Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 5803407, 12 pages
https://doi.org/10.1155/2020/5803407

Research Article

Improving Deep Learning for Forecasting Accuracy in Financial Data

Shih-Lin Lin[1,2] **and Hua-Wei Huang**[3]

[1]*Department of Mechanical Engineering, Cheng Shiu University, Kaohsiung 83347, Taiwan*
[2]*Executive Master of Business Administration (EMBA), National Cheng Kung University, Tainan 701, Taiwan*
[3]*Department of Accountancy, National Cheng Kung University, Tainan 701, Taiwan*

Correspondence should be addressed to Shih-Lin Lin; 0642@gcloud.csu.edu.tw

Received 22 November 2019; Revised 24 January 2020; Accepted 21 February 2020; Published 26 March 2020

Academic Editor: Paolo Renna

Financial forecasting is based on the use of past and present financial information to make the best prediction of the future financial situation, to avoid high-risk situations, and to increase benefits. Such forecasts are of interest to anyone who wants to know the state of possible finances in the future, including investors and decision-makers. However, the complex nature of financial data makes it difficult to get accurate forecasts. Artificial intelligence, which has been shown to be suitable for analyzing very complex problems, can be applied to financial forecasting. Financial data is both nonlinear and nonstationary, with broadband frequency features. In other words, there is a large range of fluctuation, meaning that predictions made only using long short-term memory (LSTM) are not enough to ensure accuracy. This study uses an LSTM model for analysis of financial data, followed by a comparison of the analytical results with the actual data to see which has a larger root-mean-square-error (RMSE). The proposed method combines deep learning with empirical mode decomposition (EMD) to understand and predict financial trends from financial data. The financial data for this study are from the Taiwan corporate social responsibility (CSR) index. First, the EMD method is used to transform the CSR index data into a limited number of intrinsic mode functions (IMF). The bandwidth of these IMFs becomes narrower, with regular cyclic, periodic, or seasonal components in the time domain. In other words, the range of fluctuation is small. LSTM is a good way to forecast cyclic or seasonal data. The forecast result is obtained by adding all the IMFs together. It has been verified in past studies that only the LSTM and LSTM combined with the EMD can be used. The analytical results show that smaller RMSEs can be obtained using the LSTM combined with EMD compared to real data.

1. Introduction

Recently, artificial intelligence has had a great impact on the global business environment and has found applications in many different fields. Financial data is challenging to analyze because it possesses a lot of uncertainty. Artificial intelligence can be used to classify financial data for analysis, allowing the team to screen and analyze data more quickly and to help them make more precise decisions, significantly reducing human error, bring better returns to customers, make more accurate predictions of possible outcomes, and facilitate risk control. Financial forecasting is an important part of financial data analysis. A variety of effective analytical methods have been proposed for this purpose, including short-term prediction methods such as regression analysis, exponential smoothing, and autoregressive moving average models [1–6]. Research on recurrent neural networks (RNN), one of the effective and popular artificial intelligence methodologies, began in the late 1980s, but, with recurrent neural networks, one needs to train millions of parameters, which was difficult to accomplish at that time. However, with the development of optimization methods and parallel computing in recent years, computers now have the ability to complete the training of millions of parameters, which has once again made loopy neural nets such as RNN a hot topic. Mikolov et al. and Wu et al. proposed and applied a language model based on RNN, which has achieved great success in the field of natural language processing (NLP) [7, 8]. Lipton

et al. adapted the RNN approach for speech and handwriting recognition [9], achieving an improved accuracy rate of 20% over previous results. Bengio et al. [10] found that as the memory of the historical state gradually increased, the problems of gradient disappearance and gradient divergence occurred. Their findings indicate that this type of neural network can only memorize the transient historical state. Long short-term memory (LSTM) [11] is based on using recurrent neural network, with the addition of three gate control structures, to solve the problem of gradient disappearance, thus allowing the training of a neural network model with a longer period of memory.

Generally, support vector machine (SVM) and extreme learning machine (ELM) classifiers are suitable for use in classification case studies requiring high classification accuracy [12, 13]. Predictions can be made based on financial training data to determine whether trends will go up or down but not how much the rise or fall will be. However, SVM has the following disadvantages [12]: (1) the SVM algorithm is difficult to implement for large-scale training samples because quadratic programming is used to solve the support vector, which involves the calculation of an m-th order matrix. If the m number is large, a lot of memory and computing time will be consumed for storage and calculation of the matrix. (2) SVM also has difficulty solving multiclassification problems. The classical SVM algorithm can only handle two-class classification. However, for practical applications of data collection, it is generally necessary to solve multiclass classification problems. It is difficult to analyze little data. When there is little training data, the accuracy of the training results is very high but the accuracy of the actual prediction result is very low, and the accuracy of the secondary prediction will be different. ELM reduces the complexity behind feedforward networks by generating sparse, randomly connected hidden layers. It requires less computation time, but its actual performance depends on the different tasks and data [13]. Due to the aforementioned problems, in this paper, we propose using deep learning (the LSTM model) to improve the prediction results based on financial data. The model contains multilayer networks including the input layer, output layer, hidden layer, and countless neurons and nonlinear excitation functions. It is hoped that the accuracy of the prediction can be improved using this approach. Some deep learning research methods are introduced below.

Proportional conjugate gradient backpropagation is a network training function where weights and bias values are updated according to the proportional conjugate gradient method. It can train any network as long as its weight, net input, and transfer functions have derivative functions. Backpropagation is used to calculate derivatives of performance with respect to the weight and bias variables; see Møller [14] for a more detailed discussion of the scaled conjugate gradient algorithm. The long short-term memory method is a special RNN model originally proposed to solve the problem of gradient divergence encountered with the RNN model. In the traditional RNN, the training algorithm uses backpropagation through time (BPTT). When the time is long, the residuals that need to be returned will decrease exponentially, resulting in slow network weight update, which cannot reflect the long-term memory effect of RNN. Therefore, a storage unit is needed to store the memory. The LSTM model was proposed to alleviate this problem; for related theory, please refer to [11].

In 1998, Huang proposed the Hilbert-Huang transformation, which has since received extensive attention from the academic community. This is an effective method for analyzing nonlinear, nonstationary time series, including the empirical modal decomposition method and Hilbert transform [15]. Zhang et al. [16] used the ensemble empirical mode decomposition (EEMD) method to study and analyze changes in and the characteristics of international crude oil prices, decomposing them into short-term fluctuations, medium-term fluctuations, and long-term trends. Wang et al. [17] studied the EMD-HW bagging method based on empirical mode decomposition, moving block bootstrap and Holt-Winter forecasting. Guhathakurta et al. [18] applied the EEMD method to analyze the relationship between the Indian stock market and the exchange rate and concluded that the impact models of the stock market and the exchange rate market are similar. Khalid et al. [19] found that the empirical mode decomposition method could replace the mean square error and the mean absolute error criterion of all other models for stock market returns and direction prediction. Islam et al. [20] applied the EMD method for the decomposition of data sequences, in comparison with the wavelet decomposition method, finding the EMD method to be more effective. Fang [21] applied the EEMD technique for analysis of the psychological state of investors in their study of the relationship between stock prices and investor psychology. Recently, an integrated approach using multiple models has been used for better performance in prediction problems [22–30]. For example, it was found that wind speed can be more accurately predicted by combining EMD and different prediction techniques [24]. Liu et al. proposed a neural-network-based EMD hybrid wind speed predictor, in which each IMFS/residue component is trained using the appropriate backpropagation techniques [26]. In Ren et al. [27], a combination of support vector regression (SVR) and EMD was used for accurate wind energy prediction. All the above studies show that the application of EMD technology for wind speed prediction improves the overall accuracy and prediction ability of conventional methods.

