Hindawi
Journal of Mathematics
Volume 2024, Article ID 6652218, 20 pages
https://doi.org/10.1155/2024/6652218

Research Article

A Crude Oil Spot Price Forecasting Method Incorporating Quadratic Decomposition and Residual Forecasting

Yonghui Duan,[1] Ziru Ming ,[1] and Xiang Wang [2]

[1]*Department of Civil Engineering, Henan University of Technology, No. 100, Lianhua Street, Gaoxin District, Zhengzhou 450001, China*
[2]*Department of Civil Engineering, Zhengzhou University of Aeronautics, No. 15, Wenyuan West Road, Zhengdong New District, Zhengzhou 450015, China*

Correspondence should be addressed to Ziru Ming; myq_20220926@qq.com

Received 2 December 2023; Revised 21 March 2024; Accepted 26 March 2024; Published 15 April 2024

Academic Editor: Ding-Xuan Zhou

The world economy is affected by fluctuations in the price of crude oil, making precise and effective forecasting of crude oil prices essential. In this study, we propose a combined forecasting scheme, which combines a quadratic decomposition and optimized support vector regression (SVR). In the decomposition part, the original crude oil price series are first decomposed using empirical modal decomposition (CEEMDAN), and then the residuals of the first decomposition (RES) are decomposed using variational modal decomposition (VMD). Additionally, this work proposes to optimize the support vector regression model (SVR) by the seagull optimization algorithm (SOA). Ultimately, the empirical investigation created the feature-variable system and predicted the filtered features. By computing evaluation indices like MAE, MSE, R^2, and MAPE and validating using Brent and WTI crude oil spot, the prediction errors of the CEEMDAN -RES.-VMD -SOA-SVR combination prediction model presented in this paper are assessed and compared with those of the other twelve comparative models. The empirical evidence shows that the combination model being proposed in this paper outperforms the other related comparative models and improves the accuracy of the crude oil price forecasting model.

1. Introduction

As one of the world's most important energy sources and commodities, the price trend of crude oil has a significant impact on the international political situation and the world economy. In terms of the commodity attributes of crude oil, the crude oil market is at the core of the commodity market, and the fluctuations in the price of crude oil have profoundly affected price volatility in commodity markets. For example, prices and resulting volatility spillovers between commodity markets were investigated by Ji and Fan [1] who found that the crude oil market has a significant volatility spillover effect on non-energy commodity markets. Chen et al. [2] discovered that asymmetric oil price shocks can affect the nonferrous metals market. This is due to the surge in transportation costs of raw materials and production costs caused by higher crude oil prices, which ultimately manifests itself in higher prices for consumer goods. In terms of the financial attributes of crude oil, the global macroeconomy is also increasingly affected by movements in crude oil prices. Liu et al. [3] elaborated on the factors affecting the dynamics of crude oil prices in terms of financial attributes and concluded that financial factors, speculative behavior, and other factors have a major part in oil price changes. In addition to this, crude oil prices significantly affect the exchange rate of importing countries [4]. For one thing, an increase in the trade deficit of importing countries can be caused by a spike in the price of crude oil, which in turn leads to a depreciation of the importing country's currency. Second, the price of crude oil can also hurt the stock market as well as inflation in importing countries, which in turn affects the value of currencies and changes in exchange rates. Analyzed from the energy point of view, due to the scarcity of crude oil and as the most important energy resource for

industry, crude oil has not only profoundly affected the world's economy and trade, but has also penetrated national strategies and political struggles, becoming one of the most important strategic resources in the world. In addition, many other factors have an equal impact on how volatile crude oil prices are. For example, The impact of supply and demand factors on the drop in oil prices was researched by Kim [5]. Zhou et al. [6] analyze the implication of the high-frequency linkage characteristics of the US dollar index and crude oil inventories on the volatility of crude oil prices. Bildirici et al. [7] predicted global crude oil prices under the impact of the COVID-19 anomaly that broke out. Khan et al. [8] investigate how geopolitical risks affect oil prices and freight rates. In conclusion, crude oil prices are subject to a wide range of other factors, such as supply and demand, geopolitical risks, natural disasters, and other factors, and are characterized by a high degree of uncertainty. As a result, the crude oil price series is characterized by nonlinearity and nonstationarity, making it difficult to predict future prices.

In most of the early studies scholars used traditional econometric models. The vector autoregressive (VAR) model, the generalized autoregressive conditional heteroskedasticity (GARCH) model, the autoregressive moving average (ARMA) model, the autoregressive integrated moving average (ARIMA) model and others are examples of traditional econometric models [9–12]. Due to the limitations of the traditional econometric model itself and the nonlinearity and complexity of the crude oil price time series, the forecasting results are subject to overfitting problems and poor forecasting accuracy. Artificial intelligence prediction techniques have arisen with the progress of computer technology. Compared to conventional econometric models, artificial intelligence techniques are better able to handle time series for crude oil prices. Artificial neural network model (ANN) [13], convolutional neural networks (CNN) [14], random forest model (RF) [15], support vector machines (SVMs) [16], and others are examples of artificial intelligence prediction models currently in use. Neural network models can be effectively fitted to nonlinear problems with complex relationships, but they are usually suitable for large amounts of training data and may not be able to take advantage of neural networks for problems with smaller amounts of data. On the contrary, support vector machines (SVMs) are better at dealing with small samples, nonlinearities, and sample imbalance problems. But feature selection is limited, and human parameter tuning is necessary for typical SVM models. Therefore, resolving these issues can significantly aid in practical problem resolution. For example, by introducing nonlinear kernel functions and enhancing feature selection and parameter selection strategies, Xia et al. [17] analyzed a new high-dimensional partially linear support vector machine regularized learning scheme to make up for some of the traditional SVM model's shortcomings in terms of flexibility, feature selection, and parameter selection. Second, previous studies have mostly been conducted using univariate predictive models. However, the univariate forecasting model uses only the internal characteristics of the crude oil price time series as model inputs, ignoring other external characteristics that may affect the change in crude oil prices. Multivariate predictive models are unlike univariate predictive models. Multivariate predictive models can provide more information to help models better understand and capture patterns and trends in time series data. For example, a generalized DCNNs structure was introduced by Mao et al. [18] in order to increase prediction accuracy by better capturing these complex features in time series data. In this research paper, to predict future crude oil prices more accurately, we will use a multivariate forecasting model. And this approach has been validated by previous scholars, for example, Liu and Huang [19]. Future crude oil prices were predicted by building a deep neural network that included sentiment, news events, and historical price data after a text sentiment analysis algorithm was used to extract sentiment from a vast amount of news.

A hybrid decomposition-prediction-integration prediction model is currently being developed by researchers. The main idea behind decomposition-prediction-integration is to break down the original time series signals using decomposition techniques into a series of subsequences, use each subsequence as an input to a prediction model, and then add up the prediction results to get the model's overall prediction results. Signal decomposition can be the process of breaking down complex signals into simpler predictions that can be used for machine learning, reducing noise and irrelevant information in the data, improving the computational efficiency of machine learning algorithms, and effectively preventing predictive overfitting. The signal decomposition models include wavelet transform (WT) [20], empirical modal decomposition (EMD) [21], ensemble empirical modal decomposition (EEMD) [22], variational modal decomposition (VMD) [23], etc. Wavelet transform (WT) methods are limited by their nonadaptive nature as they require artificial selection of basic functions. In contrast, the EMD decomposition technique and EEMD decomposition technique can decompose non-smooth and nonlinear signals into smooth signals with different time scales and without the need for basis function selection. However, the CEEMDAN decomposition, ICEEMDAN decomposition, and other techniques are proposed under the ongoing improvement of the latter because the EMD decomposition and EEMD decomposition techniques suffer from the phenomenon of mode aliasing. In contrast to other modal decomposition approaches like EMD, EEMD, CEEMD, and others, VMD decomposition is accomplished by generating variational issues. The usage of VMD decomposition techniques for crude oil and other domain problems has become common in recent years. Li et al. [24], for instance, suggested a novel technique for predicting the price of crude oil that combines variational mode decomposition and random sparse Bayesian learning. Zhao et al. [25] improved the variational mode decomposition to forecast crude oil prices online and in real time. To forecast monthly natural gas prices, Li et al. [26] integrated a quadratic decomposition method, merging the VMD decomposition technique with an improved backpropagation (BP) neural network.

Existing studies typically use a once-off decomposition technique to forecast crude oil prices, and there may be some that are not correctly decomposed, resulting in an incomplete signal decomposition. However, the accuracy of the overall forecast is improved by the secondary decomposition, which increases the accuracy and completeness of the primary decomposition. A lot of scholars have applied the secondary decomposition technique to carbon price prediction [27], wind speed prediction [28], and other directions. In crude oil price prediction, Zhang et al. [29] verified that quadratic decomposition can improve the prediction performance by quadratically decomposing the crude oil residual term. Li et al. [30] concluded that the quadratic decomposition is superior to the one-time decomposition model by quadratically decomposing the components with higher complexity and reconstructing the prediction results. Table 1 presents a categorization and summary of the literature on crude oil price predictions based on the aforementioned literature review.

The two-layer decomposition technique of CEEMDAN and VMD is proposed in this study, specifically, the residual term after CEEMDAN decomposition is decomposed twice by the VMD decomposition algorithm. Secondly, because there were small samples in this study and the SVR model can automatically select the features that have the greatest impact on the prediction results as well as have good generalization ability. Therefore, the SVR algorithm was adopted as the basis for the prediction model for this study. Finally, the two-layer decomposition technique is combined with the SOA-SVR prediction algorithm. To be more precise, the ultimate prediction outcomes were acquired by summing together in a linear fashion all of the simple component predictions from the original CEEMDAN decomposition and all of the simple component predictions from the VMD quadratic decomposition of the CEEMDAN residual term. To verify the prediction performance of the proposed CEEMDAN- RES.-VMD-SOA-SVR model, six hybrid models, SVR, SOA-SVR, VMD-SOA-SVR, VMD-CEEMDAN-SOA-SVR, CEEMDAN-SOA-SVR, and CEEMDAN-RES.-SOA-SVR, were used as comparison experiment. The data is selected from the monthly data of the Brent crude oil price time series and the WTI crude oil price time series to do one-step, two-step, and three-step tests of all the models to forecast the future quarterly trend of crude oil prices.

The main theoretical contributions of this paper consist of (1) to extract the remaining valid information in the primary decomposition residual term of crude oil price, a new hybrid forecasting method with a secondary decomposition residual term is proposed. It helps to improve the completeness of the decomposition results and thus the accuracy of the forecasting results. (2) An empirical study is made for the theory that the CEEMDAN decomposition technique efficacy is superior to the VMD decomposition technique. (3) Combining internal influencing factors and external influencing factors, a more comprehensive indicator system of factors affecting crude oil prices was constructed, which can be used as a reference for future research. This paper's combined prediction model is discussed in Section 2 of the article. Section 3 deals with the analysis and features screening of factors influencing crude oil prices. Section 4 describes data sources and evaluation indicators for the model. The empirical findings based on projections of the current prices for Brent and WTI crude oil are analyzed and discussed in Section 5. The study is summarized in Section 6, which also lists its flaws.

Table 1: Taxonomy of crude oil price forecasts identified through the study literature review.

