Journal of Trends in Financial and Economics
ISSN: 3007-696X
DOI: 10.61784/jtfe3005

REALIZED VOLATILITY PREDICTION WITH A HYBRID MODEL: LSTM-CEEMDAN

ZiHang Zeng
School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215400, Jiangsu, China.
Corresponding Email: Zihang.Zeng23@student.xjtlu.edu.cn

Abstract: The realized volatility (RV) in financial time series is characterized by nonlinearity, volatility, and noise. It is challenging to predict RV with a solitary forecasting model for precision. This study employs a hybrid model that integrates the Long Short-Term Memory (LSTM) network with the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) for the purpose of forecasting the returns volatility (RV) of the S&P 500 index, thereby validating its accuracy and robustness.

Keywords: LSTM; CEEMDAN; Realized volatility

1 INTRODUCTION

Volatility serves not only as a conventional metric for assessing risk within the financial markets but also as a pivotal factor in determining asset prices and constructing investment portfolios. The precise prediction of volatility has consistently been a subject of intense scholarly inquiry. In its formative stages, academic research primarily focused on the prognostication of low-frequency volatility, as delineated by seminal works such as those by Engle, Bollerslev, and Ka & Heynen. Nonetheless, the predictive outcomes of low-frequency volatility prove to be inadequate as a reliable proxy for anticipating future market risks, owing to the manifest limitations of such approaches.

It has a great range of researches when RV was first proposed, it included but not limited to, the Autoregressive (AR) model, the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model, and the Stochastic Volatility model. Besides, in recent years, there has been a burgeoning advancement in artificial intelligence, leading to a widespread deployment of deep learning algorithms across a multitude of disciplines, as evidenced by the seminal works of Lu, Que, and Cao, as well as Kodama, Pichl, and Kaizojl. When juxtaposed with conventional econometric models, deep learning methodologies exhibit superior performance, owing to their fewer constraints and enhanced feature extraction capabilities. McAleer and Medeiros introduced a nonlinear hierarchical auto-regressive (HAR) model based on neural networks. Arnerié, Poklepovié, and Teai compared two methodologies: HAR and feedforward neural networks (FNN). Furthermore, they deduced that FNN-HAR-type models exhibit superior performance in encapsulating the nonlinearity of return volatility (RV).

To address this issue, a novel hybrid model, CEEMDAN-LSTM, is introduced in this study for the purpose of RV forecasting. Furthermore, it is imperative to validate the forecasting efficacy of CEEMDAN-LSTM across both emerging and developed markets. Accordingly, we have designed an extensive analytical framework.

2 PREVIOUS LITERATURE

Currently, stock market is acknowledged as being chaotic, complicated, volatile and dynamic [1]. Thus, stock prediction has been an important topic that cannot be ignored and calls for future. According to the current literature, a variety of data-driven predictors have been developed to forecast stocks, which can be categorized into two broad approaches: those employing single models and those utilizing hybrid models. The methodology of single-model forecasting encompasses conventional statistical approaches, established machine learning algorithms, and contemporary deep learning techniques. Conversely, the hybrid model forecasting approach typically entails a synthesis of these methodologies bolstered by comprehensive feature engineering; the decomposition-integration strategy stands out as a noteworthy instance within this category. Despite its reliance on a solitary predictive model, this approach is typically categorized as hybrid.

