ISSN 2959-6157

Financial Time Series Forecasting Based on Long-Short Term Memory, Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Batch Normalization

Yixuan Xu

Ulink High School of Suzhou Industrial Park, Suzhou, China
*Corresponding author: yixuan.xu@sz-alevel.com

Abstract:

Long-short term memory (LSTM) is a state-of-art and widely used model to forecast financial time series. However, primitive LSTM networks do not perform well due to over-fitting problems of the deep learning model and non-linear and non-stationary characteristics of financial time series data. In addition, complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is an outstanding data frequency decomposition technique that can decompose original time series into several intrinsic mode functions and a residue. Thus, this paper proposed a novel hybrid network CEEMDAN-LSTM-BN based on LSTM. Specifically, to avoid over-fitting, the modified LSTM-BN network consists of two LSTM layers, two Batch Normalization (BN) layers following each LSTM layer, and a dropout layer. Each of the intrinsic mode functions and the residue would be processed by CEEMDAN-LSTM-BN and the final prediction results are obtained by reconstructing each predictive series. The advantages of the proposed CEEMDAN-LSTM-BN networks are verified by comparing them to primitive LSTM, other hybrid models, and some famous machine learning models. Moreover, the robustness of the networks is assessed by numerical experiments on different stock indices datasets.

Keywords: Financial; time series forecasting; long-short term memory; batch normalization; complete ensemble empirical mode decomposition with adaptive noise.

1. Introduction

Forecasting stock index prices is a significant issue for investors and professional researchers as it plays an important role in the economy of any region and country [1, 2]. Given the fact that stock market price data is dynamic, no-stable, non-stationary, nonlinear, noisy, and chaotic, it is tremendously difficult for researchers to analyze and forecast [3, 4]. Although precise prediction of the financial market is nearly impossible, a lot of researchers put forward various ideas and methods to resolve this problem. In fact, with the incessant progress of artificial intelligence, it is possible to forecast the financial market more precisely. However, although the statistical model has been widely used and very helpful in a variety of research areas, most machine learning models are based on stable and linear assumptions so the efficiency and accuracy of the performance in predicting stock price is poor [5].

Compared to machine learning techniques, deep learning which can process non-linear and dynamic financial time series has a better performance [6, 7]. The well-known deep learning techniques include Artificial Neural Networks (ANN) [8, 9], Recurrent Neural Networks (RNN) [10], Long Short-Term Memory (LSTM) [11], and so on. Yet despite there are certain advantages of those neural networks, they are still unable to forecast the fluctuation of the stock market accurately because there is no regression and traditional neural networks have only shallow architecture. Forecasting the financial market requires time series analysis. To be more precise, the prediction is not only related to the data at the latest time or the current time but also to the earlier time. Compared to traditional ANN, RNN, and LSTM, which are both capable of extracting noisy and non-linear data features, enable the memory of earlier information that improves forecasting accuracy. Specifically, as LSTM solves the long-term dependence problem in time series analysis, LSTM is selected to forecast the stock price index in this paper.

Although LSTM is very effective in data analysis, sometimes time series data could be so volatile and stochastic that the results of performance are still unsatisfying [12].

To tackle this problem, frequency decomposition such as Empirical mode decomposition (EMD) is used to improve the data analysis process. Moreover, to decrease the effect of noise and increase the accuracy of prediction, LSTM is combined with EMD to forecast the financial time series. However, EMD remains an unignorable problem of mode mixing which refers to oscillations of dramatically disparate scales consisting of intrinsic mode functions (IMF). To resolve it, several advanced versions have been put forward, for instance, ensemble empirical mode decomposition (EEMD) [13] and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [14] which have obvious advantages in avoiding mode mixing and reducing noise in the mode. The EEMD algorithm reduces the mode mixing effect by adding normally distributed white noise to the original series at the base of the EMD algorithm. The CEEMDAN algorithm resolves EEMD's problems of incompleteness and reconstruction error due to adding white noise.

