Computer Life
ISSN: 1819-4818 | Vol. 11, No. 1, 2023

Research on Crude Oil Price Prediction Based on Improved Long Short-Term Memory Neural Network

Jinliang Li *, Jianshe Dong, Chengxu Liang

Tianjin University of Technology and Education School of Information Technology and Engineering

* **Corresponding author**: JINLIANG LI (Email: 1530425125@qq.com)

Abstract: The prediction of financial time series has always been challenging. This study aims to improve the accuracy and robustness of crude oil price prediction by employing various analysis methods and models. Firstly, we introduce CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) and Hilbert transformation, using these methods for multiscale decomposition of crude oil price time series. The decomposed components are then forecasted using LSTM and LSTM models with attention mechanism, demonstrating superiority in multiscale features. In the reconstruction phase, we employ an intelligent reconstruction method, achieving more accurate restoration of crude oil price fluctuations compared to simple summation reconstruction. Overall, the composite model constructed in this study exhibits superior predictive performance across multiple metrics such as MSE, RMSE, MAE, and R^2. Compared to other traditional methods and single models, our model demonstrates stronger adaptability and predictive accuracy in complex market environments. This study introduces the ideas of multiscale features and intelligent reconstruction to the field of crude oil price prediction, providing a new approach to improving prediction accuracy. In the current uncertain and volatile crude oil market, our model not only has better fitting capabilities but also responds more flexibly to complex market scenarios, offering a reliable reference for relevant decision-making.

Keywords: Crude Oil Price Prediction, Long Short-Term Memory Neural Network, CEEMDAN, Hilbert Transformation, Multiscale Fusion.

1. Introduction

Crude oil is one of the crucial energy resources in the world economy, and its price fluctuations have widespread and profound effects on global supply chains, energy policies, investment decisions, and economic stability. As the crude oil market continues to evolve and the complexity of the global economy increases, accurate predictions of future crude oil prices become critically important. Traditional economic models, while to some extent capable of explaining price changes, often struggle to cope with the diversity and uncertainty of the crude oil market.

The rapid development of futures contract trading markets has closely linked energy prices with various sectors.[1] In this context, data-driven methods, such as artificial neural networks, have emerged. Neural networks, as powerful machine learning tools, have the potential to handle large-scale time series data and capture nonlinear relationships, offering new possibilities for crude oil price prediction. This approach not only leverages historical price data to forecast future price trends but also automatically adapts to different market conditions, thereby improving prediction accuracy.

The primary goal of this study is to explore and enhance the application of neural networks in crude oil price prediction. We will use historical crude oil price data and relevant factors to train deep learning models for predicting future price movements. Through this method, we aim to provide market participants, government decision-makers, and financial investors with better price prediction tools to help them better understand the dynamics of the crude oil market and mitigate potential risks.

Although neural networks hold potential in crude oil price prediction, they also face challenges, including the reliability of data sources, the quality of data processing, model complexity, and issues related to overfitting. Therefore, this study will delve into these challenges and propose methods to overcome them.

2. Related Work

Crude oil price prediction has been a subject of widespread attention due to its profound impact on the world economy and energy markets. In this field, extensive research has been conducted to develop various models and methods aimed at improving the accuracy of predictions regarding the trends in crude oil prices. Early studies in crude oil price prediction heavily relied on traditional economic models and time series analysis.

Murat and Tokat used a Random Walk (RW) model as a benchmark to forecast the price trends in the oil market and assess predictive performance.[2] Morana (2001) proposed a semi-parametric approach based on the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model to predict short-term prices of Brent crude oil.[3] Baumeister and Kilian (2012) employed a Vector Autoregression (VAR) model for short-term oil price prediction, with empirical results suggesting that VAR models outperform Autoregression (AR) and Autoregressive Moving Average (ARMA) models in directional forecasting accuracy.[4] Hou Lu (2009) based on the spot prices of Brent crude, established an ARIMA forecasting model, and analyzed the new situation and dynamic changes in the international and oil industries in 2009.[5] Zhao Sha (2013) constructed an Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) model, a Seasonal-Harmonic forecasting model, and an Exponential Smoothing model using time series analysis methods, applying these three models to international oil price prediction. Through a comparison of prediction performance, the Seasonal-Harmonic forecasting model demonstrated

relatively good results.[6]

These methods typically rely on fundamental economic indicators such as supply-demand relationships, production data, geopolitical factors, etc., to infer price trends. However, they face limitations when dealing with nonlinear relationships and the multifactorial influences on prices. For instance, they struggle to account for factors like market psychology, sentiment, and risk preferences. Figure 1 shows a comparative chart of WTI and Brent crude oil prices over the past five years (data sourced from www.investing.com).

