Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Stock Price Prediction Using ARIMA, Neural Network and LSTM Models

To cite this article: M K Ho *et al* 2021 *J. Phys.: Conf. Ser.* **1988** 012041

Stock Price Prediction Using ARIMA, Neural Network and LSTM Models

M K Ho[1], Hazlina Darman[2] and Sarah Musa[3]

[1,2,3]School of Mathematics, Actuarial and Quantitative Studies (SOMAQS), Asia Pacific University of Technology and Innovation (APU), Technology Park Malaysia, 57000 Kuala Lumpur, Malaysia

Email: [1]dr.ming.kang@apu.edu.my; [2]hazlina.darman@apu.edu.my; [3]sarah.musa@apu.edu.my

Abstract. Since the past decades, prediction of stock price has been an important and challenging task to yield the most significant profit for a company. In the era of big data, predicting the stock price using machine learning has become popular among the financial analysts since the accuracy of the prediction can be improved using these techniques. In this paper, auto-regressive integrated moving average (ARIMA), neural network (NN) and long short-term memory network (LSTM) have been used to predict Bursa Malaysia's closing prices data from 2/1/2020 to 19/1/2021. All the models will be evaluated using root mean square errors (RMSE) and mean absolute percentage errors (MAPE). The results showed that LSTM able to generate more than 90% of accuracy in predicting stock prices during this pandemic period.

1. Introduction

Coronavirus disease or COVID-19 is an infectious disease which were first discovered in Wuhan, China in December 2019. The disease has become a fastest spreading disease and the statistics has shown that more than two million people have died because of COVID-19. To date, Malaysia is currently ranked in the 29th place in the list of countries with the highest cases of COVID-19 in a two-week period, data based from US-based Johns Hopkins University. Malaysia also recorded the second highest cumulative COVID-19 cases per one million populations in ASEAN.

Figure 1 below showed the number of daily cases of COVID-19 in Malaysia from 25/1/2020 to 22/1/2021. The number of cases of COVID-19 are low in January and February 2020. In the mid of March 2020, Malaysia has shown a significant increase which has made the government to implement the first Movement Control Order (MCO). The MCO phase 1 period as announced by Prime Minister of Malaysia was begun from 18/3/2020 until 31/3/2020 [1]. On 25/3/2020, the Prime Minister announced the extension of MCO which is MCO phase 2 until 14/4/2020 [2]. Conditional Control Movement Order (CMCO) were implemented from 4/5/2020 until 9/6/2020 which allowed certain business sectors to resume operations [3]. Then, Recovery Movement Control Order (RMCO) were introduced from 10/6/2020 until 31/12/2020. As can be seen from the figure below, number of daily cases increase rapidly in October 2020 [4]. This has made the enforcement of CMCO from 14/10/2020 until 9/11/2020, CMCO in the areas with high cases from 14/12/2020 until 31/12/2020, RMCO nationwide from 1/1/2021 to 31/3/2021. High number of cases in Selangor, Kuala Lumpur, Johor, and Penang has made the enforcement of MCO until 4/3/2021, while other states are under CMCO, and Perlis is under RMCO.

Along with great impact on people's health and life, the crisis also gives significant impact on the global financial market, including Malaysia. Therefore, understanding the impact of COVID-19 on stock price has become an important issue and its prediction using appropriate time series method is immensely crucial. For instance, accurate prediction of stock price will provide investors with useful decision-making information and to maximize their profit.