Model type	Literature
Historical data + hybrid model	[12, 15]
Historical data + external factors + hybrid model	[19]
Primary decomposition + historical data + single model	[13, 20–23]
Primary decomposition + historical data + hybrid model	[10, 16]
Secondary decomposition + historical data + single model	[29]
Secondary decomposition + historical data + external factors + single model	Proposed model

2. Experiments

2.1. Data Description. The time series chart for the prices of Brent crude oil and WTI crude oil shows that since 2004, there have been six major cycles in the price of crude oil: from 2004 to 2009, 2009 to 2016, 2016 to 2019, 2019 to 2020, 2020 to 2022, and 2022 to the present. Since 1997, global crude oil prices have steadily risen, reaching a peak in 2008. The 2008 financial crisis triggered a global recession and reduced investment and consumption by businesses and individuals, which had a knock-on effect on the demand for crude oil and led to a sharp fall in crude oil prices. With the economy gradually picking up in 2009, the demand for crude oil also gradually increased, and the price of crude oil fluctuated from $40. This was immediately followed by a slowdown in the world economic recovery between 2014 and 2016, which reduced the energy demand. And yet crude oil production from OPEC countries and oil-producing countries such as Russia has increased, with crude oil supply outstripping demand, ushering in a second decline in prices. Recent events, including the global COVID-19 outbreak in 2020 and the Russian-Ukrainian conflict in 2022 have caused the price of crude oil to plummet. Over all, the time series plot shows that crude oil prices are more susceptible to short-term contingencies such as financial, natural, and geopolitical events that can cause prices to rise or plummet.

Due to the fact that WTI and Brent Crude are the two most traded crude oil prices by investors, the changes in both of these prices as well as the spreads between them are very much in the spotlight of the crude oil market. The monthly historical prices of Brent crude oil and WTI were thus selected by this paper to serve as the experimental data samples; the historical prices of Brent crude oil can be found at https://fred.stlouisfed.org/series/POILBREUSDM, and the historical prices of WTI crude oil can be found at https://fred.stlouisfed.org/series/POILWTIUSDM. A total of 312 data points covering the period from January 1997 to

December 2022 comprise the sampling. Table 2 presents the descriptive statistics of the study's data. It is evident that, for Brent crude oil, the lowest price during this period was \$10.16, and the highest price was \$133.59, while WTI crude oil had the lowest price of \$11.28 and the highest price of \$133.93. A graphical representation of the dataset partition is shown in Figure 1. Typically, the data is split into two categories: training and test. The training data is used to estimate the prediction parameters, while the test data is used to evaluate the accuracy of the predictions. The study uses 234 observations from the first 80% of the series as the training set and 78 test data points from the final 20% as the test set to predict oil prices.

2.2. Data Processing. To guarantee the accuracy of the prediction results, the data were linearly transformed to eliminate the effect of outliers, and the data samples were normalized before prediction. The min-max normalization is used in this paper to preprocess the data, as shown in the following equation:

$$\chi' = \frac{\chi - \chi_{\min}}{\chi_{\max} - \chi_{\min}}, \tag{1}$$

where χ' is the transformed data; χ is the original data; $\chi_{\min}$ is the minimum value; and $\chi_{\max}$ is the maximum value.

2.3. Evaluation Criteria. The predictability of the model is tested using the statistical methods of MAE (mean absolute error), MSE (mean square error), R^2 (coefficient of determination), and MAPE (mean absolute percentage error) to confirm the experimental effect. The formula for each evaluation indicator is as follows:

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \\ \mathrm{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2, \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2}, \\ \mathrm{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%, \end{aligned} \tag{2}$$

where y_i is the true value of the price time series, $\hat{y}_i$ is the predicted value of the price time series, and n is the number of test samples.

The models' predictive abilities can be compared using MAE, MSE, R^2, and MAPE as criteria. However, the prediction errors between the two algorithms under comparison might not be significant, and the computational outcomes of MAE, MSE, R^2, and MAPE by themselves are insufficient to show which of the two models is more or less predictive than the other. As an illustration, He and Sun [31] investigated the learning performance of regularized large margin uniform machines (LUMs) for classification problems and used the comparison theorem to obtain the error bounds and learning rate of the overclassification error for error analysis. Consequently, in order to assess the degree of predictive power between the combined model used in this work and the comparison model in terms of statistical error, the Diebold–Mariano test is employed. The Diebold–Mariano test assumes the following expression for the prediction error $u_{i,t}$ of the two comparison models:

$$u_{i,t} = \hat{y}_{i,t} - y_t, \quad i = 1, 2, \tag{3}$$

where y_t is the true value and $\hat{y}_{i,t}$ is the prediction result of model i.

The null hypothesis for the Diebold–Mariano test is H0: $E(d_t) = 0$, which states that there is no discernible difference between the two models' predictive capacities. d_t is the relative loss function difference between the models, which can be represented as $d_t = g\ (u_{1,t}) - g\ (u_{2,t})$; $g(\cdot)$is the loss function. The Diebold–Mariano test statistic is calculated as follows:

$$DM = \frac{\overline{d}}{\sqrt{2\pi \hat{f}_d(0)/T}}, \tag{4}$$

where $\overline{d} = 1/T\sum_{t=1}^{T}(\mathrm{g}(u_{1,t}) - \mathrm{g}(u_{2,t}))$ is the mean value of the loss difference; The consistent estimate of $f_d(0)$ is represented by $\hat{f}_d(0)$, which also shows the spectral line density when the loss difference frequency is zero. There are several options for loss functions in the Diebold–Mariano test; in this work, the Diebold–Mariano test is performed using the MSE and MAPE loss functions.

3. Methodology

This section briefly introduces the CEEMDAN algorithm, the VMD algorithm, the SOA optimization algorithm, and the SVR prediction model, as well as the CEEMDAN-RES.-VMD-SOA-SVR model's construction procedure and principle.

3.1. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise. EMD is a decomposition approach proposed by Huang et al. that takes into account the time scale features of the data itself to cope with nonlinear and non-smooth complex signals. However, the modal components that result from its decomposition exhibit modal aliasing and endpoint effects. EEMD is an innovative method proposed by Wu and Huang for EMD modal aliasing. The generation of modal aliasing is effectively suppressed by multiple empirical modal decompositions with superimposed Gaussian white noise. However, the introduction of EEMD noise destroys the original signal to a certain extent, the reconstruction error is large, and the introduced noise will have residuals, affecting the feature extraction information. Later, the complementary ensemble empirical modal decomposition (CEEMD) method was proposed by Yeh et al. The redundant noise of the reconstructed signal is largely phase canceled during ensemble averaging when positive and negative pairs of

Table 2: The main numerical characteristics of research data.

Research data	Time span	Size	Mean	Max	Min	Std
Brent	January, 1997 to December, 2022	312	60.0658	133.59	10.16	31.80172
WTI		312	57.37962	133.93	11.28	28.36693

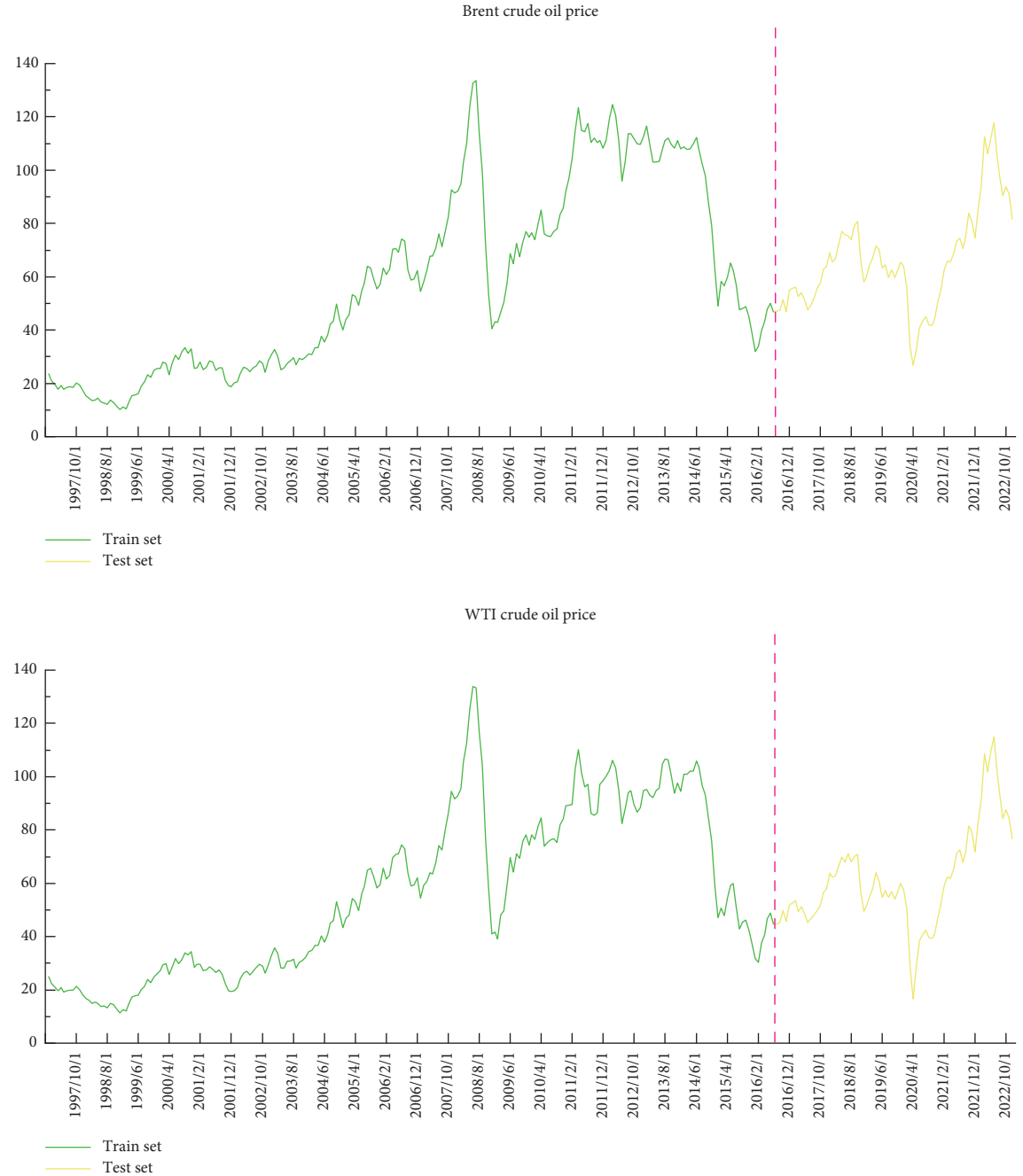

Figure 1: Training and test datasets of monthly crude oil spot prices where the green line is the training set and the yellow line is the test set.

complementary white noise are added to the original signal. This effectively improves the decomposition efficiency and solves the issues of large reconstruction error and poor decomposition completeness of EEMD. For the further ability to obtain more accurate and reliable decomposition results, a novel signal decomposition algorithm called the CEEMDAN method, has been proposed by Torres et al. The principle is to add white noise to the EMD decomposition process, and the first-order intrinsic modal components obtained by the decomposition are immediately ensemble averaged as the first IMF components obtained by the decomposition.

The crude oil price time series signal decomposition in this study is done using the CEEMDN approach. First, time series characteristics of crude oil prices are typically complex and include various levels of information such as trend, cycle, and volatility. These characteristics are important for forecasting the price of crude oil, and the CEEMDAN method can adaptively capture multiple frequency components in the time series of crude oil prices, better revealing the intrinsic structure of price volatility. Second, the CEEMDAN method successfully avoids the issue of inadequate decomposition of the first-order modal components in the decomposition process affecting the accuracy of the next-order modal components, thereby avoiding the impact on the full decomposition of the subsequent data signals. This method performs exceptionally well in eliminating noise interference. As a result, the CEEMDAN approach can offer researchers studying crude oil price forecasting a more precise intrinsic modal function (IMF), which can aid in enhancing the precision and stability of predicting outcomes by better revealing the inherent rule governing price swings. The particular decomposition of the CEEMDAN algorithm is as follows, Figure 2 visualizes the flowchart of the CEEMDAN algorithm.