Fig 1. Crude oil prices in recent years

With the rise of machine learning, researchers have begun exploring statistical and machine learning-based methods for crude oil price prediction. These methods include linear regression, support vector machines, random forests, and more. They have the capability to learn patterns from historical data, identify features and trends, thereby improving the accuracy of predictions. However, these methods still face challenges related to feature engineering and managing model complexity.

In recent years, deep learning methods, especially neural networks, have become a focal point in crude oil price prediction. Deep learning models possess the ability to handle large-scale data and capture nonlinear relationships, making them perform exceptionally well in crude oil price prediction. Researchers have explored architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to better capture patterns in price time series. Mirmirani and Li (2004) used a genetic algorithm-based artificial neural network model for oil price prediction, demonstrating its significant superiority over Vector Autoregression (VAR) models.[7] Urolagin et al. (2021) employed a Multivariate Long Short-Term Memory (LSTM) model for oil price prediction.[8] Tang et al. (2018) utilized Random Vector Functional Link (RVFL) for WTI oil price prediction, showing that RVFL without an iterative training process had shorter computation time and higher prediction accuracy.[9] These deep learning methods offer higher flexibility and accuracy, aiming to enhance the quality of crude oil price predictions.

Deep learning methods have made significant progress in crude oil price prediction. Neural network architectures like Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM) have been applied. These models can handle time series data, capturing the seasonality and cyclicality features of prices and automatically adjusting model weights, thereby improving the accuracy of long-term and short-term price trend predictions. Additionally, Convolutional Neural Networks (CNN) have been used to process images and geospatial information data related to crude oil prices, providing a more comprehensive understanding of the driving factors behind price fluctuations.

Crude oil price fluctuations are often influenced by the emotions and sentiments of market participants. In recent years, researchers have started exploring the application of sentiment analysis and social media data in crude oil price prediction. Analyzing news, social media posts, and sentiment index data can provide a better understanding of market participants' emotions and predict market reactions. Skuza and Romanowski (2015) utilized machine learning for sentiment analysis on a large volume of text data from Twitter, estimating future stock prices, proving the correlation between sentiment and stock markets, and confirming the effectiveness of the chosen method.[10] Smales and Lee (2016) studied the relationship between news sentiment, S&P 500 index returns, and changes in the Volatility Index (VIX). The results showed a significant negative correlation between volatility index changes and news sentiment as well as stock returns.[11] Chen Xiaohong et al. (2016) applied sentiment analysis techniques to Sina Weibo texts, constructing a sentiment index, examining the predictive ability of investor sentiment for the stock market. The results showed that the Weibo sentiment index had a significant predictive effect on stock prices in the short term.[12] This provides an opportunity for more comprehensive market sentiment modeling.

Deep Reinforcement Learning (DRL) has been used to address decision-making problems in crude oil price prediction. This method combines deep learning and reinforcement learning to formulate decision strategies, such as trading decisions and risk management.[13] By training agents to learn the best action strategy, DRL methods have the potential to improve the quality of crude oil price predictions in high-frequency trading and dynamic market environments.

The interpretability of deep learning models has always been a challenge as they are often seen as black-box models. Researchers have started focusing on the interpretability of deep learning models to better understand the decision-making process. In Adadi et al.'s study, they advocated for the interdisciplinary nature of the research field and introduced the main aspects and application areas of interpretability from different perspectives.[14] Zhang et al. primarily summarized the interpretability methods for neural networks and classified interpretability methods from three dimensions: involvement type, interpretation type, and interpretation scope.[15] Mohseni et al. proposed an evaluation benchmark based on images and texts for interpretability, which quantitatively evaluated the effectiveness of the method in quantitatively evaluating model interpretation.[16] Simultaneously, uncertainty analysis has become important as it helps quantify the uncertainty range of predictions, providing more information for risk management.

In conclusion, the paper provides a comprehensive review of the literature in the field of financial time series prediction. Previous research has achieved a series of significant results in revealing the driving factors of crude oil price volatility and establishing effective prediction models. However, behind these achievements, we still face some unresolved issues, such as not considering factors such as seasonality, cyclicity, and natural disasters. The rest of this paper will detail the research methods, data analysis, model training process, comparisons, research results, and discussions. Overall

3. Theoretical Methods

3.1. LSTM and GRU models

3.1.1. Basic model architecture

We chose long short-term memory networks (LSTM) as our base model for this experiment because they perform well in time series forecasting. LSTM is designed to handle long-term dependencies, which often occur in time series data. They effectively capture long-term dependencies in sequences by using gating mechanisms to decide when to retain, forget, or read information. It can adapt to different patterns, trends and periodicities, and is a recurrent neural network structure suitable for time series data. Its structure is shown in Figure 2:

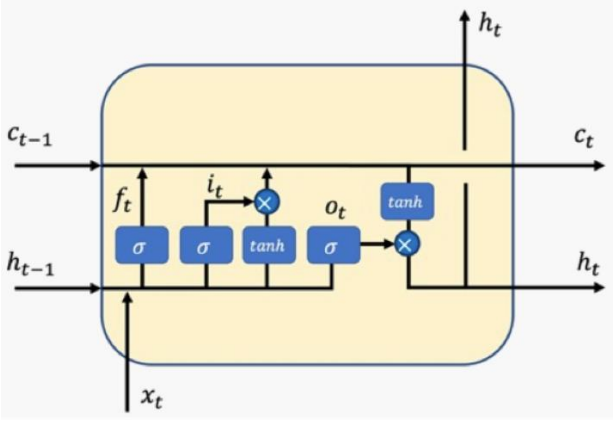

Fig 2. LSTM neural network

Figure 2 shows the internal structure of the LSTM neural network, which includes: State vector ct: It controls the state or memory of the entire LSTM unit. It is updated according to the input at each moment, thereby maintaining the memory of the LSTM unit in real time. Hidden state vector ht: It is the external output state of the current LSTM unit. It is the actual working state vector, that is, ht is generally used to do some specific tasks.

Input gate it: controls what information needs to be injected into the state vector ct based on the input information at the current moment. For example, when the input information is words that have no actual meaning, such as "的", the model may not let this information flow into the state vector, thereby maintaining the semantic expression of the model.

Forgetting gate ft: controls what information the state vector ct−1 at the previous moment needs to be masked/forgotten. For example, I climbed the Great Wall yesterday. Oh, no, it was the day before yesterday. When the model sees "No, it was the day before yesterday", it may forget "yesterday" in front of it.

Output gate ot: controls what information needs to be output by the state vector ct at the current moment. The final output information is ht.

After understanding these basic concepts, let's take a look at the specific generation process of these components. First, let's look at the generation process of these three gates, taking time t as an example:

$$i_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$

$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

It can be seen that the calculation of these three gates is to perform linear transformation using the input data xt and ht−1, and then pass the result to the sigmoid function, because the sigmoid function is a function in the value range (0,1), that is, it can transform the data Map to this fixed interval to control the flow of information.

3.1.2. input and output

We use at to represent the information to be entered.

$$a_t = \tanh(W_a x_t + U_a h_{t-1} + b_a) \quad (4)$$

The input data xt and ht−1 are also linearly transformed, and then the results are passed to the tanh function. The final result is the information to be injected into the state vector ct of the current LSTM unit. With the above components, the state vector ct of the current LSTM unit can be updated.

$$c_t = f_t \times c_{t-1} + i_t \times a_t \quad (5)$$

Obviously, the update of the LSTM unit state ct is to selectively forget the state ct−1 at the previous moment, selectively input the information at to be input at the current moment, and finally add the results of the two to indicate the direction of the input. The current LSTM unit incorporates the previous state information ct−1 and at the same time injects the latest information at. After calculating the state vector ct at the current moment, it can be output based on the state vector. That is, the current status information ct is selectively output through the output gate.

$$h_t = o_t \times \tanh(c_t) \quad (6)$$

In this article, we can process transformed oil price time series data as input, and possibly other features, etc. Try to introduce seasonal and cyclical features. The output is the crude oil price at the next point in time.

3.1.3. Training and Optimization

We train the model by minimizing the error between the predicted price and the actual price. We used loss functions and optimization algorithms to ensure that the model effectively converged on the training set.

3.2. Adaptive noise complete set empirical mode decomposition

CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) is an improvement on the traditional empirical mode decomposition (EMD) method, aiming to overcome its existing problems. In 1998, Huang Jian proposed empirical mode decomposition (EMD) as a data-based signal decomposition method, which decomposes the signal into intrinsic vibration mode functions (IMF). Although EMD has achieved certain success, there are also problems such as mode confusion, noise sensitivity and mode overlap. In 2009, Norden E. Huang and others proposed CEEMDAN, which introduced an adaptive noise adjustment mechanism and dynamically adapted to the local characteristics of the signal by constructing multiple noise auxiliary data sets to improve the accuracy and robustness of the decomposition.

Adaptive noise mechanism: CEEMDAN's adaptive noise mechanism is one of its innovations. By adding different implementations of Gaussian white noise to the signal, and adaptively adjusting the intensity of the noise through iteration, the decomposed IMF can better reflect the essence

of the signal. .

Over time, CEEMDAN has been improved and expanded, and is widely used in signal processing, geophysics, finance and other fields. Its flexibility and robustness make it a powerful tool for signal decomposition.

Overall, the introduction of CEEMDAN is of great significance in overcoming the limitations of EMD, and its adaptive noise mechanism makes it perform well in processing various signals.

