Prediction of COVID-19 Pandemic Trend by Machine Learning

Shijia Xu *

Department of Informatics, King's College London, England, WC2R 2LS, United Kingdom

* Corresponding author email: shijia.xu@kcl.ac.uk

Abstract. Since 2020, COVID-19 has had a huge impact on people's lives. Including but not limited to economic, educational, medical, and other aspects. During this period, all sectors of society and the government have intervened reasonably, so it is necessary to analyze the data on COVID-19 so far and make scientific predictions. This article starts with the analysis of raw data on COVID-19 from the World Health Organization (WHO). Then four machine learning methods, including the time series model, exponential smoothing model, XGBRegressor method, and polynomial regression model, are leveraged for trend prediction of the original data. The data, with the time ranging from January 2020 to May 2021, is taken as the training object, and then the epidemic in Jul 2021 is used for testing. The number of cases is predicted and compared with real data. It is concluded that the WHO has indeed carried out effective intervention in areas seriously affected by the epidemic and that the time series model predicts the minimum loss value.

Keywords: COVID-19; Trend Prediction; Machine Learning; WHO.

1. Introduction

COVID-19 pandemic is called a coronary toxic flow four-line disease, which is a severe acute respiratory system synthesis coronary disease (SARS-CoV-2) triggered global coronary disease toxic disease. Epidemics emerged in December 2019 in Wuhan, China. The World Health Organization declared the epidemic a public health emergency of international concern in January 2020 and became a pandemic on March 11. As of July 23, 2020, 188 more national and district reports, over 15.8 million cases of COVID-19 disease cases, 623,000 deaths, super 8.6 million cured. In January 2020, coronary disease was poisoned in the whole-ball range, and it came over. How many times have you been analyzing the epidemic situation progressing, and now the epidemic prevention is important, and because of this project, each country has been able to analyze the epidemic situation, and the future epidemic has been seen, and it is under the same training model, Selected Most Favorite Models.

This year's COVID-19 epidemic has been basically controlled, and my prediction of epidemic data can be applied to many fields. Analyzing and predicting the epidemic through machine learning is one of the most scientific and effective ways to master the COVID-19 epidemic.

With the development of biomedicine and the progress of cutting-edge technologies, these technologies can usually be used to generate more complex and high-dimensional data. In order to maximize the use of these data and unleash its full potential, machine learning is one of the indispensable factors for the advancement of clinical medicine. ML is a subordinate field of generalized artificial intelligence (AI), which refers to an algorithm that can learn directly from data and build models. This is an emerging field that covers multiple disciplines and is still developing and progressing. Computer science, statistics, and mathematics have made major contributions and are currently at the forefront of the development of life sciences. The use of ML in the clinical environment will be called "translation machine learning". It can be applied to use ML as an important decision support system in clinical medicine. The algorithm provides additional auxiliary information and can be used to help clinicians better treat patients [1, 2, 3].

The development and problem-solving ability of machine learning and artificial intelligence are well reflected in this paper. The author explains and explains from many different aspects (history, theory, type, and how to solve practical problems). Then, the evaluation matrix operation of the performance of the learning model is evaluated through specific algorithms [4]. Matteo et al. collected patient data from some Italian hospitals and then studied and analyzed the data through the CatBoost gradient boosting model. This is an advantageous way to apply machine learning to infectious disease

prediction. This paper uses other algorithms that are different from it to analyze and predict the spread of Covid-19 [5].

Machine learning is not only a core field of artificial intelligence but has become one of the most active and obvious application potentials in the computer field. It plays an increasingly important role and has been applied in recent years to predict material performance, solar cell effects, automation, and many cutting-edge technology fields. When it comes to machine learning, people will think of the advantages of accuracy, efficiency, and wide applicability [6]. In terms of experimental costs, computing power is very important when training models with a large amount of data. Therefore, high-performance computing can be used to reduce computing costs [7, 8].

In this paper, the time series model is used to analyze and predict. The research is divided into four steps: data acquisition, data analysis, model building, and model comparison.

2. Data Acquisition and Visualization

2.1 Dataset Acquisition and Pre-Processing

The data set is organized by data for COVID-19 published on Kaggle [9] collected from WHO, some examples are demonstrated in Table 1.

Table 1. Examples of the dataset.