The methodology of single-model forecasting entails the utilization of a singular predictive model for estimating carbon prices, in conjunction with certain feature engineering techniques. For single-model methods, Bhattacharjee and Bhattacharja found MLP and LSTM are the most accurate way to predict stock prices for having the least MSE and MAPE values [2]. Ariyo, Adewumi and Ayo used ARIMA model for stock price prediction and found it can guide investors to make right decisions [3]. Nevertheless, despite extensive testing and preprocessing, single model, such as, traditional statistical models struggle with utilizing non-linear, non-normal distributional stock prices for precise predictions

Classical machine learning methods, Support Vector Machine (SVM) [4] and Least Squares Support Vector Machine (LS-SVM) [5],exhibit relatively exceptional forecasting capabilities for addressing small-sample, nonlinear, and high-dimensional issues pertaining to stock market price determination. Zhu and Wei [6] discovered that after parameter optimization, the Least Squares Support Vector Machine (LS-SVM) outperforms the Auto Regressive Integrated Moving Average (ARIMA) model. However, the Support Vector Machine (SVM) merely converts the complexity of high-dimensional spaces into the challenge of identifying the optimal kernel function [7]. Deep learning methodologies have significantly advanced the cutting edge in speech recognition, visual object recognition, and object detection. Yahsi et al. [8] utilized both machine learning and deep learning methodologies and ascertained that the artificial neural network was inefficient. Given the deep sample dependency inherent in deep learning methodologies, the generalization capability of these approaches is inextricably linked to the representation of typical learning instances. It is difficult to achieve the expected performance if the sample set hard is hard to representative with contradictions and redundance.

A hybrid forecasting methodology was introduced due to the challenges encountered by researchers in accurately predicting the complex and dynamic nature of stock market price through extensive empirical research using a single model. Zhu and Wei [6] introduced a novel hybrid approach integrating the ARIMA and LS-SVM models, which was found to yield superior performance compared to their individual counterparts. Moreover, Chen, Zhou and Dai used a LSTM-based model to forecast China stock returns [9]. Convolutional Neural Networks (CNNs) are capable of detecting hierarchical structural patterns within data, while Long Short-Term Memory (LSTM) [10] networks excel at capturing long-term dependencies embedded within the dataset. Zhang[11] found ARIMA-CNN-LSTM model is better in prediction models. Decomposition-integration methodologies are extensively employed in stock market price forecasting to augment the limited dataset and enhance precision. These methods are preferred for their facile construction of predictors and time-efficiency in comparison to hybrid models, which often involve numerous intricate components. Commonly employed methods for time series decomposition include the Wavelet Transform [12], Empirical Mode Decomposition (EMD) [13], Hilbert-Huang Transform [13], Ensemble Empirical Mode Decomposition (EEMD) [14], Complementary Ensemble Empirical Mode Decomposition (CEEMD) [15], Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) [16], and Variational Modal Decomposition (VMD) [17].

In addition to the literature mentioned above, numerous scholars are dedicated to researching stock markets forecasting via the decomposition-integration approach. This study is emblematic of this trend, as it integrates CEEMDAN, VMD, and LSTM into the decomposition-integration framework for comprehensive exploration and validation. This paper concludes two fundamental frameworks, called ensemble and the respective LSTM forecasting methods and puts forward a hybrid one combined with VMD re-decomposition to predict stock market. A set of experimental supplements and comparisons are provided to verify previous literature and address the deficiencies in certain literature that are challenging to replicate during programming.

3 METHOD

3.1 The Structure of LSTM

The LSTM Flowchart can be seen in below Figure 1.

