

A comparative research of portfolio return prediction based on the ARIMA and LSTM models

Ziteng Liu*

Department of French, Electronic Science and Technology University, Chengdu, China

*Corresponding author: 2019170901022@std.uestc.edu.cn

Abstract. The ongoing development of deep learning or machine learning techniques makes time series prediction more precise. This technology has also achieved remarkable success in predicting the return of portfolio in the financial field, which means that the quantitative investment method continues to progress and shows strong applicability, and investors can obtain excess benefits in the market. This study seeks to anticipate portfolio return using deep learning and machine learning and compare and analyze the differences between them. This paper selects the stock data of six representative companies in various fields and calculates the maximum sharp ratio portfolio and then forecasts the return of the portfolio with ARIMA model and LSTM model respectively. The result indicates that, firstly, ARIMA model and LSTM model can be well applied to predict the future return of portfolio; Secondly, ARIMA model performs better in short-term prediction and stable data prediction. When it comes to long-term prediction, the LSTM model performs better. The research results may be useful for investors to choose appropriate time series models for portfolio prediction and analysis.

Keywords: ARIMA model; LSTM model; portfolio.

1. Introduction

In recent decade, with the fast progress of financial market and the increasingly close relationship with the world economy, it has brought more and greater opportunities and challenges to investors. With the change of the external political and economic environment and the impact of the COVID-19, the investment environment is becoming more and more complex. Nowadays, in the influence of the breakthrough and development of related technologies in the computer domain, deep learning and machine learning models have also achieved great success in time series prediction [1]. This makes the quantitative investment method begin to rise and show strong applicability, which can obtain huge excess return in the market [2].

In fact, numerous investigations have been done regarding this interesting field. For example, Centeno and Jackson only researched the cumulative return, daily return and the maximum sharp ratio of the best venture capital portfolio based on ARIMA model [2-3]. Wang [4] and Sen, Dutta and Mehtab [5] only studied the advantages of LSTM model in deep learning of time series prediction in the financial field based on the data such as cumulative return and sharp rate. The above studies only involve one aspect of machine learning or deep learning. In addition, Ma, Han and Wang [6] and Chen, Zhang and Mehlawat [7] studied the mixed application of LSTM model and mean-variance model in predicting portfolio returns. And Senneset and Gultvedt [8] and Choi [9] has researched the advantages of ARIMA-LSTM hybrid model in predicting portfolio returns, but the study of them rarely involves the comparison of ARIMA and LSTM. In the research of Siami-Namini, Tavakoli and Namin, they evaluated and compared the ARIMA and LSTM models and came to the conclusion that the LSTM model is superior to the ARIMA model [10].

However, the aim of this paper is to study the comparison between the ARIMA and LSTM models in forecasting the maximum sharp ratio portfolio return, and under what conditions the two models have better forecast results. The ARIMA and LSTM models which are used in this work are used to build stock forecasting models respectively. Through data preprocessing, calculation and analysis, visualization and other steps, two time series forecast models are trained to predict future portfolio returns, and to carry out a comparison. As a result, the portfolio may be accurately predicted by the ARIMA and LSTM models. It is noteworthy that ARIMA model can predict the returns of stable

portfolio data in a short time, while LSTM model can predict the multi-dimensional portfolio returns for a long time. The research of this paper has certain reference significance for investors to use deep learning and machine learning techniques to study the stock market and build a portfolio with good return performance.

This essay's remaining sections are divided into different sections. Section 2 is the main part of this paper including data, methods and results. Section 3 shows the conclusion.

2. Main Text

2.1 Data

This paper selects data including five representative companies: Apple, Amazon, Microsoft, Netflix and Tesla from Yahoofinance (<https://finance.yahoo.com>). Apple, Microsoft and Amazon are all among the world's top five technology companies. They have outstanding positions in mobile computers, e-commerce and software services, and their market capitalization has reached \$1 trillion. In addition, Netflix and Tesla are the top companies in the streaming media and new energy vehicle industries respectively. The selected data is between July 5, 2013, and July 5, 2022. Through the preliminary statistical analysis of the data in these 10 years, a total of 2265 data are collected, whose statistical data are showed in the table 1.

Through visualizing the data, TSLA has the highest mean value with 0.0021, while mean values of AMZN and MSFT are the lowest with 0.0011. Besides, TSLA also has the largest max value with 0.1990 and NFLX has the smallest min value with -0.3512.