Financial data are classified as broadband data in the frequency domain, which means that they contain a large range of fluctuations, so it is not enough just to use LSTM to make predictions. In this study, EMD is used to transform the raw nonlinear financial data into a limited number of intrinsic mode functions (IMF) and a residual. The bandwidth of these IMFs becomes narrower, with regular cyclic, periodic, or seasonal components in the time domain. LSTM is a good way to predict cycles, periods, or seasonality. The prediction result is obtained by adding all the IMFs together. Compared with using only the LSTM, the EMD-LSTM can reduce the root-mean-square-error (RMSE) compared with real data. The paper contributes to a deeper understanding of the application of deep learning combined with EMD for

financial data forecasting, significantly increasing the accuracy of the prediction models.

2. Research Theory, Data, and Methodology

2.1. Description of the Research Data. In the 21st century, corporate social responsibility (CSR) has become a factor that enterprises must take into consideration to ensure sustainable operations. Broadly speaking, CSR means that, in addition to pursuing the best interests of stockholders, companies must also take into account the interests of other stakeholders, including employees and consumers, suppliers, the community at large, and environmental concerns. Concern with CSR emerged during the extremely prosperous period of industrial development in the 20th century. When a developed country reaches a certain level of maturity in terms of industrial and commercial development, the population at large and the company begin to think about the relationships between the enterprise and the environment, community, labor force, and so forth. With the growth in economic globalization and the continued expansion of multinational corporations since the 1980s, labor relations in several countries have become extremely unbalanced. The protection of the rights and interests of labor has become a social issue of global concern, and the question of social responsibility has become more important. In western countries, where the labor force is often overseas, especially the United Kingdom and the United States, some have begun to argue that the cost and responsibility on government for social welfare should be reduced and that business enterprises should bear more of this social responsibility. The corporate social responsibility movement was initiated in developed European and American economies and gradually evolved into a worldwide trend.

Given the global emphasis, more and more companies are paying attention to CSR, demonstrating better external development than nonexecutive companies, with relatively good financial performance. Corporate governance includes mechanisms for guiding and managing enterprises and how they implement the responsibilities of business operators to protect the legitimate rights and interests of shareholders while also taking into account the interests of other stakeholders. Generally speaking, corporate governance mechanisms are divided into two types, external and internal. External governance refers to the promotion of private profits and protection of shareholders' rights through the actions of government, judicial units, and external market forces. On the other hand, internal governance aims to achieve operational objectives through the ownership structure and the functions of the board of directors and management, for the best interests of the company and all shareholders, to assist in the management of the company, and to provide effective monitoring mechanisms.

The "Taiwan Corporate Governance 100 Index" lists companies on the Taiwan Stock Exchange (including domestic listed companies and first listed companies of foreign companies, excluding Taiwan Depository Receipts). The constituent "indexes" have been selected through several quantitative criteria as outlined below (https://cgc.twse.com.tw):

(1) Sample parent company: public shares listed for public offering.

(2) Liquidity: delete stocks with a minimum average trading amount of 20% in the most recent year.

(3) Results of corporate governance evaluation: select the top 20% based on the company's corporate governance evaluation results.

(4) Financial indicators and necessary conditions: the net value per share required at the end of the previous year shall not be less than the denomination.

(5) Calculation of the market value weighting method.

First, stocks with a minimum daily trading amount of 20% during the most recent year are deleted. Then, stocks that meet the liquidity test standards for the sample are selected, that is, the stocks that meet the "20% of the results of the recent 1-year corporate governance evaluation" and "the net value of the net income per share at the end of the previous year must not be less than the denomination." Then, they are ranked according to "after-tax net profit in the most recent year" and "revenue growth rate in the most recent year," and the respective top rankings are sorted from small to large. The top 100 stocks are selected as constituent stocks. In other words, there is not a single factor list for "Corporate Governance Assessment." The "Corporate Governance 100 Index" is subject to review in July each year. It is subject to liquidity inspection, corporate governance evaluation and screening, and three financial indicators (net value per share not less than the denomination, after-tax net profit ranking, and revenue growth rate). The data for the Taiwan CSR index were sourced from https://www.taiwanindex.com.tw for the period from June 15, 2015, to December 12, 2018, to obtain a total of 863 datasets. There are daily data, five per week. The entire dataset is divided into two parts, with 90% of the total 779 datasets used for training (from June 15, 2015, to August 14, 2018) and the other 10% used for verification (a total of 84 datasets, from August 15, 2018, to December 12, 2018).

Next, the verified indicator root-mean-square-error (RMSE) is calculated using the following equation:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_1^n \left(\widehat{y}_t, -y_t\right)^2}{n}}, \tag{1}$$

where $\widehat{y}_t$ is the real data (verification); y_t is the prediction data. The smaller the RMSE, the closer the prediction data is to the real data (verification), and the larger the RMSE, the greater the difference between the predicted data and the real data (verification).

2.2. Long Short-Term Memory. The long short-term memory model is a special RNN model that is proposed to solve the problem of gradient divergence encountered with the RNN model. In the traditional RNN, the training algorithm uses backpropagation through time (BPTT). When the time is long, the residuals that need to be returned will decrease

exponentially, resulting in slow network weight update, which cannot reflect the long-term memory effect of RNN. Therefore, a storage unit is needed to store the memory. The LSTM model is proposed to alleviate this problem. For a discussion of the related theory, please refer to [11].

2.3. Empirical Mode Decomposition. The Hilbert-Huang transform is a new tool for nonstationary data analysis. Financial data is nonlinear, nonstationary, and complex and has no rules. After EMD is used for decomposition into multiple IMF bases, each IMF can be decomposed to discover inherent laws that are hidden in the data; for the related theory, please refer to [15]. The sifting procedure starts with the identification of the neighborhood minima and maxima of a periodic arrangement $X(t)$. First, recognize all the nearby maxima; then, interface with them with a cubic spline line to frame the upper envelope $e_u(t)$. Rehash the methodology for the nearby minima to deliver the lower $e_l(t)$. The local mean can be determined by

$$m_1(t) = \frac{e_u(t) + e_l(t)}{2}. \tag{2}$$

The mean is assigned in (2), and the contrast between the data and $m_1(t)$ in the main part is acquired by the accompanying condition:

$$h_1(t) = x(t) - m(t). \tag{3}$$

In the consequent sifting process, $h_1(t)$ is viewed as the information:

$$h_{1(k-1)}(t) - m_{1k}(t) = h_{1k}. \tag{4}$$

The EMD can rehash this sifting system k times, until h_{1k} is an IMF. Now

$$c_1 = h_{1k}, \tag{5}$$

which is the primary IMF part obtained from the data. The standard deviation decides when to halt the sifting procedure. This can be revised by restricting the size of the standard deviation (SD), processed from two continuous sifting results as follows:

$$\mathrm{SD} = \sum_{t=0}^{T} \frac{\left|h_{k-1}(t) - h_k(t)\right|^2}{h_{k-1}(t)^2}. \tag{6}$$

At the point when the SD can be set somewhere in the range of 0.2 and 0.3, the primary IMF c_1 is acquired, which can be composed as follows:

$$X(t) - c_1 = r_1. \tag{7}$$

Note that the buildup r_1 still contains some helpful information. We can, in this manner, treat the buildup as new information and apply the above methodology to get

$$\begin{aligned} r_1 - c_1 &= r_2, \\ &\vdots \\ r_{n-1} - c_1 &= r_n. \end{aligned} \tag{8}$$

This technique should be rehashed until the last arrangement r_n conveys no oscillation data. The rest of the arrangement is the pattern of this nonstationary information $X(t)$. Combining (6) and (7) yields the EMD for the first sign:

$$X(t) = \sum_{j=1}^{n} c_j + r_n. \tag{9}$$

In this manner, one can decompose the information into n-empirical modes and buildup r_n, which can be either the mean pattern or a steady pattern. The IMFs $c_1, c_2, \ldots, c_n$ incorporate distinctive recurrence groups extending from high to low.

