Article

Crude Oil Futures Price Forecasting Based on Variational and Empirical Mode Decompositions and Transformer Model

Linya Huang [1], Xite Yang [2], Yongzeng Lai [3] , Ankang Zou [4] and Jilin Zhang [5],*

1   Sichuan Provincial Health Information Center, Chengdu 610000, China; liny_huang@163.com
2   Business School, Sichuan University, Chengdu 610064, China; yxt@stu.scu.edu.cn
3   Department of Mathematics, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada; ylai@wlu.ca
4   Changsha Digital Cloud Chain Technology Co., Ltd., Changsha 410000, China; zakemail2021@163.com
5   School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350108, China
*   Correspondence: jlz@fjut.edu.cn

Abstract: Crude oil is a raw and natural, but nonrenewable, resource. It is one of the world's most important commodities, and its price can have ripple effects throughout the broader economy. Accurately predicting crude oil prices is vital for investment decisions but it remains challenging. Due to the deficiencies neglecting residual factors when forecasting using conventional combination models, such as the autoregressive moving average and the long short-term memory for prediction, the variational mode decomposition (VMD)-empirical mode decomposition (EMD)-Transformer model is proposed to predict crude oil prices in this study. This model integrates a second decomposition and Transformer model-based machine learning method. More specifically, we employ the VMD technique to decompose the original sequence into variational mode filtering (VMF) and a residual sequence, followed by using EMD to decompose the residual sequence. Ultimately, we apply the Transformer model to predict the decomposed modal components and superimpose the results to produce the final forecasted prices. Further empirical test results demonstrate that the proposed quadratic decomposition composite model can comprehensively identify the characteristics of WTI and Brent crude oil futures daily price series. The test results illustrate that the proposed VMD–EMD–Transformer model outperforms the other three models—long short-term memory (LSTM), Transformer, and VMD–Transformer in forecasting crude oil prices. Details are presented in the empirical study part.

Keywords: crude oil price; variational mode decomposition; empirical mode decomposition; machine learning methods; Transformer model

MSC: 91-02

1. Introduction

Crude oil is a crucial fundamental energy source and a vital strategic resource in promoting economic prosperity and ensuring political stability for countries [1,2].

Moreover, it also holds significant value as an investment instrument in financial markets. The considerable fluctuations in crude oil prices are likely to have ramifications on the general stability of the financial system, international politics, and the military [3]. In addition, the uncertainty brought about by the volatility of crude oil prices is anticipated to escalate the operating expenditure of corporations and impact investment patterns, which may impede the steadiness of economic expansion [4,5]. Traditional econometric models, machine learning technology, and hybrid models incorporating prediction methods have been broadly utilized as approaches to forecasting futures or commodity prices. Some scholars have demonstrated the use of traditional econometric models for stock and futures price forecasting, particularly the Markov models and the ARIMA models [5–9]. Hossain et al. (2006) [7] found that the ARIMA model was effective in forecasting the price of mung

beans. Traditional econometric models, derived from classical statistical frameworks, necessitate rigorous statistical analysis to examine their constituent parameters. Each estimated parameter in these models bears significant explanatory relevance. However, traditional econometric models require the forecast object to meet the hypothesis of data stationarity. Owing to the high-density, nonstationary, nonlinear, and complex characteristics of the futures price time series, the forecasting accuracy of traditional models is relatively limited [10]. Subsequently, further advancements in applying conventional models in forecasting research have been impeded.

However, the rapid development of artificial intelligence, especially machine learning technologies, broke this limit, which did not make strict assumptions on the function form of the model, the interaction between variables, and the statistical distribution of parameters [11]. Artificial intelligence and machine learning technologies demonstrate superior generalization and antinoise performance compared to traditional econometric models in forecasting complex financial time series data characterized by high density, nonlinearity, and nonstationarity [12]. Shin et al. (2013) and Yu et al. (2017) [13,14] adopted a neural network model and most minor square support vector to predict the price of light crude oil in West Texas, USA. Chen et al. (2016) [15] used artificial neural networks and support vector machines (SVMs) to improve the accuracy of gold futures price forecasting. Wang and Li (2018) [16] demonstrated that singular spectrum analysis (SSA) and a combined neural network model exhibit greater forecasting accuracy relative to conventional models, specifically in predicting the prices of corn, gold, and crude oil. Kohzadi et al. (1996) [17] found that incorporating nonlinearity into feedforward neural network models improved forecasting performance for agricultural product prices, surpassing traditional ARIMA models. However, the conventional feedforward neural network still has shortcomings, such as slow convergence speed, parameter adjustment complexity, poor generalization, antinoise capacity, etc. To avoid the shortcomings of traditional models, Wang et al. (2015) [18] used the backpropagation (BP) neural network models that can recognize nonlinear data for gold price forecasting, significantly improving accuracy. However, the BP neural network still has some defects in commodity price forecasting theory, such as not reflecting the time series relationship of gold price in the training process.

Despite the superior capabilities of artificial intelligence models in identifying nonlinear crude oil price sequences, the complexity of crude oil pricing is influenced by many factors and exhibits periodic and nonlinear characteristics. It is imperative to pursue further developments to enhance the predictive precision of data with nonlinear and high complexity; a new "decomposition and ensemble" method, along with combined models, is consequently proposed. It can be inferred that scholars are more inclined towards decomposing the initial time series into its smoother mode components primarily first, and, subsequently, the forecast of the individual decomposed mode components is conducted, followed by a comprehensive integration of all components to derive the ultimate forecasted prices [19,20]. The EMD [21], as a typical adaptive time domain signal decomposition approach, can decompose the original signal by step and generate a series of sequences with different characteristics of frequency domain characteristics.

The EMD is an adaptive time-domain signal decomposition technique that decomposes the original signal stepwise, resulting in sequences with varying frequency-domain characteristics. Due to the emphasis on the local characteristics of intrinsic mode functions (IMFs), the IMF-based forecasting analysis can more effectively use original data information. The integration of EMD and the predictive model is a combination model of financial series analysis and futures price prediction [22,23]. Zhu (2018) [24] conjoined the EMD with the minimum daily support vector machine and kernel function model to predict carbon prices. Zhang et al. (2008) [25] employed EMD and data-integration methods to predict the price of international crude oil. Azevedo et al. (2016) [26] utilized the autoregressive moving average (ARMA) model, smooth index model, and dynamic regression combined model to forecast light crude oil prices and Brent crude oil prices in Western

Texas. Safari et al. (2018) [27] demonstrated that the integrated index smooth model, the ARMA model, and the nonlinear regression neural network outperform individual models (such as the ARMA and LSTM) in forecasting oil prices. Dragomiretskiy and Zosso (2013) [28] disseminated the variational mode decomposition (VMD) methodology built upon EMD decomposition, achieving a multiresolution mode decomposition. Compared with studies on single machine learning models, those on the current hybrid model showed a considerable enhancement in predictive accuracy. Nonetheless, investigation into mixed modeling centered on the VMD decomposition technique frequently overlooks residual elements, engendering a decline in the precision of information. This paper employs the EMD technique to conduct a secondary decomposition after the VMD process to address this issue.

The artificial neural network (ANN), considered one of the most powerful techniques in the machine learning model [29], plays a significant role in future events forecast efficiently [30]. In particular, the Transformer model [31] is an efficient neural network model based on an attention mechanism, which can capture the interaction information between any pair of time steps. It has been shown to possess powerful modeling capability and finer-grained data processing. Compared with the LSTM model, the Transformer model sacrifices more parameters to utilize long-term information, thereby better capturing the long-term correlation within sequential data. As such, it is commonly employed in forecasting stock prices [32,33]. Muhammad et al. (2022) [29] proposed a Transformer-based model for predicting the daily and monthly stock prices of the Dhaka Stock Exchange (DSE). The model consists of two input layers, three Transformer layers, one pooling layer, two dropout layers, and two dense layers. The model exhibits excellent predictive performance.