Let $\mathrm{IMF}_i(t)$ represent the i th eigenmode component derived from EMD decomposition; $\overline{\mathrm{IMF}_i}(t)$ is the i th eigenmode component derived from CEEMDAN decomposition; $\mathrm{wn}^i(t)$ is a Gaussian white noise signal satisfying the standard normal distribution. $i = 1, 2, \cdots I$. is the number of times white noise was added and ε is the standard deviation of the white noise. $y(t)$ represents the signal to be decomposed, which is the gathered time series data on crude oil prices.

Step 1: Add Gaussian white noise to the signal to be decomposed $y(t)$. Get new signal $y'(t) = y(t) + \varepsilon wn^i(t)$.

Step 2: The EMD decomposition of the new signal $y'(t)$ is performed, and an overall average of the resulting modal components yields the first modal component of the decomposition $\overline{\mathrm{IMF}_1}(t) = 1/I\sum_{i=1}^{I}\mathrm{IMF}_1^i(t)$.

Step 3: Calculate the residual $r_1(t) = y(t) - \overline{\mathrm{IMF}_1}(t)$ after removing the first modal component. Continuing to add Gaussian white noise in $r_1(t)$ yields a new signal $y^{''(t)} = r_1(t) + \varepsilon wn^i(t)$.

Step 4: $y^{''(t)}$ serves as the carrier of the new signal for EMD decomposition. Overall averaging of the resulting modal components gives a second modal component $\overline{\mathrm{IMF}_2}(t) = 1/I\sum_{i=1}^{I}\mathrm{IMF}_2^i(t)$.

Step 5: Calculate the residual $r_2(t) = r_1(t) - \overline{\mathrm{IMF}_2}(t)$ after removing the second modal component.

Step 6: Until the residual signal obtained is a monotonic function, the aforementioned stages are repeated until the decomposition can no longer be carried out and the algorithm terminates. As of right now, K eigenmode components have been obtained. The original signal $y(t)$ is decomposed as $y(t) = \sum_{K=1}^{K}\overline{\mathrm{IMF}_K}(t) + r_K(t)$.

3.2. Variational Mode Decomposition. Instead of using the idea of a recursive solution, as used in traditional signal decomposition methods, VMD uses a completely nonrecursive modal decomposition to calculate IMF. By doing this, the signal decomposition process is spared from modal aliasing. The real-valued input signal $y(t)$ is to be broken down into a discrete number of modes u_k using VMD. Assuming that this mode contains frequency components that are all narrowband signals concentrated around a center frequency ω_k, the sum of the bandwidths of the center frequencies of each modal component is minimized. There are two components to the VMD algorithm: the variational problem's construction and solution. Restrained variational modeling can be expressed as follows:

$$\min_{\{u_k\},\{\omega_k\}}\left\{\sum_{k=1}^{K}\left\|\partial_t\left\{\left[\delta(t)+\frac{j}{\pi t}\right]u_k(t)\right\}e^{-j\omega_k t}\right\|_2^2\right\}, \tag{5}$$

subject to

$$\sum_k u_k = y(t), \tag{6}$$

where $y(t)$ represents the original signal, which is the residual term following CEEMDAN decomposition in this section; $\{u_k\} = \{u_1, \cdots, u_k\}$, $\{\omega_k\} = \{\omega_1, \cdots, \omega_k\}$ denote the shorthand notation for the set of modes and their corresponding center frequencies, respectively; t is the time series; k is the total number of modes; the time derivative is represented by ∂_t; the Dirac δ function; $\delta(t)$, in this study indicates the temporal localization of each IMF component; and the imaginary unit, j is used to represent the imaginary part of the resolved signal, which, when combined with the real part, forms the resolved signal in complex form.

The above constraints are transformed into an unconstrained variational problem by introducing a quadratic penalty factor and a Lagrange multiplier operator in Eq. The extended Lagrangian expression for

$$L(\{u_k\},\{\omega_k\},\quad \lambda) \coloneqq \alpha\sum_{k=1}^{K}\left\|\partial_t\left\{\left[\delta(t)+\frac{j}{\pi t}\right]u_k(t)\right\}e^{-j\omega_k t}\right\|_2^2 + \left\|y(t)-\sum_{k=1}^{K}u_k(t)\right\|_2^2 + \left\langle\lambda(t), y(t)-\sum_{k=1}^{K}u_k(t)\right\rangle, \tag{7}$$

where α is the penalization factor, also called equilibration parameter; $\lambda(t)$ is the Lagrange multiplier. The alternating direction multiplication method is used to obtain the solution formula for the mode u_k as follows:

$$\widehat{u}_k^{n+1}(\omega) = \frac{\widehat{y}(\omega) - \sum_{i\neq k}\widehat{u}_i(\omega) + (\widehat{\lambda}(\omega)/2)}{1 + 2\alpha(\omega-\omega_k)^2}. \tag{8}$$

The center frequency ω_k is calculated by solving the following equation:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega\left|\widehat{u}_k(\omega)\right|^2 d\omega}{\int_0^\infty \left|\widehat{u}_k(\omega)\right|^2 d\omega}. \tag{9}$$

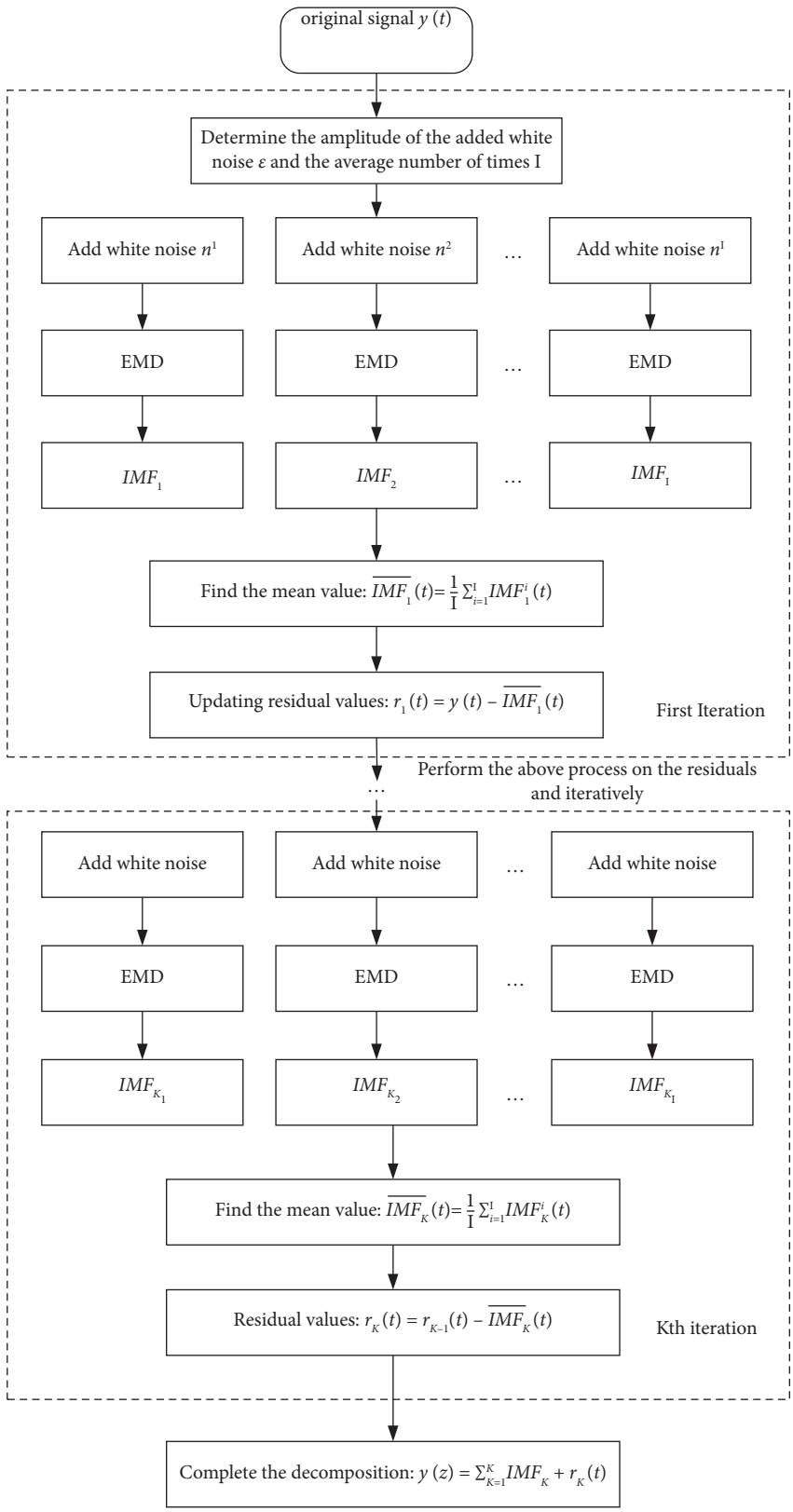

Figure 2: Flowchart of the CEEMDAN method.

Update based on the updated formulas for u_k and ω_k until the inner loop stops when the number of decompositions reaches K. Update λ according to the update formula for λ. Given the convergence accuracy ($\varepsilon > 0$), until the decision condition in the equation $\sum_k \|\hat{u}_k^{n+1} - \hat{u}_k^n\|_2^2 / \|\hat{u}_k^n\|_2^2 < \varepsilon$ is reached, the loop stops. Otherwise, return to continue (5) to continue the loop.

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \lambda\left(\hat{y}(\omega) - \sum_k \hat{u}_k^{n+1}(\omega)\right). \tag{10}$$

3.3. Seagull Optimization Algorithm. Seagulls are a common flocking seabird found all over the world that move across geographical areas as the seasons change in quest of food. A migration phase and a predation phase make up the gulls' predation cycle. During the migration phase, gulls maintain individual flight independence according to a certain pattern to keep from colliding with each other. In the attack phase, gulls attack their prey in a spiral flight.

3.3.1. Migratory (Global Search). During the migration process, the algorithm simulates how a flock of seagulls moves from one location to another. Gulls should avoid collisions at this stage. To prevent collisions with neighboring gulls, the algorithm utilizes an additional variable A to evaluate the new position of the gulls.

$$\begin{aligned} \overrightarrow{C_s}(t) &= A \times \overrightarrow{P_s}(t), \\ A &= f_c - \left(t \times \left(\frac{f_c}{\text{Max}_{\text{iteration}}}\right)\right), \end{aligned} \tag{11}$$

$\overrightarrow{C_s}(t)$ means a new position where no position conflicts exist with other gulls; $\overrightarrow{P_s}(t)$ means the current location of the seagull; t for the current iteration, $t = 0, 1, 2, \cdots$, Maxiteration; and A expresses the motion behavior of seagulls in a given search space. The frequency of the variable A, which is linearly reduced from 2 to 0, can be controlled by f_c.

(A) Direction of optimal position: after avoiding overlap with other gulls' positions, gulls move in the direction where the optimal position is located.

$$\begin{aligned} \overrightarrow{M_s}(t) &= B \times \left(\overrightarrow{P_{bs}}(t) - \overrightarrow{P_s}(t)\right) \\ B &= 2 \times A^2 \times rd. \end{aligned} \tag{12}$$

$\overrightarrow{M_s}(x)$ indicates the direction of the best position; B is the random number responsible for balancing the global and local searches; rd is a random number in the range of [0, 1]; $\overrightarrow{P_{bs}}(t)$ is the current optimal position; and $\overrightarrow{P_s}(t)$ means the current location of the seagull.