3. Results and Discussion

The up and down movement of Taiwan's Corporate Governance 100 Index is discussed below. Figure 1 shows Taiwan's CSR index. The statistical characteristics are as follows: the mean is 5412, the standard deviation is 600.1820, and the variance is 360220. This dataset contains a total of 863 points, from June 15, 2015, to December 12, 2018. The Y-axis shows the TW CSR index, that is, the TWSE Corporate Governance 100 Index. The start date is June 15, 2015, and the starting date index is 5000. After June 2015, the index fell, mainly due to the global stock market crash that occurred in August 2015. The biggest reason for this was that China's economic slowdown was worse than expected. The US Federal Reserve raised interest rates in September 2015 and international oil prices fell below $40, causing panic in international financial markets. In August 24, 2015, the Taiwan stock market fell 583 points, 7.5%, the biggest one-day drop in history, to the lowest level in 33 months. Taiwan's economy improved in January 2016, and Taiwan's CSR index also rose. According to the index Company statistics from December 2016 to December 2017 led to an increase in the Corporate Governance 100 Index by 18.39%, at a semiannual rate of 5.63%. Both increases exceed that of the simultaneous weighted index for December 2016 to December 2017 which was 15.99% at a semiannual performance rate of 4.96%. Companies showing good corporate governance were favored by investors. In October 2018, the Federal Reserve (Fed) continued to raise interest rates, while the International Monetary Fund (IMF) revised their global economic growth rate forecast for the next year. However, US economic growth has since slowed down, and the trade war that shows no sign of ending in the short term has caused the global stock market to fall. The US Dow Jones Industrial Average fell by more than 800 points and Asian stocks fell into a bear market. The Taiwan stock market-weighted index has fallen by 1,517 points since October 2018, and the TWSE CG 100 Index also fell. The statistical properties of this data are as follows: the mean is 5412, the standard deviation is 600.1820, and the variance is 360220, for a total of 863 points.

Proof of Taiwan's CSR index data is nonlinear and nonstationary [31]. The main characteristic of nonlinearity is that if there is a disproportionate relationship between the input and output for the equation describing a certain system, it is called nonlinear data. As can be seen in Table 1

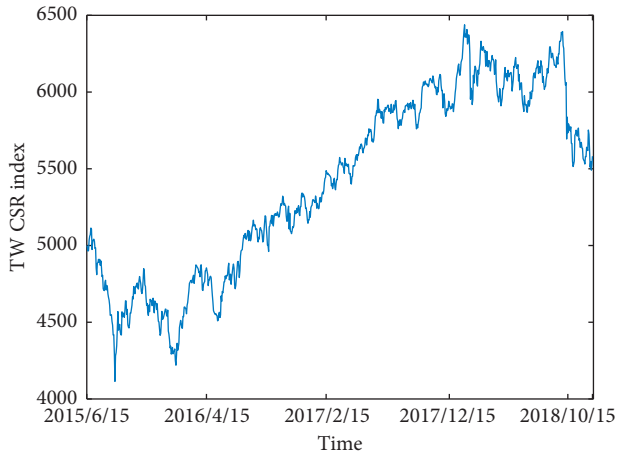

Figure 1: Taiwan CSR index. The statistical characteristics are as follows: the mean is 5412, the standard deviation is 600.182, and the variance is 360220, for a total of 863 datasets. There are 863 datasets from June 15, 2015, to December 12, 2018.

Table 1: Taiwan's CSR index.

Date	Taiwan's CSR index
2015/06/15	5000
2015/06/16	4990
2015/06/17	4964
2015/06/18	4968
2015/06/22	5035
2015/06/23	5061
2015/06/24	5071
2015/06/25	5114
2015/06/26	5107
2015/06/29	4974
2015/06/30	5024
2015/07/01	5036
2015/07/02	5042
2015/07/03	5029
2015/07/06	4981
⋮	⋮
2018/12/10	5491
2018/12/11	5524
2018/12/12	5582

and Figure 1, on June 15, 2015, Taiwan's CSR index is 7645, on June 16, 2015, Taiwan's CSR index is 8492, and on June 17, 2015, Taiwan's CSR index is 9338. This verifies the CSR index as nonlinear data.

The definition of stationarity is that the mean and variation are independent of the time point t. The values are the same at any point in time and can be expressed as follows:

$$m_x = m_{x(t)} = E[x_t],$$
$$\sigma_x^2 = E\left[(x_t - m_x)^2\right]. \tag{10}$$

Here, x_t is Taiwan's CSR index; m_x for the mean is 5412; and σ_x^2 is the variance. As can be seen in Figures 2(a) and 2(b) the monthly mean and the monthly variation of the Taiwan CSR index are different at each time point. Therefore, the Taiwan CSR index data can be defined as nonstationary.

This study uses LSTM regression networks for TW CSR index forecasting. The LSTM model parameter settings are as follows: the specified LSTM layer has 200 hidden units. First, the adaptive moment estimation optimizer is selected and trained for 250 periods. To prevent system divergence, the gradient threshold is set to 1. An initial learning rate of 0.005 is specified and the learning rate is reduced by multiplying it by a factor of 0.2 after 125 epochs.

The TW CSR index dataset includes a total of 863 datasets, from June 15, 2015, to December 12, 2018. All the data are divided into two parts, with 90% used for training (a total of 779 points, from June 15, 2015, to August 14, 2018). The other part is used for verification (10% or 84 points in total, from August 15, 2018, to December 12, 2018). The LSTM model parameters used for the regression are set as follows: the specified LSTM layer has 200 hidden units. First, the adaptive moment estimation optimizer is selected and trained for 250 periods. To prevent system divergence, the gradient threshold is set to 1. An initial learning rate of 0.005 is specified and the learning rate is reduced by multiplying it by a factor of 0.2 after 125 epochs. Figure 3 shows the LSTM's TW CSR index forecast results. Figure 4 shows the LSTM's TW CSR index forecast and the actual data verification results. The RMSE is 333.9627.

The decomposed data can reflect fluctuation information on different time scales while retaining the characteristics of the original data. The TW CSR index is first decomposed into short-, medium-, and long-term time series components. Here, there are six components, labelled IMF1 to IMF6. Figure 5 shows the results for IMF1. In terms of statistical characteristics, this is high-frequency data with an average period of 0.9279 weeks. The mean is −0.0390, the standard deviation is 30.2396, the variance is 914.4325, and the Pearson correlation coefficient is 0.0598. Figure 6 shows the LSTM prediction results for IMF1, and Figure 7 shows the prediction and actual verification results of LSTM for IMF1 with an RMSE of 2.7274. Figure 8 indicates the results for IMF2. Its statistical characteristics are as follows: the average period is 2.7839 weeks, the mean is 1.0316, the standard deviation is 41.5703, the variance is 1728.1, and the Pearson correlation coefficient is 0.0806. This is the second-highest-frequency data. Figure 9 shows the prediction results obtained with LSTM for IMF2, and Figure 10 shows the prediction and actual verification results obtained with LSTM for IMF2; the RMSE is 77.4748. Figure 11 shows the results for IMF3. The statistical characteristics are as follows: the average period is 7.3394 weeks, the mean is 3.2523, the standard deviation is 64.4050, the variance is 2.3327, and the Pearson correlation coefficient is 0.5110. Figure 12 shows the LSTM prediction results for IMF3, and Figure 13 shows the LSTM prediction and actual verification results for IMF3. The RMSE is 115.9812. IMF3 is comprised of intermediate frequency data. Figure 14 shows the results for IMF4. The