The main contributions of this study are described as follows. Given the high predictive accuracy of the Transformer model when applied to time series data, this article forecasts the value of the VMFs after the VMD decomposition and the importance of IMFs after the second decomposition of the EMD by utilizing the Transformer model. Based on the WTI and Brent crude oil price series, this study uses a secondary decomposition combined model to predict the price of WTI and Brent crude oil, respectively. Specifically, this paper employs the VMD–EMD–Transformer model, which first utilized the VMD method to decompose the original WTI and Brent crude oil price series into various subseries. The original series was then subtracted from each subseries to obtain the residual series, which was further decomposed using EMD. The Transformer model was then used to forecast all decomposed components, and the predicted values were linearly combined to form the forecast result of crude oil prices. The empirical study results clearly demonstrate that our proposed VMD–EMD–Transformer model outperforms the other three models. Details are provided in Section 3.

The organization of the remainder of this paper is as follows: Section 2 presents an overview of diverse approaches for the forecast of crude oil prices. Section 3 comprises the empirical investigation of the article, encompassing the assessment of models for forecasting crude oil futures prices. Ultimately, this paper culminates in exposing the conclusions drawn from our empirical analysis in Section 4.

2. Model Description

2.1. The VMD Model (Variational Mode Decomposition)

The VMD method, introduced by [28], provides a robust framework for decomposing nonlinear and nonstationary signals into discrete subsignals (modes) with limited bandwidth. This process isolates each mode's frequency domain characteristics iteratively while minimizing the total bandwidth.

To describe the VMD process in detail, consider a complex signal composed of multiple oscillatory components. The VMD method separates this signal into distinct modes, each representing a specific frequency range. For instance,

- Mode 1 typically captures high-frequency noise or rapid fluctuations.
- Mode 2 focuses on medium-frequency patterns, often associated with cyclic behaviors.

- Mode 3 isolates low-frequency trends, reflecting the underlying long-term behavior of the signal.

Through this decomposition, VMD enables the separation of meaningful patterns from noise, providing a more precise representation of the signal's structure and improving the subsequent modeling process.

Mathematically, the VMD method is formulated as an iterative variational framework that seeks to minimize the total bandwidth of all modes while maintaining the integrity of the original signal. This is achieved by solving the following variational problem:

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

here, w_k represents the central frequency of the k-th mode, used to identify the dominant frequency components in each mode. j denotes the imaginary unit from the Hilbert transform, essential for computing analytic signals, where t is the time; $\{u_k, k \geq 1\}$ and $\{w_k, k \geq 1\}$ are the modal components obtained after decomposition and corresponding central frequency; K represents the number of modes of decomposition; $f(t)$ indicates the time series signal of the original input; u_k represents the kth term in the subseries of $f(t)$; $\delta(t)$ is the pulse function; $\otimes$ is a convolution operator. We introduce a quadratic penalty function term α and a Lagrange multiplier λ to find the optimal solution of the constrained variational model. Therefore, the variational modal calculation formula of the constraint function is deformed as follows:

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

where λ represents the Lagrange multiplier, ensuring that the reconstructed signal from all modes matches the original signal $f(t)$, enforcing the decomposition constraint. α represents the penalty function that controls the bandwidth of each mode. Larger values of α lead to tighter bandwidths, resulting in more distinct mode separation. However, overly large values may lead to loss of signal information. Then, we transform the Lagrangian function from the time domain to the frequency domain and calculate the corresponding extreme value. We obtain the optimal solution of the constrained variational model using the alternating direction method of multipliers (ADMM). Subsequently, we transform the Lagrangian function from the time domain to the frequency domain and compute the corresponding extrema. The optimal solution $\{(u_k, w_k), k \geq 1\}$ to the constrained variational model can be obtained via the ADMM algorithm as follows:

$$\hat{u}_k = \frac{\hat{f}(w) - \sum_{i \neq k} \hat{u}_i(w) + \frac{\hat{\lambda}(w)}{2}}{1 + 2\alpha(w - w_k)^2}, \tag{3}$$

and

$$\hat{w}_k = \frac{\int_0^\infty \omega |\hat{u}_k(w)|^2 dw}{\int_0^\infty |\hat{u}_k(w)|^2 dw}. \tag{4}$$

2.2. EMD

Empirical mode decomposition (EMD), introduced by [34], is an adaptive time-domain signal decomposition technique designed to decompose a signal into intrinsic mode functions (IMFs) and a residual term. This stepwise method captures localized frequency-domain characteristics of the original time series, making it suitable for nonlinear, nonstationary data. Each IMF must satisfy two prerequisites: (1) the number of extrema and zero crossings in the signal must be equal or differ by at most one, and (2) the mean value of the envelopes formed by local maxima and minima must be zero. If these conditions are not

met, the data series undergoes stationary normalization via spline interpolation to compute upper and lower envelopes.

The EMD method adaptively decomposes a complex time series signal into a finite set of intrinsic mode functions (IMFs) and a residual term. Unlike traditional decomposition methods, EMD does not rely on predefined basis functions but, instead, extracts patterns directly from the data.

The decomposition process involves iterative sifting, where the local maxima and minima of the signal are identified, and the mean envelope is calculated. This process continues until an IMF satisfies the following two conditions:

1. The number of extrema and zero crossings must either be equal or differ by at most one.
2. The mean value of the envelope formed by local maxima and minima must be zero.

As an example, applying EMD to a crude oil price signal may yield the following: IMF1 captures high-frequency fluctuations, such as short-term market noise. IMF2 represents medium-frequency components, reflecting intermediate trends or cycles. IMF3 highlights low-frequency behavior corresponding to long-term price trends. The residual term accounts for any remaining nonperiodic behavior.

By breaking the signal into these distinct components, EMD facilitates a detailed analysis of different timescale features, enhancing forecasting accuracy and interpretability.

The mean envelope line $m_l(t)$ can be derived by computing the average of these envelopes. After that, the new data series $h_l(t)$ can be acquired by subtracting the original data series $X(t)$ from the mean envelope line $m_l(t)$:

$$X(t) - m_l(t) = h_l(t). \tag{5}$$

Typically, the signal remains nonstationary. Hence, we repeated the above procedure k times and ceased the process upon the convergence of the average envelope to zero to obtain the first component $C_l(t)$:

$$h_{1(k-1)}(t) - m_{1k}(t) = h_{1k}(t), \tag{6}$$

and

$$C_1(t) = h_{1k}(t). \tag{7}$$

The first IMF is the component $C_1(t)$ with the highest frequency. By subtracting the part $C_1(t)$ from the original data series $X(t)$, the data series $r_1(t)$ is obtained with the high frequency removed. After the above processing on $r_1(t)$, the second IMF component $C_2(t)$ can be obtained. This process is repeated until the difference series cannot be further decomposed, and the final difference series represents the trend component of the original data series:

$$r_i(t) = r_{i-1}(t) - C_i(t), \ i = 1, \ldots, n, \tag{8}$$

where n is the number of items in the dataset. Therefore, the original data series can be expressed as the sum of these IMFs and trend items:

$$X(t) = \sum_{j=1}^{n} C_j(t) + r_n(t). \tag{9}$$

The decomposition steps are summarized below:

Step 1: Initialization:

$$r_0(t) = X(t),$$

Step 2: Extract the ith IMF: First, initialize

$$h_0(t) = r_{i-1}(t), i = 1,$$

to obtain the extremum points of $h_{i-1}(t)$, perform cubic spline interpolation on the series of $h_{i-1}(t)$, gain the upper and lower envelopes of $h_{i-1}(t)$, and then define

$$h_i(t) = h_{i-1}(t) - m_i(t).$$

If $h_i(t)$ meets the standard of IMF, set

$$C_i(t) = h_i(t);$$

Otherwise, let $k = k + 1$ jump back to Step 1 to obtain the extremum points of $h_{k-1}(t)$.
Step 3: Define

$$r_i(t) = r_{i-1}(t) - C_i(t),$$

and if the number of extremum points of $r_i(t)$ is less than 2, then the decomposition ends, and the obtained $r_i(t)$ is the trend of the series $X(t)$; otherwise, set $i = i + 1$, and jump back to **Step 2**.