(B) Gulls approaching the optimal position: After moving a spot where they don't collide with one another, gulls proceed in the direction of the optimal position until they arrive at a new location.

$$\overrightarrow{D_s}(t) = \left|\overrightarrow{C_s}(t) + \overrightarrow{M_s}(xt)\right|. \tag{13}$$

$\overrightarrow{D_s}(x)$ is the new position of the gulls.

3.3.2. Attack (Local Search). During migration, gulls use their weight and wings to stay aloft, which enables them to alter the direction and velocity of their attacks. They move through the air in spiral shapes, attacking their victim. The behavior of the motion in the x, y, and z planes is described as follows:

$$\begin{aligned} x' &= r \times \cos\theta, \\ y' &= r \times \sin\theta, \\ z' &= r \times \theta, \\ r &= u \times e^{\theta v}, \end{aligned} \tag{14}$$

where r is the radius of each helix; θ is a randomized angular value in the range $[0, 2\pi]$. u and v are the correlation constants for the shape of the spiral. The Seagull's status after the attack is:

$$\overrightarrow{P_s}(t) = \overrightarrow{D_s}(t) \times x' \times y' \times z' + \overrightarrow{P_{bs}}(t), \tag{15}$$

$\overrightarrow{P_s}(t)$ is the gull's attack position.

3.4. Support Vector Regression. Support vector regression (SVR) is a regression method based on support vector machine (SVM). For the linearly differentiable situation, its model function is a linear function that fits the sample in vector space. In the case of linearly indivisible samples, a nonlinear mapping technique is used to convert the linearly indivisible samples in the low-dimensional input space into a high-dimensional feature space. This makes it possible to linearly examine the non-linear features of the samples using linear methods in high-dimensional feature spaces. The fundamental idea behind it is to build the best classification surface in the feature space by minimizing the structural risk theory based on a subset of training data patterns. This will enable the learning machine to achieve global optimization and satisfy an upper bound on the expected risk of the entire sample space with a certain degree of probability. SVR can handle high-dimensional data in nonlinear and small-sample prediction problems due to its outstanding generalization capabilities and excellent robustness to outliers. However, it also suffers from drawbacks such as high computational complexity and sensitive parameter tuning and needs to be used and tuned to the specific problem. The support vector regression (SVR) mathematical model is given below:

In the support vector regression model, external and internal characteristic variables are used to construct the input X, and the current value of the crude oil price is used as the corresponding output Y. The input X and output Y are constructed below. Based on ε-SVR, the optimization model is established as follows:

$$X=\begin{bmatrix} x_1 & \cdots & x_m \\ \vdots & \cdots & \vdots \\ x_n & \cdots & x_n \end{bmatrix}, \quad Y=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \tag{16}$$

objective function:

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$s.t.\begin{cases} y_i - (\omega x + b) \le \varepsilon + \xi_i, \\ (\omega x + b) - y_i \le \varepsilon + \xi_i^*, \\ \xi_i, \quad \xi_i^* \ge 0, \end{cases} \tag{17}$$

where $y(x) = \omega x + b$ is the prediction function; ω is the coefficient; SVR creates a spacing band on both sides of the linear function, and the maximum deviation ε between them is tolerated, calculating the loss when the deviation is larger than ε; C is a positive coefficient that affects the degree of penalty loss when training errors occur; ξ_i and ξ_i^* are slack variables, and both ξ_i and ξ_i^* are 0 for any sample x_i, if it is inside the isolation zone or on the edge of the isolation zone; if it is inside the barrier above the upper edge, then $\xi_i > 0$ and $\xi_i^* = 0$; and if it is below the lower edge of the barrier, then $\xi_i = 0$, $\xi_i^* > 0$.

For nonlinear intervals, a kernel function is required to be determined. The vector inner product space is extended by the kernel function, which allows a nonlinear regression problem to be transformed into an almost linear regression problem. Following the introduction of the Lagrange multipliers α_i, α_i^*, the SVR function model is expressed as follows:

$$y(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)k(x_i, x_j) + b, \tag{18}$$

$k(x_i, x_j)$ for the kernel function. In this paper, the radial basis function (RBF) is used as the kernel function to construct a nonlinear support vector regression (SVR) model.

3.5. The Proposed CEEMDAN-RES.-VMD-SOA-SVR Combined Forecasting Model. Due to the crude oil price as an input to the prediction model, which contains complex information, signal decomposition before prediction can better extract the features of the original signal to be passed to the prediction model, and improve the prediction accuracy. The key decomposition difficulty of discarding the remaining valid information in the residual term in earlier studies is effectively resolved by the CEEMDAN-RES.-VMD-SOA-SVR model suggested in this paper. Figure 3 illustrates the specific process of combining the CEEMDAN-RES.-VMD-SOA-SVR models.

First, the crude oil price time series is broken down into residual terms with varying frequencies and IMF subseries using CEEMDAN. The residual sequence is the part that cannot be further decomposed by the CEEMDAN decomposition technique, and the complicated information present in the residuals following modal decomposition is lost if the residual sequence is simply ignored. Second, since the VMD decomposition technique is different from the EMD recursive decomposition mode, the VMD decomposition is an adaptive and completely nonrecursive method for modal variational and signal processing, which is able to reduce the nonsmoothness of the time series with high complexity and nonlinearity and to obtain the subseries that contain multiple different frequency scales and are relatively smooth. Therefore, the VMD decomposition technique was chosen to decompose the residual sequence secondarily, yielding the VMF subsequence and another residual sequence. Finally, in the prediction step, this paper adopts the random forest technique to screen the external factors, the PACF analysis of the original series to screen the internal factors, and uses the seagull optimization algorithm to select the best parameters of the support vector regression prediction model. Figure 4 depicts the specific flow of the Seagull Optimization Support Vector Regression prediction model parameters.

4. Feature Importance Analysis

4.1. Establishment of a Primary Indicator System. One of the most significant sources of energy in the world is crude oil, and variations in its price have an impact on the macroeconomic development of various nations, inflation rates, etc., as well as the microeconomic consumption of the general public. Consequently, it is vital to take into account many different kinds of factors when making crude oil price predictions.

This paper adopts the literature analysis method to learn and summarize the analysis and research on the influence factors of crude oil price in the previous literature and establishes the junior indicator system of the influence factors of crude oil price. A search of relevant literature from CNKI based on the factors affecting crude oil prices yielded 56 papers in the initial search, and 16 core papers were obtained by eliminating irrelevant papers. According to the frequency of the influencing factors, 12 of the valid literature were analyzed in detail to establish a preliminary indicator system for the four influencing factors of crude oil prices, namely, commodity attributes, economic factors, alternative energy sources, and geopolitical factors. The preliminary indicator system is shown in Table 3.

4.1.1. Commodity Attributes. Firstly, crude oil is a fossil energy product. It has distinct commodity properties due to its substitutability, price volatility, and geographic location. As a product, market supply and demand are important factors affecting its price. OPEC's production and the OECD's member nations' consumption act as the primary drivers of supply and demand when it comes to changes in global oil prices.

4.1.2. Economic Factors. Furthermore, as a major strategic crude oil reserve, crude oil inventories are closely related to national energy security. Crude oil inventories are essential for stabilizing crude oil supply and demand. Meanwhile,

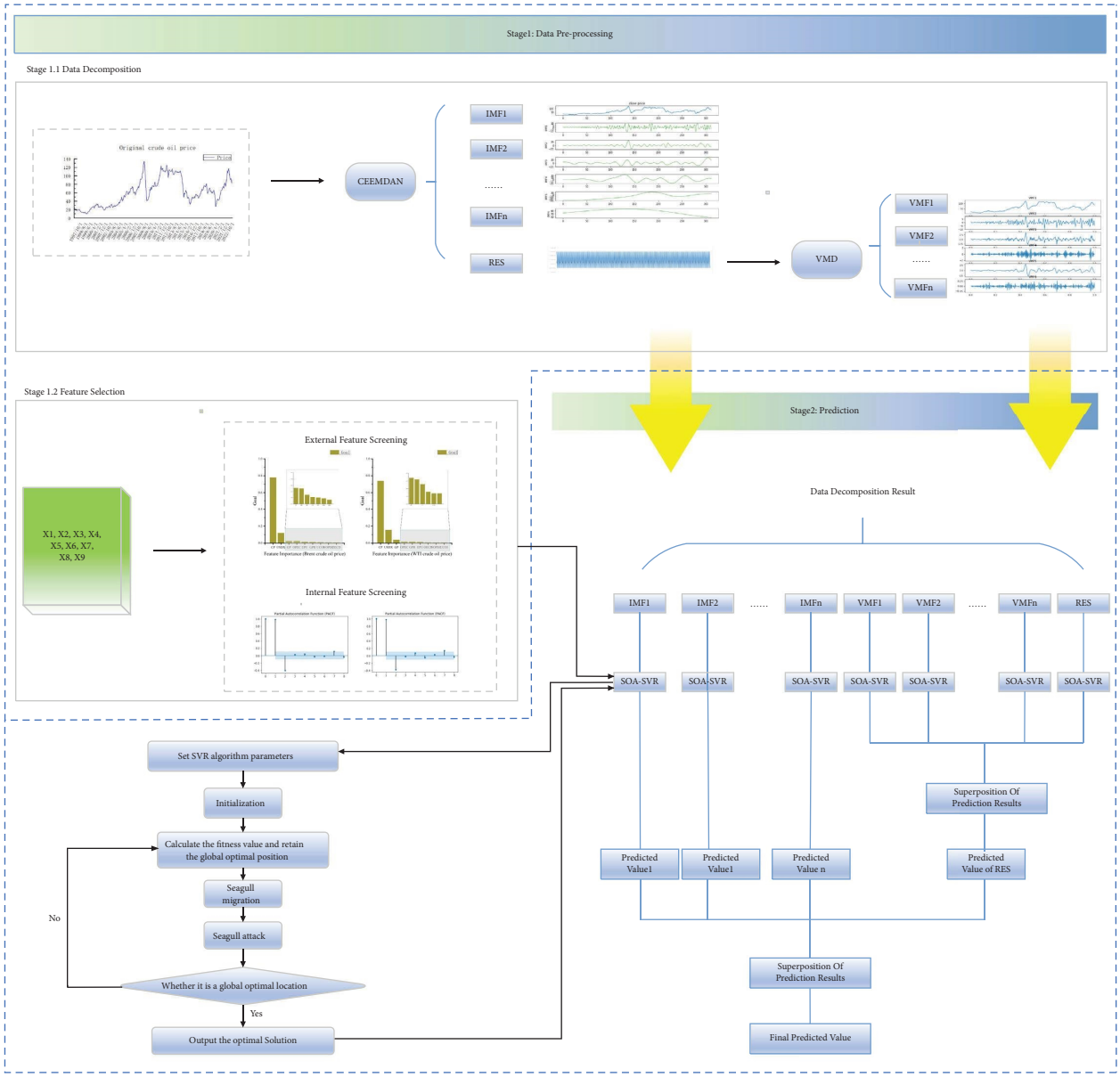

Figure 3: The process of CEEMDN-RES.-VMD-SOA-SVR model constructed.

crude oil has evolved into a financial product, incorporating more financial attributes. The operation of the global economy is one of the significant reasons for the rise, fall, and fluctuation of crude oil prices, and it has steadily become the variables that cannot be understated in impacting the price of crude oil. The major currency for trading crude oil internationally is the U.S. dollar; hence, the U.S. dollar index has a direct impact on crude oil prices. In addition, the global EPU has an impact on the financial market as well as the commodity market in numerous sectors of the worldwide economy [32], and the direction of the global economy inexorably influences the price of crude oil.

