Wiley
Discrete Dynamics in Nature and Society
Volume 2024, Article ID 8021444, 14 pages
https://doi.org/10.1155/2024/8021444

WILEY

Research Article

China Crude Oil Futures Volatility Forecasting Using LSTM Model with Optimal Noise Decomposition

Wei Jiang,[1] Wanqing Tang,[2] Huizhi Liu,[3] Yilin Zhou,[4] and Xiao Liu[4]

[1]*School of Economics, Hangzhou Normal University, Hangzhou, China*
[2]*Business School, Hohai University, Nanjing, China*
[3]*Hangzhou Metro Operation Limited Company, Hangzhou, China*
[4]*Shanghai Institute of Tourism, Shanghai Normal University, Shanghai, China*

Correspondence should be addressed to Xiao Liu; cslx2160@shnu.edu.cn

Received 18 January 2024; Revised 8 July 2024; Accepted 7 August 2024

Academic Editor: Rigoberto Medina

The long short-term memory (LSTM) recurrent neural network algorithm in deep learning has demonstrated significant superiority in predicting the realized volatility (RV) of crude oil prices. However, there is no robust and consistent conclusion regarding the handling of microstructural noise from high-frequency data during the prediction process. Therefore, this study utilizes six commonly used data decomposition methods, as documented in the literature, to address the issue of noise handling and decompose the RV series of Chinese crude oil futures. Subsequently, LSTM is integrated with the decomposed data to model and forecast RV. The empirical findings provide compelling evidence that the LSTM model based on neural networks outperforms traditional econometric models in out-of-sample forecasting. Furthermore, the LSTM model with data noise decomposition consistently exhibits superior out-of-sample prediction performance compared to the model without noise decomposition. Among the various data noise decomposition models examined, this study highlights the significant out-of-sample predictive power of variational mode decomposition (VMD), a nonrecursive signal decomposition method, that outperforms other methods. In the scenario of predicting one step ahead, the VMD-LSTM model demonstrates MAE, MSE, and HMAE values of $7.5 \times 10 - 2$, $1.10 \times 10 - 4$, and 0.423, respectively.

1. Introduction

Predicting the volatility of financial assets is an important part of risk management, derivative pricing, and portfolio investment. With the growing influence of China's crude oil futures market on a global scale, much literature has begun to focus on research into the volatility of Chinese crude oil futures (see [1–5]).

As storing and processing financial high-frequency data gets easier, intraday high-frequency data methods are replacing low-frequency data methods for predicting and estimating volatility. Numerous studies show that combining high-frequency data is better at predicting crude oil volatility and uncovering its mechanism than low-frequency data methods (see [6–9]).

Furthermore, the utilization of deep learning models has gained prominence due to their enhanced feature extraction capabilities and reduced constraints when compared to traditional econometric models [10]. Specifically, the LSTM model proposed by Graves et al. [11], which includes the CTC training criterion, has a better ability to handle time series data with long memory compared to the general RNN model. However, there is an ongoing debate about whether the LSTM model is best for predicting volatility calculated from high-frequency financial asset returns, which can be

affected by microstructural noise [12]. In pursuit of bolstering the predictive precision of LSTM models for crude oil futures, current research has predominantly centered on refining LSTM hyperparameter selection. For example, Jovanovic et al. [13] employed an enhanced seagull optimization algorithm (ISOA) to pinpoint optimal LSTM hyperparameters for forecasting crude oil prices. Furthermore, Jovanovic et al. [14] adapted the salp swarm algorithm to hone in on parameters conducive to the long short-term memory model, to enhance the performance and accuracy of WTI crude oil price prediction. Additionally, Jovanovic et al. [15] harnessed an improved Harris Hawks optimization (HHO) algorithm to identify optimal hyperparameters and leveraged the variational mode decomposition (VMD) method to grapple with the intricacies of crude oil time series price data, thereby amplifying the overarching accuracy of crude oil price forecasting. Moreover, previous research suggests that using decomposition to separate noise from the original series can improve predictive accuracy (see [16–19]). Therefore, choosing a suitable decomposition method is crucial in improving the predictive accuracy of the LSTM model.

While previous studies have leveraged data decomposition techniques to enhance time series forecasting accuracy, few have delved into identifying the most effective decomposition method to utilize when employing LSTM models for RV prediction. Thus, the primary objective of this paper is to identify the optimal data decomposition model that enhances the predictive accuracy of LSTM models in forecasting RV in the Chinese crude oil futures market. In this context, the aim is to bridge the existing research gap and justify the necessity of the approach.

In summary, the key contributions of this paper include (1) identification of the optimal data decomposition model to enhance the predictive accuracy of LSTM models for forecasting RV in the Chinese crude oil futures market and (2) bridging the existing research gap concerning the most effective decomposition method for implementing LSTM models in RV prediction.

The remainder of this paper is organized as follows. Section 2 presents a general description of the methods utilized in the empirical analysis. Section 3 describes the data and conducts the empirical research. Section 4 concludes.

2. Methods

2.1. Realized Volatility. In this paper, the intraday futures return is utilized to construct the daily realized volatility:

$$\mathrm{RV}(t) = \sum_{j=1}^{M} r_{t,j}^2, \tag{1}$$

where M is the sampling frequency and $r_{t,j}$ represents the j-th intraday return on day t.

2.2. Data Decomposition. Previous studies have shown that using data decomposition techniques to decompose the original $\mathrm{RV}(t)$ can reduce the impact of noise on RV prediction. The data decomposition techniques selected in this paper are as follows.

2.2.1. Empirical Mode Decomposition (EMD). The EMD proposed by Huang et al. [20] divides the original data into several intrinsic mode functions and a residual component:

$$L(t) = \sum_{j=1}^{n} \mathrm{IMF}_j(t) + R(t). \tag{2}$$

The EMD decomposition process is as follows: First, upper and lower envelope lines are determined based on the extreme points of the original sequence. Then, their mean value is calculated to obtain the mean envelope line. Subtracting the original sequence from this line gives an intermediate sequence. If the IMF condition is met for this sequence, an IMF component is obtained; otherwise, the process is repeated with the intermediate sequence as the new basis. This is done iteratively until all IMF components are obtained and the EMD decomposition is complete.

2.2.2. Ensemble Empirical Mode Decomposition (EEMD). Wu and Huang [21] proposed the EEMD as an improvement to EMD. The EEMD involves adding Gaussian white noise to the original sequence and repeating this step to obtain a set of new sequences:

$$M_i(t) = L(t) + \varepsilon_i(t), \tag{3}$$

where $\varepsilon_i(t)$ is a white noise sequence with mean 0 and standard deviation ε_t.

Then, the EMD decomposition is performed on this set of new sequences:

$$M_i(t) = \sum_{j=1}^{n} \mathrm{IMF}_{\mathrm{ij}}(t) + R_i(t). \tag{4}$$

Finally, the corresponding IMFs are averaged to obtain the EEMD decomposition result:

$$\begin{cases} \mathrm{IMF}_j = \dfrac{1}{m} \displaystyle\sum_{j=1}^{n} \mathrm{IMF}_{\mathrm{ij}}(t), \\ M(t) = \displaystyle\sum_{j=1}^{n} \mathrm{IMF}_j(t) + R(t). \end{cases} \tag{5}$$

EEMD leverages the property of white noise with a mean of zero to mask the inherent signal noise by adding artificially generated noise multiple times, thereby obtaining

more accurate upper and lower envelope lines. However, the introduction of white noise also introduces a new issue of unrecoverable reconstruction errors.

2.2.3. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). Torres et al. [22] proposed the CEEMDAN, which involves adding white noise to the residual terms $R_1^i(t)$ each time the $\mathrm{IMF}_1^i(t)$ component is obtained as follows:

$$\begin{cases} K_i(t) = L(t) + \varepsilon_i(t), \\ K_i(t) = \mathrm{IMF}_1^i(t) + R_1^i(t). \end{cases} \tag{6}$$

Then calculate the mean of the first-order IMF component:

$$\overline{\mathrm{IMF}_1} = \frac{1}{m}\sum_{i=1}^{n} \mathrm{IMF}_1^i. \tag{7}$$

Remove $\overline{\mathrm{IMF}_1}$ from the original sequence to get a new sequence $N_1(t) = L(t) - \overline{\mathrm{IMF}_1}$, and then $N_1(t)$. The following results are obtained through successive iterations:

$$N(t) = \sum_{j=1}^{n} \mathrm{IMF}_j(t) + R(t). \tag{8}$$

2.2.4. Improved CEEMDAN (ICEEMDAN). Colominas et al. [23] proposed ICEEMDAN to address the issues of residual noise and spurious modes in CEEMDAN. The specific formula of ICEEMDAN is as follows.

