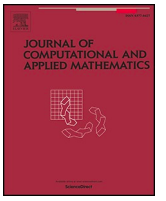
Journal of Computational and Applied Mathematics 474 (2026) 117006

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam



Deep learning enhanced energy market prediction: A robust ARIMAX–LSTM fusion for crude oil pricing

Ahmet Akusta [a], Hasan Hüseyin Yıldırım [b], Musa Gün [c,*], Şakir Sakarya [d]

[a] Rectorate, Konya Technical University, Konya, , Türkiye
[b] Faculty of Applied Sciences, Balikesir University, Balikesir, Türkiye
[c] The Faculty of Economics and Administrative Sciences, Recep Tayyip Erdogan University, Rize, Türkiye
[d] The Faculty of Economics and Administrative Sciences, Balikesir University, Balikesir, Türkiye

ARTICLE INFO

Keywords:
Deep learning
Energy markets
Forecasting
Crude oil prices
ARIMAX
LSTM
Robustness testing

ABSTRACT

Crude oil is a highly strategic global resource, and price fluctuations significantly impact nearly all economic sectors. Therefore, accurate forecasting of its prices is essential for better financial stability and decision-making. This study aims to develop a robust model using monthly data from April 2004 to January 2024 to predict the price of crude oil. We propose a novel approach that blends ARIMAX and LSTM models using a weighted combination to leverage the strengths of econometric and machine learning methods. Unlike hybrid models, which are solely designed based on a decomposition-optimization structure, in our model, an explicit ensemble with weights via grid searching is used to enhance the model's flexibility and performance. As ARIMAX is more efficient in dealing with linear relationships and exogenous variables, LSTM performs much better and effectively captures nonlinear patterns and long-range dependence. Weight hyperparameter tuning and cross-validation help reduce the risk of overfitting or underfitting in the model. Our empirical results indicate that the LSTM model provides a powerful forecasting baseline. The weighted ensemble model offers a marginal improvement on the chronological test set, and the Diebold-Mariano test confirms this advantage is statistically significant. Cross-validation reveals the standalone LSTM to be highly robust, highlighting the importance of component model selection. This study contributes to a more sophisticated framework for risk assessment in energy policy by revealing the crucial trade-off between a model's period-specific accuracy and its general robustness.

1. Introduction

Energy has long been a key driver of global change and remains essential for sustainable growth and economic development. Among energy sources, crude oil plays a particularly critical role, with West Texas Intermediate (WTI) serving as a major global commodity that significantly affects nearly all sectors of the economy. Given its central role in the global economy, WTI crude oil is widely regarded as a benchmark for pricing in both academic research and financial markets. Due to its far-reaching influence, particularly in energy-dependent countries [1], fluctuations in oil prices significantly shape policy decisions. Hence, forecasting the value of this critical asset, which has direct implications for economic stability, has become a focal point of scholarly and financial

* Corresponding author.
E-mail address: musa.gun@erdogan.edu.tr (M. Gün).

https://doi.org/10.1016/j.cam.2025.117006
Received 28 April 2025; Received in revised form 19 July 2025;
Available online 5 August 2025

inquiry.

Government policies tend to prioritize economic indicators due to their strong association with financial and social stability, which in turn reduces uncertainty and enhances overall well-being [2]. However, the substantial volatility of oil prices, compounded by the influence of numerous interrelated factors, renders accurate forecasting highly challenging. Geopolitical tensions, macroeconomic fluctuations, and evolving environmental policies often converge in complex ways, significantly disrupting the balance of supply and demand. Volatility in energy prices can pose significant challenges for national economies [3]. At the same time, these critical variables often exhibit complex, nonlinear interactions that traditional models struggle to capture, resulting in significant forecasting limitations with potentially far-reaching economic consequences. Therefore, the need for accurate forecasting models that capture market dynamics has become more important.

Recent global crises such as the COVID-19 pandemic, the Russia–Ukraine war, and the Israel–Palestine conflict have further exacerbated the fragility of global energy systems. Approximately 40 % of global oil production originates from regions characterized by heightened geopolitical risk. This makes it increasingly difficult for societies to access energy at affordable, stable, and uninterrupted prices. Fluctuations in energy prices adversely impact production costs and complicated budget planning for a wide range of economic agents, thereby underscoring the critical importance of accurate oil price forecasting.

According to Table 1, the world's energy supply increased from 536 EJ in 2010 to 642 EJ in 2023. Oil remains the most consumed energy source with 192 EJ, accounting for 30 % of the total. Fossil fuels still dominate global energy use, comprising 80 % of the primary energy supply, while renewables account for only 12 %. If current policies persist, projections show that oil will continue to hold the highest share (except renewables) in 2050, indicating its continued dominance.

