Front. Eng. Manag. 2025, 12(4): 938–951
https://doi.org/10.1007/s42524-025-5042-x

REVIEW ARTICLE

Ting YAO, Pan-Feng ZHANG, Yue-Jun ZHANG

An overview of crude oil price forecasting based on big data technology

Abstract Accurate crude oil price forecasting is critical in energy economics and energy engineering, as it informs economic policy-making and investment decisions. The emergence of big data brings both new opportunities and challenges for crude oil price forecasting. This paper systematically reviews recent advances in crude oil price forecasting in the context of big data, with a focus on the evolution of data types, predictors, and modeling techniques. In particular, it analyzes key forecasting approaches, including conventional and data-driven forecasting models, while emphasizing the growing role of emerging data sources. Promising directions for future research include the integration of multi-source data, the reconstruction of high-frequency supply and demand indicators, the development of hybrid modeling approaches, the enhancement of model interpretability, and the evaluation of the economic value of forecasting outcomes.

Keywords crude oil price forecasting, big data technology, machine learning

Received Feb. 24, 2025; revised Jun. 9, 2025; accepted Jul. 8, 2025

Ting YAO
School of Economics and Trade, Hunan University of Technology and Business, Changsha 410205, China; Center for Resource and Environmental Management, Hunan University, Changsha 410082, China

Pan-Feng ZHANG
School of Economics and Trade, Hunan University of Technology and Business, Changsha 410205, China

Yue-Jun ZHANG (✉)
Business School, Hunan University, Changsha 410082, China; Center for Resource and Environmental Management, Hunan University, Changsha 410082, China
E-mail: zyjmis@126.com

This research was supported by National Natural Science Foundation of China (Grant Nos. 72204083 and 72243003), National Social Science Fund of China (No. 22AZD128), Natural Science Foundation of Hunan Province, China (No. 2025JJ30028) and Digital Intelligence Research Foundation of Hunan University of Technology and Business, China (No. 2023SZJ20).

1 Introduction

Fluctuations in crude oil prices exert a profound impact on global economic and social stability. Oil prices directly affect the fiscal revenues of oil-exporting nations and the energy costs for oil-importing countries, and further influence broader macroeconomic indicators (Van Eyden et al., 2019). Therefore, accurate oil price forecasting is essential for formulating macroeconomic policies, guiding corporate strategic decisions, and ensuring the stability of financial markets.

To gain a comprehensive understanding of research trends in oil price forecasting, we conduct a systematic review of relevant literature. By utilizing the search terms "oil price forecasting" OR "oil price prediction," a total of 52655 articles published between 2010 and 2024 are identified in the ScienceDirect database. The annual publication trend is shown in Fig. 1. The results reveal a steady increase in the volume of research related to oil price forecasting over time. However, the number of review articles remains limited, and comprehensive comparative analyses of crude oil price forecasting models are still lacking.

In recent years, the rapid development of big data technologies has introduced new opportunities, methods, and challenges for oil price forecasting. First, the emergence of diverse data types has significantly enriched the data sources used in forecasting. Under traditional frameworks, many key variables in the crude oil market could only be indirectly estimated or updated frequently, limiting the timeliness and representativeness of predictors. In contrast, big data enables the incorporation of unstructured and real-time information, such as social media comments (Beyer Díaz et al., 2024), news texts (Bai et al., 2022), search engine data (Qin et al., 2023), satellite images (Hao and Wang, 2023), and climate data (Yao and Zhang, 2024). These unstructured data help to compensate for the deficiencies of conventional statistics and provide a broader, more dynamic view of the oil market.

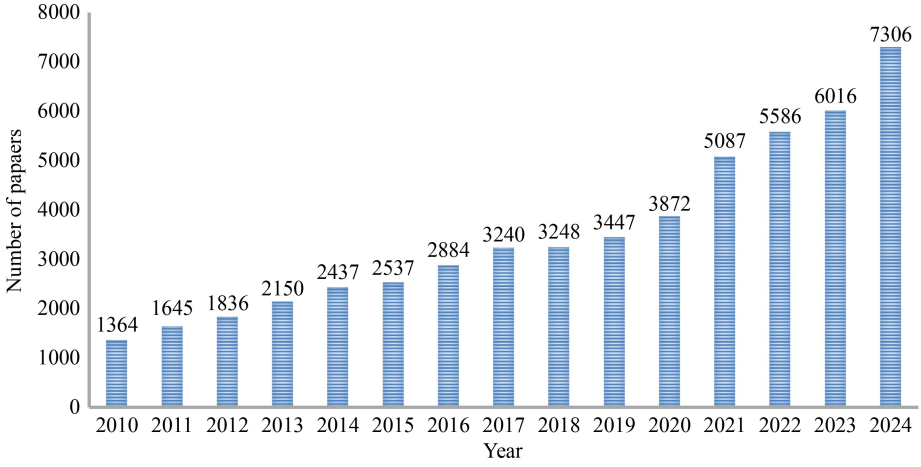

Fig. 1 The number of papers on oil price forecasting published in international journals between 2010 and 2024 (The data is derived from the ScienceDirect database).

Secondly, big data technology has transformed the construction of forecasting predictors. Traditional studies often relied on lagging or indirectly measured variables, such as trading volume, extreme returns, or survey-based sentiment, to proxy investor behavior and expectations. These approaches suffer from subjectivity and limited responsiveness. In the big data era, researchers can now track investor sentiment in real time using search engine queries and apply natural language processing (NLP) to extract attention and sentiment signals from texts, social media, and corporate announcements. This shift has made predictors more objective, timely, and reflective of actual market conditions.

Finally, oil price forecasting models have evolved to accommodate the complexity of big data environments. Conventional models—such as time series regressions or econometric approaches—are often based on rigid assumptions and struggle to capture nonlinearity, high dimensionality, and abrupt structural changes. To address these limitations, machine learning and deep learning methods have been increasingly adopted. These data-driven models excel in learning complex patterns and dynamic interactions among variables, offering improved accuracy and adaptability.

Despite these advancements, several challenges persist. The integration of heterogeneous data types remains technically difficult. Moreover, the low update frequency of fundamental indicators limits model responsiveness to rapid market shifts. Existing studies also tend to focus on accuracy improvements while overlooking issues of robustness, interpretability, and practical application. How to effectively embed forecasting outcomes into real-world decisions, such as investment, risk management, and policy design, remains an open and underexplored problem. How to translate forecasting results into actionable insights for investment, risk management, or policy design remains underexplored. Against this backdrop, this article reviews recent advances in oil price forecasting under the big data paradigm, with a particular focus on the application of machine learning techniques. It further identifies current limitations and outlines promising directions for future research.

The remainder of this article is organized as follows: Section 2 outlines the new data types and predictors for crude oil price forecasting under the background of big data; Section 3 provides a review of newly emerging crude oil price forecasting models in the context of big data, particularly the application and effectiveness comparison of machine learning models; Section 4 presents the research review and outlook.

2 The data types and predictors for crude oil price forecasting in the context of big data

2.1 New data and data processing methods for crude oil price forecasting

In the context of big data, the types of data for oil price forecasting are becoming increasingly diverse. As shown in Table 1, multi-source heterogeneous data such as text data, satellite data, search engine data, and high-frequency data are widely employed in oil price forecasting, significantly improving the accuracy of forecasting models.

2.1.1 Text data

The text data used in oil price forecasting mainly includes news reports and social media comments related to crude oil. Among them, news data are generally more formal and objective, whereas social media comments are more qualitative, unstructured, and subjective (Wong et al., 2017).