First, a set of white noise is added to the original sequence to obtain a new sequence:

$$L_i(t) = L(t) + \rho_0 E_1(\varepsilon_i(t)). \tag{9}$$

The new sequence is subjected to EMD decomposition to obtain the first set of residuals:

$$R_1(t) = \langle C(L_i(t)) \rangle. \tag{10}$$

We obtain the $\mathrm{IMF}_1(t) = L(t) - R_1(t)$ and then continue to add white noise and use local mean to obtain the second set of residuals:

$$R_2(t) = \langle C(R_1(t) + \rho_1 E_2(\varepsilon_i(t))) \rangle. \tag{11}$$

Then, we obtain the $\mathrm{IMF}_2 = R_1(t) - R_2(t)$ and repeat this process until the n-th set of residuals and the n-th IMF component are obtained as follows:

$$\begin{cases} R_n(t) = \langle C(R_{n-1}(t) + \rho_{n-1} E_n(\varepsilon_i(t))) \rangle, \\ IMF_n(t) = R_{n-1}(t) - R_n(t), \end{cases} \tag{12}$$

where ρ_n used to remove noise, $E_n(\cdot)$ denotes the IMFs component generated by the EMD decomposition, $\langle \cdot \rangle$ represents the operation of averaging, and $C(\cdot)$ is the operator of producing the local mean of the original series.

2.2.5. Variational Mode Decomposition (VMD). The VMD [24] is a nonrecursive variational mode decomposition signal analysis method. The constraints of the variational problem can be formulated as follows:

$$\begin{cases} \min\limits_{\{u_k\},\{\omega_k\}} \left\{ \sum\limits_k \left\| \partial_t \left[\left(\delta(t) + \dfrac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \\ \text{s.t.} \ \sum\limits_k u_k = f, \end{cases} \tag{13}$$

where $\{u_k\} = \{u_1, \ldots, u_k\}$ is the set of all modes and $\{\omega_k\} = \{\omega_1, \ldots, \omega_k\}$ represents the center frequency sequence. To get more optimal results, the process could be summarized as follows:

$$L(\{u_k\}, \{\omega_k\}, \lambda) \coloneqq \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle, \tag{14}$$

where $\lambda(t)$ is the Lagrange multiplier and α is a quadratic factor.

The optimal solution for equation (14) can be obtained by iteratively updating the modal components and their corresponding center frequencies using the multiplier alternating direction method. The optimal solutions for the modal components and center frequencies are as follows:

$$\begin{cases} \widehat{u_k}(\omega) \coloneqq \dfrac{\widehat{L}(\omega) - \sum_{i \neq k} \widehat{u_i}(\omega) + \widehat{\lambda}(\omega)/2}{1 + 2\alpha(\omega - \omega_k)^2}, \\ \omega_k \coloneqq \dfrac{\int_0^{+\infty} \omega \left| \widehat{u_k}(\omega) \right|^2 \mathrm{d}\omega}{\int_0^{+\infty} \left| \widehat{u_k}(\omega) \right|^2 d\omega}. \end{cases} \tag{15}$$

2.2.6. Empirical Wavelet Transform (EWT). The EWT is a combination of EMD and wavelet transform. According to Gilles [11], the empirical scale function $\widehat{\varphi_n}(\omega)$ and empirical wavelet function $\widehat{\psi_n}(\omega)$ are defined as follows:

$$\widehat{\varphi_n}(\omega) = \begin{cases} 1, & |\omega| \le (1-\gamma)\omega_n, \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}\left(|\omega| - (1-\gamma)\omega_n\right)\right)\right], & (1-\gamma)\omega_n \le |\omega| \le (1+\gamma)\omega_n, \\ 0, & \text{others}, \end{cases} \tag{16}$$

$$\widehat{\psi_n}(\omega) = \begin{cases} 1, & (1+\gamma)\omega_n \le |\omega| \le (1-\gamma)\omega_{n+1}, \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_{n+1}}\left(|\omega| - (1-\gamma)\omega_{n+1}\right)\right)\right], & (1-\gamma)\omega_{n+1} \le |\omega| \le (1+\gamma)\omega_{n+1}, \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}\left(|\omega| - (1-\gamma)\omega_n\right)\right)\right], & (1-\gamma)\omega_n \le |\omega| \le (1+\gamma)\omega_n, \\ 0, & \text{others}, \end{cases} \tag{17}$$

where ω_n represents the midpoint between two adjacent maxima of Fourier spectrum, $\gamma < \min\left(\omega_{n+1} - \omega_n / \omega_{n+1} + \omega_n\right)$.

$$\begin{cases} \beta(x) = x^4\left(35 - 84x + 70x^2 - 20x^3\right), \\ w_L^e(n,t) = \langle L, \psi_n \rangle = \left(\widehat{L(\omega)}, \overline{\psi_n(\omega)}\right)^v, \\ w_L^e(0,t) = \langle L, \varphi_1 \rangle = \left(\widehat{L(\omega)}, \overline{\varphi_1(\omega)}\right)^v, \end{cases} \tag{18}$$

where $\beta(x)$ is the conversion function, $w_L^e(n,t)$ is the detail function, and $w_L^e(0,t)$ is the approximation coefficient.

The empirical mode function is defined as follows:

$$\begin{cases} f_0(t) = w_L^e(0,t) * \varphi_1(t), \\ f_k(t) = w_L^e(k,t) * \psi_k(t). \end{cases} \tag{19}$$

2.3. LSTM Model. The setting of the cell state and gate mechanism in the LSTM model makes it easier to reset, update, and remember long-term information. The process of an LSTM unit is shown as follows.

First, the input gate i_t extracts new information from the input x_t and creates a candidate value $\widetilde{c_t}$ to update the state:

$$\begin{cases} i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \\ \widetilde{c_t} = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right). \end{cases} \tag{20}$$

Next, the forget gate f filters and retains historical information that indicates long-term trends, while discarding noncritical information:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right). \tag{21}$$

Then, by removing some information from the old cell and adding the filtered candidate value, the old cell state c_{t-1} is updated to the new cell state c_t:

$$c_t = f_t^* c_{t-1} + i_t^* \widetilde{c_t}. \tag{22}$$

Finally, the output gate o_t filters the updated c_t and calculates the final output based on the new state and the output gate state:

$$\begin{cases} o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \\ h_t = o_t \cdot \tanh\left(c_t\right). \end{cases} \tag{23}$$

2.4. Data Decomposition-LSTM Model. As previously noted, high-frequency realized volatility (RV) data are subject to significant noise. Using historical RV data directly for prediction can negatively impact forecast accuracy. However, decomposing the original RV series can reduce errors caused by noise and enhance prediction accuracy. Given the plethora of data decomposition algorithms available, this study aims to ascertain the optimal algorithm for LSTM-based RV prediction. To this end, six data decomposition-LSTM models are built to forecast RV for Chinese crude oil futures. The design of the model algorithm is inspired by Lin et al. [17], and the following steps are outlined in detail:

(1) Using six different algorithms, the original RV(t) series is decomposed into n intrinsic mode function (IMF) components and a residual term.

(2) To begin, the dataset is divided into three subsets: a training set, a validation set, and a prediction set. The training set data are then used as input for training the respective LSTM prediction model, while the validation set data are employed to evaluate the model's generalization performance. Finally, the prediction set data are utilized to generate the predicted results of $\widehat{\mathrm{IMF}_i}(t)$ and $\widehat{R(t)}$.

(3) The final prediction sequence is derived by reconstructing all the predicted results using

$$\widehat{\mathrm{RV}(t)} = \sum_{i=1}^{n} \widehat{\mathrm{IMF}_i(t)} + \widehat{R(t)}, \quad t = X + Y + 1,\ X + Y + 2, \ldots,\ X + Y + T, \tag{24}$$

where X corresponds to the length of the training set, Y represents the length of the validation set, and T indicates the length of the prediction set. The final prediction sequence is denoted as $\widehat{R(t)}$. The specific steps are shown in Figure 1.