3.3. Fusion strategy

3.3.1. Model hyperparameters

We selected appropriate hyperparameters, such as learning rate, hidden layer dimensions, time steps, etc., to ensure that each model can effectively converge during training.

3.3.2. Fusion weight strategy

We adjusted the weights during the model fusion process to ensure that each model's contribution to the final output was balanced. In order to comprehensively utilize the advantages of different models, we adopt a simple weighted average fusion strategy. First, we calculate the performance metrics of each model on the cross-validation set, including MSE, RMSE and MAE. Then, the weight of each model is determined through methods such as grid search, so that the weighted average performance on the verification set is optimal. Finally, we linearly weight the prediction results of each model according to the weight to obtain the final prediction result.

Through this fusion strategy, we expect to give full play to the strengths of each model and improve the accuracy and robustness of the overall prediction. Experimental results show that compared with a single model, the weighted average fusion method has achieved significant advantages in reducing the risk of overfitting and improving robustness.

3.4. Hilbert transform

HHT is a time-frequency analysis method based on local eigendecomposition, which includes empirical mode decomposition (EMD) and Hilbert spectrum analysis. EMD is a part of HHT, similar to CEEMD, which decomposes the signal into several intrinsic mode functions (IMFs).

3.4.1. EEMD

EMD is a data-driven decomposition method that decomposes the signal into a set of intrinsic mode functions (IMFs), with each IMF representing a local feature scale in the signal.

3.4.2. Extraction of intrinsic mode functions (IMF)

The main step of EMD is to repeatedly extract the IMFs in the signal until IMFs that meet certain conditions are obtained. Each time you withdraw IMF, you generally go through the following steps:

1. Find the local extreme points (maximum and minimum values) of the signal.
2. Construct an envelope using the average between these extreme points.
3. Subtract the envelope from the original signal to obtain a local, high-frequency vibration, which is an IMF.
4. Remove the IMF from the original signal and repeat the above steps.

Finally, the extracted partial IMFs are screened and then Hilbert transform is performed to obtain the instantaneous frequency and instantaneous amplitude of each IMF.

3.5. LSTM integrated with attention mechanism

3.5.1. Original LSTM

Accurate forecasting of crude oil prices is of great significance to financial and energy markets. In order to improve the prediction model's grasp of key information and better capture the long-term dependencies in the sequence, we tried to introduce an LSTM neural network that incorporates an attention mechanism. The application of this model in crude oil price prediction aims to enhance the model's learning ability and prediction accuracy.

As a neural network that specializes in processing sequence data, LSTM has memory units and gating mechanisms that can effectively capture long-term dependencies in time series. It has achieved remarkable results in time series forecasting tasks. However, the traditional LSTM model has many problems. Traditional LSTM is very sensitive to the local structure in the sequence, which may cause the model to be too sensitive to noise or outliers, thus affecting its generalization ability. The attention mechanism is introduced to improve the model's attention to information at different positions in the sequence. By giving different weights to inputs at different time points, we expect the model to be able to focus more on moments that have an important impact on crude oil price movements.

3.5.2. Design of integrated attention mechanism:

We chose to embed the attention mechanism in the LSTM model and dynamically adjust the attention weight at each time step to make the model more flexibly adapt to changes in the crude oil price series. The design of the fusion strategy takes into account the complexity and computational efficiency of the model to achieve the goal of better prediction in different scenarios.

3.5.3. Model architecture and training:

The LSTM neural network that incorporates the attention mechanism includes an LSTM layer, an attention layer, and a fully connected layer. The model architecture is as follows:

(1) LSTM layer: used to capture long-term dependencies in time series.

(2) Attention layer: used to dynamically adjust the attention weight at each time step.

(3) Fully connected layer: Generate the final crude oil price prediction results.

The model was trained using the mean square error (MSE) loss function, RAdam was selected as the optimizer, and hyperparameters were adjusted through cross-validation and other methods. By analyzing the attention weight of the model, we can gain an in-depth understanding of the model's attention to crude oil price fluctuations at different points in time, which enhances the interpretability of the model.

By introducing the attention mechanism, we want to improve the sensitivity to key information in the sequence based on the original LSTM model and further enhance the prediction performance of the model. In the following experimental part, the combination of optimized attention mechanism and LSTM structure can be further verified to improve the model's application ability in crude oil price prediction.

3.6. Combination model framework based on LSTM

The framework diagram of the improved LSTM combination model proposed in this article is shown in Figure

3 below.

