected by researchers with the type of approaches chosen are presented in Table 1.

3.2. ARIMA model

George E. P. Box and Gwilym M. Jenkins developed the ARIMA model in 1970, which is widely used for forecasting both stationary and non-stationary data series [33], [34]. ARIMA models measure the strength of a dependent variable in comparison to other variables that may change during the course of the analysis. The ARIMA model can deliver accurate short-term forecasts with variable assumptions and excellent forecasting performance [35]. Another rationale for using the ARIMA model is that it still provides good accuracy despite being used by numerous researchers.

Table 1. Some of the existing approaches and techniques for forecasting Covid-19

References	Models used	Types of approaches	Result (best model)	Dataset (country)
[5], [6]	ARIMA model	Single model	CI: 95% R2: 0.994 MAPE: 16.01 BIC: 4.170	India [5] Malaysia [6]
[7]	Advanced ARIMA model	Single model	CI: 95%	Top 15 countries (USA, Spain, Italy, China, and more)
[8]	ARIMA (Box-Jenkins model)	Single model	AIC: 3762.45 BIC: 3780.73 RMSE: 170.03	Malaysia
[9]	ARIMA	Single model	AIC: 818.5061	Bangladesh
[10]	SARIMA model	Single model	MAE: 39.716 RMSE: 73.374	Malaysia
[11]	ARIMA model	Single model	MAE: 52.51	Iran
[12]	ANN	Single model	MAE: 24.85	
[12]	LR	Single model	R2: 0.989	India
[12]	SVR	Single model	R2: 0.806	
[13]	GP	Single model	RMSE: 5.5574 R: 0.9999	India
[14]	SEIR model	Single model	CI: 95%	Kuwait
[15]	LSTM	Single model	MAPE: 0.99±0.51%	Istanbul, Turkey
[17]	Regression model	Single model	Average error: 5%	USA
[18]	Arima (Box-Jenkins)	Single model	MAE: 0.253 BIC: -1.819	Iran
[19]	Stacking ensemble with random forest	Single model	Accuracy: 94.39% Recall: 92%	Brazil
[20]	Fuzzy time series	Single model	MAPE: 7.94%	Indonesia
[21]	ETS, ARIMA, and ANN	Hybrid model	MAE: 1.5638 RMSE: 1.5700	Whole world Covid's data
[22]	SutteARIMA model	Hybrid model	MAPE: 0.036	Spain
[23]	ARIMA and Facebook Prophet	Hybrid model	R2: 0.96 RMSE: 10096.15	India
[24]	SOMFTS	Hybrid model	NRMSE: 0.3015	Delhi, India
[25]	LSTM and GWO	Hybrid model	MAPE: 55.74 (India) MAPE: 4.71 (USA) MAPE: 52.95 (UK)	India, USA, UK
[26]	S-LTSM and ARIMA	Hybrid model	Accuracy: 93.50% (infected) Accuracy: 87.97% (death)	Philippines, India, Brazil, and the United States

ARIMA components are included in the model as a parameter. These parameters are assigned specific integer values that indicate the type of ARIMA model. The parameters are p, d, and q notation. This notation is a standard notation that will substitute integer values for parameters in order to indicate the type of ARIMA model that has been used. These parameters must be defined by integers in order for the model to function. ARIMA can be shown in (1):

$$yt = \phi_0 + \phi_1 yt - 1 + \phi_2 yt - 2 + \dots + \phi_p yt - p + \epsilon t \quad (1)$$

where $yt, yt - 1, yt - 2, yt - p$ are stationeries while $\phi_0, \phi_1, \phi_2, \phi_p$ are constants, and ϵt is Gaussian white noise series with mean zero.