2.5. Evaluation Criteria. According to the previous research results, this paper selects the following common loss functions as evaluation criteria to evaluate the prediction performance of several models:

$$\begin{cases} \mathrm{MAE} = H^{-1} \sum_{t=1}^{H} \lceil \mathrm{RV}(t) - \widehat{\mathrm{RV}(t)} \rceil, \\ \mathrm{MSE} = H^{-1} \sum_{t=1}^{H} (\mathrm{RV}(t) - \widehat{\mathrm{RV}(t)})^2, \\ \mathrm{HMAE} = H^{-1} \sum_{t=1}^{H} \left| 1 - \dfrac{\widehat{\mathrm{RV}(t)}}{\mathrm{RV}(t)} \right|, \end{cases} \tag{25}$$

where H refers to the total length of the predicted sample, $\mathrm{RV}(t)$ represents the ground truth, and $\widehat{\mathrm{RV}(t)}$ represents the predicted value.

To obtain more reliable testing results, the MCS test is utilized to evaluate the out-of-sample predictive ability of each model. This is because relying solely on the loss function as a criterion for judging model performance may not be robust, as it cannot provide statistical information about significant differences in performance.

3. Empirical Results

3.1. Data Analysis. Existing research suggests that 5-minute high-frequency data can strike a balance between the demand for high-frequency sampling and reducing micronoise [25]. Therefore, in this paper, we select the five-minute high-frequency data of Chinese crude oil futures from March 26, 2018, to January 31, 2023, to calculate RV based on equation (1). The original data are sourced from the Tushare database.

Table 1 exhibits the descriptive statistics of the RV of Chinese crude oil futures, and Figure 2 depicts the RV time series of Chinese crude oil futures.

3.2. Data Decomposition. The frequency of data decomposition algorithms in economic and financial time prediction is increasing, but there is no comprehensive comparative analysis of which decomposition method is the best in LSTM model prediction. Therefore, before establishing the LSTM model, this paper chooses six decomposition methods to decompose the RV series.

In addition, the uniqueness of the VMD algorithm itself requires the determination of the number of decomposition layers, denoted as K, before data decomposition. Notably, setting K too small may result in incomplete decomposition and mode mixing, while setting it too large may introduce irrelevant and spurious components. Therefore, the proper determination of the K value is crucial in the VMD algorithm. In this paper, the central frequency method proposed by Huang et al. [26] is employed to determine the number K of modal decomposition.

Table 2 shows the central frequencies of IMFs under different K values. Since the similar frequency of 0.488 stabilizes after $K = 9$, K is set to 8.

The original RV is decomposed into several IMF components and a res term by six algorithms. As shown in Figure 3, all IMFs are arranged from high frequency to low frequency. Each decomposition result in the figure shows that China crude oil futures RV does contain a lot of noise.

Furthermore, descriptive statistics were performed on the subsequence components in this paper, revealing relatively small standard deviations. Additionally, the statistical significance of the Ljung–Box test results at the 1% level provided confirmation of the sequential correlation among subsequences across all decomposition methods.

3.3. Model Training. To expedite the training process of the LSTM, preprocessing of the decomposed data is necessary prior to LSTM modeling.

To strike a balance between computational efficiency and predictive performance, this study sets the learning rate of the LSTM model at 0.001 and the batch size (minibatch) at 32. It is worth noting that in-sample testing is vulnerable to data mining biases as highlighted by some scholars. To circumvent such biases and ensure robustness, the study utilizes an out-of-sample prediction approach based on rolling windows. This method enables predictions for all time points using the most up-to-date data. For evaluating the predictive performance of the LSTM model, the study specifically adopts the rolling time prediction method. In the empirical section, the subsequence data are divided into training, validation, and prediction sets in a ratio of 7 : 2 : 1. Subsequently, the training set data are utilized to train the corresponding LSTM prediction model, and the prediction results are obtained using the prediction set data.

In this paper, we utilized a conventional LSTM model configuration as our foundational architecture. However, we recognize that the selection of model architecture can significantly influence outcomes in practical applications, posing a potential limitation of our research.

To mitigate this concern and bolster the robustness and predictive efficacy of our methodology, future studies will focus on methodically optimizing our deep learning models. Insights gleaned from recent advancements, such as those by Purohit and Panigrahi [27]; Li et al. [28] and so on will guide us in refining model structures and hyperparameter configurations. This strategic enhancement aims to effectively

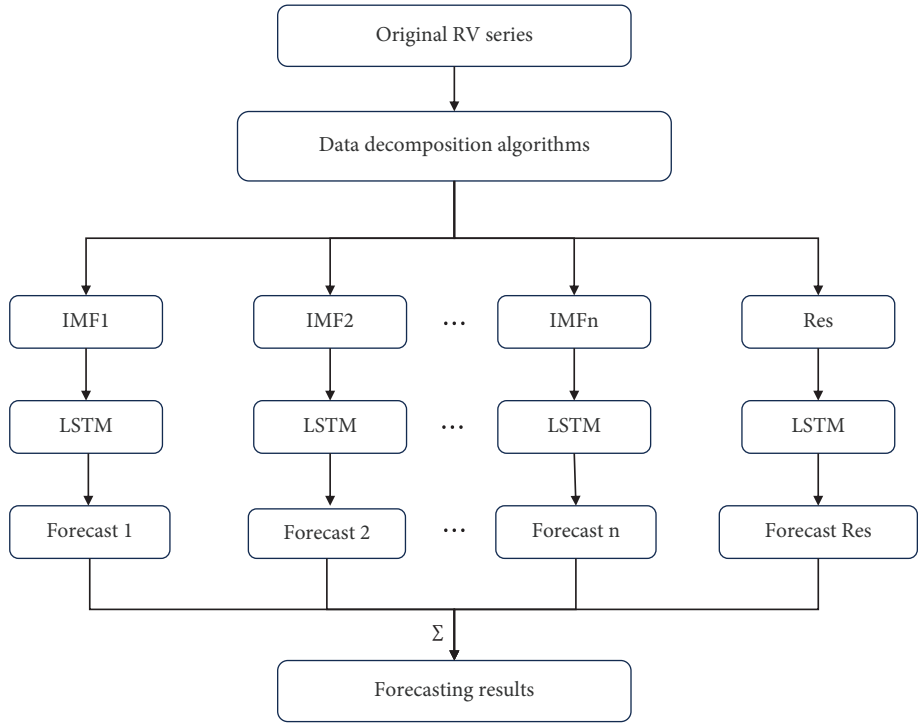


Figure 1: Flowchart of the data decomposition-LSTM model.

Table 1: Descriptive statistics of RV for China's crude oil futures[1].

Mean	Std. dev	Skewness	Kurtosis	JB	Q (5)
0.019	0.018	2.016***	6.143	2647.373***	189.848

Note. This table reports the descriptive statistics of the realized volatility of Chinese crude oil futures. JB denotes the Jarque–Bera statistic, and its null hypothesis is that the sequence follows a normal distribution. Q(5) is the Ljung–Box statistic for up to the 5th-order serial correlation. ***Rejection of the null hypothesis at the 10%, 5%, and 1% significance levels, respectively. [1]Table 1 is reproduced from [9].

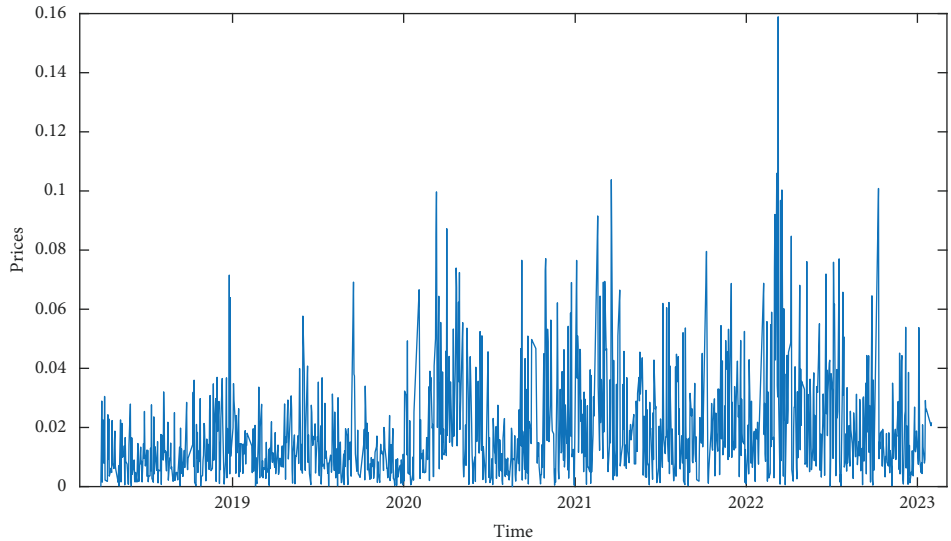


Figure 2: Time series of RV.