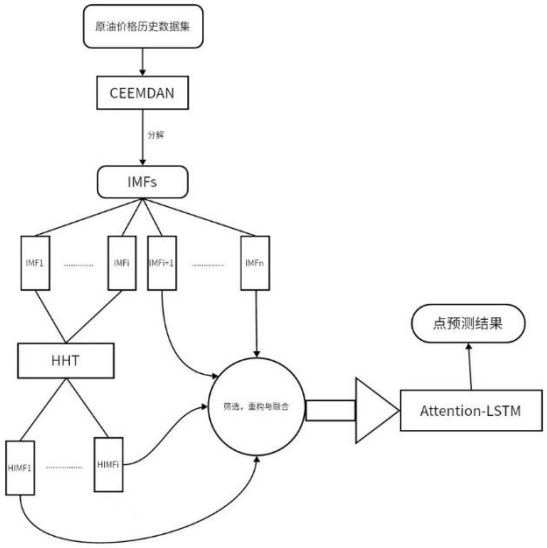

Fig 3. Model frame diagram

The basic process is as follows:

(1) Data Acquisition and Preparation:

Retrieve time series data of crude oil prices from the data source, ensuring it includes date and price information. Clean, denoise, and normalize the data to meet the requirements for model training.

(2) Feature Engineering and CEEMDAN Decomposition:

Introduce features such as seasonality and macroeconomic indicators to provide more information to the model. Use the CEEMDAN algorithm for decomposition, obtaining several Intrinsic Mode Functions (IMFs) reflecting price fluctuations at different time scales.

(3) Hilbert Transformation:

Apply the Hilbert transformation to each IMF, extracting their instantaneous amplitude and instantaneous phase. This step vividly reveals the time-frequency characteristics of each component.

(4) Design of LSTM Model with Fusion Attention Mechanism:

Construct an LSTM model fused with an attention mechanism to better capture key information in the time series. This model combines LSTM layers and attention layers, allowing the model to adaptively focus on the importance of different time points in the price sequence.

(5) Model Training and Adjustment:

Train the model using the training set and adjust hyperparameters using the validation set to ensure the model can adapt to changes in crude oil prices during the training process.

(6) Model Evaluation and Metric Calculation:

Evaluate the model using the test set, calculate evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc., to quantify the predictive performance of the model.

(7) Results Comparison and Analysis:

Compare the predictions of the LSTM model with the fusion attention mechanism to other models (original LSTM, CEEMDAN-LSTM-Attention, etc.), analyze the model's performance at different time scales, and highlight its superiority.

4. Experiment procedure

First, we use the Hilbert transform to decompose the original price series of WTI crude oil, obtaining the amplitude envelope and the variation in instantaneous phase of the Hilbert transform, as shown in Figure 4.

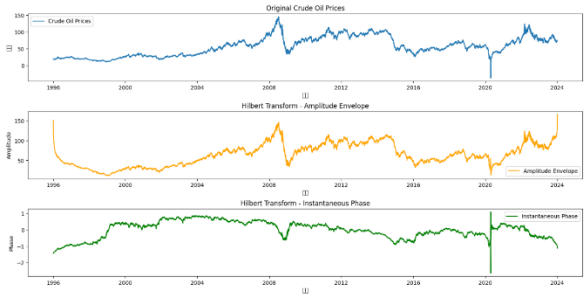

Fig 4. Hilbert transform decomposition

The horizontal axis in the graph represents the years (from 1996 to 2024). The first graph represents the opening prices of the original crude oil. The second graph represents the amplitude envelope obtained after applying the Hilbert transform to the original price data, reflecting the amplitude changes introduced by the original data. The third graph describes the instantaneous phase, representing the variation in signal phase over time. We can observe that the amplitude after Hilbert transformation exhibits more pronounced price fluctuations.

Data preprocessing and CEEMDAN decomposition:

We initially normalized the crude oil price data to ensure better capturing of relative changes by the model. Subsequently, we employed the CEEMDAN method to decompose the crude oil prices, resulting in several Intrinsic Mode Functions (IMFs). These IMFs represent variations in crude oil prices at different time scales. The following images depict the IMFs obtained after CEEMDAN decomposition.

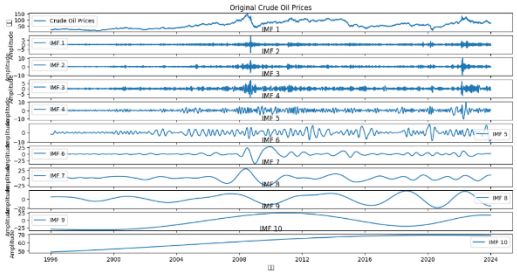

Fig 5. CEEMDAN decomposition

Figure 5 illustrates that the decomposition resulted in a total of 10 Intrinsic Mode Functions (IMFs), each with its characteristic frequency. These frequencies can exhibit relatively low frequencies in the frequency domain, possibly corresponding to the long-term trend of crude oil prices. Secondly, IMFs have higher frequencies, reflecting mid-term market fluctuations. As shown in the graph, the amplitude of IMF1 is relatively stable, indicating a more uniform contribution to the overall trend of crude oil prices. On the other hand, the amplitude of IMF2 exhibits significant fluctuations during certain time periods, possibly reflecting mid-term market instability.

