Finance Research Letters 57 (2023) 104254

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Forecasting realized volatility of Chinese crude oil futures with a new secondary decomposition ensemble learning approach

Wei Jiang [a,1], Wanqing Tang [a,1], Xiao Liu [b,2,*]

[a] School of Economics, Hangzhou Normal University, Hangzhou, China
[b] Shanghai Institute of Tourism, Shanghai Normal University, Shanghai, China

ARTICLE INFO

JEL Classification:
B41
C22
C53

Keywords:
Crude oil futures
Realized volatility
VMD model
ICEEMDAN model
Secondary decomposition

ABSTRACT

This study proposes a new VMD-ICEEMDAN-LSTM model, which combines secondary decomposition with long short-term memory neural networks (LSTM) to forecast the realized volatility (RV) of Chinese crude oil futures. The RV sequence is first decomposed into subcomponents and residuals through variational mode decomposition (VMD). Then, the iterative complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) is applied to perform a secondary decomposition on the residuals. Finally, we apply LSTM to forecast all decomposed components, and then combine all forecasting results to obtain our final forecast value. Our results show that the VMD-ICEEMDAN-LSTM model significantly outperforms existing individual and combination models.

1. Introduction

Energy is an indispensable driver of global economic development. The crude oil futures market, as the most mature and open financial energy trading market, plays an important role in international energy markets. Therefore, the accurate prediction of crude oil futures price fluctuations has become a focus of attention in academia.

The three main approaches to forecasting crude oil futures prices are traditional econometric models, artificial intelligence models, and combination models. Traditional econometric models rely on the fulfillment of specific statistical assumptions prior to modeling, including assumptions about the distribution of returns and the stationarity of the data. Moreover, these models necessitate intricate data processing and transformation, which restricts their practical application in economic and management problems (Taylan, 2017; Kim and Won, 2018). In contrast, machine learning methods do not depend on predetermined assumptions regarding data distribution or stationarity, thus making them more flexible and applicable in various contexts. Moreover, these methods struggle to capture the nonlinear relationships and long-term memory effects present in time series data, both of which are characteristics of crude oil futures price volatility, while artificial intelligence models can handle complex time series data and achieve higher predictive accuracy (Christensen et al., 2022). Neural network models, particularly the LSTM model, have been widely recognized for their ability to handle high-frequency realized volatility data with long-term memory characteristics (see, e.g., Kim and Won, 2018; Liu, 2019; Guo et al., 2022). The LSTM model has excellent descriptive capabilities for time series data with long memory, as it can effectively transmit

* Corresponding author at: No.100 Guilin Rd, Xuhui District, Shanghai 200030, China.
E-mail address: cslx2160@shnu.edu.cn (X. Liu).
[1] Postal address: No. 2318 Yuhangtang Rd, Yuhang District, Hangzhou 311121, China.
[2] Postal address: No.100 Guilin Rd, Xuhui District, Shanghai 200030, China.

https://doi.org/10.1016/j.frl.2023.104254
Received 31 May 2023; Received in revised form 6 July 2023; Accepted 23 July 2023
Available online 24 July 2023
1544-6123/

and express information in long-time sequences without "forgetting" previous information.

Although the LSTM model has greatly improved prediction accuracy, microstructural noise in high-frequency data can impact its accuracy. To overcome this problem, researchers have turned to combination models. One common approach is to decompose the original sequence using signal processing methods, apply predictive models to each component obtained from the decomposition, and then sum the predicted values of all subsequences to obtain the final prediction result of the original sequence (Li and Tam, 2017). VMD is a widely used method for data decomposition that can effectively enhance the signal-to-noise ratio of the decomposed components, thereby resulting in more comprehensive input feature information for the predictive model and significantly improved prediction accuracy (Dragomiretskiy and Zosso, 2014). Combining VMD with machine learning models has been recognized as an effective way to further improve prediction accuracy (see, e.g., He et al., 2019; Liu et al., 2020).

Many existing studies only perform a single decomposition and directly discard the residual term that encompasses numerous features. However, this practice undermines the effectiveness of data decomposition (Abdoos, 2016; Lahmiri, 2017). Moreover, other studies do not consider that the residual term, after VMD decomposition, still has a high degree of complexity and directly predicted it (e.g., Niu et al., 2020; Yang et al., 2021; Huang and Deng, 2021), thereby resulting in insufficient prediction accuracy. To improve the model's overall prediction accuracy, we aimed to enhance the prediction accuracy of the residual term. Herein, we propose a model for the secondary decomposition of VMD using the residual term that differs from those of Zhan & Tang (2022), Zhang et al. (2022) and Tang et al. (2023), who used EEMD for secondary decomposition of the VMD residual term. This paper takes into consideration the problem of reconstruction error that cannot be eliminated by EEMD, while Colominas et al. (2014) proposed ICEEMDAN, which not only improved this issue but also addressed the defects of residual noise and pseudo-modes after decomposition. Therefore, we decided to use ICEEMDAN to perform secondary decomposition on the residual term obtained from VMD decomposition and proposed the VMD-ICEEMDAN-LSTM model for the first time. The characteristics of this model are as follows.

