# Research on High-frequency stock price prediction based on Chebyshev-Stacking and Weighted LSTM neural network [1]

Yiwen Wang [1, #], Yida Zhang [2, #], Zixuan Pan [3, *, #]

[1] International Business School, Shaanxi Normal University, Xi 'an, Shaanxi, 710000

[2] School of Economics and Management, Fuzhou University, Fuzhou, Fujian, 350000

[3] Department of Economic,University of California, Berkeley , 530 Evans Hall, Berkeley, CA ,United.State of America, 94720

* Corresponding Author Email: Pzx0317@berkeley.edu

[#]These authors contributed equally.

**Abstract.** To solve the problem of high-frequency stock price prediction, this paper proposed a prediction model based on Chebyshev-Stacking and a weighted LSTM neural network. The proposed method extracts the function characteristic information of the high-frequency stock price series through Chebyshev orthogonal polynomial basis expansion. Considering that the potential model structure between each component of the function feature vector and the residual sequence predicted by the LSTM neural network is unknown and there is a certain noise, this paper used the Stacking framework to enhance the data and weighed the bias and variance of the prediction model. In addition, since the number of predictor variable periods of the LSTM neural network is a hyperparameter, the model averaging method based on distance covariance is used for optimization. The results of actual data analysis show that the proposed method is significantly better than the original LSTM neural network in terms of mean square error, absolute error, and relative error. By selecting the different number of training sets, the robustness of the improved model is verified. Finally, the proposed method can also be used in practical applications such as daily average temperature prediction, missile trajectory prediction, and real-time monitoring of atmospheric environment quality.

**Keywords:** High-frequency timing prediction, Chebyshev-Stacking, model averaging, distance covariance.

## 1. Introduction

Stock price prediction has always been a research hotspot in the field of finance. In the era of big data, due to the development of computer technology, stock investors can obtain various real-time trading data of the stock market on time. Compared with low-frequency time series data, such stock price data recorded at high frequency includes more information and is easier to capture the micro changes of the stock market. Therefore, people generally use high-frequency time series to predict the future trend of stock prices. However, there are some problems in modeling such high-frequency time series data, such as dimensional disaster, high noise, and difficulty in fully extracting effective information.

At present, there are many methods to predict stock prices, such as the ARIMA model, the GARCH model, and so on. Yang Yuyuan and Zhang Mei (2021)[2]used the ARIMA model for fitting and established linear regression expression to predict the stock opening price. This model has a small prediction error and can provide a certain basis for a stock investment decision. However, due to the characteristics of high-frequency time series data such as nonlinearity, non-stationarity, and high noise, traditional time series prediction methods have certain limitations. With the continuous development of machine learning, more and more researchers apply neural networks in stock research. For example, Li Liping (2022)[3]et al. used LSTM neural network prediction model to predict the closing price of Yunnan tourism stocks. Lin Xin and Zhu Xiaodong (2022)[4]used the LSTM model based on the Attention mechanism to predict high-frequency stock price series and compared its prediction results with MLP, RNN, and LSTM. Huang Yucheng and Fang Weiwei (2021)[5]applied

the LSTM network to fit the fluctuation rule of high-frequency stock prices, reflecting the fluctuation trend of stock prices. Peng Yan, Liu Yuhong, and Zhang Rongfen (2019)[6]combined the characteristics of the LSTM neural network and the characteristics of the stock market conducted preprocessing operations such as interpolation, wavelet denoising, and normalization on the stock price time series, and predicted the stock trend before stock investment. Sun Bingjie (2016)[7]et al. used wavelet decomposition, Elman neural network, BP neural network model, and other technologies to predict stock price series. Zhang Miaomiao (2021)[8]studied stock price prediction in the case of high dimensional variables and mixed frequency data based on LASSO variable screening, factor analysis, and LSTM neural network. Xie Xinrui, Lei Xiuren, and Zhao Yan (2020)[9]proposed a dual dimensionality reduction method combining mutual information and improved PCA to predict the actual stock data through neural networks. Zou Jie and Li Lu (2022)[10]introduced Random Forest (RF) algorithm, a dimensionality reduction processing technology, and constructed a hybrid model to predict the stock prices of 18 stocks with high circulating market values involving 18 basic industries. By using machine learning and deep learning techniques, these methods effectively deal with the dimensional disaster problem and extract the linear and nonlinear information of time series, which enriches the modeling methods of high-frequency time series data. However, most of the current high-frequency stock price prediction models do not use the function information of time series, nor do they consider how to capture the potential model structure between the components of the feature vector and the response variable after dimensionality reduction. In addition, the selection of the dimension of the predictive variable of LSTM is also a problem to be solved.

Based on this, we proposed a prediction model based on Chebyshev-Stacking and a weighted LSTM neural network. The innovation of the proposed method is mainly in the following aspects: first, it solves the problem of dimensional disaster and extracts the function information of high-frequency time series; Secondly, the latent model structure of each component of the feature vector and the response variable is captured by Stacking and the data noise is reduced. Thirdly, the model selection problem of the LSTM neural network is solved. Experimental results show that the proposed method can significantly improve the prediction accuracy of the original LSTM. The remaining content of this paper is arranged as follows: Section 2 introduces the prediction model based on Chebyshev-Stacking and weighted LSTM neural network; Section 3 will show the actual data analysis results. The fourth section is the conclusion.