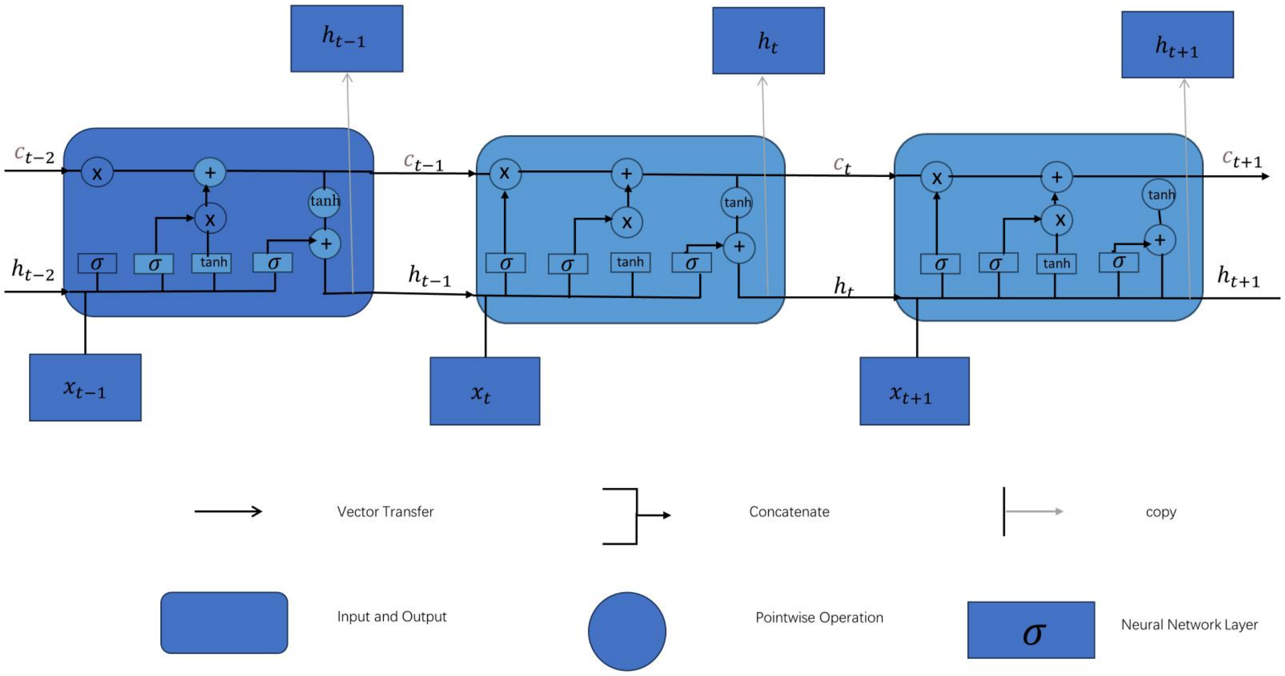

Figure1 LSTM Flowchart

3.1.1 Long Short-Term Memory (LSTM)

Hochreiter and Schmidhuber [1] initially introduced the Long Short-Term Memory (LSTM) architecture, a meticulously designed extension of the Recurrent Neural Network (RNN) that incorporates memory mechanisms to mitigate the challenges associated with long-term dependencies. LSTM, or Long Short-Term Memory, is a type of artificial recurrent neural network (RNN) architecture used in the field of deep learning. LSTMs are designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. They were introduced by Hochreiter and Schmidhuber in 1997.

Key Features of LSTM:

1) Memory Cells: The core component of an LSTM network is the memory cell, which is capable of maintaining information for long periods.

2) Gates: LSTMs use three gates to control the flow of information:

Forget Gate: Decides which information to discard from the cell state.

Input Gate: Decides which new information to store in the cell state.

Output Gate: Decides what part of the cell state to output as the hidden state.

3) Cell State: The memory cell state is updated by the gates, allowing the LSTM to maintain and modify memory over time.

Applications:

Natural Language Processing (NLP): Language modeling, text generation, machine translation.

Speech Recognition: Transcribing spoken words into text.

Time Series Prediction: Stock market prediction, weather forecasting.

Anomaly Detection: Identifying unusual patterns in data.

The function:

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

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

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{3}$$

$$c_t = f_t{}^{\circ} c_{t-1} + i_t{}^{\circ} \tilde{c}_t \tag{4}$$

$$o_t = \sigma(\ W_0 x_t + U_0 h_{t-1} + b_0) \tag{5}$$

$$h_t = o_i{}^{\circ} \tan h\,(c_t) \tag{6}$$

LSTMs are particularly effective at handling the vanishing gradient problem, which is common in traditional RNNs when dealing with long-term dependencies. This makes them suitable for tasks that require learning from and predicting sequential data over extended time periods.