(1) This paper addresses two issues in the existing research on handling residual components after VMD decomposition. Specifically, that (a) discarding residual components that contain many valuable features weakens data decomposition and causes distortion, and (b) directly predicting high-complexity residual components leads to insufficient prediction accuracy. This study improves prediction accuracy using quadratic decomposition of residual components after VMD decomposition.
(2) To address the issue of unrecoverable reconstruction error in EEMD, this paper utilizes ICEEMDAN to perform secondary decomposition on the residual components obtained from VMD decomposition. This improves the residual components' as well as the model's overall prediction accuracy. Previous research has demonstrated that ICEEMDAN is effective in addressing this issue.
(3) Compared to traditional statistical models, LSTM automatically extracts crucial feature information from crude oil futures price volatility, thus eliminating the need for manual feature selection. This enhances its effectiveness, flexibility, and user-friendliness in handling nonlinear and nonstationary data. Additionally, LSTM captures the extended dependencies of sequential data, thus leading to improved forecasting of market volatility trends. Hence, this study employs an LSTM model to predict the realized volatility of crude oil futures.

The remainder of this paper is organized as follows. Section 2 provides a general description of the methods that we use in our empirical analysis. Section 3 describes the data and provides a preliminary analysis. Section 4 conducts the empirical research. Section 5 concludes.

2. Methods

2.1. Realized volatility

In this paper, we use intraday futures returns to construct the daily realized volatility as follows:

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

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

2.2. Data decomposition

Previous research confirms that decomposing the original $RV(t)$ using data decomposition techniques can mitigate noise-induced uncertainty in RV prediction (see, e.g., Wang and Wang, 2020; Liang et al., 2022). This paper introduces a VMD-ICEEMDAN-LSTM model that utilizes data decomposition and a deep learning LSTM. The details of the VMD and ICEEMDAN techniques are as follows.

2.2.1. VMD

The VMD method is a non-recursive variational mode decomposition signal analysis method. The constraints of the variational problem can be formulated as follows:

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

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 obtain more optimal results, the process can be summarized as follows:

$$\begin{aligned} L(\{u_k\},\{\omega_k\},\lambda) := \alpha \sum_k \| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \|_2^2 \\ + \| f(t) - \sum_k u_k(t) \|_2^2 + < \lambda(t), f(t) - \sum_k u_k(t) > \end{aligned} \tag{3}$$

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

The optimal solution for Eq. (3) 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:

$$\begin{cases} \widehat{u}_k^{n+1}(\omega) := \dfrac{\widehat{f}(\omega) - \sum\limits_{i \neq k} \widehat{u}_i(\omega) + \dfrac{\widehat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \\ \omega_k^{n+1} := \dfrac{\int_0^{+\infty} \omega |\widehat{u}_k(\omega)|^2 d\omega}{\int_0^{+\infty} |\widehat{u}_k(\omega)|^2 d\omega} \end{cases} \tag{4}$$

where $\widehat{u}_k^{n+1}, \widehat{f}(\omega), \widehat{u}_k(\omega)$ and $\widehat{\lambda}(\omega)$ decompose the Fourier transforms into $u_k^{n+1}(\omega), f(\omega), u_i(\omega)$ and $\lambda(\omega)$.

2.2.2. ICEEMDAN

Colominas et al. (2014) proposed ICEEMDAN to address the issues of residual noise and spurious modes in CEEMDAN. The specific formula for ICEEMDAN is as follows:

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

$$L_i(t) = L(t) + \rho_0 E_1(\varepsilon_i(t)). \tag{5}$$

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

$$R_1(t) = \langle C(L_i(t)) \rangle. \tag{6}$$

Next, $IMF_1(t) = L(t) - R_1(t)$ is obtained. Then, white noise is continued to be added and the local mean is used to obtain the second set of residuals:

$$R_2(t) = \langle C(R_1(t) + \rho_1 E_2(\varepsilon_i(t))) \rangle. \tag{7}$$

Then, $IMF_2 = R_1(t) - R_2(t)$ is obtained. This process is repeated until the n-th set of residuals and the n-th IMF component are obtained:

$$\begin{cases} R_n(t) = \langle C(R_{n-1}(t) + \rho_{n-1} E_n(\varepsilon_i(t))) \rangle \\ IMF_n(t) = R_{n-1}(t) - R_n(t) \end{cases} \tag{8}$$

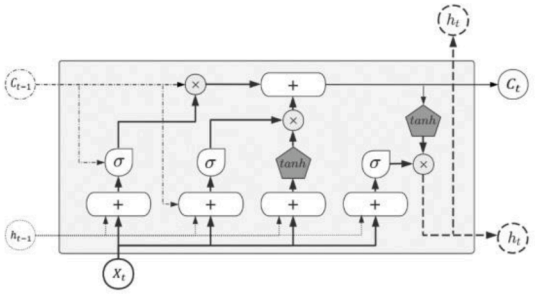

Fig. 1. One basic unit of LSTM.

where ρ_j represents the parameter for controlling the white noise energy in each iteration, $E_n(\cdot)$ denotes the *IMFs* component generated by the EMD decomposition, $\langle \cdot \rangle$ represents the operation of averaging, and $C(\cdot)$ is the operator that produces the local mean of the original series.

2.3. LSTM model

The settings of the cell state and gate mechanism in the LSTM model make it easier for it to reset, update, and remember long-term information.

In Fig. 1, C_t refers to the cell state, h_t refers to the hidden state, and X_t is the input. The process of obtaining an LSTM unit is as follows:

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

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right). \tag{9}$$

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

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{10}$$

$$\widetilde{c_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c). \tag{11}$$

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:

$$c_t = f_t^* c_{t-1} + i_t^* \widetilde{c_t}. \tag{12}$$

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:

$$\begin{cases} o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t \cdot \tanh(c_t) \end{cases}. \tag{13}$$

2.4. VMD-ICEEMDAN-LSTM model

Based on the above method, we propose a secondary decomposition technique using the residual term from the VMD decomposition. Specifically, ICEEMDAN is used to further decompose the residual term after VMD decomposition. Fig. 2 shows the detailed flow chart of the prediction process.