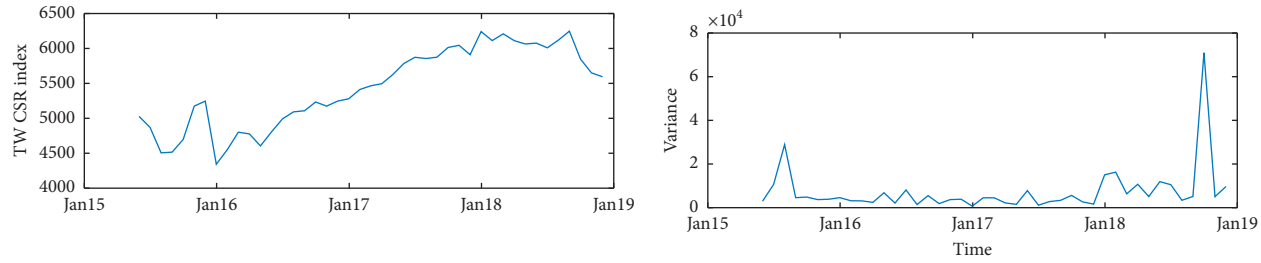

Figure 2: (a) Monthly mean of Taiwan CSR index; (b) monthly variance of Taiwan CSR index.

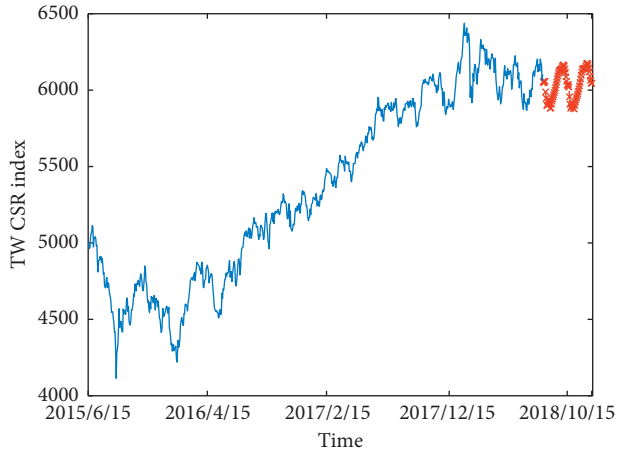

Figure 3: Taiwan CSR index as predicted by the LSTM model.

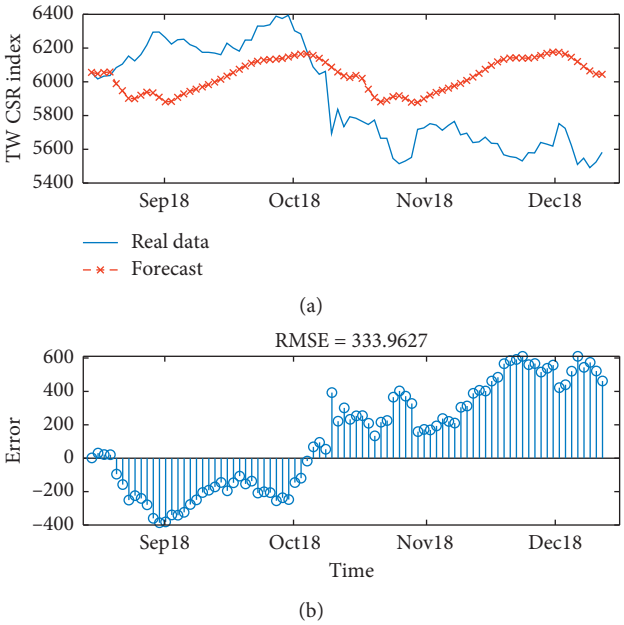

Figure 4: RMSE results from the LSTM model for Taiwan CSR index forecast and actual verification data. The RMSE is 333.9627.

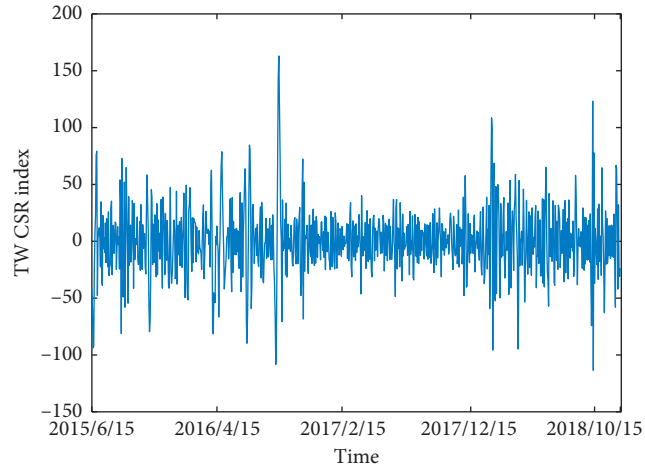

Figure 5: EMD analysis of the Taiwan CSR index used to get IMF1.

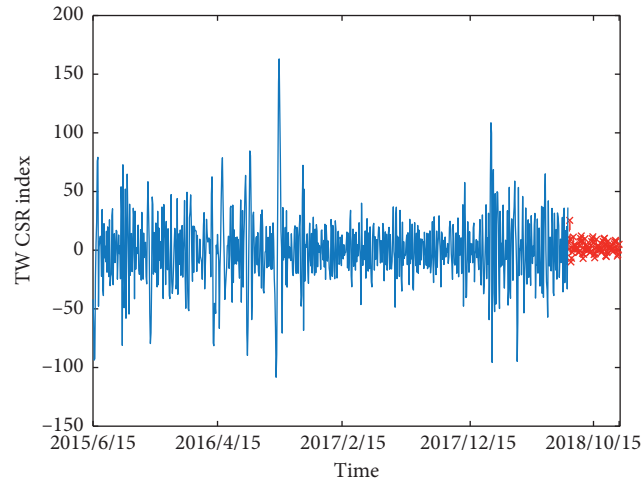

Figure 6: LSTM prediction results for IMF1.

statistical characteristics are as follows: the average period is 18.6308 weeks, the mean is 2.0298, the standard deviation is 79.5222, the variance is 6323.8, and the Pearson correlation coefficient is 0.1010. Figure 15 shows the prediction results obtained with LSTM for IMF4, and Figure 16 shows the prediction and actual verification results obtained with LSTM for IMF4. The RMSE is 51.8842. Figure 17 shows the IMF5 analysis. The statistical characteristics are as follows: the average period is 48.44 weeks, the mean is −25.8465, the standard deviation is 85.7134, the variance is 7346.8, and the Pearson

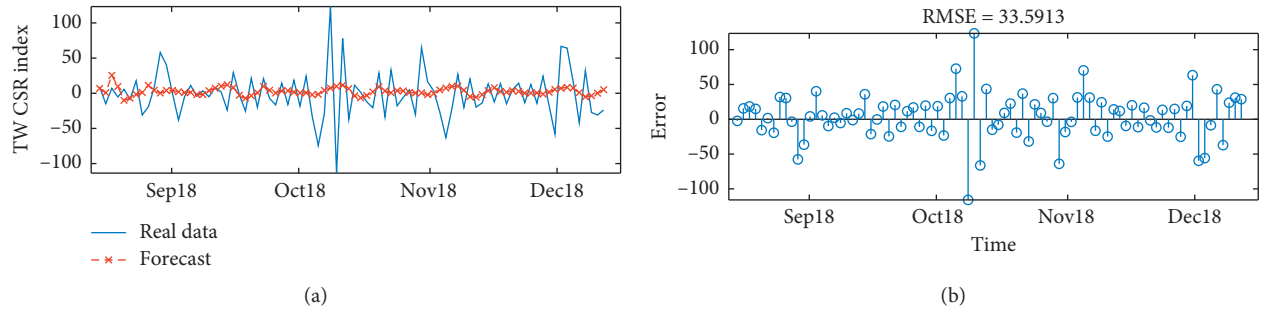

Figure 7: LSTM prediction and actual verification results for IMF1 with RMSE of 33.5913.

