

COMPARISON OF PERFORMANCE FOR FORECASTING THE NEW CASES OF COVID-19 IN MALAYSIA BY USING ARIMA, LSTM AND PROPHET

MUHAMMAD SYAFIQ ALZA BIN ALIAS¹, NORAZLIN BINTI IBRAHIM² and ZALHAN BIN MOHD ZIN³

^{1,2&3}Industrial Automation Section, UniKL Malaysia France Institute, Bangi, Malaysia
syafiq.alias@s.unikl.edu.my | norazlin@unikl.edu.my | zalhan@unikl.edu.my

Abstract

Currently, Malaysia are facing the third wave of COVID-19. The number of new cases is very alarming because this disease can bring death to affected people. Government can take early precautions to prevent the spiking of COVID-19 new cases if this disease can be detected early. Therefore, forecasting the new cases in the future can help to alert the government on the rising of COVID-19 new cases. Besides that, the proven data analysis of COVID-19 can also be used to further convince the people about the potential threat that might occur in near future. Hence, the people will know that the actions taken is not solely based on assumption, which in current situation everyone is vulnerable and pressured mainly due to the economic status. This research is conducted to compare and discover the most suitable time series forecasting model that can be used to predict the new cases of COVID-19 in Malaysia. The models used are ARIMA, LSTM and Prophet. The result shows that ARIMA model with (2,1,1) setup produced the lowest MSE and RMSE errors which indicates that this model has the highest performance compared to other models.

Keywords: COVID-19, Malaysia, ARIMA, LSTM, PROPHET

Introduction

Coronavirus Disease 2019 (COVID-19) is a contagious disease that has been widely spread throughout the world which was first identified around December 2019 in Wuhan, China. This disease was declared as a pandemic by the World Health Organization (WHO) on March 11, 2020 due to the number of cases increased is more than 118,000 that affected over 110 countries and 4,291 death[1][2]. Malaysia has no exception of being affected by the disease and currently are surviving through the third wave with 50390 of total confirmed positive cases as of November 18, 2020 [3].

According to [4], the primary transmission of the disease is by person to person through infectious droplets that occurred during coughing or sneezing, personal contact such as shaking hands, or by touching contaminated surfaces. Since the development of drugs for COVID-19 is still under research to be proven effective [5], the current prevention that is recognized to be effective is social distancing. In [6], research shows that in 10 countries that implement the social distancing, the numbers of daily confirmed cases and daily deaths displayed indications of decreasing in most of the countries after 1 to 4 weeks. Even in Malaysia, after the government has enforced the Movement Control Order (MCO) in mid-March 2020, it has returned positive results as daily new infections did not spike, while the number of cured patients continued to grow[7].

However, imposing MCO bring hefty drawbacks to the people. This is because their daily activity to generate income is restricted. Some of the people did not have enough saving to survive through the period. Although the government has given numerous of initiative to help the unfortunate people, they are still facing economic issues. Therefore, forecasting the new cases of COVID-19 in Malaysia is important to get early detection and alert on the potential increment in number of infected cases. After that, the government can implement suitable actions and precautions to prevent the spread of the disease.

In artificial intelligent field, there are several of time series prediction algorithm that are proven to work well in forecasting the future data. According to [8], forecasting is a process of predicting something future by doing calculations from previous data. It can be concluded that the historical data is very valuable and important to predict the future. However, the current data for COVID-19 has still yet to reach 1 year of data. Therefore, the methods used for forecasting are very important to ensure that the results are reliable and accurate. This is because low accuracy of prediction might cause wrong conclusion and indirectly offers bad impact due to the actions taken as precaution later such as MCO is very costly to the people. This research will focus on finding suitable time series forecasting model to predict the number of new cases for COVID 19 in Malaysia. Selected models such as Autoregressive Moving Average (ARIMA), Long Short-Term Memory (LSTM) and Prophet will be used, and the performances of each model are compared to find out the appropriate model to be used in forecasting COVID-19 cases.