3.3. SARIMA model

Seasonal autoregressive integrated moving averages, or SARIMAs, are an extension of ARIMAs that are designed to provide support for multivariate regression time series data that have seasonal characteristics. This method is extended from ARIMAs to support data that have seasonal characteristics [36], [37]. The model, designated as ARIMA (p, d, q) ARIMA (P, D, Q) S as (2):

$$\nabla d \nabla S y_t = \frac{\theta(B) \times \Theta S(B)}{\Phi(B) \times \Phi S(B)} \epsilon t \quad (2)$$

where ∇d is the difference operator: $(1 - B)d$,

$\nabla D S$ is the seasonal difference operator: $(1 - BS)D$, (B) is the moving average polynomial: $1 - \theta_1 B - \dots - \theta_q B^q$,

$S(B)$ is the seasonal moving average polynomial: $1 - \theta_1 BS - \dots - \theta_Q B^Q S$,

(B) is an autoregressive polynomial: $1 - \phi_1 B - \dots - \phi_p B^p$,

$S(B)$ is a seasonal autoregressive polynomial: $1 - \phi_1 BS - \dots - \phi_P B^P S$.

3.4. Artificial neural network

This ANN is an attempt to forecast future events utilizing simple mathematical models of the brain. The relationships between the response variable and its predictors can be complex and nonlinear. The most widely used ANNs in forecasting time series data are the multiple layer perceptron (MLP). LSTM is also one of the deep learning techniques based on artificial recurrent neural networks. Unlike feedforward neural networks, this model ensures feedback connectivity. The LSTM technology is not limited to processing individual data points, but can also process entire sequences of data [38], [39].

The model is composed of three layers: input, hidden, and output, each connected by acyclic links. There might be multiple hidden layers (Figure 1). The output of the model can be calculated using the following mathematical as (3):

$$y_t = \alpha_0 + \sum_{j=1}^q \alpha_j g(\beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i}) + \varepsilon_t \quad (3)$$

where $\alpha_j(j=0; 1; 2; \dots; q)$ and $\beta_{ij}(i=0; 1; 2; \dots; p; j=1; 2; \dots; q)$ are the model for the parameters that often been called as the connection weights; parameter p: number of input nodes and parameter q: number of hidden nodes [38].

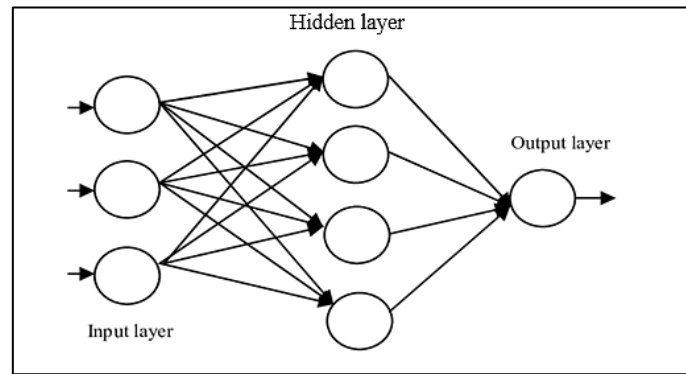


Figure 1. Example of ANN basic architecture

3.5. Regression method

Regression analysis can be defined as a statistical approach to estimating the relationship between a dependent variable and one or more independent variables, using a variety of statistical techniques to achieve this goal [40]. There are many types of regression models used for forecasting such as SVR and LR. This regression algorithm uses a supervised learning algorithm since it has two input factors (X) and one output component (Y) that it uses to learn the mapping from input to output. Below is the simple formula used for a simple LR as (4):

$$y = bx + a \quad (4)$$

where y represents the value that we are attempting to forecast, x represents the value of the independent value, and a represents the y-intercept [41].

3.6. SEIR model

The SEIR model is a more developed and widely used model for epidemic prediction than the SIR model, in which infectious diseases have an incubation period and healthy people who come into contact with patients do not become ill right away, but instead become carriers of the pathogen [42]. In short, an SEIR model is a modified version of a SIR model in which one additional parameter is considered to contribute to the quality of the output of the model as a whole, which can be considered a contributing factor to the quality [43]. The equation for the SEIR model are defined as (5) and (6).