Thus, CEEMDAN is adopted to combine with LSTM to improve the accuracy of the forecasting. In the final, this paper uses global stock price indices for practical evaluation and compares the proposed model (CEEMDAN-LSTM) with another model (EMD-LSTM, EMD-SVM, LSTM, and SVM). The original financial data comes from four major global stock indices, including the Nikkei 225 Stock Index (N225), Standard & Poor 500 Index (S&P500), Hang Seng Index (HSI), and Deutscher Aktien Index (DAX).

The remainder of the paper is organized as follows. The methodology is elaborated on in the Section 2. Section 3 discusses the experimental analysis of the prediction. The final Section 4 concludes.

2. Methodology

2.1 Empirical Mode Decomposition (EMD)

EMD is put forward by Huang et al., which decomposes the complex time series accordingly into a series of intrinsic mode functions (IMFs) by characteristic time scales of data [12]. EMD has the power to preprocess the data, transforming non-stationary and non-linear characteristics into stationary and linear.

2.2 Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

Although the EMD algorithm as a type of frequency decomposition has great advantages in processing non-stationary series not requiring basic functions in advance, it has obvious mode mixing problems. As an advanced version of EMD, EEMD largely overcomes the mode mixing problems by adding Gaussian withe noise to the original data. As a result, however, EEMD faces the problem of incompleteness which means it can not eliminate the white noise after signal reconstruction. To tackle this problem, CEEMDAN is put forward as an advanced version of EEMD. Its reconstruction error is almost zero and it has a faster calculate speed.

2.3 Long Short-Term Memory (LSTM)

With the advantage of the feedback mechanism, the RNN technique can be utilized in the one-step-ahead prediction of financial series using the latest data and previous data. However, although RNN has an advantage in dealing with long-term dependence problems, it is nearly practically useless because of the exploding or vanishing gradient problem. To tackle this problem, Hochreiter S and Schmidhuber J. put forward LSTM in 1997 [15]. Additionally, LSTM uses the gate mechanism on the base of RNN to largely solve the problem that effective historical information in the previous data can not be preserved for a long time.

LSTM includes three gates: the forget gate, the input gate, and the output gate. The structure of the LSTM unit is shown in Fig.1. For each LSTM unit at the time t, x_t is the input data, x_{t-1} is the input of the previous unit, h_t is the output of this LSTM unit, and h_{t-1} is the output of the previous unit.

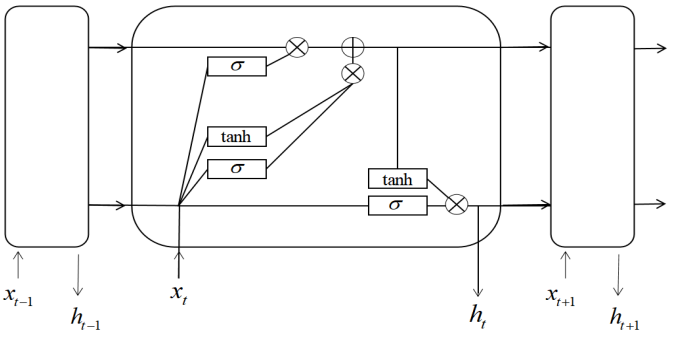


Fig. 1. Internal structure in an LSTM memory unit (Photo credit: Original)

2.4 Dropout and Batch Normalization Transform (BN)

Over-fitting is a non-ignorable issue in time series prediction, which means the trained model might have a satisfying performance in the train set, while poorly fitting the test set. To deal with this problem dropout was proposed by Hinton G E [16]. In forward propagation, some hidden nodes would be blocked with a preset rate by setting their output values as zero, and in backpropagation, the parameters in the nodes will not be updated. Another technology to solve over-fitting is Batch Normalization proposed by Ioffe and Szegegy in 2015 [17]. Batch normalization works by normalizing the input of each layer to have a mean of zero and a variance of one.

2.5 Proposed Model

2.5.1 Basic Structure of LSTM-BN:

To improve the performance in financial time series prediction, it is necessary to combine LSTM with BN layers and dropout layers. Based on this principle, the paper puts forward an LSTM-BN hybrid deep learning network whose framework is shown in Fig.2. It consists of three LSTM layers, two BN layers following the first two LSTM layers, and a dropout layer. In this LSTM-BN model, the "ReLU" function is adopted as the activation function of the fully connected layer and the paper adopts the mean square error (MSE) as the loss function. That is, $Loss = MSE = \frac{1}{N}\sum_{n=1}^{N}(d_n - y_n)^2$, Where N is the total number of days, d_n is the actual value, and y_n is the predictive value.

