Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/14/20230781

Prediction of stock price leveraging Arima model

Yeming Lin

School of Advanced Manufacturing, Nanchang University, Nanchang, Jiang Xi, 330031, China

5905122031@email.ncu.edu.cn

Abstract. In today`s world, more and more machine learning methods and prediction methods have been proposed. There are also lots of investors who choose the more difficult and complex models to use as technical analysis tools to predict stock prices. However, Rome wasn't built in a day. So, it`s necessary to learn the traditional models penetratingly. To have a more thorough understanding of classic models in the past, this paper provides a great way to analyze the ARIMA model. It explains the methods of fitting the ARIMA model step by step and improves the ARIMA model by adding seasonal parameters to fit the SARIMA model, enabling readers to better understand the advantages and disadvantages of this model. Although the results of the ARIMA model are unfortunately similar, and difficult to demonstrate its predictive power in images, the SARIMA model presents a trend prediction that conforms to people's imagine. As of today, the change in the price of Netflix's stock has been visible, which is different from the predicted price. Therefore, while making technical predictions, investors also need to combine fundamental analyses like the supply and demand balance, changes in interest rates, government regulations, macroeconomic indicators, and unique features of the industry in question, among others. Rather than relying solely on a model.

Keywords: ARIMA, machine learning, stock price prediction.

1. Introduction

Several problems with predicting in the real world relate to a temporal dimension and it`s always necessary to estimate digital sequence data called time series forecasting [1]. Time series forecasting plays a crucial role in various fields of research, and its importance cannot be overstated [2]. With the continuous advancement of modern market economies, an increasing number of individuals and organizations are showing a heightened level of financial management awareness and expertise. As the central pillar of modern economies, finance is gradually emerging as a crucial and extensively discussed topic of interest [3]. As an investment vehicle that is associated with high risk and high reward, an increasing number of individuals are beginning to predict stock prices to minimize risk. In the financial market, the fundamental goal of investing is to generate a higher return on investment capital [4]. By employing effective forecasting methods and managing risks appropriately, it is possible to achieve both reasonable and accurate predictions of stock prices and to generate remarkable returns [5]. Many investors heavily depend on time series forecasting models to inform their investment decisions [6].

The stock price is affected by a quantity of complex and interrelated elements including the supply and demand balance, changes in interest rates, government regulations, macroeconomic indicators, and unique features of the industry in question, among others [7]. The existence of anomalies can

DOI: 10.54254/2755-2721/14/20230781

significantly affect the precision of stock price predictions [8]. Forecasting stock price movements has been a challenging task due to the presence of a low signal-to-noise ratio and the ever-changing nature of the stock market [9]. At present, a variety of methods have been proposed for predicting stock prices. These can be broadly classified into two categories based on their underlying modeling theories: traditional models that rely on statistical theory (such as time series models), and machine learning models that leverage techniques such as neural networks and gray [10]. However, the models for predicting have become more and more complex and hard to explain, making investors hard to userbases the technology of machine learning and computing power to develop rapidly [11]. So, it`s important to lay a good foundation to learn them. ARIMA, as one of the most classical models, is still of great learning value.

2. Method

2.1. Dataset

Between the period from February 5, 2018, to February 4, 2022, Netflix's stock changed 1009 times, and only the closing price is analyzed here. The data set comes from the sharing of others on the Kaggle website.

2.2. Overview of the ARIMA Model

ARIMA models are a powerful tool for analyzing and predicting time series data that exhibits patterns over time. They are a combination of autoregression (AR), moving average (MA), and difference (I) models, which allows them to capture complex patterns in the data. By using ARIMA models, analysts can make forecasts based on historical trends and patterns, helping them to make informed decisions about future outcomes.

There are three parameters used to define the ARIMA models: p, d, and q.

p: Autoregressive order, indicating how many past observations are used for predicting the current value (AR component).

d: Degree of difference, indicating how many times the data is differenced to achieve stationarity, where statistical properties remain constant over time (difference component).

q: Moving average order, indicating how many past forecast errors are used to predict the current value (MA component).