4.1.3. Alternative Energy Sources. Under the situation of global climate change, energy saving and emission reduction are a global consensus, and alternative energy will be the world's mainstream. With the high price of crude oil, the market demand for research and promotion of alternative energy is even greater.

4.1.4. Geopolitical Factors. Finally, the trajectory and volatility of the price of international crude oil are strongly correlated with the political stability of the major oil-producing and crude oil-exporting countries and regions as well as the global political environment, as a key industrial raw material and strategic material. Three kinds of geopolitical considerations are distinguished based on the nation in question. The first has to do with the Middle East and OPEC countries. These areas are the site of significant geopolitical events, and these events largely impact global crude oil prices by stifling the region's supply of crude oil. The second is related to the U.S.-Russia conflict. Currently, the

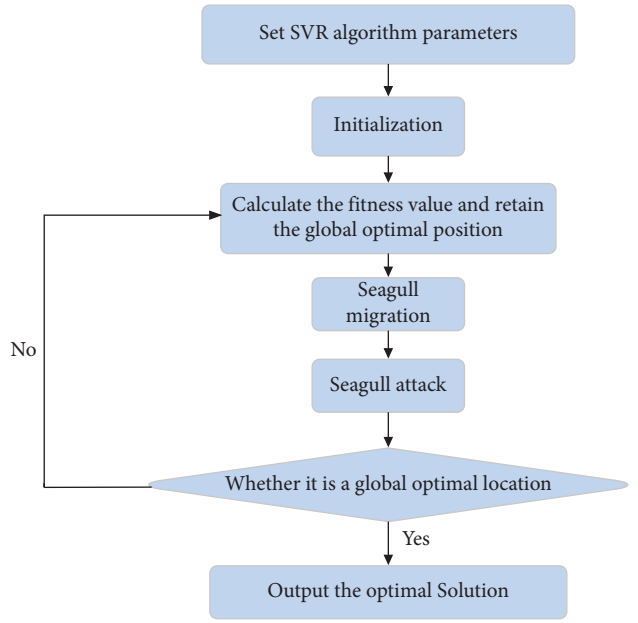

Figure 4: SOA optimization SVR parameters flowchart.

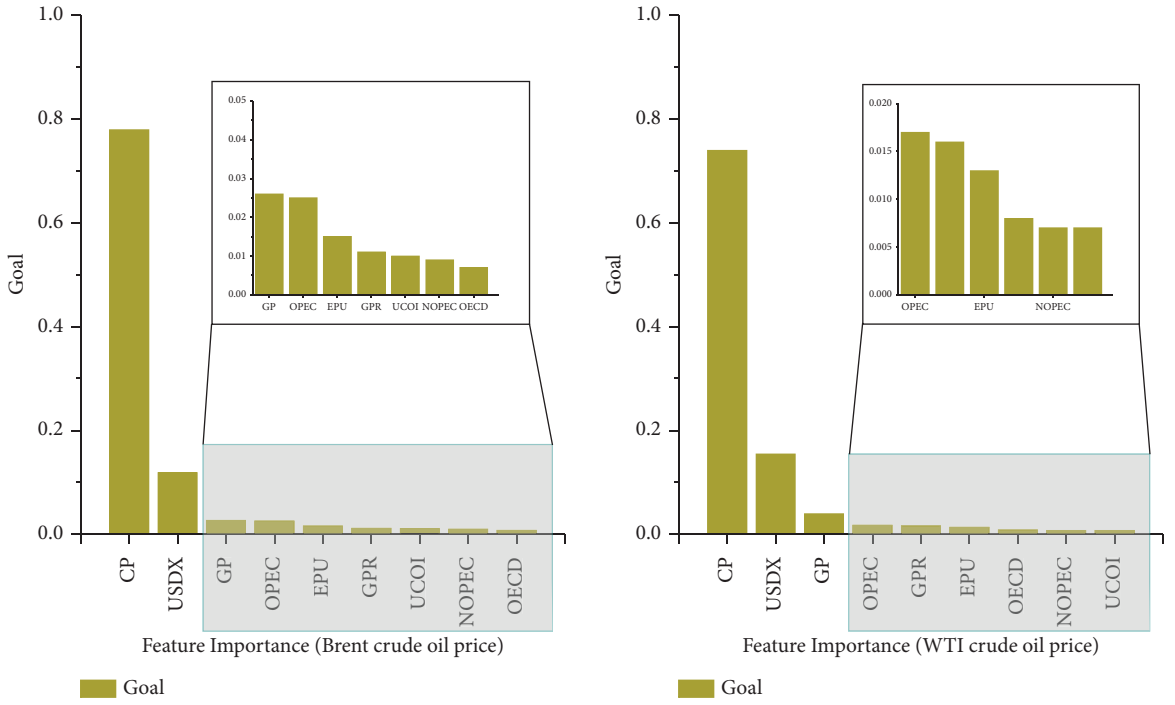

Figure 5: Feature importance.

U.S. and Russia are the world's first and third largest producers of crude oil, and both play a key role in global crude oil prices. The third is geopolitical reasons associated with trade friction between the US and China. Trade friction between the US and China primarily influences global economic growth and crude oil demand, which in turn affects global crude oil prices. In recent years, the outbreak of the Russo-Ukrainian war and the escalation of the Israeli-Palestinian conflict in 2023 have deeply affected crude oil prices.

In this paper, OPEC crude oil production and non-OPEC crude oil production are selected to represent crude oil consumption; OECD crude oil consumption is selected to represent crude oil consumption demand; U.S. crude oil inventories are selected to represent the level of crude oil stocks; power coal offshore spot price in the port of Newcastle, Australia, and natural gas spot price in the Henry Hub are selected to represent alternative energy sources; and the geopolitical risk index is selected to represent

Table 3: Indicators of factors influencing crude oil prices.

	Tier 1 indicators	Secondary indicators	Quantification of secondary indicators	Symbol representation
Factors affecting crude oil price	Commodity attributes	Crude oil supply	OPEC crude oil production	$X1$
			Non-OPEC crude oil production	$X2$
		Crude oil demand	OECD crude oil consumption	$X3$
		Crude oil inventory levels	U.S. crude oil inventories	$X4$
	Economic factors	US dollar index	USDX	$X5$
		Global economic policy uncertainty	EPU	$X6$
	Alternative energy sources	Natural gas prices	Henry hub natural gas spot price	$X7$
		Coal prices	Power coal offshore spot prices at the port of Newcastle, Australia	$X8$
	Geopolitical factors	Geopolitical risk index	GPR	$X9$

geopolitical factors. For the above variables, OPEC crude oil production, non-OPEC crude oil production, OECD crude oil consumption, U.S. crude oil inventories, Henry Hub natural gas spot price data, and power coal offshore spot price data for the Port of Newcastle, Australia, are from the U.S. Energy Information Administration (EIA) website. The data for the US dollar index is from the Wind database. Global Economic Policy Uncertainty (EPU) Index source https://www.policyuncertainty.com.Geopolitical Risk Index Source https://www.matteoiacoviello.com/gpr.htm. All indicators are based on monthly data with time intervals from January 1997 to December 2022.

4.2. Input Feature Screening. Feature selection refers to the elimination of unimportant and redundant features or the selection of effective and interacting features from a large number of candidate variables [33]. The running time of the algorithm will be significantly decreased, and the model's interpretability will be improved by only using the essential characteristics among all the features when building the model [34]. The principle of Random Forest feature importance assessment is to calculate the importance of sample feature variables by tree modeling, to quantitatively describe the degree of contribution of features to classification or regression, and to compare the magnitude of contribution between feature variables. Its benefits include the capacity to capture intricate feature interactions, increased robustness when processing high-dimensional data, and the fact that random forests' feature selection makes models simpler to understand. At the same time, however, Random forest has some drawbacks in selecting features. Random forest focuses mainly on the importance of individual features and is unable to explain the interactions between features. In this study, the Random Forest algorithm was selected to screen the raw feature variables for external influences, and the results are shown in Table 4. Figure 5 shows a visualization of the random forest feature screening. Characteristics $X1$, $X4$, $X5$, $X6$, $X7$, $X8$, and $X9$ Random Forest correlation coefficients greater than the 0.01 threshold are selected as external influences on the price of Brent crude oil in this paper. Characteristics $X1$, $X5$, $X6$, $X7$, $X8$, and $X9$ Random Forest correlation coefficients greater than the 0.01 threshold are selected as external influences on WTI crude oil prices in this paper.

Furthermore, PACF is used to determine the internal influences on crude oil prices, as crude oil prices are subject to economic and financial policy uncertainty as well as changes in geopolitical regimes with historical price lags. The autocorrelation function is the foundation of the PACF approach. It is an important statistic in time series analysis, which can be used to measure the correlation between the values of the same time series at different points in time, and can also be used in non-linear time series to analyze the biased correlation between the time series and its own lagged data. Specifically, for the crude oil price time series $\{y(t)\}$, the PACF value of $y(t)$ with $y(t-k)$ is the simple correlation coefficient between $y(t)$ and $y(t-k)$ after removing the indirect effects of $y(t-1)$, $y(t-2)$, $\cdots$, $y(t-k)$. From the results of the crude oil price PACF in Figure 6, the Brent crude oil price PACF is second-order truncated, indicating that the autocorrelation between the crude oil price and the lag of two months is strong, and the crude oil price with a one-month lag and a two-month lag of the Brent crude oil price is chosen as the internal influences. The WTI crude oil price PACF results are significant in the first, second, and seventh orders, and the crude oil prices with a one-month lag, two-month lag, and seven-month lag are chosen as the internal influences.

4.3. Feature Screening Comparison Experiment. A comparison was made with the initial feature variable prediction trials to confirm the efficacy of the filtered features. Tables 5 and 6 display the prediction performance results both before and after feature screening. The results of the feature comparison test demonstrate the effectiveness of the feature screening process. Dollar index ($X5$) and coal price ($X9$) get the greatest scores. The value of the coal price ($X9$) is considered to reflect the fact that there is a strong correlation between the volatility of crude oil prices and the substitution effect of fossil fuels; a fall in the price of coal would increase demand for coal, which in turn will impact the price of crude oil. Secondly, the score of the dollar index ($X5$) suggests that the changes in the value of the dollar have an equally significant impact on the price of crude oil. The price of crude oil increases in importing nations using their currencies when the dollar appreciates, and the demand for crude oil decreases.

5. Experimental Results and Discussion

Based on the criteria of representativeness and diversity, we used thirteen alternative models, including seven single models, VAR, BPNN, RF, ARIMA, ELM, XGBoost, SVR, and five combination models, SOA-SVR, VMD-RES.-SOA-SVR, VMD-CEEMDAN-SOA-SVR, CEEMDAN-SOA-SVR and CEEMDAN-RES.-SOA-SVR to illustrate the prediction performance of the combined prediction models proposed in this study. Table 7, Figures 7 and 8 displays the evaluation metrics findings for Brent and WTI crude oil, while Figure 9 illustrate the thirteen model's predictive power. The results show that the CEEMDAN-RES.-VMD-SOA-SVR combined prediction model developed in this paper outperforms the other models in all the evaluation indexes.