The news data used in crude oil price forecasting

Table 1 Summary of new data types used for oil price forecasting research in the context of big data

Category	Data	Literature	Data sources
Text data	News headlines	Bai et al. (2022)	Investing.com
Text data	News articles	Li et al. (2019b)	Investing.com
Text data	News articles	Gong et al. (2022)	Oilprice.com
Text data	News headlines	Wu et al. (2021a)	Oilprice.com
Text data	News headlines	Jiang et al. (2022a)	Oilprice.com
Text data	News articles	Loughran et al. (2019)	The DJES news database
Text data	Social media comments	Chen et al. (2021)	Sina Weibo and Twitter
Text data	Social media comments	Elshendy et al. (2018)	Twitter
Text data	Social media comments	Jiang et al. (2022b)	Eastmoney
Text data	Social media comments	Beyer Díaz et al. (2024)	Twitter
Image data	Cloud cover	Mukherjee et al. (2021)	NOAA (National Oceanic and Atmospheric Administration)
Image data	Cloud cover	Hao and Wang (2023)	Calculated from NASA's MODIS cloud cover data
Image data	OHLC charts	Ren et al. (2024)	Constructed from price and trading volume data from Investing.com
Search engine data	Google Trends	Wang et al. (2018)	Google
Search engine data	Google Trends	Qin et al. (2023)	Google
Search engine data	Google Trends	Li et al. (2020)	Google
Search engine data	Google Trends	Tang et al. (2020)	Google
High-frequency data	Financial assets, bulk commodities and macro assets	Degiannakis and Filis (2018)	TickData
High-frequency data	High-frequency stock index	Zhang and Wang (2019)	Yahoo Finance and the Wind database
High-frequency data	Crude oil futures trading data in five minutes	Huang et al. (2021)	The Thompson Reuter Tick History (TRTH) database
High-frequency data	Financial markets and US energy market data	Baumeister et al. (2015)	Bloomberg, Yahoo Finance, and EIA

primarily originates from platforms such as Investing.com and Oilprice.com. Investing.com is a comprehensive financial information service platform offering diverse market information. Bai et al. (2022) and Li et al. (2019b) demonstrate the effectiveness of mining news text from Investing.com to predict crude oil prices. In contrast, Oilprice.com focuses more on in-depth reporting on the energy sector. Gong et al. (2022) and Jiang et al. (2022a) find that news data from Oilprice.com helps improve the accuracy of oil price forecasting.

The headlines and articles of news texts are valuable sources for oil price forecasting. News headlines typically encapsulate the main theme and key viewpoint of the news, often reflecting the prevailing sentiments, which makes them widely utilized in oil price forecasting. For example, Bai et al. (2022) employ semantics-assisted nonnegative matrix factorization (SeaNMF) to extract topic features from news headlines for crude oil price forecasting. Wu et al. (2021a) apply Convolutional Neural Network (CNN) to extract text features from social media news headlines, demonstrating that such information enhances prediction accuracy for oil prices, production, and consumption. In contrast, news articles provide more detailed background information and analysis, typically expressing emotions in a more objective and neutral tone. Consequently, some researchers, such as Gong et al. (2022) and Loughran et al. (2019), have applied news articles to forecast oil prices.

Social media comments are increasingly valued for their timely updates and capacity to reflect real-time user sentiment and emerging trends. Among these platforms, Twitter has emerged as a prominent source for forecasting international oil prices. Chen et al. (2021) and Elshendy et al. (2018) employ Twitter data to predict oil prices. Moreover, some scholars use Chinese social media comments to predict Chinese crude oil futures prices. For example, Jiang et al. (2022b) utilize Chinese-language comment data from Eastmoney to forecast prices in the Chinese crude oil futures market. Similarly, Chen et al. (2021) incorporate both Chinese and English comment data from Sina Weibo and Twitter to predict crude oil contract prices on the Shanghai International Energy Exchange and the New York Mercantile Exchange, respectively. Moreover, social media commentary by politicians on oil-related topics has also been shown to contribute to oil price forecasting. Beyer Díaz et al. (2024) find that these tweets from US President Donald Trump significantly improve the predictive power of daily oil price forecasting models.

2.1.2 Image data

In addition to text data, recent research has increasingly incorporated image data to enhance oil price forecasting. In recent years, satellite imagery data, as a type of image data, has been used for research on oil price forecasting. For example, Mukherjee et al. (2021) report that government announcements have a significant impact on oil prices during periods of predominantly cloudy weather at major oil storage sites. Hao and Wang (2023) utilize satellite data on cloud cover to predict oil returns from the perspective of information uncertainty. Their findings reveal that that higher cloud cover within a given week correlates with increased information uncertainty, which in turn leads to lower oil returns in the subsequent week. Ren et al. (2024) employs CNN to analyze technical chart patterns (OHLC charts generated from daily crude oil price and trading volume data) for forecasting crude oil price movements.

2.1.3 Search engine data

Search engine data, particularly Google Trends, has emerged as a powerful tool for gauging investor attention due to its real-time capabilities and extensive reach. By analyzing fluctuations in search volumes for specific keywords, researchers can accurately gauge public interest in oil prices and the oil market. Studies have shown that investor attention index derived from search engine data can significantly enhance the accuracy of oil price forecasting models and help identify short-term market trends.

For example, Wang et al. (2018), Qin et al. (2023), Tang et al. (2020), and Li et al. (2020) use Google Trends data to predict crude oil prices. Specifically, Wang et al. (2018) construct a market attention index using Google Trends, along with two event indices: climate attention index and war attention index. Qin et al. (2023) develop a forecasting model that incorporates a broader range of online information by utilizing Google Trends and a stacking strategy. Their research demonstrates that Google Trends data significantly enhances the forecasting accuracy of both single-model and multi-model machine learning methods. Similarly, Tang et al. (2020) employ search engine data covering various factors related to crude oil prices and apply Multivariate Empirical Mode Decomposition (MEMD) to capture the intricate relationships between oil prices and multi-factor search engine data. Meanwhile, Li et al. (2020) construct an oil price forecasting model utilizing multilingual index and find that search engine data with multilingual keywords exhibits superior forecasting performance compared to data with monolingual keywords.

2.1.4 High-frequency data

High-frequency data provides the advantage of capture real-time market information, overcoming the limitations of monthly fundamental data in characterizing intraday dynamics within commodity and financial market (Degiannakis and Filis, 2018). With advancements in computer technology, the acquisition and storage costs of high-frequency data have significantly decreased, facilitating its widespread application in the field of oil price forecasting (Wen et al., 2016). For example, Zhang and Wang (2019) utilize high-frequency stock index data to predict low-frequency monthly crude oil prices. Similarly, Huang et al. (2021) explore the nonlinear dynamic relationship between geopolitical risk and oil prices using 5-minute trading data from crude oil futures.

However, there is ongoing debate about whether high-frequency data improves forecasting performance. Baumeister et al. (2015) utilize high-frequency data from US financial and energy markets to refine oil-price forecasts but find no consistent improvement in predictive accuracy. In contrast, Degiannakis and Filis (2018) incorporate high-frequency financial information with fundamental oil-market variables and find a significant enhancement in forecast performance. This indicates that the forecasting performance of high-frequency data largely depends on the selected variables and forecasting models.

2.1.5 Data processing methods

Effective data processing plays a key role in oil price forecasting in the era of big data. As shown in Table 2, signal decomposition technology and text data processing are commonly used in oil price forecasting research.

(1) Signal decomposition technology

Signal decomposition technology can decompose the complex oil price fluctuations into simpler and more easily analyzable components, providing clearer features and more accurate inputs for oil price forecasting models. By minimizing noise interference, these techniques not only extract key features but also enhance the accuracy and stability of forecasting models. Commonly used signal decomposition techniques in oil price forecasting include Empirical Mode Decomposition (EMD), Variational Mode Decomposition (VMD), and Singular Spectrum Analysis (SSA).

EMD is an adaptive signal decomposition technique that generates intrinsic mode functions based on the characteristics of oil price data. This approach facilitates the extraction of components such as trends, periodic fluctuations, and noise. For example, Fang et al. (2023a) use EMD based on Improved Slope-Based Method (ISBM) to decompose the oil price time series into Intrinsic Mode Functions (IMFs) and residual terms, subsequently employing Feedforward Neural Network (FNN) for ensemble analysis. Several variants of EMD, such as Ensemble Empirical Mode Decomposition (EEMD), Bivariate Empirical Mode Decomposition (BEMD), and

Table 2 Summary of data processing methods used for oil price forecasting in the context of big data

Category	Methods	Typical literature	Results
Signal decomposition technology	EMD, VMD, SSA, and SR	Fang et al. (2023a), Tang et al. (2015), Sun et al. (2018), Zhang et al. (2021), Li et al. (2021), Huang and Deng (2021), Zhang et al. (2023), Lin et al. (2022), Yu et al. (2017), Li et al. (2024)	Signal decomposition technology can decompose complex price fluctuations, enabling effective feature extraction and noise reduction, which can help improve the accuracy and stability of forecasting models.
Text data processing	Dictionary-based methods and deep learning-based methods	Li et al. (2016), Lucey and Ren (2021), Li et al. (2019b), Jiang et al. (2022b), Fang et al. (2023b), Bai et al. (2022)	Text data containing investor sentiment can improve the performance of forecasting models.