The general equation for an ARIMA model is:

$$Yt = \mu + \varphi 1(Yt-1-\mu)+\ldots+\varphi p(Yt-p-\mu)+\varepsilon t+\theta 1\varepsilon t-1+\ldots+\theta q\varepsilon t-q \quad (1)$$

2.3. Mathematical modeling of ARIMA

A time series records the historical behavior of a system, where the individual data points may not be distinguishable, but their collective values exhibit a discernible pattern. During the 1970s, statisticians Box and Jenkins introduced the Autoregressive Integrated Moving Average (ARIMA) model [12]. The classical form of the ARIMA model is as follows:

$$\begin{cases} X_t = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad \emptyset_p \neq 0, \theta_q \neq 0, \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0 \quad s \neq t, \\ E(x_s \varepsilon_t) = 0 \quad s < t, \end{cases} \quad (2)$$

where $\emptyset_k(1 \leq k \leq p)$ and $\emptyset_k(1 \leq k \leq q)$ real parameters and in this scenario, there is a purely random time series that satisfies the stationarity conditions. However, for time series analysis using ARIMA, it is essential to transform a non-stationary series, such as stock prices, into a stationary series.

2.4. Modeling process of ARIMA

The modeling process of ARIMA is demonstrated in Figure 1. Initially, it is essential to execute a stability test on the stock price. If the data fails this test, it must undergo a differential operation until it

DOI: 10.54254/2755-2721/14/20230781

passes, and only then can it proceed to the next stage of White Noise Verification. If it fails the latter, it becomes necessary to fit an ARIMA model to estimate the p/q values. Conversely, if it passes the White Noise Verification, it can progress to the final stage of model fitting. Following the completion of model fitting and training, the final predictions can be made.

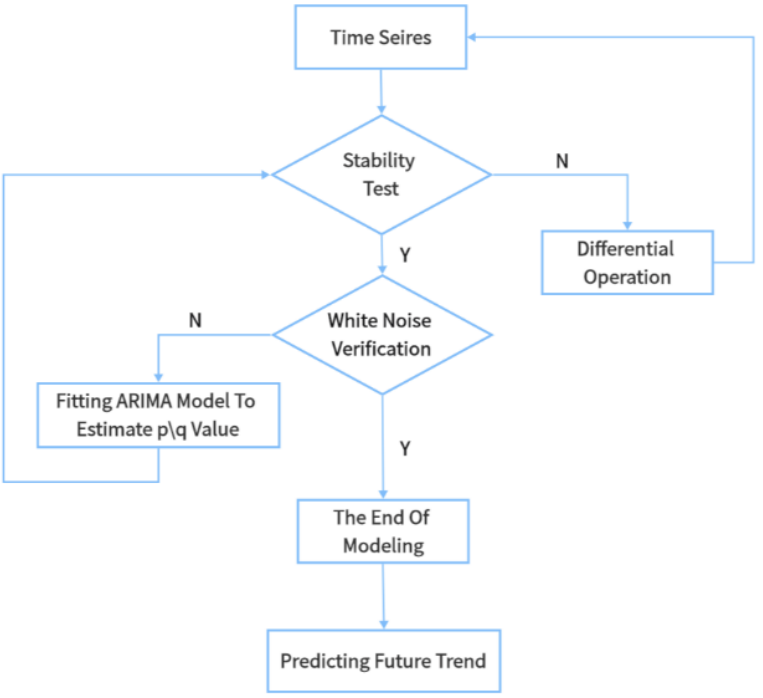

Figure 1. Process of modeling.

2.4.1. Stability test. It is evident from Figure 2 that the stock price exhibits insufficient stability, thus implying that it may be a non-stationary series. To verify this initial assumption, this work utilized the Augmented Dickey-Fuller (ADF) test, which yielded a p-value greater than 0.05, thereby supporting the non-stationarity of the series.