3.2 CEEMDAN

CEEMDAN, or Complete Ensemble Empirical Mode Decomposition with Adaptive Noise, is an advanced signal processing technique used for decomposing a complex signal into a set of simpler components called Intrinsic Mode Functions (IMFs). It is an enhancement of the Empirical Mode Decomposition (EMD) and Ensemble Empirical Mode Decomposition (EEMD) methods.

Key Features of CEEMDAN:

1) Adaptive Noise Addition: CEEMDAN adds adaptive noise to the signal multiple times to address mode mixing issues, which occur when different oscillatory modes are combined into a single IMF.

2) Ensemble Approach: By averaging the results of multiple decompositions with different noise realizations, CEEMDAN provides more stable and accurate IMFs.

3) Iterative Process: CEEMDAN iteratively refines the IMFs by subtracting the noise-adapted mean from the original signal.

Steps in CEEMDAN

1) Add Adaptive Noise: Add different realizations of white noise to the original signal.

2) Decompose with EMD: Apply EMD to each noisy signal to obtain the IMFs.

3) Compute Ensemble Mean: Calculate the mean of the corresponding IMFs from all noisy signals.

4) Iterative Refinement: Subtract the ensemble mean from the original signal and repeat the process on the residual signal to extract the next IMF

Applications:

Signal Processing: Decomposing non-stationary and nonlinear signals in fields like geophysics, biomedicine, and engineering.

Fault Diagnosis: Identifying faults in mechanical systems by analyzing vibration signals.

Financial Time Series: Analyzing and forecasting stock prices and other economic indicators.

CEEMDAN is particularly useful for its robustness and ability to handle complex signals, providing a clearer and more accurate decomposition compared to traditional methods like EMD and EEMD.