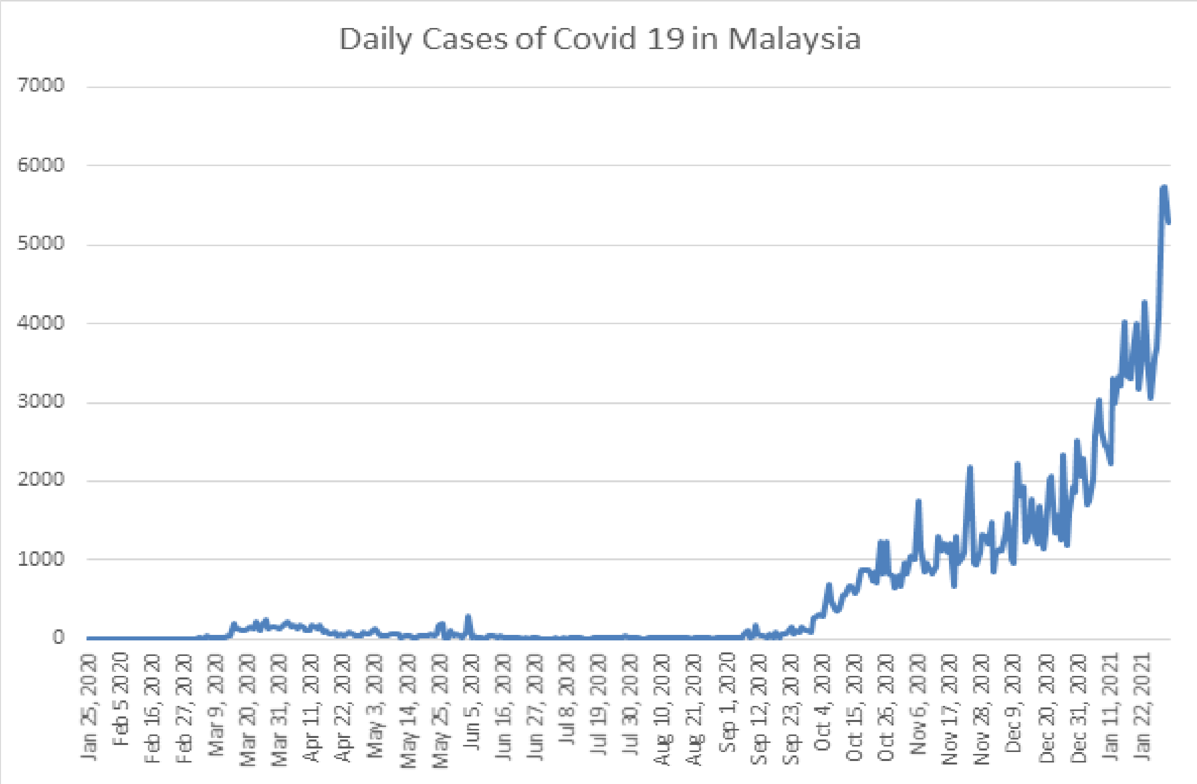

Figure 1. Daily Cases of COVID-19 in Malaysia.

Stock exchange of Malaysia, Bursa Malaysia, which was previously known as Kuala Lumpur Stock Exchange (KLSE) plays an important role to allow the trading of Malaysian securities including clearing, trading, listing and settlement services. It comprises 30 largest listed companies and accounted about 60 percent in total of Bursa Malaysia Stock Exchange, and often being used as the benchmark of the overall market performance of Malaysian stocks [5]. On 26 March 2020, Bursa Malaysia announces additional relief measures targeted to a broader group of participants within the capital market. These measures are designed to help lessen the financial burden and provide greater flexibility in navigating the challenging period posed by the COVID-19 pandemic [6].

In the next section, closing stock prices of Bursa Malaysia is predicted using various machine learning methods. Classical forecasting techniques such as moving average and exponential smoothing are not considered in this study since stock prices of Bursa Malaysia for the period 2020-2021 is highly volatile, dynamic and complex. Historical data (2/1/2020 – 19/1/2021) which downloaded from *Yahoo Finance* is collected and used three models to predict Bursa Malaysia's closing price. The models are Auto-Regressive Integrated Moving Average (ARIMA), Neural Network (NN), and Long Short-Term Memory network (LSTM).

2. Related Works

Stock market is one of the main indicators for economic condition. Despite claims that stock market is an efficient market that follows a random walk process and therefore is not predictable, there are some evidence that from previous studies that stock market can be predicted [7] [8]. [9] believed that accurate model and designing of appropriate variables able to predict the stock prices and stock price movement patterns. Their paper proposed an approach of hybrid modelling for stock price prediction using machine learning and deep learning-based models. Four deep learning-based regression models using long-and short-term memory (LSTM) network are constructed and found that LSTM-based univariate model that used one-week prior data as input for predicting is the most accurate model.

[10] developed a model using Recurrent Neural Network (RNN) and LSTM to predict future stock market. They compared the accuracy of ARIMA and LSTM methods to forecast the time series data and the results showed that LSTM is far more superior to ARIMA. In [8], eight classification models are presented for short-term stock price movement forecasting, namely logistic regression, LSTM, Decision Tree, Bootstrap Aggregation (Bagging), Boosting, Random Forest, ANN, and Support Vector Machines (SVM) classifications. Among all these classifications, ANN on average, produced the highest level of accuracy, while LSTM provided a very accurate prediction on stock price movement in a short interval of time. It is undeniable that many studies believed that LSTM method improves the accuracy in predicting the stock prices, as shown in papers by [11], [12], and [13]. Irrefutably, many papers used ARIMA method to predict the stock prices due to its simplicity and wide acceptability of the model, despite many other forecasting models give more accurate prediction, such as [14], [15], and [16].

In Malaysia, many studies have been conducted in predicting Bursa Malaysia stock price using various methods such as Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), Log-Normal Distribution, Geometric Brownian motion and Fuzzy Time Series. [17] predicted the individual stock movements in Bursa Malaysia using online news where they combined Automatic Text Extraction and Neural Networks (NN) prediction in their study. [18] compared two models for forecasting stock prices in Bursa Malaysia, namely Fuzzy time series and Geometric Brownian motion. While, study of [7] has proved the ability of Artificial Neural Network (ANN) method, an artificial intelligence system, in predicting the movement of the FTSE Bursa Malaysia. Previously, [5] in their study also used two methods, namely ARIMA and ANN to predict Bursa Malaysia stock prices where their findings revealed that ANN method outperformed the ARIMA. Meanwhile, [19] has compared multiple linear regression (MLR) method and ANN to predict the Malaysia stock exchange market and the result is the prediction model using ANN is more accurate than MLR.