## 2. Theory and Method

### 2.1. Chebyshev orthogonal polynomial basis expansion

Chebyshev orthogonal polynomial expansion is the expansion of smooth functions on finite intervals with Chebyshev polynomials as the basis.

$$X(t) = \sum_{i=1}^{I} a_i \varphi_i(t) \tag{1}$$

Where is the Chebyshev orthogonal polynomial of order I, which can represent an eigenvector. The expansion process of Chebyshev orthogonal polynomials can be reduced to fitting infinite-dimensional data with finite-dimensional (i-dimensional) coefficient vectors. In this process, the number of basis functions can determine the degree to which the data is smoothed. When the dimension of I is very small, smoothing is very effective, but the problem of underfitting will occur, which requires us to strengthen the investigation of the bias and variance of equation (1).

### 2.2. Stacking framework

Stacking is through the model of the original data fitting model of the stack, first by machine learning the original data, then these base learning will take place on the original data output, the

second will this a few of the model output under the list of the stack in the form of composition (m, p) d of new data, m representative sample, p represents the number of learning, Finally, the new sample data is handed over to the second layer model for fitting. The stacking process framework is shown in Figure 1 below:

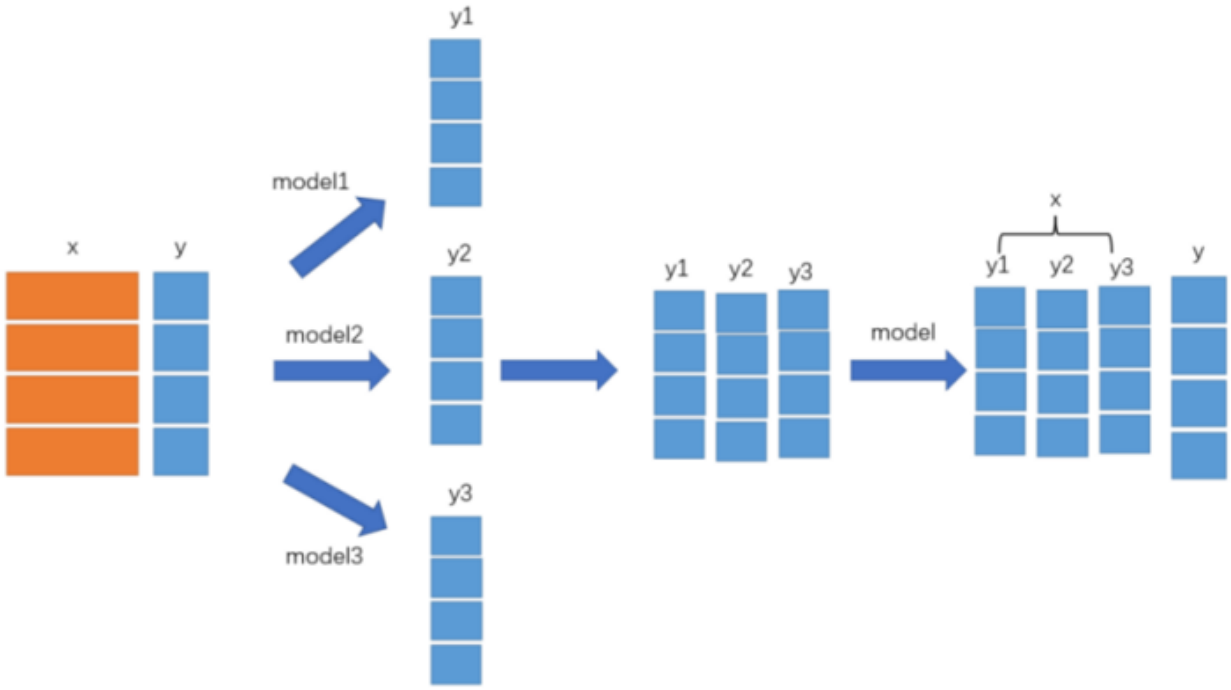



**Figure 1.** Stacking process.

Firstly, the feature x and label y are input into three models respectively, and the three models are learned separately. Then, the predicted value is given for X, and the output values of the three models are stacked in the way of columns to form new sample data. The new sample data is regarded as the label x, and the label of the new data is still the label y of the original data. The x and y of the new data are handed to the second layer model for fitting, which is used to fuse the results of the three models in the previous round.

### 2.3. LSTM Neural Network

LSTM is a special kind of recurrent neural network. This kind of network is different from the general feedforward neural network. LSTM can analyze the input using time series.LSTM has been used to solve the common long-term dependence problem in the general recursive neural networks since its design. LSTM can effectively transmit and express the information in the long time series and will not lead to the forgotten useful information of a long time ago. At the same time, LSTM can also solve the gradient disappearance/explosion problem in RNN, so it has better performance in processing time series data.The calculation of Cells in the LSTM model can be represented by the following Figure 2:

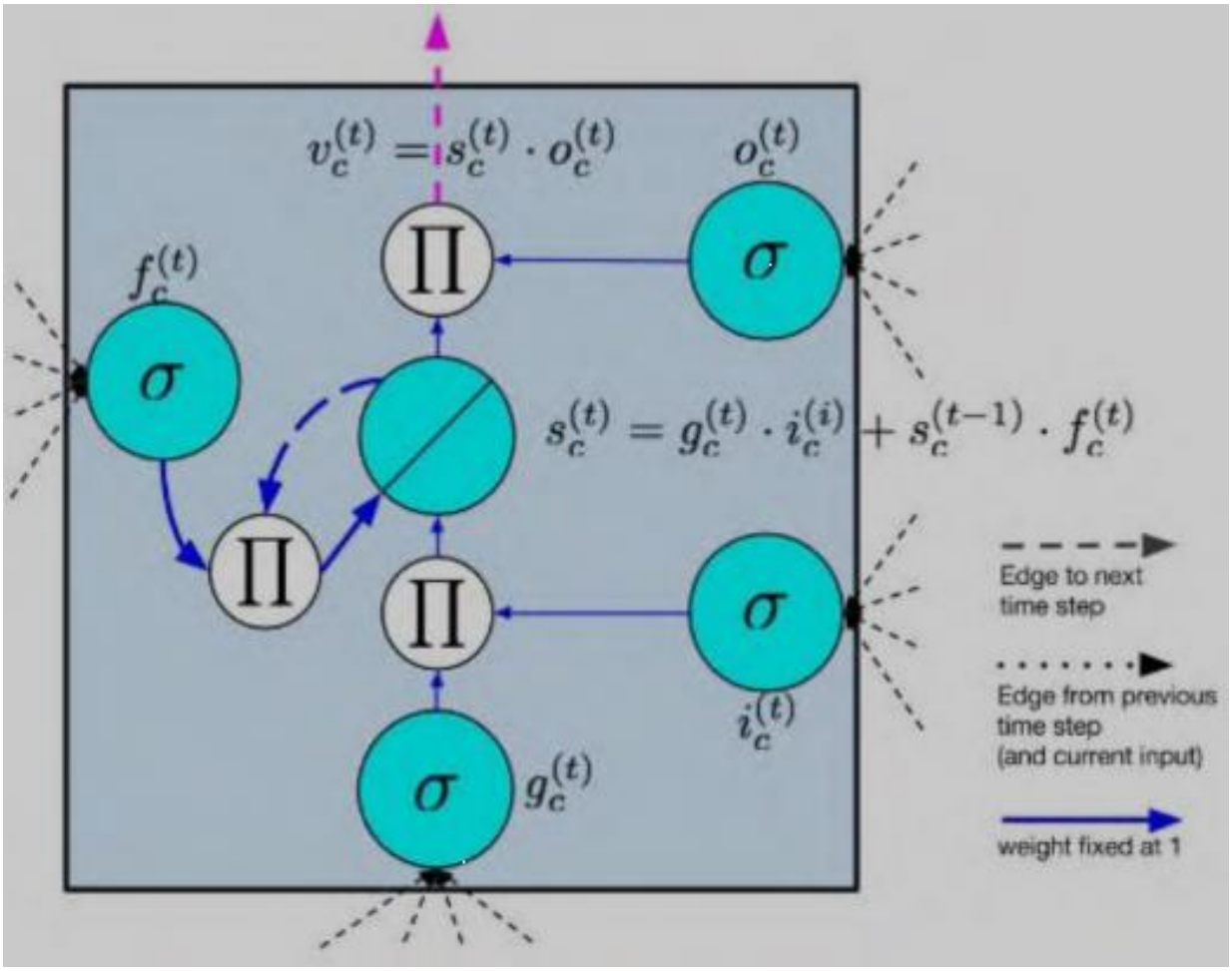



**Figure 2.** Memory unit of LSTM.

In the Cell, is the input node, is the input gate, is the forgetting gate, is the output gate, is the internal state node, and the input is the output of the current input and the internal state node at the previous time point after being filtered by the input gate. $g_c i_c f_c o_c s_c$ The computation of cells in the LSTM layer can be expressed as the following grouping:

$$g^{(t)} = \varphi(W_{gx}x^{(t)} + W_{gx}h^{(t-1)} + b_g) \quad (2)$$

$$i^{(t)} = \sigma(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + b_i) \quad (3)$$

$$f^{(t)} = \sigma(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + b_f) \quad (4)$$

$$o^{(t)} = \sigma(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + b_o) \quad (5)$$

$$s^{(t)} = g^{(t)} \odot i^{(t)} + s^{(t-1)} \odot f^{(t)} \quad (6)$$

$$h^{(t)} = s^{(t)} \odot o^{(t)} \quad (7)$$

Where W and b are the coefficient and bias of linear relationship, are TANh activation function, are Sigmod activation function, and $\odot$ represent dot product. φσBecause memory units in LSTM can have good effects on processing time series data, the LSTM model is selected to process stock price data in this paper.

### 2.4. LSTM neural network based on Chebyshev-Stacking and distance covariance

This paper is a research on the prediction of high-frequency stock prices, but considering the non-stationarity of high-frequency stock price series, LSTM neural network is adopted to replace the

traditional ARIMA model. The known first N period stock prices are used as the training set (X), and the N +1 period stock prices (Y) that need to be predicted are used as the validation set. However, as the stock prices as the time point data lack the overall smooth function feature, the Chebyshev orthogonal polynomial basis expansion is used to extract the function feature information of the high-frequency stock price series.

Secondly, the improved principle in this paper is to use an additive model (GAM) to predict, specifically, the function feature information of high-frequency stock price series is used to predict the information (residual sequence) that is not captured by LSTM. The specific prediction framework is the Stacking method. In addition, the model averaging method based on distance covariance is used to optimize how to determine the number of predictive variables of LSTM. [11]The distance correlation coefficient is calculated as follows:[12]

$$\hat{d}corr(u,v) = \frac{\hat{d}\,\mathrm{cov}(u,v)}{\sqrt{\hat{d}\,\mathrm{cov}(u,u)\hat{d}\,\mathrm{cov}(v,v)}} \tag{8}$$

$\hat{d}\,\mathrm{cov}^2(u,v) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$, $\hat{S}_1$, $\hat{S}_2$, $\hat{S}_3$ respectively is:

$$\hat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|u_i - u_j\right\|_{d_u}\left\|v_i - v_j\right\|_{d_u} \tag{9}$$

$$\hat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|u_i - u_j\right\|_{d_u}\frac{1}{n^2}\left\|v_i - v_j\right\|_{d_u} \tag{10}$$

$$\hat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\left\|u_i - u_l\right\|_{d_u}\frac{1}{n^2}\left\|v_i - v_l\right\|_{d_u} \tag{11}$$