3.2 Analysis of Experimental Results

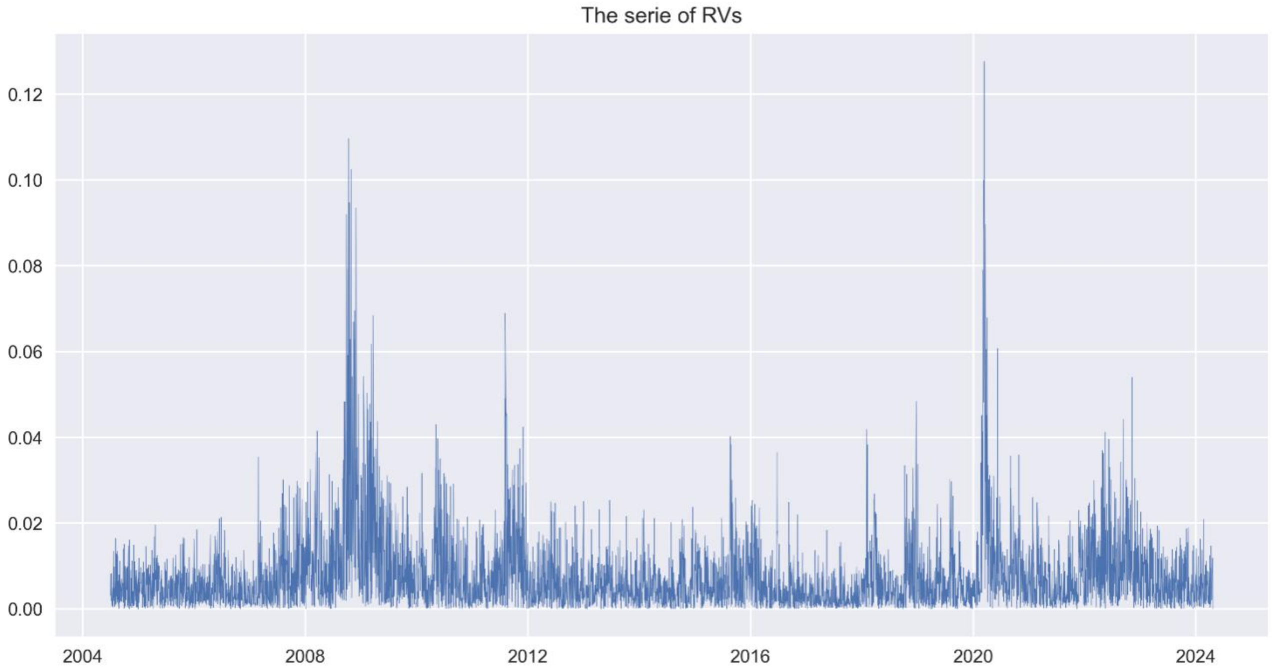

Figure 2 The Series of RVs

From this Figure 2, the stock price is deeply influenced by the international issue, such as: in 2020, the COVID-19 outbroke, the stock price was in serious decline. Therefore, it accorded with the prediction of this model.

