2023 IEEE 6th International Conference on Information Systems and Computer Aided Education (ICISCAE) | 979-8-3503-1344-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICISCAE59047.2023.10392876

Short-Term Stock Price Prediction Based on Long-Short Term Memory Neural Network Model

Baofeng Li
Business School
Henan University of Science and Technology
Luoyang, China
bfli2008@163.com

Yangyang Li*
Business School
Henan University of Science and Technology
Luoyang, China
*Corresponding author: 2053738255@qq.com

***Abstract*—The rapid development of artificial intelligence technology affects every aspect of our lives and shows the characteristics of computing everywhere. As one of the closest branches of AI technology, deep learning is widely used in various aspects. Neural network models have a prominent advantage in stock prediction due to their powerful learning and computational capabilities. This research adopts the long-short term memory model (LSTM) to make short-term forecasts of the maximum and closing prices of the stock indexes using the relevant data of the CSI 300 index from July 31, 2019 to July 28, 2023 as a sample. The results show that the LSTM model has a good performance on stock price prediction and can achieve a better fitting effect and prediction accuracy.**

Keywords*—LSTM, neural network, stock price prediction, artificial intelligence

I. INTRODUCTION

Stock price fluctuations can reflect the performance of the national economy, the correct judgment of the stock price trend, not only for individual retail investors and institutional investors to effectively allocate funds, make the right decision, to provide a useful reference, but also conducive to the stability and health of the market economic environment. Therefore, stock price forecasting has always been one of the key issues for scholars to study and pay attention to. In the research related to stock price prediction, according to the classification of prediction methods, it can be mainly categorized into time series analysis, machine learning and deep learning. Time series models can be used for stock price forecasting mainly because of the time continuity of stock prices. Zhijun Yu and Shanlin Yang (2013) used a generalized autoregressive conditional heteroskedasticity forecasting model (GARCH) and introduced an error correction method to forecast the SSE index, and concluded that the corrected model improved the forecasting accuracy and the model was valid [1]. In addition to this, by establishing ARIMA model to analyze and forecast the time series of GEM stock prices, it proves that ARIMA model can also have a better prediction effect on short-term stock prices [2]. Xiao, D. and Su, J. (2022) used both ARIMA and LSTM models to forecast a sample of New York Stock Exchange stocks and stated that although the ARIMA model achieves accurate predictions, the LSTM model works better [3]. Stock prices are characterized by non-linearity and non-stationarity, and are affected by other factors in addition to the time factor. Therefore, traditional time series models may have difficulty in making accurate forecasts when faced with stock price data that are highly noisy, non-stationary, and have multiple influencing factors. With the development of computer technology, machine learning algorithms are gradually being applied to predict stock prices. Yiyue He et al. (2019) proposed the support vector regression method under EMD decomposition for stock price prediction, and concluded that EMS-SVRF has smaller prediction and fitting errors and higher stock price prediction accuracy compared to stock price prediction methods such as ARMA-GARCH [4]. Xiangfeng Yin et al. (2021) argued that twinned support vector regression with two types of kernel functions is informative for stock price prediction [5], but SVR often suffers from parameter tuning difficulties as well as kernel function selection problems. Zhengxu Yan et al. (2021) added Pearson coefficients to the Random Forest model and proved that the improved Random Forest model can be applied to short-term forecasting of stock prices and reduce the effect of noise [6].