Complementary Ensemble Empirical Mode Decomposition (CEEMD) (Tang et al., 2015), are also widely used to enhance oil price forecasting accuracy. Sun et al. (2018) apply BEMD to decompose complex valued signals and then use multi-layer perceptron (MLP) and interval exponential smoothing method to predict the upper and lower limits, as well as residual components of IMFs, respectively. Zhang et al. (2021) employ EEMD for quadratic decomposition of the residual term after VMD decomposition and apply a particle swarm optimization kernel extreme learning machine to predict oil prices, achieving higher forecast accuracy.

Compared with EMD, VMD offers superior robustness and accuracy in oil price forecasting. Unlike EMD, VMD decomposes signals into multiple IMFs with fixed center frequencies, effectively addressing the mode aliasing problem found in EMD. Li et al. (2021) decompose oil price series with VMD, followed by random sparse Bayesian learning to predict each subsequence separately and combine the results to form the final predicted price. Huang and Deng (2021) optimize VMD parameter selection with an improved signal-energy-based (ISE) rule and employ a moving window strategy to refine the decomposition, then apply Long Short-Term Memory (LSTM) for forecasting. By integrating VMD decomposition and sample entropy (SE), Zhang et al. (2023) propose the VMD-SE-GRU model to improve oil price forecast accuracy. Li et al. (2024) proposes a novel secondary decomposition method that combines Twitter sentiment data to reduce the difficulty of crude oil price prediction through BEMD and genetic algorithm-optimized VMD.

SSA decomposes signals into multiple orthogonal components, effectively revealing different characteristics of oil price signals. Wang and Li (2018) apply SSA to decompose the time series of corn, gold, and crude oil prices, and subsequently develop an integrated forecasting model. This model combines a wavelet neural network, a backpropagation neural network, and a radial basis function neural network to predict the future prices of these commodities.

In addition, sparse representation (SR) is employed in oil price forecasting models. Yu et al. (2017) introduce coupled SR as a decomposition method for oil price time series, followed by individual forecasting using FNN. The final forecast is obtained by aggregating the individual predictions through a summation approach.

(2) Text data processing technology

Text data, including news reports and social media comments, offer valuable insights into market sentiment and public reactions to oil price fluctuations. These emotional signals serve as important supplementary predictors that can help to improving the forecasting accuracy. Sentiment analysis, a key area in NLP, focuses on extracting opinions, emotions, and attitudes from unstructured text (Chen et al., 2022). It enables researchers to quantify public sentiment and incorporate it into forecasting models.

Sentiment analysis was initially grounded in the lexicon-based methods, which can accurately reflect the unstructured features of the text (Liang et al., 2023). These methods use a set of vocabulary with pre-set emotional tendencies which are typically categorized into three types: negative, positive, and neutral. The emotional tendency of the text is quantified by scoring the frequency of words from these three categories. For instance, Li et al. (2016) apply sentiment analysis techniques based on Henry's Financial-Specific Dictionary (Henry, 2008) to extract news sentiment, employing these emotional signals as predictors in models such as logit regression, support vector machines, and decision trees. Lucey and Ren (2021) conduct a comparison of the out-of-sample forecasting power of Oil-Specific Dictionary (Loughran et al., 2019) and Henry's (2008) Financial-Specific Dictionary, concluding that Oil-Specific Dictionary can significantly improve forecasting performance.

Some scholars have turned to use deep learning methods to analyze text sentiment to uncover more complex emotional patterns. For example, Li et al. (2019b) utilize CNN to extract emotional features from news texts, combining them with a latent Dirichlet allocation (LDA) topic model to classify news topics, thus improving oil price forecast accuracy. Jiang et al. (2022b) leverage LSTM and Bidirectional Encoder Representations from Transformers (BERT) to analyze the sentiment of real-time online financial forum comments. Similarly, Fang et al. (2023b) employ the Financial Bidirectional Encoder Representations from Transformers (FinBERT) for sentiment analysis of financial news and integrate the model with VMD, attention mechanism, and bidirectional gated recurrent unit (BiGRU) to optimize the accuracy of oil

price forecasting.

To further improve the accuracy of sentiment analysis, some studies integrate dictionary-based methods with deep learning techniques. Bai et al. (2022) apply Global Vectors (GloVe) to convert documents into mathematical representations, subsequently extracted topic features using the short text SeaNMF topic modeling tool. Finally, they compute sentiment scores using TextBlob, which includes a built-in dictionary for sentiment analysis.

2.2 Predictors and predictor selection methods for crude oil price forecasting

This section further explores methods for extracting effective predictors from big data and apply them to improve the performance of oil price forecasting models. In the context of big data, with the diversification of data sources, researchers have increasingly introduced various novel predictors. These include investor sentiment extracted from social media comments and news texts and investor attention generated from search engine data. Such predictors enrich the dimensionality of forecasting models, enhancing their forecasting capabilities. Moreover, effective predictor selection plays a crucial role in identifying the most relevant variables, thus reducing risks of overfitting and multicollinearity, which in turn enhances the generalization capability of the forecasting model. Table 3 summarizes the main predictors and predictor selection methods in oil price forecasting research.

2.2.1 Investor attention

With the advancement of behavioral finance, numerous studies have examined the impact of investor attention on oil price fluctuations. Commonly used indicators for measuring investor attention include news and headline news (Narayan et al., 2017; Yuan, 2015), price limiting events (Seasholes and Wu, 2007), extreme returns (Barber and Odean, 2008), and trading volume (Hou et al., 2009). Although these proxy variables provide useful insights for oil price forecasting, they also present certain limitations. Specifically, some indicators depend on statistical data, introducing a lag, and the selection of these indicators can be subjective (Yao et al., 2017).

In recent years, researchers have increasingly leveraged real-time data, particularly search engine data, to measure investor attention, thereby mitigating the lag associated with traditional proxy variables. The investor attention index, constructed based on Google search volume, can more accurately and timely reflect market attention and trend changes. Numerous studies demonstrate that predictors derived from search engine data significantly improve the accuracy of oil price forecasting models. For example, Guo and Ji (2013) utilize Google search volume and Google Insights for Search to measure public attention toward the oil market. Yao et al. (2017) apply principal component analysis to aggregate Google search volume and construct the Google Search Volume Index as a proxy for investor attention. Qu and Li (2023) construct a multi-perspective investor attention index based on a Google search index consisting of 25 search terms related to alternative energy, macroeconomics, and geopolitics. In terms of forecasting performance, Li et al. (2015), Han et al. (2017), Wang et al. (2018) and Li et al. (2020) confirm that the forecasting factors constructed based on Google search data significantly improve the performance of oil price forecasting models.

Table 3 Predictors and predictor selection methods used for the oil price forecasting

Category	Proxy variables and methods	Typical literature	Results
Investor attention	News and news headlines, price-limiting events, extreme returns, trading volumes, and	Narayan et al. (2017), Yuan (2015), Seasholes and Wu (2007), Barber and Odean (2008), Hou et al. (2009)	There is a lag in these variables, and the process of selecting the indicators is subjective.
	Google search volume index	Guo and Ji (2013), Yao et al. (2017), Qu and Li (2023), Li et al. (2015), Han et al. (2017), Wang et al. (2018), Li et al. (2020)	The Google search volume index can fully reflect the investor attention to keywords in real time and can improve the performance of the forecasting model.
Investor sentiment	Market variables	Deeney et al. (2015), He and Casey (2015), Li et al. (2022)	This indicator is only the result of a combination of many economic factors, and it is difficult to fully capture investor sentiment solely through market variables.
	Text-based sentiment index	Li et al. (2019b), Zhao et al. (2023b), Wu et al. (2024), Li et al. (2016)	Deep learning techniques improve the ability of text sentiment analysis, thereby enhancing the performance of oil price forecasting models.
Financial market variables	Economic survey results	Qadan and Nama (2018)	This indicator is generally monthly and quarterly indicators, with a lag.
	Speculative activity and technical indicators	Manera et al. (2016), Guo et al. (2022a), Morana (2013), Alquist and Gervais (2013), Yin and Yang (2016), Wang et al. (2020)	This predictor can complement macroeconomic variables, and even technical indicators have stronger forecasting performance than macroeconomic variables.
	High-frequency financial data	Baumeister et al. (2015), Degiannakis and Filis (2018), Zhang and Wang (2019)	It contains more microscopic transaction information and can provide higher forecasting performance based on fundamental information.
Predictor selection	LASSO, Elastic network, ridge regression, and penalty regression	Ma et al. (2018), Zhang et al. (2019), Hao et al. (2020), Xing and Zhang (2022), Fu et al. (2024)	Variable selection tools can select more effective predictors by constraining model parameters, thereby improving forecasting performance.