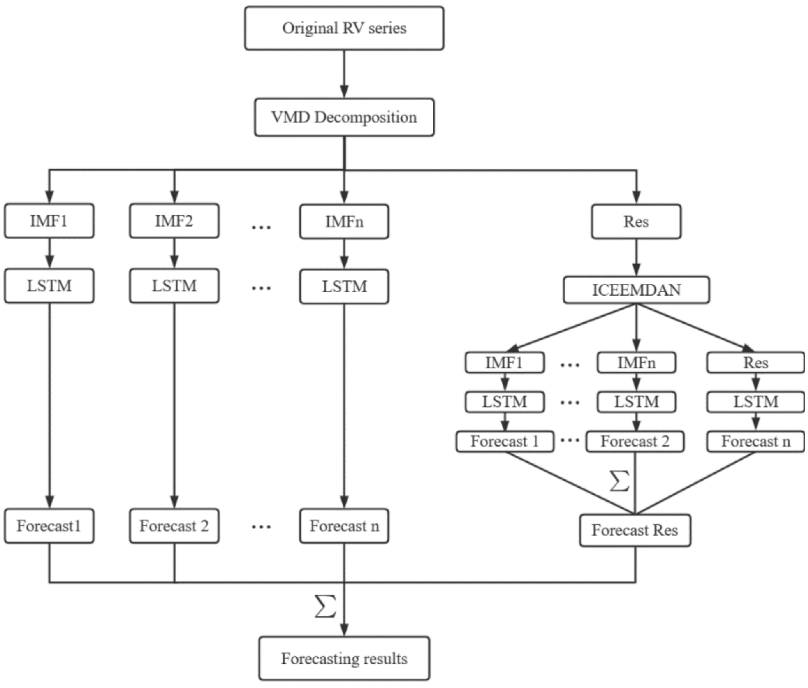

Fig. 2. Flow chart of the VMD-ICEEMDAN-LSTM model.

2.5. Evaluation criteria

This study selects four common loss functions:

$$\begin{cases} MAE = H^{-1}\sum_{t=1}^{H}|y-\hat{y}| \\ MSE = H^{-1}\sum_{t=1}^{H}(y-\hat{y})^2 \\ HMAE = H^{-1}\sum_{t=1}^{H}\left|1-\frac{\hat{y}}{y}\right| \\ HMSE = H^{-1}\sum_{t=1}^{H}\left(1-\frac{\hat{y}}{y}\right)^2 \end{cases} \tag{14}$$

where H refers to the total length of the predicted sample, y represents the ground truth, and $\hat{y}$ represents the predicted value.

To obtain more robust test results, we use the MCS test (Hansen et al., 2011) as a method to test the out-of-sample predictive ability of each model. In this study, we choose range statistic T_R and semi-quadratic statistic T_{SQ} as the test criteria:

$$\begin{cases} T_R = \max_{i,j\in M}\frac{|\overline{d_{ij}}|}{\sqrt{var(\overline{d_{ij}})}} \\ T_{SQ} = \max_{i,j\in M}\frac{(\overline{d_{ij}})^2}{\sqrt{var(\overline{d_{ij}})}} \end{cases} \tag{15}$$

where $\overline{d_{ij}} = H^{-1}\sum_{t=1}^{H} d_{ij,t}$ represents the average value of the relative loss function values of the RV predictions of models i and j.

3. Empirical results

3.1. Data analysis

Andersen & Bollerslev (1998) pointed out that five-minute high-frequency data can strike a balance between meeting the high frequency requirements of the sample and reducing micro noise. This study selects the five-minute high-frequency data of Chinese crude oil futures from March 26, 2018 to January 31, 2023 to calculate RV based on Eq. (1). The original data are sourced from the Tushare database.

Table 1 exhibits the descriptive statistics of the RV of Chinese crude oil futures. The data show that the RV of Chinese crude oil futures has significant features such as being leptokurtic and skewed, while Jarque–Bera statistic indicates that it has a significant nonnormal distribution and the Ljung–Box statistic shows the existence of a long-memory feature. Fig. 3 depicts the RV time series of Chinese crude oil futures.

3.2. Data decomposition

The original RV is first decomposed into several components and a residual term using the VMD, and then the residual term is secondary decomposed by ICEEMDAN. Before performing VMD decomposition, it is necessary to determine the number K of modal decomposition. This paper chooses the central frequency method to determine K.

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.

Fig. 4 shows the decomposed subsequences, where each *IMF* component in the figure is sorted from high to low frequency, Res represents the residual term, and all subsequences fluctuate around zero, which makes them easier to predict. The complexity of the residual term obtained by VMD decomposition is still relatively high, thus making it difficult to ensure prediction accuracy by directly predicting it. Therefore, there is a need for secondary decomposition. Additionally, the result of VMD decomposition indicates that the

Table 1
Descriptive statistics of the RV of Chinese crude oil futures.

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. *, **, and *** denote rejection of the null hypothesis at the 10%, 5%, and 1% significance levels, respectively.

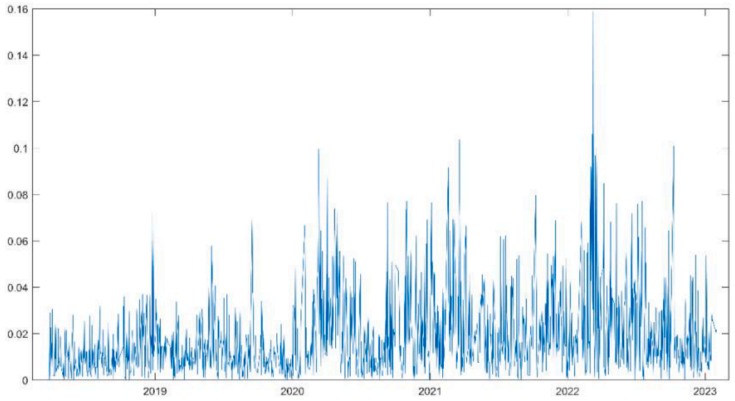

Fig. 3. Time series of RV.

