Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30962

Machine Learning-Driven Multi-model Ensemble for Crude Oil Price Prediction: A Comprehensive Review

Haiming Wang

Qiongtai Normal University, Haikou, China
2212765388@qq.com

Abstract. Crude oil has become an indispensable resource for ensuring the normal operation of society, and accordingly, crude oil price forecasting has emerged as a research area. However, crude oil prices often fluctuate due to various human or natural factors, making accurate prediction challenging. In recent years, an increasing number of scholars have adopted ensemble models instead of single models for oil price forecasting. In view of this trend, this paper collects and organizes eight representative studies (selected from high-impact literature on machine learning-based ensemble methods in the past five years) and conducts a comprehensive analysis. These models are categorized into three groups based on their technical cores: traditional machine learning ensembles, cross-domain hybrid models, and deep learning-based core models. The research results show that under the same conditions, ensemble models tend to achieve more accurate oil price forecasts than traditional single models. Additionally, this paper analyzes the current limitations of these ensemble models and proposes targeted improvement measures, providing feasible insights for their future development and practical application in related fields.

Keywords: Oil price, Ensemble model, Artificial Intelligence, Machine learning

1. Introduction

Crude oil is of vital importance to the economic development of modern society, and its price directly influences and determines many business and political decisions [1]. However, oil prices are often affected by various factors and fluctuate, such as supply and demand, geopolitics, inventory and logistics. Therefore, crude oil price prediction has become one of the most paramount research problems in the energy field [2]. In fact, many scholars have begun to use machine learning-based methods to build oil price prediction models and have achieved remarkable results [3].

For many years, experts have used a single model to predict oil prices. For example, Lu et al. [4] adopted the MS-MIDAS model to study the impact of the Chicago Board Options Exchange crude oil Volatility Index on the prediction of China's oil futures. Cen Z. et al. [5] utilized LSTM network based on CNN by alleviating the influence from historical data and enhancing that of current data. Similarly, Siham AKIL et al. [6] combined LSTM and Temporal Convolutional Networks (TCN), the model effectively captures temporal patterns and economic effects. The results show its superior performance over traditional methods and highlight the significant impact of specific economic indicators on oil prices. Furthermore, Cui Z et al. [7] innovatively adopted the pinball loss function

DOI: 10.54254/2755-2721/2026.TJ30962

to enhance the reliability of VMD (variational mode decomposition)-GRU (gated recurrent unit) model.

Even if these single models employ various functional components, there is still a significant flaw when it comes to the problem of oil price prediction with obvious volatility.

In this background, this paper systematically examines the research on the prediction of oil prices utilizing ensemble models. Through professional analysis and comparison, it summarizes the advantages and disadvantages of different methods. Moreover, the innovative direction of integrating signal decomposition methods with neural networks is introduced.

More importantly, the main contribution of this paper is that it provides a comprehensive framework for the field of oil price prediction. On the one hand, it plays a significant role in real-world scenarios such as industrial development (e.g., energy industry planning), business decision-making (e.g., pricing strategies in oil trade), and even offering references for the analysis of geopolitical situations. On the other hand, in academic areas, it bridges the gap between theoretical research and practical application in oil price prediction. It not only enriches the methodological system of integrated models but also inspires future scholars to explore more interpretable and robust algorithms for oil price forecasting. Moreover, its framework sets a benchmark for interdisciplinary studies, encouraging collaborations among economics, data science, and energy policy research.

In conclusion, the research of this article is guided by the following questions:

RQ1: How do ensemble models improve prediction performance?

RQ2: How are the interpretability and practical application value of the ensemble model in oil price prediction? How to balance the two?

RQ3: What are the performance boundaries and limitations of existing ensemble models in oil price prediction?

2. Methodology

To ensure the reliability and representativeness of the final selected literature, this article employs a very effective process to screen papers. The process references methods utilized by H. S. Al Nuaimi et al. [8], who summarized studies about carbon emission estimation based on machine learning and provided feasible insights. Moreover, Kodjo Abel Odah et al. [9] used the similar method to summarize the best-performing models in 57 studies for predicting or estimating tomato yields. It cannot be ignored that the screening process adopted in this article is very similar to those used in these two experiments. Therefore, it can be proved that the method adopted in this article is very effective.

2.1. The processes of screening

We first screen papers through keyword queries from the three databases of Google Scholar, ScienceDirect, and MDPI. The key words selected in this article are "machine learning", "oil price", "forecasting", "prediction", "ensemble", "integrated". The screening requirement is that the above keywords must appear in the title, abstract or keywords of the literature.

It is worth noting that through the listed keywords, it can be found that we have taken into account the factor of synonyms. Therefore, the articles after screening will be more comprehensive and will not wrongly filter out the relevant papers. In addition, this article only considers open access and English research articles to ensure authority and universality.

DOI: 10.54254/2755-2721/2026.TJ30962

The results of the first round of screening are as follows, Google Scholar (6,030 articles, accounting for 96.6%) dominates, while MDPI (10 articles, 0.2%) and ScienceDirect (200 articles, 3.2%) account for relatively small proportions. However, merely considering quantity is not enough. Since many of the articles presented on Google Scholar and MDPI have not been reviewed and certified by authoritative institutions, this article only adopts the literature from ScienceDirect.