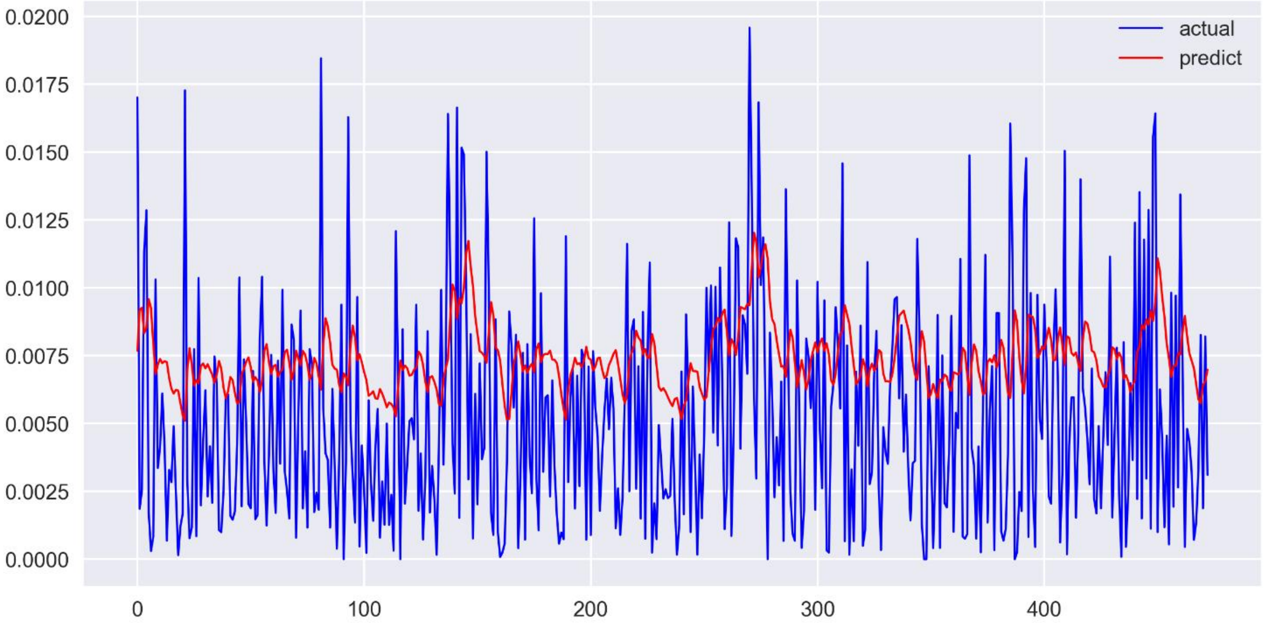

Figure 3 Actual and Prediction Comparison of CEEMDAN

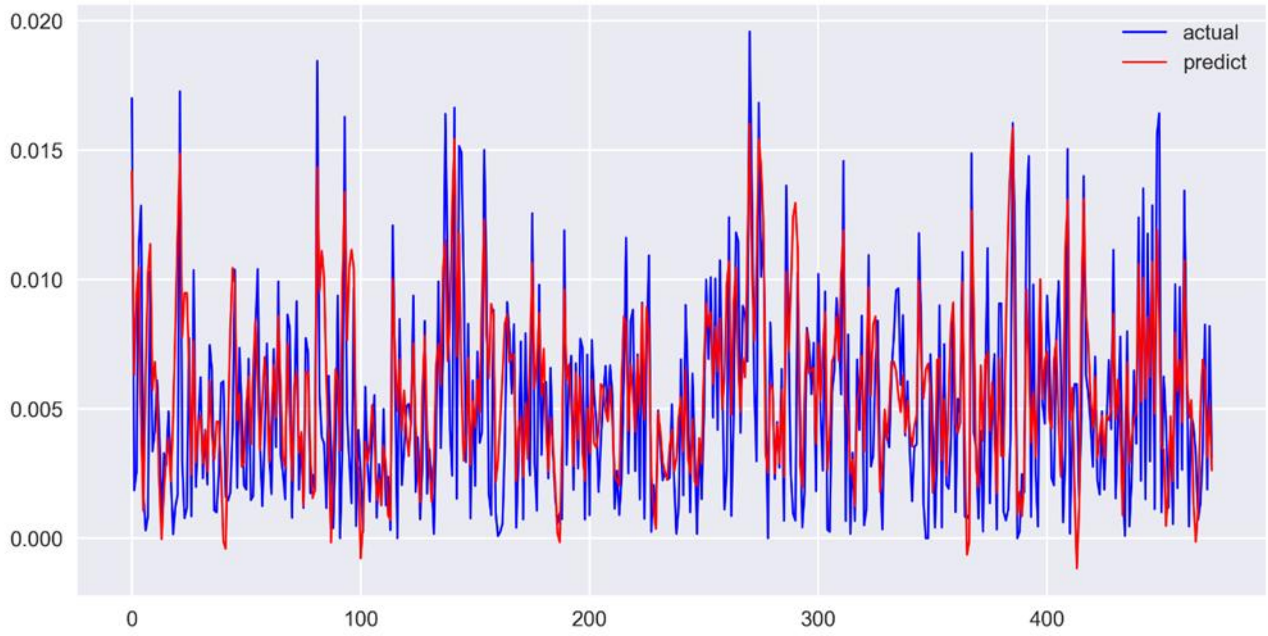

Figure 4 Actual and Prediction Comparison of CEEMDAN LSTM

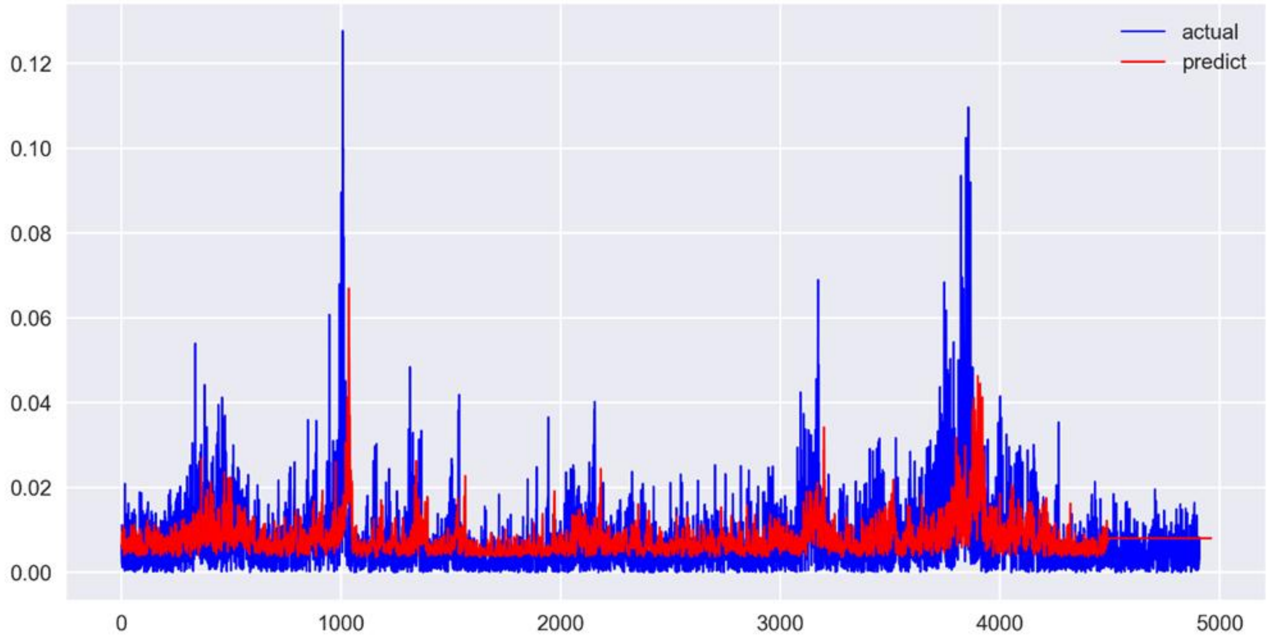

Figure 5 Actual and Prediction Comparison of HAR

Actual and Prediction Comparison of CEEMDAN, CEEMDAN LSTM and HAR can be seen in Figure 3-5.

MAE (Mean Absolute Error) is a metric used to measure the difference between predicted values and actual values, commonly used in regression analysis. Unlike MSE (Mean Squared Error), MAE uses absolute errors instead of squared errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{7}$$

where $\boldsymbol{y_i}$ represents the actual values, $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ represents the predicted values, and n is the number of data points. A lower MAE indicates better model performance. The advantage of MAE is that it is less sensitive to outliers compared to MSE, as it uses absolute values rather than squared values.

MSE (Mean Squared Error) is a metric used to measure the average squared difference between the predicted values and the actual values in a dataset. It's commonly used in statistics and machine learning to evaluate the performance of a regression model.

$$\text{MSE} = \frac{n}{\sum_{i=1}^{n} \frac{1}{(y_i - \widehat{y}_i)^2}} \tag{8}$$

where $\boldsymbol{y_i}$ represents the actual values, $\boldsymbol{\widehat{y}_i}$ represents the predicted values, and n is the number of data points. A lower MSE indicates a better fit of the model to the data.