2.3. Transformer Model

The Transformer model is a well-known, recently developed machine learning technique with robust parallel processing capabilities and superior performance compared to conventional neural network models for fitting and forecasting real-world problems. The Transformer model consists of an attention mechanism and uses position coding, a self-attention mechanism, and a feedforward neural network to self-identify, self-adjust, and calculate the whole connection layer. First of all, in the absence of a temporal relationship of the input sample series, we need to use sine and cosine functions for positional encoding of the input, where each positional code only corresponds to a sine signal or cosine signal:

$$PE(P_{os}, 2i) = \sin\left(\frac{P_{os}}{10,000^{\frac{2i}{d}}}\right), \tag{10}$$

and

$$PE(P_{os}, 2i + 1) = \cos\left(\frac{P_{os}}{10,000^{\frac{2i}{d}}}\right), \tag{11}$$

where P_{os} is the position of elements in the sample for predicting crude oil price series; i is the dimension of elements in the sample for predicting crude oil price series; d is the dimension of the position vector in the model. Subsequently, the crude oil price series is input into the matrix and mapped to three different vector matrices, including the query matrix Q, the key matrix K, and the value matrix V through three different linear transformations, as shown in Formula (12). The output of the multihead attention mechanism is obtained by linear transformation and the application of the Softmax function, as depicted in Formula (13).

$$Q = XW^Q,$$
$$K = XW^K,$$
$$V = XW^V, \tag{12}$$

where X is the sample input matrix, and W^Q, W^K, and W^V are the corresponding linear transformation weight matrices, respectively.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{13}$$

where d_k is the gradient factor.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V),$$
$$MultiHead(Q, K, V) = Concat(head_1, head_2, K, head_h)w^0, \tag{14}$$

where W_i^Q, W_i^K, and W_i^V are the linear transformation weight matrices of the ith-head of Q, K, and V, respectively; w^0 is the linear transformation weight matrix after the multihead attention matrix is spliced; Concat is used to splice multiple results of multiple attention. Subsequently, the crude oil price forecasting samples, processed by multihead attention mechanisms, are fed into a fully connected feedforward neural network layer for computation, which is output as follows:

$$F_{FN}(Z) = \max(0, ZW_1 + b_1)w_2 + b_2, \tag{15}$$

where b_1 and b_2 represent offsets; ultimately, the two sublayers of multihead attention and feedforward neural network are coded and normalized by residual connection, and then we can obtain the predicted value of crude oil price as follows:

$$S_{out} = LN(x + S_{out}(x)),$$
$$LN(x_i) = \alpha \times \frac{x_i - u_L}{\sqrt{\alpha_L^2 + \varepsilon}} + \beta, \tag{16}$$

where LN is layer normalization; u_L and α_L^2 are the mean and variance, respectively; α and β represent gain and bias, respectively; ε is the adjustment parameter. Further details can be found in [35].

The attention mechanism is a core component of the Transformer model, enabling it to capture relevant patterns in time series data like crude oil prices. Specifically, the self-attention mechanism computes a weighted representation of the input sequence by evaluating the relationships between each time step. This process allows the model to identify and focus on the most influential time points, even when they are far apart in the sequence.

In the context of crude oil price forecasting, the attention mechanism operates as follows: 1. Temporal dependencies: The mechanism assigns higher weights to significant historical time points, such as previous price spikes or dips, that are most relevant for predicting the current price. For example, during periods of high volatility, the model focuses on recent rapid changes to forecast short-term fluctuations. 2. Trend recognition: The attention mechanism can recognize gradual trends, such as seasonal patterns or persistent upward/downward price movements, by examining long-term dependencies. 3. Noise reduction: The weighted representation helps the model filter out less relevant data, effectively reducing the influence of noise and improving prediction accuracy.

Mathematically, the attention mechanism computes a set of attention scores using the query (Q), key (K), and value (V) matrices derived from the input sequence. The attention scores are calculated as follows:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where

Q: represents the current time step's query vector,
K: encodes all other time steps as keys,
V: contains the values corresponding to each key, and
softmax: ensures the scores form a probability distribution, highlighting the most relevant time steps.

Through this process, the Transformer dynamically captures both short-term and long-term dependencies in crude oil price data, adapting its focus based on the data's

volatility and structure. This capability makes it particularly effective for handling crude oil price series' nonlinear and nonstationary characteristics".

2.4. VMD–EMD–Transformer Model

This subsection presents the combined model prediction construction method based on the VMD decomposition and the associated implementation steps. We employ a multistep approach integrating the VMD and EMD decompositions with a Transformer model to forecast the WIT and Brent crude oil prices. As shown in Figure 1, the proposed Transformer method based on VMD consists of the following four steps:
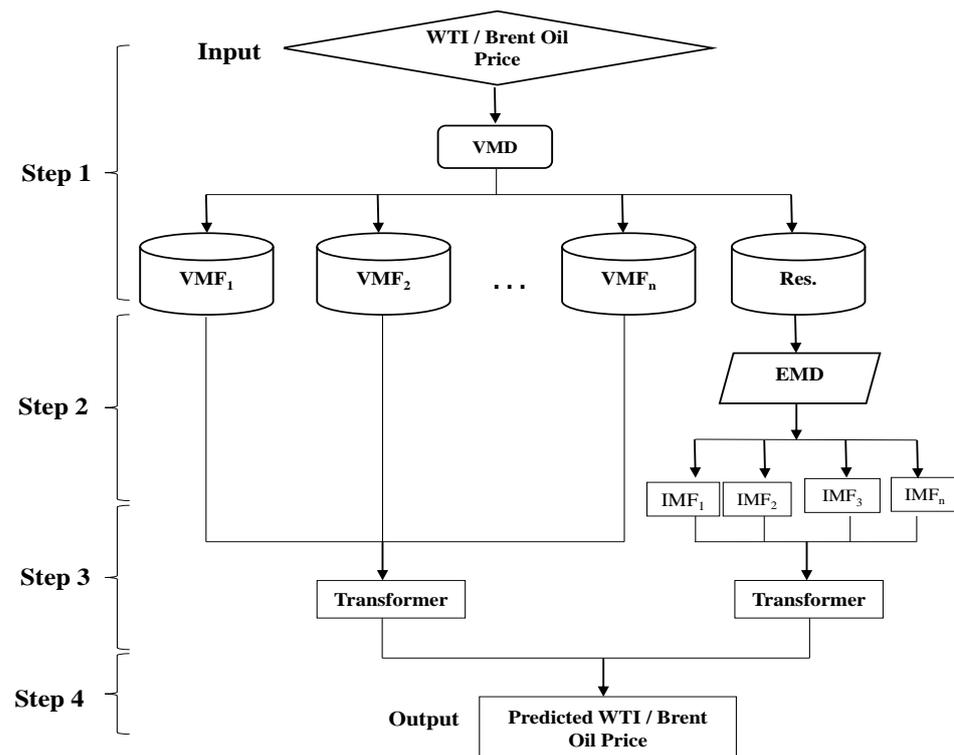
Figure 1. The proposed VMD–EMD–Transformer prediction approach.

