

Stock Price Prediction Using ARIMA and LSTM Models: An Application to CSI 300 Closing Prices

Yifan Zhang *

Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, USA

* Corresponding Author Email: yifanz91@usc.edu

Abstract. In the development of the real economy, the performance of the stock market can well reflect its development status, so it is of great value to select effective forecasting methods to make a relatively accurate prediction of stock prices. The research and forecast object of this paper is the daily adjusted closing price of CSI 300 index. The sample data is derived from China Securities Data Website and is composed of six characteristic data for 261 days from July 1, 2022 to July 26, 2023. This paper compares the prediction ability of Autoregressive Integrated Moving Average (ARIMA) model and Long Short Term Memory (LSTM) networks. For the deviation function, this paper chooses MAE and RMSE two methods. By comparing the prediction ability of the models, the following conclusions are drawn: The daily closing price of CSI 300 has autocorrelation and long-term memory, which enables the prediction of future price based on historical data. Compared with ARIMA model, LSTM has stronger prediction ability, but the prediction has a lag, which is not conducive to reflecting the drastic trend changes in the short term. More characteristic information needs to be added to characterize to obtain more accurate results.

Keywords: ARIMA model; LSTM networks; CSI 300, stock prediction.

1. Introduction

Stock is the main investment product in the securities investment market, and it is the main source of income for investors in the securities market. To analyze the historical stock price and make a reasonable prediction of the future trend of stock change is the key to make profits in stock investment. With the development of the stock market, investors have a deeper understanding of the changing rules of the stock market, and the methods of stock analysis and prediction are gradually systematized.

In recent years, the factor investment theory, which quantifies the stock related factors and makes analysis of the quantitative factors, has been widely concerned by investors. Through quantitative analysis and prediction of stocks, not only can all factors be considered, but also can eliminate the negative impact of subjective factors on stock selection decisions. The ARIMA model represents the Autoregressive Integrated Moving Average, which is proposed in 1970 as the first time series model.[1]. Since the model can capture the trend, seasonality and randomness in the data, it is soon widely used in the forecast of stocks. Therefore, it can better describe the future direction of financial markets and predict the prices in the future, so it is widely used to analyze stock prices, exchange rate, and financial derivatives pricing. Zhang et al. employed ARIMA model to set up stock portfolio [2]. Ahmar et al. compared different variants of ARIMA model and tested prediction ability respectively [3]. Babu and Reddy used ARIMA model to predict Indian stock market, which shown that hybrid ARIMA model with multi-layer construct is able to increase the prediction accuracy of stock prices [4].

With the development of artificial neural network theory, reinforcement learning approaches are also applied in the financial field. According to the literature review published by Hambly in 2021, deep reinforcement learning has been widely used in the financial field in decision making, portfolio optimization and asset arbitrage [5]. As an important method of deep reinforcement learning, Artificial Neural Networks (ANN) are widely used in predictive analysis of various stock data. Kristjanpollar et al. combined the artificial neural network with the GARCH model to effectively predict the volatility of gold [6].

Long Short-Term Memory network (LSTM) is an essential enhancement of ANN, since it has the ability to deal with long-term dependencies and long-term time series and can effectively learn nonlinear relations and multiple characteristic factors, which allows it suitable for analyzing and forecasting stock data. Nafia et al. utilized LSTM based prediction to construct equity-market-neutral stock portfolio and confirmed that using LSTM model could effectively predict future prices and hedge market risk [7]. Usmani and Shamsi employed LSTM to predict stock prices by analyzing financial news [8].

At present, many scholars have conducted comparative studies on ARIMA model, RNN, and LSTM networks to test the prediction ability of these models. Moghar and Hamiche demonstrated that RNN and LSTM could both predict stock prices, but for the series with long-term memory, LSTM had a better performance under every condition [9]. Ho, Darman, and Musa conducted research on LSTM, Neural networks, and ARIMA model and shown that LSTM could generate more than 90% of accuracy [10].