2.2. AI-driven screening

It is noteworthy that this study refers to the experiment conducted by S. Kamra et al. [10] and adopts an AI-based approach for the second round of screening and this approach has achieved very authoritative and reliable results. There is no doubt that with the continuous development of society and the advancement of technology, artificial intelligence will play an increasingly significant role in the field of literature collection and screening.

Hence, we used ChatGPT to review 200 research articles in ScienceDirect and selected 11 of the most effective ones to ensure that the main body of the articles closely revolved around the several key words proposed in this article. Finally, after manual review and elimination of one article that was published too far ago (this paper only adopts the literature published from 2017 to 2025) and two that did not match the content, eight of the most representative articles were summarized [11-18], as shown in Table 1. It should be noted that the data used in the empirical part of the eight articles are all from the two databases, the West Texas Intermediate (WTI) and Brent Oil.

Table 1. Summary of related studies on crude oil price forecasting

year	author	title	journal	Ensemble model
2025	Kumari, Pritty; Chaudan, G. Y.; Kumar M, Satish	Transforming oil market analysis: A novel GAN + LSTM predictive framework	Next Energy	GAN +LSTM
2025	Zhu, Shuijie; Xu, Mei; Wu, Jie; Wang, Yanheng; Jiang, Xinsheng; Huang, Zhuangzhuang; Zhu, Wangyang	A study on crude oil price forecasting model integrating CEEMDAN-VMD multiscale decomposition with CNN-BiLSTM	Results in Engineering	CEEMDAN-VMD + CNN-BiLSTM
2024	Hoang, Christian; Hadas, Christian; Budin, Constantin; d'Arcy, Anne	How to select oil price prediction models — The effect of statistical and Financial performance metrics and sentiment scores	Energy Economics	Vader + Dictionary + Watson
2024	Luo, Hua; Yu, Yue	A novel hybrid forecasting system for crude oil futures prices: A dual perspective of deterministic forecasting and uncertainty analysis	Heliyon	RF +XGB+LGBM
2024	Yang, Xiong; Zhang, Zihang; Xu, Huihua	RV-FELM: Futures commodity price forecasting based on RIME-VMD algorithm coupled with FA-ELM	Heliyon	RIME-VMD + FA-ELM
2024	Zhang, Chen; Zhou, Xinmiao	Forecasting value-at-risk of crude oil futures using a hybrid ARIMA-SVR-POT model	Heliyon	ARIMA-SVR-POT
2022	Ding, Xinsheng; Fu, Lianlian; Ding, Yuehui; Wang, Yinglong	A novel hybrid method for oil price forecasting with ensemble thought	Energy Reports	Proposed + CM
2017	Chen, Yanhui; He, Kaijian; Tso, Geoffrey K.F.	Forecasting Crude Oil Prices: a Deep Learning based Model	Procedia Computer Science	ARMA + Deep Learning

DOI: 10.54254/2755-2721/2026.TJ30962

3. Literature review

In this section, we will categorize the eight papers that have undergone strict screening into three types based on the normal process: integrated models based on deep learning, Cross-domain hybrid models, and ensemble framework models. The reason for this approach lies in the fact that each category of models differs in technical principles, applicable scenarios, strengths, and limitations, while there are also commonalities across different studies. Explicit classification not only renders the structure of the paper clear at a glance but also facilitates other scholars in summarizing and inducing the differences and consensus within the field.

3.1. Three integrated models based on deep learning

With the continuous advancement of hardware and software, the computing power of computers has been steadily enhanced, which has made it feasible to integrate deep learning-based models on the basis of single models. This section will elaborate in detail on three models that integrate deep learning.

3.1.1. Deep learning technique

Deep learning, a core subset of artificial intelligence (AI) and machine learning (ML), employs multi-layered neural networks to automatically learn hierarchical representations from large-scale data via statistical techniques, enabling accurate recognition, prediction, and decision-making for complex tasks. More importantly, deep learning offers notable advantages for crude oil price forecasting: it effectively captures complex nonlinear and time-varying dynamics inherent in oil price movements, outperforms traditional linear models in handling high-dimensional, noisy, and non-stationary data, and enables adaptive learning from large-scale historical data to enhance prediction accuracy and robustness.

3.1.2. The processes of the three models

Yanhui Chen et al. [18] created a brand-new integrated model by fusing the autoregressive moving average model (ARMA) and the deep learning model in 2017, including the deep belief network (DBN) and the long short-term memory network (LSTM) and its equation is as follows:

$$y_t = \omega_{lm}\widehat{r_{lm}} + \omega_{nlm}\widehat{r_{nlm}}$$

Where (ω) denotes the weights of different forecasts; lm represents linear models, $(\widehat{r_{lm}})$ denotes the estimate from linear models; nlm represents nonlinear models employed, and $\left(\widehat{r_{nlm}}\right)$ denotes the estimate from nonlinear models. The experimental results are shown in Table 2.

It can be seen that the ARMA model effectively captures the linear features in oil price fluctuations, while the deep learning component also exploits nonlinear features (e.g., those induced by market sentiment, political factors, etc.).