Related Works

A. ARIMA

One of the most popular prediction methods in time series prediction is called ARIMA model which was pioneered by Box and Jenkins[9]. Based on[10], this model has 92.1% accuracy which significantly outperforms the simple moving average method. This method has been implemented in many areas such as agriculture[11], price prediction

[12], etc. Although, the Seasonal ARIMA (SARIMA) can outperform the ARIMA model, there are studies conducted that shows ARIMA model can achieve higher accuracy than SARIMA and Early Aberration Reporting System (EARS) [13]. Besides that, due to the limitation of data, the ARIMA model will fit nicely since it can work well within short time frame.

ARIMA model can be divided into 3 categories which are the AR, Autoregression part which is used to do the prediction based on past values or lags, I is the Integrated part which is used to alter the data point to make the time series stationary and MA is the Moving Average which is almost similar as AR but depend on the error terms[14]. Generally, the notation for ARIMA model can be written as ARIMA (p, d, q). The parameters of p, d and q can be described as follows:

- 'p' is the number of previous value or lag to make the prediction.
- 'd' is the number of times to make the data stationary by differencing the data point.
- 'q' is the number of past error terms included in the model.

B. LSTM

Another common approach in forecasting time series is by using LSTM model. It is a branch of Recurrent Neural Network (RNN) [15]. The dropout element in LSTM can solve the problem of gradient disappearance and excessive gradient. As one of the deep learning category of algorithm, LSTM has performed slightly better compared to ARIMA in research made by [16]. However, the computation cost is very high. Based on the research, the training time for LSTM is almost 47 times longer compared to ARIMA model. LSTM is able to memorize time series in their cell unit. This cell unit holds three logic gates based on sigmoid neural network layer that are called input gate, output gate and forgetting gate in which data can chose either to passed or processed [17]. The Illustration of an LSTM memory cell is displayed in Figure 1.

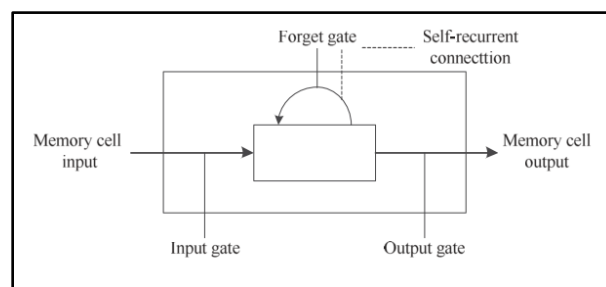


Figure 1: LSTM memory cell [18].

C. Prophet

Prophet is an open-source time series prediction model developed recently by Facebook's Core Data Science[19]. It is available in Python and R which consists of its own special data frame to predict the time series easily. The two main columns in the data frame are called "ds" which stores the date time series and "y" which is used to store the corresponding value of the date time. According to [20], the advantages of Prophet are it is strong to missing data, capturing the shifts in the trend and large outliers. In [15], this model is described to has a good processing ability for predicting highly seasonal data with long-term non-stationary trends or for missing data. Besides that, it is also said by [21] that the model can be used to predict high quantity of abnormal and irregular data pattern in solving the problem of estimating the telecom systems with time series data. Other than that, [22] study also reveals that Prophet algorithm can deliver better result than ARIMA in his work of predicting the building power consumption. In general Prophet is an additive model that can be formulated as follow [19]:

$$m(t) = n(t) + o(t) + p(t) + \epsilon$$

where

$n(t)$ is the trend(s) which foresees long-term increase or decrease in data.

$o(t)$ is the Fourier series with seasonality that indicatethe effect of season related factor(s).

$p(t)$ is holidays or large event that have impacts on the time series.

ϵ is the error term that is irreducible.