The objective of this study is to collect the daily closing prices of FTSE Bursa Malaysia over one-year period and develop a robust forecasting framework for forecasting Bursa Malaysia closing prices using three forecasting models, namely ARIMA, LSTM and Neural Network (NN). Later, the best predicting model will be determined based on their accuracy. Data from January 2020 to January 2021 is chosen as this is the period where the pandemic of Covid-19 affected Malaysia's stock prices the most. It is believed the best model is likely to learn from the past movement pattern of the daily closing prices of Bursa Malaysia and effectively exploit the past data in forecasting the future closing prices of Bursa Malaysia.

3. Dataset and Features

In this paper, historical trading data from 2/1/2020 to 19/1/2021 for Bursa Malaysia closing prices is taken. The data contains trading information on high price, low price, open price, close price, adjusted close price and trading volume but only adjusted close price is used in this study. Based on the time series plot for Bursa Malaysia closing prices in figure 2, the performance of stock price was stable at the beginning of 2020 because the number of COVID-19 cases in Malaysia remained low [16]. Starting from March 2020, the number of COVID-19 cases in Malaysia had increased dramatically which affect a significant downward trend of stock price in March-April 2020. This dramatic stock market crash also due to the rapid grow of COVID-19 cases around the world. Thereafter, KLCI closing price had come back to the level of 1600 on January 2021.

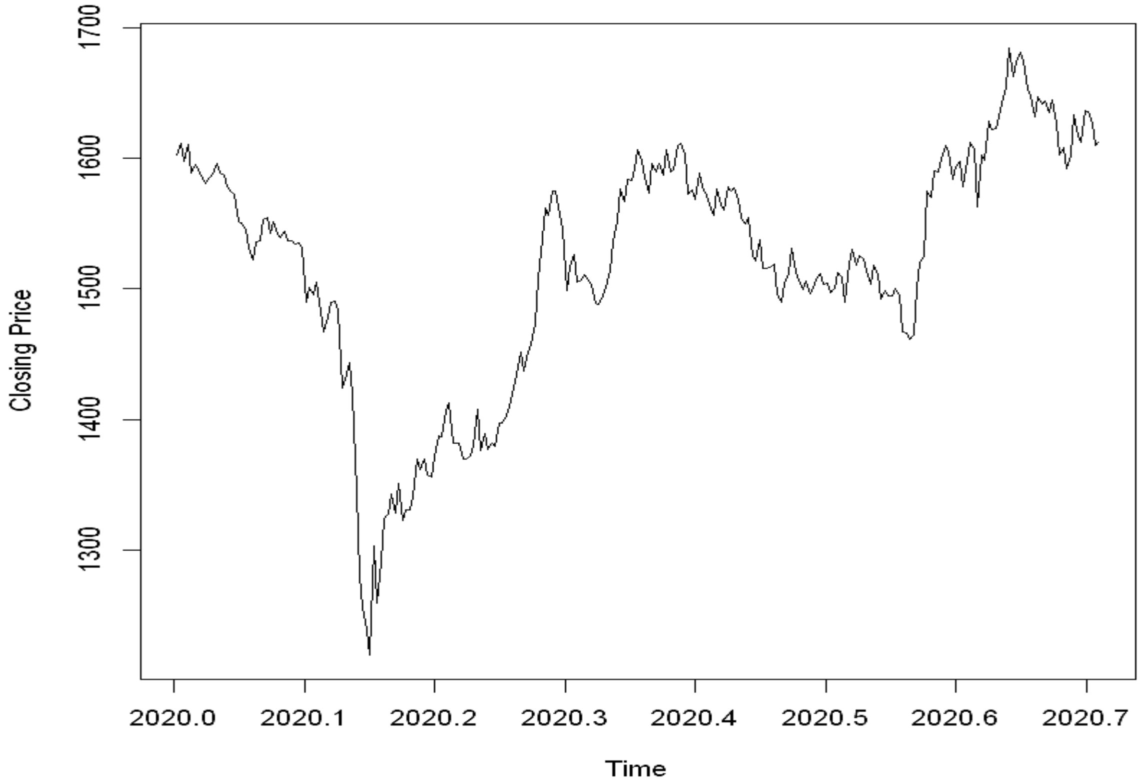

Figure 2. Adjusted closing price for Bursa Malaysia closing prices from 2/1/2020 to 19/1/2021.

4. Methodology

To evaluate the performance of each forecasting model, the data is divided into two parts: 70% of the data (from 2/1/2020 to 28/9/2020) is used to train and optimize the parameters in the models meanwhile the other 30% (from 29/9/2020 to 19/1/2021) is used to test the performance of the models on "unseen" dataset. Three forecasting models are applied in this study, which are autoregressive integrated moving average (ARIMA) model, neural network (NN) and long short-term memory model (LSTM).