DOI: 10.54254/2755-2721/2026.TJ30962

Table 2. Comparison of Mean Squared Error (MSE) among different models

Statistics	RW	ARMA	DBN	RW-DBN	ARMA-DBN	LSTM	RW-LSTM	ARMA-LSTM
MSE×10⁻⁴	5.3235	5.4753	5.3236	5.3224	5.3811	5.5219	5.3868	5.4504

Moreover, P. Kumari et al. [11] developed a novel model that integrates the generative adversarial network (GAN) and LSTM in 2025. Through experimental analysis, this combination is proven to be highly effective. Specifically, GANs excel at learning complex data distributions and generating high-fidelity images, texts, and other modalities, with mature applications in both academic and industrial domains. The equation used in the experiment is as follows.

$$\min_{G}\max_{D} E_{x\sim P_{\mathrm{dt}}(x)}\left[\log D\left(x\right)\right]+E_{z\sim P_{z}(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right]$$

Here, (G) (LSTM-based generator) attempts to generate synthetic data to deceive (D) (LSTM-based discriminator), which seeks to distinguish real data (x) (from the true data distribution) from $(G(z))$ (synthetic data generated by (G) from a random noise vector $(z\sim P_{z}(z))$).

In P. Kumari's study, GANs can to a certain extent mitigate the overfitting of LSTMs caused by excessive learning of training data, enabling the model to achieve favorable prediction performance on unseen crude oil price data. The specific experimental results are shown in Table 3.

The GAN-LSTM hybrid model outperforms other competing models on the test dataset, demonstrating its superior capability in modeling the underlying functional relationships.

Table 3. Performance comparison of GAN-LSTM ensemble model with LSTM, GRU and ANN

Accuracy measures	GAN+LSTM	LSTM	GRU	ANN
MSE	0.0013	0.0076	0.0070	0.0043
MAE	0.0292	0.0726	0.0696	0.0557
MAPE	4.6399	10.8732	10.3165	8.4798
SMAPE	4.7341	11.6714	11.0292	8.9499
NRMSE	0.0574	0.1381	0.1333	0.1046
R-squared	0.9430	0.5548	0.5852	0.7445

S Zhu et al. [12] proposed a hybrid forecasting model (complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN)- VMD- convolutional neural network (CNN)-bidirectional long short-term memory (BiLSTM)) to improve the accuracy of crude oil forecasting.

The CEEMDAN algorithm addresses modal aliasing by injecting adaptive Gaussian white noise into the original signal and performing multiple Empirical Mode Decomposition (EMD) decompositions. In WTI price forecasting, it extracts multiscale fluctuation patterns via Intrinsic Mode Functions (IMFs), constructs a multiscale feature matrix, and integrates with models like LSTM to predict each IMF component and reconstruct prices. Compared with traditional methods, CEEMDAN improves the clarity of feature decomposition and noise robustness, thereby better capturing the nonlinear relationships between crude oil prices and macroeconomic variables, and

DOI: 10.54254/2755-2721/2026.TJ30962

ultimately providing more stable and accurate forecasting results. The experimental results are shown in Table 4.

Table 4. Performance evaluation of different models

Model	MSE	RMSE	MAE	MAPE	R-squared
BILSTM	53.7746	7.3331	5.2950	9.2550	64.9215
CNN-BILSTM	30.7731	5.5474	4.2957	4.6710	78.9198
VMD-CNN-BILSTM	31.4605	5.6090	4.0583	6.2542	90.5280
CEEMDAN-VMD-CNN-BILSTM	15.6319	3.9537	2.8250	3.6648	95.9386

3.1.3. Summary of three integrated models based on deep learning

All three approaches integrate linear/nonlinear, time-domain/frequency-domain, and local/global feature extraction capabilities, addressing the limitations of single traditional models (e.g., ARMA's inability to capture nonlinearity, standalone LSTM's weakness in frequency-aware decomposition). They effectively model crude oil prices' complex dynamics driven by geopolitics, market sentiment, and macroeconomic factors. They can deliver more stable and reliable forecasts.

However, all approaches involve multi-stage integration, requiring extensive hyperparameter tuning and high computational resources. This increases the threshold for implementation and reproducibility. Meanwhile, the integration of deep learning and multi-step fusion leads to poor transparency in the forecasting process. It is difficult to quantify the contribution of individual components, hindering practical interpretation for policy- and decision-makers. Finally, all methods rely heavily on high-quality, large-scale datasets. Insufficient historical data, noisy inputs (e.g., incomplete macroeconomic indicators), or unbalanced samples (e.g., overrepresentation of stable market periods) can significantly degrade performance.

3.2. Cross-domain hybrid models

Traditional statistical models (e.g., ARMA, linear regression) excel at characterizing linear patterns in data (e.g., trends in time series, linear correlations between variables), while machine learning models (e.g., LSTM, random forests) can excavate complex nonlinear patterns (e.g., nonlinear impacts of market sentiment and geopolitical events on oil prices). Integrating these two types of models compensates for the deficiencies in reliability and accuracy of traditional single models.

3.2.1. The processes of two efficient models

A hybrid model, namely ARIMA-SVR-POT, is proposed by C. Zhang [13] in 2024, which integrates the autoregressive integrated moving average (ARIMA), support vector regression (SVR), and the Peak Over Threshold (POT) method derived from extreme value theory. Among them, Value at Risk (VaR), a financial term pioneered by J.P. Morgan, denotes the maximum potential loss an investor may suffer from holding a financial asset or portfolio over a specified horizon at a given confidence level [19]. The following equation can represent the hybrid model used in VaR calculations.

$$\widehat{\mathscr{H}_t} = \widehat{\mathscr{L}_t} + \widehat{N_t} + \widehat{\Sigma_t}$$

DOI: 10.54254/2755-2721/2026.TJ30962

where $\widehat{\mathscr{L}_t}$, $\widehat{N_t}$ and $\widehat{\Sigma_t}$ are the forecast values for linearity, non-linearity, and extreme using the ARIMA and SVR models and POT, respectively.