Finally, the mean square error, absolute error, and relative error were compared with the original LSTM model. We call this method the LSTM neural network based on Chebyshev-Stacking and distance covariance weighting, and the specific process is shown in Figure 3 below[13]:

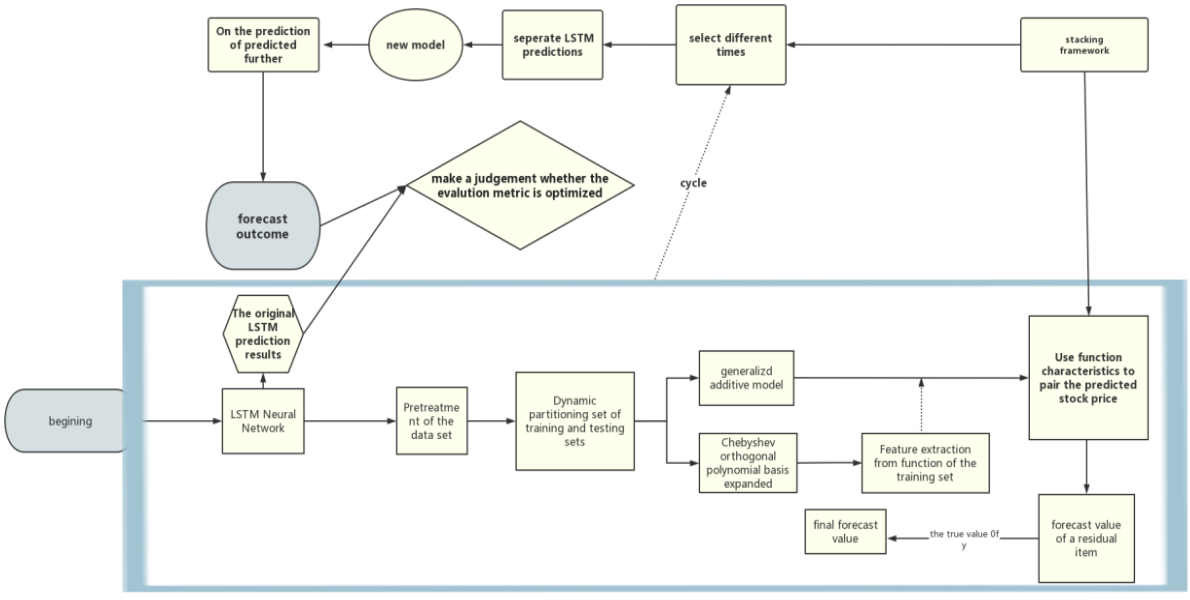



**Figure 3.** Model Construction.

# 3. Data Analysis

## 3.1. Data Exploration

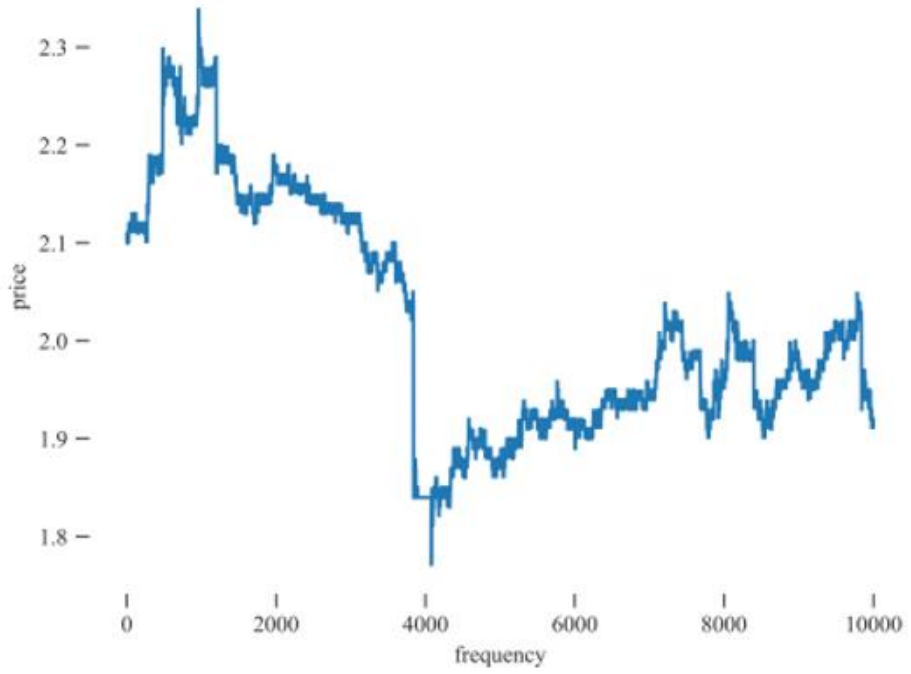
The stocks are randomly selected through the Wind database, and we use the daily trading price data of three of the randomly selected stocks to construct the dataset for subsequent experiments. These three stocks belong to different industries, among which stock code SH600777 Xinchao Energy belongs to the energy industry, stock code SH600811 Dongfang Group belongs to the food industry, and stock code SH603833 Oppai Home is in the furniture manufacturing industry. Since the stock price data belongs to a stochastic system, reflecting the market economic data and the trading dynamics, various factors and there influence on stock prices are complex internal patterns. We want to investigate whether our proposed method has advantages for stock price prediction in different situations when the stochastic system is subject to varying levels of complexity. Therefore, we randomly select stocks to avoid the influence of other factors. And we will test the model through data analysis to test our method's accuracy and the operation's specific feasibility. How to objectively evaluate the accuracy of a model needs to introduce three indicators: mean relative error (MRE), mean square error (MSE), posterior error (BE), and their corresponding calculation formulas are as follows:

$$MRE = \frac{1}{n}\sum_{k=1}^{n}\frac{|e_k|}{Y_k} \tag{12}$$

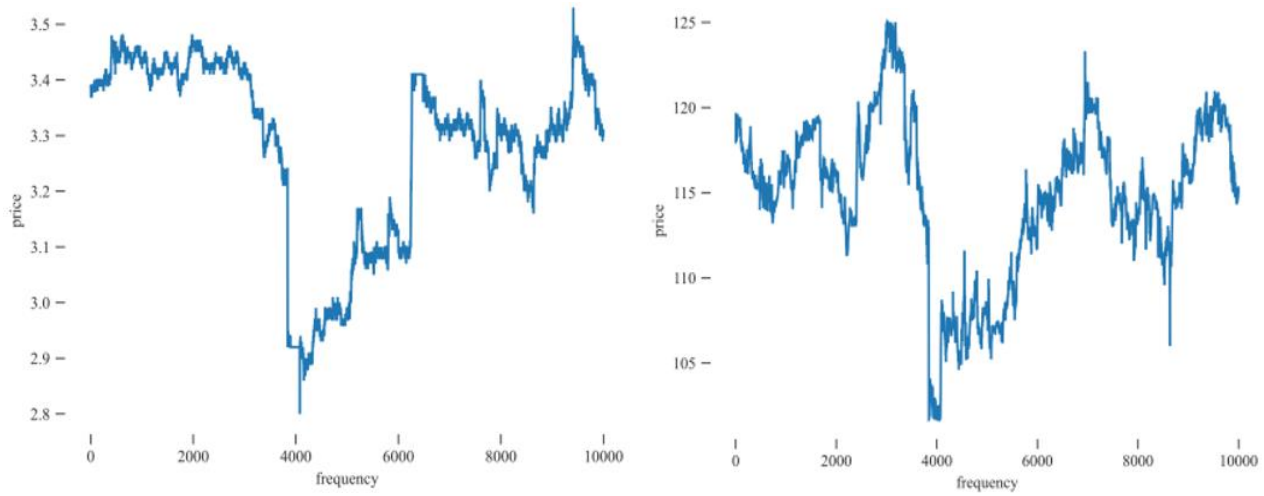
$$MRE = \frac{1}{n}\sum_{k=1}^{n}\frac{|e_k|}{Y_k} \tag{13}$$

$$MRE = \frac{1}{n}\sum_{k=1}^{n}\frac{|e_k|}{Y_k} \tag{14}$$

Where $e_k$ is the residual sequence, $Y_k$ is the true value, $S_1$ is the standard deviation of the original sequence, $S_2$ is the standard deviation of the relative value sequence. The smaller the three indicators of the model, the higher the prediction accuracy. The following Figure 4 is the particular trending charts of the three stocks.




(a)




(b) (c)

**Figure 4.** (a b, c) Stock price trend of SH600811, SH600777, SH603833.

### 3.2. Comparison of results

In the process of building the model, we used Python as the computational language for the model construction, the set of hyperparameters of the LSTM neural network is [2, 3, 4, 5, 6].[14] Based on high-frequency trading data, we select the time points at 10,000, we divided the data set into different training and testing sets according to 6000, 6500, 7000, 7500, 8000, 8500, and 9000 to test the robustness of the model. The specific results are shown in the following three tables and Figure 5:

**Table 1.** Comparison results on SH600777.

| | SMLSTM-MSE | SMLSTM-MRE | SMLSTM-BE | LSTM-MSE | LSTM-MRE | LSTM-BE |
|---|---|---|---|---|---|---|
| 6000 | 0.00006 | 0.00239 | 0.06338 | 2.52174 | 0.79119 | 0.77667 |
| 6500 | 0.00006 | 0.00257 | 0.06012 | 2.51359 | 0.78992 | 0.77994 |
| 7000 | 0.00004 | 0.00214 | 0.05117 | 2.51184 | 0.78965 | 0.78804 |
| 7500 | 0.00003 | 0.00192 | 0.04831 | 2.52095 | 0.79109 | 0.75028 |
| 8000 | 0.00003 | 0.00181 | 0.04378 | 2.55443 | 0.79631 | 0.68575 |
| 8500 | 0.00003 | 0.00172 | 0.04169 | 2.54303 | 0.79454 | 0.72271 |
| 9000 | 0.00002 | 0.00159 | 0.03796 | 2.54335 | 0.79458 | 0.72440 |

**Table 2.** Comparison results on SH600811.

| | SMLSTM-MSE | SMLSTM-MRE | SMLSTM-BE | LSTM-MSE | LSTM-MRE | LSTM-BE |
|---|---|---|---|---|---|---|
| 6000 | 0.00010 | 0.00179 | 0.05806 | 6.91464 | 0.80506 | 0.38562 |
| 6500 | 0.00005 | 0.00140 | 0.04275 | 6.93062 | 0.80598 | 0.36953 |
| 7000 | 0.00005 | 0.00134 | 0.04223 | 6.91739 | 0.80521 | 0.35803 |
| 7500 | 0.00004 | 0.00126 | 0.03997 | 6.89179 | 0.80374 | 0.37124 |
| 8000 | 0.00004 | 0.00120 | 0.03774 | 6.87198 | 0.80258 | 0.36996 |
| 8500 | 0.00004 | 0.00115 | 0.03668 | 6.93068 | 0.80599 | 0.36354 |
| 9000 | 0.00003 | 0.00111 | 0.03528 | 6.81748 | 0.79944 | 0.39970 |