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)I(t)}{N} - \alpha S(t) \quad (5)$$

$$\frac{dE(t)}{dt} = \beta \frac{S(t)I(t)}{N} - \gamma E(t) \quad (6)$$

As part of the SEIR model, the total population will be classified into four classes based on the degree to which they have been infected with Covid-19, such as susceptible, $S(t)$, exposed, $E(t)$, infected-infectious, $I(t)$ and recovered, $R(t)$, where t is the time variable. Λ and μ correspond to births and natural deaths independent of the disease, and α is the fatality rate (Figure 2) [44].

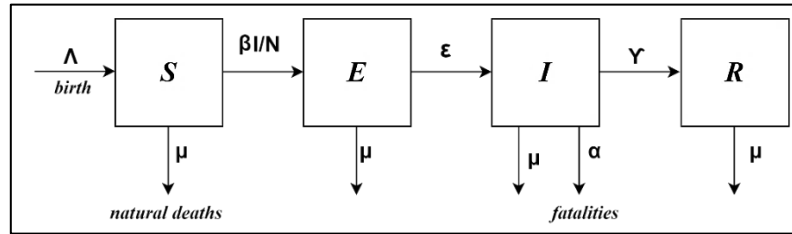


Figure 2. Typical SEIR model

3.7. Facebook prophet

The Facebook Prophet identifies time-series data making use of an additive model, whereby nonlinear trends are fit with respect to yearly, weekly, daily-seasonality, and holiday effects [43]. This algorithm is a free open-source tool known for its ability to work well with time-series data that appears seasonal. The algorithm developed by Facebook's data science team's sole focus is business forecasting [45]. They are combined in (7):

$$y(t) = g(t) + s(t) + h(t) + \varepsilon t \quad (7)$$

where $g(t)$ stands for logistic growth curved that is used in statistics for non-periodic model variations, while $s(t)$ is seasonal changes, $h(t)$ stands for user-provided and t stands for error terms estimated for any substantial changes that are not implemented by the model. By utilizing time as an independent variable, Facebook Prophet attempts to readjust numerous linear and nonlinear functions of time.

3.8. Genetic programming-based

There are many types of algorithms established for GP, but GP algorithms are algorithms based on the principles of nature, and the representation of a program is based on the principles of tree structure, selection, crossover, and mutation [46]. This style of programming can be classified as an evolutionary algorithm because the primary principle is the same as in any of them, but the unique feature here is that people in a population are represented by tree structures [47]. Iteratively, GP, in particular, creates a new generation of computer programs based on analogs of naturally occurring genetic processes as shown in Figure 3. In GP, the program is mostly expressed as syntax trees rather than lines of code. A tree consists of nodes (also called points) and links (Figure 4) [48].

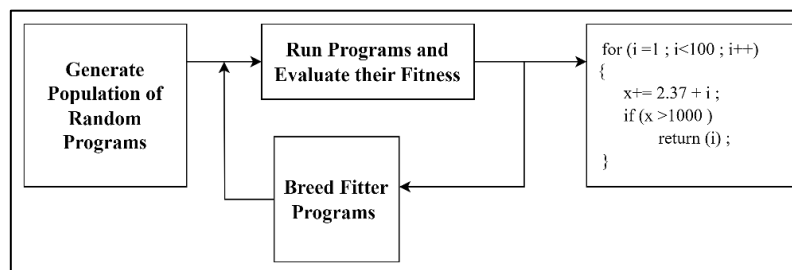


Figure 3. Main loop of GP

3.9. Dataset

The choice of dataset in research is frequently directly related to the geographical location under study. Researchers frequently choose datasets that are specific to the geographical area of interest, tailoring their data selection to reflect the study area's unique nuances and characteristics. Furthermore, many researchers use publicly available datasets to improve the transparency and accessibility of their work.