Deep learning is a new branch and research area of machine learning that typically employs artificial neural networks to learn and interpret data. In stock price forecasting, neural network models are also widely used, which mainly include BP neural networks, convolutional neural networks (CNN), recurrent neural networks (RNN), and long-short term memory (LSTM). Since BP neural network has a large error in stock price prediction, fangzhong Qi et al. (2020) introduced the PCA method and the improved Drosophila algorithm into the BP neural network model, proved that PCA-IFOA-BP neural network is effective in stock price prediction [7]. However, BP neural network, as a shallow neural network, has a single structure and is prone to overfitting, gradient vanishing and gradient explosion problems [8]. A large number of scholars have conducted more research on the application of LSTM model in the field of stock price forecasting, moreover, most of them believe that LSTM model performs well in stock price prediction. Yuxu Feng and Yumei Li (2019) found that the LSTM model was able to obtain lower RMSE values and the LSTM model outperformed the SVR and Adaboost models in their study of stock index maximum prices [9]. Optimization and improvement of LSTM model, using optimization or combination model for stock price prediction [10-12], also reflects the feasibility and advantages of LSTM model in stock price forecasting. Yiqiu Fang et al. (2022) united the CNN model with the LSTM model to establish an LSTM-CNN model based on RMSE, and at the same time established the BP and CNN models as a control, and concluded that the joint model has better performance and feasibility, and outperforms the BP and CNN models [13]. Yuwen Hu (2021) optimized the LSTM model adopting PCA and LASSO methods respectively and compared them, stated that optimizing the LSTM model by using principal component analysis can make the model have strong generalization ability and higher prediction accuracy [14]. In short-term forecasting of stock prices, Liu, H. et al. (2022) proposed a CAE-LSTM combination model and concluded that the combination model has good predictive ability [15]. In addition to this, Abdul Quadir Md et al. (2022) argued that LSTM model can overcome the obstacles to achieve accurate prediction of stock prices and concluded that the optimized LSTM model can accurately predict the stock prices by building a multilayer sequential LSTM model [16]. However, some scholars have argued that it is not the case that the deeper the neural network level, the better the results achieved [17]. Similarly, B. Gülmez (2023) argued that the LSTM model can accurately predict stock prices, moreover, proved that the LSTM-ARO model is effective in stock price prediction by optimizing the LSTM model using the artificial rabbit optimization algorithm (ARO) with the data of the Dow Jones Industrial Average stocks [18].

Summarizing the above analysis, the volatility of stock prices is affected by all sorts of other factors in addition to the time factor, so the time series analysis method is generally difficult to make effective forecasting for highly noisy, nonlinear, and non-stationary stock price data. With the advent and maturity of machine learning algorithms plus neural network models, more and more studies have denoted that the LSTM model has a better performance in stock price forecasting. Existing researches based upon the pure LSTM model have made relevant predictions on the stock price of individual stocks or the price of stock indexes, or have used the improved and joint model to make stock price predictions, which has verified the effectiveness of the LSTM model in stock price forecasting. In this research, the LSTM model is adopted to predict the maximum and closing prices of the CSI 300 index. Unlike some studies, which are devoted to the optimization of the underlying model itself, only a single data processing method, normalization, is used to quantify the raw data. In this paper, both normalization and standardization methods are used to process the data and form an experimental control to find a more suitable data processing method for the model, which in turn enhances the prediction accuracy of the model, so as to achieve a better stock price prediction. This research volunteers a reference basis for the effectiveness of LSTM model on the application of stock price prediction and provides a useful reference for investors to allocate funds effectively and make correct decisions.

II. LSTM MODEL CONSTRUCTION

The LSTM model can be understood as an improvement and extension of the recurrent neural network model RNN. RNN models use the same weight matrix, and when the weight matrices are multiplied cyclically, multiple combinations of the same function tend to produce extreme nonlinear behavior, which leads to the gradient vanishing and gradient explosion problems. As the amount of information increases, RNN models suffer from poor long-range dependence and cannot effectively handle long sequential data. Therefore, the LSTM model came into being, which has two layer structures and three gate mechanisms, namely, cell layer, hidden layer, input gate, forgetting gate, output gate, respectively. The hidden layer stores the short-term state, the cell layer stores the long-term state, and the information between the two is interacted through three kinds of gates, and the cell structure of the LSTM model is clearly presented in Fig. 1.