In the result detection section, he used the Kupiec test which is a statistical method used to validate the reliability of a result. As shown in Table 5 and Table 6, each confidence level (95%, 99%, 99.5%, 99.9%) corresponds to a reference interval. The likelihood ratio (LR) statistic falls entirely within the reference interval, indicating that the deviation between the loss probability predicted by the model and the actual loss probability is small, and the VaR prediction is valid. For example, ARIMA-SVR-POT has an LR of 0.4775 under the long position at the 95% confidence level, which falls within the interval (0.000982, 5.02) and passes the test. Conversely, if the LR statistic exceeds the reference interval, it indicates a large deviation between the model prediction and the actual situation, and the VaR is unreliable. For example, ARIMA-EGARCH has an LR of 13.8199 under the long position at the 95% confidence level, which is much larger than 5.02 and fails the test. Obviously, in the comparison between long and short positions, the ARIMA-SVR-POT model demonstrates superior performance in the short position scenario. Furthermore, this model passes all validation tests, with an extremely small discrepancy between the actual loss probability and the predicted probability, making it the most outstanding model in the evaluation.

Table 5. Long position metrics of different models

Long Position	95%	99%	99.5%	99.9%
ARIMA-EGARCH	13.8199	0.6769	0.6982	4.3691
ARIMA-SVR	1.8101	4.6320	1.4739	2.3704
ARIMA-EGARCH-POT	1.0402	1.6364	0.3366	0.0670
ARIMA-SVR-POT	0.4775	0.4219	0.2099	0.0670

Table 6. Short position metrics of different models

Short Position	95%	99%	99.5%	99.9%
ARIMA-EGARCH	11.9601	1.6364	0.6982	4.3691
ARIMA-SVR	2.7811	3.7578	3.2859	4.3691
ARIMA-EGARCH-POT	1.8101	1.1084	0.3366	0.8783
ARIMA-SVR-POT	1.0402	0.1533	0.0096	0.3034

Another highly innovative integrated model was created by Christian Haas et al. [14] and proposed in 2024. They offer a novel perspective for crude oil price forecasting models. Beyond conventional statistical metrics, to gauge the practical utility of the proposed models, they further evaluate the potential financial implications of their predictions through the simulation of a straightforward trading strategy. To elaborate, the authors demonstrate that the integration of qualitative information into forecasting models via sentiment analysis is capable of yielding enhancements in both statistical accuracy and financial performance.

Table 7 illustrates that in terms of statistical forecasting accuracy, the Vector Autoregression (VAR) models integrated with Vader or Watson sentiment analysis exhibit prominent performance. Specifically, the VAR models incorporating Vader or Watson sentiment analysis achieve the lowest Root Mean Square Error (RMSE) values, which are standard assessment indicator in oil price forecasting [20,21] —for instance, the RMSE of VAR-Vader stands at 2.097, while that of VAR-Watson is 2.125. Furthermore, these models pass the Diebold-Mariano (DM) test [22], with

DOI: 10.54254/2755-2721/2026.TJ30962

relatively high values in the DM column indicating that their forecasting errors are statistically significantly lower than those of other competing models.

From the perspective of financial returns, the Neural Network-A (NN-A) model combined with Watson sentiment analysis (NN-A-Watson) delivers the highest Return on Investment (ROI), which verifies its capability to generate positive returns in trading strategies based on long-term forecasting.

Table 7. Evaluation of model performance via RMSE, DM, MAE, and ROI

Model type	sentiment	RMSE	DM	DA	MAPE	Invested capital	ROI	Trade frequency
NNA	Watson	2.164	2	0.64	2.996	488,090	0.162	56
VAR	Valer	2.097	12	0.60	2.911	3,482,120	0.123	95
NNA	Valer	2.138	3	0.58	2.853	1,554,500	0.112	87
NNA	-	2.231	1	0.57	3.158	4,954,570	0.111	111
VAR	Watson	2.128	12	0.52	2.950	2,601,070	0.106	94
NNA	Dict	2.390	0	0.64	3.383	703,000	0.089	70
NNA	Valer	2.197	1	0.59	3.007	3,182,300	0.083	81
VAR	-	2.454	0	0.67	3.499	454,560	0.057	83
VAR	All	1.974	1	0.54	3.084	1,084,000	0.072	90
VAR	Dict	2.271	12	0.59	3.177	3,062,180	0.037	101
Spread	-	2.322	1	0.53	3.375	993,820	0.021	39
NNA	-	2.274	1	0.63	3.155	1,454,500	-0.038	88

3.2.2. Summary of the two ensemble models

The advantages of the two models are remarkably prominent. The ARIMA-SVR-POT model can effectively characterize the extreme fluctuations of crude oil and other financial assets (e.g., price slumps induced by black swan events). It only requires time series data such as price, eliminating the need for collecting and processing sentiment text data, which results in low costs for data acquisition and preprocessing. Notably, capturing extreme risks constitutes its core advantage.

On the other hand, the ensemble model integrated with sentiment analysis inherits the neural network's capability to capture nonlinear and sentiment-related correlations while retaining the linear interdependence and stability of the VAR model. It exhibits outstanding performance in both statistical accuracy (achieving the lowest RMSE and Mean Absolute Percentage Error (MAPE)) and financial returns (delivering the highest ROI), thereby functioning as an "all-round" model. Furthermore, it maintains stable performance across both short-term trading and long-term forecasting, and successfully passes validity tests (e.g., the Kupiec test) in both long and short position strategies.