Table 2
Central frequencies of *IMFs* under different *K* values.

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

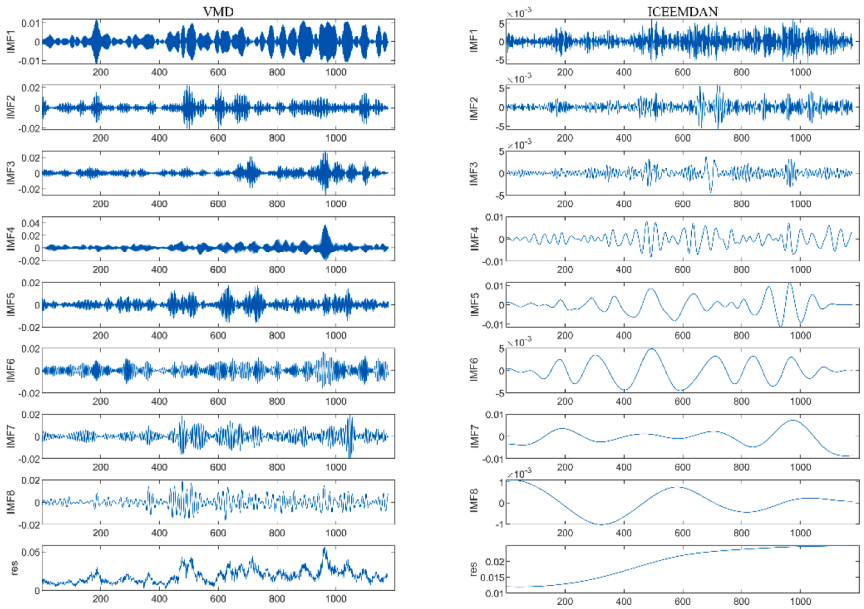

Fig. 4. Decomposition results of RV.

RV of Chinese crude oil futures contains a large amount of noise.

Table 3 shows that the subsequences obtained through VMD decomposition and the residual term after ICEEMDAN secondary decomposition both have relatively small standard deviations.

Table 3
Descriptive statistics of the decomposition results of RV.

	Mean	Std. Dev	Skewness	Kurtosis	Q(5)
Panel A: VMD					
IMF1	1.72×10^{-8}	4.80×10^{-3}	0.002	−0.334***	5353.113***
IMF2	1.99×10^{-8}	5.41×10^{-3}	0.002	1.814***	1953.451***
IMF3	2.64×10^{-8}	5.37×10^{-3}	0.028	4.585***	2869.072***
IMF4	3.55×10^{-8}	5.43×10^{-3}	1.198***	7.794***	2246.508***
IMF5	5.31×10^{-8}	4.55×10^{-3}	0.008	1.433***	2379.843***
IMF6	1.11×10^{-7}	4.55×10^{-3}	0.020	0.206***	2493.936***
IMF7	2.27×10^{-7}	4.68×10^{-3}	0.043	1.989***	2558.132***
IMF8	9.16×10^{-7}	4.99×10^{-3}	0.213***	0.996***	1847.668***
Res	1.86×10^{-2}	8.79×10^{-3}	1.004***	1.210***	5709.886***
Panel B: ICEEMDAN					
IMF1	9.00×10^{-6}	2.00×10^{-3}	−0.006	0.029	943.156***
IMF2	2.50×10^{-5}	1.28×10^{-3}	0.032	2.757***	1002.750***
IMF3	9.00×10^{-6}	9.34×10^{-4}	−0.146***	2.772***	1583.658***
IMF4	-7.10×10^{-5}	2.77×10^{-3}	0.005	0.270***	4171.207***
IMF5	-3.24×10^{-4}	3.93×10^{-3}	0.108***	0.615***	5586.800***
IMF6	-1.51×10^{-4}	2.28×10^{-3}	0.032	−0.745***	5808.612***
IMF7	-5.99×10^{-4}	3.47×10^{-3}	−0.211***	0.224***	5801.039***
IMF8	3.40×10^{-5}	5.53×10^{-4}	−0.051	−0.626***	5846.810***
Res	1.97×10^{-2}	4.87×10^{-3}	−0.443***	−1.454***	5855.489***

Note: Q(5) is the Ljung–Box statistic for up to the 5th-order serial correlation. *, **, and *** denote rejection of the null hypothesis at the 10%, 5%, and 1% significance levels, respectively.

3.3. Hyperparameter selection

To balance computational efficiency and predictive accuracy, this article sets the LSTM model's learning rate to 0.001 and batch size (mini-batch) to 32 based on the research findings of Keskar et al. (2016). In the empirical section, we employ the rolling time prediction method to test the predictive performance of the above model and divide each subsequence into a training set, test set, and prediction set following a 7:2:1 split. Then, the training set data are fed into the corresponding LSTM prediction model as input for training, and finally the prediction results are obtained using the testing set data.

Different hyperparameters will affect the accuracy of the hybrid prediction model. To obtain the best prediction results for different component sequences, optimal hyperparameters were selected through experimentation. The optimal hyperparameter set for the proposed prediction model is shown in Table 4.

3.4. Forecasting performance

Fig. 5 shows the predicted results of the RV decomposition sequence of Chinese crude oil futures in the LSTM model. The blue dashed line represents the actual value, and the red solid line represents the predicted value.