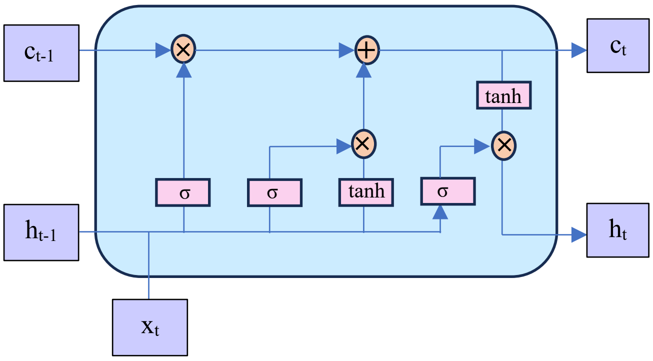

Fig. 1. LSTM cell structure

The function of the gate mechanism is to remove or add data information to cell layer state. Specifically, input gates, also called update gates, determine the updating of information in the cell state, generating a temporary new state or updating the old state; the forgetting gates controls what information should be discarded; and the output gates determine the output of the information. The formulas for the individual gates of the LSTM model are as follows:

Forgotten gate:

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

Input gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\bar{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \bar{c}_t \tag{4}$$

Output gate:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

In the above formulas, W_f, W_i, W_c, W_o respectively denotes the corresponding weights, xt represents the input feature variable at the moment t, b_f, b_i, b_c, b_o separately denote the corresponding bias values, ft, it, ct, ot, ht represent the output or state values of the individual neural network units, and σ is the logistic sigmoid activation function, tanh represents the tangent hyperbolic function.

III. DATA PROCESSING AND SELECTION OF INDICATORS

A. Sample Selection

For ensuring the validity and applicability of the model, this research selects the relevant data of CSI 300 stock index as the sample to forecast the closing price and maximum price of the stock index, and the constituent stocks of this index are mainstream investment stocks in the market with high liquidity, large scale, and good representativeness, which can effectively reflect the overall trend of the market; The selected sample interval is from July 31, 2017 to July 28, 2023, and the daily data of opening price, closing price, volume, turnover, high price and low price of CSI 300 index with a total of 6 features are used as input features, and the dataset is divided into a training set and a test set in the retio of 8:2, so as to predict the index maximum price and closing price; The data involved are taken from the Cathay Pacific database.

B. Data Processing Methods

Because the selected data sets have different magnitudes, in an effort to avoid that the different magnitudes of the data would interfere with the predictive model, so before the data enter the model for training, it is essential to fulfil the dimensionless processing of the feature data, moreover, selecting the two methods of normalization and standardization, and experimental control.

Normalization formula:

$$\bar{X}_k = \frac{X_k - X_{k_{min}}}{X_{k_{max}} - X_{k_{min}}} \tag{7}$$

Standard method formula:

$$\bar{X}_k = \frac{X_k - X_{k_{mean}}}{X_{k_{std}}} \tag{8}$$

In the above equation, X_k represents the kth feature in the original dataset, X_{kmin} represent the minimum values, on the contrary, X_{kmax} denotes maximum values, in addition, X_{kmean} is the mean of data, as well as X_{kstd} is standard deviation, and $\bar{X}_k$ denotes the data after quantization. After normalization, the eigenvector values are all mapped between the intervals [0,1], and after normalization, the mean of the data becomes 0 as well as the standard deviation becomes 1. Some of the data after normalization and standardization are presented in Table I and Table II.