4.1. Autoregressive Integrated Moving Average (ARIMA)

An autoregressive integrated moving average (ARIMA) is a model that uses time series data to make predictions. It is characterized by 3 terms: p, d, q where p represents the order of AR term, q represents the order of MA term and d represents order of differencing to transform a non-stationary series to stationary. The methodology starts with identification of models based on behaviour of autocorrelation (ACF) and partial autocorrelation (PACF) plots. Once a model is specified, the parameters of the model are estimated and a diagnostic checking for the adequacy of the model is done by using Ljung-Box statistic test. If the model is adequate, the model can proceed to forecast.

4.2. Neural Network (NN)

Neural Network (ANN) is a computing system that inspired by biological neurons of human brain. A human brain acts like a processor which can process information and compute output from inputs. Similar with biological neurons of human brain, ANNs composed of multiple nodes which are connected by links and interact with each other. Input data can take by the nodes and simple operations on the data are performed by the nodes. The result is passed to the other neurons where the output of each node is known as activation value. Each of the nodes sum the activation values and modifies the value. The process will continue through the network, through hidden layers, until it reaches the output nodes.

A neural network can be thought of as a network of 'neurons' which are organized in layers. The predictors (inputs) form the bottom layer, and the forecasts (or outputs) form the top layer. They may also be intermediate layers containing 'hidden neurons'. Linear regressions are equivalent to the simplest networks with no hidden layers in machine learning. The neural network becomes non-linear once we add an intermediate layer with hidden neurons. This is a multilayer feed-forward network, where each layer of nodes receives inputs from the previous layers. For example, the inputs into hidden neuron *j* are combined linearly to give

$$z_j = b_j + \sum_{i=1}^{4} w_{i,j} x_i \tag{1}$$

In the hidden layer, this is then modified using sigmoid function to give the input for the next layer.

$$s(z) = \frac{1}{1+e^{-z}} \tag{2}$$

4.3. Long Short-Term Memory Network

Long Short-Term Memory (LSTM) network is a type of recurrent neural network capable of learning order dependence in sequence prediction problems. All recurrent neural networks have the form of a chain of repeating modules of neural network. LSTMs also have this chain like structure, but the repeating module has a different structure. There are four neural network layers, interacting as below.

4.3.1. Front Gate

The first step in LSTM is to determine or to control what information we are going to throw away from the cell state or memory. It is decided by the first sigmoid layer called the 'forget gate layer':

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

4.3.2. Input Gate

The next step is to determine what new information we are going to store in the cell state from current input. There are two parts called 'input gate layer' and 'tanh layer'. The input gate is another sigmoid layer which outputs are between 0 and 1 and decides which values we will update. Tanh layer will create a vector of new candidate values, $\tilde{C}_t$ which will be used to update the cell state. These two layers are combined to create an update to the state.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{4}$$

$$\tilde{C}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) \tag{5}$$

Next step is to update the old cell state, C_{t-1} into the new cell state, C_t.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{6}$$

4.3.3. Output Gate

To decide what parts of the cell state we are going to output from the memory, we must run a sigmoid layer first, where the output will be based on our cell state. Then, a tanh-layer is used to put the cell state to push the values to be between -1 and 1 and multiply it by the output of the sigmoid gate.

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1,} x_t + b_o\right]\right) \tag{7}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{8}$$

5. Results and Discussion

To evaluate the performances of each forecasting model in predicting Bursa Malaysia closing price, root mean square error (RMSE) and mean absolute percentage error (MAPE) are chosen. The equation of each indicator is shown:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2} \tag{9}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{10}$$

where Y_i is the actual observations, $\hat{Y}_i$ is the predicted value obtained proposed forecasting model and N is the number of forecast values. Model with the minimum RMSE and MAPE will be chosen as the best forecasting model.

5.1. ARIMA models

Figure 3 shows the autocorrelation (ACF) and partial autocorrelation (PACF) plots for the training data after performing first non-seasonal differencing (I(1)). Based on the behaviour of ACF and PACF plots, random walk model and random walk with drift are proposed since a phenomena of white noise process presented in both plots. Besides, ARIMA(1,1,1) and ARIMA(2,1,2) are also proposed since a phenomena of dies down exponentially with sine-wave presented in both plots.