Province /State	Country /Region	Lat	Long	Date	Con-firm	Death	Recover	Active	WHO Region
Nan	Afghanistan	33.93	67.70	2020/1/22	0	0	0	0	Eastern Mediterranean
Nan	Albania	41.15	20.16	2020/1/22	0	0	0	0	Europe
Nan	Algeria	28.03	1.65	2020/1/22	0	0	0	0	Africa
Nan	Andorra	42.50	1.52	2020/1/22	0	0	0	0	Europe
Nan	Angola	-11.20	17.87	2020/1/22	0	0	0	0	Africa
Nan	Antigua and Barbuda	17.06	-61.79	2020/1/22	0	0	0	0	Americas
Nan	Argentina	-38.41	-63.61	2020/1/22	0	0	0	0	Americas
Nan	Armenia	40.06	45.03	2020/1/22	0	0	0	0	Europe

Through observation of the Table 1, it can be found that some data are nan, which means there are missing data and require further data cleaning before learning. So, these data need to be pre-processed. The strategy used for data cleaning is to delete these data with the nan value. The cleaned data set is demonstrated in Table 2. They will be grouped according to the national city, and the daily number of cures confirmed cases, deaths, and the total number of confirmed cases in the same national cities will be grouped.

Table 2. Examples of cleaned dataset.

	Country/Region	Confirmed	Deaths	Recovered	Active
0	Afghanistan	36263	1269	25198	9796
1	Albania	4880	144	2745	1991
2	Algeria	27973	1163	18837	7973
3	Andorra	907	52	803	52
4	Angola	950	41	242	667

2.2 Data Visualization

To better understand the dataset, the data is visualized as shown in Figure 1. In this figure, the top 20 countries with the highest confirmed cases are demonstrated. For example, the US is with the highest confirmed cases at 4290259. The Brazil is the second one with 2442375 cases and the India is the third one with 1480073 cases [10].

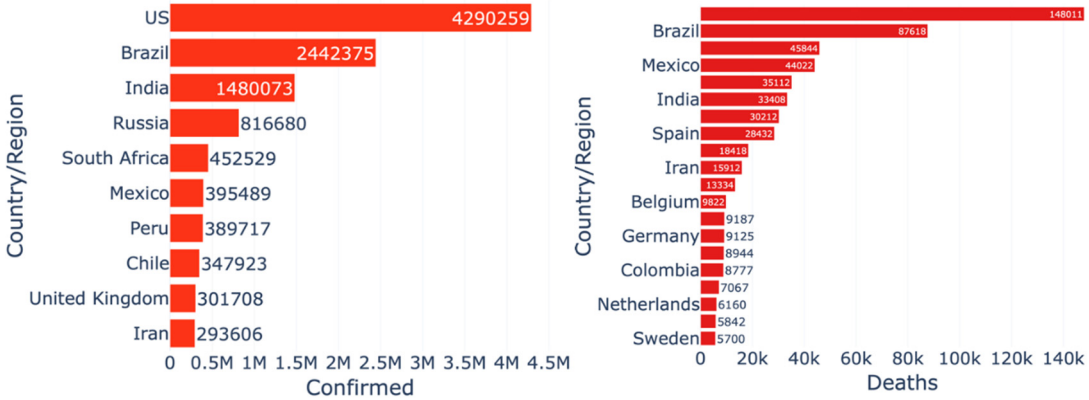

Fig 1. 20 countries with the highest confirmed death cases in the world.

The data shown in Figure 1 merely displays the COVID-19 cases at a specific time point, where the tendency of the increasing of COVID-19 could not be reflected. To observe the development trend of the world COVID-19 epidemic, a curve plot is drawn as shown is Figure 2. In this figure, the development trend of the world epidemic is calculated as drawn as the line charts. From the tendency, readers could observe that COVID-19 cases are getting more and more, without the occurrence of the turning point.