2.2.2 Investor sentiment

In oil price forecasting research, proxy variables for measuring investor sentiment mainly include economic survey results, market variables, and text-based sentiment indices.

Economic survey results primarily originate from data released by government and related agencies. Commonly used indicators include the Economic Policy Uncertainty Index, Consumer Confidence Indices (e.g., those published by the Conference Board and the University of Michigan), and sentiment surveys from the American Association of Individual Investors. Qadan and Nama (2018) construct an investor sentiment proxy for oil market based on these survey data and Google search volume index. While these indicators are typically released on a monthly or quarterly basis by official agencies and are subject to time lags (Li et al., 2019b).

Market variables include factors such as closing price, highest price, lowest price, futures trading volume, oil price fluctuations, speculative trading, and the put call ratio of options. For instance, Deeney et al. (2015) apply principal component analysis (PCA) to extract common signals from trading volume of oil futures, historical volatility of oil prices, put call ratio of oil options, speculative trading, and implied volatility of local stock market indices. These extracted signals are then used as proxies for investor sentiment. He and Casey (2015) develop an emotional endurance index using closing prices to predict the return rate of oil service stocks and crude oil prices. Although market variables are highly objective and reflect multiple economic factors comprehensively, they face inherent limitations in fully characterizing investor sentiment (Li et al., 2022).

With advances in NLP, sentiment analysis techniques based on text data, such as news (Li et al., 2019b) and social media comments (Zhao et al., 2023b; Wu et al., 2024), are widely adopted to measure investor sentiment. For example, Li et al. (2016) employ sentiment analysis techniques based on Henry's (2008) Financial-Specific Dictionary to extract sentiment features from news articles. Wu et al. (2021b) proposes a hybrid prediction model based on CNN and Google Trends data, which significantly improves the accuracy of crude oil price prediction by analyzing online crude oil news text features and market search trends. By leveraging large volumes of text data, sentiment indices are able to accurately captures real-time fluctuations in investor sentiment.

2.2.3 Financial market variables

Crude oil is not only a vital commodity but also a widely traded asset in global financial markets, exhibiting pronounced financial characteristics. As a result, fluctuations in crude oil prices are closely linked to changes in financial markets. A substantial body of research employs financial market variables, such as speculative activity, technical indicators, and direct financial market data, to enhance the accuracy of oil price forecasting models.

Speculative activity is a widely utilized factor in oil price forecasting. Manera et al. (2016) employ three indicators to measure the speculative level in the energy futures market: market share of non-commercial traders, Working's (1960) T-index, and the percentage of non-commercial traders' net long positions to total open contracts in the future market. Guo et al. (2022a) emphasize that climate policy uncertainty and financial speculation have a significant impact on crude oil and natural gas prices, using Working's (1960) T-index as a proxy for financial speculation. However, the role of financial speculation in driving oil prices remains controversial. Morana (2013) applies both Working's (1960) T-index and futures basis to measure speculative shocks in the futures market. This study concludes that financial shocks, such as financial speculation, have been important drivers of oil price fluctuations since 2000. In contrast, Alquist and Gervais (2013) contend that changes in financial firm positions cannot predict changes in oil prices, but changes in oil prices can predict changes in positions.

Some studies have examined the role of technical indicators in oil price forecasting. Yin and Yang (2016) and Wang et al. (2020) argue that technical indicators offer better predictive power for oil prices than traditional economic variables. Yin and Yang (2016) utilize 18 technical indicators based on three trading rules, including moving average rule, momentum rule, and balanced trading volume rule, to predict oil prices. Wang et al. (2020) employ five trading rules i.e., momentum, filtering, moving average, oscillator trading, and support resistance, to generate 105 technical indicators for forecasting oil prices.

Additionally, some studies directly use high-frequency data from various financial markets, including futures, spots, stocks, foreign exchange, and bonds, as predictors for oil prices. These predictors are advantageous due to the high frequency, real-time, and easy access, making them widely used in oil price forecasting. For example, Baumeister et al. (2015), Degiannakis and Filis (2018), and Zhang and Wang (2019) use high-frequency financial data from different markets to predict oil prices. Specifically, Baumeister et al. (2015) utilize high-frequency predictors such as gasoline and crude oil spot price differences, crude oil futures and spot price differences, as well as returns and excess returns of oil company stocks. Degiannakis and Filis (2018) combine fundamental data with high-frequency financial data to predict oil prices, and the results show that high-frequency financial data significantly improves the forecasting accuracy. Similarly, Zhang and Wang (2019) select MSCI World Index,

S&P 500 Index, American Express Oil Index, and FTSE 100 Index as predictors. This study concludes that high-frequency stock market indices are more effective than low-frequency data in predicting monthly crude oil prices.

2.2.4 Predictor selection methods

In the context of big data, the expanding availability of data has led to a substantial increase in the number of predictors for oil price forecasting. However, the relevance and effectiveness of these predictors are often time-varying. Incorporating an excessive number of predictors can lead to overfitting, reduce the robustness of the model, and introduce multicollinearity issues. Consequently, predictor selection plays a crucial role in improving the performance and generalization ability of forecasting models. By selecting the most relevant predictors and constraining unnecessary ones, predictor selection methods help oil price forecasting models adapt better to the time-varying nature of oil price drivers and improve the efficiency of the model.

Least Absolute Shrinkage and Selection Operator (LASSO) and elastic networks are widely used in oil price forecasting for predictor selection and play a crucial role in enhancing forecasting performance. These methods are particularly effective when dealing with high-dimensional data, as they automatically select the most influential predictors while shrinking the coefficients of less relevant variables. Ma et al. (2018) investigate the effectiveness of LASSO regression in predicting oil price fluctuations and find that LASSO can improve the prediction accuracy and economic significance of the model by automatically selecting variables and removing redundant variables. Zhang et al. (2019a) highlight the application of LASSO and elastic networks for oil price forecasting, noting that these methods can effectively identify powerful and complementary factors that contribute to accurate forecasts. Hao et al. (2020) combine a robust loss function with regularization techniques such as LASSO, ridge regression, and elastic networks to predict crude oil prices, demonstrating that the optimized variable set significantly improves forecasting performance. Xing and Zhang (2022) investigate the effectiveness of four shrinkage methods: LASSO, elastic network, ridge regression, and penalty regression based on non-convex penalties and the Huber loss function. This study indicates that careful predictor selection improves forecast accuracy across different time horizons, particularly for long-term oil price forecasting where stability is paramount. Fu et al. (2024) compare various predictors and machine learning models in INE crude oil futures volatility forecasting, and find that ensemble tree models outperform traditional approaches.

3 The main forecasting models for crude oil prices in the context of big data

In the era of big data, oil price forecasting models have evolved from traditional econometric approaches to advanced machine learning techniques. To further enhance prediction accuracy and foresight, hybrid models have emerged by integrating the strengths of diverse methods. Table 4 outlines the forecasting models in oil price forecasting in the context of big data.