Table 2 shows that the USA is both the largest oil producer and consumer, followed by countries like Saudi Arabia, Russia, China, and India. The top 10 oil-producing countries meet 73 % of the global output, while the top 10 consuming countries account for 61 % of demand. Changes in oil prices affect the economies of both oil-exporting and oil-importing countries. While rising oil prices represent a revenue increase for oil exporters, they also represent a cost factor for oil importers [6]. This highlights the strategic importance of WTI crude oil in global energy trade and macroeconomic forecasting.

Recognizing the vital influence of crude oil on the global economy, extensive efforts have been dedicated to accurately modeling and forecasting its price dynamics. Early approaches predominantly relied on statistical and econometric methods, typically assuming linear and time-invariant relationships. While these traditional models provided a foundation for understanding market dynamics, they often failed to capture the complex, nonlinear, and evolving nature of oil price behavior. For example, the time-varying inefficiencies in short-term oil price movements identified by Alvarez-Ramirez et al. [7] highlight the inherent limitations of linear modeling frameworks. In a related attempt, Shambora and Rossiter [8] employed artificial neural networks (ANNs) to detect market inefficiencies. However, despite moving beyond linearity, their models demonstrated only modest performance improvements, highlighting the need for more sophisticated approaches.

In response to these limitations, artificial neural networks have gained increasing attention for their ability to handle complexity and nonlinearity in data. ANNs are computational architectures inspired by biological neural systems, capable of learning patterns and performing predictive tasks autonomously [9]. Structurally, ANNs consist of interconnected processing units—neurons—arranged in layers: an input layer that receives data, one or more hidden layers that process the data, and an output layer that delivers predictions. Their adaptability and generalization capabilities have facilitated their application across various domains, including energy production forecasting [10] and financial market analysis [11,12].

The growing interest in ANN-based models is further supported by advances in artificial intelligence (AI) and machine learning, which have significantly expanded the toolkit available for forecasting tasks [13,14]. The increasing computational power of modern systems, along with algorithmic developments, has contributed to the widespread adoption of machine learning techniques across fields such as economics, finance, energy, and renewable energy studies [15,16]. Particularly when dealing with long-term, high-frequency, or nonlinear time-series data, machine learning algorithms often outperform conventional econometric methods by uncovering hidden structures and learning from data without explicit programming [17].

As machine learning techniques continue to evolve, more sophisticated and hybrid approaches have been developed to address the specific challenges posed by oil price forecasting. Nonlinear and ensemble-based models have shown considerable promise in this regard. For instance, Yu et al. [18] integrated Empirical Mode Decomposition (EMD) with feed-forward neural networks (FNNs) to enhance predictive accuracy, while Ghaffari and Zare [19] introduced data-filtered neural networks to mitigate noise and improve

Table 1
World Energy Supply (EJ).

Energy Type/Years	2010	2022	2023	2030	2035	2040	2050
Oil	173	187	192	195	189	182	176
Coal	153	172	175	156	131	114	94
Natural Gas	115	144	145	153	153	152	152
Renewables	43	74	78	120	153	185	241
Nuclear	30	29	30	36	41	45	49
Traditional use of biomass	21	19	19	15	13	12	10
Total	536	629	642	676	682	691	722

Note: The values of 2010, 2021, and 2022 show realizations. 2030, 2040, and 2050 values are the IEA's estimates of the levels that will be reached in these years if current policies are maintained.
Source: [4].

Table 2
The Top Producer and Consumer Countries of Oil.

	Oil Producer			Oil Consumer	
Country	MB/D	Share	Country	MB/D	Share
United States	21.91	22 %	United States	20.01	20 %
Saudi Arabia	11.13	11 %	China	15.15	15 %
Russia	10.75	11 %	India	5.05	5 %
Canada	5.76	6 %	Russia	3.68	4 %
China	5.26	5 %	Saudi Arabia	3.65	4 %
Iraq	4.42	4 %	Japan	3.38	3 %
Brazil	4.28	4 %	Brazil	3.03	3 %
United Arab Emirates	4.16	4 %	South Korea	2.55	3 %
Iran	3.99	4 %	Canada	2.41	2 %
Kuwait	2.91	3 %	Germany	2.18	2 %
Total top 10	74.59	73 %	Total top 10	61.08	61 %
World total	101.81		99.95		

Note: MB/*D* = million barrels per day, Share = Share of World Total.
Source: [5].

learning efficiency. Similarly, hybrid models that combine wavelet transforms with recurrent neural networks (RNNs) [20], as well as EMD-augmented FNNs proposed by Xiong et al. [21], represent efforts to better capture the multifaceted nature of financial time series. Furthermore, Bildirici and Ersin [22] applied neural network-augmented GARCH models to account for volatility clustering. However, their framework lacked the integration of exogenous macroeconomic variables and did not incorporate robustness testing, thereby limiting the generalizability of their findings.