Table 1. Descriptive statistics of data

	AAPL	AMZN	MSFT	NFLX	TSLA
Mean	0.0012	0.0011	0.0011	0.0012	0.0021
Max	0.1198	0.1413	0.1422	0.1903	0.1990
Min	-0.1286	-0.1405	-0.1473	-0.3512	-0.2106

Figure 1 shows the range of returns of each stock, and some outliers which will affect the analysis should be cleared. In addition, figure 2 shows that the cumulative returns of stocks is also different which can be observed on the y-axis.

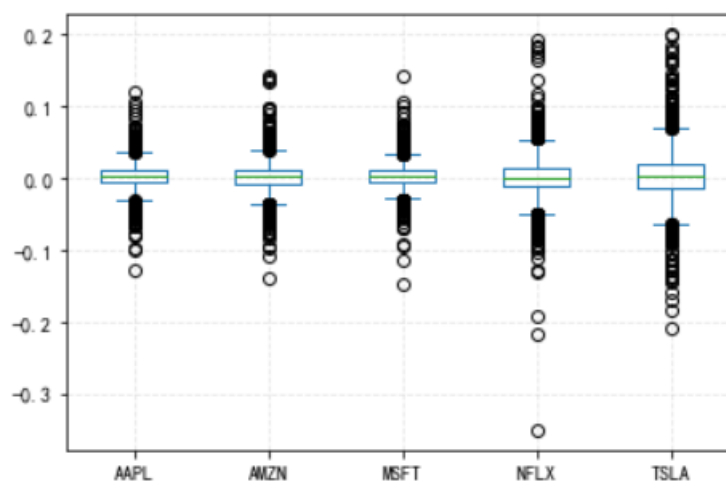


Fig. 1 The box plot of data

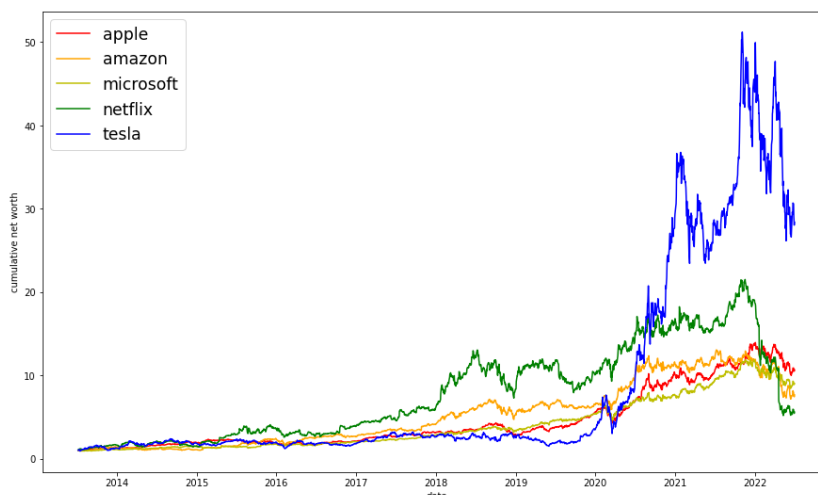


Fig. 2 The cumulative returns of five stocks

2.2 Method

2.2.1 Sharp ratio

If the Sharp ratio is less than one, the fund's operational risk exceeds its rate of return; if it is more than one, the fund's return exceeds its volatility risk. The Sharp ratio demonstrates that for every additional risk point that investor take on, they may receive an excess return. This allows for the calculation of the Sharp ratio, or the ratio of investment return to risk-taking, for each portfolio. The ratio which is higher represents that the portfolio will be better. The formula is shown as:

$$\text{Sharp Ratio} = \frac{E(R_p) - R_f}{\sigma_p} \quad (1)$$

Where $E(R_p)$ is the anticipated annualized yield on a financial portfolio, R_f indicates the annualized interest rate with no risk and σ_p denotes the annualized return standard deviation.

2.2.2 ARIMA

The article employs an autoregressive integrated moving average (ARIMA) model. "AR" means for "autoregressive," "p" stands for the number of autoregressive terms, "Ma" represents for moving average, "q" stands for the number of moving average terms, and I stands for the number of differences orders required to make the sequence stationary in the formula ARIMA (p, i, q).

ARIMA model requires the data to be stable. From a statistical point of view, data distribution is stable, which means it does not vary as time passes. Therefore, non-stationary data shows fluctuations due to trends, which must be converted before analysis.