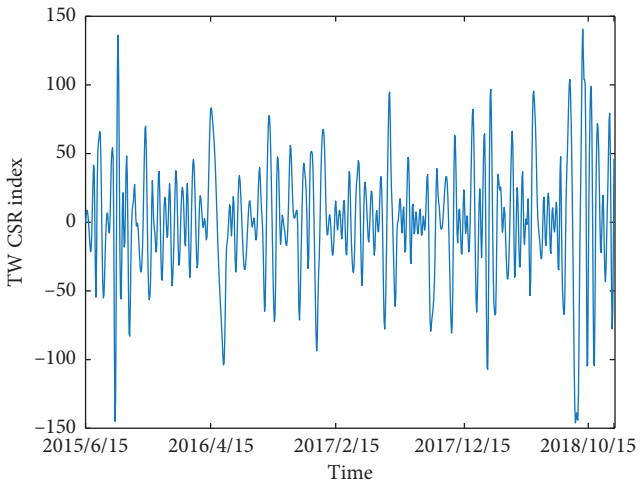

Figure 8: EMD analysis of the Taiwan CSR index used to get IMF2.

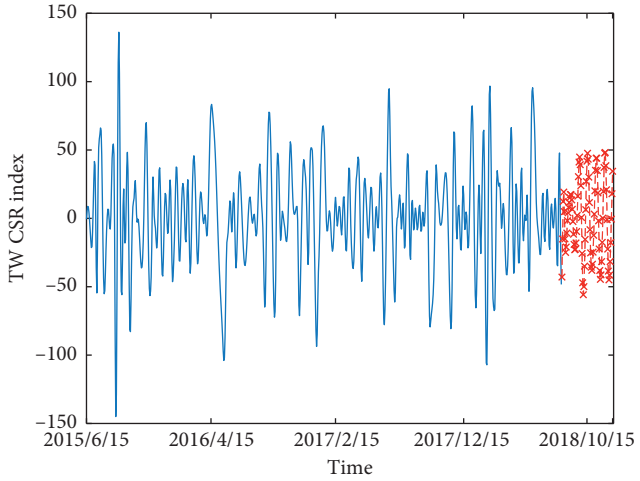

Figure 9: LSTM prediction results for IMF2.

correlation coefficient is 0.5485. Figure 18 shows the LSTM prediction results for IMF5, and Figure 19 shows the prediction and actual verification results obtained with LSTM for IMF5. The RMSE is 35.5218. Figure 20

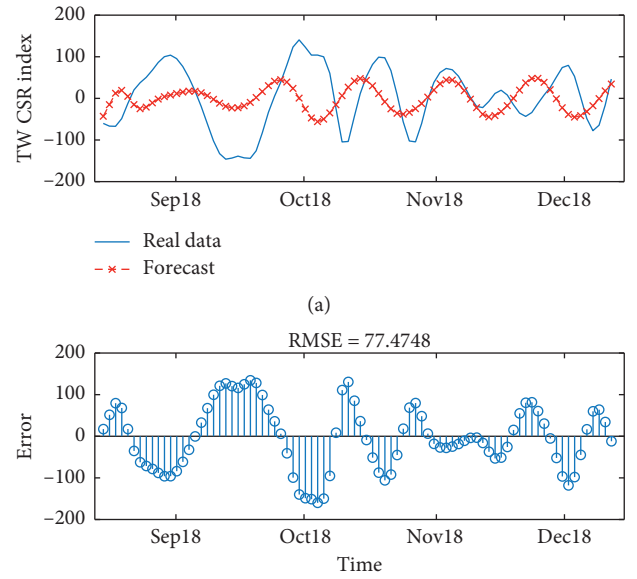

Figure 10: LSTM prediction and actual verification results for IMF2 with RMSE of 77.4748.

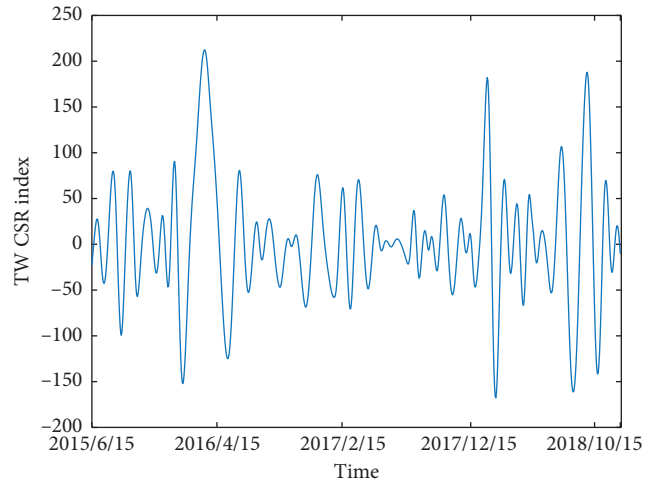

Figure 11: EMD analysis of the Taiwan CSR index to get IMF3.

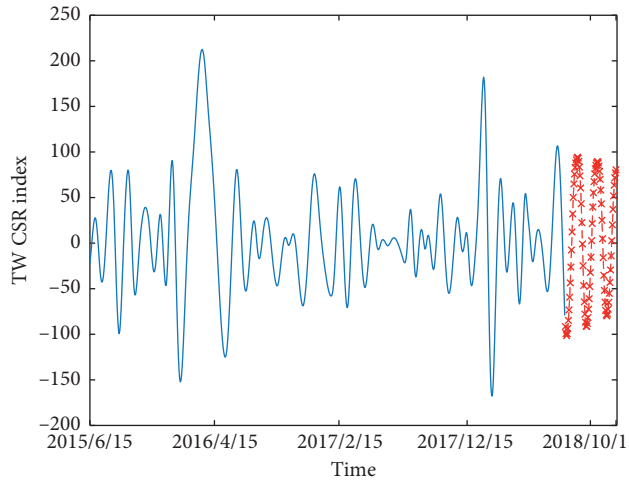

Figure 12: LSTM prediction results for IMF3.

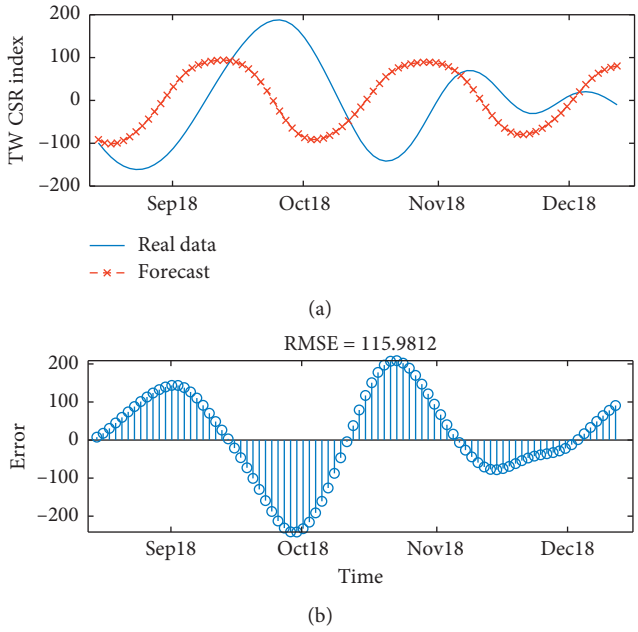

Figure 13: LSTM prediction and actual verification results for IMF3 with RMSE of 115.9812.