In summary, the evolution from traditional linear econometric models toward sophisticated machine learning techniques marks a fundamental transformation in how crude oil price dynamics are modeled and analyzed. As the complexity of energy markets increases, so too does the need for more adaptive, nonlinear, and data-driven methodologies. ANN-based and hybrid models, supported by growing computational capabilities, offer promising avenues for achieving greater predictive accuracy and robustness.

Building on this paradigm shift, recent literature has increasingly focused on integrating optimization techniques and decomposition methodologies to enhance model performance more effectively. These approaches aim to refine data inputs and model architectures by minimizing noise and optimizing learning processes. For example, Li et al. [23] combined EEMD with Relevance Vector Machines (RVM) and Particle Swarm Optimization (PSO) to improve forecasting precision. Miao et al. [24] employed Lasso regression for feature selection, demonstrating the importance of variable relevance in predictive modeling. In addition, the incorporation of alternative data sources, such as Li et al.'s [25] use of the Google search volume index illustrates the growing role of big data in capturing market sentiment. Meta-heuristic algorithms, including the Bat algorithm [26], have also been utilized to integrate multiple model outputs and enhance overall forecasting performance. However, despite these advancements, many hybrid approaches still suffer from weaknesses related to model robustness, overfitting, and inadequate validation across different datasets [27]. Ensemble models often lack clarity in how component models contribute to the final prediction, and they rarely undergo detailed sensitivity analysis. This limits their practical applicability in policy and investment contexts.

To address the persistent challenge of forecasting highly volatile WTI crude oil prices, this study proposes a structured ensemble model that combines the strengths of both traditional econometric methods and modern deep learning techniques. Specifically, we develop a hybrid Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) –Long Short-Term Memory (LSTM) framework that integrates the linear models with the predictive capacity of nonlinear architectures. By incorporating exogenous macroeconomic indicators and conducting rigorous robustness testing, including sensitivity analysis and cross-validation, the proposed model aims to enhance both forecasting accuracy and reliability under complex market conditions. In summary, our study proposes a robust ARIMAX–LSTM forecasting model that integrates exogenous macroeconomic indicators and performs comprehensive robustness testing. It aims to address three main research questions: (i) whether machine learning models outperform traditional econometric ones in robustness and accuracy; (ii) whether hybrid models provide superior forecasting; and (iii) how sensitive model accuracy is to parameter variations. The model's performance is benchmarked through various robustness checks and statistical tests.

The novelty of our model lies in a methodological division of labor between ARIMAX and LSTM networks, chosen for their complementary strengths. The ARIMAX component captures interpretable linear relationships with key exogenous macroeconomic indicators (e.g., the Dollar Index and oil inventories), grounding the model in economic theory. The LSTM component models complex nonlinear and long-memory dynamics that arise from market sentiment and geopolitical shocks.

This study makes a significant contribution to the literature by addressing a key gap through the development of a hybrid forecasting framework that integrates economic covariates and combines both linear and nonlinear dynamic structures. First, the model explicitly incorporates exogenous economic variables—often neglected in conventional time series approaches—while capturing complex patterns through the combination of ARIMAX and LSTM components. This integration enables the model to capture both short-term linear effects and long-memory nonlinear dynamics, particularly episodes of heightened market volatility. Second, although the ensemble model achieves the lowest point-forecast errors, the Diebold–Mariano test suggests its advantage over the stand-alone LSTM is not statistically significant. Nevertheless, this ranking remains stable across a comprehensive set of robustness

checks, including five-fold cross-validation, multiple ARIMAX specifications, diverse LSTM architectures, and an exhaustive grid of ensemble weights, underscoring the model's reliability and generalizability. Finally, the study reveals the practical trade-off between model complexity and forecasting accuracy, presenting a flexible framework that can be tailored to real-world applications such as policy design, risk management, and capital allocation—domains where reliable and interpretable forecasts are of critical importance.

The research is arranged as follows: Following the introduction, we explain the literature review in Section 2. Section 3 sets out the specification of the analysis method and the test types. The results are reported in Section 4. The last part discusses the findings and presents concluding remarks.

2. Literature review

With global economies, energy markets, and financial decisions at stake, precise forecasting of crude oil prices has become a key concern in both academic and policy domains. The volatile, complex, and nonlinear nature of oil prices has prompted researchers to explore a range of modeling approaches, from classical statistical methods to advanced hybrid machine learning techniques, each aiming to improve predictive accuracy while maintaining robustness. Nevertheless, the inherent trade-off between accuracy and robustness presents a critical obstacle to modeling reliability.