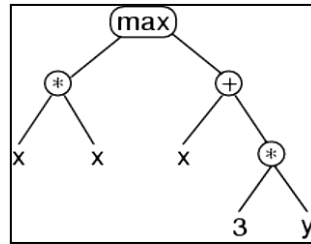


Figure 4. Basic tree-like program presentation used in GP

For example, researchers conducting health-related studies in Malaysia may use datasets obtained directly from the country's health ministry. In this context for Covid-19 researchs, researchers can use these datasets from the Ministry of Health (MOH) website to use their dataset. Alternatively, researchers may use datasets from reliable international sources in order to obtain a more comprehensive global perspective. For example, researchers can use datasets from the WHO, which provides a comprehensive and open-source dataset for Covid-19 as well as other diseases.

3.10. Evaluation metrics

To evaluate model performance and determine which models are the most accurate and dependable, a wide range of evaluation metrics are available. In the iterative process of improving models and increasing their predictive power, these metrics are essential. Of all the assessment metrics available, RMSE, and MAE are the most commonly used benchmarks across a range of research fields. By averaging the absolute values of errors, MAE provides a more straightforward representation than RMSE, which emphasizes the significance of larger errors with a comprehensive measure of the average magnitude of errors.

In general, the RMSE is a useful indicator of forecasting accuracy; however, it should only be applied when comparing the forecasting errors of various models for a particular variable, not when comparing forecasts for different variables. Mean squared error can be expressed as the square root of MSE. While taking the root produces a metric with the same units as y , it has no effect on the relative ranks of the models. This metric conveniently represents the typical or "standard" error for normally distributed errors. Therefore, choosing RMSE as one of the evaluation metrics for forecasting model research is the most appropriate option [49].

MAE signifies the average utter difference between the actual and estimated values in theory. Due to the ease of interpretation of the error value, MAE is also one of the most commonly used evaluation metrics. For both evaluation metrics, if the MAE and RMSE are close to 0, it indicates that the model is more accurate. The lower the value, the better the model is. The model with the lower value of MAE and RMSE will be used as the final model for this study [50], [51].

4. DISCUSSION

As previously described, multiple machine-learning techniques have been found for forecasting and monitoring the Covid-19 pandemic. Regardless of their intrinsic differences, it is critical to recognize that each model has its own set of benefits and drawbacks, as painstakingly detailed in Table 2. This comprehensive examination emphasizes the importance of scrutinizing the complexities of individual algorithms, providing researchers and practitioners with valuable insights to make informed forecasting and monitoring decisions in the dynamic pandemic landscape.

Table 2. Advantages and limitations of the machine learning model

Models	Advantages	Limitations
SVR	The model's complexity does not depend on the data's dimensions	Not suitable for large data sets
LR	Simple to implement and easier to interpret	Sensitive to outliers
ARIMA	Perform well for short-term forecast	Model unstable because of changes in observations and changes in model specifications [52]
SIR	Simple and easy to use [53]	It is deterministic and has constant parameters
Neural network	Great efficiency. Provide flexible nonlinear modeling [54], [55]	Require large datasets
GP	It is capable of optimizing a wide variety of problems, including discrete functions, multi-objective problems, and continuous functions [56]	Several parameters should be set by the decision-makers

5. CONCLUSION

As a result, there is no limit to the variety of machine learning algorithms and approaches utilized for predicting Covid-19 outbreaks. Predictive models help track epidemics and disease transmission. These frameworks and models could help to predict Covid-19 cases; however, the accuracy of predictions may depend on data quality, testing availability, and public health activities.

Although these models may be useful in capturing a wide range of factors that contribute to disease transmission, it is essential to be aware that they may have limitations. Data availability and quality, public health activities, and human behavior are only some of the unknowns and assumptions that might affect the model's accuracy. Thus, it is crucial to include these concepts and models in a more all-encompassing strategy for monitoring and preventing the Covid-19 pandemic. The study of these research findings can provide insight into the different approaches taken by researchers and the outcomes they have achieved. It is also possible to conduct a further review of this progress for the next research by taking into account new machine-learning techniques or improved versions of tried-and-true techniques. We anticipate that future research will establish a framework that can help the government, health authorities, and the general public be ready for future pandemic disease control by utilizing machine learning models.

ACKNOWLEDGEMENTS

This work was supported by the MMU IR Fund (MMUI/220102).