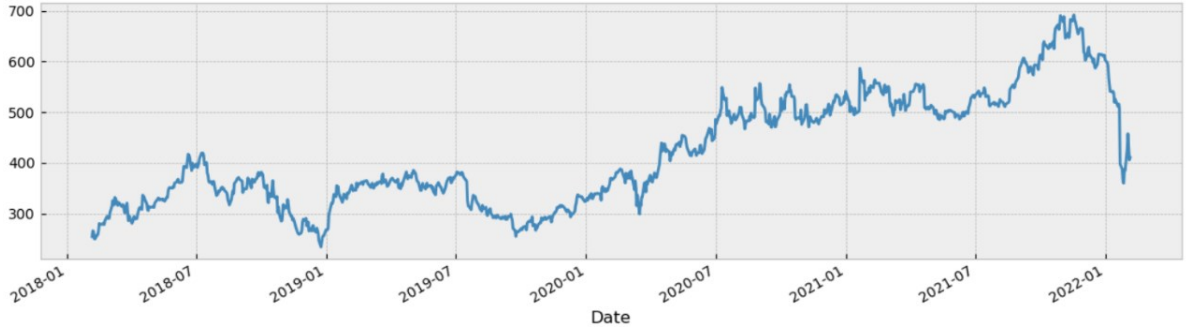

Figure 2. Stock price demonstration.

2.4.2. First-order difference. Figure 3 displays the relatively stable image obtained after performing the first-order difference. Nevertheless, it is essential to note that the ADF test results remain a crucial component of the analysis. Following the differential operation, the ADF value was found to be less than 0.05, indicating that the series has become stationary. Moreover, as the stationary sequence was attained through only one differential operation, the value of d can be established as 1.

DOI: 10.54254/2755-2721/14/20230781

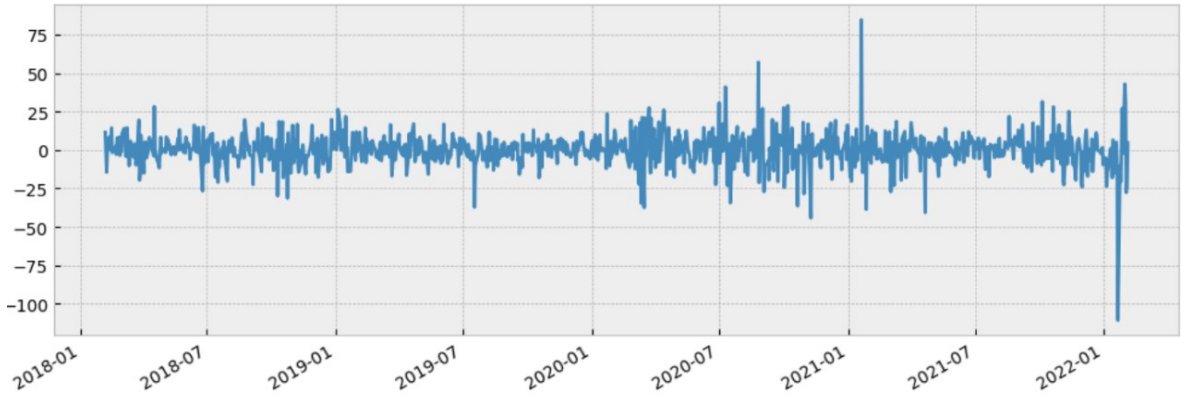

Figure 3. Stock price after conducting first-order difference.

$$\text{When d=1,}(1-B)Y_t = \nabla Y_t = Y_t - Y_{t-1}$$
$$\therefore \begin{array}{l}(Y_t - Y_{t-1}) - \sum_{i=1}^{p} \emptyset_i (Y_{i-1} - Y_{(t-1)-i}) = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-1} \Leftrightarrow \\ (Y_t - Y_{t-1}) - \emptyset_1 (Y_{t-1} - Y_{(t-2)}) - \emptyset_2 (Y_{(t-2)} - Y_{(t-3)}) - \cdots \backslash \\ -\emptyset_p (Y_{t-p} - Y_{(t-1)-p}) = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}\end{array} \quad (3)$$

It displays the ARIMA formula for a time series when the value of d is set to 1.