The aim of this paper is testing the prediction ability of ARIMA model and LSTM networks and analyzing model performance based on the data of CSI 300 adjusted closing prices from 1/7/2022 to 26/7/2023 in which Chinese stock market had experienced a great turning point after the adjustment of COVID-19 prevention policy.

2. Methods

2.1. Data Source

When choosing the predicted data set, the data of a single stock may have subjective and uncertain bias. In order to test the application effect of ARIMA model and LSTM networks in the Chinese stock market, the breadth and comprehensiveness of the data should be considered.

The data for this study is sourced from the CSI 300 index, which represents a comprehensive snapshot of the Chinese stock market. The CSI 300 index is specifically chosen due to its pivotal role in reflecting the overall dynamics of the Chinese stock market. The index comprises 300 of the largest and most liquid stocks traded on the Shanghai and Shenzhen stock exchanges, providing a robust foundation for volatility prediction analysis.

The daily data of closing prices, adjusted closing prices, opening prices, highest prices, lowest prices, and the volume from July 2022 to June 2023 are chosen. Over this time period, Chinese stock market has experienced an enormous turning point due to the adjustment of the COVID-19 prevention policy. Constructing models with the data in this period can effectively test the prediction ability of different models against drastic changes. For ARIMA model, only adjusted closing prices are used to construct the model, but for LSTM networks, all sorts of data are put into the model as different features in order to obtain more exact results.

In this paper, the training set consists of the first 80% of the data (from July 2022 to May 2023), which is viewed as a complete cycle of fluctuations. The remaining data is used as a test set and compared with the predicted results.

2.2. Model Introduction

2.2.1 ARIMA model

ARIMA which stands for AutoRegressive Integrated Moving Average, is a widely used time series model to capture hidden information in data and predict future values. In an ARIMA(p , d , q) model, the parameter p represents the order of the AutoRegressive component which capture the linear relationship between the current value and its p historical data, and the parameter q represents that the Moving Average component extracts the linear relationship between the current value and past q white noise errors.

An important assumption of data used to construct ARIMA model is the stationarity, which means that the data cannot contain seasonality or any trends. The parameter d represents the order of difference required to make the series stationary.

Given X_i ($i = 1, 2, \dots, t - p$) representing the value of data at time i in a time series after the difference of d order, by an ARIMA (p, d, q) model, the value of X_t (with difference of d order) can be represented as below:

$$X_t = \sum_{i=1}^{t-p} \alpha_i X_i + \sum_{j=1}^{t-q} \beta_j \epsilon_j + \epsilon_t \quad (1)$$

2.2.2 LSTM networks

Recurrent Neural Networks (RNN) has been widely employed for time series with temporal dependencies, since an unfolded RNN are capable of processing current data by using previous data, which is similar to the principle of ARIMA model. Meanwhile, RNN has the shortage of training time series with a long-term dependency due to the neglect of past information. As an important variant and development of RNN, LSTM has the capability to remember long term information by saving information as memory during each iteration.

LSTM is constructed by three parts named forget gate, input gate, and output gate. The forget gate is able to determine which of information in the memory should be forget in the current state. The input gate is used to qualify the weighted importance of new information added into the model. The output gate can provide the result calculated by new information and previous information. All gates are defined by the sigmoid function demonstrated below:

$$F(t) = \text{sigmoid}(W_f X_t + W_{hf} H_{t-1} + b_f) \quad (2)$$

$$I(t) = \text{sigmoid}(W_i X_t + W_{hi} H_{t-1} + b_i) \quad (3)$$

$$O(t) = \text{sigmoid}(W_o X_t + W_{ho} H_{t-1} + b_o) \quad (4)$$

In the Eqs. (2), (3), and (4), $F(t)$, $I(t)$, $O(t)$ represent results given by forget gate, input gate, and output gate at state t . W_f , W_i , W_o , W_{hf} , W_{hi} , W_{ho} are the weights of three gates, and b_f , b_i , and b_o represents bias variables. X_t is new information at state t , and H_{t-1} is previous information at $t-1$.