However, both models suffer from partial lack of interpretability—they can only explain certain components or individual sub-models, while the overall integration rules (e.g., weight assignment) remain difficult to decompose. Secondly, the models exhibit a certain degree of complexity, posing challenges in aspects such as model combination, parameter tuning, and weight allocation.

DOI: 10.54254/2755-2721/2026.TJ30962

3.3. Three ensemble framework models

This section will introduce three novel ensemble models, which construct comprehensive forecasting systems through “multi-model ensemble”, “multi-perspective analysis”(e.g., deterministic + uncertainty) or “multi-algorithm coupling”, in order to leverage the complementarity of different methods.

3.3.1. Processes of three ensemble framework models

In 2022, X. Ding et al. [17] designed an ensemble model which employed the Random Forest (RF), XGBoost (XGB), and LightGBM (LGBM). Each of the three base models independently predicts input samples, generating their respective "Predicted Value 1", "Predicted Value 2", and "Predicted Value 3". Subsequently, the prediction results of each base model are assigned equal weights. Finally, the integrated outcome is output as the final forecasting result (Final Forecasting).

The most critical component of this method lies in the adoption of the equal weight approach, which endows each base model with an "equal status and weight". As a weight assignment strategy, it is characterized by its simplicity and efficiency. The model structure is shown in Figure 1.

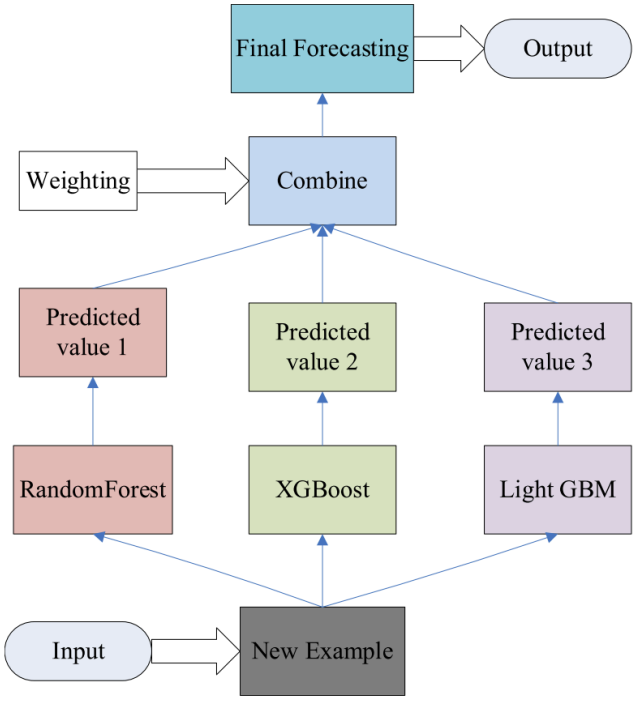

Figure 1. The structure of the model

In this experiment, Mean Absolute Error (MAE), MAPE, and Root Mean Squared Logarithmic Error (RMSLE) were employed to evaluate the proposed model alongside five other traditional single models. Table 8 clearly demonstrates that the proposed method achieves superior performance, ranking first across three evaluation metrics and outperforming the models from previous experiments significantly in every indicator.

DOI: 10.54254/2755-2721/2026.TJ30962

Table 8. The evaluation metrics of six models

Model	MAE	MAPE	RMSLE	Prediction Speed (second)
SVR	20.4213	0.0562	0.0795	3.58
RF	14.2699	0.0385	0.0583	8.02
XGB	14.5111	0.0387	0.0566	5.98
LGBM	14.4022	0.0384	0.0552	4.27
ARIMA-BP	15.8601	0.0447	0.0506	300.26
RF-XGB-LGBM	13.7417	0.0368	0.0538	11.61

H. Luo et al. [15] constructed a hybrid oil price prediction system from the dual perspectives of deterministic forecasting and uncertainty analysis in 2024. The ensemble nature of the proposed model is reflected in its "hierarchical integration of multi-technology stacks". By integrating three categories of technologies—denoising, multi-model prediction, and optimization strategies—it achieves a comprehensive transcendence in prediction accuracy. Figure 2 shows the five main processes of the model's operation.

(1) Denoising: Crude oil futures data (WTI, Brent) are denoised via EMD, CEEMD, SSA, VMD to obtain high-SNR sequences.

(2) Forecasting: Denoised data is input into a benchmark model pool (ARIMA, LSTM, SVM, etc.) for prediction.

(3) Screening: Top 4 optimal models are selected based on accuracy, with their weights determined.

(4) Optimization: Parameter initialization, position update, and fitness calculation are performed to output Pareto optimal solutions.

(5) Uncertainty Analysis: Interval prediction, multi-category analysis, and FCM are integrated for final analysis.

The test results of the proposed model (PROPOSED-CM) are illustrated in Table 9. It achieves the optimal performance across all evaluation metrics: the MAPE is 1.0761, the MAE is 1.0133, and the RMSE is 1.3733 (with negligible prediction errors). Additionally, the coefficient of determination (R^2) reaches 0.9891 (approaching 1, indicating extremely strong explanatory power), and the Directional Accuracy (DACC) is 85.0174 (surpassing 85% accuracy in predicting price movement directions). Overall, the model exhibits significantly superior out-of-sample generalization ability compared to all traditional counterparts.