HMAE (Harmonic Mean Absolute Error) is a less common error metric used to evaluate the performance of predictive models. Unlike MAE (Mean Absolute Error), HMAE uses the harmonic mean of the prediction errors instead of the arithmetic mean. The specific formula is:

$$HMAE = \frac{n}{\sum_{i=1}^{n} \frac{1}{|y_i - \widehat{y}_i|}} \tag{9}$$

where $\boldsymbol{y_i}$ represents the actual values, $\boldsymbol{\widehat{y}_i}$ represents the predicted values, and n is the number of data points.

The advantage of HMAE is that it is more sensitive to large errors (outliers) since the harmonic mean is more affected by extremely small values. In certain situations, this can provide a more useful performance evaluation compared to MAE.

HMSE (Harmonic Mean Squared Error) is a less common error metric used to evaluate the performance of predictive models. Unlike MSE (Mean Squared Error), HMSE uses the harmonic mean of the prediction errors instead of the arithmetic mean. The specific formula is:

$$HMSE = \frac{n}{\sum_{i=1}^{n} \left(\frac{1}{y_i - \widehat{y}_i}\right)^2} \tag{10}$$

where y_i represents the actual values, $\widehat{y}_i$ represents the predicted values, and n is umber of data points.

The advantage of HMSE is that it is more sensitive to outliers since the harmonic mean is more affected by extremely small values. In certain situations, this can provide a more useful performance evaluation compared to MSE.