Early studies to forecast oil prices typically relied on traditional time series models, such as Autoregressive Integrated Moving Average (ARIMA). While these models offer strong baseline predictions, their linear structure limits their ability to capture complex market dynamics. In response, hybrid approaches that combine classical models with machine learning components have gained traction. Wang and Wu [28], for instance, combined ARIMA with Backpropagation (BP) neural networks to improve accuracy beyond that of any standalone model. Likewise, Lyu and Chang [29] combined ARIMA and BP networks to forecast WTI crude oil prices, demonstrating that such hybrid models enhance confidence in forecasts within global markets. Alquist and Kilian [30] further underscored the importance of carefully selecting sample periods and model specifications, demonstrating that these decisions substantially affect forecast performance.

Efforts to extend these classical methods led to hybrid formulations. Aamir et al. [31] introduced a hybrid framework incorporating Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), ARIMA, and Kalman Filter techniques, aiming to capture both linear and structural dynamics to improve prediction accuracy for investment and policy-making contexts. In comparative studies, Ramyar and Kianfar [32] showed that Multilayer Perceptrons (MLPs) outperformed traditional Vector Autoregressive (VAR) models, highlighting the latter's inability to fully capture the intricacies of oil price behavior.

These limitations in traditional modeling approaches have stimulated interest in decomposition-based techniques, which address issues of non-stationarity and noise in time series data [33]. Decomposition frameworks separate a complex signal into more manageable subcomponents, each of which can be modeled more effectively. Yu et al. [18] introduced an EMD–FNN model to exploit multiscale features, and Li et al. [23] integrated EEMD with RVM, using PSO for kernel tuning. Similarly, Lin and Sun [34] combined CEEMDAN with Multilayer Gated Recurrent Units (ML-GRU) neural networks to model the nonlinear and volatile behavior of oil prices with greater fidelity. Qin et al. [35] proposed a decomposition-ensemble model based on EEMD and Local Linear Prediction (LLP) for predicting energy price time series with nonlinearity and non-stationarity features, showing that it consistently outperformed other techniques across varying parameter configurations and test conditions. Another research [36] utilized multi-scale forecasting to explore the relationship between oil prices and search engine data, revealing that multivariate decomposition methods can detect joint patterns across heterogeneous time series where traditional models fail. These decomposition frameworks were particularly effective in reducing noise, enhancing signal clarity, and improving performance during short-term volatility bursts. The general applicability of such decomposition methods is further underscored by their successful use in other fields for analyzing complex non-stationary signals, such as in biomedical engineering for high-accuracy voice pathology detection [37]. Later, robustness checks the temper of these gains. Lin and Sun [34] show their CEEMDAN–MLGRU edge vanishes in rolling-window tests, Yu et al. [18] find boundary effects, double error for their wavelet–RNN beyond four weeks, and Zhang et al. [38] report SVM–PSO accuracy swinging with each optimizer restart, signaling that many decomposition–optimization hybrids sacrifice stability for point accuracy.

In parallel with these developments, the adoption of machine learning and deep learning techniques has substantially advanced crude oil price forecasting. These approaches, particularly ANN, LSTM, Gated Recurrent Units (GRU), and Multiple Kernel Learning (MKL), have proven effective at capturing the nonlinearities and complex dependencies inherent in energy markets. Villada et al. [39] constructed an ANN based on the daily oil price, incorporating macro-financial indicators such as the Dollar Index and S&P 500 to capture investor risk aversion and nonlinear market responses. The empirical results confirmed that ANN models not only yield accurate forecasts but also capture the nonlinear characteristics of oil prices. Deng and Sakurai [40] applied MKL-based regression models for predicting spot oil prices to outperform traditional SVR techniques, further substantiating the value of non-linear models. Cheng et al. [41] introduced a hybrid strategy that fused a Vector Error Correction (VEC) model with a Nonlinear Autoregressive (NAR) neural network, thereby accommodating both lagged price dependencies and nonlinear dynamics. The VEC model allows for both endogenous and exogenous factors through crude oil price lags and their effects on corresponding variables, and the NAR model captures the non-linear dynamics of the prices. Likewise, Ramyar and Kianfar [32] and Hadjira et al.[42]. also emphasized the superiority of neural networks such as MLPs and fuzzy time series models over traditional linear approaches, particularly under volatile and uncertain market conditions.

Given that no single neural model can fully encapsulate the multifaceted behavior of crude oil prices, researchers have increasingly turned to hybrid frameworks that integrate multiple models. Abdollahi [43], for instance, proposed a composite model involving Support Vector Machine, CEEMD, Complete Ensemble Empirical Mode Decomposition (CEEMD), Markov-Switching Generalized Autoregressive Conditional Heteroscedasticity (MS-GARCH), and PSO. This architecture was designed to handle both volatility

clustering and regime shifts more effectively. Similarly, Lin and Sun [34] demonstrated that their CEEMDAN-MLGRU model more accurately captured price fluctuations compared to standalone methods. These findings reinforce the utility of hybrid approaches in modeling financial time series with nonlinear and nonstationary characteristics.