REFERENCES




- [1] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges," *International Journal of Antimicrobial Agents*, vol. 55, no. 3, p. 105924, Mar. 2020, doi: 10.1016/j.ijantimicag.2020.105924.
- [2] Y. Shi *et al.*, "An overview of COVID-19," *Journal of Zhejiang University: Science B*, vol. 21, no. 5, pp. 343–360, May 2020, doi: 10.1631/jzus.B2000083.
- [3] A. Elengoe, "COVID-19 outbreak in Malaysia," *Osong Public Health and Research Perspectives*, vol. 11, no. 3, pp. 93–100, Jun. 2020, doi: 10.24171/j.phrp.2020.11.3.08.
- [4] R. Güner, İ. Hasanoglu, and F. Aktaş, "Covid-19: prevention and control measures in community," *Turkish Journal of Medical Sciences*, vol. 50, no. SI-1, pp. 571–577, 2020, doi: 10.3906/sag-2004-146.
- [5] N. Darapaneni, D. Reddy, A. R. Paduri, P. Acharya, and H. S. Nithin, "Forecasting of COVID-19 in India Using ARIMA Model," *2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2020*, 2020, pp. 0894–0899, doi: 10.1109/UEMCON51285.2020.9298045.
- [6] S. Singh *et al.*, "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models," *Journal of Infection in Developing Countries*, vol. 14, no. 9, pp. 971–976, Sep. 2020, doi: 10.3855/JIDC.13116.
- [7] R. K. Singh *et al.*, "Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model," *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19115, May 2020, doi: 10.2196/19115.
- [8] S. S. Januri, I. Ab Malek, N. Nasir, and Z. A. M. Md Yasin, "Forecasting the spread of daily confirmed Covid-19 cases in Malaysia," *International Journal of Academic Research in Business and Social Sciences*, vol. 12, no. 2, 2022, doi: 10.6007/ijarbss.v12-i2/12114.
- [9] J. A. Ratu, M. A. Masud, M. M. Hossain, and M. Samsuzzaman, "Forecasting the COVID-19 Pandemic in Bangladesh Using ARIMA Model," in *3rd International Conference on Sustainable Technologies for Industry 4.0, STI 2021*, Dec. 2021, pp. 1–6, doi: 10.1109/STI53101.2021.9732576.
- [10] 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, 2022, doi: 10.3390/ijerph19031504.
- [11] L. Moftakhar, M. Seif, and M. S. Safe, "Exponentially increasing trend of infected patients with Covid-19 in Iran: a comparison of neural network and arima forecasting models," *Iranian Journal of Public Health*, vol. 49, pp. 92–100, Jul. 2020, doi: 10.18502/ijph.v49is1.3675.
- [12] A. U. Mandayam, A. C. Rakshith, S. Siddesha, and S. K. Niranjan, "Prediction of Covid-19 pandemic based on regression," in *Proceedings - 2020 5th International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2020*, Nov. 2020, pp. 1–5, doi: 10.1109/ICRCICN50933.2020.9296175.
- [13] R. Salgotra, M. Gandomi, and A. H. Gandomi, "Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming," *Chaos, Solitons and Fractals*, vol. 138, p. 109945, Sep. 2020, doi: 10.1016/j.chaos.2020.109945.
- [14] A. A. Al-Shammari *et al.*, "The impact of strict public health measures on COVID-19 transmission in developing countries: the case of Kuwait," *Frontiers in Public Health*, vol. 9, 2021, doi: 10.3389/fpubh.2021.757419.
- [15] S. S. Helli, Ç. Demirci, O. Çoban, and A. Hamamci, "Short-term forecasting COVID-19 Cases in Turkey using long short-term memory network," in *TIPTEKNO 2020 - Tip Teknolojileri Kongresi - 2020 Medical Technologies Congress, TIPTEKNO 2020*, Nov. 2020, pp. 1–4, doi: 10.1109/TIPTEKNO50054.2020.9299235.
- [16] X. Yuan, J. Xu, S. Hussain, H. Wang, N. Gao, and L. Zhang, "Trends and prediction in daily new cases and deaths of COVID-19 in the United States: an internet search-interest based model," *Exploratory Research and Hypothesis in Medicine*, pp. 1–6, Apr. 2020, doi: 10.14218/erhm.2020.00023.
- [17] W. Zou, "The COVID-19 pandemic prediction in the US based on machine learning," in *Proceedings - 2020 International Conference on Public Health and Data Science, ICPHDS 2020*, Nov. 2020, pp. 283–289, doi: 10.1109/ICPHDS51617.2020.00062.
- [18] H. I. Mustafa and N. Y. Fareed, "COVID-19 cases in Iraq; forecasting incidents using Box-Jenkins ARIMA model," in *Proceedings - 2nd Al-Noor International Conference for Science and Technology, NICST*, Aug. 2020, pp. 22–26, doi: 10.1109/NICST50904.2020.9280304.