Methodology

In this research, 3 time series models which are known as ARIMA, LSTM and Prophet will be used to forecast the new cases of COVID-19 in Malaysia. All the experiment is conducted in Python using the Jupyter Notebook. The overview of this research methodology is shown in Figure 2.

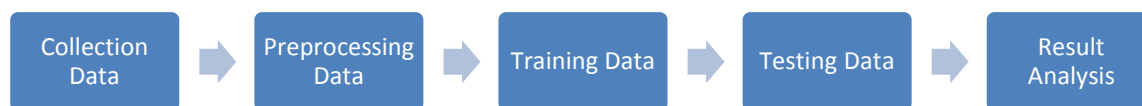


Figure 2: Overview of research methodology.

Basically, the development of this research begins with the collection of data. The data is taken from WHO website [23]. This data comprises of daily cases of the new confirmed COVID-19 globally starting from January 3, 2020 to November 18, 2020. Although the data has 9 columns in total (Figure 3), the focus of this research is only on 2 columns which are the "Date_reported" and the "New_cases" columns. The

“Date_reported” column is used to define the time series of the data and “New_cases” column is the output of the forecast. Besides that, the data is also filtered based on Country in which Malaysia is the main interest in this research.

	A	B	C	D	E	F	G	H	I
1	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths	
40089	3/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40090	4/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40091	5/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40092	6/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40093	7/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40094	8/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40095	9/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40096	10/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40097	11/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40098	12/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40099	13/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40100	14/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40101	15/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40102	16/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40103	17/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40104	18/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40105	19/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40106	20/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40107	21/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40108	22/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40109	23/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40110	24/1/2020	MY	Malaysia	WPRO	0	0	0	0	
40111	25/1/2020	MY	Malaysia	WPRO	7	7	0	0	
40112	26/1/2020	MY	Malaysia	WPRO	1	8	0	0	

Figure 3: Sample of the data for COVID-19 cases in excel format.

In the pre-processing phase, the dataset will be checked for any missing values. After that, the data is split into training and testing data. All the rows that are available will be used as training data except for the last 30 days which is split to become the testing data.

The training data is used to develop the models for forecasting while the testing data is separated to validate the accuracy of the forecast. After that, the prediction of 3 forecasting models will be performed. Only ARIMA model is experimented with different configurations to further analyse on the performance of the model. The performance metric of the testing data for comparison that is used in this research is Mean Square Error (MSE) and Root Mean Square Error (RMSE) methods.

Result and Discussion

The general overview of the dataset is shown in Figure 4. Based on the figure, the number of new cases increase significantly around mid-September until end of the data. This will be very challenging for the model to successfully predict the forecast value of new cases.