The memory will be updated by results from forget gate and input gate expressed mathematically as in Eq. (5), in which C_t represent the memory at state t , and the $\otimes$ function is used to select which of new information is important and should be remembered. $\otimes$ is defined as element wise multiplier for elements in C_{t-1} and $F(t)$ or $I(t)$ and $\tanh$. After this process, some previous information in memory will be forget, and some new information will be added into memory. Meanwhile, the new result at state t will be calculated by $O(t)$ and C_{t-1} in Eq. (6).

$$C_t = C_{t-1} \otimes F(t) + I(t) \otimes \tanh(W_c X_t + W_{hc} H_{t-1} + b_c) \quad (5)$$

$$C_t = C_{t-1} \otimes F(t) + I(t) \otimes \tanh(W_c X_t + W_{hc} H_{t-1} + b_c) \quad (6)$$

2.3. Performance indices

To comprehensively assess the predictive accuracy of our models, we utilize two key error analysis metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These two functions are shown as follow:

$$MAE = \frac{1}{T} \sum |v_t - RV_t| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{T} \sum (v_t - RV_t)^2} \quad (8)$$

3. Results and Discussion

3.1. Result of ARIMA Model

The historical data of Adjusted Closing prices of CSI 300 is drawn as below (figure 1), and it can be found that the data is not stable and has a downward trend. Meanwhile, the autocorrelation diagram of data (figure 2) shows that the data are autocorrelated. The ARMA model requires the stationarity of data, so the original data needs to be further processed to eliminate the influence of the current trend.