**Table 3.** Comparison results of SH603833.

| | SMLSTM-MSE | SMLSTM-MRE | SMLSTM-BE | LSTM-MSE | LSTM-MRE | LSTM-BE |
|---|---|---|---|---|---|---|
| 6000 | 0.07115 | 0.00138 | 0.05660 | 1.3060E+4 | 0.99522 | 0.95767 |
| 6500 | 0.06296 | 0.00130 | 0.05330 | 1.3061E+4 | 0.99528 | 0.95833 |
| 7000 | 0.06162 | 0.00126 | 0.05273 | 1.3062E+4 | 0.99521 | 0.95768 |
| 7500 | 0.05791 | 0.00122 | 0.05111 | 1.3059E+4 | 0.99520 | 0.95734 |
| 8000 | 0.05420 | 0.00116 | 0.04945 | 1.3060E+4 | 0.99521 | 0.95790 |
| 8500 | 0.04726 | 0.00110 | 0.04618 | 1.3061E+4 | 0.99525 | 0.95840 |
| 9000 | 0.04235 | 0.00106 | 0.04372 | 1.3060E+4 | 0.99521 | 0.95746 |

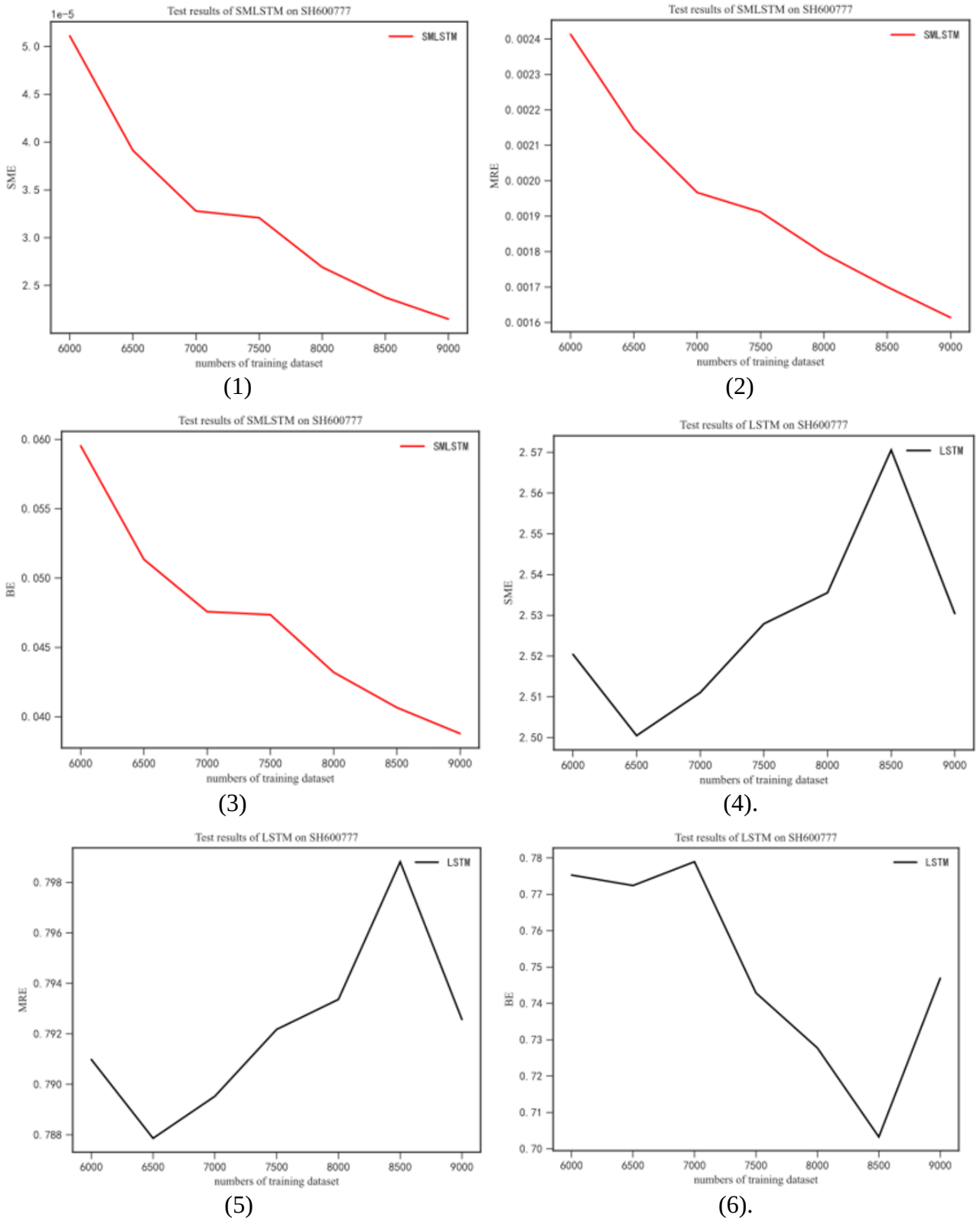



**Figure 5.** (1, 2, 3, 4, 5, 6) SME, MRE and BE in SMLSTM and LSTM for SH600777.

From the experimental results, the proposed method performs better than the original LSTM in terms of mean square error, relative absolute error and posterior error. And in some experiments, it was three or four orders of magnitude lower than the original LSTM. In addition, it can be seen from Figure 2-4 that the performance of the original LSTM shows a certain fluctuation under the condition of different number of samples in the training set. On the contrary, the error of the improved method consistently decreases with the increase of the number of samples, which indicates that the improved method is more robust than the original LSTM. Proposed method in the experiment of cosco than the

original LSTM high precision because of using the additive model and average method, effectively balance the prediction model of variance and deviation, although most of the error is under the condition of training set and testing set and calculated, but it can also demonstrate the proposed method can well time sequence curve fitting. In conclusion, combined with Figure 6 and Figure 7, it can be seen that the proposed method can effectively improve the prediction accuracy of the original LSTM and has certain robustness.