2.4.3. White noise verification. To satisfy the prerequisite conditions for further research, the original time series must exhibit white noise characteristics, while the first-order differential series must not exhibit such characteristics. The white noise verification test was conducted on the original series, resulting in a p-value greater than 0.05, indicating that the original series satisfies the white noise criterion. Similarly, the white noise test was carried out on the first-order differential series, resulting in a p-value very close to 0, significantly lower than the threshold of 0.05, indicating that the differential series meets the non-white noise criterion. It is only when these conditions are met that the data can be considered suitable for further research.

2.5. Fitting ARIMA model

Once the time series data set has been confirmed to meet the stationarity and white noise criteria, the next step is to determine the appropriate values for p and q in the ARIMA model and to fit the model to the data.

2.5.1. Determine p/q value. The ACF decides the value of q. Figure 4 is the picture of ACF which was made after the first-order difference with lags=30. So, q=1. Similarly, the PACF decides the value of p. Figure 5 is the picture of PACF which was made after the first-order difference with lags=30. So, q=1.

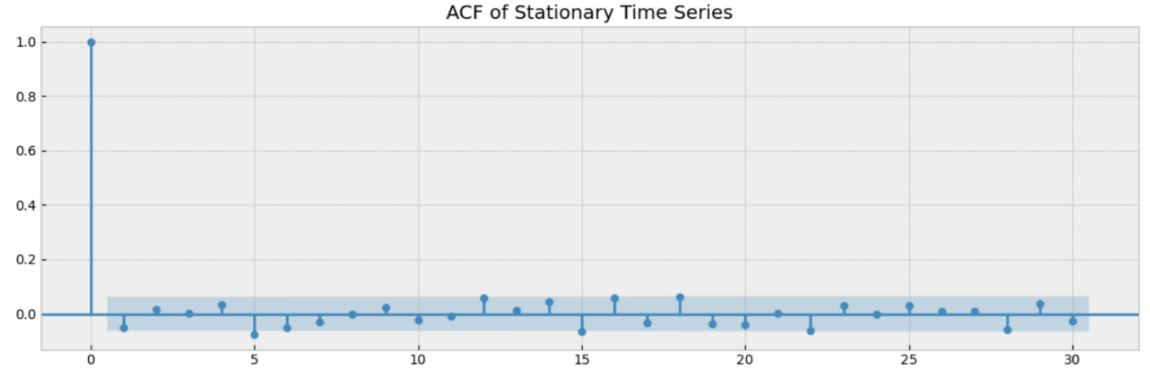

Figure 4. ACF result.