3.1 Conventional and data-driven forecasting models

With the rapid development of big data technology, oil price forecasting models have evolved from traditional econometric models to more complex machine learning and deep learning models. Big data provides vast amounts of high-frequency information, which can significantly improve the accuracy of oil price forecasting. However, effectively utilizing the data across different forecasting models remains an important research topic. This section reviews the application of econometric

Table 4 The main models and methods for oil price forecasting in the context of big data

Category	Models and methods	Typical literature	Results
Conventional and Data-driven forecasting models	Econometric models	Mohammadi and Su (2010), Narayan and Narayan (2007), Hou and Suardi (2012), Zhang et al. (2019b), Huang et al. (2009), Zhang and Zhang (2023a), Pan et al. (2017), Zhao (2022), Wu et al. (2023), Salisu et al. (2022), Zhang and Zhang (2023b), Baumeister et al. (2015)	Although econometric models have good interpretability, they lack the ability to predict nonlinear time series.
	Machine learning models	Godarzi et al. (2014), Xu et al. (2023), Jammazi and Aloui (2012), Li et al. (2019a), Fan et al. (2016), Karasu et al. (2020), Ghaffari and Zare (2009), Chiroma et al. (2015), Yu et al. (2016), Wang et al. (2018), Cen and Wang (2019), Urolagin et al. (2021), Wang and Wang (2020), Jiang et al. (2022a), Lin et al. (2022), Mohsin and Jamaani (2023), Liang et al. (2025),	Machine learning models have stronger nonlinear time series processing capabilities than econometric models, and also have better adaptability to adapt to complex crude oil markets.
Hybrid and ensemble forecasting models	Weighting methods	Safari and Davallou (2018), Abdollahi and Ebrahimi (2020), Naser (2016)	Hybrid models can effectively integrate the strengths of individual models, compensate for their respective shortcomings, and thus improve the interpretability and robustness of models.
	Stacking and bagging methods	Zhao et al. (2017), Qin et al. (2023)	

models, traditional machine learning models, and deep learning models in oil price forecasting in the context of big data.

3.1.1 Econometric models

Traditional econometric models continue to play a fundamental role in oil price forecasting. Commonly used models include Autoregressive (AR) models (Guo et al., 2022b), Autoregressive Integrated Moving Average (ARIMA) models (Mohammadi and Su, 2010), Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models (Narayan and Narayan, 2007; Hou and Suardi, 2012; Zhang et al., 2019b), Vector Autoregressive (VAR) models (Huang et al., 2009), Heterogeneous Autoregressive (HAR) models (Zhang and Zhang, 2023a). Due to the solid theoretical foundation and relatively simple structure, traditional econometric models are frequently used as benchmark models and are widely applied in oil price forecasting.

However, traditional econometric models demand uniform data frequency across all inputs. As a result, low-frequency data (quarterly or annual) are commonly employed in modeling and forecasting tasks using traditional econometric models. In contrast, high-frequency data (such as daily and hourly data) are rich in real-time information, which can provide valuable insights into predicting low-frequency variables (such as quarterly and annual oil prices). Yet, incorporating high-frequency data into traditional econometric models poses significant methodological challenges.

To address the limitations of traditional models in handling data of differing frequencies, mixed frequency models have been widely used in oil price forecasting. These models process data at multiple frequencies to improve the accuracy of forecasting. In oil price forecasting, the mixed-data sampling (MIDAS) model is often combined with GARCH to form the GARCH-MIDAS model. Pan et al. (2017) employ a regime-switching GARCH-MIDAS model to study the impact of macroeconomic factors on oil price volatility and find that the multi-regime GARCH-MIDAS model outperforms the single-regime model in out-of-sample forecasting. Zhao (2022) combines the Lasso-adaptive method for variable selection with the GARCH-MIDAS framework and finds that a multi-factor model performs better than a single-factor model in predicting oil price fluctuations. Wu et al. (2023) adopt single factor and two factor GARCH-MIDAS-GPR models to study the impact of geopolitical risk on oil prices and find that geopolitical risk has a positive effect on oil price fluctuations. Salisu et al. (2022) utilize the GARCH-MIDAS framework to examine the forecasting power of global financial cycles on oil market volatility and find a positive correlation between oil market volatility and global financial cycles. Zhang and Zhang (2023b) compare the performance of various GARCH models based on structural changes in predicting crude oil market volatility. This study reveals that flexible Fourier form (FFF) GARCH-type models considering smooth shift were superior to the GARCH model with Markov mechanism switching (MRS) in predicting crude oil price fluctuations and portfolio performance. Besides the GARCH-MIDAS model, Mixed Frequency Vector Autoregressive (MF-VAR) models are another type of frequency-mixing model. Baumeister et al. (2015) find that the forecasting performance of MF-VAR is inferior to both VAR and MIDAS models. As a result, MF-VAR models have seen limited application in oil price forecasting research.

3.1.2 Machine learning models

Most machine learning models are data-driven, and their performance is limited by the scale and quality of available data sets (Zhou et al., 2023). With the advent of big data era, the availability of massive and diverse data sets has greatly enhanced the capabilities of these models. In oil price forecasting, such data-rich environments enable forecasting models to better capture complex nonlinear patterns and temporal dynamics, thereby improving prediction accuracy. The commonly used machine learning models mainly include Artificial Neural Network (ANN), Support Vector Regression (SVR), Backpropagation Neural Network (BPNN), Extreme Learning Machine (ELM), Adaptive Neuro-Fuzzy Inference System (ANFIS), and AdaBoost.

ANN is a foundational neural network model, widely applied in oil price forecasting, due to its ability to learn complex patterns in data by simulating neuron structures. Godarzi et al. (2014) develop a dynamic ANN model to predict oil price movements and demonstrate its superior performance compared to static ANN models. Xu et al. (2023) propose an improved model that can adjust the bias and weight of ANN hidden layers, and the results show that the neural network model is superior to the econometric model in forecasting accuracy.

BPNN extends ANN by applying backpropagation to optimize neural network weights, thereby enhancing nonlinear pattern capture in oil price time series. Jammazi and Aloui (2012) combine multi-layer BPNN and wavelet decomposition to predict crude oil prices and find that the model outperforms BPNN in forecasting performance. Li et al. (2019a) employ a genetic algorithm to optimize BPNN and Support Vector Machine (SVM), and combine VMD to predict WTI and Brent crude oil spot prices.

SVR is a widely used machine learning method for forecasting tasks, particularly effective in capturing nonlinear relationships in time series data. SVR is derived from the SVM framework, which is primarily effective in classification tasks (Zhao et al., 2023a). SVR shows strong generalization ability in crude oil price

forecasting. Fan et al. (2016) apply independent component analysis to decompose oil prices into three independent components, then predict these components using SVR, and finally integrate the results. Karasu et al. (2020) propose a crude oil price forecasting model that combines SVR with wrapper-based feature selection using multi-objective particle swarm optimization. This approach effectively captures the nonlinear and fractal characteristics of crude oil time series, leading to improved forecasting accuracy.

Models based on iterative algorithms, such as ANFIS, have also been used for oil price forecasting. For instance, Ghaffari and Zare (2009) employ ANFIS to predict intraday crude oil price fluctuations and combine ANN and fuzzy logic to predict WTI crude oil spot prices. Similarly, Chiroma et al. (2015) propose a GA-NN model combining neural networks and genetic algorithms for predicting WTI crude oil prices. This study demonstrates that the computational efficiency and forecast accuracy of the model are superior to the benchmark model.

Machine learning models based on iterative algorithms, such as ANN, SVM, and BPNN, outperform traditional econometric models in handling nonlinear time series. However, these models also exhibit drawbacks, including time-consuming, sensitive parameters, slow convergence, and susceptibility to getting stuck in local optima. For example, ANN relies on gradient descent and iterative algorithms to adjust parameters, which can be time-consuming and prone to issues such as local optima and slow convergence. Similarly, SVM depends on grid search methods and iterative learning algorithms to adjust parameters, which also suffers from the drawbacks of long processing time and excessive sensitivity to parameters (Tang et al., 2018; Yu et al., 2016).

To overcome the limitations, non-iterative algorithms such as ELM have been applied. For example, Yu et al. (2016) propose an ensemble learning model that integrates EEMD with extend ELM for predicting WTI crude oil spot prices. This study demonstrates that the model has better forecasting performance than the benchmark model. Wang et al. (2018) combine the ELM with BEMD to analyze the forecasting effect of Internet attention on oil price fluctuations.