TABLE I. SELECTED DATA AFTER NORMALIZATION

Dates	Open	High	Low	Close	Vloume	Money
2017/7/31	0.2623	0.2546	0.2766	0.2719	0.2749	0.1774
2017/8/1	0.2678	0.2627	0.2850	0.2834	0.2865	0.1986
2017/8/2	0.2777	0.2675	0.2929	0.2800	0.3421	0.2267
2017/8/3	0.2733	0.2614	0.2783	0.2684	0.2685	0.1840
2017/8/4	0.2631	0.2519	0.2743	0.2613	0.3003	0.1903
2017/8/5	0.1903	0.2492	0.2682	0.2680	0.2068	0.1314
2017/8/6	0.2630	0.2515	0.2785	0.2699	0.2434	0.1524
2017/8/7	0.2642	0.2520	0.2766	0.2695	0.2080	0.1364
2017/8/8	0.2632	0.2547	0.2676	0.2642	0.2232	0.1551

TABLE II. SELECTED DATA AFTER STANDARDIZATION

Dates	Open	High	Low	Close	Vloume	Money
2017/7/31	-0.7467	-0.7527	-0.7172	-0.7241	0.3392	-0.6117
2017/8/1	-0.7184	-0.7115	-0.6761	-0.6677	0.4237	-0.4882
2017/8/2	-0.6674	-0.6875	-0.6373	-0.6842	0.8290	-0.3247
2017/8/3	-0.6900	-0.7185	-0.7088	-0.7415	0.2925	-0.5731
2017/8/4	-0.7427	-0.7661	-0.7288	-0.7766	0.5247	-0.5370
2017/8/5	-0.7834	-0.7798	-0.7587	-0.7433	-0.1575	-0.8795
2017/8/6	-0.7434	-0.7685	-0.7081	-0.7339	0.1095	-0.7571
2017/8/7	-0.7370	-0.7656	-0.7173	-0.7359	-0.1483	-0.8505
2017/8/8	-0.7423	-0.7521	-0.7618	-0.7622	-0.0373	-0.7417

C. Indicator for Assessing the Accuracy of Forecasts

For measuring the prediction effect of LSTM model, this research chooses the mean square error MSE as the evaluation index, the smaller its value is, the higher the prediction accuracy is and the better the selected model is, and the MSE is calculated by the formula:

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(\tilde{y}_k - y_k)^2 \tag{9}$$

In the above equation $\tilde{y}_k$ and y_k respectively represents the predicted and actual values, when the difference between the two is smaller, the MSE value is also smaller, and it also indicates that the prediction accuracy of the model is better.

IV. Experimental results

A. Optimization of the Number of Neurons

When using LSTM model to train historical data, the relevant setup parameters of the model, such as time step, number of neurons, and the times of iterations, etc., will have an impact on the model's fitting and prediction, and for this reason, it is necessary to find a suitable parameter to avoid the model's large prediction error, bad prediction and overfitting.

Regarding the number of neuron nodes, before performing the model training, it is the essential and critical parameter that needs to be considered. Different number of neuron nodes will have an impact on the prediction of model. When the number of neuron nodes is small, the ability of the neural network to learn and process data information will be weakened, and may not be able to complete the necessary information processing, however, the number of neurons is not the more the better, if it is too much, will lead to an increase in the complexity of the network structure, which may result in the emergence of the problem of overfitting and making the neural network model learning slower, which also weakens the predictive power of the model. Therefore, a suitable number of neuron nodes needs to be looked for. Fixing the step size and testing the number of better neurons, the results of the optimization search are shown in Table III and Table IV. This experiment uses both normalization and standardization to process the data to form an experimental control, which further verifies that the use of different methods of data dimensionless processing will have an effect on the forecasting capability of the model.

TABLE III. OPTIMIZATION OF NEURON NUMBER SEARCH UNDER NORMALIZATION

Number of Neurons	High		Close	
	MSE	*R²*	*MSE*	*R²*
8	0.15872064	-1.88431697	0.14209249	-1.40831242
16	0.11353314	-1.06315678	0.123239	-1.08876641
32	0.06566998	-0.19337371	0.05806067	0.01593509
64	0.03112795	0.43433388	0.02765632	0.53125555
128	0.01697003	0.69161575	0.01674239	0.71623482
256	0.01748033	0.68234249	0.01607527	0.72754174