5.1. Forecasting Result. The models utilized in this research, Model 1 through 8, are prediction models that do not consider a decomposition technique. Model 9 utilizes the VMD technique to decompose the original price time series then performs SOA-SVR forecasts for each modal component of the decomposition separately, and then accumulates the forecasts for each sub-series. The specific steps of Model 10 are: firstly, the original price time series are decomposed by using the VMD technique, then decompose the decomposed modal components again using CEEMDAN technology, and finally forecast the modal components of the secondary decomposition using the SOA-SVR model,

Table 4: Random forest feature importance values.

Indicator ranking	Variable	Correlation factor-Brent	Variable	Correlation factor-WTI
1	**X8**	**0.779**	**X8**	**0.739**
2	**X5**	**0.118**	**X5**	**0.154**
3	**X7**	**0.026**	**X7**	**0.039**
4	**X1**	**0.025**	**X1**	**0.017**
5	**X6**	**0.015**	**X9**	**0.016**
6	**X9**	**0.011**	**X6**	**0.013**
7	**X4**	**0.01**	*X3*	0.008
8	*X2*	0.009	*X2*	0.007
9	*X3*	0.007	*X4*	0.007

Bold values represent factors with random forest correlation coefficients greater than 0.01, which are treated as external influences on crude oil prices.

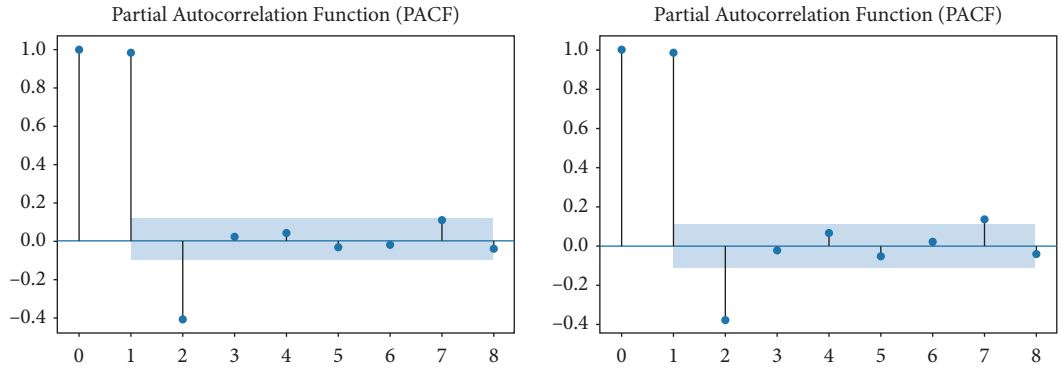

Figure 6: The PACF results of crude oil price data.

Table 5: Results of feature screening experiments (Brent).

Variables	MAE	MSE	R^2	MAPE
Original feature variables	0.258547543	0.067252309	0.999808462	0.426413042
Feature variables for screened	0.253209173	0.065713872	0.999812843	0.416608602

Table 6: Results of feature screening experiments (WTI).

Variables	MAE	MSE	R^2	MAPE
Original feature variables	0.126115703	0.017915159	0.999948721	0.227421878
Feature variables for screened	0.105476793	0.014019929	0.99995987	0.19782242

and the predictions of the components are summed up to arrive at the final prediction results. Model 11 utilizes the CEEMDAN technique to decompose the original price time series then performs SOA-SVR forecasts for each modal component of the decomposition without the residual term, respectively, and then accumulates the forecasts of each subseries. Model 12 differs from Model 11 in that the predictions of all modal components including the residual term produced by a first decomposition are summed up to form the final prediction.

In this paper, a combined forecasting model named CEEMDAN -RES.- VMD -SOA-SVR is developed. First, the original price time series are decomposed using the CEEMDAN decomposition technique to obtain each simple modal component. Second, the residual terms of the CEEMDAN decomposition are decomposed using the VMD variational decomposition technique. Finally, the overall prediction results of the VMF_n modal components of RES from the secondary decomposition of the VMD decomposition technique are added to the prediction results of the IMF_{n-1} modal components of the CEEMDAN decomposition to obtain the final prediction results. Table 8 shows the main terms covered in this paper. Table 9 demonstrates the characteristics of the different models. The advantage of the combined model presented in this paper is illustrated in Figure 9.

The following conclusions can be made by contrasting each model experiment's outcomes for the Brent and WTI datasets. When compared to Models 2–6, Model 7 performs better overall based on the benchmark predictive model evaluation result. Model 1's prediction effect on the dataset for Brent crude oil is superior to that of model 7, but it is inferior to that of model 7 on the dataset for WTI crude oil, suggesting that model 1's generalization is inadequate. Model 8 outperforms Model 7 after additional parameter adjustment of Model

Table 7: Comparison of forecasting performance of different models.

Prediction horizon	Model	Error indicator			
		MAE	MSE	R^2	MAPE
Brent	Model 1	0.75640808	7.235322874	0.979393405	1.499175621
	Model 2	3.084721757	15.02700232	0.957202277	4.834750621
	Model 3	2.552414749	11.26246556	0.967923883	4.14440152
	Model 4	4.327111644	32.68631888	0.906907579	6.858666476
	Model 5	7.540650463	54.22654482	0.674676565	5.111353319
	Model 6	3.048104627	16.97328049	0.951659171	4.119864135
	Model 7	2.176896112	7.885804418	0.977540799	3.496302775
	Model 8	0.935671962	0.875608146	0.997506377	1.541913286
	Model 9	0.923817867	0.853850097	0.997568188	1.522745834
	Model 10	0.277028123	0.109454043	0.999688269	0.457206033
	Model 11	0.266495464	0.071089735	0.999797533	0.439362401
	Model 12	0.266489663	0.071086592	0.999797542	0.439354093
	Model 13	**0.253209173**	**0.065713872**	**0.999812843**	**0.416608602**
WTI	Model 1	1.727588906	23.19526068	0.933607765	3.549475148
	Model 2	2.751387662	14.0782922	0.959703437	4.888496881
	Model 3	2.607948718	15.06218041	0.956887235	4.723907996
	Model 4	4.602267573	36.04344152	0.896832172	8.305570286
	Model 5	6.065383929	41.74257749	0.794649857	4.10834166
	Model 6	3.094695824	20.77134049	0.940545797	5.078767711
	Model 7	1.857406191	5.993485395	0.982844733	3.407799122
	Model 8	1.08142759	1.453631162	0.995814806	1.961160398
	Model 9	0.990683307	1.003060623	0.997128921	1.791215044
	Model 10	0.358825088	0.134193663	0.999615895	0.647166427
	Model 11	0.112444806	0.012667775	0.999963741	0.203303061
	Model 12	0.112436505	0.012665842	0.999963746	0.20328829
	Model 13	**0.105476793**	**0.012019929**	**0.999964732**	**0.19782242**

Bold values indicate that model 13 is the best experimental result among the 13 compared models.

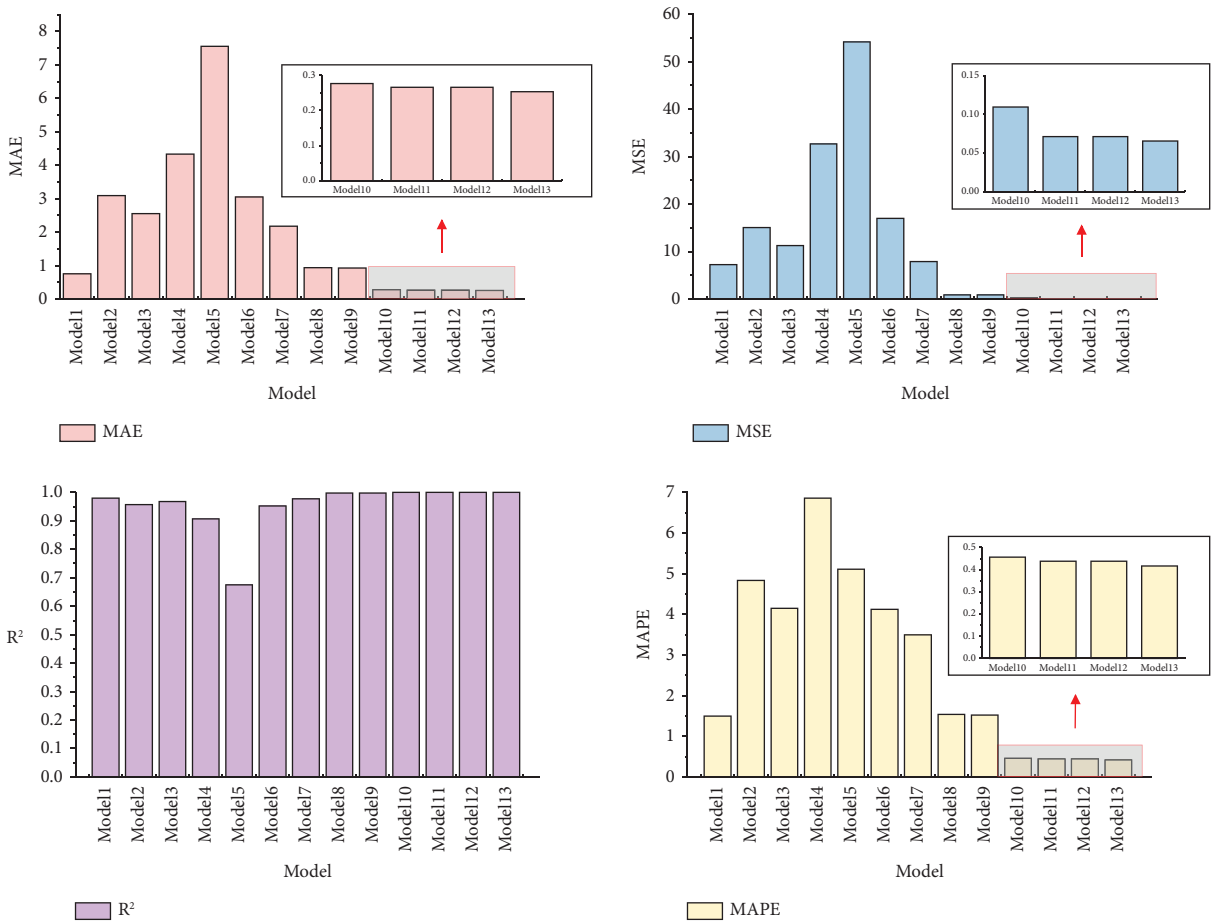

Figure 7: Performance of evaluation indicators for Brent crude oil prices under ten models.