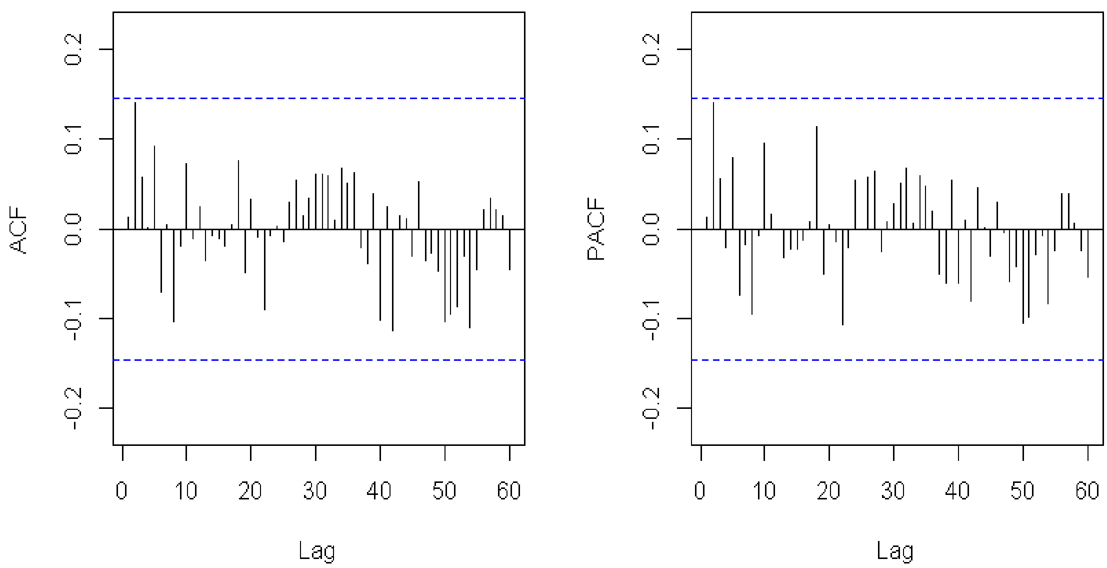

Figure 3. ACF and PACF plots for the training data after differencing.

To check the adequacy of the proposed ARIMA models, Ljung-Box test is applied to examine the autocorrelation of the residuals. For instance, the model is adequate if the p-value of the test is greater than $\alpha = 0.05$. Table 1 showed the results of random walk, random walk with drift models, ARIMA(1,1,1) and ARIMA(2,1,2) on the training and testing data. Based on the results, ARIMA(2,1,2) is preferred because the RMSE and MAPE are the lowest. All the ARIMA models are adequate since the p-values in Ljung-Box test are greater than 0.05. However, all the models are prone to overfit to the data because of large RMSE and MAPE values in the testing data compared to training data. In figure 4, it is clearly showed that none of the models able to produce a good forecast on the testing data because the k-step forecast follows exactly the last forecast value (a horizontal straight line), except for ARIMA(2,1,2) which able to produce the forecast follows a downward line.

Table 1. Summary of results for four ARIMA models.

Models	RMSE		MAPE		Ljung-Box test
	Training	Testing	Training	Testing	(p-value)
Random walk	18.0254	90.3542	0.8818	4.6015	0.6904
Random walk with drift	18.0184	111.6326	0.8836	5.6071	0.6904
ARIMA(1,1,1)	17.9427	89.1495	0.8787	4.5480	0.7260
ARIMA(2,1,2)	17.8178	88.3096	0.8602	4.5148	0.7092