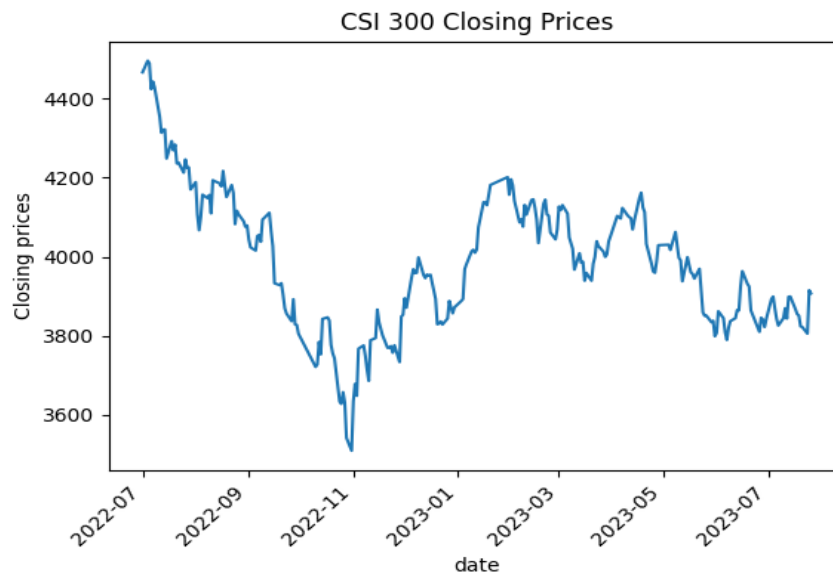


Fig. 1 Closing prices from 2022.07 to 2023.07

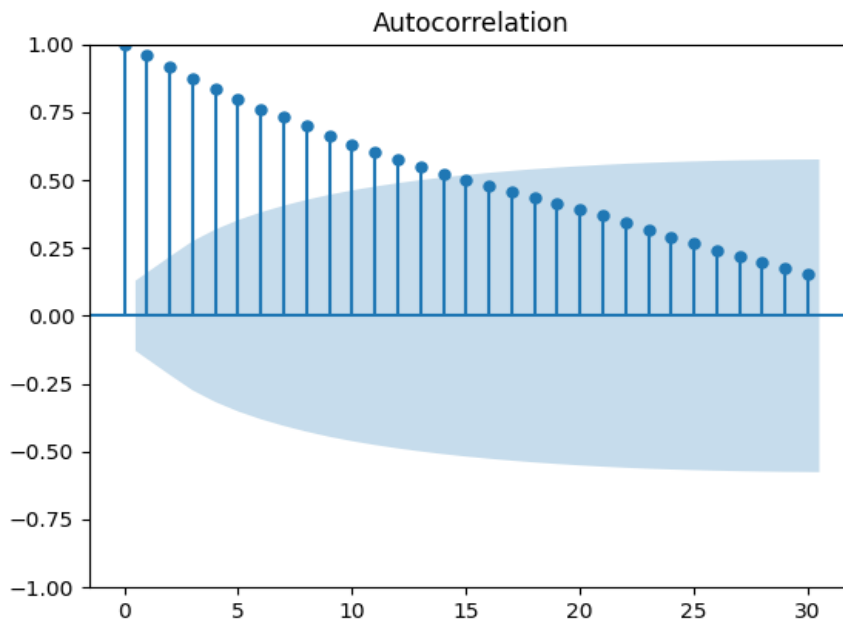


Fig. 2 Autocorrelation plot of closing prices

First of all, we try to carry out first-order difference processing on the data, and the line chart (figure 3) shows obvious stationarity. Besides, according to the autocorrelation graph of the first-order difference data, it can also be seen that the data is stationary data. Therefore, a time series model can be constructed for the first-order differential data of Adjusted Closing prices.

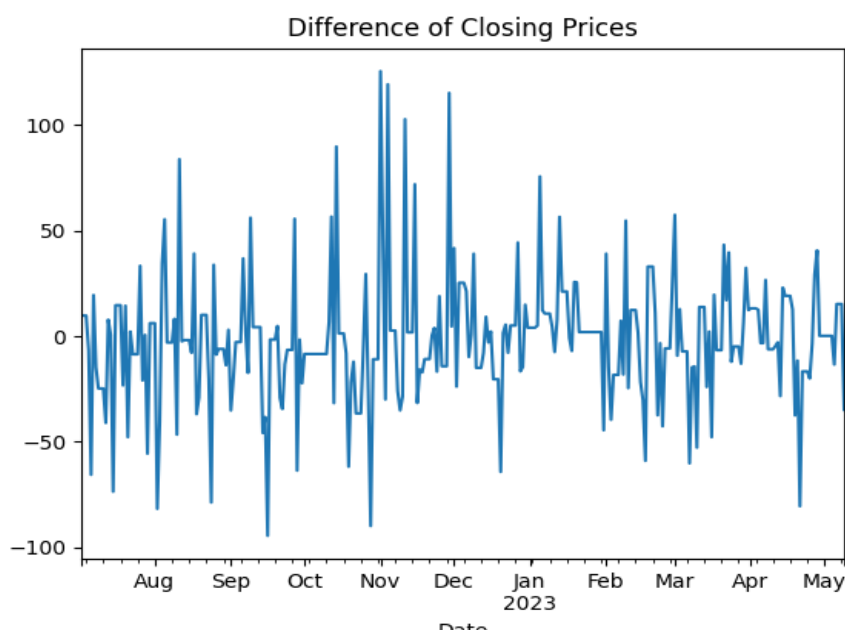


Fig. 3 Closing prices data after one order difference

In the division of training set and test set, this paper uses the first 80% data as training set data to train the ARIMA model, and the last 20% data as test set data to compare with the predicted value of the model. Firstly, an AR (1) model is constructed on the data and the result of model diagnosis (Table 1) shows that according to the z test, $P > 0.05$, that is, all coefficients cannot be guaranteed to be significant at the 5% confidence level, so the AR (1) model is overthrown.