DOI: 10.54254/2755-2721/2026.TJ30962

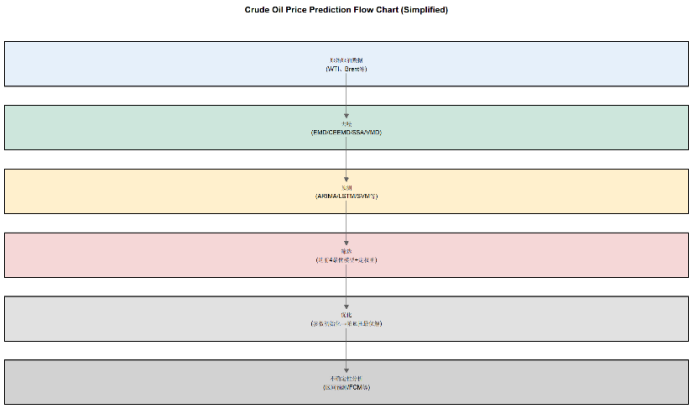

Figure 2. The five main processes of the model's operation

Table 9. The test results of the proposed model (Proposed-CM)

	MAPE	MAE	RMSE	MSE	R^2	DACC
ARIMA	3.6372	21.6037	26.8400	720.3851	0.8457	44.4056
BPNN	2.6996	15.9455	20.8864	436.2430	0.9066	50.3497
GRNN	3.9779	23.6585	30.0419	902.5186	0.8067	49.3007
SVM	2.4447	14.5506	18.5252	343.1812	0.9265	47.5524
LSTM	2.7162	16.1524	20.9142	437.4030	0.9063	46.8531
Proposed-CM	1.0761	1.0133	1.3733	1.8860	0.9891	85.0174

X. Yang et al. [16] proposed a new RV-FELM ensemble learning method based on VMD decomposition, Fourier attention prediction (FA) model, extreme learning machines (ELM) model and RIME optimization algorithm (ROA) in 2024. The specific construction algorithm of RV-FELM is as follows:

(1) Decompose the highly volatile nonlinear price series into several reasonable subsequences via the VMD algorithm based on the RIME optimization algorithm.

(2) Deconstruct and separately forecast the trend term and seasonal term using the ELM model and FA model; frequency-domain Attention can eliminate high-frequency noise in seasonal data, ensuring greater robustness.

(3) Integrate the prediction results in the time domain and frequency domain respectively and compare them with the original series.

For the purpose of comparison, they employed other traditional forecasting methods as benchmark models and compared their predictive performance with the RV-FELM ensemble learning approach. These traditional methods include commonly used futures forecasting models such as ARIMA, SVR, and ANN. Additionally, to examine the decomposition effect of VMD on crude oil prices, the EEMD decomposition method was incorporated for comparison with VMD, specifically in the form of EEMD-ELM and RIME-EEMD-ELM. The experimental results are shown in Figure 3.

DOI: 10.54254/2755-2721/2026.TJ30962

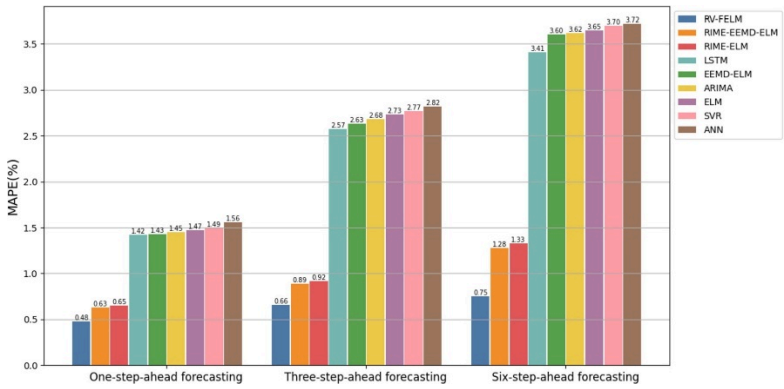

Figure 3. The experimental results

The RV-FELM model exhibits significant advantages in prediction accuracy for one-step-ahead, three-step-ahead, and six-step-ahead forecasts, with MAPE values of 0.48%, 0.66% and 0.75% respectively. These values outperform those of other benchmark models across the prediction horizons, highlighting the superior accuracy of the proposed method in commodity futures price forecasting. Furthermore, the predictive capability of RV-FELM surpasses that of RIME-EEMD-ELM, indicating that the decomposition effect of VMD on commodity futures prices is superior to that of EEMD. Meanwhile, it is also verified that prediction models adopting hybrid decomposition methods demonstrate higher stability in the long-term forecasting of commodity futures prices.

3.3.2. Summary of the three ensemble framework models

All three methods adopt the idea of "multi-model/multi-step integration". By integrating the advantages of different technologies (such as decomposition algorithms, prediction models, and optimization strategies), they effectively reduce the bias and variance of single models and exhibit stronger stability in complex nonlinear scenarios. For instance, the RF-XGB-LGBM ensemble via multi-model voting, the hierarchical integration of multi-technology stacks proposed by Luo et al., and the decomposition-prediction-fusion process of RV-FELM all maintain excellent accuracy across different prediction horizons. However, all three methods require fine-tuning of a large number of hyperparameters, such as the structural parameters of the models, the control parameters of the decomposition algorithms, and the weight parameters of the integration strategies. Unreasonable hyperparameter settings will seriously affect model performance. Meanwhile, due to the complexity of their internal decision-making logic, it is difficult to intuitively explain the generation mechanism of prediction results, which limits their application in fields with high interpretability requirements (e.g., financial regulation, medical diagnosis).