To further boost predictive performance, ensemble strategies have emerged as a powerful paradigm. These models combine the outputs of multiple learners—often via weighted schemes or regularized regression—to achieve improved accuracy and generalization. Zhang et al. [38] employed heterogeneous autoregressive models with shrinkage techniques such as Elastic Net and Lasso to outperform individual forecasting models, demonstrating the benefit of combining diverse predictive views. Despite their promise, many ensemble models lack comprehensive robustness assessments. Gunarto et al. [44], for example, introduced an AR(1)-GARCH(1, 1) model with remarkably low prediction error; however, its sensitivity to different parameter settings and sample conditions was not thoroughly examined. This gap points to a broader issue in the literature—namely, the insufficient evaluation of models' robustness and sensitivity to input variation. Mishra [45] and Jha and Cucculelli [46] also underlined that weighted ensemble models significantly outperform single models by giving more prominence to stronger learners. Yet, despite their effectiveness, such ensemble models are rarely stress-tested under different parameter configurations or market regimes.

Robustness and sensitivity analyses are essential for the practical applicability of forecasting models, as they assess the model's stability under varying conditions. Without such an evaluation, the generalizability and reliability of forecasting models remain questionable. In this context, Qin et al. [35] showed that their EEMD-LLP model maintained accuracy across different input configurations. These efforts emphasize the increasing significance of validating models not only for accuracy but also for their reliability under varying conditions. Other researchers have tackled robustness from alternative perspectives. Ren and Xu [47] proposed an adversarial robustness verification scheme for machine learning-based dynamic spectrum access, yielding more reliable results under live testing conditions. Chuah et al. [48] introduced a diagnostic toolkit to detect parameter sensitivity and robustness-related anomalies in ML classifiers, while Gosch et al. [49] demonstrated how incorporating label semantics into Graph Neural Networks can mitigate over-robustness without sacrificing accuracy in adversarial environments.

Taken together, the literature reveals a pressing need for forecasting models that strike a balance between predictive power and robustness. Although hybridization and ensemble learning approaches have shown promise, only a few studies have systematically tested these models across different regimes, parameter configurations, or data scenarios. Additionally, many hybrid models remain "black boxes," failing to leverage the interpretability of econometric methods or evaluate sensitivity to external shocks.

This study addresses that gap directly by proposing a structured ensemble framework that combines linear models like ARIMAX with the predictive strength of nonlinear deep-learning architectures such as LSTM. Embedded within this framework are rigorous robustness testing and detailed sensitivity analysis, aiming both to boost forecasting performance and to enhance the reliability of models in real-world applications.

Looking ahead, future research should focus on developing integrated forecasting models that harness both linear and nonlinear paradigms within well-optimized, weighted ensembles supported by comprehensive validation. By carefully designing and testing ARIMAX–LSTM combinations, studies can significantly improve the credibility and practical utility of crude-oil price forecasts.

3. Methodology

3.1. Data collection and preprocessing

Data collection and preprocessing are the backbone for building a robust predictive data model. Therefore, the study collects accurate data from trusted sources and good transformations that preserve statistics. This section describes the data sources (Table 3), transformation types, and how the data will be partitioned into training and testing.

To ensure the model performs effectively across a dataset characterized by heterogeneous scales and measurement units, standardized monthly data spanning from April 2004 to January 2024 were employed. Standardization is particularly crucial when

Table 3
Data Source.

Variable Name	Literature Reference	Source
WTI Crude Oil Price	Used as the primary target variable for forecasting models. Its time series is analyzed for complex characteristics like nonlinearity and chaos, and advanced machine learning models are developed to improve prediction accuracy [50, 51].	Yahoo Finance
Dollar Index	Used as a crucial external economic and financial indicator in forecasting models. It is incorporated as an input variable to enhance the predictive performance and accuracy of models for crude oil and other commodity prices [52–55].	Yahoo Finance
Crude Oil Imports	Used as a fundamental supply-demand indicator and a key input variable in various forecasting models. It helps enhance the accuracy of crude oil price predictions by capturing market dynamics and supply chain pressures [52,56, 57].	eia.gov
Crude Oil Exports	Serves as a critical input variable in forecasting models, representing a key component of global supply. Incorporating export data helps improve the predictive accuracy of both traditional and advanced machine learning models for crude oil prices [52,54,55].	eia.gov
Crude Oil Stocks (Inventories)	Used as a fundamental indicator of the supply-demand balance and a key input for price prediction. This variable is particularly significant for modeling price volatility, as inventory announcements can cause notable market reactions, and its inclusion enhances the accuracy of advanced forecasting frameworks [52,58,59].	eia.gov

working with multivariate time-series data, as it enhances comparability between features and stabilizes the learning process across different periods.