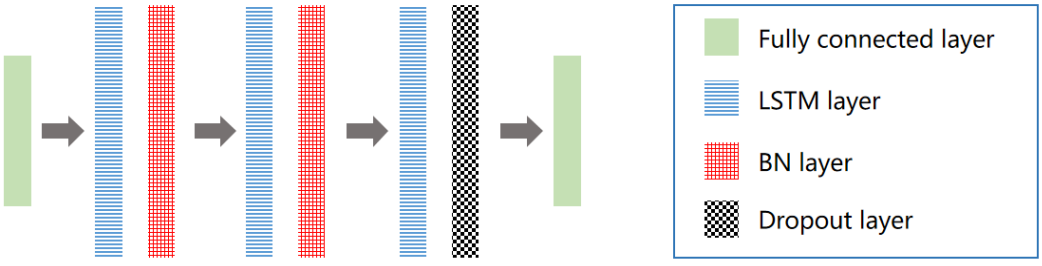


Fig. 2. Basic structure of proposed model (Photo credit: Original)

2.5.2 Structure of CEEMDAN-LSTM-BN

The CEEMDAN technique is used to decompose the original series in advance to get smooth sequences. The proposed CEEMDAN-LSTM-BN model is applied to forecast some notable stock price indices and the implementation steps are shown in the following:

1) Use CEEMDAN to decompose the original series $S(t)$ into IMFs sequences $C_i(t)(i=1,2,...,M)$ and a residue $R_M(t)$. 2) The IMFs obtained and residue are used as the inputs of the LSTM-BN model for training and get the predicted results respectively. The predicted result of the test set is $\overline{C}_i(t)(i=1,2,...,M)$ and $\overline{R}_M(t)$. 3) Get the final predicted result by the formula $S_i(t)=\sum_{i=1}^{M}\overline{C}_i(t)+\overline{R}_M(t)(t=1,2,..,L)$, Where $S_i(t)$ is the final predictive series of the test set and L is the length of the test series, according to each IMF and the residue obtained.

3. Analysis of Experimental Results

3.1 Data Preparation

Compared to a single stock, stock market indexes are generally regarded as the best performance indicator of the financial market. To demonstrate the robustness of the proposed hybrid model, the daily closing prices of the S&P500, N255, HSI, and DAX are selected as the original data. Moreover, for better performance of the prediction, realized volatility (RV) would be calculated from the original data and used as input data for the hybrid model. RV is a measure used in financial time series analysis to quantify the amount of variation or fluctuation in the prices of an asset over a specific period. Incorporating RV into deep learning models involves using it as an input variable to capture the intrinsic market dynamics better. This approach helps in training models to learn from historical volatility patterns and improve the accuracy of predictions. The method for calculating RV is defined as $r_{t,i}=\ln\left(p_{t,i}/p_{t,i-1}\right)$. $RV_t=\sqrt{\sum_{i=1}^{n}r_{t,i}^2}$, Where is the close price at interval i on day t, $p_{t,i-1}$ is the price at the previous interval on the day t and n is the number of intervals

within the day.

Aiming to accelerate the training speed and strengthen the generalization ability of the model, the following standardization method is applied to the RV: $P = (RV - \mu_{RV}) / \sigma_{RV}$, Where μ_{RV} and σ_{RV} are the mean value and standard deviation of RV and RV is the data set of standardized realized volatility as the input data to the proposed model.

Fig. 3 shows the normalized realized volatility data of financial time series and the statistical analysis of the normalized RV data is shown in Table 1. The data of all indices are from October 1, 2004, to April 4, 2024. Especially, the RVs of all four indices are relatively higher in 2008 and 2019 respectively due to the financial crisis and COVID-19 epidemic. Moreover, the top 90% of data would be selected as a training set and the rest 10% would serve as a test set.