Compared with traditional machine learning models, deep learning models can better capture long-term dependencies and potential features in oil price time series. The commonly used deep learning techniques in oil price forecasting models mainly include LSTM, CNN, Gated Recurrent Unit (GRU).

LSTM addresses the gradient vanishing problem in traditional recurrent neural networks (RNNs) by introducing gating mechanisms and performs more stably in long-term sequence forecasting. Cen and Wang (2019) extend the training data based on prior knowledge data transmission algorithm, decompose the oil price time series using ensemble empirical mode decomposition, and finally use LSTM for oil price forecasting. Urolagin et al. (2021) propose a multivariate LSTM model with Markov and Z-score transformations for predicting WTI oil prices. Li et al. (2024) use VMD to decompose crude oil prices into high- and low-frequency components, which are then predicted using the ACI model and iLSTM, respectively, before being combined into the final forecast.

GRU adds Reset Gate and Update Gate on the basis of RNNs (Jiang et al., 2021), making it a simplified version of LSTM that improves the training speed and computational efficiency of the model by reducing the complexity of the gating mechanism. Busari and Lim (2021) construct AdaBoost-LSTM and AdaBoost-GRU models for oil price forecasting and find that AdaBoost-GRU has better forecasting performance. Wang and Wang (2020) apply EMD to decompose oil price time series and then use GRU with stochastic time effective weights. Jiang et al. (2022a) combine sentiment analysis and ensemble analysis and use EEMD and GRU optimized by seagull algorithm to predict oil prices. Liang et al. (2025) propose a GRU-based nonlinear Granger Causality model (GRU-GC) for crude oil price forecasting and causality determination.

CNN uses a multi-level feature extraction structure to identify potential patterns in oil price time series, thereby improving forecast accuracy. Lin et al. (2022) decompose oil prices into high-frequency and low-frequency sequences using WT and then predict the decomposed time series using bidirectional LSTM, attention mechanism, and CNN model. Mohsin and Jamaani (2023) develop a CNN model to predict crude oil prices based on the historical prices of five precious metals. The results show that the deep learning model provides better forecasting performance than econometric models and traditional machine learning models.

3.2 Hybrid and ensemble forecasting models

Hybrid models integrate different types of models to leverage their respective strengths and mitigate the limitations of individual models, thereby improving the accuracy and robustness of oil price forecasting. In oil price forecasting, hybrid models commonly utilize weighting and ensemble methods. The weighting method optimizes the overall forecasting performance by assigning different weights to various sub-models and integrating the forecasting results. The ensemble method further enhances the accuracy and reliability of forecasting by training multiple sub-models and fusing the forecasting results. The selection of appropriate weighting or ensemble methods has a significant impact on the performance of hybrid models.

The weighting method combines the forecasting results of multiple models to improve the overall forecasting

accuracy and robustness. Common weighting methods include time-varying weighting and constant weighting based on genetic algorithm. These methods aim to integrate the strengths of diverse models by optimally allocating weights according to their individual forecasting performance. For example, Safari and Davallou (2018) employ exponential smoothing models, ARIMA, and nonlinear autoregressive neural networks as separate forecasting models, and apply Kalman filtering to determine time-varying weights for the forecasting of these three models. Abdollahi and Ebrahimi (2020) utilize ANFIS, Autoregressive Fractionally Integrated Moving Average, and Markov regime-switching models to predict crude oil prices and integrate the prediction results using three weighting methods: genetic algorithm weighting, equal weights, and error-value-based weights. The results show that the hybrid model based on genetic algorithm weighting is superior to those based on other weighting algorithms. Additionally, Dynamic Model Average (DMA) is another time-varying weighting method that dynamically adjusts model weights and improves forecast accuracy and robustness by combining the outputs of multiple candidate models. Naser (2016) apply DMA to a large data set containing 149 time series variables to predict WTI crude oil prices.

The ensemble method improves overall forecasting performance by combining the outputs of multiple forecasting models. Common ensemble methods include bagging and stacking. These methods typically improve the stability and accuracy of forecasting by constructing multiple sub-models and integrating the outputs. Zhao et al. (2017) employ Stacked Denoising Autoencoders (SDAE) and bagging to predict WTI spot prices. This study indicates that the model has better forecasting performance than SDAE and other benchmark models. Qin et al. (2023) develop an oil price forecasting model that can select indicators from a wider range of online information and has a stacking strategy by introducing Google Trends.

4 Conclusions

In the context of big data, substantial advancements have been achieved in oil price forecasting, particularly with respect to data sources, predictors, and forecasting models. At the data level, the widespread application of multi-source heterogeneous data and high-frequency data provides rich information and valuable new inputs for forecasting modeling. Consequently, a variety of novel data sources—such as social media comments, news texts, search engine queries, satellite imagery, and climate data—have been introduced to expand the range of predictors, significantly enhancing both the accuracy and timeliness of forecasting models. Regarding forecasting models, while traditional econometric models continue to play a vital role, the increasing adoption of machine learning techniques is pivotal in addressing the complex patterns and nonlinearities of oil prices. Additionally, the development of hybrid models, which combine various approaches through weighting or ensemble methods, further enhances forecasting accuracy and robustness by overcoming the limitations of relying on a single model.

Despite notable advancements, current research on oil price forecasting still faces challenges. First, although sentiment analysis has progressed, the integration and utilization of multi-source heterogeneous data remain limited. Data types such as audio and videos have yet to be widely applied in oil price forecasting. Secondly, the supply and demand data for oil prices are often low-frequency and time-delayed, resulting in benchmark models that cannot respond to market fluctuations in a timely manner. Thirdly, most existing research focuses on improving forecasting accuracy, with less emphasis on the robustness and generalization ability of models. Finally, there is a notable gap in research on the practical application of oil price forecasting outcomes. While much of the existing research focuses on improving forecast accuracy, there is insufficient attention on how to enhance model interpretability and operational usability for real-world applications, such as portfolio management, risk control, and policy formulation.

Based on the limitations of existing research, future crude oil price forecasting can be explored from five directions: data, predictors, forecasting models, forecasting results and economic value.

(1) Integrating multi-source heterogeneous data

With the rapid advancement of technologies for processing text, audio, video, and satellite imagery, future research should leverage cross-modal learning to integrate heterogeneous data. This can help to extract richer features relevant to crude oil pricing and enhance forecasting accuracy.

(2) Reconstructing high-frequency supply and demand indicators

Traditional supply–demand predictors suffer from low temporal resolution and lag in capturing real-time market dynamics. The development of big data and IoT technologies has improved access to high-frequency information. Future studies should explore constructing expectation-based, high-frequency supply and demand indicators, using a fusion of textual, visual, and structured data, to improve the timeliness of predictions.

(3) Enhancing hybrid modeling approaches

Hybrid models integrate the strengths of multiple forecasting techniques and exhibit greater adaptability across varying market conditions. Future work can deepen hybrid model design by introducing more flexible architectures and ensemble strategies to better handle complex and volatile markets.

(4) Improving model interpretability

As forecasting models become more complex and involve more predictors, they face challenges like the curse of dimensionality and overfitting. Future research should incorporate advanced dimensionality reduction techniques and integrate interpretability tools to simplify models while preserving predictive power. Balancing accuracy and transparency will support informed decision-making by policymakers and investors.

(5) Improving the economic value of predictions

Future research should go beyond prediction accuracy to explore the practical utility of forecasting models in areas such as risk management, hedging strategies, and policy formulation. By quantifying the role of forecasts in reducing investment risks and supporting intervention decisions, the economic relevance of oil price models can be better demonstrated.

Competing Interests The authors declare that they have no competing interests.