In ARIMA model, 'AR' is an autoregressive model that firstly needs to determine a value of the order p, which indicates that the historical value across a number of time periods predicts the current value. The p-order autoregressive model's formula is as follows:

$$y_t = \mu + \sum_{i=1}^p \gamma_i y_{t-i} + \epsilon_t \quad (2)$$

Where ϵ_t is the error, p is the order, γ_i is the autocorrelation coefficient, μ is the constant and y_t is the current value. The autoregressive model's buildup of error components is highlighted by the moving average model "Ma." The q-order autoregressive process's formula is as follows:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^q \theta_i \epsilon_{t-i} \quad (3)$$

In ARIMA model, 'I' is differential. For non-stationary series, the data may be made more stable by using the difference approach. The first-order difference technique is the most used approach. The formula for first order difference is:

$$\Delta y = y_t - y_{t-1} \quad (4)$$

2.2.3 LSTM model

LSTM algorithm is an important time series algorithm which is used most at present. Long-term dependencies can be learned via a specific type of RNN (recurrent neural network). Its major objective is to address the gradient vanishing and explosion problems that arise during extensive timeline training. Shortly said, LSTM outperforms regular RNN in lengthier sequences.

In addition, LSTM adds three neural network layers: "forgetting gate", "memory gate" and "output gate". The forgetting gate can decide which information to discard and output a value from 0 to 1. 1 means fully reserved, while 0 means fully discarded.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

where f_t is forgetting gate. Memory gates can determine what information to remember. The closer the output value is to 1, the more information will be remembered.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (7)$$

where i_t is memory gate, c is memory cell and C_t is cell state. The output result of the output gate combines the status information with the current memory information.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t * \tanh(C_t) \quad (9)$$

Where o_t is output gate and h_t is hidden state.

2.3 Results

2.3.1 Portfolio optimization

Through the figure 3, Tesla has the highest variance among all stocks, followed by Netflix. The one with the least variance is MSFT. In addition, AAPL and AMZN had the lowest covariance, followed by that of AAPL and NFLX.

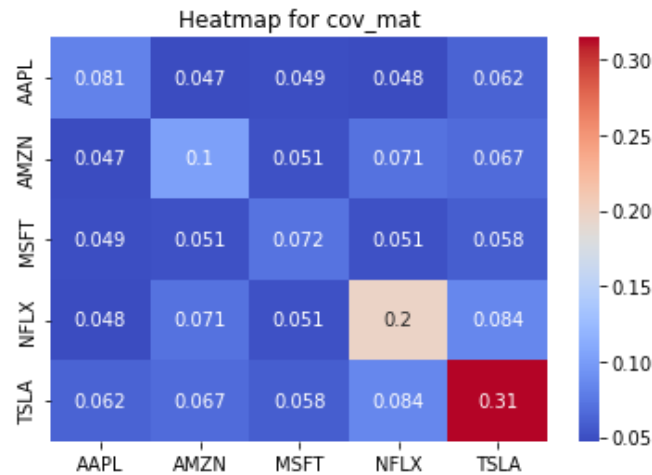


Fig. 3 The heat map of covariance matrix

Next, mean-variance investment portfolio can be constructed according to the 5 stocks screened out. The vector of return rate is $X = (X_1, \dots, X_5)^T$, the covariance matrix is

$$\Sigma_5 = (v_{ij})_{5 \times 5}, \quad i, j = 1, 2, \dots, 5 \quad (10)$$

Let $P_5 = \sum_{i=1}^5 \omega_i X_i$ represent the return rate of the portfolio composed of five risky assets. According to Markowitz theory, the linear programming model is

$$\min w^T \Sigma_5 w \quad (11)$$

$$s. t. \quad r_p = w^T \mu, \quad (12)$$

$$w^T I = 1 \quad (13)$$

Where $r_p = E(P_5)$, $\mu = E(X)$, I is the 5×1 unit column vector. The efficient frontier in the figure 4 can help analyze the maximum sharp ratio portfolio. The horizontal axis in the figure 3 represents the volatility of the portfolio, and the vertical axis represents the expected return of the portfolio. According to the legend on the right, the point which is greener means that the ratio of expected return to volatility that is sharp ratio is greater. Then, through the table 2, the maximum sharp ratio is 127.46%. And the weight are shown in table 3, where AAPL has the largest weight with 37.42%, while NFLX has the smallest weight with 1.78%.