DOI: 10.54254/2755-2721/14/20230781

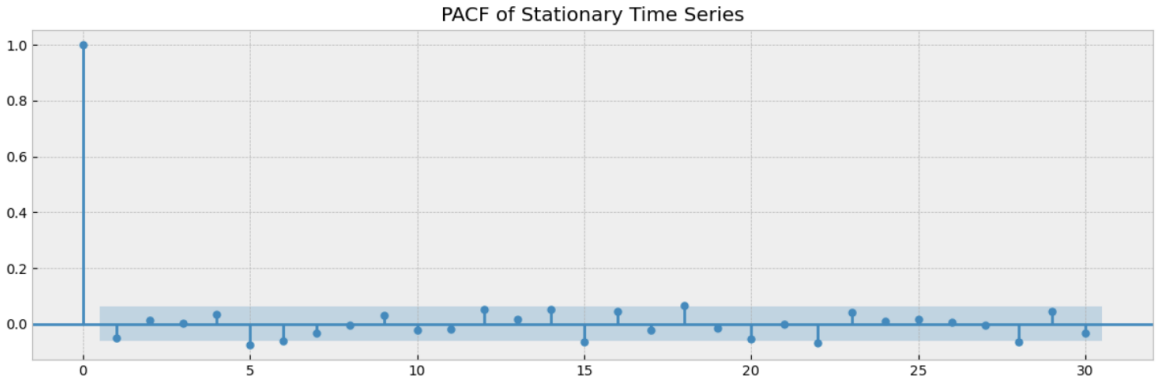

Figure 5. PACF result.

2.5.2. Fitting. After a rigorous process of testing, selection, and fitting of the p and q values, the design was able to successfully fit an ARIMA model to the dataset. The model is presented in Table 1, and its quality was assessed using the AIC and BIC criteria. The results indicated that the model performed adequately in forecasting stock prices, suggesting its potential usefulness for future analyses.

Table 1. ARIMA model setting.

Dep. Variable	Close	No.Observations	1008
Model	ARIMA(1,1,1)	Log Likelihood	-3850.678
Ljung-Box(L1)(Q)	0.00	AIC	7707.356
Covariance Type	opg	BIC	7722.100

3. Result

3.1. Performance of ARIMA model

After the aforementioned procedures of tests, parameter selection, and model fitting, it is imperative to perform an analysis of the fitted model to validate its accuracy.

The data set was separated into a training set (80%) and a test set (20%) to verify the accuracy of this model. The results in Figure 6 show that the predicted values (orange) closely match the test set data, indicating that the model has the good predictive ability.

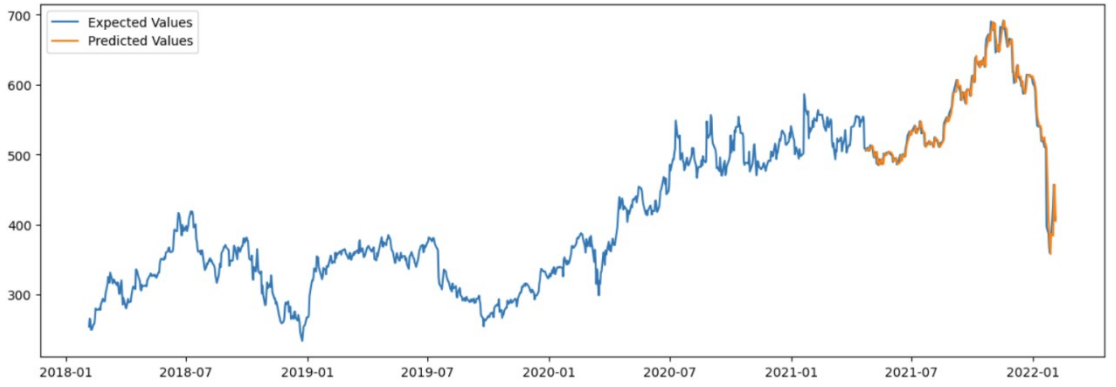

Figure 6. Visualization result of prediction.

To prove the availability of the model, it should choose the Mean Square Error (MSE) and Mean Relative Error (MRE) to be the target. These formulas are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (4)$$

DOI: 10.54254/2755-2721/14/20230781

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \quad (5)$$

The results of the fitted model are displayed in Table 2. This result is acceptable for a stock forecast model. By comparing the MSE and MRE values of the model with those of other models or the naive prediction, the superiority of the ARIMA model in predicting stock prices could be observed.

Table 2. Quantitative performance.

MAE	8.07924228400611
RMSE	13.09631326458214

The data for the next 60 days is forecasted using the model and the corresponding plot is shown in Figure 7. The predicted values are shown in orange.