To visually compare the predictive performance of the combined model, we plot them together in Fig. 6, which shows that the decomposed prediction results improve the lagging problem of single LSTM prediction. Moreover, the VMD-ICEEMDAN-LSTM model proposed in this paper has better predictive performance for the RV of Chinese crude oil futures.

Table 5 shows that under three evaluation indicators and different lead times, the VMD-ICEEMDAN-LSTM model has the best predictive performance. In the single prediction models, LSTM has better predictive performance in the short, medium and long terms than the HAR and ARIMA models, thus indicating that LSTM has better time series predictive capability in deep learning algorithms than traditional econometric models. In addition, the predictive performance of the three combined prediction models is significantly better than that of the single models.

Table 4
Hyperparameters of VMD-ICEEMDAN-LSTM.

	VMD		ICEEMDAN	
	Window	Epoch	Window	Epoch
IMF1	3	420	3	460
IMF2	5	420	3	460
IMF3	5	360	3	460
IMF4	4	300	3	440
IMF5	4	300	4	360
IMF6	4	240	5	120
IMF7	5	240	5	90
IMF8	5	180	2	90
Res	–	–	2	80

Note: "Window" is the length of each input. "Epoch" is the number of trainings.

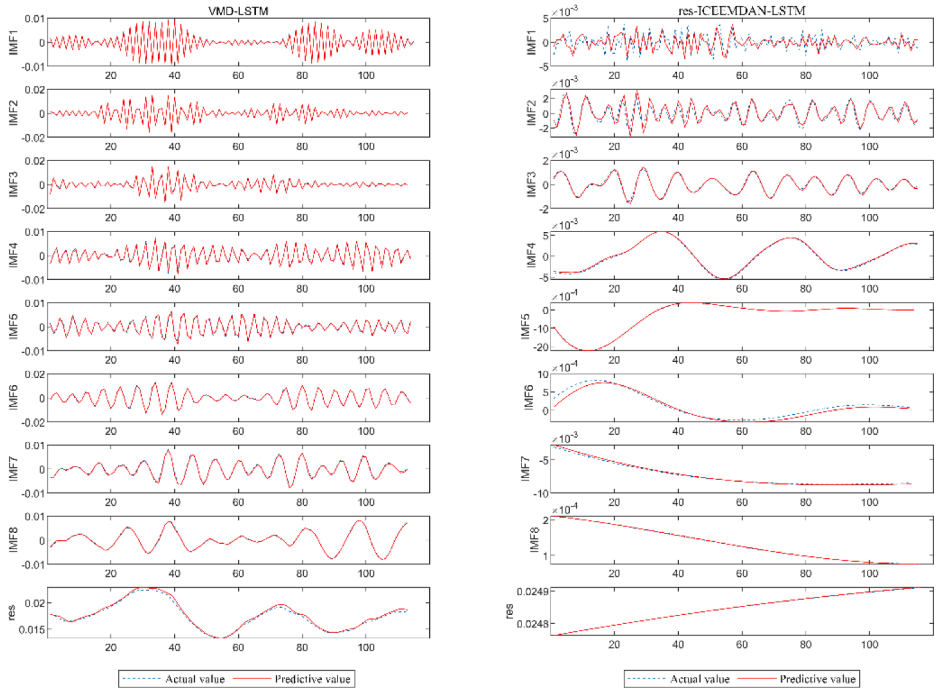

Fig. 5. Forecasting results of subseries of RV.

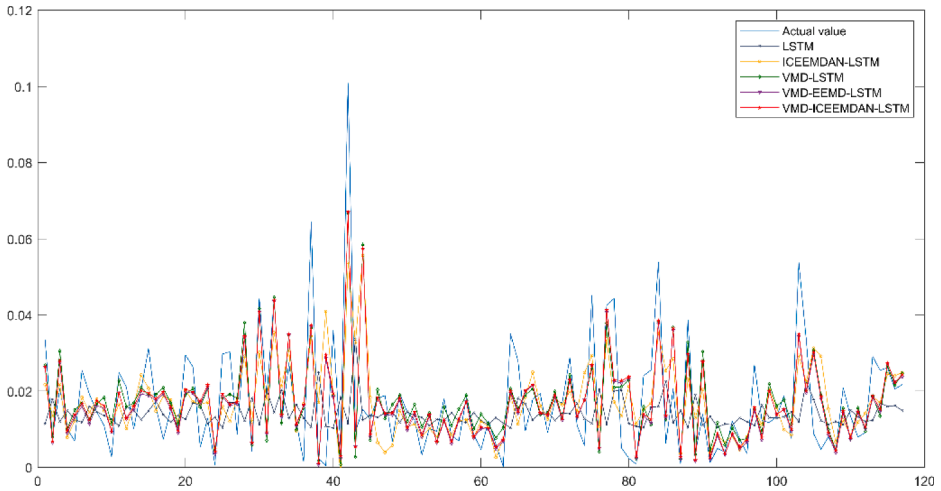

Fig. 6. Forecasting results of RV.

The numerical values of all results show that the VMD-ICEEMDAN-LSTM model has better performance than the VMD-LSTM model. This suggests that performing a secondary decomposition of the residual components obtained from VMD can effectively improve the predictive accuracy of the model. The reason for this is that performing a secondary decomposition of the residual components before prediction can greatly reduce the complexity of the original data and thus improve the overall predictive accuracy of the model.

All results confirm the excellence of the VMD-ICEEMDAN-LSTM model in predicting the realized volatility of Chinese crude oil futures as well as its prediction stability and reliability in different time periods.