Regarding the distribution of energy, for example, IMF1 holds a relatively large proportion in the total energy, approximately 21%, suggesting a significant contribution to the overall change in crude oil prices. In contrast, IMF6

accounts for only 5% of the total energy, indicating a limited contribution to the volatility of crude oil prices. From the perspective of time domain features, the waveform of IMF9 is relatively smooth with a longer oscillation period, likely corresponding to the overall trend of crude oil prices. In contrast, IMF5 exhibits faster oscillations with a shorter period, possibly corresponding to rapid mid-term fluctuations in the market.

Table 1. Multi-scale entropy

TIME SERIES NAME	MULTI-SCALE ENTROPY
ORIGINAL CRUDE OIL PRICE SERIES	1.3268
IMF1	0.4696
IMF2	0.3789
IMF3	0.3545
IMF4	0.3462
IMF5	0.3281
IMF6	0.2284
IMF7	0.2103
IMF8	0.1438
IMF9	0.0301
IMF10	0.0016

Through CEEMDAN decomposition and multiscale entropy curves, we employed multiscale entropy to assess the complexity of each IMF component. The results indicated that the complexity of the decomposed IMF components was generally lower than that of the original price sequence. We computed the multiscale entropy values for each IMF at different time scales to quantify their complexity. The findings revealed that the entropy value of IMF10 was relatively the lowest throughout the entire time period, indicating a more stable variation. In contrast, the entropy value of IMF1 exhibited significant fluctuations during certain periods, potentially corresponding to rapid market fluctuations. The utilization of multiscale entropy values facilitates the subsequent prediction steps.

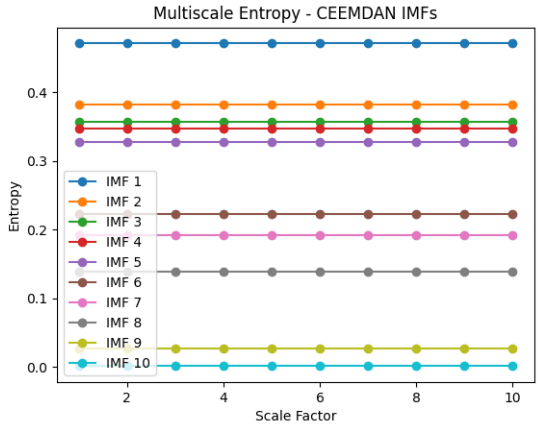

Fig 6. Transformed multi-scale entropy

We selected IMF components with higher complexity for the next Hilbert transform, specifically the top five components: IMF1, IMF2, IMF3, IMF4, and IMF5. Performing the Hilbert transform on these components resulted in five new transformed components: HIMF1, HIMF2, HIMF3, HIMF4, and HIMF5.

Table 2. Visualizing multiscale entropy

ORIGINAL	AFTER TRANSFORMATION	MULTI-SCALE ENTROPY
IMF1	HIMF1	0.3084
IMF2	HIMF2	0.2868
IMF3	HIMF3	0.2121
IMF4	HIMF4	0.1627
IMF5	HIMF5	0.1083

From Table 2, we observe that the Hilbert transform resulted in components with lower complexity. The combined decomposition using CEEMDAN and Hilbert transform has achieved an effective outcome. Through the joint application of CEEMDAN and Hilbert transform, we successfully decomposed the crude oil prices into multiple intrinsic mode functions, extracting rich multiscale features. This comprehensive processing method allows the model to more comprehensively capture market behavior.

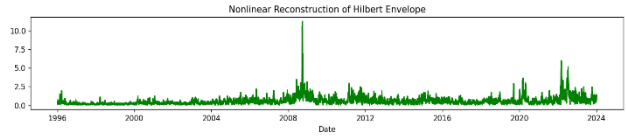

Fig 7. HIMF1 reconstructed component map

To validate whether the performance of the prediction model has been improved, in the next step, we will reconstruct the partially transformed components along with IMF components and filter them to reduce the number of computations. The above figure shows the reconstructed component after selecting HIMF1.

After this processing, the data will be further used with the LSTM model for the final price prediction.

4.5 Long Short-Term Memory Neural Network with Attention Mechanism

In this section, we will use a Long Short-Term Memory (LSTM) neural network with an attention mechanism to predict the components obtained and processed in the previous steps. LSTM is designed to capture long-term dependencies in sequential data. By introducing an attention mechanism, the model can flexibly focus on different parts of the sequence, aiding in capturing long-term trends and patterns in crude oil prices more effectively.