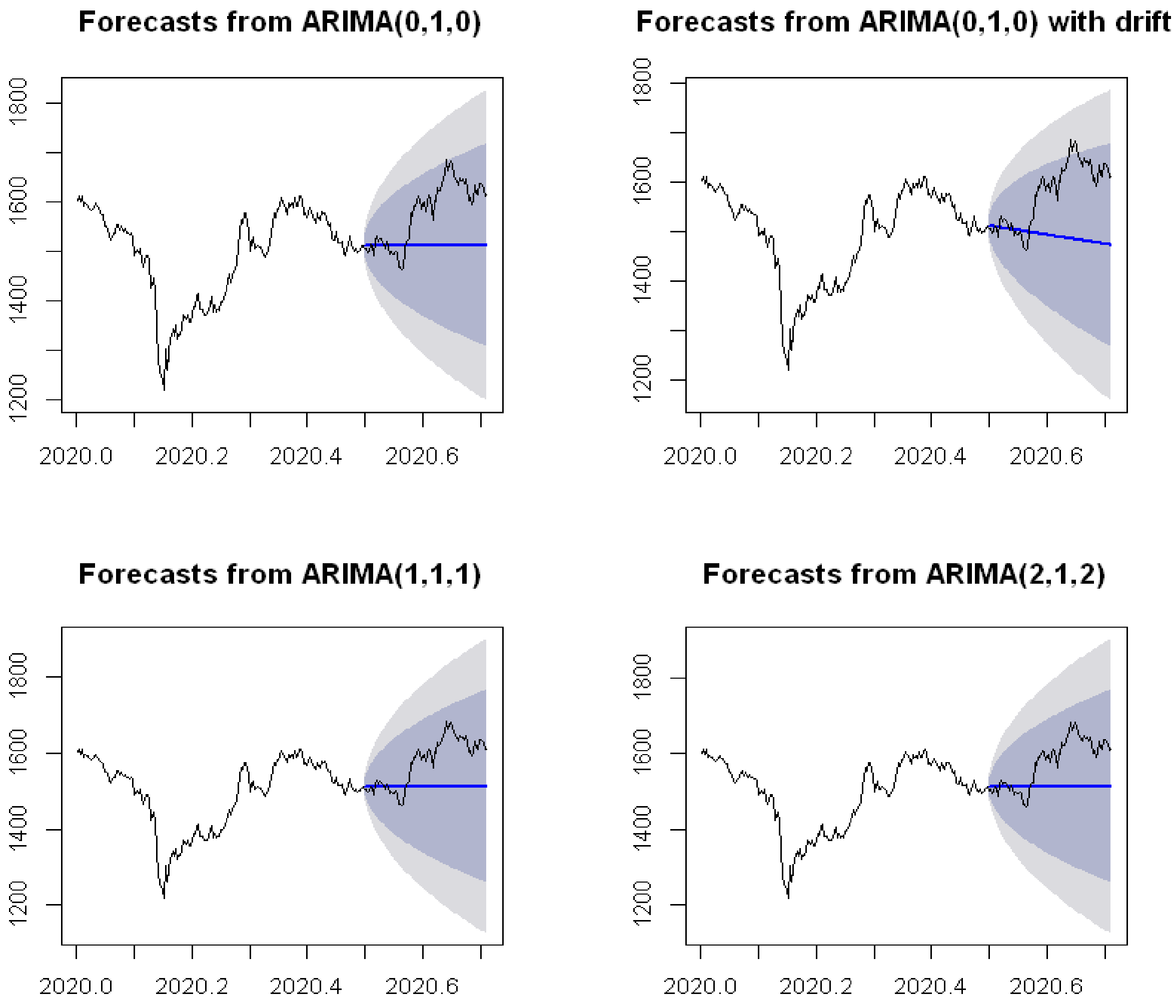

Figure 4. Predicted values (indicated using blue line) on the testing data using 4 ARIMA models.

5.2. NN Models

In this study, two feedforward neural networks are used. First NNAR model uses 5 lagged values and 3 nodes in the hidden layer. It has an average of 20 networks and each of which is a 5-3-1 network with 22 weights. Second NNAR model uses 4 lagged values and 2 nodes in the hidden layer. It also has an average of 20 networks and each of which is a 4-2-1 network with 13 weights. Table 2 shows the results of NNAR(5, 3) model on training and testing data.

Table 2. Summary of results for NNAR(5, 3) and NNAR (4, 2) models

Model	RMSE		MAPE	
	Training	Testing	Training	Testing
NNAR(5,3)	14.3501	93.8031	0.7258	4.7656
NNAR(4,2)	15.8855	70.8142	0.7913	3.7341

The result in table 2 showed that both NN models are overfitted to the stock price data because the RMSE and MAPE values in testing data are comparatively higher than in training data. The predicted values using NNAR(5,3) and NNAR(4,2) failed to capture high variations of stock price from September 2020 to January 2021. As shown in figure 5, predicted values using NNAR models only able to capture an upward trend if using NNAR(4,2) and a downward trend if using NNAR(5,3) during the testing period. This is mainly because the model used in this study is a simple neural network with only one hidden layer.

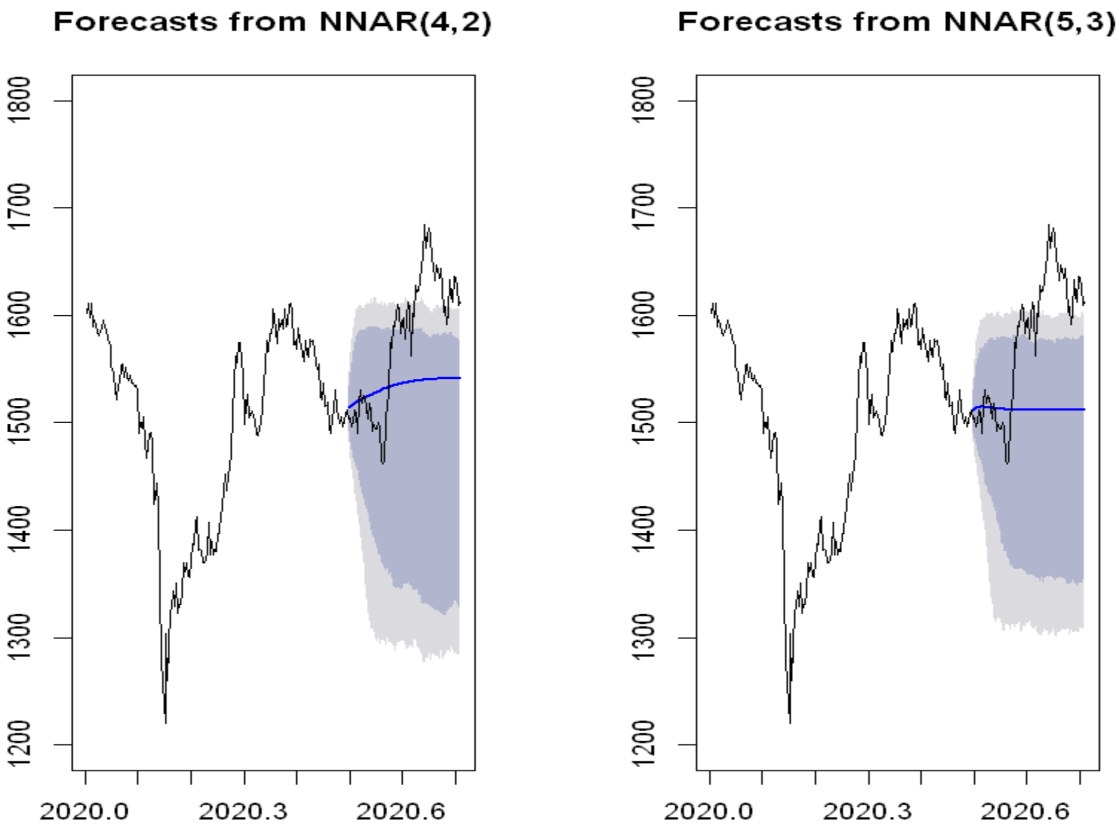

Figure 5. Predicted values (indicated using blue line) on the testing data using NNAR models.