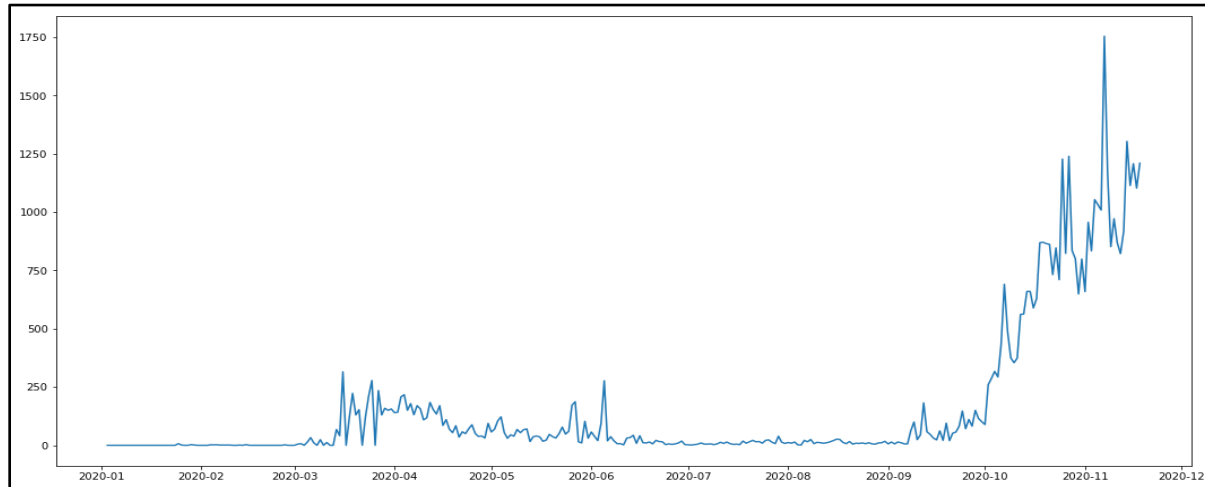


Figure 4: General overview of COVID-19 new cases in Malaysia.

Figure 5 displays the patterns of the data for new cases of COVID-19 in Malaysia. The common approach of analysing the pattern is by illustrating the Trend, Seasonality and Residual. This data shows small increment in Trend around mid-March and starts to drop at the end of June. However, the data shows a sudden spike in the number of new cases around mid-September until the last dated data. In terms of seasonality, it occurs regularly along the timeline. After the trend and seasonal is removed, the residuals represent the remaining data from the time series. Ideally, the smallest value of residuals is good for the models to achieve high performance in forecasting.

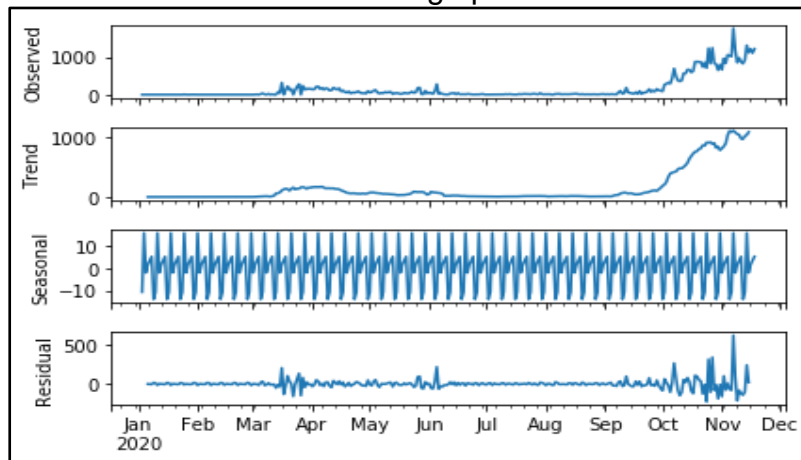


Figure 5: Pattern of the data for COVID-19 new cases in Malaysia.

After the models are trained, the results of each model in predicting the number of new cases are presented in Figure 6. The figure shows the overview of predicted number of new cases starting from October 20, 2020 until November 9, 2020.