TABLE IV. OPTIMIZATION OF THE NUMBER OF NEURONS UNDER STANDARDIZATION

Number of Neurons	High		Close	
	MSE	*R²*	*MSE*	*R²*
8	0.00555597	0.89903526	0.00781454	0.86755213
16	0.00516038	0.90622409	0.00633917	0.892558
32	0.00556142	0.89893631	0.00710937	0.87950399
64	0.00583425	0.89397839	0.00752908	0.87239035
128	0.00897468	0.83690963	0.01114941	0.81102972
256	0.00515764	0.9062739	0.00713534	0.8790638

From the results, it is clear that the model works better when the data are processed using the standardized method. Under the standardized method, a smaller MSE value can be achieved, and R^2 is closer to 1, which makes the prediction error smaller and the fit better, so the subsequent parameter optimization searching, all use the standardized method to process the data; Under the normalization method, a neuron node number of 16 leads to optimal of the LSTM model. With this result, it is set to 16 and optimization is sought for the step size of the stock time series data.

B. Time Step Optimization

The neural network model predicts according to a certain length of time series data, the longer the time step, the more historical data it contains, but too long a time step will lead to an increase in the invalid information input into the model, increasing the difficulty of model training, and its prediction effect will become worse, and too short a time step will directly reduce the effective information in the input data, which will also make the model's prediction ability decline, so it is necessary to determine a more reasonable time step.

TABLE V. TIME STEP OPTIMIZATION

Pacemaker	Number of Neuron Nodes(16)			
	High		Close	
	MSE	***R²***	***MSE***	***R²***
8	0.00547263	0.89659019	0.00722309	0.87275715
10	0.0042943	0.91873889	0.00622671	0.89031528
12	0.00425897	0.91952333	0.00601247	0.89408354
14	0.00361496	0.93159406	0.00581809	0.89751317
16	0.00464346	0.91286829	0.00858915	0.85014763
18	0.00817346	0.8455559	0.0102893	0.81874232
20	0.00516038	0.90622409	0.00633917	0.892558
22	0.00609253	0.8894012	0.00764532	0.86968828

Table V shows that when the number of neuron nodes is 16, choosing a time step of 14 minimizes the forecasting error of the LSTM model, at which time the MSE values of the maximum and closing prices are 0.00361496 and 0.00581809, and the R^2 is 0.93159406 and 0.89751317, respectively.

C. Determination of the Times of Model Iterations

The times of iterations will also affect the final prediction results, the number of iterations has no relationship with the neural network model itself, but it will affect the prediction results performance, but also an important parameter of neural network model, it is too small, can not achieve a better learning effect, it will make the model prediction accuracy decline. In general, train the model until the MSE value converges or the model prediction accuracy is no longer improved can stop training, according to the actual need to determine the number of iterations required for the experiment.

From Fig. 2 and Fig. 3, with the increase of the number of iterations, the MSE values of both training set and test set first decrease rapidly, then gradually stabilize. When the model is iterated about 80 times, the MSE value gradually converges and stabilizes, and the number of iterations is determined to be 100 according to the needs of this experiment.