Table 2: Central frequencies of IMFs under different K values[2].

K	Central frequencies									
1	0.000									
2	0.001	0.172								
3	0.001	0.327	0.448							
4	0.001	0.171	0.333	0.448						
5	0.001	0.070	0.252	0.347	0.442					
6	0.001	0.163	0.287	0.374	0.439	0.489				
7	0.001	0.134	0.228	0.325	0.382	0.441	0.488			
8	0.000	0.054	0.138	0.228	0.322	0.381	0.440	0.487		
9	0.000	0.065	0.131	0.187	0.271	0.332	0.384	0.441	0.488	
10	0.000	0.056	0.109	0.166	0.226	0.287	0.336	0.385	0.441	0.488

[2]Table 2 is reproduced from [9].

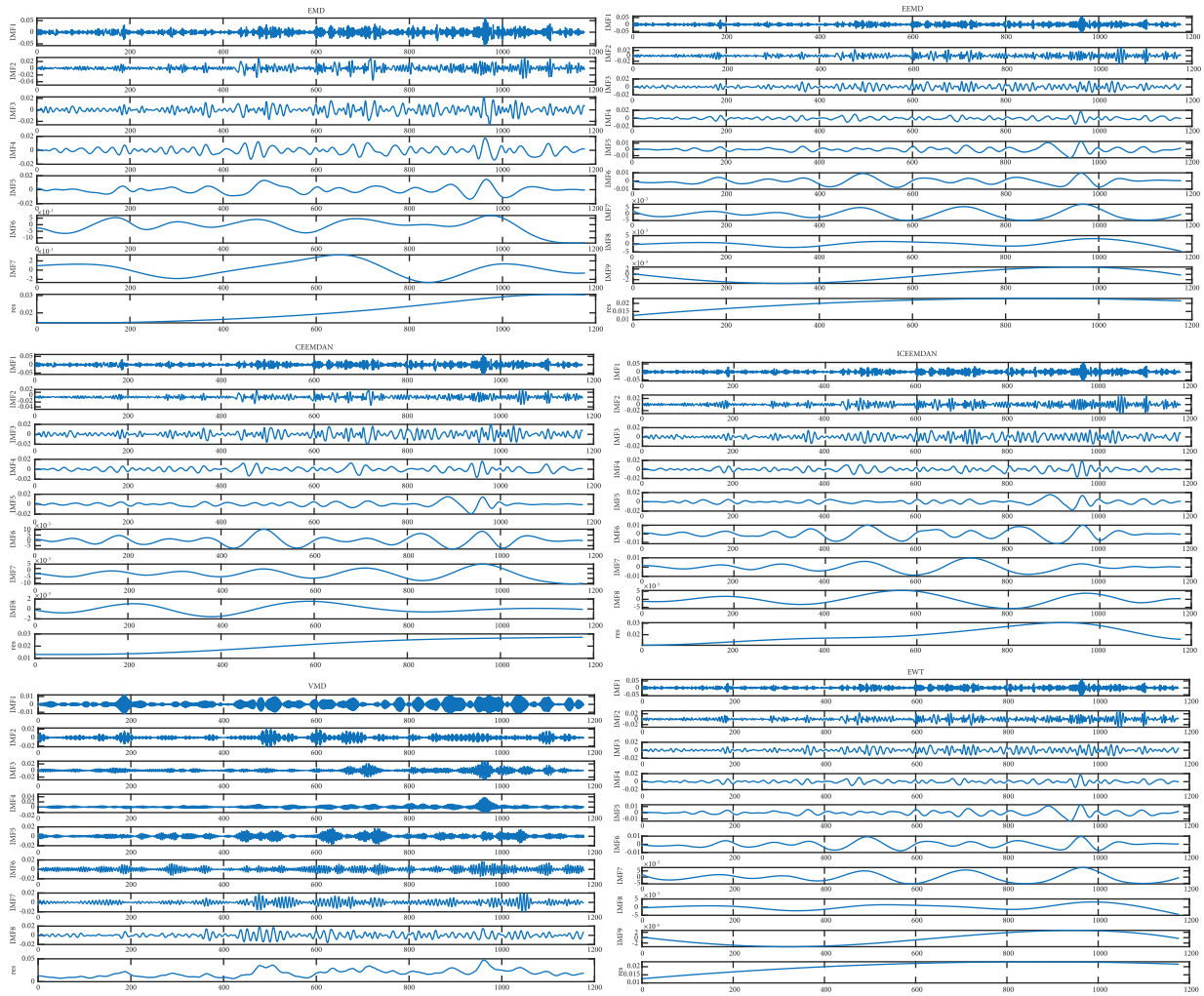


Figure 3: Decomposition results of RV.

elevate both the performance and generalizability of our models.

Specifically, we intend to explore methodologies such as grid search and Bayesian optimization to systematically fine-tune critical parameters of the LSTM model, including hidden unit size, learning rate, and regularization techniques. By doing so, we aim not only to enhance the performance of our model but also to fortify its relevance and reproducibility across diverse datasets and real-world applications.

3.4. Forecasting Performance. To contrast the effectiveness of the selected data decomposition model in improving LSTM prediction accuracy of RV, the predictive results are compared and analyzed based on three error evaluation

indicators: MAE, MSE, and HMAE. In addition, three different time windows are used to evaluate the predictive performance of the model in different time ranges: 1-step ahead (short-term), 5-step ahead (medium-term), and 20-step ahead (long-term).

In addition to the LSTM model, this paper conducts a comparative analysis of two widely used econometric models, namely, the ARIMA model and the HAR model, which have been commonly employed for RV prediction in recent years.

Table 3 shows the results of the nine prediction models across three error evaluation metrics. The findings demonstrate that the data-decomposed LSTM model outperforms the single LSTM model. Moreover, the LSTM model consistently exhibits lower evaluation metric values compared to the ARIMA and HAR models, indicating its superior predictive performance. Notably, the VMD-LSTM model demonstrates the lowest metric values among the three models at various lead times, followed by the EWT-LSTM model, while the EMD-LSTM model performs worst. Specifically, at a lead time of 1 step, the VMD-LSTM model shows a significant reduction in MAE, MSE, and HMAE values by 34.58%, 58.11%, and 46.10%, respectively, in comparison with the single LSTM model. Similarly, the EWT-LSTM model achieves a decrease of 31.39%, 55.68%, and 41.04% in the respective metrics. These results emphasize the effectiveness of the VMD algorithm in enhancing LSTM's predictive accuracy.

Moreover, the EMD-LSTM model, which displayed the least improvement in predictive performance, demonstrated reductions of 17.71%, 39.04%, and 32.75% in the respective loss values. These findings provide compelling evidence for the substantial enhancement of LSTM's predictive efficacy achieved through data decomposition. This phenomenon can be attributed to the direct learning of a unified representation by the LSTM model for time series forecasting, encompassing both noise and trend components derived from the observed data. The amalgamation of these factors may impede the model's capacity to capture genuinely pertinent features, resulting in an ambiguous learning process and consequent overfitting issues that diminish prediction accuracy.

Table 3 also provides additional noteworthy insights. It shows that as lead times increase, the improvement in the predictive performance of all data-decomposed LSTM models also increases. For example, in comparison with the single LSTM model, the VMD-LSTM model exhibits decreases of 76.75% and 81.42% in MAE values at lead times of 5 steps and 20 steps, respectively. These findings emphasize the robustness and reliability of the data-decomposed LSTM model in producing accurate predictions across diverse time intervals.