	New_cases	LSTM_Predictions	Prophet_Predictions	ARIMA_Predictions111	ARIMA_Predictions112	ARIMA_Predictions211	ARIMA_Predictions212
2020-10-20	865	808.328627	394.346257	808.669417	829.893761	794.455887	851.494419
2020-10-21	862	823.816725	412.816107	805.657232	864.024177	798.132571	876.992267
2020-10-22	732	843.287169	406.001121	807.912177	897.875463	809.765799	913.487938
2020-10-23	847	862.767217	418.550245	810.634813	931.450657	812.342749	952.674733
2020-10-24	710	882.492047	428.630245	813.398978	964.752705	814.123319	992.523541
2020-10-25	1228	902.219577	427.759542	816.166830	997.784761	816.925215	1032.538699
2020-10-26	823	922.598649	449.812556	818.935009	1030.549587	819.805634	1072.599132
2020-10-27	1240	944.102109	438.566570	821.703218	1063.050151	822.570989	1112.675262
2020-10-28	835	966.547112	457.036420	824.471429	1095.289331	825.328750	1152.759864
2020-10-29	801	989.803971	450.221434	827.239640	1127.269973	828.099461	1192.851175
2020-10-30	649	1014.080973	462.770558	830.007852	1158.994893	830.870888	1232.948763
2020-10-31	799	1039.305747	472.850557	832.776063	1190.466875	833.640859	1273.052522
2020-11-01	659	1065.227748	471.979854	835.544274	1221.688672	836.410765	1313.162427
2020-11-02	957	1092.233124	494.032869	838.312486	1252.663010	839.180834	1353.278471
2020-11-03	834	1120.286159	482.786882	841.080697	1283.392583	841.950909	1393.400653
2020-11-04	1054	1149.331509	501.256732	843.848908	1313.880056	844.720566	1433.528971
2020-11-05	1032	1179.369902	494.441746	846.617120	1344.128064	847.491022	1473.663427
2020-11-06	1009	1210.459690	506.990870	849.385331	1374.139217	850.261081	1513.804020
2020-11-07	1755	1242.481987	517.070870	852.153542	1403.916092	853.031139	1553.950749
2020-11-08	1168	1275.531798	516.200167	854.921754	1433.461242	855.801197	1594.103614
2020-11-09	852	1309.604244	538.253181	857.689965	1462.777188	858.571255	1634.262617

Figure 6: The forecast values of the new cases using the 3 models against the real values.

Each of the predicted number of new cases for the tested data in the last 30 days of the overall dataset is plotted in Figure 7. This figure will help in visualizing the performance of each models used in this research. Based on figure, most of the models is plotted around the correct value of number of new cases in blue line. Only Prophet that seems to predict the value lower than the real value of the new cases. Besides that, all the graph shows a steady increase of number of predicted new cases. However, the ARIMA (1,1,1) shows almost a constant line without any increment or drops in number of predicted new cases.

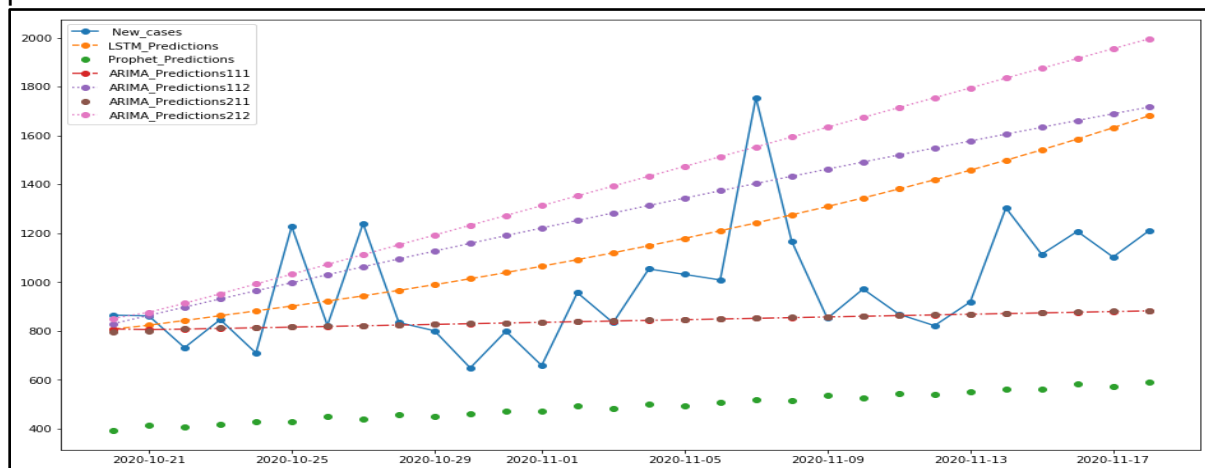


Figure 7: The graph of forecast values of the new cases against the real values.