4. Discussion

This section summarizes the advantages and disadvantages of the methods presented above, answers the three questions raised in the introduction, and also analyzes the current research trends while proposing possible improvement measures. Meanwhile, we also describe the limitation of this paper.

4.1. Performance of ensemble models based on machine learning

On the whole, all three categories of methods adopt the integration framework. By fusing the advantages of multiple technologies or models, they effectively reduce the bias and variance of

DOI: 10.54254/2755-2721/2026.TJ30962

individual models. In complex, nonlinear, and highly volatile scenarios such as crude oil price forecasting, these methods can better capture the multi-scale features and underlying patterns of data, thereby maintaining superior prediction stability across different forecast horizons or market conditions. Each category has achieved breakthroughs in overcoming the difficulties of crude oil price forecasting along its respective technical route.

However, precisely due to the inherent characteristics of ensemble models, all three categories involve the configuration of a large number of hyperparameters, which require fine-tuning based on specific datasets. Otherwise, overfitting or underfitting is likely to occur, increasing the threshold for model application and the associated time costs. Additionally, whether it is deep learning integration, traditional machine learning integration, or cross-domain hybrid fusion, the internal decision-making logic of these models is relatively complex. It is thus difficult to intuitively explain the generation mechanism of prediction results, and their interpretability is insufficient for application in certain fields.

4.2. Answers to the three main questions

RQ1: Ensemble models mainly improve prediction performance through three dimensions: technical complementarity, error cancellation, and multi-dimensional feature mining.

RQ2: In terms of interpretability, the traditional machine learning ensemble group exhibits the strongest interpretability, the cross-domain hybrid models demonstrate moderate interpretability, and the deep learning core group has the weakest interpretability. From a practical perspective, all three categories of methods have demonstrated value in different scenarios. To balance these two performance aspects, the approach of "modular decomposition + transparency of key processes" can be adopted. For instance, the multi-technology stack integration proposed by Luo et al. modularizes each process, including denoising, forecasting, optimization, and uncertainty analysis, and documents the input-output relationships and technical logic of each module in detail.

RQ3: The optimal performance of the three categories of models is concentrated in scenarios characterized by sufficient data, stable conditions, and prediction horizons matching the technical characteristics of the models. However, the three categories of models have limited adaptability to extreme scenarios (e.g., sudden geopolitical conflicts and policy shifts), exhibit high dependence on hyperparameters with insufficient generalization ability, and incur higher computational costs compared to single models.

4.3. Possible improvement measures

Practitioners can draw on the idea of multi-technology stack integration to decompose the core processes of the three categories of models into independent modules such as data preprocessing, feature extraction, and prediction output. They should document the input-output relationships and logical rules of each module, thereby making the decision-making chain traceable and facilitating users to quickly locate key decision-making links, which can improve the interpretability of the models. In addition, researchers can reduce noise interference through data cleaning, adopt methods such as SMOTE to address the problem of class imbalance, and perform data augmentation operations (e.g., rotation, cropping, and noise injection) to expand the diversity of training data, so as to tackle the issue of insufficient generalization.

DOI: 10.54254/2755-2721/2026.TJ30962

4.4. Limitations of this paper

Although we have summarized eight valid papers by combining AI technology with manual screening, other databases such as Scopus and Web of Science have not been included in the reference scope due to access restrictions. This may result in some advanced ensemble model methods not being considered in this paper.

5. Conclusion and future work

This paper selects eight articles on crude oil price forecasting using machine learning-based ensemble models through an effective screening process. For a more systematic presentation, the eight articles are categorized into three groups, with their respective applicable scenarios and limitations summarized in each section.

Most importantly, the research results indicate that the performance of ensemble models is often more accurate and reliable than that of single models. Notably, this paper proposes several improvement methods, which theoretically address the existing shortcomings of current ensemble models (e.g., poor interpretability and high computational complexity). However, the empirical validation aspect remains to be explored.

Future research should focus on enhancing the interpretability of ensemble models through state-of-the-art explanation frameworks and modular transparency designs. Additionally, exploring cross-domain applications and real-time adaptation capabilities should further reveal the practical value of these models in complex scenarios like crude oil price forecasting and beyond.

References

[1] H. J. Ali Ahmed, O. H. M. N. Bashar, and I. K. M. M. Wadud, “The transitory and permanent volatility of oil prices: What implications are there for the US industrial production?, ” Applied Energy, vol. 92, pp. 447–455, Apr. 2012, doi: 10.1016/j.apenergy.2011.11.013.

[2] L. Yu, W. Dai, L. Tang, and J. Wu, “A hybrid grid-GA-based LSSVR learning paradigm for crude oil price forecasting, ” Neural Computing and Applications, vol. 27, no. 8, pp. 2193–2215, Aug. 2015, doi: 10.1007/s00521-015-1999-4.

[3] A. Rao, G. D. Sharma, A. K. Tiwari, M. R. Hossain, and D. Dev, “Crude oil Price forecasting: Leveraging machine learning for global economic stability, ” Technological Forecasting and Social Change, vol. 216, p. 124133, Jul. 2025, doi: 10.1016/j.techfore.2025.124133.