Enhancing the clarity of comparing the predictive performance of various decomposition models and furnishing additional evidence supporting the superior performance of VMD within data decomposition algorithms, the comparative results of predictions derived from six decomposition models are illustrated alongside those from a singular LSTM model in Figure 4. Overall, decomposing the data and using it as input to the LSTM model significantly improves prediction accuracy. The VMD-LSTM model improves lagging predictions and peak value predictions significantly better than other models. This may be because the VMD considers the narrowband properties of the components, resulting in a more focused filtering frequency band and higher signal-to-noise ratio of the obtained IMFs components. As a result, the feature information input to the LSTM model is more sufficient than others.

The MCS test is selected to further verify whether the VMD is the best model among the six decomposition models. The results are shown in Table 4.

Panel A shows that in the short term, the VMD-LSTM model possesses the highest p value, followed by the EWT-LSTM model. It is noteworthy that only the EWT-LSTM model passed the MCS test solely based on its MAE value, while the other models failed in all indicators. These results imply that the VMD-LSTM model exhibits superior performance compared to the other models in the short term.

Moreover, as illustrated in Panel B and Panel C, the VMD-LSTM model demonstrates a p value of 1 for all indicators. Notably, no other models passed the MCS test in the medium term and long term. These findings suggest that the VMD algorithm leads to a more substantial enhancement in the LSTM model's long-term predictive performance for RV than the other algorithms.

Since the LSTM models are stochastic, the nonparametric WSRT is conducted and the results are presented in Table 5. Table 5 confirms that among the six data decomposition-LSTM models, the VMD-LSTM model provides statistically better MAE, MSE, and HMAE than all other models except the EWT-LSTM model. Compared to the EWT-LSTM model, the VMD-LSTM model provides significant improvements in HMAE. However, the EWT-LSTM model provides statistically equivalent MAE and MSE to the VMD-LSTM model.

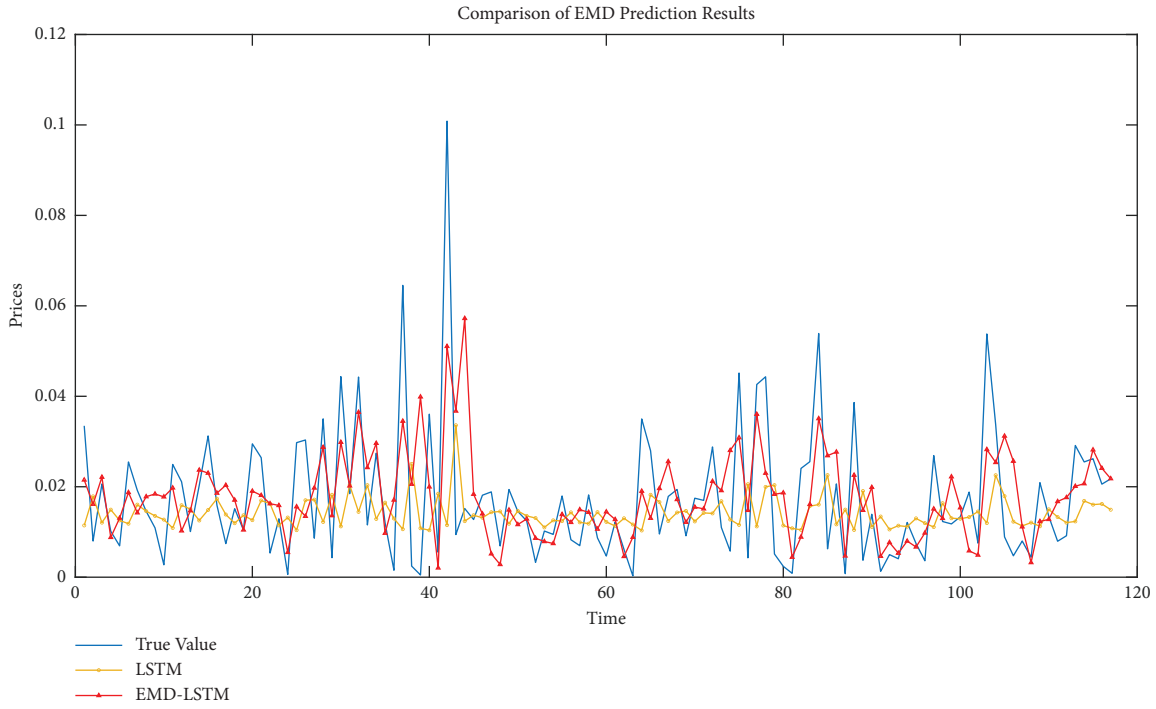
Furthermore, the potential applications of the VMD-LSTM model extend beyond predicting the volatility of Chinese crude oil futures. This versatile approach shows promise for forecasting various other time series datasets across different domains, such as financial indicators like stock prices, exchange rates, or cryptocurrency prices, as well as nonfinancial fields, including energy forecasts like solar power output, electricity demand, or wind power generation. To apply the VMD-LSTM model to different time series, researchers need to identify relevant input variables of interest in specific domains and understand the unique characteristics and dynamics of new time series data, which are crucial for effective model adjustment. This may involve tailored preprocessing steps based on the data features and could potentially necessitate modifications to the model architecture to capture different patterns and trends.

The reason the VMD-LSTM model can be used to forecast time series beyond the realized volatility of Chinese crude oil futures lies in its multifaceted advantages. First, VMD (variational mode decomposition) can effectively decompose complex time series data into several intrinsic mode functions, making the data's features more discernible.

Table 3: Forecasting evaluation results under three loss functions.

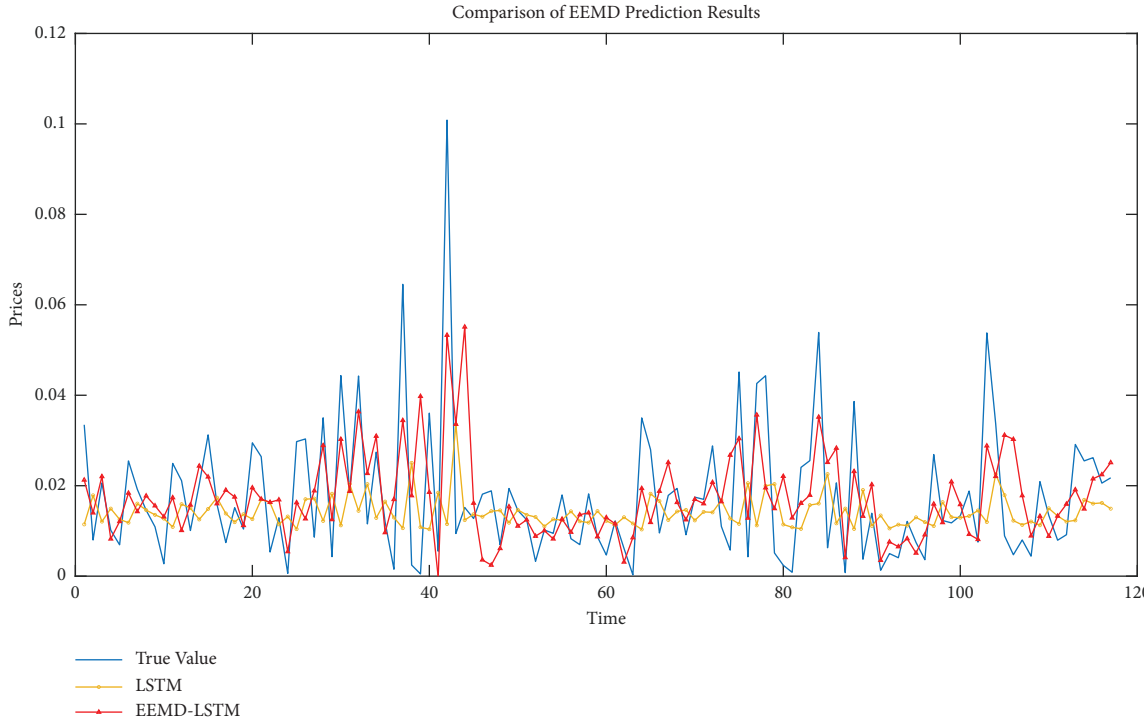
Models	MAE	MSE	HMAE
Panel A: $H = 1$			
LSTM	1.12×10^{-2}	2.64×10^{-4}	0.784
LSTM-EMD	9.25×10^{-2}	1.61×10^{-4}	0.527
LSTM-EEMD	9.01×10^{-2}	1.51×10^{-4}	0.519
LSTM-CEEMDAN	9.25×10^{-2}	1.61×10^{-4}	0.545
LSTM-ICEEMDAN	8.90×10^{-2}	1.49×10^{-4}	0.528
LSTM-VMD	$\mathbf{7.50 \times 10^{-2}}$	$\mathbf{1.10 \times 10^{-4}}$	**0.423**
LSTM-EWT	7.71×10^{-2}	1.17×10^{-4}	0.462
HAR	1.34×10^{-2}	4.00×10^{-4}	1.293
ARIMA	1.16×10^{-2}	2.38×10^{-4}	0.644
Panel B: $H = 5$			
LSTM	1.15×10^{-2}	2.49×10^{-4}	0.688
LSTM-EMD	7.64×10^{-2}	1.04×10^{-4}	0.506
LSTM-EEMD	6.99×10^{-2}	8.70×10^{-5}	0.471
LSTM-CEEMDAN	7.82×10^{-2}	1.07×10^{-4}	0.544
LSTM-ICEEMDAN	6.65×10^{-2}	8.16×10^{-5}	0.444
LSTM-VMD	$\mathbf{2.66 \times 10^{-2}}$	$\mathbf{1.08 \times 10^{-5}}$	**0.140**
LSTM-EWT	3.89×10^{-2}	2.40×10^{-5}	0.201
HAR	1.34×10^{-2}	4.02×10^{-4}	1.286
ARIMA	1.12×10^{-2}	2.34×10^{-4}	0.630
Panel C: $H = 20$			
LSTM	1.22×10^{-2}	2.76×10^{-4}	0.732
LSTM-EMD	7.56×10^{-2}	1.01×10^{-4}	0.413
LSTM-EEMD	7.15×10^{-2}	8.58×10^{-5}	0.385
LSTM-CEEMDAN	7.54×10^{-2}	1.00×10^{-4}	0.403
LSTM-ICEEMDAN	6.86×10^{-2}	8.05×10^{-5}	0.373
LSTM-VMD	$\mathbf{2.26 \times 10^{-2}}$	$\mathbf{8.25 \times 10^{-6}}$	**0.126**
LSTM-EWT	3.55×10^{-2}	2.01×10^{-5}	0.204
HAR	1.42×10^{-2}	4.46×10^{-4}	1.349
ARIMA	1.03×10^{-2}	2.25×10^{-4}	0.581