We conducted the MCS test to verify the superiority of the VMD-ICEEMDAN-LSTM prediction model. The results of the MCS test are shown in Table 6. It is worth noting that the *p*-values of the VMD-ICEEMDAN-LSTM model under the three loss indicators are all the maximum value of 1 in the long, medium, and short terms. This again confirms the robustness of the proposed secondary decomposition prediction model in improving the predictive accuracy of the realized volatility of Chinese crude oil futures.

Table 5
Forecasting results under four loss functions.

Models	MAE	MSE	HMAE	HMSE	MAE	MSE	HMAE	HMSE	MAE	MSE	HMAE	HMSE
	Panel A: $H = 1$				Panel B: $H = 5$				Panel C: $H = 20$			
LSTM	1.12×10^{-2}	2.64×10^{-4}	0.838	1.705	1.15×10^{-2}	2.49×10^{-4}	0.696	0.914	1.22×10^{-2}	2.76×10^{-4}	0.744	1.043
ICE-LSTM	8.90×10^{-3}	1.49×10^{-4}	0.571	0.648	6.54×10^{-3}	7.95×10^{-5}	1.849	1.189	6.89×10^{-3}	8.18×10^{-5}	0.617	2.338
VMD-LSTM	7.48×10^{-3}	1.10×10^{-4}	0.494	0.672	2.65×10^{-3}	1.07×10^{-5}	0.231	0.108	2.30×10^{-3}	8.43×10^{-6}	0.429	1.808
VMD-EEMD-LSTM	7.27×10^{-3}	1.06×10^{-4}	0.473	0.347	2.28×10^{-3}	7.68×10^{-5}	0.191	0.080	1.72×10^{-3}	4.29×10^{-6}	0.685	2.191
VMD-ICE-LSTM	**7.23×10^{-3}**	**1.05×10^{-4}**	**0.438**	**0.272**	**2.25×10^{-3}**	**7.37×10^{-6}**	**0.191**	**0.079**	**1.70×10^{-3}**	**4.11×10^{-6}**	**0.252**	**0.312**
HAR	1.34×10^{-2}	4.00×10^{-4}	7.314	6.902	1.34×10^{-2}	4.02×10^{-4}	7.451	7.138	1.42×10^{-3}	4.46×10^{-4}	8.304	8.205
ARIMA	1.16×10^{-2}	2.38×10^{-4}	2.619	5.588	1.16×10^{-2}	2.43×10^{-4}	2.679	5.762	1.23×10^{-2}	2.69×10^{-4}	2.945	6.582

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

Table 6
Forecasting results using the MCS test.

Forecasting model	MAE		MSE		HMAE		HMSE	
	Range	SeimQ	Range	SeimQ	Range	SeimQ	Range	SeimQ
Panel A: $H = 1$								
LSTM	0.000	0.000	0.007	0.008	0.000	0.000	0.017	0.030
ICE-LSTM	0.004	0.001	0.007	0.008	0.020	0.018	0.141	0.085
VMD-LSTM	0.039	0.015	0.007	0.008	0.020	0.054	0.141	0.085
VMD-EEMD-LSTM	0.097	0.097	0.072	0.072	0.020	0.054	0.141	0.085
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.007	0.007	0.000	0.000	0.017	0.003
ARIMA	0.000	0.000	0.007	0.007	0.000	0.000	0.017	0.006
Panel A: $H = 5$								
LSTM	0.000	0.000	0.005	0.004	0.000	0.000	0.002	0.003
ICE-LSTM	0.001	0.000	0.005	0.004	0.000	0.000	0.002	0.003
VMD-LSTM	0.006	0.005	0.005	0.004	0.002	0.003	0.002	0.003
VMD-EEMD-LSTM	0.684	0.684	0.231	0.231	0.930	0.930	0.844	0.844
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.005	0.004	0.000	0.000	0.002	0.002
ARIMA	0.000	0.000	0.005	0.004	0.000	0.000	0.002	0.002
Panel A: $H = 22$								
LSTM	0.000	0.000	0.000	0.001	0.000	0.006	0.027	0.123
ICE-LSTM	0.000	0.000	0.000	0.001	0.039	0.084	0.027	0.123
VMD-LSTM	0.000	0.000	0.000	0.001	0.039	0.084	0.027	0.123
VMD-EEMD-LSTM	0.540	0.540	0.101	0.101	0.039	0.084	0.027	0.119
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.001	0.000	0.000	0.023	0.018
ARIMA	0.000	0.000	0.000	0.001	0.000	0.000	0.023	0.018

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 underlined. $H = 1$, 5, 20 respectively represent 1 step ahead, 5 steps ahead, and 20 steps ahead. Range and SeimQ represents range statistic, and semi-quadratic statistic, respectively (Hansen et al., 2011).

3.5. Robustness test

To ensure that robustness of the results, we perform several robustness checks in this section, including changing the sample frequency, changing the ratio of the training set, test set, and prediction set, and eliminating the data during the COVID-19 outbreak in China.

3.5.1. Changing the sample frequency

In this section, a replication of the research was conducted by utilizing 1-minute data intervals. The resulting findings are presented in Table 7. Notably, despite fluctuations in the frequency of high-frequency data, the model proposed in this paper still demonstrates the best performance.

3.5.2. Changing the sample ratios

Considering that the data proportion setting in the LSTM model may influence the prediction results, we have adjusted the proportions of our training sets, testing sets, and prediction sets. The new distribution ratio is set to 6:2:2, compared to the previous distribution of 7:2:1. The MCS test results of the robustness check are shown in Table 8, where it can be seen that the VMD-ICEEMDAN-LSTM model still exhibits the best performance.