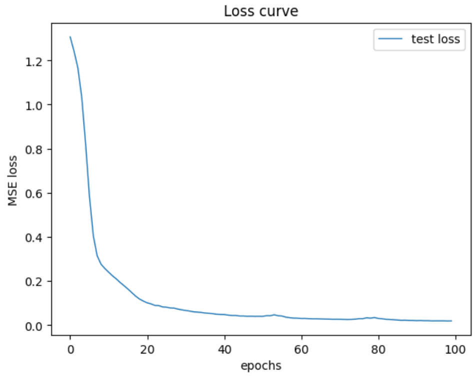

Fig. 2.MSE loss function for the test set

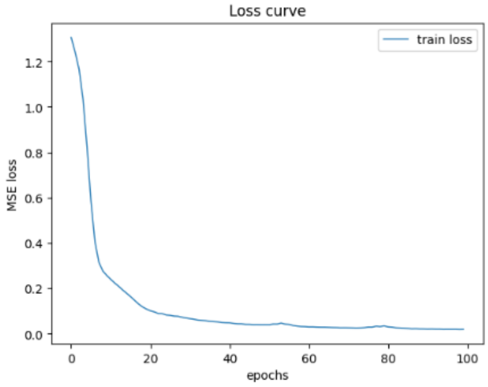

Fig. 3.MSE loss function for training set

D. LSTM Model Prediction Results

According to the outcome of the previous parameter optimization search, the main parameters of this experiment are set, which are presented in Table VI. On this basis, this paper applies the LSTM model to train the data in the training set, and then validate it using the data in the test set, adopting the historical data of stock price to realize the short-term prediction of the future day's stock price. The experimental results indicate that when the LSTM model is adopted to predict the stock price of CSI 300 index, the MSE value of the maximum price is 0.00361496, and the closing price is 0.00581809. In addition to this, the data of the test set is put into the LSTM model to forecast the maximum and closing prices of the CSI 300 index in the coming day, and then the predicted values are compared with the actual values, and plotted for visualization. From the graphs, there is a high degree of overlap between the two, which indicates that the LSTM model performs well and achieves good fitting effect and prediction accuracy, and the prediction effect of the maximum price and closing price is shown in Fig. 4 and Fig. 5.

TABLE VI. MAIN PARAMETERS

Model	Parameters	Value
LSTM	pacemaker	14
	number of neuron nodes	16
	times of iterations	100
	number of hidden layers	2
	activation function	tanh
	optimization function	Adam

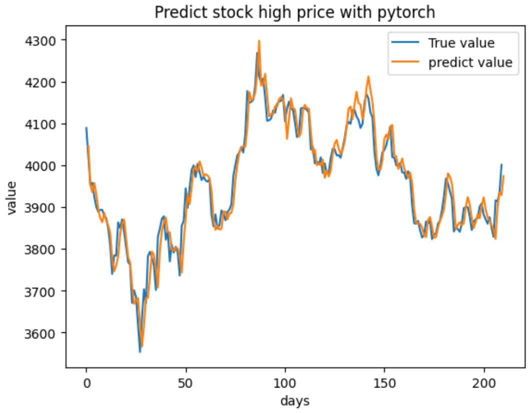

Fig. 4. Forecast of the highest price of the stock index

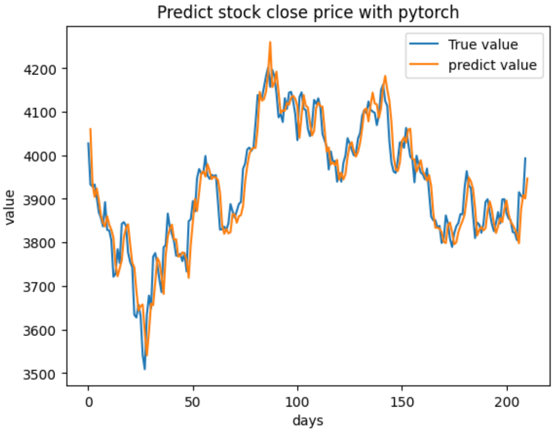

Fig. 5. Forecast of the closing price of the stock index

V. Conclusion

In this paper, we constructed an LSTM neural network model to forecast the maximum and closing prices of CSI 300 index in the short term, and compared the effects of normalization and standardization on the prediction performance of LSTM model, and came to the following conclusions through the theoretical analysis and experimental results in the previous paper:

(1) Constructing LSTM model for short-term prediction of CSI 300 index stock price has applicability and validity, and the model can achieve better fitting effect and prediction accuracy, which also indicates the feasibility of using LSTM model for stock price prediction research.