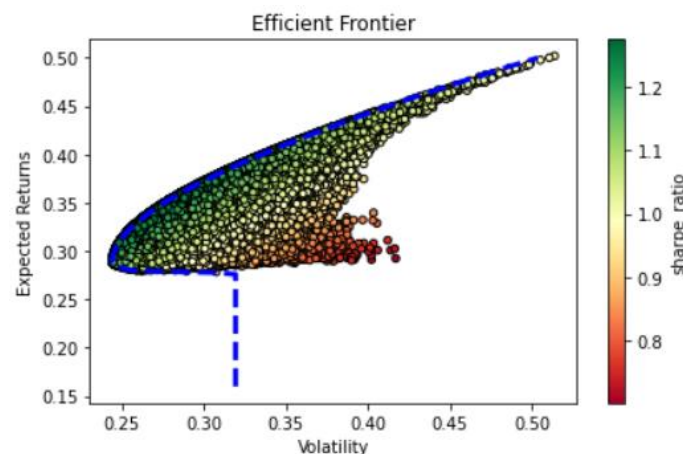


Fig. 4 The efficient frontier of five stocks

Table 2. The performance of Maximum Sharp Ratio portfolio

	Return	Volatility	Sharpe Ratio
Performance	33.72%	26.46%	127.46%

Table 3. Weights of Maximum Sharp Ratio portfolio

	AAPL	AMZN	MSFT	NFLX	TSLA
Weights	37.42%	9.24%	32.36%	1.78%	19.20%

The returns of the maximum sharp ratio portfolio are shown in a time series diagram in Figure 5. ARIMA model requires data to be stable so that it is necessary to use Dickey-Fuller test to check the data stability.

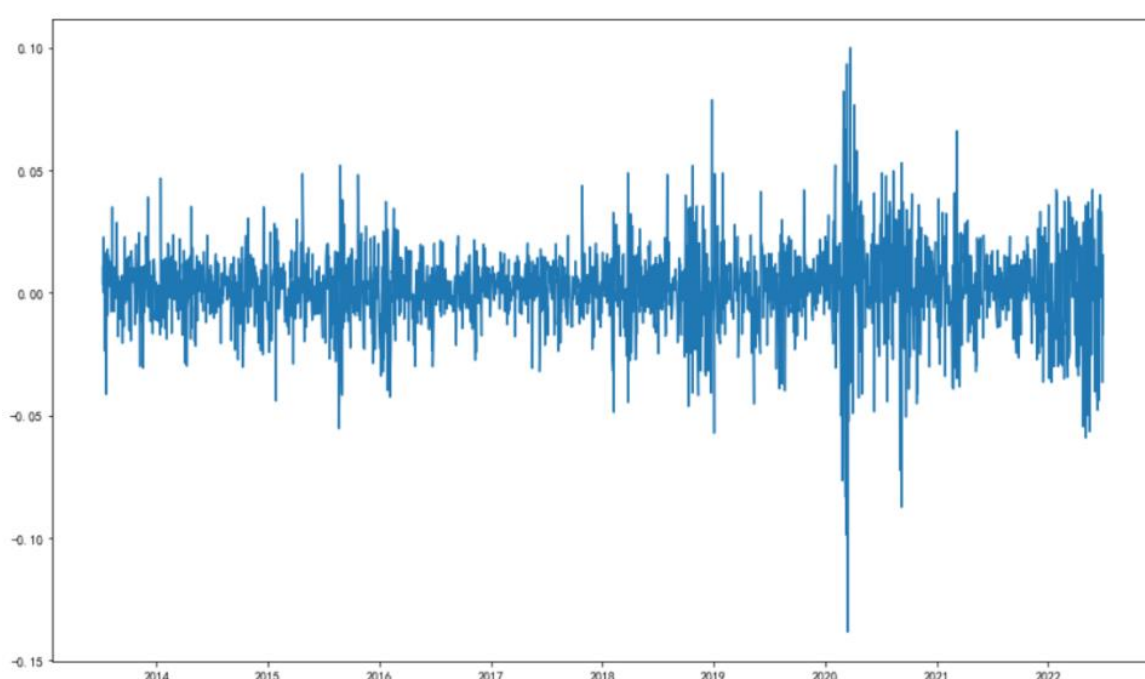


Fig. 5 Time series of the Maximum Sharp Ratio portfolio

Under a certain confidence level, for time series data, make a Null Hypothesis: time series data is unstable. Time series data will be stable if the Null Hypothesis can be rejected with a specific degree of confidence. The p value is significantly below 0.05, and the test statistic value is significantly below the critical values in the three scenarios of 1 percent, 5 percent, and 10 percent, according to data in table 4. Therefore, the data is stable, and the difference is not necessary.