Step 1: The VMD decomposition method is utilized to decompose the crude oil price series into mode components (VMFi), followed by the derivation of the residual series through the subtraction of VMFi from the original price series.

Step 2: Given the intricate and nonlinear characteristics of the residual series, while the EMD decomposition method can decompose the series into relatively stable components, we employ EMD to decompose the residual neglected in previous studies to extract an additional set of series (IMFi).

Step 3: We use the Transformer model to solve the prediction task of VMFi modal components after VMD decomposition and IMFi modal components after EMD decomposition residual.

Step 4: We integrate the forecast results. The forecast results of the IMFi are incorporated as the residual term predictions. Subsequently, the forecast results of the VMFi are combined with the residual term predictions to generate the ultimate forecasted futures prices. See Figure 1 for the specific process.

2.5. Assessment Criteria for Forecasting Performance

Referring to the existing literature [8,36,37], this paper uses three indicators to assess the forecasting performance of the model: the mean absolute error (MAE), the root mean square error (RMSE), and the mean absolute percentage error (MAPE), which are commonly

employed to ascertain the accuracy of the prediction model. Specifically, the lower the MAE, RMSE, and MAPE values, the higher the prediction accuracy. The MAE is defined as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i|, \tag{17}$$

where N, x_i, and $\hat{x}_i$ are the number of test sets, the actual, and the predicted values, respectively. The root mean square error (RMSE) shares the same unit with the predicted variable, which makes the error results more intuitive and easily interpretable. Moreover, RMSE is highly sensitive to the instances where predicted values deviate significantly from actual values, making it a widely adopted metric in prediction assessment. Compared with RMSE, MAPE is unaffected by dimensions, which is beneficial to the direct comparison between different models and is an important index to measure the prediction accuracy of models. Using the two indicators at the same time can evaluate the prediction accuracy of the model more comprehensively. The formulas to calculate RMSE and MAPE are as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2}, \tag{18}$$

and

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|x_i - \hat{x}_i|}{|x_i|}, \tag{19}$$

respectively.

3. Empirical Study

3.1. Data

In this study, the daily closing prices of West Texas Intermediate (WTI) crude oil futures and the Brent crude oil futures traded in the New York Mercantile Exchange (NYME) are used and the data are obtained from the US Energy Information Administration (EIA). The sample period extends from 4 January 2000 to 30 December 2022. The sample data are divided into the training set and the test set in the ratio of 7:3. That is, the data in the interval from 4 January 2000 to 31 December 2015 are used as the training set, and those in the interval from 4 January 2016 to 30 December 2022 as the test set. The WTI crude oil futures price series contains 5773 sample points, while the Brent crude oil futures price series contains 5769 sample points. The descriptive statistical results of the WTI and Brent crude oil price series are shown in Table 1. This paper utilizes the sequential validation method, as in the field of machine learning, cross-validation is commonly applied for assessing models based on static data. In contrast, sequential validation tends to be employed for time series data.

Table 1. Descriptive statistics for the oil futures prices.

Dataset		Statistic Description					
		Min.	Median	Mean	Max.	Std.	Length
WTI	Train dataset	17.50	62.31	63.85	145.31	28.57	4018
	Prediction dataset	−36.98	56.47	59.77	123.64	19.27	1755
Brent	Train dataset	16.51	61.65	65.86	143.95	32.77	4061
	Prediction dataset	9.12	62.57	63.79	133.18	20.63	1780

The historical price trend charts of crude oil are shown in Figure 2. It can be observed that the WTI crude oil prices are highly volatile. Additionally, the WTI crude oil prices had an unprecedented dip into negative values in 2020, and the overall pattern of price fluctuations for WTI crude oil remained indistinct. The Brent crude oil price remained positive but fluctuated violently, demonstrating a lack of discernable change patterns. In

addition, the skewness values of WTI and Brent are 0.33 and 0.36, respectively, and the kurtosis values are 0.71 and −0.9, respectively. From the statistical values of skewness and kurtosis, it is evident that the distribution of crude oil prices exhibits asymmetry.
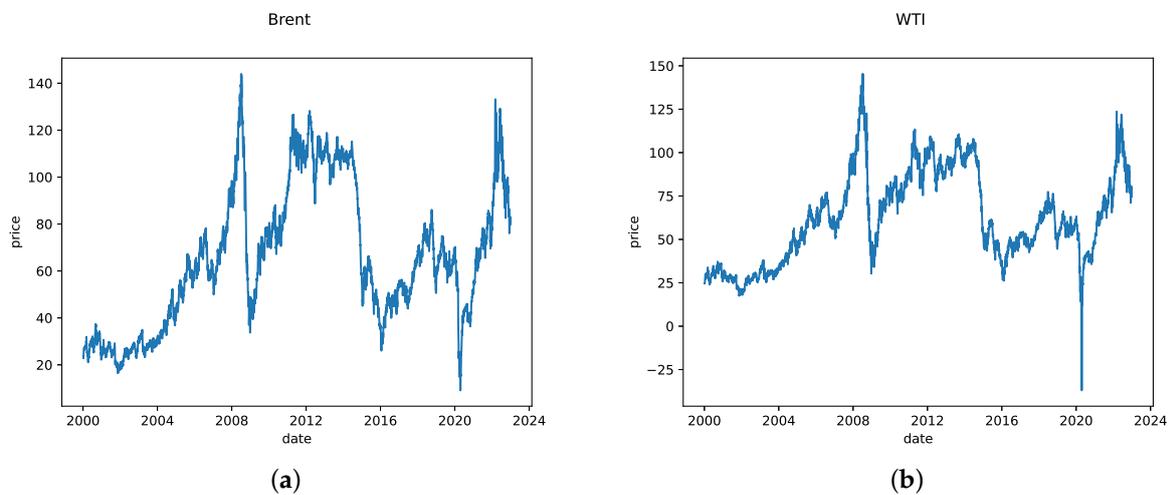

Figure 2. Brent and WTI crude oil historical prices.

Furthermore, the observed negative kurtosis value indicates that the distribution of crude oil prices shows a fat-tailed property relative to the standard normal distribution. Additionally, the positive skewness implies a right-biased tendency in the distribution of crude oil prices. Due to the presence of nonstationary and nonlinear characteristics in the crude oil futures price series, it is precise and reasonable to employ the VMD–EMD–Transformer model to forecast the prices of WTI crude oil and Brent crude oil.

The significant volatility observed in both the WTI and Brent crude oil price series in Figure 2 presents a major challenge for accurate forecasting. Traditional econometric models like ARCH and GARCH are widely used for volatility modeling by capturing linear dependencies in variance. However, their effectiveness declines due to nonlinear and high-frequency fluctuations, as seen in these series. The VMD–EMD–Transformer model addresses this issue through its dual decomposition strategy. First, the VMD method isolates dominant frequency components, filtering out noise while retaining meaningful patterns. Then, the EMD further refines residual components to capture subtle oscillations, making it well suited for nonlinear time series. These decomposed components, when input into the Transformer model, enable robust forecasting that effectively handles both short-term fluctuations and long-term trends.