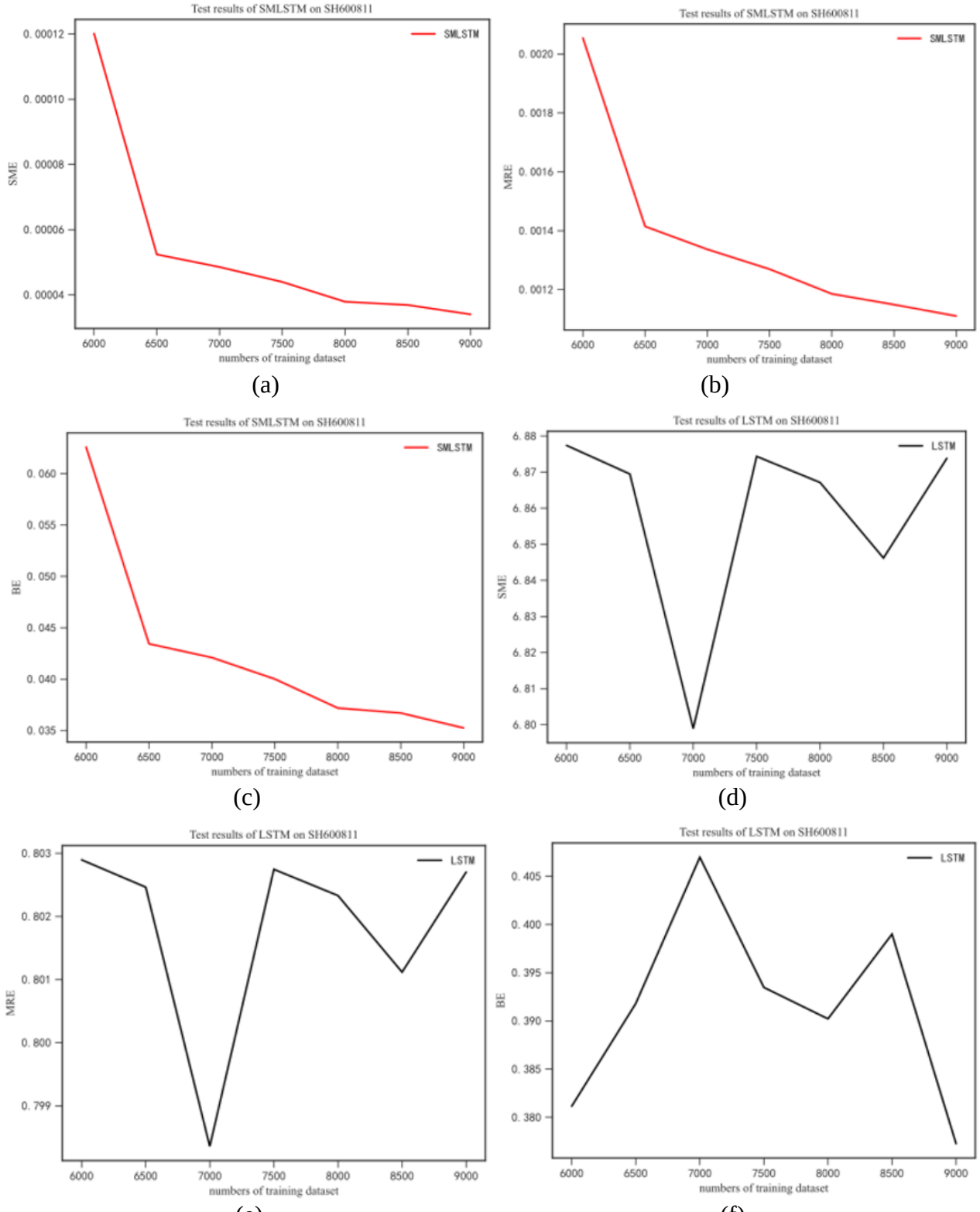



**Figure 6.** SME, MRE and BE in SMLSTM and LSTM for SH600811.

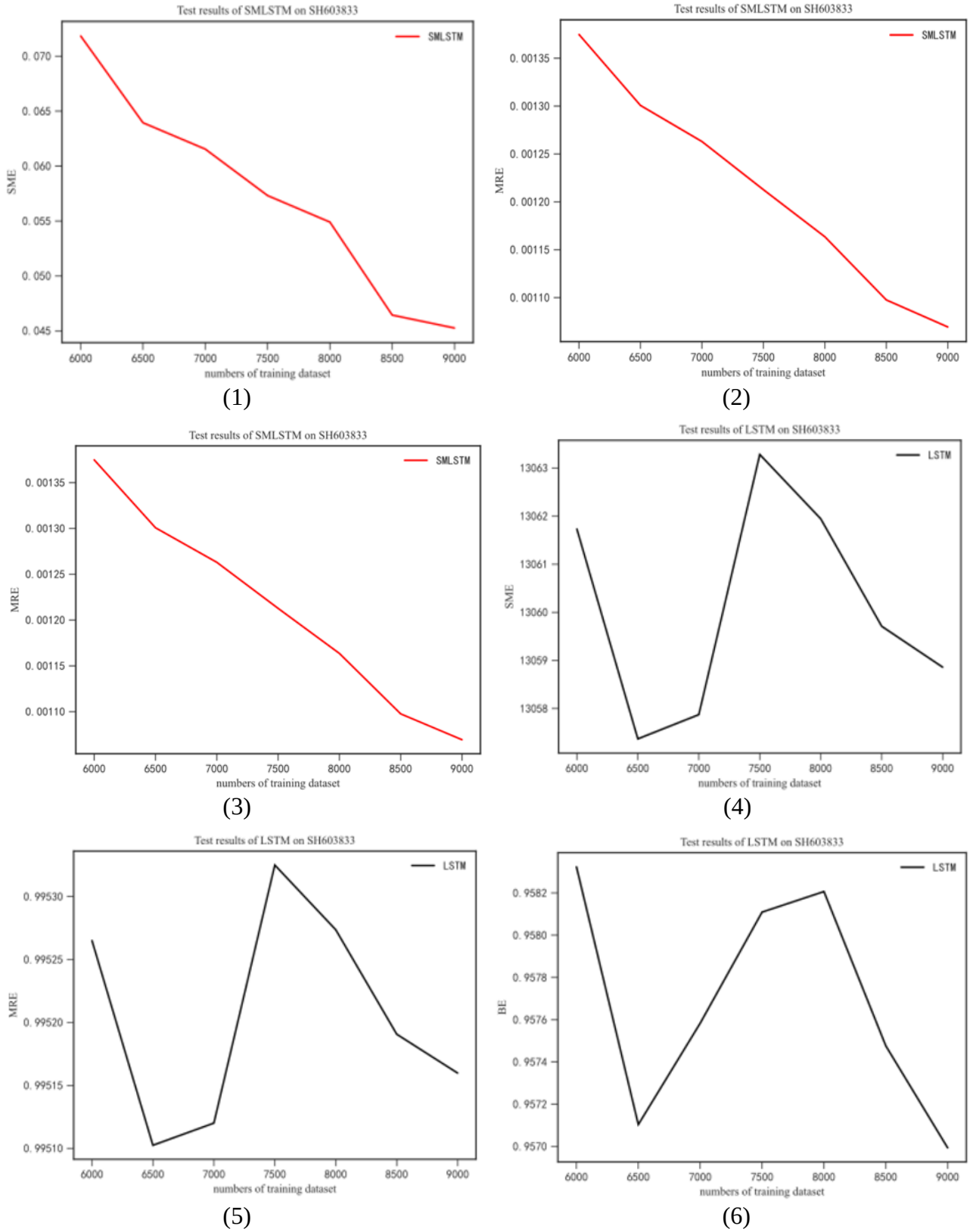



**Figure 7. (**1, 2, 3, 4, 5, 6**)** SME, MRE and BE in SMLSTM and LSTM for SH603833.

## 4. Conclusion

In this paper, a prediction model based on Chebyshev-Stacking and a weighted LSTM neural network is proposed. The proposed method extracts the function information of high-frequency time series while reducing the dimensionality. The potential model structure of Chebyshev orthogonal

polynomial base expansion coefficient vectors and response variables was captured by the Stacking framework and the data noise was reduced. In addition, the model selection problem of the LSTM neural network is solved by the model averaging method based on distance covariance weighting. For the future research contents: first, it is necessary to continue to explore the influence of different basis function expansion on the prediction results; Second, the utility of the proposed method in predicting traffic flow and air quality; Third, some stock fundamental data can be integrated into the prediction model to make the prediction results more accurate. Therefore, how to design an information fusion framework is worthy of further research.

## References

[1] Zai-jun wang. Using the base function expansion method solving one-dimensional finite deep square potential well model [J]. Journal of college physics, 2016, 35 (08): 39-43.

[2] Yang yuyuan, zhang mei. Empirical analysis of stock prices based on ARIMA model [J]. Science and technology information,2021,19(29):121-123+127.

[3] Li Li-Ping, ZENG Li-fang, JIANG Shao-ping, HE Wen-Qian. Stock Price Prediction based on LSTM Neural network [J/OL]. Journal of Yunnan Minzu University (Natural Science Edition):1-10[2022-07-07 18:48].