Notes. Numbers in bold imply that the corresponding model has the lowest loss function among all models. $H = 1$, 5, and 20 represent 1 step ahead, 5 steps ahead, and 20 steps ahead, respectively.

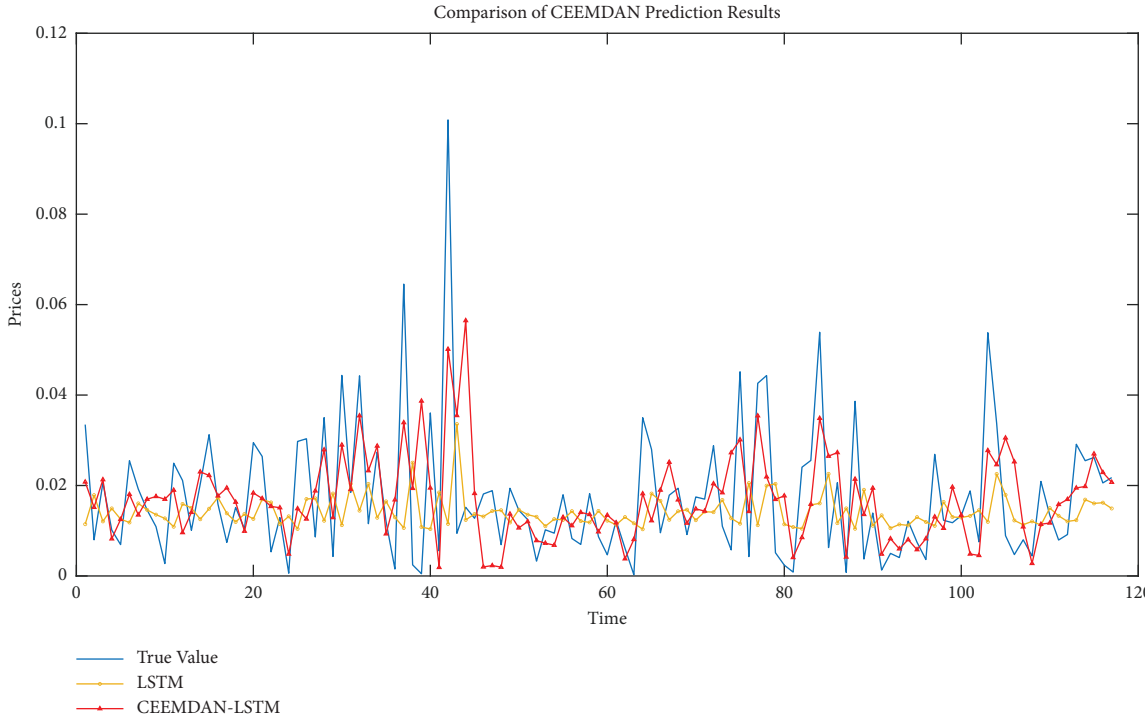


(a)

Figure 4: Continued.

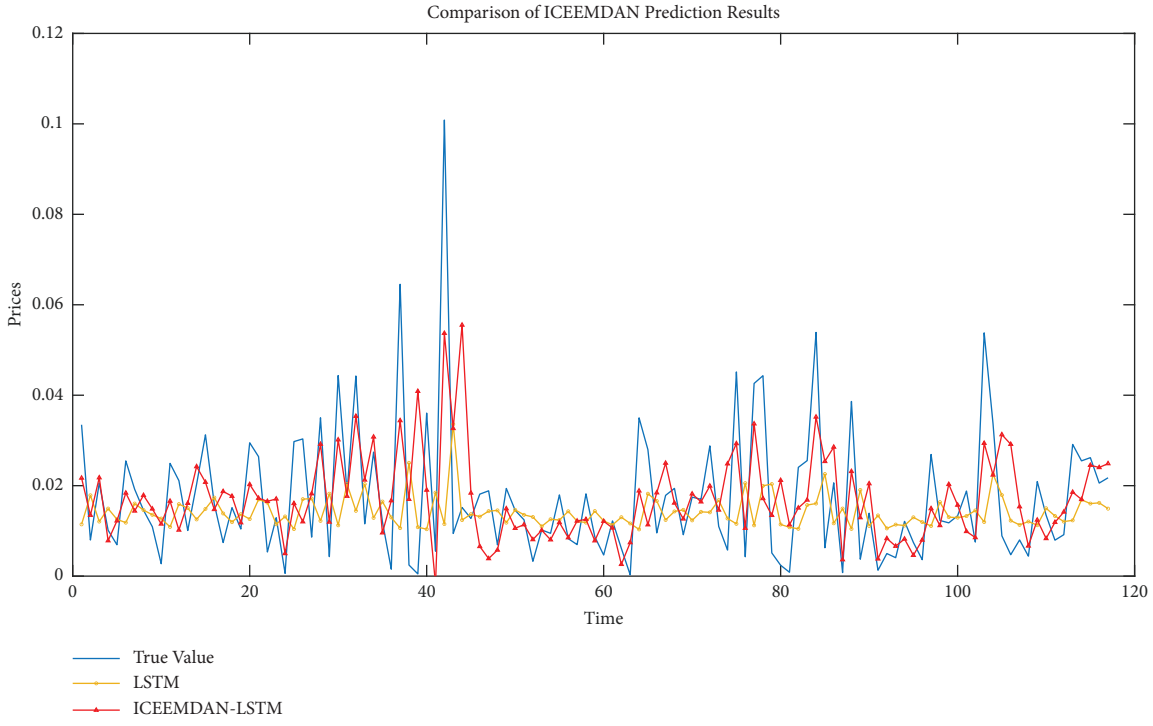


(b)