The overall performance based on MSE value and RMSE value is tabulated in Table 1. It can be decided that the ARIMA model for (2,1,1) setup has achieved the best

performance among other models. This is because ARIMA model works well in predicting non-seasonal data with small amount of training data. ARIMA (2,1,1) has recorded the lowest value of MSE errors and RMSE errors in which the values are 66331.53 and 257.55 respectively followed by ARIMA (1,1,1) and LSTM. The lowest performance is attained by Prophet model with MSE value of 277708.81 and almost double the RMSE errors value which is 526.98. Unless the exogenous variables are included and more data is recorded, the models like LSTM and Prophet might perform better.

Table 1: The RMSE and MSE results based on the experimented models.

Models	RMSE Errors	MSE Errors
LSTM	328.067570	107628.330516
Prophet	526.980845	277708.810945
ARIMA111	257.801239	66461.478664
ARIMA112	412.833655	170431.626668
ARIMA211	257.549079	66331.528179
ARIMA212	545.899559	298006.328888

Conclusion

In this paper, the performances of 3 time series forecasting models in predicting the number of COVID-19 new cases are presented. The result shows that ARIMA (2,1,1) has the lowest MSE errors and RMSE errors value compared to LSTM, Prophet and other ARIMA models configuration. Therefore, it can be summarized that ARIMA (2,1,1) works well with satisfying result in forecasting the number of new cases for COVID-19 in Malaysia. This model can work with limited amount of data and suitable to help the government predicting the new cases that might occur and the necessary precaution actions. Besides that, the algorithm is also very simple and suitable for non-seasonality data.

Implementing this model in forecasting the COVID-19 new cases will help the government in preparing for safety measures to prevent the disease from spreading. They can take actions not solely based on the assumption but to show the proven data analysis to further convince the people on the potential upcoming threat. During this period, everyone is not in stable condition. Thus, we need to take care of each other as the hash tag promoted by the government, #kitajagakita.

In future works, this research will be conducted by including the exogenous variables such as the effect of MCO, holidays, etc. As proven, people activities will affect the social distancing that might spread the disease and give impact to the number of new cases. Implementing this study will also help to discover the influence of MCO in preventing this disease.