- [19] P. Podder and M. R. H. Mondal, "Machine learning to predict COVID-19 and ICU requirement," in *Proceedings of 2020 11th International Conference on Electrical and Computer Engineering, ICECE*, Dec. 2020, pp. 483–486, doi: 10.1109/ICECE51571.2020.9393123.
- [20] S. Saadah and M. A. Permana, "Fuzzy time series using Chen and Lee model to predict COVID-19 in Aceh Indonesia," in *Proceedings-2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021*, Sep. 2021, pp. 79–84, doi: 10.1109/IC2IE53219.2021.9649283.
- [21] S. K. Safi and O. I. Sanusi, "A hybrid of artificial neural network, exponential smoothing, and ARIMA models for COVID-19 time series forecasting," *Model Assisted Statistics and Applications*, vol. 16, no. 1, pp. 25–35, Mar. 2021, doi: 10.3233/MAS-210512.
- [22] A. S. Ahmar and E. B. del Val, "SutteARIMA: short-term forecasting method, a case: Covid-19 and stock market in Spain," *Science of the Total Environment*, vol. 729, p. 138883, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138883.
- [23] V. Poleneni, J. K. Rao, and S. A. Hidayathulla, "COVID-19 prediction using ARIMA model," in *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, pp. 860–865, Jan. 2021, doi: 10.1109/Confluence51648.2021.9377038.
- [24] A. Kumar and K. Kaur, "A hybrid SOM-fuzzy time series (SOMFSTS) technique for future forecasting of COVID-19 cases and MCDM based evaluation of COVID-19 forecasting models," in *Proceedings - IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems, ICCIS 2021*, Feb. 2021, pp. 612–617, doi: 10.1109/ICCIS51004.2021.9397216.
- [25] S. Prasanth, U. Singh, A. Kumar, V. A. Tikkiwal, and P. H. J. Chong, "Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach," *Chaos, Solitons and Fractals*, vol. 142, Jan. 2021, doi: 10.1016/j.chaos.2020.110336.
- [26] R. R. Maaliw, M. A. Ballera, Z. P. Mabunga, A. T. Mahusay, D. A. Dejelo, and M. P. Seno, "An ensemble machine learning approach for time series forecasting of COVID-19 cases," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2021*, pp. 633–640, Oct. 2021, doi: 10.1109/IEMCON53756.2021.9623074.
- [27] N. Zheng *et al.*, "Predicting COVID-19 in China using hybrid AI model," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 2891–2904, Jul. 2020, doi: 10.1109/TCYB.2020.2990162.
- [28] A. Kurniawan and F. Kurniawan, "Time series forecasting for the spread of Covid-19 in Indonesia using curve fitting," in *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIconCIT 2021*, pp. 45–48, Apr. 2021, doi: 10.1109/EIconCIT50028.2021.9431936.
- [29] A. Al Mamun, M. Sohel, N. Mohammad, M. S. Haque Sunny, D. R. Dipta, and E. Hossain, "a comprehensive review of the load forecasting techniques using single and hybrid predictive models," *IEEE Access*, vol. 8, pp. 134911–134939, 2020, doi: 10.1109/ACCESS.2020.3010702.
- [30] D. Berberich, "Hybrid methods for time series forecasting," Inovex, 2020, Online [Available]: <https://www.inovex.de/de/blog/hybrid-time-series-forecasting/>. (Accessed: Dec. 30, 2022).
- [31] A. I. Elwasify, "A combined model between artificial neural networks and ARIMA models," *International Journal of Recent Research in Commerce Economics and Management (IJRRCM)*, vol. 2, no. 2, pp. 134–140, 2015.
- [32] L. B. Sina, C. A. Secco, M. Blazevec, and K. Nazemi, "Hybrid forecasting methods—a systematic review," *Electronics (Switzerland)*, vol. 12, no. 9, p. 2019, Apr. 2023, doi: 10.3390/electronics12092019.
- [33] E. Stellwagen and L. Tashman, "ARIMA: The models of box and Jenkins," *Foresight: The International Journal of Applied Forecasting*, no. 30, pp. 28–34, 2013.
- [34] M. M. A. Alfaki and S. B. Masih, "Modeling and forecasting by using time series ARIMA models," *International Journal of Engineering Research and*, vol. V4, no. 03, 2015, doi: 10.17577/ijertv4is030817.
- [35] S. L. Ho and M. Xie, "The use of ARIMA models for reliability forecasting and analysis," *Computers and Industrial Engineering*, vol. 35, no. 1–2, pp. 213–216, 1998, doi: 10.1016/s0360-8352(98)00066-7.
- [36] J. Liu, F. Yu, and H. Song, "Application of SARIMA model in forecasting and analyzing inpatient cases of acute mountain sickness," *BMC Public Health*, vol. 23, no. 1, p. 56, Jan. 2023, doi: 10.1186/s12889-023-14994-4.
- [37] A. F. Mohamad, A. M. Jasin, A. Asmat, R. Canda, J. Ismail, and A. B. M. Soom, "Sales Analytics Dashboard with ARIMA and SARIMA Time Series Model," in *13th IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE*, May 2023, pp. 106–112, doi: 10.1109/ISCAIE57739.2023.10165270.
- [38] S. G. Fard, H. M. Rahimi, P. Motie, M. A. S. Minabi, M. Taheri, and S. Nateghinia, "Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review," *Heliyon*, vol. 7, no. 10, p. e08143, Oct. 2021, doi: 10.1016/j.heliyon.2021.e08143.
- [39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [40] S. Taylor, "Regression analysis," *Corporate Finance Institute*, Accessed: Dec. 24, 2022. Online [Available]: <https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/>.
- [41] P. G. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.
- [42] Y. Ma, Z. Xu, Z. Wu, and Y. Bai, "COVID-19 spreading prediction with enhanced SEIR model," in *Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE*, pp. 383–386, Oct. 2020, doi: 10.1109/ICAICE51518.2020.00080.
- [43] J. M. Carcione, J. E. Santos, C. Bagaini, and J. Ba, "A simulation of a COVID-19 epidemic based on a deterministic SEIR model," *Frontiers in Public Health*, vol. 8, May 2020, doi: 10.3389/fpubh.2020.00230.
- [44] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of COVID-19 in China by dynamical modeling," *International Conference on Artificial Intelligence and Computer Engineering*, 2020.
- [45] S. Mohan, A. Rajendran, K. Ajith, K. Varma, and A. Ashok, "A Covid-19 study and forecasting," in *Proceedings of the 2022 3rd International Conference on Intelligent Computing, Instrumentation and Control Technologies: Computational Intelligence for Smart Systems, ICICICT 2022*, pp. 315–321, Aug. 2022, doi: 10.1109/ICICICT54557.2022.9917761.
- [46] M. T. Ahvanooy, Q. Li, M. Wu, and S. Wang, "A survey of genetic programming and its applications," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, Apr. 2019, doi: 10.3837/tiis.2019.04.002.
- [47] W. Banzhaf, "Genetic programming: first European workshop, EuroGP'98: Paris, France," *Lecture notes in computer science, Berlin ; New York: Springer*, 1998.
- [48] V. Bojtar and J. Botzheim, "Queen bee based genetic programming method for a hive like behavior," *20th IEEE International Symposium on Computational Intelligence and Informatics, CINTI 2020 - Proceedings*, pp. 127–132, 2020, doi: 10.1109/CINTI51262.2020.9305824.




