Based on GCT-RFE Modeling Research on Driving Mechanisms of Crude Oil Prices

Guangjie Liu, Junjie Zhao, Jinping Zhang *

School of Mechanical and Power Engineering, Shenyang University of Chemical Technology, Shenyang, China, 110142

* Corresponding Author Email: zhangjinping140206@syuct.edu.cn

Abstract. This paper discusses the shortcomings of traditional econometric methods in analyzing the nonlinear characteristics of the oil market, and proposes a new strategy based on AI and machine learning. We construct an analytical framework covering eight dimensions of finance, macroeconomics, inventory, policy uncertainty, technical indicators, supply and demand, and geopolitics, involving a total of 52 predictors. Through ADF test and differential processing, we ensure the stability of the data. Using the two-stage model, we screened out the key factors in 52 indicators through GCT, and further screened out 9 main indicators by RFE algorithm. Through KFold cross validation, we verify the robustness of this model and prove that it has a high degree of accuracy in predicting future oil prices and market fluctuations.

Keywords: Crude Oil Prices, Drivers, Granger Causality Test, Two-stage Model.

1. Introduction

The international crude oil market is currently facing a challenging VUCA era, marked by high volatility, uncertainty, complexity, and ambiguity. Ha et al. used the factor-augmented vector autoregressive model to assess global and domestic factors influencing oil prices, concluding that global shocks have a greater impact on inflation [1]. Huang et al. developed a 2SVBMA model that extended oil-related features to external variables, demonstrating good robustness and predictive performance [2]. Obadi et al. used Granger causality to examine the relationship between oil price, drilling quantity, inventory changes, and speculators, finding a bidirectional causality between oil prices and speculators [3].

Ha et al.'s focus on domestic shocks limits their analysis, while He et al.'s simplistic model lacks sufficient treatment of nonlinearities, reducing its applicability [1]-[4]. Both studies overlooked key predictors. In contrast, Cheng et al.'s GCT-SFA two-stage model offers valuable insights but becomes complex with many variables [5].

This study uses daily WTI crude oil price data from January 1, 2010, to December 31, 2023, and introduces 52 predictive indicators. By developing a multi-indicator framework, it aims to identify robust correlations driving crude oil prices. The GCT-RFE model improves Cheng's approach by filtering essential drivers with RFE, enhancing simplicity and interpretability, while integrating core drivers into a machine learning model for prediction.

2. Data description and preliminary analysis

2.1. Supply-and-demand balance

In the international crude oil market, demand is primarily driven by global economic activities. As the world economy grows, particularly with industrialization and urbanization, energy demand, especially for crude oil, has significantly increased [6]. Crude oil supply is influenced by factors such as political stability in producing countries, production changes, and technological advancements. Political turmoil in the Middle East, home to the largest oil reserves, notably impacts global supply. For example, OPEC production and non-OPEC oil output directly affect the global supply [6].The World crude oil supply and demand indicators are shown in Table.1.

Table 1. World crude oil supply and demand indicators Table

Level 1 indicators	Level 2 indicators	Symbol	Author	Source
Demand (economics)	USA Real Estate Index	Y_1	Yi et al. (2023)	www.eia.gov/
	EU GPA price index	Y_2	Liu et al. (2023)	Wind Data Service
	USA People's Consumer Price Index	Y_3	Drachal et al. (2023)	www.eia.gov/
	USA crude oil consumption	Y_4	Lu et al. (2023)	www.fred.stlouisfed.org/
	USA GPA index	Y_5	Yi et al. (2018)	www.fred.stlouisfed.org/
	USA Energy Producer Price Index	Y_6	Nonejad et al. (2020)	Wind Data Service
	Crude oil consumption by international organizations	Y_7	Drachal et al. (2023)	www.eia.gov/
	USA People's Energy Consumption Price Index	Y_8	Drachal et al. (2023)	Wind Data Service
Supply (as in supply and demand)	USA crude oil product supply	Y_{11}	Xu et al. (2023)	www.eia.gov
	USA shale oil and gas production	Y_{12}	Kilian et al. (2013)	
	World crude oil production	Y_{13}	Xu et al. (2023)	
	USA crude oil production	Y_{14}	Xu et al. (2023)	
	Non-OPEC crude oil production	Y_{15}	Li et al. (2019a)	
	OPEC crude oil production	Y_{16}	Drachals et al. (2017)	
	USA Refinery utilization rate	Y_{17}	Ma et al. (2013)	
	Port Henry Natural Gas Spot Prices	Y_{18}	Yi et al. (2013)	