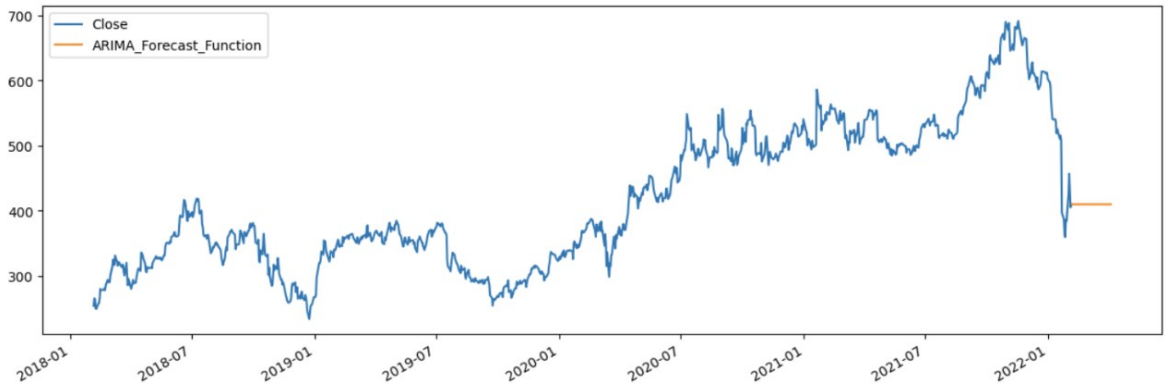

Figure 7. Prediction result of the next 60 days.

3.2. Performance of the improved model

The model after adding the parameters of seasonal to the ARIMA model is called the SARIMA model. Most stock prices are seasonally priced on an annual basis. Similarly, after testing, seasonal can be determined to be optimal at 12. As observed in Figure 8 the stock price data did not show significant changes, which could be due to the limited data volume or the inability of the model to directly handle time series with seasonal components. To address this issue, a seasonal ARIMA (SARIMA) model was employed for further improvement. After selecting the appropriate seasonal parameters, the resulting model was fitted and the results are presented.

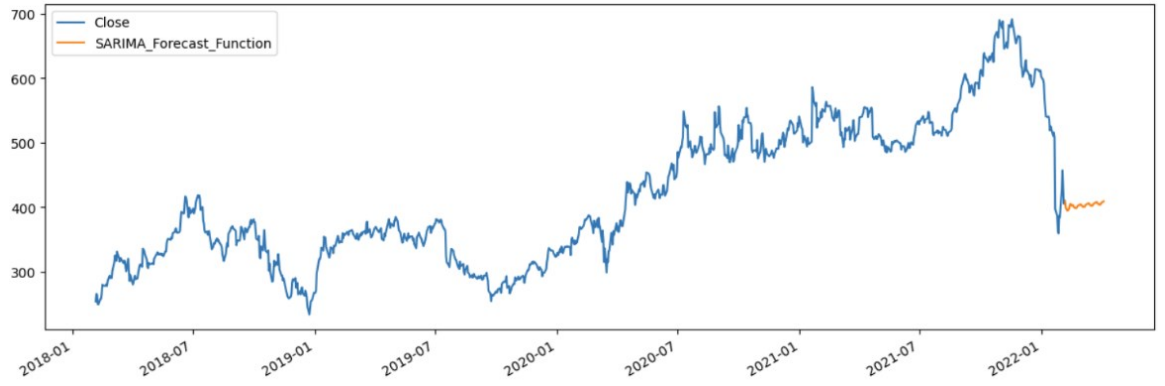

Figure 8. The prediction result in the SARIMA model.

DOI: 10.54254/2755-2721/14/20230781

4. Discussion

The seasonally adjusted data were fitted using the SARIMA model. As shown in Figure 8 a clear pattern of fluctuations has emerged. It must be stable data when the ARIMA model is used to predict time series data. If the data of stock price is unstable, it is impossible to grasp the rules well. While stable data is not seasonal. The stock prices are seasonal, so SARIMA`s result looks better. However, when considering the overall stock price over this period, the observed price changes do not appear to align with the fluctuations. Nevertheless, the overall trend is in line with the predicted direction. Given that stock prices are subject to changes in a company's fundamental factors, this observation is understandable.