3.2. Preprocessing Original Price Data

This paper uses VMD decomposition to process the original series according to the determined number of decomposition modes, and the residual series is obtained by subtracting the modal components after VMD decomposition from the original series. According to the mode components VMF0-VMF2 listed in Table 2 and the statistical information of the residual, we can find that the Pearson correlation value and Kendall correlation value of the residual series are higher than those of VMF2, which further proves that the residual contains pertinent information regarding the original signal. The EMD is utilized in this paper to prevent information loss and forecast the previously disregarded residual series due to its intricacy. The mode components of the post-EMD residual series are illustrated in Figures 3 and 4. The decomposition results reveal that after EMD, the residual series is decomposed into three IMF signals (IMF1–IMF3). The IMF mode components are arranged from low to high frequency, with IMF1 exhibiting the lowest fluctuation frequency and IMF3 displaying the highest. After the secondary decomposition, the residual mode components show lower nonstationarity and complexity; thus, utilizing these modal components for forecasting crude oil prices can result in higher accuracy.
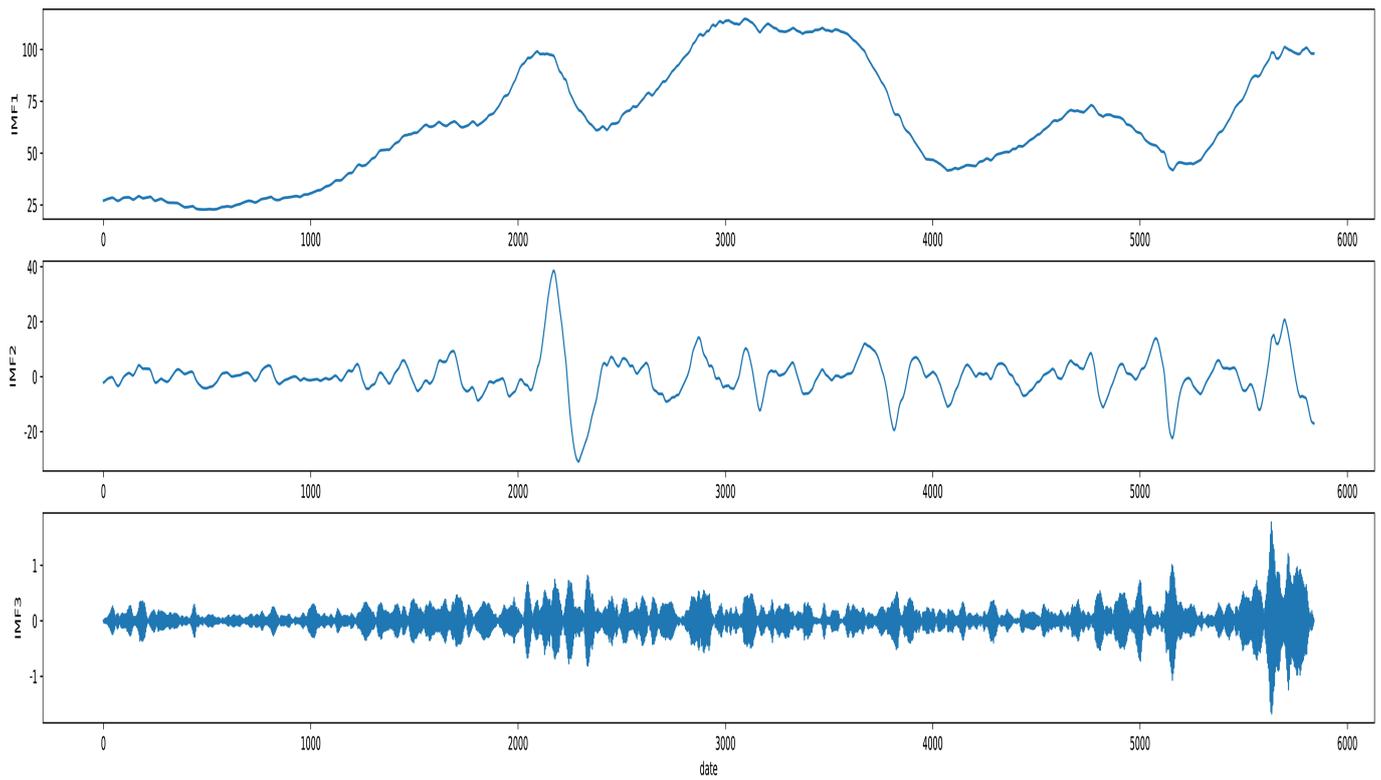
Figure 3. Intrinsic mode functions (IMFs) derived from the Brent crude oil price residue using EMD, highlighting different frequency components.
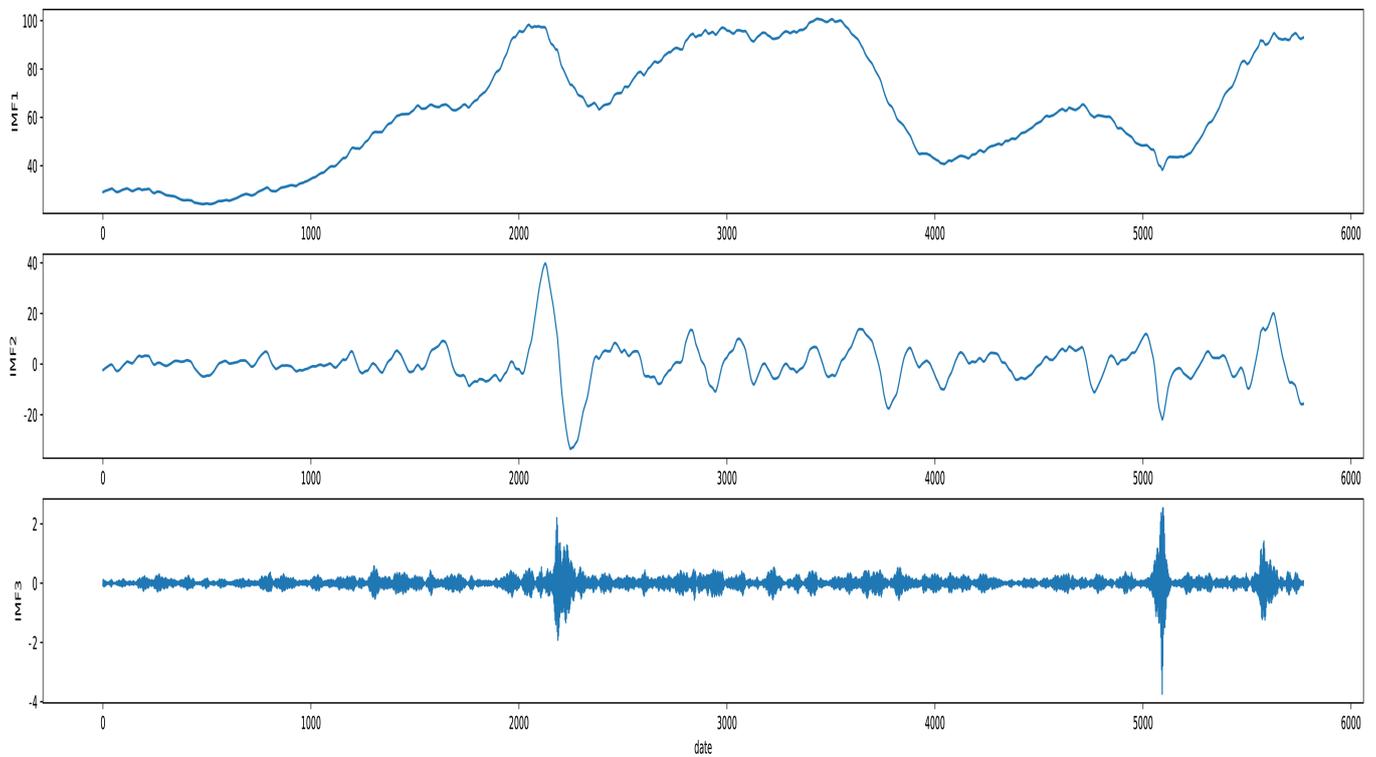
Figure 4. Intrinsic mode functions (IMFs) derived from the WTI crude oil price residue using EMD, illustrating low- and high-frequency trends.