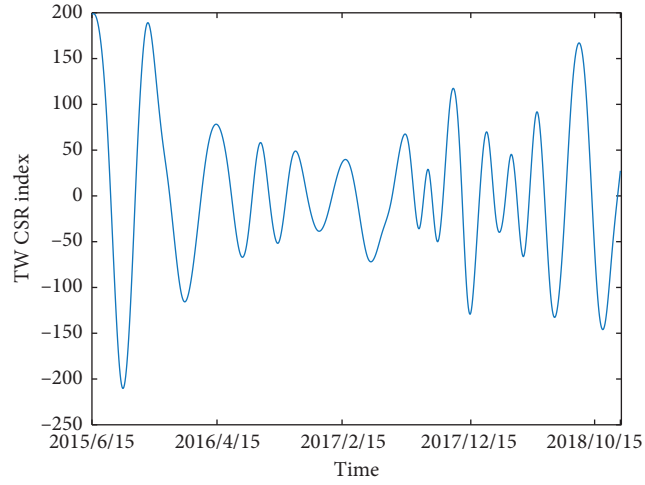

Figure 14: EMD analysis of the Taiwan CSR index used to get IMF4.

shows the trend of the TW CSR index data in IMF6. The statistical characteristics are as follows: the mean is 5431.6, the standard deviation is 552.7848, the variance is 305570, and the Pearson correlation coefficient is 0.9710. In the IMF6 data, we can observe that the index had a minimum of 4,693 on January 5, 2016, and a maximum of 6161 on April 18, 2018. During this period, the index rose by 1468, with the highest fall after April 18, 2018. Figure 21 shows the prediction results obtained with LSTM for IMF6. Figure 22 shows the prediction and actual verification results obtained with LSTM for IMF6. The RMSE is 52.6335. These IMFs are added to restore the predicted data. Figure 23 shows decomposition by the EMD followed by predictions by the LSTM method, after all the

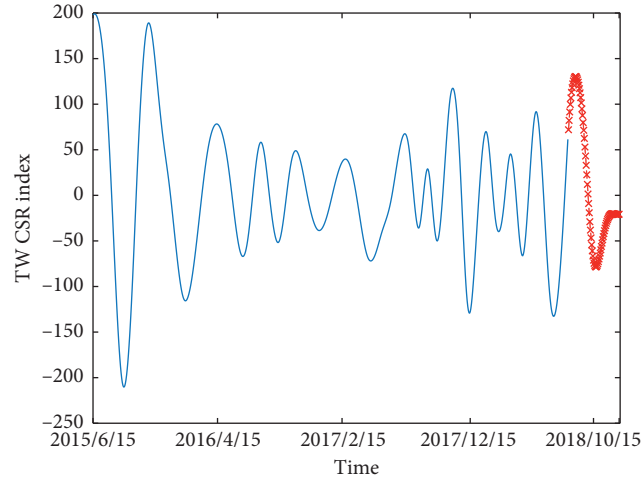

Figure 15: LSTM prediction results for IMF4.

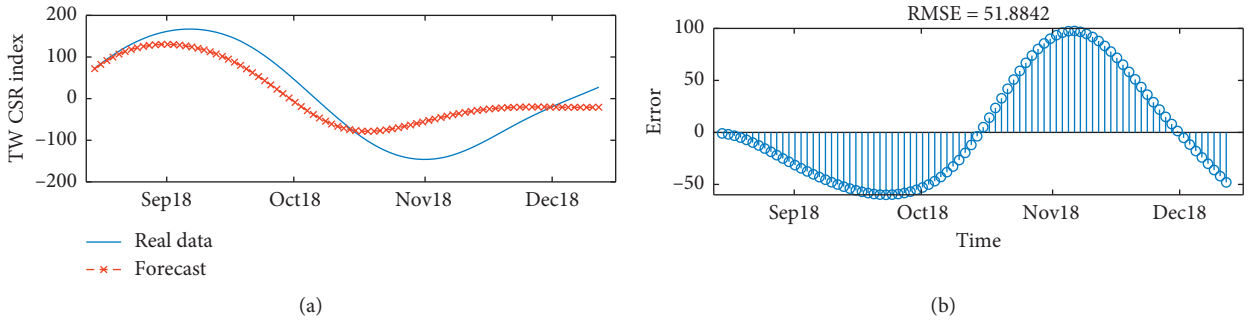

Figure 16: LSTM prediction and actual verification results for IMF4 with RMSE of 51.8842.

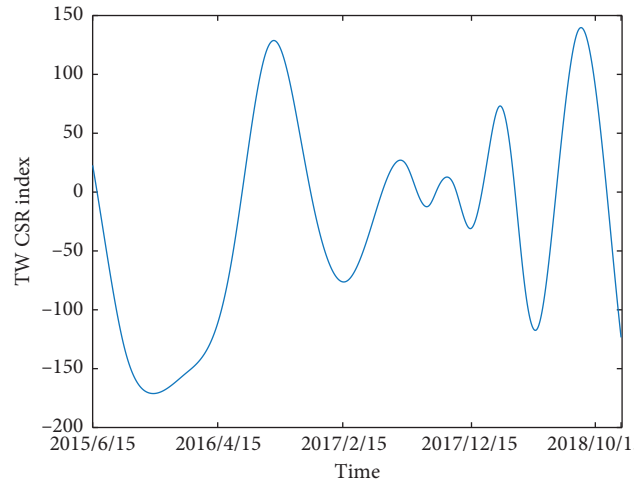

Figure 17: EMD analysis of the Taiwan CSR index used to get IMF5.

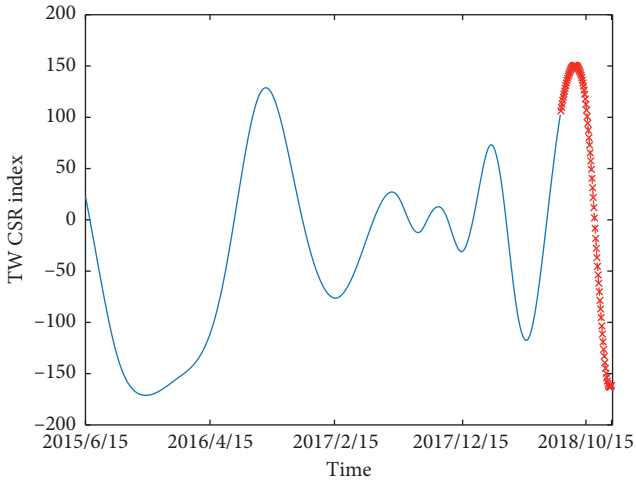

Figure 18: LSTM prediction results for IMF5.