5.3 LSTM model

LSTM models with two and three hidden layers are used for comparison. For instance, unit of each hidden layer in both LSTM models is designed to be 100 and one neuron in the output state. Other parameters used to build LSTM included Rectified Linear (*ReLU*) activation function and *Adam* optimization algorithm. Table 3 shows the results of LSTM models on training and testing data.

Table 3. Summary of results for LSTM models.

Model	RMSE		MAPE	
	Training	Testing	Training	Testing
LSTM with 2 hidden layers	16.7985	17.5289	0.8322	0.8373
LSTM with 3 hidden layers	16.7170	16.8410	0.8192	0.8184

The result in table 3 showed that the two LSTM models are free from overfitting problem because they have good predicting skills on both training and "unseen" testing data. By referring to RMSE and MAPE values, the best model is LSTM with 3 hidden layers. As shown in figure 6, LSTM model not only able to predict the future stock prices well with minimum errors but also able to recover the pattern and behaviour of Bursa Malaysia closing prices in the testing period. This result also indicted that LSTM model is the best model in predicting Bursa Malaysia stock prices because it has the lowest RMSE and MAPE values compared to random walk and NNAR models.

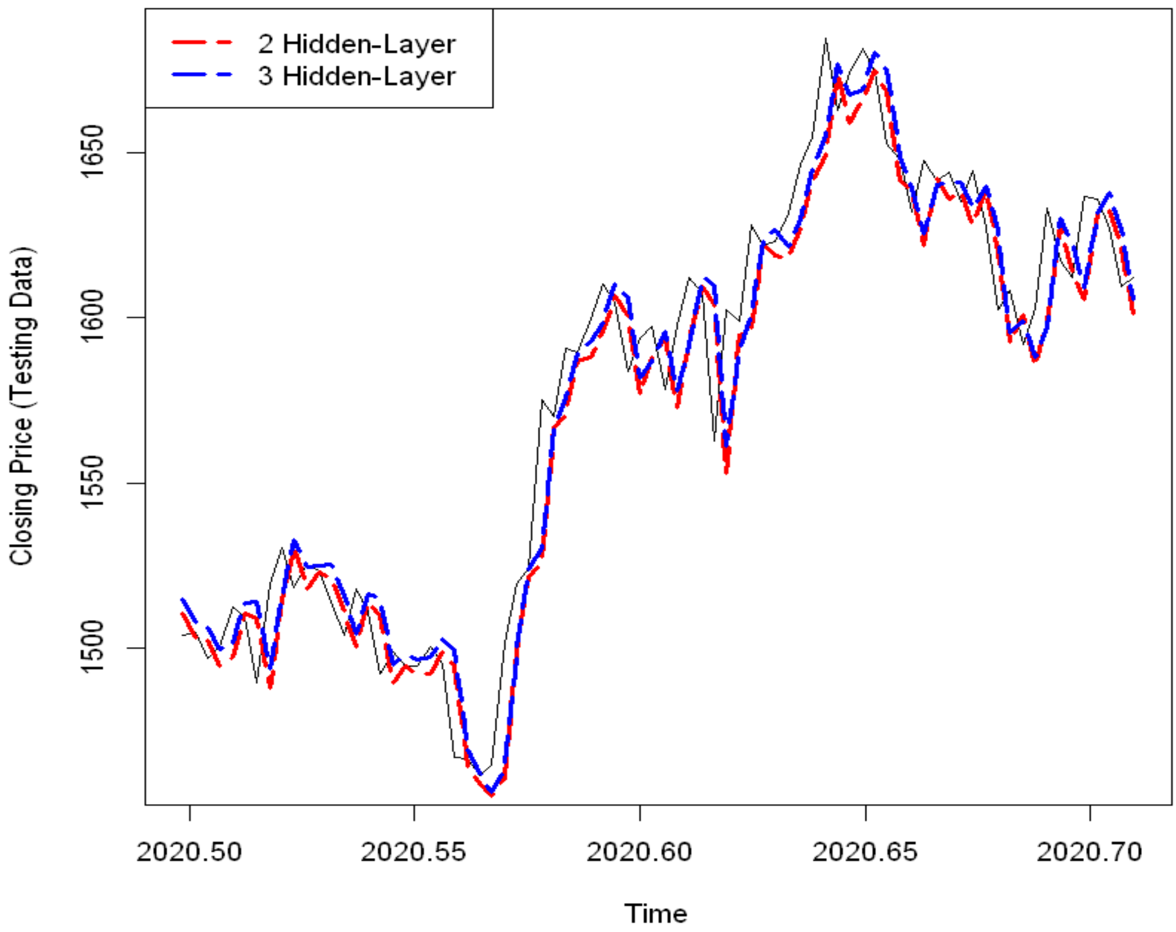

Figure 6. Predicted values on the testing data using LSTM models.