2.2. Stock levels

Changes in inventory levels are an important predictor of crude oil prices[7]. OECD crude oil inventories, total crude oil inventories of USA, strategic stocks and commercial stocks are key data for assessing the supply and demand situation in the global and USA crude oil markets.

2.3. Financial market developments

The increasing diversity of participants in financial markets, including investment banks, hedge funds, commodity index funds, etc., makes crude oil price movements more complex and unpredictable[8]. The Financial market indicators are shown in Table.2.

Table 2. Collection of financial market-type indicators for world crude oil prices

Level 1 indicators	Level 2 indicators	Symbol	Author	Source
Financial market	Gold Closing Price Data	Y_{23}	Liu et al. (2023b)	Wind Data Service
	Copper Closing Price Data	Y_{24}	Baumeister et al. (2019)	Wind Data Service
	SSE Composite Data	Y_{25}	Lu et al. (2023)	Wind Data Service
	US/European Currency Rates	Y_{26}	Chang et al. (2023)	https://stlouifeds.org/econ/
	US Dollar Data	Y_{27}	Chang et al. (2023)	Wind Data Service
	Trade Dollar Data	Y_{28}	Ma et al. (2019)	www.cboer.com/micro/vix/
	Hang Seng Data	Y_{29}	Chang et al. (2023)	Wind Data Service
	USA Currency Stock	Y_{30}	Lu et al. (2023)	Wind Data Service
	S&P 500 Data	Y_{31}	Drachl et al. (2023)	https://stooqs.com
	VIX Data	Y_{32}	Drachl et al. (2023)	https://stooqs.com
	Dow Jones Industrial Data	Y_{33}	Xu et al. (2023)	Wind Data Service
	NASDAQ Data	Y_{34}	Xu et al. (2023)	Wind Data Service
	Non-Energy Commodity Data	Y_{35}	Li et al. (2023)	www.worldbank.org/

2.4. Macroeconomic situation

Macroeconomic factors play an integral role in analyzing the crude oil market and its price trends. Recent studies, such as the work of Balcilar M et al. (2017), have emphasized the importance of the macroeconomic environment in forecasting crude oil prices [9]. The Macroeconomic indicators are shown in Table.3.

Table 3. Collection of macroeconomic indicators of world crude oil prices

Level 1 indicators	Level 2 indicators	Symbol	Author	Source
Macro-economic	LIBOR Data	Y_{36}	Chang et al. (2018)	https://research.stlouifeds.org/econ/
	USA Unemployment Data	Y_{37}	Nonejd et al. (2020)	https://research.stlouifeds.org/econ/
	USA Short-term Interest Rate Data	Y_{38}	Xiu et al. (2018)	https://research.stlouifeds.org/econ/
	Government Bond Yield Data	Y_{39}	Drachl et al. (2016)	https://research.stlouifeds.org/econ/
	Non-Agricultural Workers	Y_{40}	Miao et al. (2017)	https://research.stlouifeds.org/econ/
	USA Inflation Rate	Y_{41}	Xiu et al. (2018)	www.stats.oecd.org
	USA National Debt Interest Rate	Y_{42}	Chang et al. (2018)	https://research.stlouifeds.org/econ/
	USA Manufacturing Capacity Utilization Data	Y_{43}	Nonejd(2020)	https://research.stlouifeds.org/econ/
	USA Oil and Gas Extraction Capacity Utilization Data	Y_{44}	Chang et al. (2019)	https://research.stlouifeds.org/econ/
	Chicago Fed National Activity Data	Y_{45}	Yi et al. (2018)	https://research.stlouifeds.org/econ/
	Price-Earnings Ratio Data	Y_{46}	Chang et al. (2018)	https://research.stlouifeds.org/econ/
	University of Michigan Inflation Expectations Index	Y_{47}	Yi et al. (2018)	https://research.stlouifeds.org/econ/
	USA Industrial Production Data	Y_{48}	Nonejd et al. (2020)	https://research.stlouifeds.org/econ/
	USA Crude Oil Industrial Production Data	Y_{49}	Welchs et al. (2008)	www.hecl.unil.ch/agoyal/
	Book-to-market ratio	Y_{50}	Welchs et al. (2008)	www.hecl.unil.ch/agoyal/