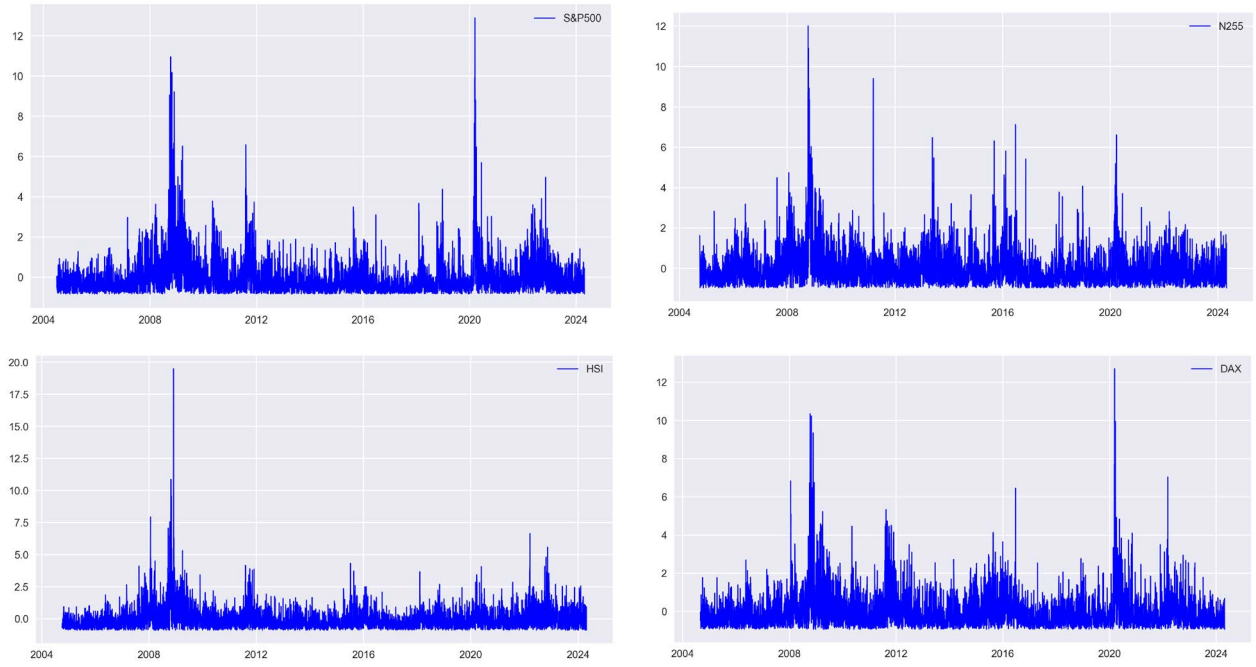


Fig. 3. The normalized Realized Volatility of the indices (Photo credit: Original)

Table. 1 The statistical analysis result normalized Realized Volatility of the four indices

Index	Count	Mean	Min	Max	Standard deviation
S&P500	4987	0	-0.833	12.889	1
N255	4816	0	-0.970	12.016	1
HSI	4850	0	-0.884	19.493	1
DAX	4999	0	-0.934	12.711	1