6. Conclusion

In this paper, three forecasting machine learning models are used including ARIMA, NN model and LSTM models to predict Bursa Malaysia closing stock prices from 2/1/2020 to 19/1/2021. This range of time is chosen because it was the period where the number of COVID-19 cases increased dramatically in Malaysia, which also trigger an unpredictable movement of stock prices in 2020. To choose the best machine learning model, MAPE and RMSE are chosen. Among the three models, LSTM model has the best performance in Bursa Malaysia stock price prediction because it has the smallest MAPE and RMSE values. For instance, LSTM model not only able to generate more than 90% of accuracy but also able to explain unpredictable movement of stock prices during this pandemic period. In the future, financial indicators can be included to improve the accuracy of prediction since these indicators may influence the movement of the stock prices. It is also recommended to use sentiment analysis with LSTM model to make a better prediction.

References

[1] Wikipedia Ensiklopedia Bebas. 27 February 2021. Garis masa pandemik COVID-19 di Malaysia. https://ms.wikipedia.org/wiki/Garis_masa_pandemik_COVID-19_di_Malaysia
[2] Adib Povera, Hana Naz Harun 23 April 2020. “MCO Phase 4 from April 29 to May 12”. New Straits Times. Archived from the original 28 February 2021. Retrieved 28 February 2021.
[3] Zakiah Koya 10 May 2020. “Conditional MCO extended for another four weeks to June 9”. The Star. Archived from the original 28 February 2021. Retrieved 28 February 2021.
[4] Statista. 27 February 2021. Number of daily confirmed cases of the novel coronavirus infection COVID-19 in Malaysia as of February 22, 2021. https://www.statista.com/statistics/1110785/malaysia-covid-19-daily-cases/
[5] Chan T and Sy A 2015 Prediction of Bursa Malaysia Stock Index using Autoregressive Integrated Moving Average and Artificial Neural Network *My Stats 2015.*
[6] Bursa Malaysia. 27 February 2021. Bursa Malaysia announces additional relief measures to alleviate the impact of Covid-19 on Capital Market Players. https://www.bursamalaysia.com/about_bursa/media_centre/bursa-malaysia-announces-additional-relief-measures-to-alleviate-the-impact-of-covid-19-on-capital-market-players
[7] Chan T, Sy A and Vc Y 2017 Prediction of Bursa Malaysia Equity Index using Artificial Neural Network *Malaysian Finance Association Annual Conference 2017.*
[8] Sen J 2018 Stock Price Prediction Using Machine Learning and Deep Learning Frameworks *International Conference on Business Analytics and Intelligence.*
[9] Mehtab S, Sen J and Dutta A 2020 Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models *Second Symposium on Machine Learning and Metaheuristics Algorithms.*
[10] Moghar A and Hamiche M 2020 Stock Market Prediction Using LSTM Recurrent Neural Network *Procedia Computer Science* **170** pp 1168-73.
[11] Roondiwala M, Patel H and Varma S 2017 Predicting Stock Prices Using LSTM. *International Journal of Science and Research (IJSR)* **6**.
[12] Chen K, Zhou Y and Dai F 2015 A LSTM-based method for stock returns prediction: A case study of China stock market *2015 IEEE International Conference on Big Data (Big Data),* Santa Clara, CA, USA, pp. 2823-24.
[13] Ghosh A, Bose S, Maji G, Debnath N and Sen S 2019 Stock Price Prediction Using LSTM on Indian Share Market *Proceeding of 32nd International Conference on Computer Applications in Industry and Engineering* **63** pp 101 - 10
[14] Jung-Hua W and Jia-Yann L 1996 Stock market trend prediction using ARIMA-based neural networks *Proceedings of International Conference on Neural Networks (ICNN'96)* **4** Washington, DC, USA pp. 2160-65.
[15] Alwadi S, Almasarweh M and Alsaraireh A 2018 Predicting Closed Price Time Series Data Using ARIMA Model. *Modern Applied Science*. **12**.

[16] Mondal P, Shit L and Goswami S 2014 Study of Effectiveness of Time Series Modeling (Arima) in Forecasting Stock Prices. *International Journal of Computer Science, Engineering and Applications* **4** pp 13-29.
[17] Seng L and Chan P P 2006 Prediction of Individual Stock Movements in Bursa Malaysia using Online News *SME-Enterpreneurship Global Conference 2006.*
[18] Shafii N, Ramli N, Alias R and Fauzi N 2019 Fuzzy Time Series and Geometric Brownian Motion in Forecasting Stock Prices in Bursa Malaysia *Jurnal Intelek* **14** pp 240-50.
[19] Abdu M S and Saratha S 2017 The Use Of Artificial Neural Network And Multiple Linear Regression For Stock Market Forecasting *Matematika* **33.**