2.5. Economic policy uncertainty

In recent years, there has been a significant increase in economic policy uncertainty, which is reflected in various aspects from tax policy, trade policy to monetary policy [10]. The Indicators in the category of economic policy uncertainty are shown in Table.4.

Table 4. Collection of indicators for the economic policy uncertainty category of world crude oil

Level 1 indicators	Level 2 indicators	Symbol	Author	Source
EPU	USA EPU Data	Y_{41}	Lei et al. (2023)	www.policyunscertaintys.com
	Russia EPU Data	Y_{42}	Lei et al. (2023)	www.policyunscertaintys.com
	United Kingdom EPU Data	Y_{43}	Lei et al. (2023)	www.policyunscertaintys.com
	China EPU Data	Y_{44}	Lei et al. (2023)	www.policyunscertaintys.com
	Canada EPU Data	Y_{45}	Lei et al. (2023)	www.policyunscertaintys.com
	Japan EPU Data	Y_{46}	Lei et al. (2023)	www.policyunscertaintys.com
	EPU-global-GDP	Y_{47}	Lei et al. (2023)	www.policyunscertaintys.com
	EPU-global-PPP	Y_{48}	Lei et al. (2023)	www.policyunscertaintys.com

2.6. Geopolitical risks

Geopolitical risk plays a crucial role in the crude oil market because the production, transportation, and consumption of crude oil are widely distributed globally and these activities are often vulnerable to geopolitical conflicts and tensions [11].

2.7. Technological advances

Technological advancements are reshaping the crude oil market. Improved extraction has made unconventional resources viable, increasing supply, efficiency, and reducing costs, affecting global prices. Advances in energy consumption technology are also shifting demand, potentially reducing reliance on traditional oil. Key indicators are shown in Table.5.

Table 5. Collection of indicators in the category of technological progress in world crude oil prices

Level 1 indicators	Level 2 indicators	Symbol	Author	Source
Technical indicators	Difference between futures price and WTI crude oil spot	Y_{50}	Bameister et al. (2016)	www.eia.gov
	Difference between Brent Crude Oil Spot and WTI	Y_{51}	Bameister et al. (2016)	
	Difference between WTI Crude Oil Spot and Gasoline	Y_{52}	Bameister et al. (2016)	

Based on the above 7 dimensions a structural relationship is obtained as shown in Figure.1.