IMF prediction results are added together. Figure 24 shows that the LSTM adds all the EMD prediction results to the actual verification RMSE of 175.7331. Table 2 shows the statistical recognition of the decomposition

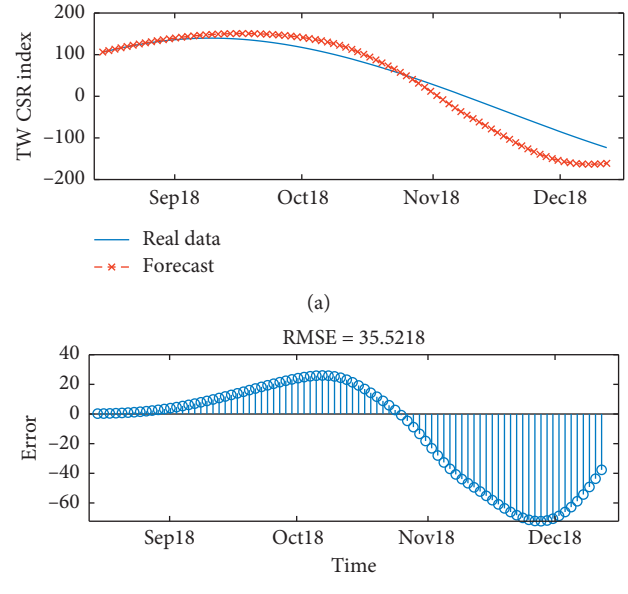

Figure 19: LSTM prediction and actual verification results for IMF5 with RMSE of 35.5218.

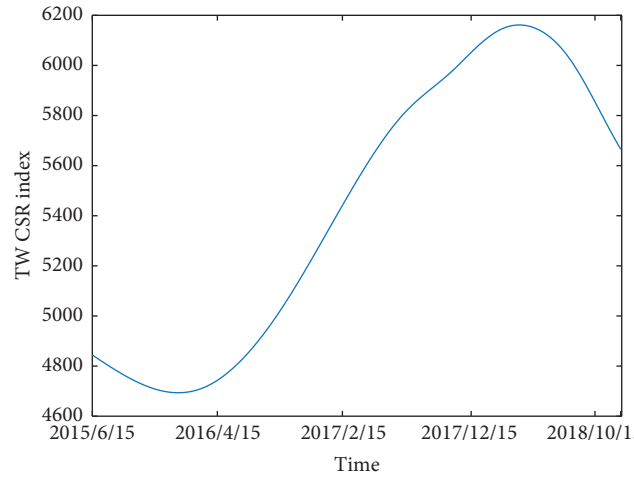

Figure 20: EMD analysis of the Taiwan CSR index used to get IMF6.

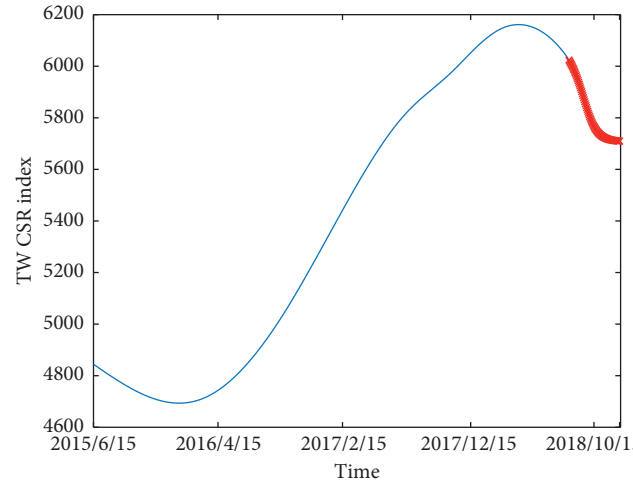

Figure 21: LSTM prediction results for IMF6.

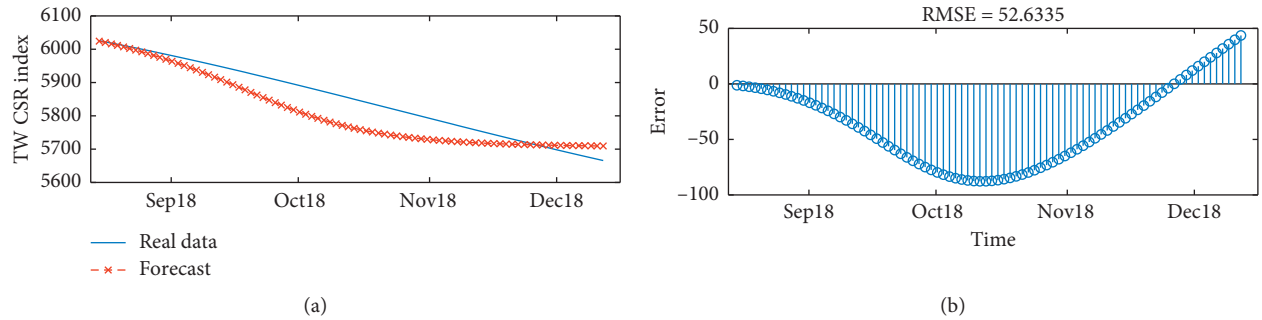

Figure 22: LSTM prediction and actual verification results for IMF6 with RMSE of 52.6335.

sequences. EMD is used to decompose the data to different IMFs with simpler statistical properties for LSTM prediction according to the characteristics of different IMFs. Table 3 indicates the difference in the RMSE between the real data and the predicted results for the two methods. It can be seen that the LSTM plus EMD RMSE is 175.7331, better than the LSTM predicted RMSE, which is 333.9627.

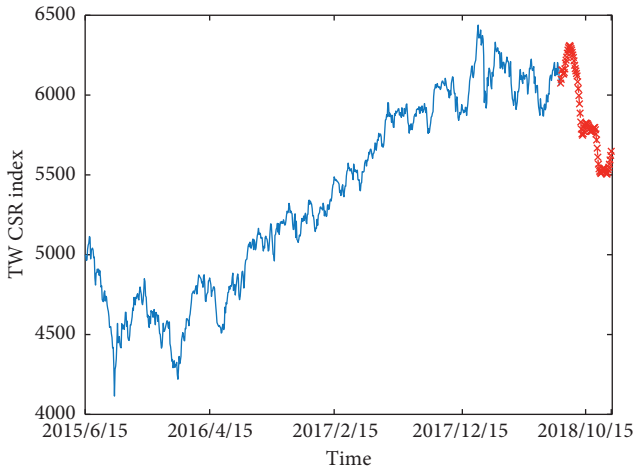

Figure 23: Addition of all prediction results for IMF1 to IMF6.

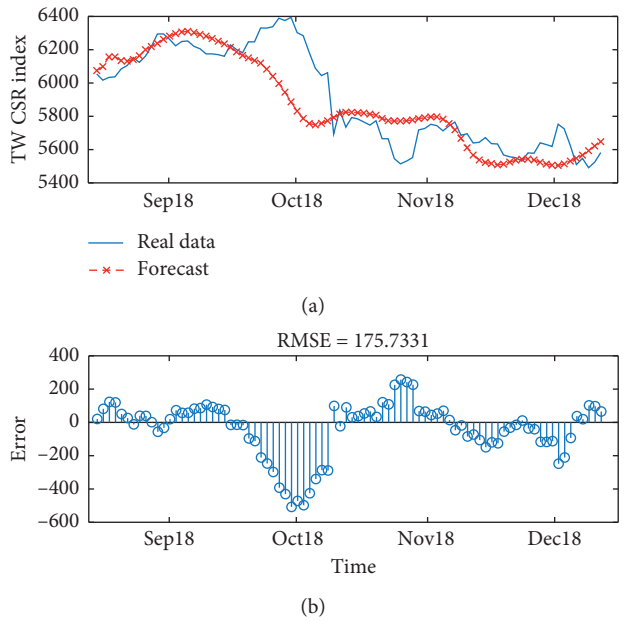

Figure 24: Addition of all LSTM EMD prediction results compared to the actual verification RMSE of 175.7331.

Table 2: Statistical recognition of the decomposition sequences.