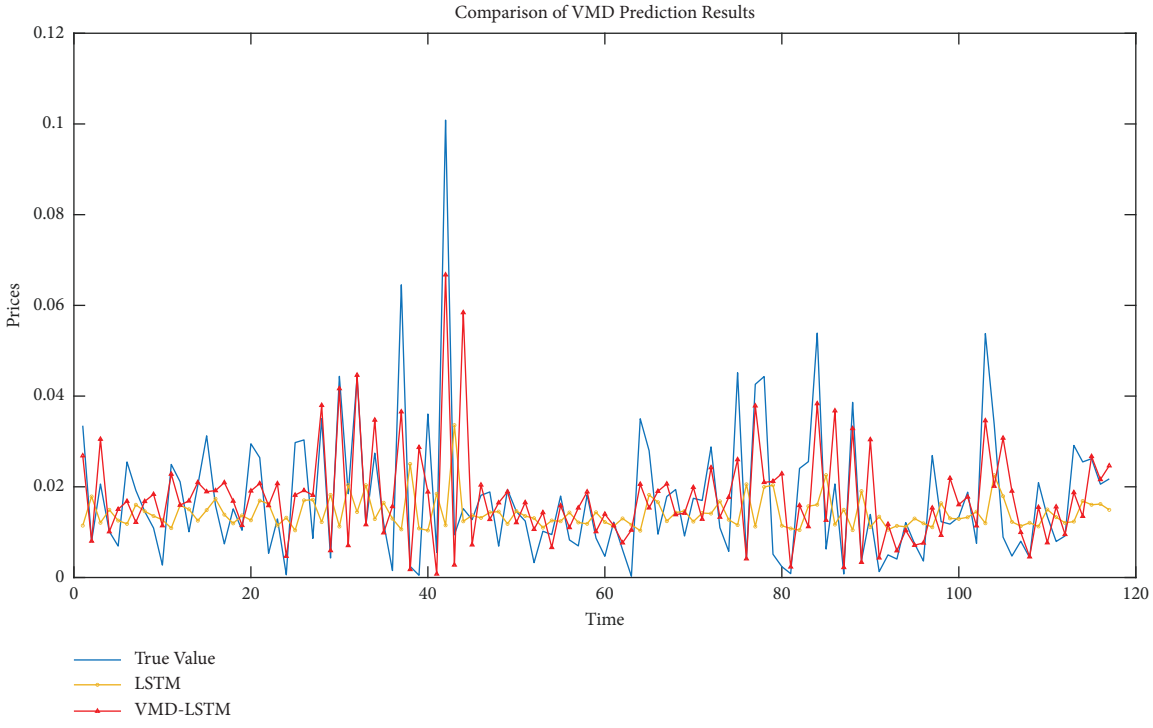


(c)

Figure 4: Continued.

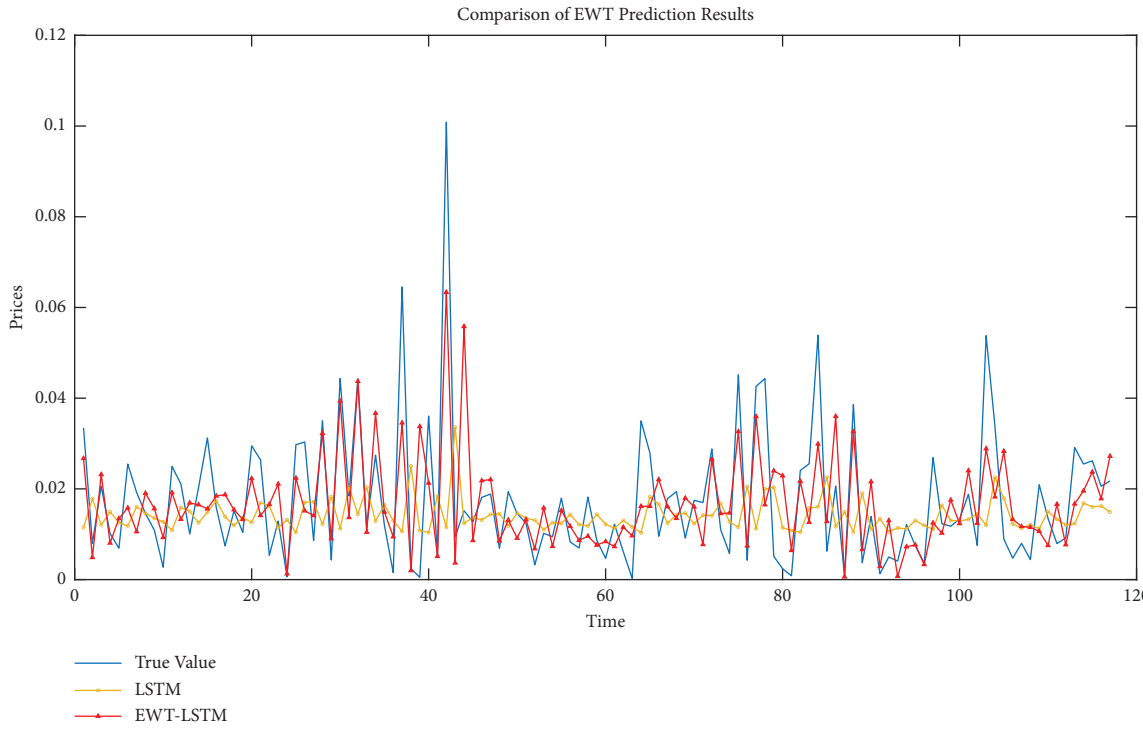


(d)



(e)

Figure 4: Continued.



(f)

Figure 4: Predictive results of different decomposition models.

Table 4: Forecasting evaluation results using the MCS test.

Forecasting models	MAE		MSE		HMAE	
	Range	SeimQ	Range	SeimQ	Range	SeimQ
Panel A: $H = 1$						
LSTM	0.001	0.002	0.055	0.021	0.000	0.000
EMD-LSTM	0.025	0.019	0.136	0.050	0.002	0.004
EEMD-LSTM	0.025	0.019	0.136	0.050	0.002	0.004
CEEMDAN-LSTM	0.025	0.019	0.136	0.050	0.000	0.002
ICEEMDAN-LSTM	0.025	0.021	0.136	0.050	0.000	0.002
VMD-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
EWT-LSTM	*0.292*	*0.292*	0.150	0.150	0.002	0.004
HAR	0.000	0.000	0.017	0.012	0.000	0.000
ARIMA	0.000	0.000	0.027	0.013	0.000	0.000
Panel A: $H = 5$						
LSTM	0.000	0.000	0.007	0.003	0.000	0.000
EMD-LSTM	0.001	0.000	0.013	0.007	0.000	0.000
EEMD-LSTM	0.000	0.000	0.013	0.006	0.000	0.000
CEEMDAN-LSTM	0.000	0.000	0.013	0.006	0.000	0.000
ICEEMDAN-LSTM	0.001	0.000	0.013	0.007	0.000	0.000
VMD-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
EWT-LSTM	0.001	0.000	0.013	0.007	0.050	0.050
HAR	0.000	0.000	0.013	0.004	0.000	0.000
ARIMA	0.000	0.000	0.006	0.003	0.000	0.000

Table 4: Continued.

Forecasting models	MAE		MSE		HMAE	
	Range	SeimQ	Range	SeimQ	Range	SeimQ
Panel A: $H = 22$						
LSTM	0.000	0.000	0.004	0.004	0.000	0.000
EMD-LSTM	0.000	0.000	0.004	0.004	0.000	0.002
EEMD-LSTM	0.000	0.000	0.004	0.004	0.000	0.002
CEEMDAN-LSTM	0.000	0.000	0.004	0.004	0.000	0.002
ICEEMDAN-LSTM	0.000	0.000	0.004	0.004	0.000	0.002
VMD-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
EWT-LSTM	0.000	0.000	0.004	0.004	0.000	0.002
HAR	0.000	0.000	0.004	0.004	0.000	0.000
ARIMA	0.000	0.000	0.004	0.004	0.000	0.000

Notes. The numbers in bold indicate that the corresponding models have best forecasting performance under the MCS criterion. The numbers with p values larger than 0.25 are italic. $H = 1$, 5, 20 represent 1 step ahead, 5 steps ahead, and 20 steps ahead, respectively. Range and SeimQ represent range statistic and semiquadratic statistic, respectively [29].

Table 5: Wilcoxon signed-rank test results.

	MAE	MAE	HMAE
EMD-LSTM	−	−	−
EEMD-LSTM	−	−	−
CEEMD-LSTM	−	−	−
ICEEMD-LSTM	−	−	−
EWT-LSTM	≈	≈	−

Notes. Wilcoxon signed-rank test results indicating the worse (−), better (+), and equivalent (≈) than the VMD-LSTM model in predicting China crude oil futures volatility.