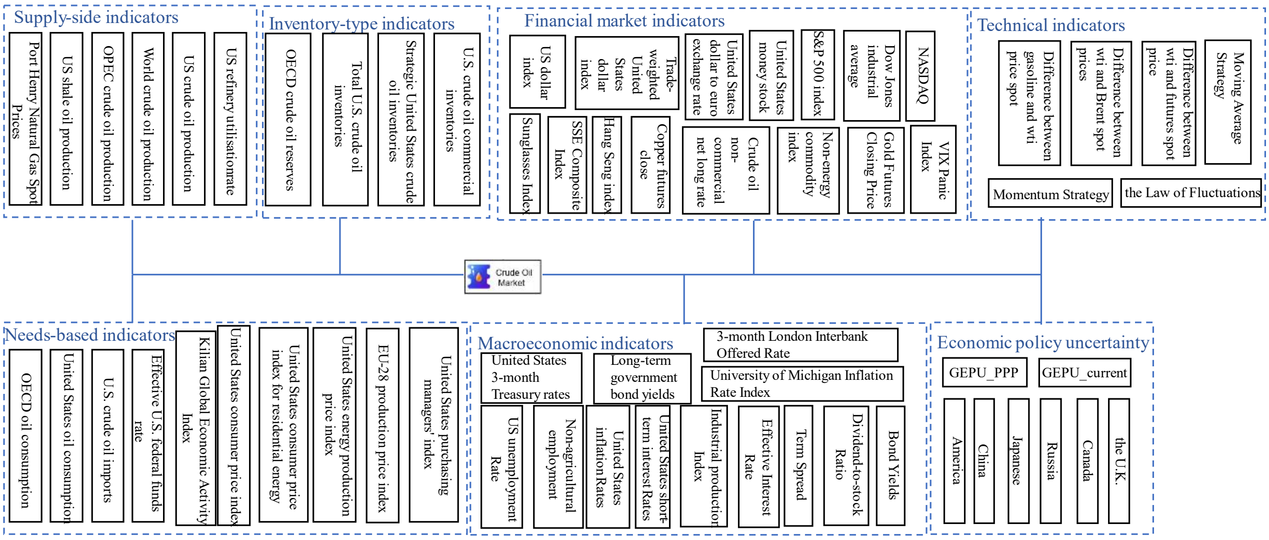

Figure 1. Structure of oil price forecasting indicators

3. Design and construction of the GCT-RFE model

To understand the complexity of crude oil price fluctuations, it is essential to identify the key driving factors. Building on the multi-index framework established in Section 2, this section further examines which factors offer the highest predictive power and information value for crude oil price prediction [5].

3.1. Two-stage model design based on efficiency

This paper proposes a two-stage crude oil price prediction model. Eight-dimensional predictors are integrated with WTI crude oil price as the dependent variable. First, a Granger causality test filters predictors with causal relationships, removing weak indicators. Then, a gradient boosting regression tree with recursive feature elimination selects the most accurate drivers, reducing overfitting and improving accuracy. Model effectiveness is evaluated using NMSE and K-fold cross-validation. The research framework is shown in Figure.2.

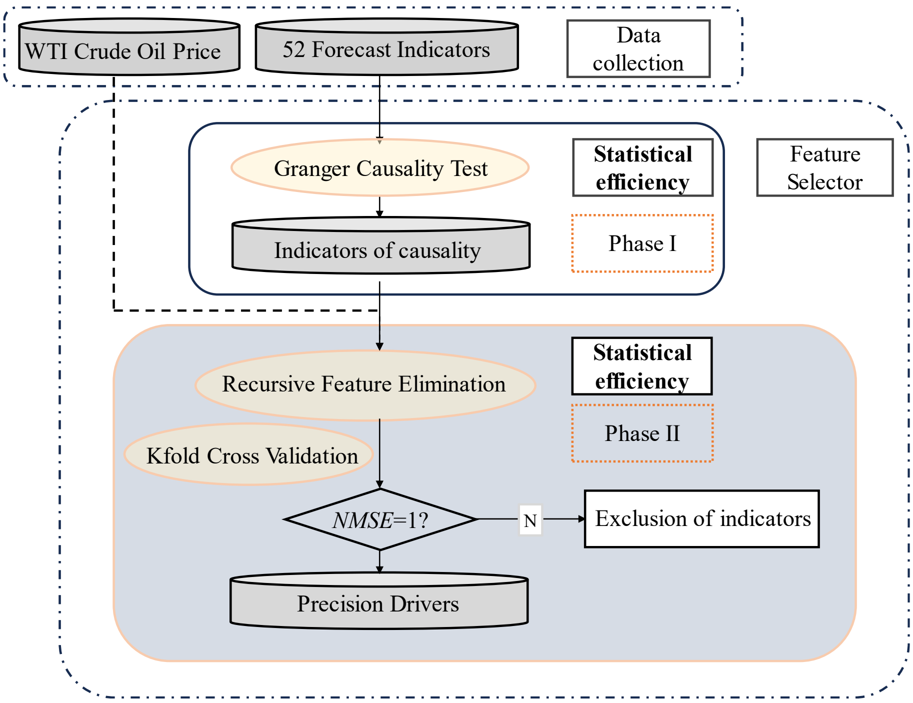

Figure 2. Efficiency-based two-stage modeling framework

The Granger Causality Test (GCT) compares the predictive ability of two models: one using only the target variable's historical values, and the other including both the target and potential causal variables. If the model with the causal variables significantly improves prediction, these variables are considered Granger causes of the target variable [12]. The relationship is expressed by the following equation:

$$Y_t = \beta_0 + \sum_{i=1}^{m} \beta_i Y_{t-1} + \sum_{i=1}^{m} \alpha_i X_{t-1} + \mu_t \tag{1}$$

$$X_t = \delta_0 + \sum_{i=1}^{m} \delta_i X_{t-1} + \sum_{i=1}^{m} \lambda_i Y_{t-1} + \upsilon_t \tag{2}$$

In testing Granger causality between *X* (predictor) and *Y* (WTI crude oil spot price), if $\sum_{i=1}^{m} \alpha_i = 0$ and $\sum_{i=1}^{m} \lambda_i = 0$, then *X* and *Y* are independent; otherwise, a bidirectional effect exists.

Recursive Feature Elimination (RFE) identifies key features by iteratively reducing feature count, enhancing model performance and simplifying interpretation, particularly useful for high-dimensional data where some features may be irrelevant or redundant. The classic RFE algorithm uses a linear kernel with ordered coefficients.

$$\begin{cases} Rank(i) = (w_i)^2 \\ w = \sum_{i=1}^{l} \alpha_i x_i y_i \end{cases} \tag{3}$$

When nonlinear, Guyon assumes that the value of α in the second planning remains unchanged, i.e., the classifier remains unchanged, after a feature is removed from the training sample matrix. When this assumption is satisfied, the feature ranking factor is:

$$\begin{cases} Rank(\,i) = \frac{1}{2}\alpha^T Q\alpha - \frac{1}{2}\alpha^T Q(-i)\alpha \\ Q_{ij} = K\left(x_i \cdot x_j\right) = \Phi(x_i)^T \Phi\left(x_j\right) \end{cases} \tag{4}$$

3.2. Data Selection and Preprocessing

This study uses daily data from January 1, 2010, to December 31, 2023, as high-frequency data captures more detailed information on volatility. Daily data aligns with the frequency of most selected indicators, which are updated daily, ensuring that short-term volatility details are not overlooked. For variables with lower frequencies, the data is adjusted by populating daily entries with the same observation across the relevant period (e.g., weekly data is applied daily within that week).

4. Experimental Results

4.1. Granger causality test

By using regression analysis, the model that includes only the lag term of Y and the model that includes both the lag term of Y and the lag term of X are built separately. Table.6 shows the results obtained by Python.

Table 6. Results of Granger causality test

Variable	Min p-value
US Oil Consumption	0.097423073
US Oil Imports	0.041363106
US Effective Federal Funds Rate	0.176596257
US Consumer Price Index	0.095693509
US Producer Price Index	0.000000073
US Shale Gas Reserves	0.610880787
US Crude Oil Production	0.074963873
US Refinery Utilization Rate	0.091298146
US Crude Oil Inventory	0.000011952
US Commercial Crude Oil Inventory	0.015379273
US Strategic Crude Oil Inventory	0.006592655
Broad Dollar Index	0.012561727
Dollar to Euro Exchange Rate	0.028574023
US M2 Money Supply	0.000013647
VIX Volatility Index	0.002443694
NYMEX Crude Oil Futures Price	0.000000000
US Unemployment Rate	0.003451076
US Non-farm Employment	0.002413778
US Inflation Rate	0.000037160
US Short-term Interest Rate	0.176596257
Long-term Government Bond Yield	0.000010474
US Industrial Production Index	0.018151287
US Manufacturing Capacity Utilization	0.069578965
US Oil and Gas Extraction Capacity Utilization	0.006593104
US EPU Index	0.522217027
China EPU Index	0.052345832
Japan EPU Index	0.343451509
Russia EPU Index	0.028012502
Canada EPU Index	0.075519351
UK EPU Index	0.000087700
Global EPU Index	0.051397846

4.2. Recursive Feature Elimination Algorithm Based on Gradient Boosted Regression Trees and Cross Validation

To optimize the feature set based on the previous Granger causality test, this paper uses a recursive feature elimination algorithm with gradient regression trees to identify the best subset of features. Cross-validation is applied to evaluate the model and compare results after each recursion to find the optimal solution [8].