Importantly, this two-decade period encompasses several major global events that have had a profound impact on oil price dynamics. As discussed earlier, oil markets are acutely responsive to macroeconomic shocks and geopolitical developments. Within this timeframe, the 2008 global financial crisis, the Russia-Ukraine war, and the COVID-19 pandemic have each introduced considerable volatility into oil price movements [60,61]. These exogenous disruptions highlight the necessity for robust and adaptive modeling techniques—such as machine learning and hybrid approaches—that effectively capture nonlinearities and structural breaks in oil price behavior.

The MinMaxScaler and StandardScaler from the scikit-learn library make the features comparable. MinMax scaling is the process by which the range of variables in a dataset is normalized. This step is crucial for algorithms like Stochastic Gradient Descent to work with features on the same scale [62,63]. Model performance has dramatically increased with StandardScaler, a popular scaling technique [64].

The StandardScaler provides some benefits like better classifier performance and an automatic framework for meta-learning, selecting optimal scaling methods [65]. But it also implies some drawbacks. For instance, scale-invariant parameters coupled with momentum-based gradient descent optimizers may make you suffer slightly suboptimal performance again, showing why we must do it right regarding scale. In real-world applications, e.g., healthcare, the model interpretability needs to be kept or enhanced to gain trust in the predictions [66].

The data is split chronologically between the training and test sets, with 80 % of the data used for training and 20 % reserved for the test. The data is partitioned chronologically, so the models are tested using unseen data in a simulated manner for real-world forecasting based on historical observations.

3.2. Model development

3.2.1. ARIMAX

The ARIMA offers a unified framework for analyzing and forecasting nonstationary time series data, capturing deterministic and stochastic components. Three building blocks define the model [67]: autoregressive (AR) dynamics, integrated (I) differencing for stationarity, and a moving average (MA) component. Let z_t denote the observed time series at time t, and B is the backshift operator such that $Bz_t = z_{t-1}$. The series is differenced d times to remove nonstationary trends:

$$\nabla^d z_t = (1 - B)^d z_t \tag{1}$$

This transformed series is then fitted by a combination of AR and MA operators. The autoregressive component of order p is represented by the polynomial:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2}$$

while the moving average component of order q is expressed as:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{3}$$

By combining these elements, the general ARIMA (p, d, q) model is formulated as:

$$\phi(B)(1 - B)^d z_t = \theta(B) a_t \tag{4}$$

where a_t is a white noise with zero mean and constant variance σ_a^2. This general ARIMA formulation allows the model to accommodate different time-dependent behaviors, making it ideal for research and practical forecasting in time series analysis [67].

The choice of algorithms has become the key to achieving good prediction performance in a prediction model. In this part, we apply the ARIMAX model, which considers exogenous covariates in the analysis. We describe the procedure for selecting the parameters, training them, and evaluating the model performance, starting with the stationarity of the time series using the Augmented Dickey-Fuller (ADF) test.

ADF test, in the Table 4, $d = 1$ (first difference) confirms stationarity at 5 % level: the p-value is < 0.05 and a test statistic (-3.3649) lower than the 5 % critical value (-2.8739). In model search, we fix $d = 1$ to maintain model simplicity and uphold the principle of

Table 4
Augmented Dickey-Fuller Test Results.

Metric	Value
Test Statistic	−3.3649
p-value	0.0122
Lags Used	1
Number of Observations	236
Critical Value (1 %)	−3.4584
Critical Value (5 %)	−2.8739
Critical Value (10 %)	−2.5733

parsimony. Since there is a $d = 1$ differencing, it will improve the accuracy metrics, which warrants its inclusion in the ARIMAX model.

In an ARIMAX model, the autoregressive (AR) terms account for the association between the dependent variable and its lagged values [68,69]. The integrated (I) part, which handles non-stationarity of time series by permitting a variety of ways to attain stationarity [70]. The moving average (MA) model deals with the relationship between the dependent variable and past residual errors [69]. On the other hand, exogenous variables are independent inputs not included in the time series but superimposed on the model due to their possible effect on the dependent variable [68,70].

The best set of hyperparameters from Table 5 that gives the least MSE validation score is determined using three metrics (Table 6): Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The ARIMAX model is also able to add more exogenous variables that may be affecting the modeled ones [70]. However, the ARIMA model has an exclusive dependence on historical time series and a failure to consider external facts [71].

The ARIMAX model is fitted to the training data with exogenous variables such as the crude oil import/export data, crude oil stocks, and the Dollar Index as hyperparameters. During this process, the data is scaled to estimate model coefficients that minimize the MSE and maximize model fitness.

In Fig. 1, the actual price (green) and the predicted price (red) do not align: the green line has pikes around 2020—2021 that are not observed in the red line. It indicates that the model is not well-suited to capturing the sharp volatility in crude oil prices during this period.