Through this table 1, LSTM CEEMDAN perform the best compared to the other four models, the prediction is more accurate. Besides, the prediction of hybrid model is most consistent with actual.

Table 1 Model Performance of Five Measures

	CEEMDAN	CEEMDAN LSTM	SVR	HAR	AR
MAE	0.004	0.002	0.059	0.004	0.004
MSE	2.289	1.027	0.003	2.673	2.528
HMAE	0.553	0.430	0.921	1.296	0.545
HMSE	0.420	0.321	0.852	2.920	0.391

4 CONCLUISION

Stock price forecasting is vital to maintain a practical and stable financial market and offer practical guidance for production, operation, and investments. Through python 3.9.13 and various models. This manuscript delineates two fundamental CEEMDAN-LSTM frameworks and introduces a hybrid model integrating CEEMDAN and LSTM. Extensive validations and comparisons have established their efficacy and robustness. The main conclusions are as follows:

By combining LSTM and CEEMDAN, it is possible to more effectively handle the complexity and variability of stock price data, thereby improving predictive performance.

The amalgamation of CEEMDAN and LSTM models endows stock price forecasting with enhanced predictive capabilities. CEEMDAN adeptly decomposes complex, non-linear and non-stationary financial time series data into a set of Intrinsic Mode Functions (IMFs), which are more stable and simpler representations. This decomposition process filters out noise, highlights relevant features, and thus refines the input data for LSTM processing. Subsequently, the LSTM model is better equipped to capture the long-term dependencies and short-term fluctuations inherent in the financial markets, due to the provision ofIMFs at varying scales.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

[1] Singh R, Srivastava S. Stock prediction using deep learning. Multimedia Tools and Applications, 2017, 76: 18569-18584.

[2] Bhattacharjee I, Bhattacharja P. Stock price prediction: a comparative study between traditional statistical approach and machine learning approach. 2019 4th international conference on electrical information and communication technology (EICT), 2019: 1-6

[3] Ariyo AA, Adewumi AO, Ayo CK. Stock price prediction using the ARIMA model. 2014 UKSim-AMSS 16th international conference on computer modelling and simulation, 2014: 106-112.

[4] Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20: 273–97.

[5] Suykens JA, Van Gestel T, De Brabanter J, De Moor B, Vandewalle JP. Least squares support vector machines. World scientific, 2002.

[6] Zhu BZ, Wei YM. Carbon price prediction based on integration of GMDH, particle swarm optimization and least squares support vector machines. Syst Eng-Theory Pract, 2011, 31(12): 2264–71.

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–44

[8] Yahşi M, Çanakoğlu E, Ağralı S. Carbon price forecasting models based on big data analytics. Carbon Manage, 2019, 10(2):175–87.

[9] Holthausen RW, Larcker DF. The prediction of stock returns using financial statement information. Journal of accounting and economics, 1992, 15(2-3): 373-411.

[10] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9(8): 1735–80.

[11] Zhang Z. Research on stock index prediction based on ARIMA-CNN-LSTM model. 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024), 2024, 558-565.

[12] Daubechies I. Ten lectures on wavelets. Society for industrial and applied mathematics, 1992.

[13] Huang NE, Shen Z, Long SR, et al. The empirical mode decomposition and the Hilbert spectrum for non-linear and nonstationary time series analysis. Proc R Soc Lond Ser A Math Phys Eng Sci, 1998, 454(1971): 903–95.

[14] Wu Z, Huang NE. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal, 2009, 1(01): 1–41.

[15] Yeh JR, Shieh JS, Huang NE. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. Adv Adapt Data Anal, 2010, 2(02): 135–56.

[16] Torres ME, Colominas MA, Schlotthauer G, Flandrin P. A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE international conference on acoustics, speech, and signal processing (ICASSP). , 2011, 4144–7.

[17] Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process, 2013, 62(3): 531–44.