Table 2. Measures of the VMFs and the residue of the WTI and Brent oil prices.

Observed	Pearson Correlation	Kendall Correlation	Spearman
WTI crude oil price VMD decomposition			
WTI_VMF0	0.918	0.781	0.931
WTI_VMF1	0.521	0.319	0.411
WTI_VMF2	0.033	0.013	0.017
WTI_Residue	0.215	0.103	0.141
Brent crude price VMD decomposition			
Brent_VMF0	0.925	0.801	0.948
Brent_VMF1	0.505	0.284	0.365
Brent_VMF2	0.030	0.016	0.023
Brent_Residue	0.217	0.115	0.159

3.3. Forecasting Results

Our test results verify the superiority of the VMD–EMD–Transformer model over the other three models—LSTM, Transformer, and VMD–Transformer models. Specifically, this paper employs the Transformer model to predict the decomposed modal components. It accumulates the predicted values of each modal component through VMD and EMD to obtain the forecast of crude oil prices. Table 3 shows that in the prediction of WTI crude oil price, the values of MAE, MSE, and MAPE for the VMD–EMD–Transformer model are 2.88, 0.092, and 5.602, respectively, while those for the same model in predicting Brent crude oil price are 2.813, 0.088, and 4.946, respectively. In both cases, the VMD–EMD–Transformer model performs much better than the other three models.

Table 3. Performance of VMD–EMD–Transformer across forecasting horizons.

Horizon	MAE	RMSE	MAPE
Short-term (1 week)	2.25	0.085	5.12%
Long-term (6 months)	3.45	0.120	7.84%

To further analyze the model's robustness across different forecasting horizons, we evaluate its performance for both short-term (1 week) and long-term (6 month) predictions. Table 3 summarizes the error metrics for these intervals.

As shown in Table 3, the model achieves lower errors for short-term predictions due to the relatively stable patterns in shorter horizons. However, the accuracy slightly decreases for long-term forecasts as the volatility in crude oil prices accumulates over extended periods. Despite this, the VMD–EMD–Transformer model consistently outperforms traditional econometric models (e.g., ARCH and GARCH) and other machine learning models (e.g., LSTM) across both horizons, demonstrating its robustness in diverse forecasting scenarios.

To evaluate the effectiveness of the proposed VMD–EMD–Transformer model, we compare its performance with three commonly used forecasting methods: LSTM, Transformer, and VMD–Transformer. The results are listed in Table 4. These methods are briefly introduced below:

1. LSTM (long short-term memory): LSTM is a type of recurrent neural network (RNN) designed to handle sequential data by capturing both short-term and long-term dependencies. It utilizes gated units to control the flow of information, making it effective for time series forecasting tasks. However, LSTM may struggle with capturing complex patterns in nonlinear and nonstationary data, like crude oil prices.

2. Transformer: The Transformer model is a deep learning architecture based on self-attention mechanisms. Unlike RNNs, it processes the entire input sequence simultaneously, enabling it to capture long-range dependencies efficiently. Transformers are particularly effective in handling large-scale time series data but may require additional preprocessing to manage noise in volatile data.

3.   VMD–Transformer: This hybrid approach combines variational mode decomposition (VMD) with a Transformer model. VMD decomposes the original time series into several modes, isolating meaningful frequency components. The Transformer then predicts each mode separately, with the results aggregated to form the final forecast. While this method improves performance by addressing data complexity, it does not account for residual patterns in the decomposed signal.

By comparing these methods, we aim to highlight the advantages of the VMD–EMD–Transformer model in capturing both modal components and residual patterns, thereby enhancing forecasting accuracy for crude oil price data.

Table 4. Performance comparison of four different models, where the numbers with * are the smallest of "errors" among the forecasting methods.

Model	MAE	MSE	MAPE
WTI Crude Oil price forecasting			
VMD–EMD–Transformer	**2.880 ***	**0.092 ***	**5.602 ***
VMD–Transformer	9.580	0.275	18.027
Transformer	10.394	0.331	21.189
LSTM	13.161	0.423	28.501
Brent Crude Oil price forecasting			
VMD–EMD–Transformer	**2.813 ***	**0.088 ***	**4.946 ***
VMD–Transformer	3.525	0.101	6.431
Transformer	6.815	0.183	12.030
LSTM	8.197	0.264	18.063