This preprocessing method helps the LSTM model better capture the long-term dependencies and nonlinear dynamic characteristics within the time series. Second, LSTM, as a deep learning model suitable for sequence data modeling, can automatically learn and memorize long-term temporal dependencies during training, thus being suitable for a variety of time series forecasting tasks. Consequently, the integration of VMD and LSTM models for predicting time series beyond the realized volatility of Chinese crude oil futures can better handle the complexity and dynamic characteristics of the data, thereby improving the accuracy and robustness of forecasts.

4. Conclusions

This paper aims to enhance research on the impact of utilizing data decomposition algorithms to enhance LSTM model predictions for RV. With limited research regarding RV forecasting for Chinese crude oil futures, the performance of six frequently utilized data decomposition algorithms is empirically compared. The findings indicate the following.

First, in the context of RV prediction, the LSTM model yields superior outcomes compared to traditional econometric models such as ARIMA as well as more recent models like the widely used HAR. Nonetheless, it is worth noting that the prediction results of all three models are subject to significant lag effects.

Second, the inclusion of decomposed sequences into the LSTM prediction model significantly improves its forecasting efficacy on the RV series of Chinese crude oil futures. This highlights the ability of data decomposition to effectively reduce noise and mitigate errors caused by noise in the prediction model. Furthermore, unlike simplistic noise elimination approaches, data decomposition breaks down the original sequence into multiple IMF components and a residual term. This approach not only smooths the noise but also retains the fundamental market characteristics inherent in the original financial series. The retention of these features contributes significantly to the enhancement of the prediction accuracy.

Thirdly, among the six frequently utilized decomposition algorithms listed in this paper, the VMD algorithm exhibits superior efficacy in enhancing LSTM prediction accuracy and sustains a consistent effect as the number of forecasting steps increases. The EWT algorithm ranks second in terms of effectiveness, while the EMD algorithm shows comparatively inferior performance. However, it is crucial to note that meticulous consideration should be given to the selection of the penalty factor and the number K of modal decomposition for VMD during practical applications, as these choices can significantly influence the efficacy of VMD decomposition. Currently, an optimal parameter combination has not been identified.

Finally, there are still aspects of this study that requires further improvement. Although the study has filled the gap in research on improving the accuracy of LSTM model predictions of RV by comparing data decomposition algorithms, it only uses a single RV sequence as input data. For future endeavors aimed at enhancing the accuracy of Chinese crude oil futures RV predictions, we will consider incorporating additional variables, such as the climate factors that have received widespread attention recently.

Data Availability

The data used to support the findings of this study will be made available on request.

Conflicts of Interest

The authors declare that they have no known conflicts of financial interests or personal relationships that could have appeared to influence the work reported in this study.

Acknowledgments

This work was supported by the Zhejiang Province Philosophy and Social Science Planning Project (24NDJC222YBM), the National Natural Science Foundation of China (72171170), and School-level Cultivation Program of Young Interdisciplinary Innovation Team for Humanities and Social Sciences Research at Shanghai Normal University (310AWO20323005411).

References

[1] X. Lu, F. Ma, J. Wang, and J. Wang, "Examining the predictive information of CBOE OVX on China's oil futures volatility: evidence from MS-MIDAS models," *Energy*, vol. 212, 2020.

[2] Z. Niu, Y. Liu, W. Gao, and H. Zhang, "The role of coronavirus news in the volatility forecasting of crude oil futures markets: evidence from China," *Resources Policy*, vol. 73, 2021.

[3] D. Jin, M. He, L. Xing, and Y. Zhang, "Forecasting China's crude oil futures volatility: how to dig out the information of other energy futures volatilities?" *Resources Policy*, vol. 78, 2022.

[4] X. Yan, J. Bai, X. Li, and Z. Chen, "Can dimensional reduction technology make better use of the information of uncertainty indices when predicting volatility of Chinese crude oil futures?" *Resources Policy*, vol. 75, 2022.

[5] Y. Huang, W. Xu, D. Huang, and C. Zhao, "Chinese crude oil futures volatility and sustainability: an uncertainty indices perspective," *Resources Policy*, vol. 80, 2023.

[6] F. Ma, Y. Wei, W. Chen, and F. He, "Forecasting the volatility of crude oil futures using high-frequency data: further evidence," *Empirical Economics*, vol. 55, pp. 653–678, 2018.

[7] M. Liu and C. C. Lee, "Capturing the dynamics of the China crude oil futures: markov switching, co-movement, and volatility forecasting," *Energy Economics*, vol. 103, 2021.

[8] X. Li, Y. Liao, X. Lu, and F. Ma, "An oil futures volatility forecast perspective on the selection of high-frequency jump tests," *Energy Economics*, vol. 116, 2022.

[9] W. Jiang, W. Tang, and X. Liu, "Forecasting realized volatility of Chinese crude oil futures with a new secondary decomposition ensemble learning approach," *Finance Research Letters*, vol. 57, 2023.

[10] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models," *Expert Systems with Applications*, vol. 103, pp. 25–37, 2018.

[11] A. Graves, "Generating sequences with recurrent neural networks," 2013, https://arxiv.org/abs/1308.0850.

[12] L. Zhang, P. A. Mykland, and Y. Aït-Sahalia, "A tale of two time scales: determining integrated volatility with noisy high-frequency data," *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1394–1411, 2005.

[13] L. Jovanovic, N. Bacanin, A. Jovancai et al., "Oil price prediction approach using long short-term memory network tuned by improved seagull optimization algorithm," in *International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering and Technology*, pp. 253–265, Springer Nature Singapore, Singapore, 2022.

[14] L. Jovanovic, D. Jovanovic, N. Bacanin et al., "Multi-step crude oil price prediction based on lstm approach tuned by salp swarm algorithm with disputation operator," *Sustainability*, vol. 14, no. 21, 2022.

[15] L. Jovanovic, M. Antonijevic, M. Zivkovic et al., "Long short-term memory tuning by enhanced Harris hawks optimization algorithm for crude oil price forecasting," *Advances in Computers*, vol. 135, 2024.

[16] B. Wang and J. Wang, "Deep multi-hybrid forecasting system with random EWT extraction and variational learning rate algorithm for crude oil futures," *Expert Systems with Applications*, vol. 161, 2020.

[17] Y. Liang, Y. Lin, and Q. Lu, "Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM," *Expert Systems with Applications*, vol. 206, 2022.

[18] W. Sun, H. Chen, F. Liu, and Y. Wang, "Point and interval prediction of crude oil futures prices based on chaos theory and multiobjective slime mold algorithm," *Annals of Operations Research*, vol. 31, 2022.

[19] M. Ling and G. Cao, "Carbon trading price forecasting based on parameter optimization VMD and deep network CNN–LSTM model," *International Journal of Financial Engineering*, vol. 11, 2024.

[20] 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, pp. 903–995, 1998.

[21] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[22] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4144–4147, Prague, Czech Republic, May 2011.

[23] M. A. Colominas, G. Schlotthauer, and M. E. Torres, "Improved complete ensemble EMD: a suitable tool for biomedical signal processing," *Biomedical Signal Processing and Control*, vol. 14, pp. 19–29, 2014.

[24] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, pp. 531–544, 2014.

[25] T. G. Andersen and T. Bollerslev, "Answering the skeptics: yes, standard volatility models do provide accurate forecasts," *International Economic Review*, vol. 39, pp. 885–905, 1998.

[26] N. Huang, Y. Wu, G. Cai et al., "Short-term wind speed forecast with low loss of information based on feature generation of OSVD," *IEEE Access*, vol. 7, pp. 81027–81046, 2019.

[27] S. K. Purohit and S. Panigrahi, "Novel deterministic and probabilistic forecasting methods for crude oil price employing optimized deep learning, statistical and hybrid models," *The Information of the Science*, vol. 658, 2023.

[28] Y. Li, J. Liu, and Y. Teng, "A decomposition-based memetic neural architecture search algorithm for univariate time series forecasting," *Applied Soft Computing*, vol. 130, 2022.

[29] P. R. Hansen, A. Lunde, and J. M. Nason, "The model confidence set," *Econometrica*, vol. 79, no. 2, pp. 453–497, 2011.