- [49] J. R. Koza and R. Poli, "A genetic programming tutorial," *Introductory Tutorials in Optimization Search and Decision Support*, pp. 55-80, 2003.
- [50] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481-5487, 2022, doi: 10.5194/gmd-15-5481-2022.
- [51] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.*, vol. 30, pp. 79-82, 2005, doi: 10.3354/cr030079.
- [52] 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, 2023, doi: 10.1007/s00521-020-05626-8.
- [53] P. Harjule, V. Tiwari, and A. Kumar, "Mathematical models to predict COVID-19 outbreak: an interim review," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 2, pp. 259-284, Feb. 2021, doi: 10.1080/09720502.2020.1848316.
- [54] "Nonlinear model predictive control using neural networks," *IEEE Control Systems*, vol. 20, no. 3, pp. 53-62, Jun. 2000, doi: 10.1109/37.845038.
- [55] A. Muthée, "The Basics of genetic algorithms in Machine Learning," *Engineering Education Community Website*, 2021. Accessed on July 15, 2023. Online. [Available]: <https://www.section.io/engineering-education/the-basics-of-genetic-algorithms-in-ml>.
- [56] R. Katarya *et al.*, "A review of various mathematical and deep learning based forecasting methods for COVID-19 pandemic," in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Mar. 2021, pp. 874-878, doi: 10.1109/ICACCS51430.2021.9441966.