3.5.3. Excluding the data during the COVID-19 outbreak in China

Additionally, to enhance the reliability of the empirical research findings, we excluded data from the period during the COVID-19 outbreak in China. The results in Table 9 demonstrate that the proposed model in this paper still outperforms all other models.

All MCS test results thus confirm the robustness of the VMD-ICEEMDAN-LSTM model on different data sets.

4. Conclusions

Accurate prediction of the realized volatility of Chinese crude oil futures is of great significance for all parties involved in international energy transactions as well as the relevant policy makers. To overcome the deficiencies of the existing research on crude oil futures price fluctuations prediction using decomposition technology, this paper proposes for the first time a new hybrid model: VMD-ICEEMDAN-LSTM.

In the existing research, although VMD performs well in enhancing the signal-to-noise ratio, some studies have overlooked the fact that the residual components after VMD decomposition still exhibit high complexity. As a result, directly predicting the residual components may affect the prediction accuracy of the model. Therefore, it is necessary to perform secondary decomposition on the residual components. Currently, models used for secondary decomposition such as EEMD suffer from defects that include residual

Table 7
Forecasting results under 1-minute frequencies with the MCS test.

Forecasting models	MAE		MSE		HMAE		HMSE	
	Range	SeimQ	Range	SeimQ	Range	SeimQ	Range	SeimQ
Panel A: $H = 1$								
LSTM	0.000	0.000	0.005	0.003	0.000	0.001	0.012	0.020
ICE-LSTM	0.009	0.005	0.005	0.003	0.023	0.018	0.074	0.056
VMD-LSTM	0.040	0.022	0.005	0.003	0.023	0.050	0.074	0.057
VMD-EEMD-LSTM	0.088	0.088	0.121	0.121	0.023	0.050	0.074	0.057
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.005	0.003	0.000	0.000	0.012	0.006
ARIMA	0.000	0.000	0.005	0.003	0.000	0.000	0.012	0.006
Panel A: $H = 5$								
LSTM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ICE-LSTM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VMD-LSTM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VMD-EEMD-LSTM	0.221	0.221	0.012	0.012	0.122	0.122	0.045	0.045
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ARIMA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel A: $H = 22$								
LSTM	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.023
ICE-LSTM	0.000	0.000	0.000	0.001	0.065	0.071	0.001	0.023
VMD-LSTM	0.000	0.000	0.000	0.001	0.103	0.103	0.001	0.023
VMD-EEMD-LSTM	0.029	0.029	0.067	0.067	0.033	0.071	0.274	0.274
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.007
ARIMA	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.007

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 underlined. $H = 1$, 5, 20 respectively represent 1 step ahead, 5 steps ahead, and 20 steps ahead. Range and SeimQ represents range statistic, and semi-quadratic statistic, respectively (Hansen et al., 2011).

Table 8
Forecasting results under the ratio of 6:2:2 with the MCS test.

Forecasting models	MAE		MSE		HMAE		HMSE	
	Range	SeimQ	Range	SeimQ	Range	SeimQ	Range	SeimQ
Panel A: $H = 1$								
LSTM	0.000	0.000	0.008	0.002	0.000	0.000	0.001	0.003
ICE-LSTM	0.000	0.000	0.037	0.010	0.026	0.032	0.117	0.070
VMD-LSTM	0.193	0.119	0.037	0.020	0.070	0.066	0.001	0.008
VMD-EEMD-LSTM	0.193	0.119	0.285	0.285	0.604	0.604	0.315	0.315
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.004	0.001	0.000	0.000	0.001	0.000
ARIMA	0.000	0.000	0.007	0.001	0.000	0.001	0.001	0.008
Panel A: $H = 5$								
LSTM	0.000	0.000	0.001	0.003	0.000	0.000	0.000	0.004
ICE-LSTM	0.000	0.000	0.001	0.009	0.001	0.002	0.000	0.004
VMD-LSTM	0.000	0.000	0.001	0.009	0.000	0.000	0.000	0.004
VMD-EEMD-LSTM	0.003	0.003	0.066	0.066	0.047	0.047	0.000	0.004
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
ARIMA	0.000	0.000	0.000	0.001	0.000	0.000	0.055	0.055
Panel A: $H = 22$								
LSTM	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
ICE-LSTM	0.000	0.000	0.000	0.000	0.004	0.001	0.001	0.001
VMD-LSTM	0.000	0.000	0.000	0.000	0.360	0.309	0.358	0.354
VMD-EEMD-LSTM	0.000	0.000	0.000	0.000	0.360	0.336	0.737	0.737
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000
ARIMA	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.003

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 underlined. $H = 1$, 5, 20 respectively represent 1 step ahead, 5 steps ahead, and 20 steps ahead. Range represents range statistic, and SeimQ represents semi-quadratic statistic.

noise and pseudo modes, while ICEEMDAN has successfully overcome these issues and is thus a better choice.

The empirical results first indicate that the LSTM model performs better than several traditional econometric models. Moreover, after the original sequence is decomposed by VMD, the fitting effect of the LSTM model for the predicted values is significantly

Table 9
Forecasting results under excluding the data during the COVID-19 outbreak with the MCS test.