Table 4. The Dickey-Fuller test results

	Test statistics	P-value	Lags Used	Number of Observation Used	Critical Value(1%)	Critical Value(5%)	Critical Value(10%)
Results	-15.586	0	8	2255	-3.433	-2.863	-2.567

After that, by viewing figures 6 and 7, it is possible to identify the order of the model, which is represented by the values of p and q in the ARIMA (p, I q), so that p equals 3 and q equals 5. In addition, the value of i is 0, because there is not difference.

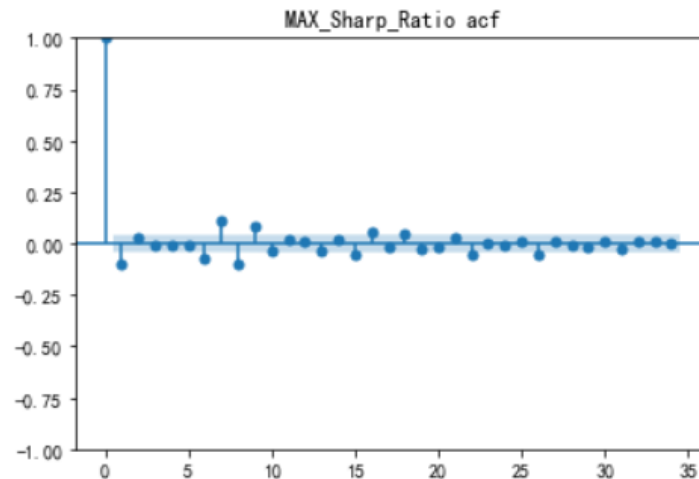


Fig. 6 ACF of the Maximum Sharp Ratio portfolio

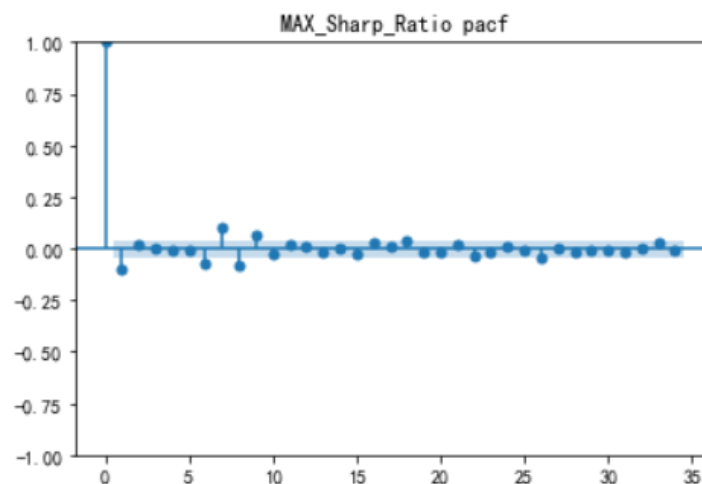


Fig. 7 PACF of the Maximum Sharp Ratio portfolio

After predicting the return rate of the portfolio using the correspondingly trained ARIMA model and LSTM model, through figure 8, the result is that ARIMA model performed better in the prediction of the first 10 days, but the performance has deteriorated in the prediction of longer time, which means that the predicted value begins to tend to the average. Besides, according to the figure 9, the LSTM model can predict the complete 30 days and has a better performance in long-term prediction.