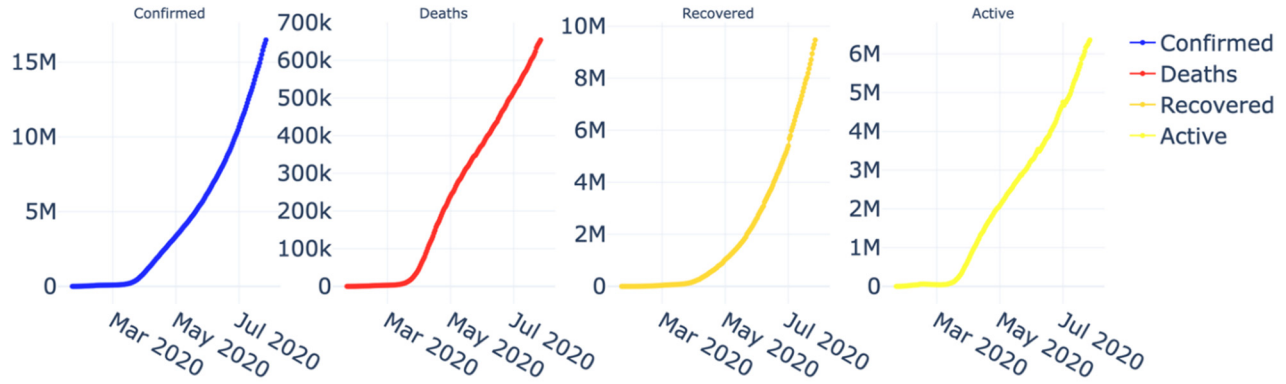

Fig 2. The development trend of COVID-19 worldwide

The growth trend of the COVID-19 epidemic is still obvious worldwide. Although the number of cures is also rising, the number of confirmed cases and deaths is also rising at a higher slope. In short, the number of confirmed cases and deaths is still greater than the number of cures, so the current situation is not optimistic. People should be more vigilant to prevent the epidemic from being swept over again.

After the basic data visualization analysis, it is of great importance of understanding the possibilities of analyzing the factors affecting the COVID-19 epidemic. Among these factors, one of the most important one is the whether the interference of the WHO works. To meet this goal, the data of countries interference by the WHO is collected and visualized as demonstrated in Figure 3.

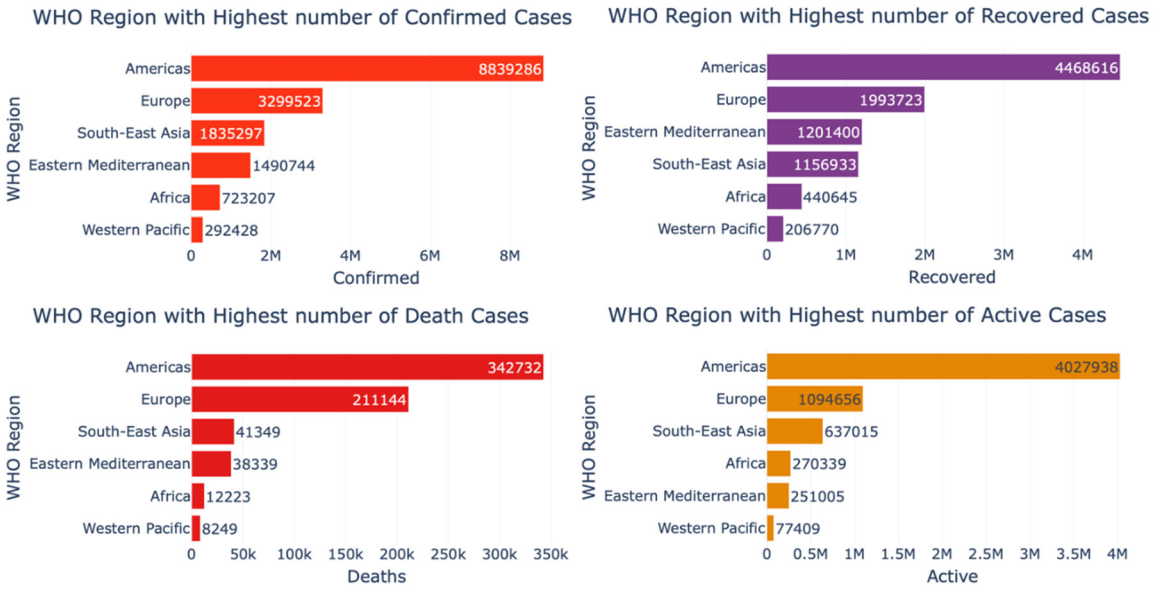

Fig 3. Epidemic data in various regions with WHO intervention

In summary, the population can be analyzed. The WHO has a certain connection with the spread of the COVID-19 epidemic. In addition, many other factors will affect the outbreak of the COVID-19 epidemic, such as climatic factors such as the humidity and temperature of the air, or other human factors such as national preventive measures.

3. Method

3.1 Model Selection

In this section, four different regressive models are introduced for the prediction of the COVID-19 pandemic trend. The input of the models is the previous population of patients suffering from COVID-19 and the output is the prediction of future data. Four different models, including Prophet model, exponential smoothing model, XGBRegressor, and Polynomial regression model, are leveraged for the prediction.