Forecasting models	MAE		MSE		HMAE		HMSE	
	Range	SeimQ	Range	SeimQ	Range	SeimQ	Range	SeimQ
Panel A: $H = 1$								
LSTM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ICE-LSTM	0.005	0.005	0.023	0.023	0.001	0.000	0.000	0.000
VMD-LSTM	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
VMD-EEMD-LSTM	0.000	0.000	0.000	0.000	0.119	0.119	0.275	0.275
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ARIMA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel A: $H = 5$								
LSTM	0.000	0.000	0.000	0.000	0.711	0.431	0.448	0.454
ICE-LSTM	0.000	0.000	0.000	0.000	0.403	0.745	0.448	0.454
VMD-LSTM	0.454	0.454	0.255	0.255	0.720	0.150	0.448	0.454
VMD-EEMD-LSTM	0.016	0.076	0.001	0.005	0.054	0.745	0.094	0.142
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.000	0.010	0.009	0.094	0.049
ARIMA	0.000	0.000	0.000	0.000	0.720	0.621	0.094	0.142
Panel A: $H = 22$								
LSTM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ICE-LSTM	0.000	0.000	0.000	0.000	0.005	0.022	0.018	0.030
VMD-LSTM	0.000	0.000	0.005	0.005	0.547	0.512	0.610	0.610
VMD-EEMD-LSTM	0.000	0.000	0.000	0.000	0.547	0.512	0.513	0.499
VMD-ICE-LSTM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ARIMA	0.000	0.000	0.000	0.000	0.001	0.000	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 underlined. $H = 1$, 5, 20 respectively represent 1 step ahead, 5 steps ahead, and 20 steps ahead. Range and SeimQ represents range statistic, and semi-quadratic statistic, respectively (Hansen et al., 2011).

improved. Then, the secondary decomposition models display better predictive ability than the VMD-LSTM model. This implies that secondary decomposition of the residual values of VMD can effectively improve the model's overall predictive ability. The VMD-ICEEMDAN-LSTM model proposed in this paper outperforms the other models in terms of the evaluation indicators and MCS tests, thus demonstrating its superiority in RV prediction.

Moreover, the test results when data from the COVID-19 period are included also confirm that the model proposed in this paper still has excellent predictive performance in the face of extreme conditions.

CRediT authorship contribution statement

Wei Jiang: Conceptualization, Methodology, Software, Writing – original draft. **Wanqing Tang:** Formal analysis, Investigation, Writing – review & editing, Visualization. **Xiao Liu:** Data curation, Validation, Resources, Supervision.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

Acknowledgments

The authors wish to express their thanks to the two anonymous reviewers for their valuable comments in the earlier version of this paper. This work was supported by the National Social Science Foundation of China [22BJY256].

References

Abdoos, A.A., 2016. A new intelligent method based on combination of VMD and ELM for short term wind power forecasting. Neurocomputing 203, 111–120.
Andersen, T.G., Bollerslev, T., 1998. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. Int. Econ. Rev. (Philadelphia) 885–905.
Christensen, K., Siggaard, M., Veliyev, B., 2022. A machine learning approach to volatility forecasting. J. Financ. Econom.

Colominas, M.A., Schlotthauer, G., Torres, M.E., 2014. Improved complete ensemble EMD: a suitable tool for biomedical signal processing. Biomed. Signal Process. Control 14, 19–29.
Dragomiretskiy, K., Zosso, D., 2014. Variational mode decomposition. IEEE Trans. Signal Process. 62, 531–544.
Guo, W., Liu, Q., Luo, Z., Tse, Y., 2022. Forecasts for international financial series with VMD algorithms. J. Asian Econ. 80, 101458.
Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.
He, F., Zhou, J., Feng, Z., Liu, G., Yang, Y., 2019. A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with Bayesian optimization algorithm. Appl. Energy.
Huang, Y., Deng, Y., 2021. A new crude oil price forecasting model based on variational mode decomposition. Knowl. Based Syst. 213, 106669.
Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T., 2016. On large-batch training for deep learning: generalization gap and sharp minima. ArXiv abs/1609.04836.
Kim, H.Y., Won, C.H., 2018. Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models. Expert Syst. Appl. 103, 25–37.
Lahmiri, S., 2017. Comparing variational and empirical mode decomposition in forecasting day-ahead energy prices. IEEE Syst. J. 11, 1907–1910.
Li, Z., Tam, V.W., 2017. Combining the real-time wavelet denoising and long-short-term-memory neural network for predicting stock indexes. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8.
Liang, Y., Lin, Y., Lu, Q., 2022. Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM. Expert Syst. Appl. 206, 117847.
Liu, Y., 2019. Novel volatility forecasting using deep learning-long short term memory recurrent neural networks. Expert Syst. Appl. 132, 99–109.
Liu, Y., Yang, C., Huang, K., Gui, W., 2020. Non-ferrous metals price forecasting based on variational mode decomposition and LSTM network. Knowl. Based Syst. 188.
Niu, H., Xu, K., Wang, W., 2020. A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network. Appl. Intellig. 50, 4296–4309.
Tang, H., Bhatti, U.A., Li, J., Marjan, S., Baryalai, M., Assam, M., Mohamed, H.G., 2023. A new hybrid forecasting model based on dual series decomposition with long-term short-term memory. Int. J. Intellig. Syst. 2023.
Taylan, O., 2017. Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. Atmos. Environ. 150, 356–365.
Wang, B., Wang, J., 2020. Deep multi-hybrid forecasting system with random EWT extraction and variational learning rate algorithm for crude oil futures. Expert Syst. Appl. 161, 113686.
Yang, Y., Yang, Y., Zhou, W., 2021. Research on a hybrid prediction model for stock price based on long short-term memory and variational mode decomposition. Soft Comput. 25, 13513–13531.
Zhan, L., Tang, Z., 2022. Natural gas price forecasting by a new hybrid model combining quadratic decomposition technology and LSTM model. Math. Probl. Eng. 2022.
Zhang, T., Tang, Z., Wu, J., Du, X., Chen, K., 2022. Short term electricity price forecasting using a new hybrid model based on two-layer decomposition technique and ensemble learning. Electric Power Syst. Res. 205, 107762.