3.2.2. LSTM

LSTM is a gated RNN variant that can alleviate the vanishing gradient problem and capture long-term dependencies of sequences. That makes it very attractive for time series prediction when historical patterns and lag effects are important, precisely in crude oil price forecasting.

At each timestep, the LSTM cell maintains a cell state C_t and a hidden state h_t regulated by a gating mechanism. The following equations describe the input gains of an LSTM unit at a time step t [72]

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \quad \text{(forget gate)} \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \tag{6}$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(candidate cell state)} \tag{7}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \quad \text{(cell state update)} \tag{8}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \tag{9}$$

$$h_t = o_t \odot \tanh(C_t) \quad \text{(hidden state output)} \tag{10}$$

Here, $\sigma(\cdot)$ is the sigmoid activation, $\tanh(\cdot)$ is the hyperbolic tangent, and $\odot$ is element-wise multiplication. x_t denotes the input at time t, where W_* and b_* are the learnable weights and biases for each gate.

These equations allow LSTM to control how information flows over time and determine what to remember, update, and forget. It has shown excellent results for modeling non-linear, temporal dynamics of financial time series like crude oil prices. An LSTM is just a more controllable version of an RNN — it can remember long-term dependencies in sequences as well as regular ones [73]. It can be used for time series predictions [62] and effectively addresses the common issue of gradient fading/explosion, which often hinders standard RNNs [74].

We use LSTM because it can learn the long-term trend of WTI crude oil prices as an output model. As shown in Table 7, the model includes an LSTM layer with 50 units (return sequences), another layer with 50 units and returns sequences at the output, dropout layers afterward for each LSTM to prevent overfitting by randomly deactivating 40 % of the neurons during training, and a final dense layer for production.

Reshaping enables LSTM to learn the time series and capture the temporal dependencies, increasing productivity. The Adam optimizer trains faster deep learning models and uses an MSE loss function to measure prediction error (Table 8). We found that the optimal configuration for the LSTM is 50 units and a dropout rate of 0.4.

The LSTM model is trained for 100 epochs for adequate learning and convergence. A batch size of 32 is selected to optimize the trade-off between computational efficiency and model performance. The model's performance is validated on the test dataset to monitor overfitting and ensure generalization. After training, the model's predictions are tested on both the training and test datasets. The inverse transformation of predicted values is employed to facilitate comparison with actual WTI crude oil prices, enabling a

Table 5
Grid Search Parameter Ranges.

Parameter	Range
p	[1, 2, 3, 4]
d	[1]
q	[1, 2, 3, 4]

Table 6
Performance Metrics for the ARIMAX Model (Training and Testing).

Parameters	Train MSE	Train RMSE	Train MAE	Test MSE	Test RMSE	Test MAE
(4, 1, 4)	0.07901	0.28109	0.20818	0.55603	0.74568	0.59872

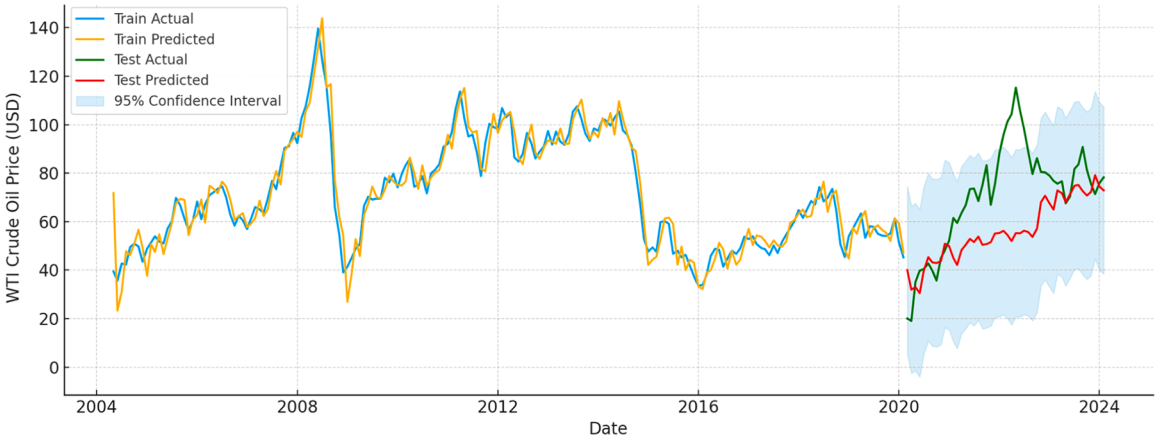


Fig. 1. Actual vs Predicted WTI Crude Oil Price from ARIMAX.

Table 7
Model Architecture.

Layer	Configuration
First LSTM Layer	50 units, keeps returning sequences
Second LSTM Layer	50 units, stop returning sequences
Dropout Layer	40 % of the neurons are turned off temporarily after each LSTM layer
Dense Layer	1 unit for the final output

Table 8
Compilation Details.