3.1.1 Prophet Model

The Prophet of Time Series Model program is an additive regression model consisting of four main components A piecewise linear or logistic growth curve trend. prophet automatically detects changes in the trend by selecting change points from the data. This method applies a statistical approach to predict future developments based on past trends, which in most cases usually fits the pattern of how things develop. While considering the trends, it is also necessary to pay attention to the impact of seasonal and cyclical changes on specific points in time to achieve greater predictive accuracy. But in the end, it is also necessary to acknowledge the impact that random variables may have on the results.

3.1.2 Exponential Smoothing

The exponential smoothing method is divided into simple exponential, quadratic exponential, and triple exponential smoothing models. In this example, we use the quadratic exponential model and the cubic exponential smoothing model. The exponential smoothing model is a time series-based prediction method specifically for single variable data, which can be expanded to support fragmented data with systematic trend overall data or seasonal components. It is a powerful and widely applicable prediction method that can be used as an alternative to the popular Box-Jenkins ARIMA series of methods. This method is relatively flexible, universal, and has a little subjective randomness in the

selection of smooth indicators. The advantage is that if less data is required, the required results can be predicted.

3.1.3 XGBRegressor Method

XGBoost stands for "extreme gradient boosting", which is one of the implementations of the gradient boosting tree algorithm. XGBoost is commonly used for supervised machine learning models, which are characterized by fast computation, parallelization, and high performance. XGBoost is an implementation of gradient-boosted decision trees. Use the model to predict the results first, then use the error estimation model to carry out the error estimation model, repeat this process, and integrate the new error estimation model into the model. The accuracy of the initial estimate is not high, because it can be made up in subsequent error estimates, and the input must be data of the Data Frame type.

3.1.4 The Polynomial Regression Model.

The polynomial regression model is one of the forms of linear regression models and is also known as a special case of multiple linear regression, which estimates the relationship as an nth-order polynomial. Polynomial regression is sensitive to outliers, so the presence of one or two outliers can seriously affect its performance. The use of polynomial regression methods should be used when the need is to build a model suitable for dealing with nonlinearly separable data. In this regression technique, the best-fit line is not straight, but a curve that fits the data points. For a polynomial regression, the index of some independent variables is greater than 1. Some variables have exponentials, while others do not. However, choosing the exact index of each variable naturally requires some prior knowledge of the current data set and the final output.

3.2 Evaluation Matrix

The root means square error (RMSE) is leveraged to evaluate the loss of the model, train as a data set before October 2020, and evaluate the training results of the model as a test set from October 1 to October 20. The formula to find the root mean square error, often abbreviated RMSE, is as follows:

$$RMSE = \sqrt{\sum (P_i - O_i)^2 / n} \quad (1)$$

Where the Σ is a fancy symbol that means “sum”. Pi is the predicted value for the i-th observation in the dataset. Oi is the observed value for the i-th observation in the dataset. n is the sample size

4. Result

In this section, qualitative demonstrations of the prediction performances are sequentially displayed from section 4.1 to section 4.4. Afterward, the quantitative performances are compared among the four methods. In these experiments, the COVID-19 data achieved from WHO is taken as that material. Epidemic data ranging from January 2020 to May 2021is taken as the training samples, and the epidemic in Jul 2021 is used for testing.

4.1 Result of Model Prophet

Figure 4 demonstrate the performances of the Prophet model. The line in purple color is the COVID-19 training data. The line in red color is the actual COVID-19 cases and the line in green denotes the predicted results. It could be observed that the prediction is quite accurate which has similar tendency with the ground truth. Moreover, readers could notice that the number of predicted cases is larger than the actual confirmed cases, indicating that there are some factors helpful for decreasing the confirmed cases during this period.