BIOGRAPHIES OF AUTHORS






Md Shohel Sayeed    has been a member of Multimedia University since 2001 and now he serves as a Professor of the Faculty of Information Science and Technology. His core research interest is in the area of biometrics, information security, image and signal processing, pattern recognition, and classification. Till date, he has published over 90 research papers in international peer-reviewed journals and international conference proceedings as a result of his research work. His research works have been accepted by journals such as IEEE transactions on pattern analysis and machine intelligence (TPAMI), International Journal of Pattern Recognition and Artificial Intelligence (IJPRI), discrete dynamics in nature and society (DDNS). These papers have been cited in various international journals and conferences a number of times. Several of his findings have been presented in a number of well recognized IEEE conferences such as ICSP2006, ICIAS2007, ITSIM2008, CSECS2009, and ITSIM2010. He has been appointed technical paper reviewer for Journal of Pattern Recognition Letters, IEEE Transaction on Neural Networks, IEEE Transactions on Automation Science and Engineering, Journal of Computer Methods and Programs in Biomedicine and International Journal of Computer Theory and Engineering. He has also been invited to review technical papers for several international conferences. In recognition of his professional contribution, he has obtained recognition as a senior member of IEEE Computer Society. He can be contacted at email: shohel.sayeed@mmu.edu.my.



Siti Najihah Hishamuddin    is a postgraduate student at Multimedia University Melaka. She received a B.Sc. degree in computer science from the National University of Malaysia (UKM). She worked in the finance industry as an IT Back-end developer before she decided to pursue a master's in Multimedia University. She majored in data science during her degree and decided to do research under her supervisor, Prof Shohel about machine learning. She can be contact at email: 1221400121@student.mmu.edu.my.



Ong Thian Song    works in Faculty of Information Sciences and Technology (FIST), Multimedia University. His research interests include Machine learning and biometrics security. He holds M.Sc. degree in 2001 from University of Sunderland, UK and Ph.D. degree in 2008 from Multimedia University, Malaysia. He has published more than 70 international refereed journals and conference articles. He is a senior member of IEEE and has also served as the editorial board for IEEE Biometric Council Newsletter from year 2013 to 2015. He can be contact at his email: tsong@mmu.edu.my.