Figures 5 and 6 display curves of the predicted and actual prices for the four models on the test set. The closer the curve of predicted prices by a model to the curve of the actual prices, the better the model; so, from these figures, one can observe that the VMD–EMD–Transformer model exhibits the highest prediction accuracy among the four models.
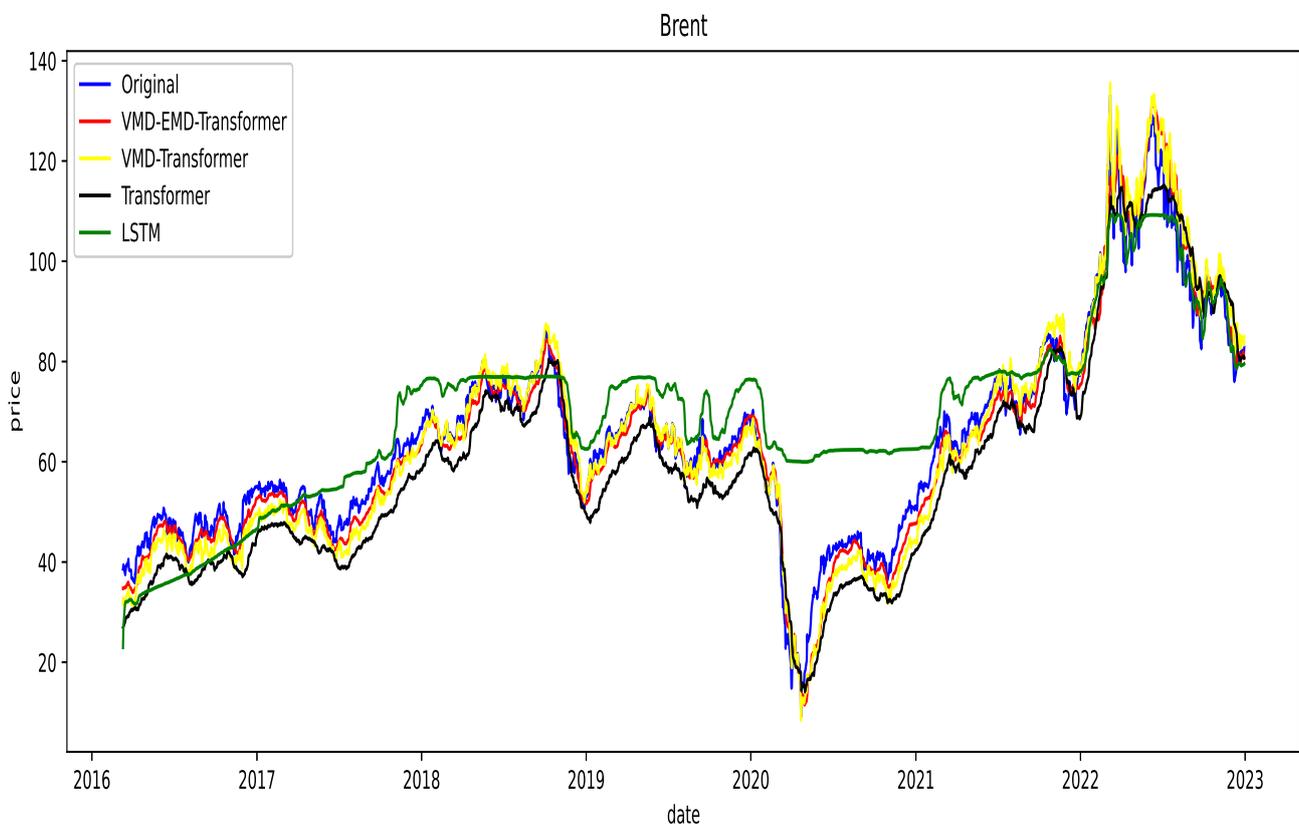
Figure 5. Comparison of actual and predicted Brent crude oil prices using the VMD–EMD–Transformer and baseline models.
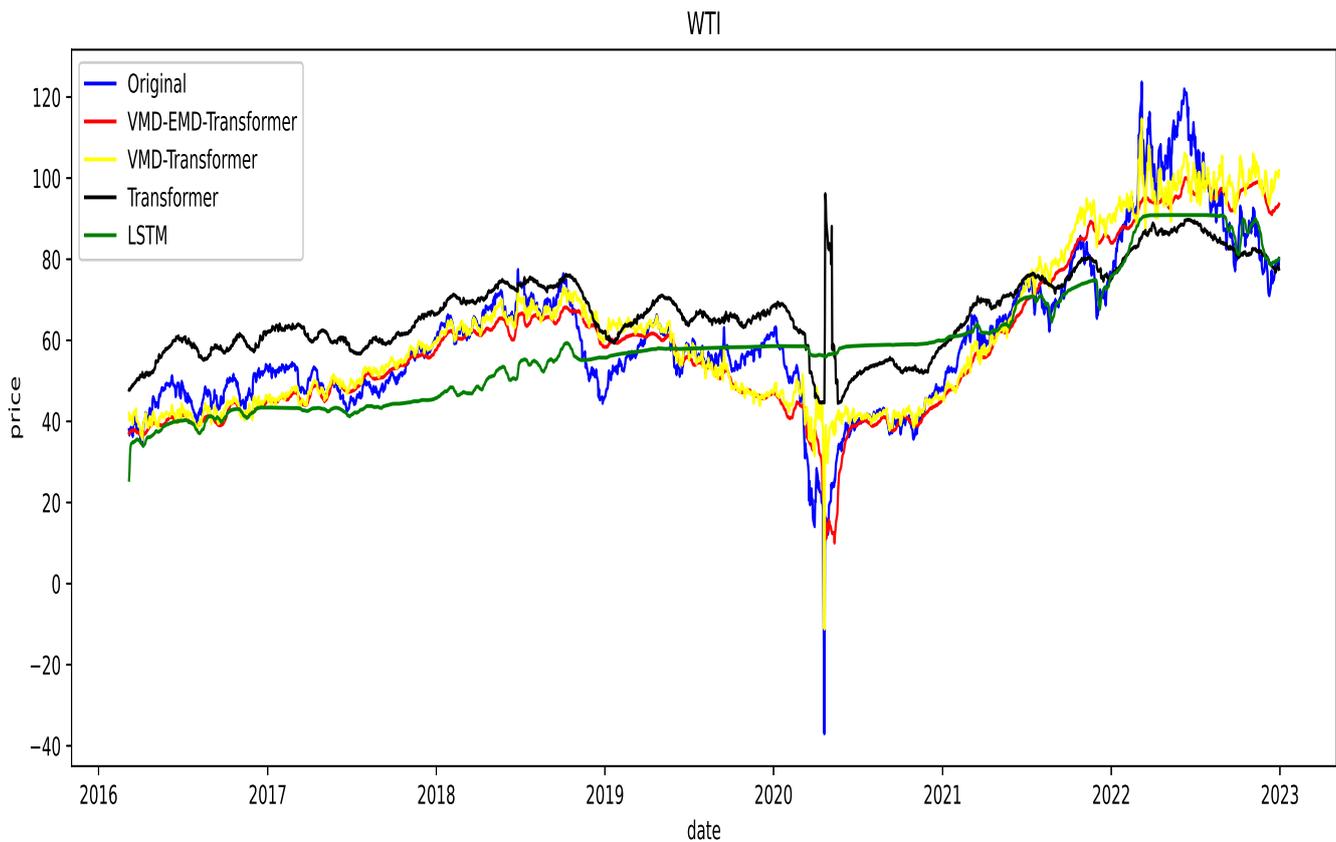
Figure 6. Comparison of actual and predicted WTI crude oil prices using the VMD–EMD–Transformer and baseline models.

3.4. Comparison with Traditional Econometric Models

In practical forecasting applications, parsimony is often preferred, with simpler models like ARIMA and GARCH widely utilized due to their ease of estimation and interpretability. To assess the added value of the VMD–EMD–Transformer model, we conducted a quantitative comparison with ARIMA and GARCH models using the same dataset for crude oil price forecasting. The results are summarized in Table 5.

As shown in Table 5, while ARIMA and GARCH models perform adequately for linear or low-volatility scenarios, their accuracy significantly declines when applied to the nonlinear, high-frequency fluctuations of crude oil prices. In contrast, the VMD–EMD–Transformer model substantially reduces errors across all metrics, particularly in RMSE, indicating superior handling of complex, nonlinear patterns. This improvement justifies the added complexity of the VMD–EMD–Transformer model, especially in scenarios where high forecasting accuracy is critical, such as in financial investment or risk management.

Table 5. Performance comparison of forecasting models, where the numbers with * are the smallest of "errors" among the forecasting methods.

Model	MAE	RMSE	MAPE
ARIMA	8.12	2.67	18.45%
GARCH	7.98	2.54	17.80%
VMD–EMD–Transformer	**2.88 ***	**0.092 ***	**5.60% ***

4. Conclusions

The proposed VMD–EMD–Transformer model in this paper amalgamates the strengths of VMD, EMD, and Transformer models known for their superior generalization ability,

resulting in a significant enhancement in the forecasting accuracy of the model. Given the nonstationary and nonlinear traits of crude oil futures' daily closing price series, we utilized the VMD method to decompose the original price sequence into multimode components. Furthermore, we decomposed the residual sequence through the EMD method to generate a more refined line, thus augmenting the forecasting accuracy of the Transformer model. Ultimately, we obtained the ultimate forecasted price series by accumulating the predicted values of each mode component post-decomposition. This paper addressed the problem of information loss in traditional time series models that employ VMD decomposition due to the neglect of residuals. Consequently, this study proposed a combined model that fully captures the advantages of models. An empirical evaluation was conducted on the daily closing prices of crude oil futures to test the proposed forecasting model's effectiveness, demonstrating remarkable forecasting accuracy. The significant volatility in crude oil prices poses challenges for traditional econometric models like ARCH and GARCH, which are limited by their linear assumptions. The proposed VMD–EMD–Transformer model effectively addresses these challenges by leveraging dual decomposition techniques to capture nonlinear, high-frequency fluctuations and employing a Transformer model to identify temporal dependencies. Quantitative comparisons show that the VMD–EMD–Transformer significantly outperforms ARCH and GARCH in forecasting accuracy, highlighting its potential for real-world financial applications where precision is critical.