REFERENCE

- [1] J. Ducharme, "World Health Organization Declares COVID-19 a 'Pandemic.' Here's What That Means," 2020. [Online]. Available: <https://time.com/5791661/who-coronavirus-pandemic-declaration/>. [Accessed: 19-Nov-2020].
- [2] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020." [Online]. Available: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>. [Accessed: 19-Nov-2020].
- [3] "Terkini | COVID-19 MALAYSIA." [Online]. Available: <http://covid-19.moh.gov.my/terkini>. [Accessed: 19-Nov-2020].
- [4] V. Bhavana, P. Thakor, S. B. Singh, and N. K. Mehra, "COVID-19: Pathophysiology, treatment options, nanotechnology approaches, and research agenda to combating the SARS-CoV2 pandemic," *Life Sci.*, vol. 261, no. August, p. 118336, 2020.
- [5] B. M. Bakadia *et al.*, "Prevention and treatment of COVID-19: focus on interferons, chloroquine/hydroxychloroquine, azithromycin, and vaccine," *Biomed. Pharmacother.*, 2020.
- [6] T. P. B. Thu, P. N. H. Ngoc, N. M. Hai, and L. A. Tuan, "Effect of the social distancing measures on the spread of COVID-19 in 10 highly infected countries," *Sci. Total Environ.*, vol. 742, p. 140430, 2020.
- [7] S. Salim, "Health DG urges public to comply with MCO to reduce number of positive COVID-19 cases in Malaysia | The Edge Markets." [Online]. Available: <https://www.theedgemarkets.com/article/health-dg-urges-public-comply-mco-reduce-number-positive-covid19-cases-malaysia>. [Accessed: 19-Nov-2020].
- [8] B. Siregar, E. B. Nababan, A. Yap, and U. Andayani, "Forecasting of Raw Material Needed for Plastic Products Based in Income Data Using ARIMA Method," *Int. Conf. Electr. Electron. Inf. Eng.*, pp. 135–139, 2017.
- [9] Y. Wang and Y. Guo, "Forecasting Method of Stock Market Volatility in Time Series Data Based on Mixed Model of ARIMA and XGBoost," pp. 205–221, 2019.
- [10] Y. Pan, M. Zhang, Z. Chen, M. Zhou, and Z. Zhang, "An ARIMA Based Model for Forecasting the Patient Number of Epidemic Disease," pp. 31–34, 2015.
- [11] S. Noureen, S. Atique, V. Roy, and S. Bayne, "Analysis and application of seasonal ARIMA model in Energy Demand Forecasting : A case study of small scale agricultural load," pp. 521–524, 2019.
- [12] L. Wang and Z. Zhang, "Research on Shanghai Copper Futures Price Forecast Based on X12-ARIMA-GARCH Family Models," *Int. Conf. Comput. Inf. Big Data Appl.*, pp. 304–308, 2020.
- [13] L. E. Jeronimo-Martinez, R. E. Menendez-Mora, and H. Bolivar, "Forecasting Acute Respiratory Infection cases in Southern Bogota : EARS vs . ARIMA and SARIMA," 2017.
- [14] B. Singh, P. Kumar, and D. N. Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling," *First Int. Conf. Power, Control Comput. Technol.*, vol. 1, pp. 38–43, 2020.
- [15] W. X. Fang, P. C. Lan, W. R. Lin, H. C. Chang, H. Y. Chang, and Y. H. Wang, "Combine Facebook Prophet and LSTM with BPNN Forecasting financial markets: The Morgan Taiwan Index," *Proc. - 2019 Int. Symp. Intell. Signal Process. Commun. Syst. ISPACS 2019*, pp. 2019–2020, 2019.
- [16] A. Essien and C. Giannetti, "A Deep Learning Model for Smart Manufacturing Using Convolutional LSTM Neural Network Autoencoders," *IEEE Trans. Ind. Informatics*, vol. 16, no. 9, pp. 6069–6078, 2020.
- [17] Y. Wang, S. Zhu, and C. Li, "Research on Multistep Time Series Prediction Based on LSTM," *3rd Int. Conf. Electron. Inf. Technol. Comput. Eng.*, pp. 1155–1159, 2019.
- [18] S. Liu, G. Liao, and Y. Ding, "Stock transaction prediction modeling and analysis based on LSTM," *Proc. 13th IEEE Conf. Ind. Electron. Appl. ICIEA 2018*, pp. 2787–2790, 2018.

- [19] C. R. Madhuri, M. Chinta, and V. V. N. V. P. Kumar, "Stock market prediction for time-series forecasting using prophet upon ARIMA," *2020 7th Int. Conf. Smart Struct. Syst. ICSSS 2020*, pp. 0–4, 2020.
- [20] I. Yenidoğan, A. Çayır, O. Kozan, T. Dağ, and Ç. Arslan, "Bitcoin Forecasting Using ARIMA and PROPHET," *3rd Int. Conf. Comput. Sci. Eng.*, pp. 6–9, 2018.
- [21] G. Jain and R. R. Prasad, "Machine learning, Prophet and XGBoost algorithm: Analysis of Traffic Forecasting in Telecom Networks with time series data," *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.)*, pp. 893–897, 2020.
- [22] F. Gong, N. Han, D. Li, and S. Tian, "Trend Analysis of Building Power Consumption Based on Prophet Algorithm," *2020 Asia Energy Electr. Eng. Symp. AEEES 2020*, pp. 1002–1006, 2020.
- [23] "WHO Coronavirus Disease (COVID-19) Dashboard." [Online]. Available: <https://covid19.who.int/>. [Accessed: 20-Nov-2020].