3.2 CEEMDAN of RV

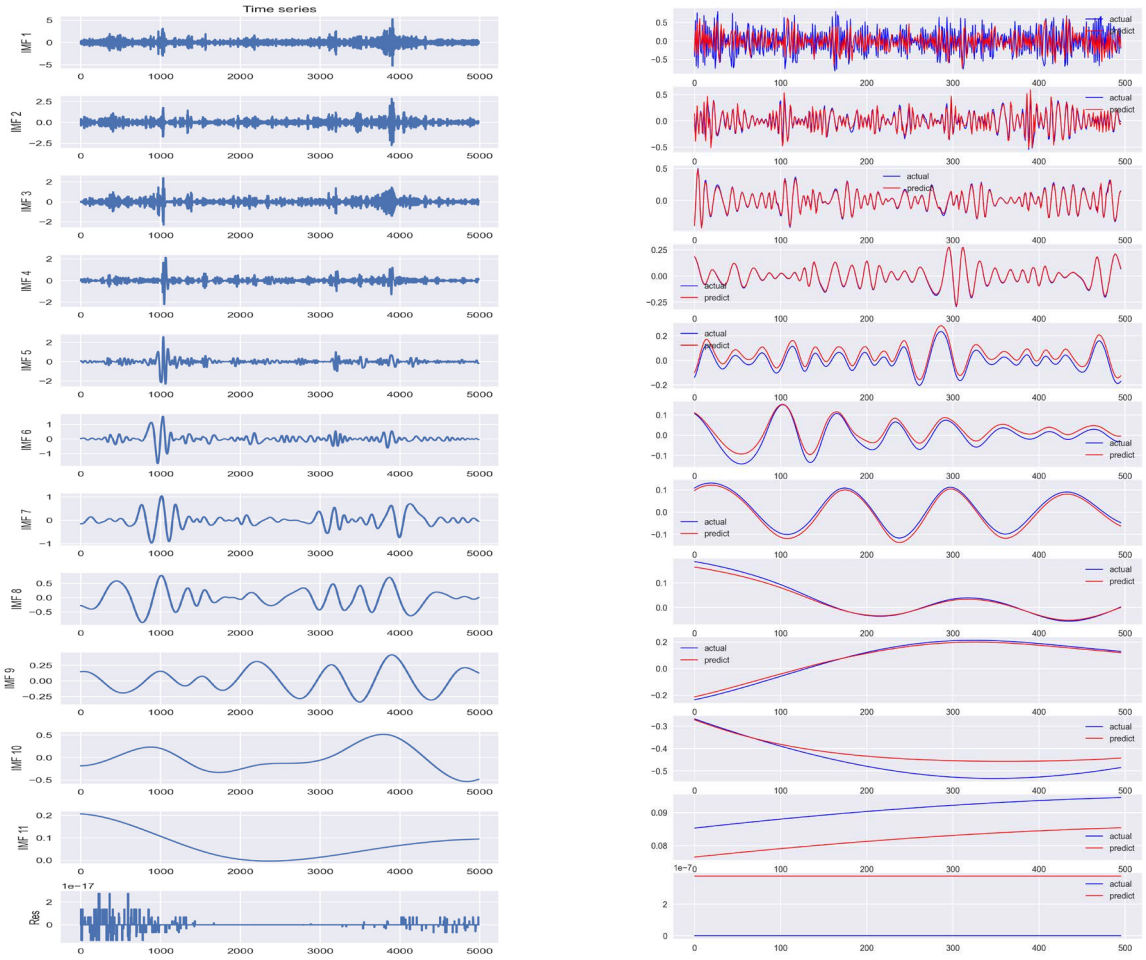
The normalized RV time series is decomposed by CEEMADAN into several IMFs and one residue. Fig. 4(a) shows the decomposition results of the S&P500 index series and each IMF is arranged from high frequency to low frequency.

3.3 Training process and prediction results

After decomposition, each sub-series consisting of IMFs and residue is divided into a training set and a test set. Fig. 4(b) shows the results of sub-series for S&P500 index data. The prediction performance of high-frequency IMF1 and IMF2 is relatively low due to the high amplitude of the components.

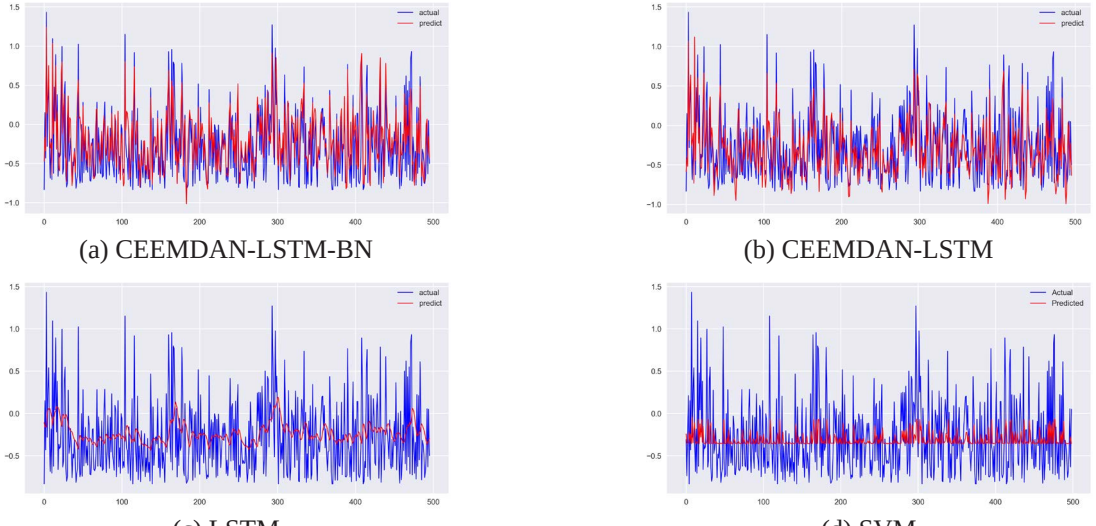
The forecast results of the S&P500 index by the proposed CEEMDAN-LSTM-BN model are shown in Fig. 5(a). By this graph, the prediction accuracy of the proposed models is outstanding. To evaluate the performance with more accuracy, two error measures are adopted, including Mean Squared Error (MSE) and Mean Absolute Error (MAE). MAE is calculated as follows: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and