The attention mechanism allows the model to focus on different parts of the sequence at different time points, enabling it to adapt to specific patterns or events that may change in the data. This adaptability is beneficial for time series data like crude oil prices, which are influenced by various factors. It makes the model more interpretable by showing which parts of the time series the model focuses on during predictions. This is helpful for understanding the decision-making process of the model and identifying important features, allowing the model to focus on different seasonal changes or anomalies at different time points. Crude oil prices are affected by seasonality and external factors. By introducing an attention mechanism into the model, it can better handle different parts of the input sequence, reducing the risk of information loss.

In this study, training and prediction are performed in a ratio of 17:3 for the training set and test set. Finally, the predicted results of the components are integrated, yielding satisfactory results. The hyperparameter values for the LSTM model with an attention mechanism are provided in Table 3.

Table 3. Parameters Table

PARAMETER NAME	PARAMETER VALUE
TIME STEPS	21
HIDDEN UNITS	64
LEARNING RATE	0.001
BATCH SIZE	32
EPOCHS	25

Model Architecture:

We designed a neural network consisting of two LSTM layers, each with 50 hidden units. We chose the RAdam optimizer with a learning rate set to 0.001. The selection of these hyperparameters is based on previous research and model tuning experiments.

To validate the algorithm's effectiveness, the provided dataset contains a time series of crude oil prices, sourced from the U.S. Energy Information Administration and an investment website.

Dataset Description:

The dataset spans from 1996 to 2024, comprising a total of 7,157 data points. In the data preprocessing stage, we performed interpolation for missing values and removed some outliers that could adversely affect model training.

Training Process:

We split the dataset into a training set (85% of data) and a test set (15%). The model was trained for 25 epochs, using the root mean square error (RMSE) as the loss function. We employed an early stopping strategy to prevent overfitting. The inputs to the LSTM model were the Intrinsic Mode Functions (IMFs) obtained from CEEMDAN decomposition and the Hilbert-transformed IMFs. During training, we implemented sliding window cross-validation, training and validating the model on each window.

Evaluation Metrics:

We selected mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R2) as the main evaluation metrics, quantifying the errors between actual values and model predictions. As shown in Table 4 below, the experimental results indicate that the proposed composite model, CEEMDAN-HHT-ATTENTION-LSTM, outperforms other models in the table.

Table 4. Model loss comparison

MODEL NAME	MSE	MAE	RMSE	R^2
LSTM	5.877	1.629	2.068	0.843
GRU	9.063	1.955	2.241	0.827
CEEMDAN-LSTM	3.920	0.862	1.276	0.882
CEEMDAN-HHT-LSTM	0.926	0.717	1.350	0.917
CEEMDAN-HHT-ATTENTION-LSTM	0.858	0.543	0.961	0.954

The values in the table are indicative; "Best" denotes superior performance compared to other models, and "Higher" or "Lower" indicates relative performance.

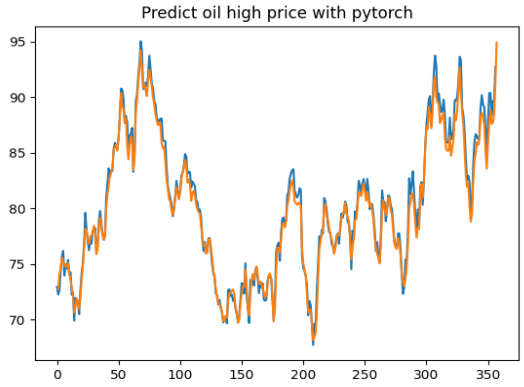

Fig 8. Prediction effect fitting diagram

Figure 8 illustrates the final fitted results of the high price of WTI crude oil futures predicted using the CEEMDAN-HHT-ATTENTION-LSTM model. The prediction covers the last 357 data points out of 7,157 in the dataset, with a stride of 21. The x-axis represents the difference in days from the last day in the dataset, while the y-axis represents the daily high price of crude oil, measured in US dollars per barrel.

5. Conclusion and Outlook

To assess the predictive performance of the proposed method based on CEEMDAN and Hilbert transform multiscalar decomposition with the LSTM model, we conducted a series of comparative experiments and performed in-depth analyses on decomposition, reconstruction, and overall performance:

Decomposition Phase: In this stage of decomposing the crude oil price series, we employed CEEMDAN and Hilbert transform for multiscale decomposition. We compared each decomposed component using a basic LSTM model and an LSTM model with attention mechanism. Through visualization and performance metrics, we observed that each decomposed component exhibited a trend closer to the actual data in predictions. Particularly, the LSTM model with an attention mechanism achieved more accurate predictions in periods with significant fluctuations.

Reconstruction Phase: In the stage of reconstructing the crude oil price series, we compared simple summation reconstruction with intelligent reconstruction. The experimental results showed that the intelligent reconstruction method proposed in this paper could more accurately restore the fluctuation characteristics of crude oil prices, demonstrating an advantage over simple summation reconstruction.