[4] X. Lu, F. Ma, J. Wang, and J. Wang, “Examining the predictive information of CBOE OVX on China’s oil futures volatility: Evidence from MS-MIDAS models, ” Energy, vol. 212, p. 118743, Dec. 2020, doi: 10.1016/j.energy.2020.118743.

[5] Z. Cen and J. Wang, “Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer, ” Energy, vol. 169, pp. 160–171, Feb. 2019, doi: 10.1016/j.energy.2018.12.016.

[6] S. AKIL, S. SEKKATE, and A. ADIB, “Multimodal Deep Learning for Oil Price Forecasting Using Economic Indicators, ” Procedia Computer Science, vol. 236, pp. 402–409, 2024, doi: 10.1016/j.procs.2024.05.047.

[7] Z. Cui, T. Li, Z. Ding, X. Li, and J. Wu, “Probabilistic oil price forecasting with a variational mode decomposition-gated recurrent unit model incorporating pinball loss, ” Data Science and Management, vol. 8, no. 3, pp. 237–247, Sep. 2025, doi: 10.1016/j.dsm.2024.10.003.

[8] H. S. Al Nuaimi, A. Acquaye, and A. Mayyas, “Machine learning applications for carbon emission estimation, ” Resources, Conservation & Recycling Advances, vol. 27, p. 200263, Sep. 2025, doi: 10.1016/j.rcradv.2025.200263.

[9] K. A. Odah, S. C. A. Houetohossou, V. R. Houndji, and R. L. Glèlè Kakaï, “Machine learning techniques for tomato yield prediction: A comprehensive analysis, ” Smart Agricultural Technology, vol. 12, p. 101067, Dec. 2025, doi: 10.1016/j.atech.2025.101067.

[10] S. Kamra et al., “P23 Evolving Use of Artificial Intelligence and Machine Learning in Systematic Literature Reviews (SLRs), ” Value in Health, vol. 26, no. 12, p. S6, Dec. 2023, doi: 10.1016/j.jval.2023.09.032.

DOI: 10.54254/2755-2721/2026.TJ30962

[11] P. Kumari, G. Y. Chandan, and S. Kumar M, “Transforming oil market analysis: A novel GAN + LSTM predictive framework, ” Next Energy, vol. 7, p. 100303, Apr. 2025, doi: 10.1016/j.nxener.2025.100303.
[12] S. Zhu et al., “A study on crude oil price forecasting model integrating CEEMDAN-VMD multiscale decomposition with CNN-BiLSTM, ” Results in Engineering, vol. 27, p. 106391, Sep. 2025, doi: 10.1016/j.rineng.2025.106391.
[13] C. Zhang and X. Zhou, “Forecasting value-at-risk of crude oil futures using a hybrid ARIMA-SVR-POT model, ” Heliyon, vol. 10, no. 1, p. e23358, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23358.
[14] C. Haas, C. Budin, and A. d’Arcy, “How to select oil price prediction models — The effect of statistical and financial performance metrics and sentiment scores, ” Energy Economics, vol. 133, p. 107466, May 2024, doi: 10.1016/j.eneco.2024.107466.
[15] H. Luo and Y. Yu, “A novel hybrid forecasting system for crude oil futures prices: A dual perspective of deterministic forecasting and uncertainty analysis, ” Heliyon, vol. 10, no. 21, p. e39818, Nov. 2024, doi: 10.1016/j.heliyon.2024.e39818.
[16] X. Yang, Z. Zhang, and H. Xu, “RV-FELM: Futures commodity price forecasting based on RIME-VMD algorithm coupled with FA-ELM, ” Heliyon, vol. 10, no. 17, p. e36631, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36631.
[17] X. Ding, L. Fu, Y. Ding, and Y. Wang, “A novel hybrid method for oil price forecasting with ensemble thought, ” Energy Reports, vol. 8, pp. 15365–15376, Nov. 2022, doi: 10.1016/j.egyr.2022.11.061.
[18] Y. Chen, K. He, and G. K. F. Tso, “Forecasting Crude Oil Prices: a Deep Learning based Model, ” Procedia Computer Science, vol. 122, pp. 300–307, 2017, doi: 10.1016/j.procs.2017.11.373.
[19] D. Duffie and J. Pan, “An Overview of Value at Risk, ” The Journal of Derivatives, vol. 4, no. 3, pp. 7–49, Feb. 1997, doi: 10.3905/jod.1997.407971.
[20] C. Baumeister and L. Kilian, “Real-Time Forecasts of the Real Price of Oil, ” Journal of Business & Economic Statistics, vol. 30, no. 2, pp. 326–336, Apr. 2012, doi: 10.1080/07350015.2011.648859.
[21] C. Baumeister, L. Kilian, and X. Zhou, “ARE PRODUCT SPREADS USEFUL FOR FORECASTING OIL PRICES? AN EMPIRICAL EVALUATION OF THE VERLEGER HYPOTHESIS, ” Macroeconomic Dynamics, vol. 22, no. 3, pp. 562–580, 2018. doi: 10.1017/S1365100516000237
[22] F. X. Diebold and R. S. Mariano, “Comparing Predictive Accuracy, ” Journal of Business & Economic Statistics, vol. 13, no. 3, pp. 253–263, Jul. 1995, doi: 10.1080/07350015.1995.10524599.