n is the number of observations.



(a) decomposed by CEEMDAN (b) predicted by CEEMDAN-LSTM-BN

Fig. 4 Decomposition and its prediction results of S&P500 (Photo credit: Original)



(a) CEEMDAN-LSTM-BN (b) CEEMDAN-LSTM

(c) LSTM (d) SVM

Fig. 5 The prediction results of S&P500 (Photo credit: Original)

The smaller the value of MSE and MSA, the smaller the deviation between the predicted value and the original

value, suggesting better performance of the model. The MSE value is more stable and solid, so it is used as the main evaluation criterion. Table 2 shows the prediction error of the models. According to Table 2, the prediction performance of CEEMDAN-LSTM-BN is better than LSTM, LSTM-BN, and CEEMDAN-LSTM since CEEMDAN extracts more effective features from original data deemphasizing the effect of noise and BN improves the prediction accuracy of LSTM network.

Table. 2 Comparison of the prediction error of different models

(a) Results of S&P500

Model	CEEMDAN -LSTM-BN	CCEMDAN -LSTM	LSTM-BN	LSTM	HAR	AR	SVM	Random Forest
MSE	0.095	0.104	0.204	0.207	0.260	0.453	0.423	1.029
MAE	0.233	0.253	0.368	0.375	0.430	0.454	0.545	0.661

(b) Results of N255

Model	CEEMDAN -LSTM-BN	CCEMDAN -LSTM	LSTM-BN	LSTM	HAR	AR	SVM	Random Forest
MSE	0.216	0.240	0.468	0.499	0.537	0.539	0.714	1.147
MAE	0.343	0.371	0.548	0.567	0.599	0.619	0.602	0.514

(c) Results measures of HSI

Model	CEEMDAN -LSTM-BN	CCEMDAN -LSTM	LSTM-BN	LSTM	HAR	AR	SVM	Random Forest
MSE	0.114	0.123	0.227	0.229	0.314	0.380	0.799	1.506
MAE	0.259	0.253	0.399	0.404	0.480	0.549	0.586	0.728

(d) Results of DAX

Model	CEEMDAN -LSTM-BN	CCEMDAN -LSTM	LSTM-BN	LSTM	HAR	AR	SVM	Random Forest
MSE	0.168	0.170	0.367	0.367	0.419	0.449	0.549	1.222
MAE	0.283	0.300	0.477	0.480	0.548	0.578	0.523	0.741

3.4 Comparison with other models

To verify the performance of the proposed hybrid deep learning method, this paper compared the prediction results of several famous machine learning models, including Support Vector Machine (SVM), Random Forest, AutoRegressive (AR), and Heterogeneous Autoregressive (HAR). All models use the same data set. SVM is widely used for time series prediction [18]. SVM casts the input data to high-dimensional space and applies the Linear Regression model in high-dimension space to predict the nonlinear data. This paper uses the radial basis function as the kernel function and implements the SVM through the "sklearn" library in Python. The AR model is a famous time series model that uses the dependencies between an observation and several past values to make predictions. HAR model is an extension of the AR model that includes components to capture the heterogeneous nature of financial time series data and HAR is widely used in finance to model and predict volatility. Random Forests have good performance in various fields including financial market prediction Random Forest is an ensemble learning method that constructs a multitude of decision trees during train-

ing and outputs the mode of the classes or mean prediction of the individual trees. Table 2 shows the prediction measures of the four indices using these models.