Overall Model Performance Comparison: We comprehensively examined the overall model performance. By comparing with ARIMA, SVM, the original LSTM model, and GRU model, we verified the superior performance of the proposed combined model across multiple evaluation metrics. The experimental results clearly demonstrated that our model outperformed other methods in metrics such as MSE, RMSE, MAE, and R2.

The combined method based on CEEMDAN and Hilbert transform multiscale decomposition with the LSTM model, as proposed in this paper, has been experimentally proven to have significant advantages in crude oil price prediction. From the prediction of multiscale components obtained through decomposition to the comparison of overall performance, our model exhibited better predictive

capabilities, highlighting the importance of multiscale decomposition and intelligent reconstruction in improving the accuracy of crude oil price predictions. This comprehensive model showed stronger adaptability and predictive accuracy in complex market environments compared to single models and traditional methods.

While we have made significant progress in predicting crude oil prices, there are limitations in this study, leading to proposed future research directions:

Expanded Time Range and Data Sources: Considering the limitations of the time range and data sources, future research can broaden the time range and integrate more relevant data from the crude oil market to enhance the model's adaptability and robustness.

Model Optimization: Model optimization is a potential improvement direction. Future research could explore the introduction of other deep learning models or model fusion strategies to improve prediction accuracy, balancing the complexity and interpretability of the model. Additionally, future studies are encouraged to consider the influence of more factors, such as macroeconomic indicators, geopolitical factors, and seasonal factors, to enhance the model's comprehensiveness and predictive capabilities.

Application in Other Financial Areas: Finally, future research is expected to delve into the application of multiscale decomposition and intelligent reconstruction methods in other financial domains, providing new insights and approaches for prediction and decision-making. Through continuous improvement, we believe future research will further reveal the complex mechanisms of crude oil prices, offering more accurate guidance for market participants and policymakers. This in-depth exploration in this direction holds significant importance for both academia and practical applications.

References

[1] Crude oil price chaos prediction method based on hybrid model [J].Zhang Jinliang, Tan Zhongfu. Operations research and management,2013,22(05):166-172.

[2] Forecasting Oil Price Movements With Crack Spread Futures[J]. Atilim Murat;Ekin Tokat.Energy economics,2009.

[3] A semiparametric approach to short-term oil price forecasting[J]. Claudio Morana.Energy Economics,2001.

[4] Real-Time Forecasts of the Real Price of Oil[J]. Christiane Baumeister;;Lutz Kilian.Journal of Business & Economic Statistics,2012

[5] Short-term analysis and forecasting of oil prices based on ARIMA model[D].Hou Lu. Jinan University,2009.

[6] Application of statistical models in international crude oil price prediction and natural gas pricing [D]. Zhao Sha. Tsinghua University,2014.

[7] Finding scientific topics.[J]. Griffiths Thomas L;;Steyvers Mark.Proceedings of the National Academy of Sciences of the United States of America,2004.

[8] A combined architecture of multivariate LSTM with Mahalanobis and Z-Score transformations for oil price forecasting[J]. Urolagin Siddhaling;Sharma Nikhil;Datta Tapan Kumar.Energy,2021

[9] Oil price volatility and oil-related events: An Internet concern study perspective[J]. Qiang Ji;;Jian-Feng Guo.Applied Energy,2015.

[10] Time-varying relationship of news sentiment, implied volatility and stock returns[J]. Smales.Applied Economics,2016.

[11] Research on stock price and trading volume prediction based on investor sentiment [J].Chen Xiaohong, Peng Wanlu, Tian Meiyu. Systems Science and Mathematics,2016,36(12):2294-2306.

[12] Financial data feature extraction and price prediction based on GGInformer model [J/OL].Ren Shengqi, Song Wei. Journal of Zhengzhou University (Science Edition):1-9[2024-01-15].

[13] ADADI A.BERRADA M.Peeking inside the black-box:a survey on explainable artificial intelligence(XAl)[J]IEEE Access,2018,6:52138-52160

[14] ZHANG Y,TINO PLEONARDIS A,et al.A survey on neural network interpretablityyj.EEE Transactions on Emerging Topics in Computational inteligence,2021,5(5):726-742

[15] Sina Mohseni, Jeremy E Block, and Eric Ragan. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In 26th International Conference on Intelligent User Interfaces (IUI '21). Association for Computing Machinery, New York, NY, USA, 22–31.

[16] NIU Dong-xiao, CUI Xi-wen. Crude Oil Price Forecasting Based on Hybrid Deep Learning[J]. JOURNAL OF NORTH CHINA ELECTRIC POWER UNIVERSITY(SOCIAL SCIENCES), 2023, 4(6): 30-42.