(2) Optimizing the number of neurons, the time step and the times of iterations, as well as selecting the appropriate model parameters, which has a certain impact upon the performance of model prediction effect. From the experimental results, when the time step is 14 as well as the number of neurons is 16, the LSTM model predicts the best performance, and the number of iterations is 100, which is more suitable for this experiment. This also reflects that not the longer the length of the data incorporated into the model, the more effective information the model learns, and the blind increase in the number of neurons may also damage the model accuracy. Therefore, when using the LSTM model for stock price forecasting, it is necessary to determine the appropriate model parameters.

(3) When dimensionless processing of raw data, different processing methods, will have a certain impact on prediction accuracy of the model. For this reason, in the use of LSTM model, it is quite indispensable to test the prediction performance of the model under different data processing methods, so as to determine the appropriate method, in order to enhance the model prediction accuracy.

References

[1] Z. Yu and S. Yang, “A model for stock price forecasting based on error correction,” Chinese Journal of Management Science, vol. 21, pp. 341-345, 2013.

[2] Y. Wu and X. Wen, “Short-term stock price forecasting based on ARIMA modeling,” Statistics and Decision, vol. 23, pp. 83-86, 2016.

[3] D. Xiao and J. Su, “Research on stock price time series prediction based on deep learning and autoregressive integrated moving average,” Scientific Programming, vol. 2022, pp. 4758698, 2022.

[4] Y. He, N. Gao, F. Wang, S. Ru, and J. Han, “Research on integrated forecasting of stock price based on EMD and support vector regression,” Journal of Northwest University, vol. 49, pp. 329-336, 2019.

[5] X. Yin, H. Cui, and X. Wen, “Comparison of TSVR based on two kinds of kernel functions in stock price forecasting,” Statistics and Decision, vol. 37, pp. 43-46, 2021.

[6] Z. Yan, C. Qin, and G. Song, “Random forest model stock price prediction based on pearson feature selection,” Computer Engineering and Applications, vol. 57, pp. 286-296, 2021.

[7] F. Qi, S. Lin, and T. Yu, “A prediction model of stock price based on PCA and IFOA-BP neural network,” Computer Applications and Software, vol. 37, pp. 116-121+156, 2020.

[8] Z. Zhou and X. He, “Stock price prediction method based on optimized LSTM model,” Statistics & Decision, vol. 39, pp. 143-148, 2023.

[9] Y. Feng and Y. Li, “A research on the CSI 300 index prediction model based on LSTM neural network,” Mathematics in Practice and Theory, vol. 49, pp. 308-315, 2019.

[10] S. Chen and L. Ge, “Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction,” Quantitative Finance, vol. 19, pp. 1507-1515, 2019.

[11] W. Liu, J. Li, Y. Li, A. Sun, and J. Wang, “A CNN-LSTM-based model to forecast stock prices,” Complexity, vol. 2020, pp. 6622927, 2020.

[12] F. Liu, P. Qin, J. You, and Y. Fu, “Sparrow search algorithm-optimized long short-term memory model for stock trend prediction,” Computational Intelligence and Neuroscience, vol. 2022, pp. 3680419, 2022.

[13] Y. Fang, Z. Lu, and J. Ge, “Forecasting stock prices with combined RMSE loss LSTM-CNN model,” Computer Engineering and Applications, vol. 58, pp. 294-302, 2022.

[14] Y. Hu, “Stock forecast based on optimized LSTM model,” Computer Science, vol. 48, pp. 151-157, 2021.

[15] H. Liu, L. Qi, and M. Sun, “Short-Term stock price prediction based on CAE-LSTM method,” Wireless Communications and Mobile Computing, vol. 2022, pp. 4809632, 2022.

[16] A.Q. Md et al., “Novel optimization approach for stock price forecasting using multi-layered sequential LSTM,” Applied Soft Computing, vol. 134, pp. 109830, 2023.

[17] J. Zhao, D. Zeng, S. Liang, H. Kang, and Q. Liu, “Prediction model for stock price trend based on recurrent neural network,” Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 745-753, 2021.

[18] B. Gülmez, “Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm,” Expert Systems with Applications, vol. 227, pp. 120346, 2023.