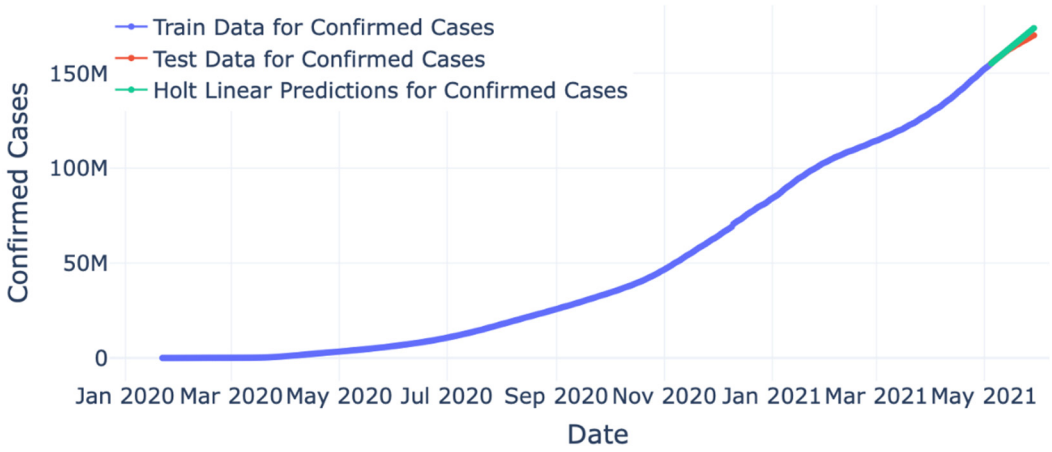

Fig 4. Prediction result of Prophet model.

4.2 Exponential Smoothing Model

The plot in Figure 5 shows the performances achieved by the exponential smoothing model. The meaning of color of different lines is the same with that in Figure 4. Figure 5 demonstrates that the exponential smoothing model performs worse than the Prophet model. Similarly, the predicted number of cases is larger than the ground truth.

Fig 5. Prediction result of exponential smoothing model.

4.3 XGBRegressor Model

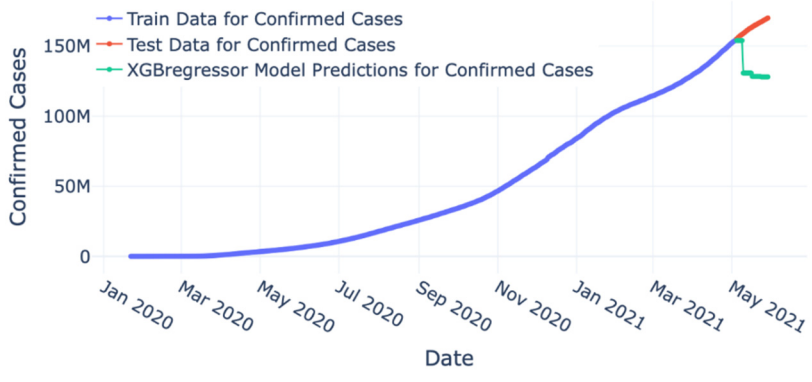

Fig 6. Prediction result of XGBRegressor model.

XGBoost is one of the Boosting algorithms. The idea of Boosting algorithm is to integrate many weak classifiers to form a strong classifier. Because XGBoost is an ascending tree model, it integrates many tree models to form a strong classifier. The regression tree model used is the CART regression tree model. The performance is demonstrated in Figure 6. It shows that the prediction is not that practical. There is an obvious drop in the number of cases.

4.4 Polynomial Regression

Polynomial regression model is one of the most widely used model, which could be attributed to is simplicity and effectiveness. The quantitative performances are shown in Figure 7.

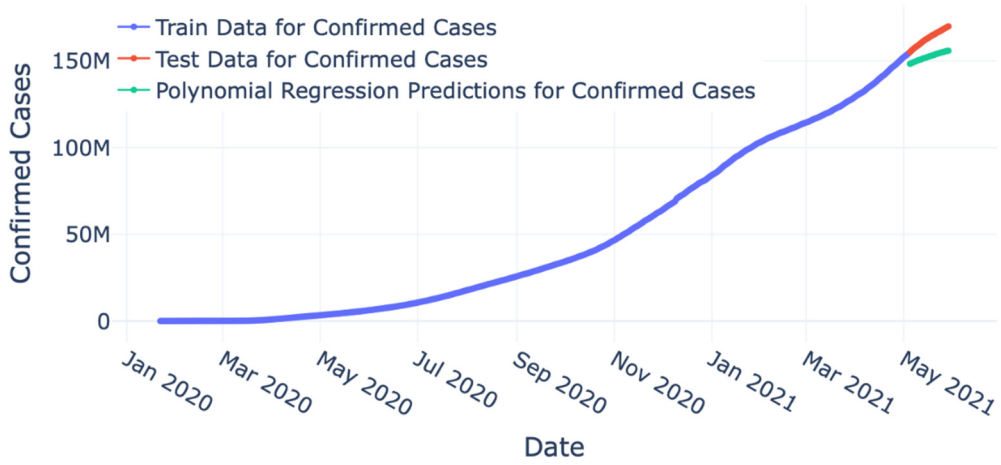

Fig 7. Prediction result of polynomial regression model.