Table 1. Model diagnosis of ARIMA (1,1,0)

Coefficients	Values	Standard Deviation	z	$P > z $
AR.L1	-0.0123	0.077	-0.159	0.872
Constant	1530.98	127.09	12.04	0.00

As this paper focuses on empirical analysis and comparison of model prediction effects, it does not carry out theoretical analysis on parameter selection of ARIMA model, but only uses grid search method to find the best model based on Akaike information. By using the grid search method, ARIMA (2,1,2) model has the highest AIC value. By using the ARIMA (2,1,2) to predict Closing prices data during more than 300 days, the model has been summarized below as Table 2. According to the model diagnosis, it can be seen that the P value of all parameters is less than 0.05, which can reject the null hypothesis and ensure the significance of all coefficients at the 5% confidence level. The model diagnosis passed. By drawing the QQ plot (figure 4), the model residuals basically follow normal distribution and can be regarded as white noise sequence, which means that there is no more information in the data that can be extracted by the model ARIMA (2,1,2).

Table 2. Model diagnosis of ARIMA (1,1,0)

Coefficients	Values	Standard Deviation	z	$P > z $
AR.L1	1.49	0.027	55.47	0.00
AR.L2	-0.96	0.023	-41.04	0.00
MA.L1	-1.47	0.024	-61.73	0.00
MA.L1	0.97	0.024	40.90	0.00
Constant	840.98	46.218	18.19	0.00

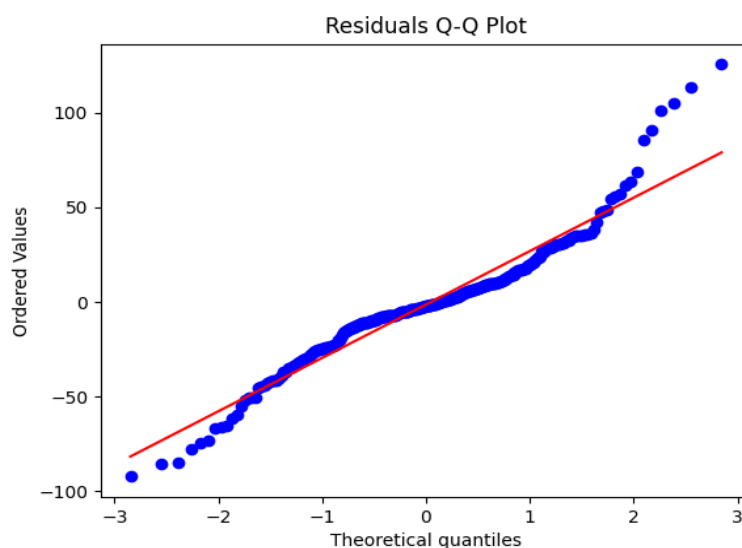


Fig. 4 Q-Q plot of residuals

By comparing the prediction results and real data, the average price of prediction results is higher than it of test data by about 3%, but the standard deviation of prediction results is far away lower than it of test data. The RMSE between prediction results and real prices is 145.06 and the MAE is 131.17.

3.2. Result of LSTM Networks

When LSTM networks were used to predict the trend of Adjusted Closing prices of CSI 300, the scale of the training set is the same as that of ARIMA model. Because LSTM networks is still supervised learning, the model training step size cannot adjust automatically but needs to be set manually. The LSTM networks used in this paper aim at learning short-term characteristics of stock prices to predict stock trends, so the training step should not be too long or too short, otherwise it is difficult to extract the main features of the model. According to the grid search method, 5 days are finally selected as the training step length of the model.