References

1. Wang, P.; Lü, Y.J. Extreme risk measurement models of the international oil market and backtesting analysis. *Financ. Res.* **2018**, *9*, 192–206.
2. Sun, J.; Zhao, P.; Sun, S. A new secondary decomposition-reconstruction-ensemble approach for crude oil price forecasting. *Resour. Policy* **2022**, *77*, 102762. [CrossRef]
3. Moya-Martínez, P.; Ferrer-Lapeña, R.; Escribano-Sotos, F. Oil price risk in the Spanish stock market: An industry perspective. *Econ. Model.* **2014**, *37*, 280–290. [CrossRef]
4. El Anshasy, A.A.; Bradley, M.D. Oil prices and the fiscal policy response in oil-exporting countries. *J. Policy Model.* **2012**, *34*, 605–620. [CrossRef]
5. Sévi, B. Forecasting the volatility of crude oil futures using intraday data. *Eur. J. Oper. Res.* **2014**, *235*, 643–659. [CrossRef]
6. Shi, S.; Liu, W.; Jin, M. Stock price forecasting using a hybrid ARMA, BP neural network, and Markov model. In Proceedings of the 2012 IEEE 14th International Conference on Communication Technology, Chengdu, China, 9–11 November 2012; pp. 981–985.
7. Hossain, Z.; Abdus Samad, Q.; Ali, Z. ARIMA model and forecasting with three types of pulse prices in Bangladesh: A case study. *Int. J. Soc. Econ.* **2006**, *33*, 344–353. [CrossRef]
8. Ji, L.; Zou, Y.; He, K.; Zhu, B. Carbon futures price forecasting based on the ARIMA-CNN-LSTM model. *Procedia Comput. Sci.* **2019**, *162*, 33–38. [CrossRef]
9. Wang, L.; Zhang, Z. Research on Shanghai copper futures price forecast based on X12-ARIMA-GARCH family models. In Proceedings of the 2020 International Conference on Computer Information and Big Data Applications (CIBDA), Guiyang, China, 17–19 April 2020; pp. 304–308.

10. Lin, K.P.; Pai, P.F.; Yang, S.L. Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. *Appl. Math. Comput.* **2011**, *217*, 5318–5327. [CrossRef]

11. Ghoddusi, H.; Creamer, G.G.; Rafizadeh, N. Machine learning in energy economics and finance: A review. *Energy Econ.* **2019**, *81*, 709–727. [CrossRef]

12. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]

13. Shin, H.; Hou, T.; Park, K.; Park, C.K.; Choi, S. Prediction of movement direction in crude oil prices based on semi-supervised learning. *Decis. Support Syst.* **2013**, *55*, 348–358. [CrossRef]

14. Yu, L.; Xu, H.; Tang, L. LSSVR ensemble learning with uncertain parameters for crude oil price forecasting. *Appl. Soft Comput.* **2017**, *56*, 692–701. [CrossRef]

15. Chen, H.H.; Chen, M.; Chiu, C.C. Integrating artificial neural networks and text mining to forecast gold futures prices. *Commun. Stat. Simul. Comput.* **2016**, *45*, 1213–1225. [CrossRef]

16. Wang, J.; Li, X. A combined neural network model for commodity price forecasting with SSA. *Soft Comput.* **2018**, *22*, 5323–5333. [CrossRef]

17. Kohzadi, N.; Boyd, M.S.; Kermanshahi, B.; Kaastra, I. A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing* **1996**, *10*, 169–181. [CrossRef]

18. Wang, Y.; Zhang, L.; Liu, Y.; Guo, J. Gold price prediction method based on improved PSO-BP. *Int. J. u-e-Serv. Sci. Technol.* **2015**, *8*, 253–260. [CrossRef]

19. Yu, L.; Wang, S.; Lai, K.K. We forecast crude oil prices with an EMD-based neural network ensemble learning paradigm. *Energy Econ.* **2008**, *30*, 2623–2635. [CrossRef]

20. Yu, L.; Wang, Z.; Tang, L. A decomposition–ensemble model with data-characteristic-driven reconstruction for crude oil price forecasting. *Appl. Energy* **2015**, *156*, 251–267. [CrossRef]

21. Lin, C.S.; Chiu, S.H.; Lin, T.Y. Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Econ. Model.* **2012**, *29*, 2583–2590. [CrossRef]

22. Liu, H.; Yin, S.; Chen, C.; Duan, Z. Data multi-scale decomposition strategies for air pollution forecasting: A comprehensive review. *J. Clean. Prod.* **2020**, *277*, 124023. [CrossRef]

23. Tian, Z.; Gai, M. New PM2.5 forecasting system based on combined neural network and an improved multi-objective optimization algorithm: Taking the economic belt surrounding the Bohai Sea as an example. *J. Clean. Prod.* **2022**, *375*, 134048. [CrossRef]

24. Zhu, B.; Ye, S.; Wang, P.; He, K.; Zhang, T.; Wei, Y.M. A novel multiscale nonlinear ensemble learning paradigm for carbon price forecasting. *Energy Econ.* **2018**, *70*, 143–157. [CrossRef]

25. Zhang, X.; Lai, K.K.; Wang, S.Y. A new approach for crude oil price analysis based on empirical mode decomposition. *Energy Econ.* **2008**, *30*, 905–918. [CrossRef]

26. Azevedo, V.G.; Campos, L.M. Combination of forecasts for the price of crude oil on the spot market. *Int. J. Prod. Res.* **2016**, *54*, 5219–5235. [CrossRef]

27. Safari, A.; Davallou, M. Oil price forecasting using a hybrid model. *Energy* **2018**, *148*, 49–58. [CrossRef]

28. Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2013**, *62*, 531–544. [CrossRef]

29. Muhammad, T.; Aftab, A.B.; Ahsan, M.; Muhu, M.M.; Ibrahim, M.; Khan, S.I.; Alam, M.S. Transformer-Based Deep Learning Model for Stock Price Prediction: A Case Study on Bangladesh Stock Market. *arXiv* **2022**, arXiv:2208.08300. [CrossRef]

30. Diligenti, M.; Roychowdhury, S.; Gori, M. Integrating prior knowledge into deep learning. In Proceedings of the 2017 the 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 920–923.

31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–15.

32. Yang, L.; Ng, T.L.J.; Smyth, B.; Dong, R. Html: Hierarchical Transformer-based multi-task learning for volatility prediction. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 441–451.

33. Li, C.; Qian, G. Stock Price Prediction Using a Frequency Decomposition Based GRU Transformer Neural Network. *Appl. Sci.* **2022**, *13*, 222. [CrossRef]

34. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond.* **1998**, *454*, 903–995. [CrossRef]

35. Jakaša, T.; Andročec, I.; Sprčić, P. Electricity price forecasting—ARIMA model approach. In Proceedings of the 2011 8th International Conference on the European Energy Market (EEM), Zagreb, Croatia, 25–27 May 2011; pp. 222–225.

36. Guha, B.; Bandyopadhyay, G. Gold price forecasting using the ARIMA model. *J. Adv. Manag. Sci.* **2016**, *4*, 117–121.

37. Mozetič, I.; Torgo, L.; Cerqueira, V.; Smailović, J. How to evaluate sentiment classifiers for Twitter time-ordered data? *PLoS ONE* **2018**, *13*, e0194317. [CrossRef] [PubMed]