4.5 Comparison of Models

In previous experiments, the performances of different models are shown. However, it is difficult to subjectively evaluate the effectiveness of each model. In Table 3, the qualitative results, measured by RMSE is summarized and compared. The best result is achieved by the Prophet model, followed by the exponential smoothing model, polynomial regression, and XGBRegressor model. Moreover, it could be observed that traditional machine learning algorithms perform poorly in this time series-based project, which also tells us that the specific problem is analyzed, and the selection of models should be based on the needs of the project [11].

Table 3. RMSE results of different models

Models	RMSE
Prophet	1.028587e+06
Exponential Smoothing	2.692388e+06
Polynomial Regression	10.94552e+06
XGBRegressor	32.27250e+06

5. Discussion

It can be observed that WHO's interventions in the epidemic have achieved considerable results in all regions. As can be seen from Figure 3, WHO's interventions in developed countries and regions such as the Americas and Europe are the most effective. However, for developing countries such as Asia, Africa, and third world countries and regions, the epidemic has not achieved obvious results. However, in addition to WHO interventions, there are many factors that affect the spread of the epidemic, including but not limited to people's living habits, epidemic prevention measures in local communities, and relevant policies of various governments. These factors will affect the results of this data analysis. As for the experiments of the machine learning part, four different methods are

evaluated and compared. It could be concluded that people should choose different regressive models according to different tasks and it is extremely difficult to develop a uniform model which is suitable for all various regression tasks. However, there is still some limitations. For example, the data of the training model in the experiment are limited and cannot accurately predict the development trend of the epidemic. There could be some other factors affecting the number of confirmed cases but could not be reflected merely using previous confirmed cases.

6. Conclusion

This article starts with the capture, collation, and cleaning of raw data, and uses four different regressive models to analyze it. Then use the WHO intervention as a variable to compare the four-dimensional data of the epidemic to obtain the effectiveness and limitations of WHO intervention. To reduce the spread and spread of COVID-19 and help governments and organizations make reasonable and correct decisions, the ability to predict future cases is essential. The author chose the prophet algorithm because it has high accuracy. For the author, the accuracy of the model can be proved by this experiment. In this work, the epidemic data ranges from January 2020 to May 2021used for training, and the epidemic in Jul 2021 is used for testing. The results show that the Prophet and exponential smoothing model achieve satisfactory results, while the polynomial regression model and the XGBRegressor model perform poorly. It could be concluded that people should select a different regressive model according to different tasks. Moreover, the effective models predict that the number of confirmed cases is larger than the real one, indicating that WHO interference is helpful in decreasing the growth rate of confirmed cases and further demonstrates the WHO intervention is effective. In the future, to control the epidemic trend of COVID-19, the WHO should play more important role in controlling the growth of the confirmed cases.

References

[1] Couckuyt A, Seurinck R, Emmaneel A, et al. Challenges in translational machine learning[J]. Human Genetics, 2022: 1-16.

[2] Wuest T, Weimer D, Irgens C, et al. Machine learning in manufacturing: advantages, challenges, and applications. Production & Manufacturing Research, 2016, 4(1): 23-45.

[3] L'heureux A, Grolinger K, Elyamany H F, et al. Machine learning with big data: Challenges and approaches. Ieee Access, 2017, 5: 7776-7797.

[4] Hayati N, Fauziah P, Wandi D. Trend of the spread of COVID-19 in Indonesia using the machine learning prophet algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 2021: 1780-1788.

[5] Chieregato M, Frangiamore F, Morassi M, et al. A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. Scientific reports, 2022, 12(1): 1-15.

[6] Khanday A M U D, Rabani S T, Khan Q R, et al. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights, 2022, 2(2): 100120.

[7] Niroshan L, Carswell J D. Machine Learning with Kay. AGILE GIScience Series, 2022, 3: 1-11.

[8] Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial intelligence in medicine, 2020, 104: 101822.

[9] Devakumar K.P., COVID-19 Dataset, 2020, URL: https://www.kaggle.com/datasets/imdevskp/corona-virus-report.

[10] Mishra S, Bordin C, Taharaguchi K, et al. Predictive analytics beyond time series: Predicting series of events extracted from time series data. Wind Energy, 2022.

[11] Svetunkov I. Complex exponential smoothing. Lancaster University (United Kingdom), 2016.