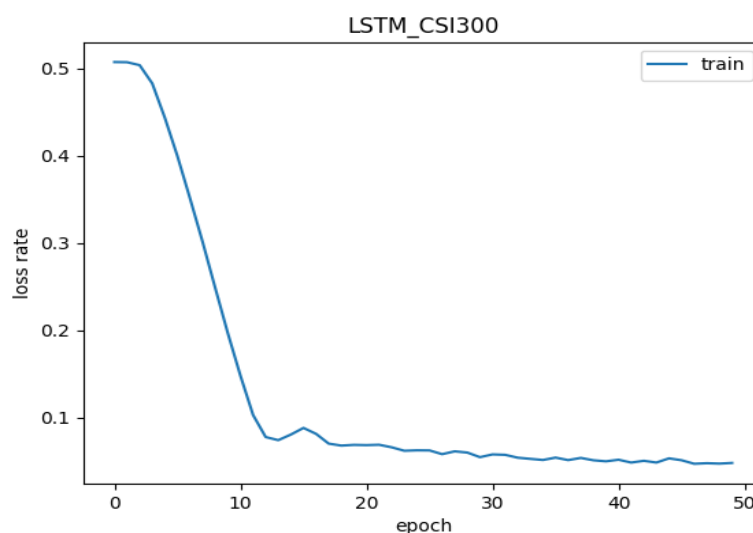


Fig. 5 Loss rate during training

LSTM networks have various parameters should be chosen manually, including the number of hidden neurons, activation function, optimizer, dropout ratio, and model learning rate. Different values of parameters have different effects on the model. The number of hidden neurons will directly affect the model prediction results. If the number of hidden neurons is too small, the data information cannot be completely extracted. Otherwise, it will lead to overfitting problems. By using grid search method, the number of neurons is 100, the dropout ratio is 0.2, and the model learning rate is equal

to 0.01. The activation function and optimizer are ReLU and Adam functions, respectively. In terms of model feature data, as mentioned in 2.1, this paper selects the Opening price, closing price, Adjusted Closing Prices, Highest prices, Lowest prices and Volume of CSI 300 as feature data and puts them into LSTM networks to forecast the Adjusted Closing prices in the future. Figure 5 and figure 6 demonstrate model training loss functions and prediction results respectively. According to the loss function, the training loss gradually approaches 0 after 10 times, which indicates that the model training is effective.

The average price of prediction results is 3893.18(CNY), and the average price of empirical data is 3871.63(CNY), where is only a 0.5% difference. The standard deviation of prediction results is 46.6, being closed to that of test data as 51.8. According to the figure 6, the LSTM model is capable of obviously capturing overall features of prices within each interval, though with a slight lag. The RMSE between prediction results and real prices is 58.4 and the MAE is 49.7.