Gradient boosting trees, proposed by Friedman in 1999, improve model performance by iteratively adding weak learners (usually decision trees) and fitting the residuals of previous models. The model update formula is:

$$F_{m+1}(x) = F_m(x) + \nu \cdot h_m(x) \tag{5}$$

Cross-validation is a technique for evaluating the generalisation performance of a model by dividing the dataset into a smaller number k of subsets and then using these subsets one by one as the validation set and the remaining k-1 subsets as the training set.

For each round, the model is trained on the training set and evaluated on the validation set, and the average of the k rounds of evaluation metrics is calculated as the model performance metric. In this paper, negative mean square error (NMSE) is used for scoring with the formula:

$$NMSE = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

4.3. Setting of model parameters

In the gradient boosted tree regression model, the weak learner was set to 100 trees, with a learning rate of 0.1, tree depth of 3, and a minimum of 2 samples required to split internal nodes. The loss function used was least squares error. The KFold cross-validation method with 5-fold validation and random data shuffling was applied. The recursive feature elimination algorithm removed one feature per iteration, and model performance was evaluated using the NMSE value, where a value closer to 0 indicates better performance.

4.4. Empirical evidence of the quadratic screening model

The number of optimal features obtained from the results of the cross-validation experiments is nine, as shown in Figure.3. The best subset of features is finally selected, as shown in Table 7.

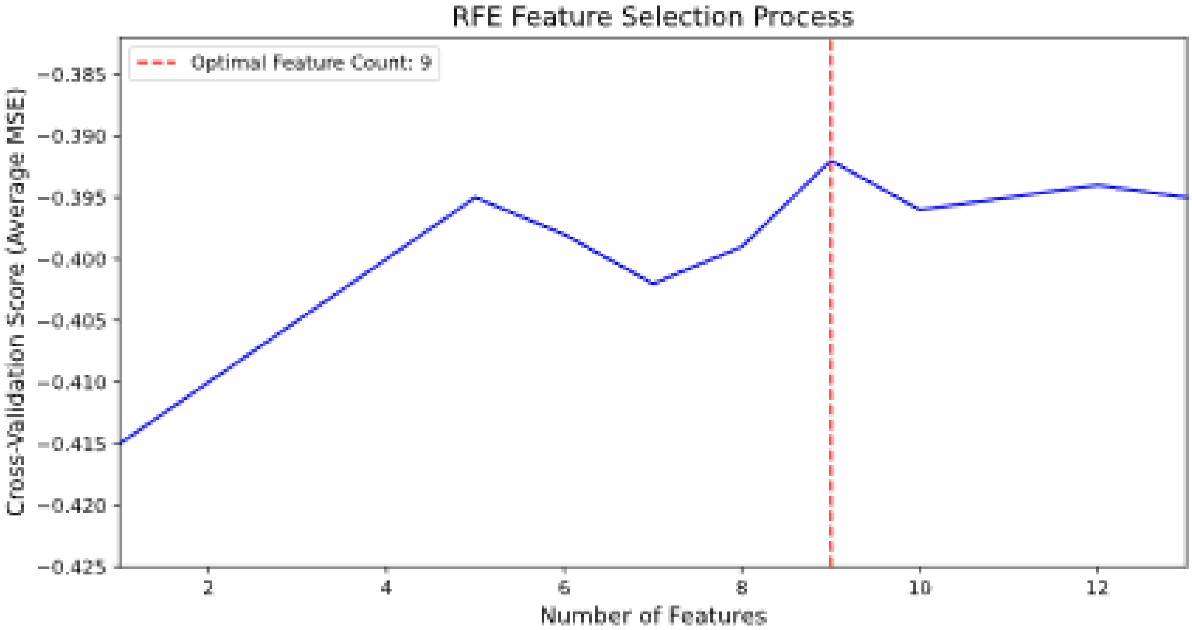

Figure 3. Cross-validation to select the number of features

Table 7. Subset of best features

Importance Ranking	Indicators
1	NYMEX Crude Oil Futures Price
2	VIX Panic Index
3	Broad Dollar Index
4	UK EPU Index
5	USA Crude Oil Inventories
6	USA Crude Oil Commercial Stocks
7	USA Inflation Rate
8	USA Producer Price Index
9	Non-Agricultural Employment