References

Abdollahi H, Ebrahimi S B (2020). A new hybrid model for forecasting Brent crude oil price. Energy, 200: 117520

Alquist R, Gervais O (2013). The role of financial speculation in driving the price of crude oil. Energy Journal, 34(3): 35–54

Bai Y, Li X, Yu H, Jia S (2022). Crude oil price forecasting incorporating news text. International Journal of Forecasting, 38(1): 367–383

Barber B M, Odean T (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies, 21(2): 785–818

Baumeister C, Guérin P, Kilian L (2015). Do high-frequency financial data help forecast oil prices? The MIDAS touch at work. International Journal of Forecasting, 31(2): 238–252

Beyer Díaz S, Coussement K, De Caigny A, Pérez L F, Creemers S (2024). Do the US president's tweets better predict oil prices? An empirical examination using long short-term memory networks. International Journal of Production Research, 62(6): 2158–2175

Busari G A, Lim D H (2021). Crude oil price prediction: A comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. Computers & Chemical Engineering, 155: 107513

Cen Z, Wang J (2019). Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer. Energy, 169: 160–171

Chen W, Lai K K, Cai Y (2021). Exploring public mood toward commodity markets: a comparative study of user behavior on Sina Weibo and Twitter. Internet Research, 31(3): 1102–1119

Chen X, Zhang W, Xu X, Cao W (2022). A public and large-scale expert information fusion method and its application: Mining public opinion via sentiment analysis and measuring public dynamic reliability. Information Fusion, 78: 71–85

Chiroma H, Abdulkareem S, Herawan T (2015). Evolutionary neural network model for West Texas Intermediate crude oil price prediction. Applied Energy, 142: 266–273

Deeney P, Cummins M, Dowling M, Bermingham A (2015). Sentiment in oil markets. International Review of Financial Analysis, 39: 179–185

Degiannakis S, Filis G (2018). Forecasting oil prices: High-frequency financial data are indeed useful. Energy Economics, 76: 388–402

Elshendy M, Colladon A F, Battistoni E, Gloor P A (2018). Using four different online media sources to forecast the crude oil price. Journal of Information Science, 44(3): 408–421

Fan L, Pan S, Li Z, Li H (2016). An ICA-based support vector regression scheme for forecasting crude oil prices. Technological Forecasting and Social Change, 112: 245–253

Fang T, Zheng C, Wang D (2023a). Forecasting the crude oil prices with an EMD-ISBM-FNN model. Energy, 263: 125407

Fang Y, Wang W, Wu P, Zhao Y (2023b). A sentiment-enhanced hybrid model for crude oil price forecasting. Expert Systems with Applications, 215: 119329

Fu T, Huang D, Feng L, Tang X (2024). More is better? The impact of predictor choice on the INE oil futures volatility forecasting. Energy Economics, 134: 107540

Ghaffari A, Zare S (2009). A novel algorithm for prediction of crude oil price variation based on soft computing. Energy Economics, 31(4): 531–536

Godarzi A A, Amiri R M, Talaei A, Jamasb T (2014). Predicting oil price movements: A dynamic Artificial Neural Network approach. Energy Policy, 68: 371–382

Gong X, Guan K, Chen Q (2022). The role of textual analysis in oil futures price forecasting based on machine learning approach. Journal of Futures Markets, 42(10): 1987–2017

Guo J, Long S, Luo W (2022a). Nonlinear effects of climate policy uncertainty and financial speculation on the global prices of oil and gas. International Review of Financial Analysis, 83: 102286

Guo J F, Ji Q (2013). How does market concern derived from the Internet affect oil prices? Applied Energy, 112: 1536–1543

Guo Y, Ma F, Li H, Lai X (2022b). Oil price volatility predictability based on global economic conditions. International Review of Financial Analysis, 82: 102195

Han L, Lv Q, Yin L (2017). Can investor attention predict oil prices? Energy Economics, 66: 547–558

Hao X, Wang Y (2023). Cloud cover and expected oil returns. Humanities & Social Sciences Communications, 10(1): 605

Hao X, Zhao Y, Wang Y (2020). Forecasting the real prices of crude oil using robust regression models with regularization constraints. Energy Economics, 86: 104683

He L T, Casey K M (2015). Forecasting ability of the investor sentiment endurance index: The case of oil service stock returns and crude oil prices. Energy Economics, 47: 121–128

Henry E (2008). Are investors influenced by how earnings press releases are written? The Journal of Business Communication (1973), 45(4): 363–407

Hou A, Suardi S (2012). A nonparametric GARCH model of crude oil price return volatility. Energy Economics, 34(2): 618–626

Hou K, Peng L, Xiong W (2009). A tale of two anomalies: The implications of investor attention for price and earnings momentum. Available at SSRN 976394

Huang D, Yu B, Fabozzi F J, Fukushima M (2009). CAViaR-based forecast for oil price risk. Energy Economics, 31(4): 511–518

Huang J, Ding Q, Zhang H, Guo Y, Suleman M T (2021). Nonlinear dynamic correlation between geopolitical risk and oil prices: A study based on high-frequency data. Research in International Business and Finance, 56: 101370

Huang Y, Deng Y (2021). A new crude oil price forecasting model based on variational mode decomposition. Knowledge-Based Systems, 213: 106669

Jammazi R, Aloui C (2012). Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling. Energy Economics, 34(3): 828–841

Jiang F, Wang K, Dong L, Pan C, Xu W, Yang K (2021). AI driven heterogeneous MEC system with UAV assistance for dynamic environment: Challenges and solutions. IEEE Network, 35(1): 400–408

Jiang H, Hu W, Xiao L, Dong Y (2022a). A decomposition ensemble based deep learning approach for crude oil price forecasting. Resources Policy, 78: 102855

Jiang Z, Zhang L, Zhang L, Wen B (2022b). Investor sentiment and machine learning: Predicting the price of China's crude oil futures market. Energy, 247: 123471

Karasu S, Altan A, Bekiros S, Ahmad W (2020). A new forecasting model with wrapper-based feature selection approach using multiobjective optimization technique for chaotic crude oil time series. Energy, 212: 118750

Li J, Qian S, Li L, Guo Y, Wu J, Tang L (2024). A novel secondary decomposition method for forecasting crude oil price with twitter sentiment. Energy, 290: 129954

Li J, Tang L, Wang S (2020). Forecasting crude oil price with multilingual search engine data. Physica A, 551: 124178

Li J, Xu Z, Yu L, Tang L (2016). Forecasting oil price trends with sentiment of online news articles. Procedia Computer Science, 91: 1081–1087

Li J, Zhu S, Wu Q (2019a). Monthly crude oil spot price forecasting using variational mode decomposition. Energy Economics, 83: 240–253

Li T, Qian Z, Deng W, Zhang D, Lu H, Wang S (2021). Forecasting crude oil prices based on variational mode decomposition and random sparse Bayesian learning. Applied Soft Computing, 113: 108032

Li X, Ma J, Wang S, Zhang X (2015). How does Google search affect trader positions and crude oil prices? Economic Modelling, 49: 162–171

Li X, Shang W, Wang S (2019b). Text-based crude oil price forecasting: A deep learning approach. International Journal of Forecasting, 35(4): 1548–1560

Li Z, Huang Z, Failler P (2022). Dynamic correlation between crude oil price and investor sentiment in China: Heterogeneous and asymmetric effect. Energies, 15(3): 687

Liang Q, Lin Q, Guo M, Lu Q, Zhang D (2025). Forecasting crude oil prices: A Gated Recurrent Unit-based nonlinear Granger Causality model. International Review of Financial Analysis, 102: 104124

Liang W, Chen X, Huang S, Xiong G, Yan K, Zhou X (2023). Federal learning edge network based sentiment analysis combating global COVID-19. Computer Communications, 204: 33–42

Lin Y, Chen K, Zhang X, Tan B, Lu Q (2022). Forecasting crude oil futures prices using BiLSTM-Attention-CNN model with Wavelet transform. Applied Soft Computing, 130: 109723

Loughran T, McDonald B, Pragidis I (2019). Assimilation of oil news into prices. International Review of Financial Analysis, 63: 105–118

Lucey B, Ren B (2021). Does news tone help forecast oil? Economic Modelling, 104: 105635

Ma F, Liu J, Wahab M I M, Zhang Y (2018). Forecasting the aggregate oil price volatility in a data-rich environment. Economic Modelling, 72: 320–332

Manera M, Nicolini M, Vignati I (2016). Modelling futures price volatility in energy markets: Is there a role for financial speculation? Energy Economics, 53: 220–229

Mohammadi H, Su L (2010). International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models. Energy Economics, 32(5): 1001–1008

Mohsin M, Jamaani F (2023). A novel deep-learning technique for forecasting oil price volatility using historical prices of five precious metals in context of green financing–A comparison of deep learning, machine learning, and statistical models. Resources Policy, 86: 104216