[4] Lin Xin, Zhu Xiaodong. LSTM stock price prediction model based on Attention mechanism [J]. Journal of chongqing university (natural science edition), 2022, 33 (02) 6:75-82.

[5] Huang Yucheng, Fang Weiwei. Research on stock price forecast based on LSTM network [J]. Modern computer,2021,27(34):51-55+60.

[6] Peng Yan, Liu Yuhong, Zhang Rongfen. Modeling and analysis of stock price prediction based on LSTM [J]. Computer engineering and applications,2019,55(11):209-212.

[7] Sun Bingjie, Tang Rui, Zuo Yi, Huang Minghe. Research on neural network stock prediction based on wavelet analysis [J]. Computer & digital engineering,2016,44(06):1031-1034+1106.

[8] ZHANG M M. Research on price forecasting of Shanghai Composite Index based on LASSO dimension reduction, LSTM and mixing model [D]. Donghua University,2021.

[9] Xie Xinrui, Lei Xiuren, Zhao Yan. Application of MI and improved PCA dimensionality reduction algorithm in stock price prediction [J]. Computer engineering and applications,2020,56(21):139-144.

[10] Zou Jie, Li Lu. Research on Stock Price Prediction of RF-SA-GRU Model [J/OL]. Computer Engineering and Applications :1-20[2022-09-1913:54].

[11] Li Guoyu, Zhang Jian, Meng Yongliang. Fault diagnosis method of Subway power supply System based on Bagging algorithm [J]. Automation Technology and Application,202,41(09):110-112.

[12] Zhu Mingmin, Liu Sanyang. A Sensitivity Analysis method for covariance Matrix of Gaussian Networks Based on Improved Bhattacharyya Distance [J]. Journal of zhejiang university (natural science),2019,46(01):9-14+21.

[13] Ge Yang, Ma Jia-xin, REN Yong, QIN Jian-Cong. Mechanical and electrical equipment based on improved LSTM conditions remaining life prediction [J]. Journal of changshu institute of technology, 2022, 4 (5): 65-72.

[14] Ding Jie. Study on Web Development and Application of Python Script Language [J]. Digital Communication World,2021, (10):163-164.