It is plausible that, as the company's fundamentals recover over the year, the stock prices will eventually approach the previously predicted values. In addition to the model, other factors such as news events, market sentiment, and economic indicators can also influence stock prices. Therefore, it is important to consider these factors when forecasting stock prices. Incorporating these external variables into the model could potentially improve the accuracy of the predictions. Furthermore, the predictive power of the model may vary depending on the time frame of the analysis.

5. Conclusion

The process and basic formulas of the ARIMA model are well explained in the paper. Took the stock price of Netflix as an example, and fit the model by the dataset step by step. Although the results obtained are not satisfactory, they can also reflect the characteristics of the ARIMA model. For such data, the ARIMA model may not be very suitable, However, research such as this can lay a foundation for exploring the combination of various models and ARIMA models to improve the predictive ability of other models. Combining ARIMA and SARIMA models with currently popular models is a promising approach to time series forecasting. This could involve using machine learning models such as random forests, gradient boosting, or deep learning architectures like recurrent neural networks or transformers. These models have shown significant improvements in forecasting accuracy over traditional time series models in recent years.

References

[1] Khaldi, R., El Afia, A., Chiheb, R., & Tabik, S. (2023). What is the best RNN-cell structure to forecast each time series behavior?. Expert Systems with Applications, 215, 119140.

[2] Garai, S., & Paul, R. K. (2023). Development of MCS based-ensemble models using CEEMDAN decomposition and machine intelligence. Intelligent Systems with Applications, 18, 200202.

[3] Lin, S., & Feng, Y. (2022) Research on Stock Price Prediction Based on Orthogonal Gaussian Basis Function Expansion and Pearson Correlation Coefficient Weighted LSTM Neural Network, Advances in Computer, 5(6), 23-30.

[4] Chaudhari, K., & Thakkar, A. (2023). Neural network systems with an integrated coefficient of variation-based feature selection for stock price and trend prediction. Expert Systems with Applications, 119527.

[5] Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. International Journal of Crowd Science, 5(1), 55-72.

[6] Xing, J., & Li, Y. (2022). An Optimization Framework for Stock Price Prediction Based on Statistical Information and Recursive Model Average--Taking ARIMA Model as an Example. Cloud and Service-Oriented Computing, 2(1), 21-27.

[7] Wang, M. X., Xiao, Z., Peng, H. G., Wang, X. K., & Wang, J. Q. (2022). Stock price prediction for new energy vehicle enterprises: An integrated method based on time series and cloud models. Expert Systems with Applications, 208, 118125.

[8] Naidoo, V., & Du, S. (2022). A Deep Learning Method for the Detection and Compensation of Outlier Events in Stock Data. Electronics, 11(21), 3465.

[9] Yang, J., Zhang, W., Zhang, X., Zhou, J., & Zhang, P. (2023). Enhancing stock movement prediction with market index and curriculum learning. Expert Systems with Applications, 213,

DOI: 10.54254/2755-2721/14/20230781

118800.

[10] Chen, Q., Ma, S., & Yang, R. (2022) The effectiveness of stock prediction models: evidence from time series analysis and machine learning scenarios. 2022 5th International Conference on Financial Management, Education and Social Science, 519-526.

[11] Chen, J., Wen, Y., Nanehkaran, Y. A., Suzauddola, M. D., Chen, W., & Zhang, D. (2023). Machine learning techniques for stock price prediction and graphic signal recognition. Engineering Applications of Artificial Intelligence, 121, 106038.

[12] Vieira, A., Sousa, I., & Dória-Nóbrega, S. (2023). Forecasting daily admissions to an emergency department considering single and multiple seasonal patterns. Healthcare Analytics, 3, 100146.