	Average period (weeks)	Mean	Standard deviation	Variance	Pearson correlation coefficient
Real data	—	5412	600.1820	360220	1
IMF1	0.9279	−0.0390	30.2396	914.4325	0.0598
IMF2	2.7839	1.0316	41.5703	1728.1	0.0806
IMF3	7.3394	3.2523	64.4050	4148	0.0496
IMF4	18.6308	2.0298	79.5222	6323.8	0.1010
IMF5	48.44	−25.8465	85.7134	7346.8	0.5485
Residual term	—	5431.6	552.7848	305570	0.9710

Table 3: RMSE for real data and predicted results for the two methods.

Method	RMSE
Deep learning	333.9627
EMD + deep learning	175.7331

4. Conclusion

This paper proposes an empirical modal decomposition method to improve deep learning for the prediction of financial trends and financial data. Deep learning technology (e.g., LSTM) is suitable for big data prediction, but it can also be used for small data prediction with only poor accuracy. In fact, there are many practical situations where big data cannot be obtained, and only small data can be obtained for prediction. Data from Taiwan's CSR index were used for this study starting on June 15, 2015, with a total of 863 datasets. It was found that deep learning technology alone (in this case LSTM) is not good at predicting accuracy with small data. The EMD is used in this study to improve the accuracy. The standard method for dividing a dataset is 70% for training and 30% for testing. The more training data, the more accurate the results obtained. In many cases, the amount of data used for training is determined based on the characteristics of the data. In the study, the best results were obtained when 90% of the TW CSR index dataset was used for training and 10% for testing. In MF6, we can observe that the index is at its lowest on January 5, 2016 (4693), and at its highest on April 18, 2018 (6161). During this period, the total index rose by 1468, with the highest fall after April 18, 2018. Verification and comparison of the two methods show that EMD plus LSTM produces less error than prediction results obtained with only the LSTM model. The advantages of the proposed model are as follows: 1. EMD does not require complex mathematical operations. 2. EMD can analyze the frequency of data changes over time, disassembling complex financial data into components with multiple simple characteristics, and predictions made based on these components can improve prediction accuracy. 3. This research model is suitable for trending data such as economics or finance. Many new and improved EMDs have been proposed. The latest EMD prediction results for comparison could be a good direction for future research.

Data Availability

The Taiwan CSR index data used to support the findings of this study have been deposited in the Taiwan Corporate Governance 100 Index repository (https://www.taiwanindex.com.tw).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Hua-Wei Huang was supported by the NSC under Grant no. MOST 107-2410-H-006-017-MY3. Shih-Lin Lin was supported by the NSC under Grant no. MOST 109-2222-E-230 -001 -MY2.

References

[1] B. Nicoletti, *The Future of FinTech: Integrating Finance and Technology in Financial Services*, Springer, Berlin, Germany, 2017.

[2] T. Lynn, J. G. Mooney, P. Rosati, and M. Cummins, *Disrupting Finance FinTech and Strategy in the 21st Century*, Springer, Berlin, Germany, 2019.

[3] R. C. Cavalcante, R. C. Brasileiro, V. L. F. Souza, J. P. Nobrega, and A. L. I. Oliveira, "Computational intelligence and financial markets: a survey and future directions financial markets: a survey and future directions," *Expert Systems with Applications*, vol. 55, pp. 194–211, 2016.

[4] L. B. Van, "Machine learning: a revolution in risk management and compliance?" *Journal of Financial Transformation*, vol. 45, pp. 60–67, 2017.

[5] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3–12, 2017.

[6] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," *Artificial Intelligence Review*, vol. 50, no. 1, pp. 49–73, 2018.

[7] T. Mikolov, M. Karafiat, L. Burget, J. Černocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 1045–1048, Chiba, Japan, September 2010.

[8] Y. Wu, X. Lu, H. Yamamoto, S. Matsuda, C. Hori, and H. Kashioka, "Factored language model based on recurrent neural network," in *Proceedings of the COLINO 2012*, pp. 2835–2850, Mumbai, India, December 2012.

[9] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, https://arxiv.org/abs/1506.00019.

[10] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficultficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] D. Anguita, A. Ghio, N. Greco, L. Oneto, and S. Ridella, "Model selection for support vector machines: advantages and disadvantages of the machine learning theory," in *Proceedings of the International Joint Conference on Neural Networks*, Barcelona, Spain, EEE, July 2010.

[13] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[14] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[15] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[16] X. Zhang, K. K. Lai, and S. Y Wang, "A new approach for crude oil price analysis based on Empirical Mode Decomposition," *Energy Economics*, vol. 30, no. 3, pp. 905–918, 2008.

[17] A. M. Wang, M. T. Ismail, and S. A. L. Wadi, "Improving forecasting accuracy for stock market data using EMD-HW bagging," *PLoS One*, vol. 13, no. 7, Article ID e0199582, 2018.

[18] K. Guhathakurta, I. Mukherjee, and A. R. Chowdhury, "Empirical mode decomposition analysis of two different financial time series and their comparison," *Chaos, Solitons & Fractals*, vol. 37, no. 4, pp. 1214–1227, 2008.

[19] M. Khalid, M. Sultana, and F. Zaidi, "Stock market returns and direction prediction: an empirical study on karachi stock exchange," *Mathematical Theory and Modeling*, vol. 5, no. 3, pp. 106–111, 2015.

[20] M. R. Islam, M. Rashed-Al-Mahfuz, S. Ahmad, and M. K. I. Molla, "Multiband prediction model for financial time series with multivariate empirical mode decomposition," *Discrete Dynamics in Nature and Society*, vol. 2012, Article ID 593018, 21 pages, 2012.

[21] Y. Fang, "A study on the correlations between investor sentiment and stock index and macro economy based on EEMD method," *Journal of Financial Risk Management*, vol. 4, no. 3, pp. 206–215, 2015.

[22] S. Kv, G. N. Pillai, and B. Peethambaran, "Prediction of landslide displacement with controlling factors using extreme learning adaptive neuro-fuzzy inference system (ELANFIS)," *Applied Soft Computing*, vol. 61, pp. 892–904, 2017.

[23] K. V. Shihabudheen and B. Peethambaran, "Landslide displacement prediction technique using improved neuro-fuzzy system," *Arabian Journal of Geosciences*, vol. 10, no. 22, p. 502, 2017.

[24] G. N. Pillai and K. V. Shihabudheen, "Wind speed forecasting using empirical mode decomposition and regularized ELANFIS," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, IEEE, Honolulu, HI, USA, November 2017.

[25] K. V. Shihabudheen and G. N. Pillai, "Wind speed and solar irradiance prediction using improved neuro-fuzzy systems," in *Proceedings of the IEEE World Congress on Computational Intelligence*, pp. 1–7, IEEE, Rio de Janeiro, Brazil, July 2018.

[26] H. Liu, C. Chen, H.-G. Tian, and Y.-F. Li, "A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks," *Renewable Energy*, vol. 48, pp. 545–556, 2012.

[27] Y. Ren, P. N. Suganthan, and N. Srikanth, "A novel empirical mode decomposition with support vector regression for wind speed forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 8, pp. 1793–1798, 2016.

[28] H.-F. Yang and Y. P. P. Chen, "Hybrid deep learning and empirical mode decomposition model for time series applications," *Expert Systems with Applications*, vol. 120, pp. 128–138, 2019.

[29] Ü. Ç. Chen and Ş. Ertekin, "Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition," *Neurocomputing*, vol. 361, pp. 151–163, 2019.

[30] W. C. Hong and G. F. Fan, "Hybrid empirical mode decomposition with support vector regression model for short term load forecasting," *Energies*, vol. 12, no. 6, pp. 1–16, 2019.

[31] J. H. Stock, "Unit roots, structural breaks and trends," in *Handbook of Econometrics*, R. F. Engle and D. L. McFadden, Eds., vol. IV, pp. 2739–2841, Elsevier, Amsterdam, Netherlands, 1994.