The error measures MSE and MAE values of the S&P500 index by the proposed hybrid CEEMDAN-LSTM-BN model are 0.095 and 0.233 respectively, which is the smallest among the models. Additionally, the result of CEEMDAN-LSTM is highly better than LSTM and other models, and BN layers also have an important positive effect on hybrid models. The proposed model also performs better than the other models in predicting the N255, HSI, and DAX, and the predictive value is very close to the original value.

4. Conclusion

The paper develops a novel CEEMDAN-LSTM-BN hybrid deep learning model to forecast the realized volatility of the stock index in financial time series data. In this study, the BN layer and dropout layer are adopted to avoid over-fitting problems in primitive LSTM networks. In addition, the CEEMDAN signal decomposition technique is combined with LSTM-BN to further enhance the performance. The robustness and effectiveness of the hybrid models are verified by a set of numerical experiments using different stock indices. At the same time, this paper compares the proposed model with primitive LSTM, LSTM-BN, CCEMDAN-LSTM, and other four famous machine-learning models. Although the proposed hybrid models have a satisfying performance in forecasting financial indices, there are still some places to improve in the future. For instance, data fusion methods can be studied, involving not only stock prices as input data but also trading volume, different time scale series, as well as macroeconomic and microeconomic data. In addition, more advanced forecasting models will be studied by introducing state-of-the-art deep learning methods such as xLSTM and Transformer.

References

[1] Introduction to time series and forecasting [M]. New York, NY: Springer New York, 2002.

[2] Poon S H, Granger C W J. Forecasting volatility in financial markets: A review [J]. Journal of economic literature, 2003, 41(2): 478-539.

[3] Abu-Mostafa Y S, Atiya A F. Introduction to financial forecasting [J]. Applied intelligence, 1996, 6: 205-213.

[4] Qin S J, Chiang L H. Advances and opportunities in machine learning for process data analytics [J]. Computers & Chemical Engineering, 2019, 126: 465-473.

[5] Chen Y, Hao Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction [J]. Expert Systems with Applications, 2017, 80: 340-355.

[6] Guresen E, Kayakutlu G, Daim T U. Using artificial neural network models in stock market index prediction [J]. Expert systems with Applications, 2011, 38(8): 10389-10397.

[7] Rezaei H, Faaljou H, Mansourfar G. Stock price prediction using deep learning and frequency decomposition [J]. Expert Systems with Applications, 2021, 169: 114332.

[8] Guresen E, Kayakutlu G, Daim T U. Using artificial neural network models in stock market index prediction [J]. Expert systems with Applications, 2011, 38(8): 10389-10397.

[9] Kim K, Han I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index [J]. Expert systems with Applications, 2000, 19(2): 125-132.

[10] Kamijo K, Tanigawa T. Stock price pattern recognition-a recurrent neural network approach [C]//1990 IJCNN international joint conference on neural networks. IEEE, 1990: 215-221.

[11] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions [J]. European journal of operational research, 2018, 270(2): 654-669.

[12] Nelson D M Q, Pereira A C M, De Oliveira R A. Stock market's price movement prediction with LSTM neural networks [C]//2017 International joint conference on neural networks (IJCNN). Ieee, 2017: 1419-1426.

[13] Wu Z, Huang N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method [J]. Advances in adaptive data analysis, 2009, 1(01): 1-41.

[14] Colominas M A, Schlotthauer G, Torres M E. Improved complete ensemble EMD: A suitable tool for biomedical signal processing [J]. Biomedical Signal Processing and Control, 2014, 14: 19-29.

[15] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[16] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv:1207.0580, 2012.

[17] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// International conference on machine learning. pmlr, 2015: 448-456.

[18] Sapankevych N I, Sankar R. Time series prediction using support vector machines: a survey [J]. IEEE computational intelligence magazine, 2009, 4(2): 24-38.