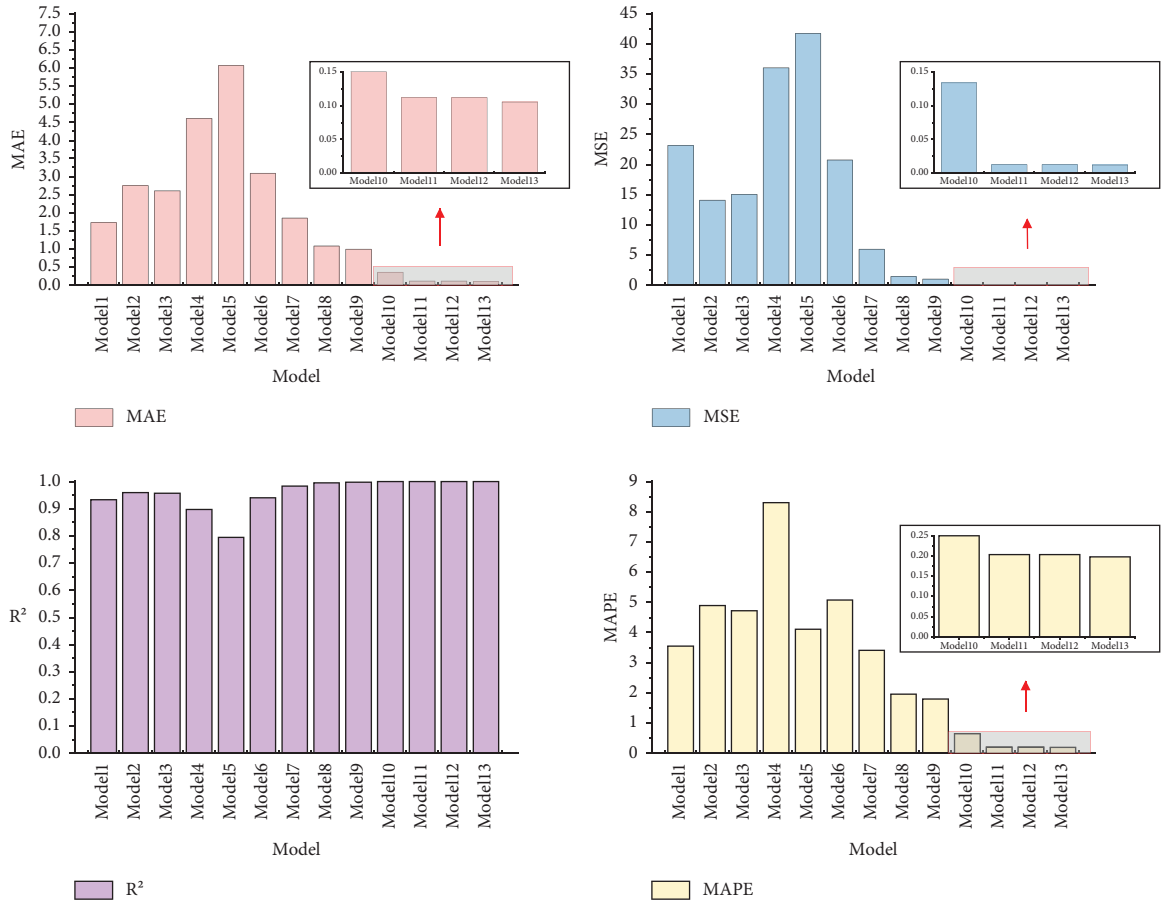

Figure 8: Performance of evaluation indicators for WTI crude oil prices under ten models.

7. This is due to the optimization of the SVR algorithm parameters using the seagull optimization algorithm, which automates the traversal of the SVR parameter selection. The optimized SVR model has improved convergence speed, better performance when dealing with high-dimensional data, reduced the occurrence of overfitting phenomenon, and made the SVR algorithm more stable and reliable when dealing with complex problems. Compared to a single model, Model 9 and Model 11 demonstrate that the inclusion of decomposition breaks down the original time series into several simple components, which helps to deal with the complexity and non-stationarity of the time series and improves the accuracy of the forecasting model. Furthermore, the prediction performance of Model 10 with the addition of a secondary decomposition beats the prediction performance of Model 9 with a primary decomposition. However, its prediction performance is still inferior to the primary decomposition of Model 11. This indicates that the CEEMDAN decomposition technique is more effective than the VMD decomposition technique for nonlinear and nonsmooth signal decomposition, and is more capable of restoring the accuracy of the original signal and preserving the energy decomposition of the signal, making the decomposition results more accurate and stable. Specifically, the advantage of the CEEMDAN technique in signal decomposition is that the weighting process of the noise is introduced before the decomposition, which can make CEEMDAN more robust in processing signals with higher noise levels. Meanwhile, CEEMDAN can better remove the noise part of the signal and decompose the residuals many times, to better capture the details and local features in the signal, which makes CEEMDAN well adapted to dealing with complex non-linear and non-smooth signals.

Comparing the prediction performance of Model 11 and Model 12, it can be found that Model 12 predicts better than Model 11 after adding the residual term. This indicates that the decomposed residual term contains the same complex information as the original time series, and the prediction by adding the residual term can improve the accuracy of the prediction results.

In comparison to previous models, the proposed combined model 13 has the best MAE, MSE, R^2, and MAPE values. It not only outperforms Model 12 but also predicts better results than Model 10 with quadratic decomposition. The results illustrate that Model 10 uses a quadratic decomposition-integration strategy, which improves the prediction accuracy. And the quadratic decomposition and prediction of the residual terms of the original time series can help to improve the model performance and enhance the model interpretability.

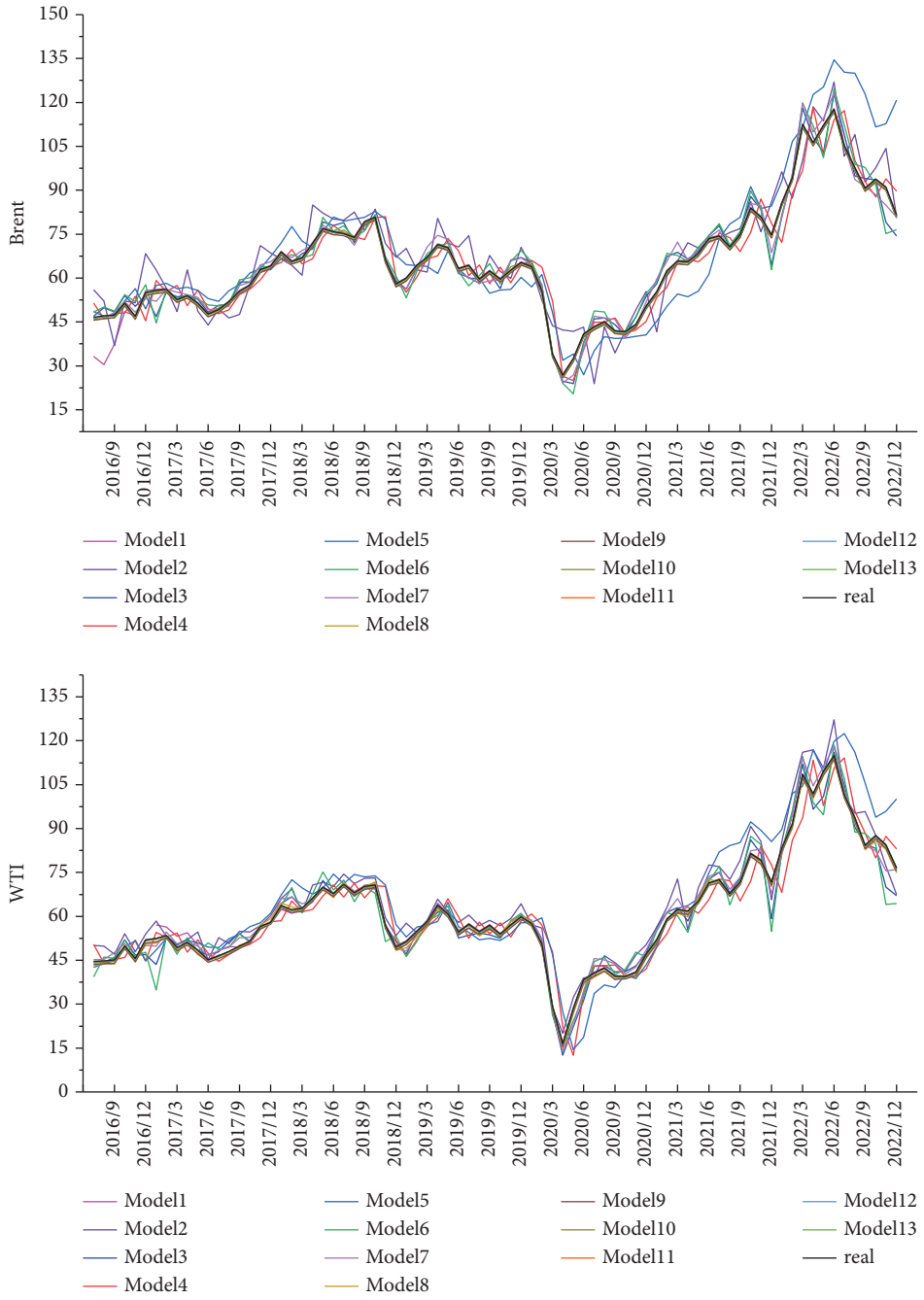

Figure 9: Impact of different model predictions.

Overall, as can be shown from the results of forecasting crude oil prices in two separate benchmark crude oil markets, the combined CEEMDAN -RES.-VMD -SOA-SVR forecasting model described in this study has the most satisfactory forecasting performance.

5.2. Analysis of Diebold–Mariano Test Results. This section employs the Diebold–Mariano test approach to examine the prediction performance between the comparison model and the CEEMDAN -RES.- VMD -SOA-SVR combination model suggested in this work from the standpoint of statistical error. Table 10 presents the results of the Diebold-Mariano test *P*-value for the CEEMDAN -RES.- VMD -SOA-SVR model in comparison to the other models discussed in this research for the two datasets, where the loss functions are MSE and MAPE, respectively. In the case of the Diebold–Mariano test, there is a significant difference in the predictive capacity of the two tested comparative models when the *P* value is less than 0.05, meaning that the original hypothesis is rejected at the 5% confidence level. The Diebold–Mariano test findings between the CEEMDAN -RES.-VMD -SOA-SVR model and each of the comparator models are analyzed as follows.

Table 8: Main terms covered in this paper.

Terminology	Specific meaning
CEEMDAN	Complete ensemble empirical mode decomposition with adaptive noise
VMD	Variable mode decomposition
VAR	Vector autoregressive model
BPNN	Back-propagation neural network
RF	Random forest
ARIMA	Autoregressive integrated moving average model
ELM	Extreme learning machine
XGBoost	Extreme gradient boosting
SOA	Seagull optimization algorithm
SVR	Support vector regression
MAE	Mean absolute error
MSE	Mean square error
MAPE	Mean absolute percentage error

Table 9: Characteristics of different models.

	Decomposition technique	Two-layer decomposition technique
VAR (model 1)		
BPNN (model 2)		
RF (model 3)		
ARIMA (model 4)		
ELM (model 5)		
XGBoost (model 6)		
SVR (model 7)		
SOA-SVR (model 8)		
VMD-SOA-SVR (model 9)	√	
VMD-CEEMDAN-SOA-SVR (model 10)	√	√
CEEMDAN-SOA-SVR (model 11)	√	
CEEMDAN- RES.-SOA-SVR (model 12)	√	
CEEMDAN -RES.- VMD -SOA-SVR (model 13)	√	√

Ps: model 12 and model 13 are decompositions that include the residual term after the first decomposition.

Table 10: Summary of DM test results.

Model	Brent crude oil dataset		WTI crude oil dataset	
	MSE	MAPE	MSE	MAPE
Model 1	$1.0261e-06$	$9.1364e-04$	$1.9866e-10$	$3.0922e-09$
Model 2	$2.9071e-09$	$7.5103e-13$	$1.6597e-07$	$2.9054e-10$
Model 3	$5.9466e-05$	$3.1781e-11$	$3.8538e-08$	$1.3185e-10$
Model 4	$2.5103e-10$	$4.8014e-15$	$4.3583e-08$	$1.3803e-12$
Model 5	$4.9128e-08$	$6.1358e-11$	$2.8343e-08$	$3.7800e-10$
Model 6	$7.0954e-07$	$1.4833e-11$	$6.6489e-04$	$2.1575e-08$
Model 7	$3.1893e-05$	$4.4891e-05$	$1.4838e-04$	$8.5191e-05$
Model 8	$2.5304e-04$	$4.4998e-04$	$1.4401e-03$	$2.9092e-03$
Model 9	0.0001	$9.6604e-03$	0.0007	$1.9009e-03$
Model 10	0.0090	0.0061	0.0073	0.0044
Model 11	0.0291	0.0184	0.0277	0.0150
Model 12	0.0493	0.0436	0.0376	0.0374

The P value of the Diebold–Mariano test for the CEEMDAN -RES.- VMD -SOA-SVR model under MSE and MAPE loss functions in comparison to the other models is less than 0.05 in both the Brent and WTI datasets. It is thus demonstrated that the CEEMDAN -RES.- VMD -SOA-SVR model presented in this research greatly outperforms the comparison models in the vast majority of situations with statistical significance in both the Brent and WTI datasets.