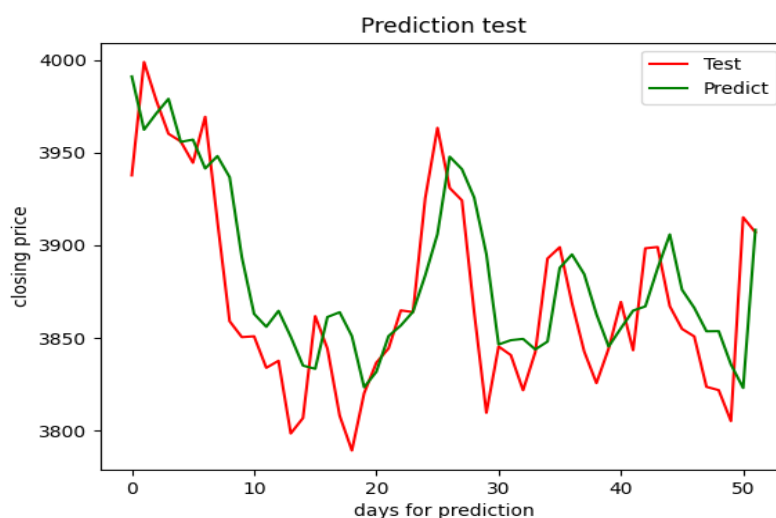


Fig. 6 Comparison between prediction results and real data

3.3. Discussion

The results shown above are able to illustrate the prediction ability of ARIMA model and LSTM networks. For the ARIMA model, although it has extracted all valid information in the data, the prediction results are still far away from the empirical data, which may be derived from the heteroscedasticity of the Closing prices of CSI 300. As introduced in sector 2.2, the key assumption of time series capable to be fitted by ARIMA model is that the time series data have the same variance during the whole time period. Through the ARCH effect test, the Adjusted Closing prices data of CSI 300 have obvious heteroscedasticity, explaining why the ARIMA model cannot predict the volatility observed in the test set. Therefore, for stock prices, the ARIMA model can be effective only within short interval, otherwise the difference of volatility for long term would be ignored by the model.

For the LSTM networks, by learning the change pattern, it can highly estimate the trend of stock prices, without considering the stationery and heteroscedasticity. However, the prediction results are strongly affected by the training step length used to construct the LSTM networks, which is manifested by the lag with the empirical data. Thus, LSTM networks may hardly capture the sudden change during a short period.

4. Conclusion

According to the research in this paper, the Adjusted closing prices of CSI 300 are heteroscedasticity time series with autocorrelation and long-term memory. Therefore, it can be analyzed and predicted using ARIMA model and LSTM networks.

Closing prices show an obvious downward trend in the time period, so when using the ARIMA model to forecast, first order difference should be performed on the data, and then the data after the difference should be modeled. According to the research, ARIMA (2,1,2) has a good performance, but due to the heteroscedasticity of the data, the model can only reflect the change trend of the model in the short term and reflects the long-term volatility of the price poorly. On the basis of this model, the GARCH family model should be further introduced to describe the volatility change of the model.

Compared with ARIMA model, LSTM networks show better performance. Thanks to the analysis of long-term memory information, it not only significantly reduces the error between the predicted value and the true value (RMSE and MAE), but also effectively reflects the future price fluctuation trend. However, the prediction results of LSTM are lagging and cannot reflect the drastic changes in the short term. In the application, the parameters in the algorithm should be manually adjusted for different stocks, so as to obtain better prediction results.

References

- [1] Box G E P, Jenkins G M, Reinsel G C. Time Series Analysis: Forecasting and Control, Fourth Edition. Wiley, 2008.
- [2] Zhang, Xiaoduo, Zhai Rui, and Gao Wen. Analysis and Economic Prediction of Stock Index Based on Index Tracking and ARIMA Mode. Journal of Physics: Conference Series, 2021.
- [3] Ahmar A S, et al. Prediction of BRIC stock price using ARIMA, SutteARIMA, and holt-winters. Computers, Materials & Continua, 2022, 70(1): 523–534.
- [4] Narendra Babu C, Eswara Reddy B. Prediction of selected Indian stock using a partitioning–interpolation based ARIMA–GARCH model. Applied Computing & Informatics, 2015, 11(2): 130–143.
- [5] Hambly B, Xu R, Yang, H. Recent advances in reinforcement learning in finance. Mathematical Finance, 2023, 33(3): 437–503.
- [6] Kristjanpoller W, Minutolo M C. Gold price volatility: A forecasting approach using the Artificial Neural Network–GARCH model. Expert Systems with Applications, 2015, 42(20): 7245–7251.
- [7] Nafia A, Yousfi A, Echaoui A. Equity-Market-Neutral Strategy Portfolio Construction Using LSTM-Based Stock Prediction and Selection: An Application to S&P500 Consumer Staples Stocks. International Journal of Financial Studies, 2022, 11(2).
- [8] Usmani S, Shamsi J A. LSTM based stock prediction using weighted and categorized financial news. PloS One, 2023, 18(3).
- [9] Moghar A, Hamiche M. Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 2020, 170: 1168–1173.
- [10] Ho M K, Darman H, Musa S. Stock Price Prediction Using ARIMA, Neural Network and LSTM Models. Journal of Physics: Conference Series, 2021.