5. Conclusions

Crude oil serves as both a critical resource for development and a national strategic asset. Amid shifting global energy demands and the economic impact of oil price fluctuations on China, this paper introduces a combination model to identify and forecast key drivers of crude oil prices. The GCT-RFE-based model shows robust predictive performance, offering valuable insights for energy policy adjustments.

The model's practical applicability is significant, as it can be used to inform energy pricing strategies, optimize inventory management, and guide investment decisions in the oil and gas industry. Additionally, it helps policymakers better anticipate the economic implications of price changes, enhancing their ability to respond proactively.

While promising, future research could explore advanced feature selection algorithms, model averaging techniques for greater accuracy, and machine learning with extensive data to enhance prediction precision. The GCT-RFE model’s two-stage approach is particularly effective for high-dimensional data, enabling applications like quantifying crude oil inventory impacts and informing recovery management post-economic events. Accurate oil price forecasts can support governments and businesses in managing energy supply chains, preparing for volatility, and improving long-term strategic planning.

References

[1] Ha J, Kose M A, Ohnsorge F, et al. Understanding the global drivers of inflation: How important are oil prices? [J]. Energy Economics, 2023, 127: 107096.

[2] Huang B, Sun Y, Wang S. A new two-stage approach with boosting and model averaging for interval-valued crude oil prices forecasting in uncertainty environments [J]. Frontiers in Energy Research, 2021, 9: 707937.

[3] Obadi S M, Korcek M. The crude oil price and speculations: Investigation using granger causality test [J]. International Journal of Energy Economics and Policy, 2018, 8 (3): 275-282.

[4] He Y, Han A, Hong Y, et al. Forecasting crude oil price intervals and return volatility via autoregressive conditional interval models [J]. Econometric Reviews, 2021, 40 (6): 584-606.

[5] Cheng X, Wu P, Liao S S, et al. An integrated model for crude oil forecasting: Causality assessment and technical efficiency [J]. Energy Economics, 2023, 117: 106467.

[6] Indupurnahayu I, Setiawan E B, Agusinta L, et al. Changes in demand and supply of the crude oil market during the COVID-19 pandemic and its effects on the natural gas market [J]. International Journal of Energy Economics and Policy, 2021, 11 (3): 1-6.

[7] Chen Y C, Huang W C. Constructing a stock-price forecast CNN model with gold and crude oil indicators [J]. Applied Soft Computing, 2021, 112: 107760.

[8] Sun C, Min J, Sun J, et al. The role of China's crude oil futures in world oil futures market and China's financial market [J]. Energy Economics, 2023, 120: 106619.

[9] Balcilar M, Bekiros S, Gupta R. The role of news-based uncertainty indices in predicting oil markets: a hybrid nonparametric quantile causality method [J]. Empirical Economics, 2017, 53: 879-889.

[10] Su C W, Huang S W, Qin M, et al. Does crude oil price stimulate economic policy uncertainty in BRICS? [J]. Pacific-Basin Finance Journal, 2021, 66: 101519.

[11] Li S, Tu D, Zeng Y, et al. Does geopolitical risk matter in crude oil and stock markets? Evidence from disaggregated data [J]. Energy Economics, 2022, 113: 106191.

[12] Masaadeh A H, Dundar B, Samuelson M I, et al. Bilateral Somatically Derived Germ Cell Tumors With 12p Gains Arising in High-grade Serous Carcinomas of the Ovary: A Case Report and Review of the Literature [J]. International Journal of Gynecological Pathology, 2024: 10.1097.