Morana C (2013). Oil price dynamics, macro-finance interactions and the role of financial speculation. Journal of Banking & Finance, 37(1): 206–226

Mukherjee A, Panayotov G, Shon J (2021). Eye in the sky: Private satellites and government macro data. Journal of Financial Economics, 141(1): 234–254

Narayan P K, Narayan S (2007). Modelling oil price volatility. Energy Policy, 35(12): 6549–6553

Narayan P K, Ranjeeni K, Bannigidadmath D (2017). New evidence of psychological barrier from the oil market. Journal of Behavioral Finance, 18(4): 457–469

Naser H (2016). Estimating and forecasting the real prices of crude oil: A data rich model using a dynamic model averaging (DMA) approach. Energy Economics, 56: 75–87

Pan Z, Wang Y, Wu C, Yin L (2017). Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model. Journal of Empirical Finance, 43: 130–142

Qadan M, Nama H (2018). Investor sentiment and the price of oil. Energy Economics, 69: 42–58

Qin Q, Huang Z, Zhou Z, Chen C, Liu R (2023). Crude oil price forecasting with machine learning and Google search data: An accuracy comparison of single-model versus multiple-model. Engineering Applications of Artificial Intelligence, 123: 106266

Qu H, Li G (2023). Multi-perspective investor attention and oil futures volatility forecasting. Energy Economics, 119: 106531

Ren X, Jiang W, Ji Q, Zhai P (2024). Seeing is believing: Forecasting crude oil price trend from the perspective of images. Journal of Forecasting, 43(7): 2809–2821

Safari A, Davallou M (2018). Oil price forecasting using a hybrid model. Energy, 148: 49–58

Salisu A A, Gupta R, Demirer R (2022). Global financial cycle and the predictability of oil market volatility: Evidence from a GARCH-MIDAS model. Energy Economics, 108: 105934

Seasholes M S, Wu G (2007). Predictable behavior, profits, and attention. Journal of Empirical Finance, 14(5): 590–610

Sun S, Sun Y, Wang S, Wei Y (2018). Interval decomposition ensemble approach for crude oil price forecasting. Energy Economics, 76: 274–287

Tang L, Dai W, Yu L, Wang S (2015). A novel CEEMD-based EELM ensemble learning paradigm for crude oil price forecasting. International Journal of Information Technology & Decision Making, 14(1): 141–169

Tang L, Wu Y, Yu L (2018). A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting. Applied Soft Computing, 70: 1097–1108

Tang L, Zhang C, Li L, Wang S (2020). A multi-scale method for forecasting oil price with multi-factor search engine data. Applied Energy, 257: 114033

Urolagin S, Sharma N, Datta T K (2021). A combined architecture of multivariate LSTM with Mahalanobis and Z-Score transformations for oil price forecasting. Energy, 231: 120963

Van Eyden R, Difeto M, Gupta R, Wohar M E (2019). Oil price volatility and economic growth: Evidence from advanced economies using more than a century's data. Applied Energy, 233–234: 612–621

Wang B, Wang J (2020). Energy futures and spots prices forecasting by hybrid SW-GRU with EMD and error evaluation. Energy Economics, 90: 104827

Wang J, Athanasopoulos G, Hyndman R J, Wang S (2018). Crude oil price forecasting based on internet concern using an extreme learning machine. International Journal of Forecasting, 34(4): 665–677

Wang J, Li X (2018). A combined neural network model for commodity price forecasting with SSA. Soft Computing, 22(16): 5323–5333

Wang Y, Liu L, Wu C (2020). Forecasting commodity prices out-of-sample: Can technical indicators help? International Journal of Forecasting, 36(2): 666–683

Wen F, Gong X, Cai S (2016). Forecasting the volatility of crude oil futures using HAR-type models with structural breaks. Energy Economics, 59: 400–413

Wong T C, Chan H K, Lacka E (2017). An ANN-based approach of interpreting user-generated comments from social media. Applied Soft Computing, 52: 1169–1180

Working H (1960). Speculation on hedging markets. Food Research Institute Studies, 1(2): 185–220

Wu B, Wang L, Lv S X, Zeng Y R (2021b). Effective crude oil price forecasting using new text-based and big-data-driven model. Measurement, 168: 108468

Wu B, Wang L, Wang S, Zeng Y R (2021a). Forecasting the US oil markets based on social media information during the COVID-19 pandemic. Energy, 226: 120403

Wu J, Zhao R, Sun J, Zhou X (2023). Impact of geopolitical risks on oil price fluctuations: Based on GARCH-MIDAS model. Resources Policy, 85: 103982

Wu W, Xu M, Su R, Ullah K (2024). Modeling crude oil volatility using economic sentiment analysis and opinion mining of investors via deep learning and machine learning models. Energy, 289: 130017

Xing L M, Zhang Y J (2022). Forecasting crude oil prices with shrinkage methods: Can nonconvex penalty and Huber loss help? Energy Economics, 110: 106014

Xu Z, Mohsin M, Ullah K, Ma X (2023). Using econometric and machine learning models to forecast crude oil prices: Insights from economic history. Resources Policy, 83: 103614

Yao T, Zhang Y J (2024). The impact of air pollution on crude oil futures market. Journal of Futures Markets, 44(6): 1055–1068

Yao T, Zhang Y J, Ma C Q (2017). How does investor attention affect international crude oil prices? Applied Energy, 205: 336–344

Yin L, Yang Q (2016). Predicting the oil prices: Do technical indicators help? Energy Economics, 56: 338–350

Yu L, Dai W, Tang L (2016). A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. Engineering Applications of Artificial Intelligence, 47: 110–121

Yu L, Zhao Y, Tang L (2017). Ensemble forecasting for complex time series using sparse representation and neural networks. Journal of Forecasting, 36(2): 122–138

Yuan Y (2015). Market-wide attention, trading, and stock returns. Journal of Financial Economics, 116(3): 548–564

Zhang S, Luo J, Wang S, Liu F (2023). Oil price forecasting: A hybrid GRU neural network based on decomposition–reconstruction methods. Expert Systems with Applications, 218: 119617

Zhang T, Tang Z, Wu J, Du X, Chen K (2021). Multi-step-ahead crude oil price forecasting based on two-layer decomposition technique and extreme learning machine optimized by the particle swarm optimization algorithm. Energy, 229: 120797

Zhang Y, Ma F, Wang Y (2019a). Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors?. Journal of Empirical Finance, 54: 97–117

Zhang Y J, Wang J L (2019). Do high-frequency stock market data help forecast crude oil prices? Evidence from the MIDAS models. Energy Economics, 78: 192–201

Zhang Y J, Yao T, He L Y, Ripple R (2019b). Volatility forecasting of crude oil market: Can the regime switching GARCH model beat the single-regime GARCH models? International Review of Economics & Finance, 59: 302–317

Zhang Y J, Zhang H (2023a). Volatility forecasting of crude oil futures market: Which structural change-based HAR models have better performance? International Review of Financial Analysis, 85: 102454

Zhang Y J, Zhang H (2023b). Volatility forecasting of crude oil market: which structural change based GARCH models have better performance? Energy Journal, 44(1): 175–194

Zhao J (2022). Exploring the influence of the main factors on the crude oil price volatility: An analysis based on GARCH-MIDAS model with Lasso approach. Resources Policy, 79: 103031

Zhao J, Hosseini S, Chen Q, Jahed Armaghani D (2023a). Super learner ensemble model: A novel approach for predicting monthly copper price in future. Resources Policy, 85: 103903

Zhao L T, Xing Y Y, Zhao Q R, Chen X H (2023b). Dynamic impacts of online investor sentiment on international crude oil prices. Resources Policy, 82: 103506

Zhao Y, Li J, Yu L (2017). A deep learning ensemble approach for crude oil price forecasting. Energy Economics, 66: 9–16

Zhou X, Zheng X, Cui X, Shi J, Liang W, Yan Z, Yang L T, Shimizu S, Wang K I (2023). Digital twin enhanced federated reinforcement learning with lightweight knowledge distillation in mobile networks. IEEE Journal on Selected Areas in Communications, 41(10): 3191–3211