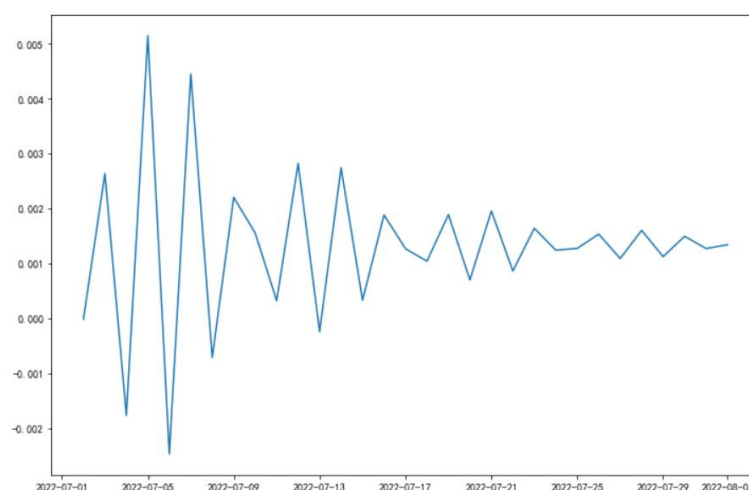


Fig. 8 The predict of ARIMA model in 30 days

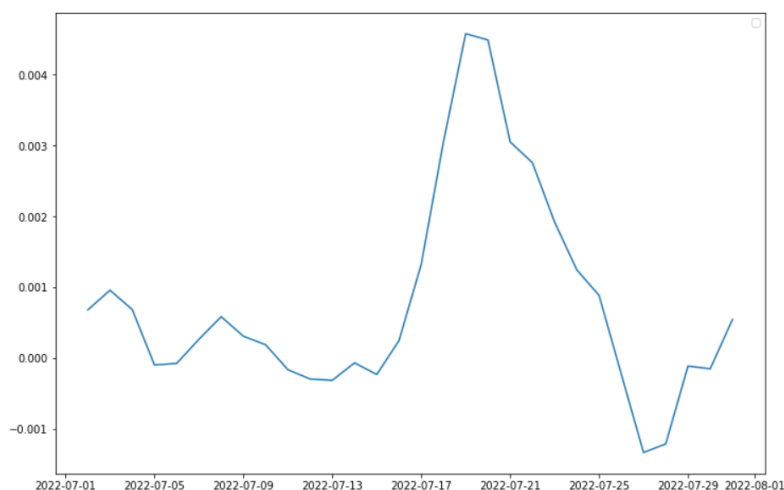


Fig. 9 The predict of LSTM model in 30 days

3. Conclusion

At present, most relevant studies only research the application of ARIMA and LSTM models in the financial field separately. The purpose of this paper is to analyze the difference between ARIMA and LSTM models in predicting the return of portfolio in order to help investors to choose more appropriate time series prediction models to obtain greater returns. This paper selects the maximum sharp ratio portfolio formed by the stock data of several companies from 2013 to 2022 and predicts the future return. It is found that ARIMA model performs better in short-term prediction and stable data prediction. In contrast, LSTM model has higher accuracy in long-term prediction. However, whether these models are universal needs further study, and the data source can be extended.

References

- [1] Thuankhonrak P, Rattagan E, Phoomvuthisarn S. Machine trading by time series models and portfolio optimization//2019 4th International Conference on Information Technology (InCIT). IEEE, 2019: 217-222.
- [2] Jackson J C, Prassanna J, Quadir M A, et al. Stock market analysis and prediction using time series analysis. Materials Today: Proceedings, 2021.
- [3] Centeno V, Georgiev I R, Mihova V, et al. Price forecasting and risk portfolio optimization//AIP conference proceedings. AIP Publishing LLC, 2019, 2164(1): 060006.
- [4] Wang W, Li W, Zhang N, et al. Portfolio formation with preselection using deep learning from long-term financial data. Expert Systems with Applications, 2020, 143: 113042.
- [5] Sen J, Dutta A, Mehtab S. Stock portfolio optimization using a deep learning LSTM model//2021 IEEE Mysore Sub Section International Conference (MysuruCon). IEEE, 2021: 263-271.
- [6] Ma Y, Han R, Wang W. Portfolio optimization with return prediction using deep learning and machine learning. Expert Systems with Applications, 2021, 165: 113973.
- [7] Chen W, Zhang H, Mehlatat M K, et al. Mean-variance portfolio optimization using machine learning-based stock price prediction. Applied Soft Computing, 2021, 100: 106943.
- [8] Senneset K, Gultvedt M. Something old, something new: a hybrid approach with ARIMA and LSTM to increase portfolio stability, 2020.
- [9] Choi H K. Stock price correlation coefficient prediction with ARIMA-LSTM hybrid model. arXiv preprint arXiv:1808.01560, 2018.

- [10] Siامي-Namini S, Tavakoli N, Namin A S. A comparison of ARIMA and LSTM in forecasting time series//2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2018: 1394-1401.