5.3. Comparison with Related Models. To enhance the demonstration of the combined model's accuracy in predicting crude oil prices in this work, a comparative experiment was carried out using methods suggested in the literature [35, 36]. way to carry out the comparative study. A proposed PHM model in the literature [35] comprises a linear autoregressive neural network (NAR), an autoregressive integral moving average model (ARIMA), and an

Table 11: Comparative results with existing advanced crude oil forecasting models.

Model	Literature [35]	Proposed model	Literature [36]	Proposed model
MAE	4.69	**0.4304**	—	—
RMSE	5.68	**0.4313**	0.0311	**0.0306**
MAPE	8.31	**0.4639**	5.44	**0.027**

Bold values indicate that the results of the evaluation metrics of the model proposed in this study are better than those of the comparison model.

exponential smoothing model (ESM). The data range is the monthly change in OPEC crude oil prices over the period January 2003 to August 2016, giving a total of 165 observations. The first 132 data are the training set and the last 32 data are the test set. Literature [36] used a combination of conventional economic data and GSVI data to suggest a hybrid K-means + KPCA + KELM approach based on the divide and conquer doctrine. The scope of the dataset is the monthly changes in WTI crude oil prices during the period from January 2004 to December 2018, where the data from January 2004 to December 2017 is the training set and the data from January 2018 to December 2018 is the test set. The combined model suggested in this work has the best prediction performance in MAE, RMSE, and MAPE, according to the experimental data, which are displayed in Table 11. This suggests that the model has excellent generalization ability and robustness in addition to improved prediction performance.

6. Conclusions

Following the decomposition-integration idea, this paper combines the scheme of the secondary decomposition technique with artificial intelligence predictive modeling. The following conclusions are drawn:

(1) The proposed CEEMDAN -RES.- VMD -SOA-SVR combined model is compared with the benchmark models VMD-SOA-SVR, VMD-RES.-SOA-SVR, VMD-CEEMDAN-SOA-SVR, CEEMDAN-SOA-SVR, CEEMDAN- RES.-SOA-SVR, and so on. In comparison, the CEEMDAN -RES.-VMD -SOA-SVR combined model has better predictive and robust predictive performance using MAE, MSE, R^2, and MAPE statistical methods to test the predictive performance of the model.

(2) Crude oil prices are affected by many complex factors, to avoid the impact of low correlation indicators on the model prediction effect, based on the establishment of the primary indicator system, the use of Random Forest Correlation Analysis and PACF to establish the characteristics of the variables of Brent crude oil prices and WTI crude oil prices. The influencing factor variables presented in this paper can be used to inform future crude oil price forecasting studies.

The shortcomings of this study are firstly, the lack of generalized rules for selecting VMD parameters and secondly, the single choice of methods for screening the characteristic variables. What's more, although the combined model this study suggests performs better in terms of prediction, its complexity and the multitude of parameter options cause it to take longer to run. To enhance the overall forecast, future considerations will be given to streamlining the model run procedures and refining the VMD algorithm's parameters.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 81973791).

References

[1] Q. Ji and Y. Fan, "How does oil price volatility affect non-energy commodity markets?" *Applied Energy*, vol. 89, no. 1, pp. 273–280, 2012.

[2] Y. Chen, X. Zhu, and H. Li, "The asymmetric effects of oil price shocks and uncertainty on non-ferrous metal market: based on quantile regression," *Energy*, vol. 246, Article ID 123365, 2022.

[3] Z. Liu, Z. Ding, T. Lv, J. S. Wu, and W. Qiang, "Financial factors affecting oil price change and oil-stock interactions: a review and future perspectives," *Natural Hazards*, vol. 95, no. 1-2, pp. 207–225, 2018.

[4] W. Qiang, A. Lin, C. Zhao, Z. Liu, M. Liu, and X. Wang, "The impact of international crude oil price fluctuation on the exchange rate of petroleum-importing countries: a summary of recent studies," *Natural Hazards*, vol. 95, no. 1-2, pp. 227–239, 2018.

[5] M. S. Kim, "Impacts of supply and demand factors on declining oil prices," *Energy*, vol. 155, pp. 1059–1065, 2018.

[6] J. Zhou, M. Sun, D. Han, and C. Gao, "Analysis of oil price fluctuation under the influence of crude oil stocks and US dollar index— based on time series network model— based on time series network model," *Physica A: Statistical Mechanics and its Applications*, vol. 582, Article ID 126218, 2021.

[7] M. Bildirici, N. Guler Bayazit, and Y. Ucan, "Analyzing crude oil prices under the impact of COVID-19 by using LSTAR-GARCHLSTM," *Energies*, vol. 13, no. 11, p. 2980, 2020.

[8] K. Khan, C.-W. Su, R. Tao, and M. Umar, "How do geopolitical risks affect oil prices and freight rates?" *Ocean & Coastal Management*, vol. 215, Article ID 105955, 2021.

[9] Y. Chen, C. Zhang, K. He, and A. Zheng, "Multi- step- ahead crude oil price forecasting using a hybrid grey wave model," *Physica A: Statistical Mechanics and its Applications*, vol. 501, pp. 98–110, 2018.

[10] J. Wu, Y. Chen, T. Zhou, and T. Li, "An adaptive hybrid learning paradigm integrating CEEMD, ARIMA and SBL for crude oil price forecasting," *Energies*, vol. 12, no. 7, p. 1239, 2019.

[11] D. P. Dash, N. Sethi, and D. P. Bal, "Is the demand for crude oil inelastic for India? Evidence from structural VAR analysis," *Energy Policy*, vol. 118, pp. 552–558, 2018.

[12] W. Kristjanpoller and M. C. Minutolo, "Forecasting volatility of oil price using an artificial neural network-GARCH model," *Expert Systems with Applications*, vol. 65, pp. 233–241, 2016.

[13] Y. Ding, "A novel decompose-ensemble methodology with AIC-ANN approach for crude oil forecasting," *Energy*, vol. 154, pp. 328–336, 2018.

[14] J. Wang, T. Zhang, T. Lu, and Z. Xue, "A hybrid forecast model of EEMD-CNN-ILSTM for crude oil futures price," *Electronics*, vol. 12, no. 11, p. 2521, 2023.

[15] X. Ding, L. Fu, Y. Ding, and Y. Wang, "A novel hybrid method for oil price forecasting with ensemble thought," *Energy Reports*, vol. 8, pp. 15365–15376, 2022.

[16] Y. Cheng, J. Yi, X. Yang, K. K. Lai, and L. Seco, "A CEEMD-ARIMA-SVM model with structural breaks to forecast the crude oil prices linked with extreme events," *Soft Computing*, vol. 26, no. 17, pp. 8537–8551, 2022.

[17] Y. Xia, Y. Hou, and S. Lv, "Learning rates for partially linear support vector machine in high dimensions," *Analysis and Applications*, vol. 19, no. 01, pp. 167–182, 2021.

[18] T. Mao, Z. Shi, and D. Zhou, "Approximating functions with multi-features by deep convolutional neural networks," *Analysis and Applications*, vol. 21, no. 01, pp. 93–125, 2023.

[19] J. Liu and X. Huang, "Forecasting crude oil price using event extraction," *IEEE Access*, vol. 9, pp. 149067–149076, 2021.

[20] K. He, L. Yu, and K. K. Lai, "Crude oil price analysis and forecasting using wavelet decomposed ensemble model," *Energy*, vol. 46, no. 1, pp. 564–574, 2012.

[21] T. Fang, C. Zheng, and D. Wang, "Forecasting the crude oil prices with an EMD-ISBM-FNN model," *Energy*, vol. 263, Article ID 125407, 2023.

[22] Y. X. Wu, Q. B. Wu, and J. Q. Zhu, "Improved EEMD-based crude oil price forecasting using LSTM networks," *Physica A: Statistical Mechanics and its Applications*, vol. 516, pp. 114–124, 2019.

[23] Y. Huang and Y. Deng, "A new crude oil price forecasting model based on variational mode decomposition," *Knowledge-Based Systems*, vol. 213, Article ID 106669, 2021.

[24] T. Li, Z. Qian, W. Deng, D. Zhang, H. Lu, and S. Wang, "Forecasting crude oil prices based on variational mode decomposition and random sparse Bayesian learning," *Applied Soft Computing*, vol. 113, Article ID 108032, 2021.

[25] Y. Zhao, W. Zhang, X. Gong, and C. Wang, *A Novel Method for Online Real-Time Forecasting of Crude Oil price*, Elsevier BV, Amsterdam, Netherlands, 2021.

[26] J. Li, Q. Wu, Y. Tian, and L. Fan, "Monthly Henry Hub natural gas spot prices forecasting using variational mode decomposition and deep belief network," *Energy*, vol. 227, Article ID 120478, 2021.

[27] J. Li and D. Liu, "Carbon price forecasting based on secondary decomposition and feature screening," *Energy*, vol. 278, Article ID 127783, 2023.

[28] W. Sun, B. Tan, and Q. Wang, "Multi-step wind speed forecasting based on secondary decomposition algorithm and optimized back propagation neural network," *Applied Soft Computing*, vol. 113, Article ID 107894, 2021.

[29] T. Zhang, Z. Tang, J. Wu, X. Du, and K. Chen, "Multi-step-ahead crude oil price forecasting based on two-layer decomposition technique and extreme learning machine optimized by the particle swarm optimization algorithm," *Energy*, vol. 229, Article ID 120797, 2021.

[30] G. Li, S. Yin, and H. Yang, "A novel crude oil prices forecasting model based on secondary decomposition," *Energy*, vol. 257, Article ID 124684, 2022.

[31] X. He and H. Sun, "Error analysis of classification learning algorithms based on lums loss," *Mathematical Foundations of Computing*, vol. 6, no. 4, pp. 616–624, 2023.

[32] Y. Feng, D. Xu, P. Failler, and T. Li, "Research on the time-varying impact of economic policy uncertainty on crude oil price fluctuation," *Sustainability*, vol. 12, no. 16, p. 6523, 2020.

[33] G. Wei, J. Zhao, Y. Feng, A. He, and J. Yu, "A novel hybrid feature selection method based on dynamic feature importance," *Applied Soft Computing*, vol. 93, Article ID 106337, 2020.

[34] X. Liu, H. Tang, Y. Ding, and D. Yan, "Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings," *Energy and Buildings*, vol. 273, Article ID 112408, 2022.

[35] A. Safari and M. Davallou, "Oil price forecasting using a hybrid model," *Energy*, vol. 148, pp. 49–58, 2018.

[36] Y. Yang, J. Guo, S. Sun, and Y. Li, "Forecasting crude oil price with a new hybrid approach and multi-source data," *Engineering Applications of Artificial Intelligence*, vol. 101, Article ID 104217, 2021.