Parameter	Configuration
Optimizer	Adam
Loss Function	Mean Squared Error

rigorous assessment of model performance (Table 9).

Fig. 2 shows the actual versus predicted prices using the LSTM model. The blue line indicates actual prices, the orange line shows predictions on the training data, and the green line indicates predictions on the test data. As predicted, the LSTM model performs well in the training set, and actual values are very close. However, predictions on test sets are far from actual values.

During the testing period, the green line tends to move in the same direction but shows different amplitudes and delays in the actual price, mainly regarding high volatility. It indicates that our model learns some general patterns but needs to be fine-tuned and trained on more extensive data to enhance precision, especially for sharp price movements.

The LSTM model shows convergence in long-term dependencies and temporal patterns. This is an interesting characteristic for forecasting time series like WTI crude oil prices, as they are subject to external influences. Although the model presents a few observed discrepancies in the test set, our robust base will allow better predictions.

3.2.3. Ensemble model

Forecast combination techniques seek to enhance predictive performance by aggregating multiple model forecasts into a unified

Table 9
Performance Metrics for the LSTM Model (Training and Testing).

Units	Dropout	Train MSE	Train RMSE	Train MAE	Test MSE	Test RMSE	Test MAE
50	0.4	0.005603	0.074855	0.057068	0.006808	0.082509	0.069445

Fig. 2. Actual vs Predicted WTI Crude Oil Price from LSTM Model.

estimate. The underlying premise is that individual forecasting models often capture different aspects of the data-generating process, and their combination can mitigate model-specific biases and reduce forecast variance. As originally discussed by Timmermann [75], a linear forecast combination approach can be formulated as:

$$\widehat{y}^{\text{combined}}_{t+h|t} = \omega_1 \widehat{y}_{t+h|t,1} + \omega_2 \widehat{y}_{t+h|t,2}, \text{ with } \omega_1 + \omega_2 = 1 \tag{11}$$

where $\widehat{y}_{t+h|t,i}$ denotes the forecast from model i for horizon h, and ω_i are the corresponding combination weights. This formulation ensures unbiasedness under the assumption of unbiased component forecasts and allows for a diversification of model-specific risks.

In this study, we generalize a strategy of combining a linear statistical model – ARIMAX– with a nonlinear deep learning model – LSTM– to enhance predictive performance. The ensemble forecast at time t is specified as:

$$\widehat{y}_t = w_A \widehat{y}_{t,\text{ARIMAX}} + w_L \widehat{y}_{t,\text{LSTM}}, \text{ where } w_A + w_L = 1 \tag{12}$$

This model hybridizes the temporal intuitiveness of the ARIMAX model and the deep pattern recognition abilities of LSTM networks. The weights w_A and w_L are optimized to minimize the out-of-sample forecasting error using the MSE loss function. Such an ensemble method takes advantage of the complementary advantages between the two models, providing a more robust and precise prediction scheme.

In this study, we combine ARIMAX and LSTM models as an ensemble model to improve prediction accuracy and robustness. We implement different weight intensities to enhance the accuracy of price predictions. The ensemble is based on the weighted sum of ARIMAX and LSTM predictions, and the model adjusts the weights automatically in real time according to the samples in different data spaces. The model then classifies it with multivariate logistic regression, the top weights that trump singular models. The input of multiple subsets in an ensemble model aggregates its forecast for a final decision, enhancing performance with time series forecasting [75].

Mishra [45] explains that the weighted ensemble model enhances forecast accuracy since the best-performing model will have more prominence when predicting new data points (inferior forecasting performance). Jha and Cucculelli [46] point out that this results in less dramatic predictors than unweighted procedures. A wide range of studies proved that weighted ensemble models surpass individual models [76].

We determine the weights, in Table 10, for combining ARIMAX and LSTM model predictions by conducting a grid search, selecting values from 0 to 1 in increments of 0.05.

The evaluation uses MSE, RMSE, and MAE measures for every weight set. The minimum of RMSE helps us find the optimal weight (Table 11).

Fig. 3 compares actual WTI crude oil prices to those predicted using the ensemble algorithm. The blue line shows the exact prices, and the orange and green lines represent the training/test data predictions. The ensemble model has excellent predictive performance on the training set, as the actual and predicted values are closely linked by eye observation.

As we can see, the ensemble model's estimates are much more consistent with the actual prices and improve slightly over time during testing when aligned with the red line, compared to individual models. These predictions should have captured not just the average up or downward movement but price fluctuations too; that means learning has been more efficient. This progression shows the

Table 10
Grid Search Weight Range.

Weight for ARIMAX Predictions	Weight for LSTM Predictions
0, 0.05, 0.10, …, 1.0	1 - Weight for ARIMAX

Table 11
Performance Metrics for the Ensemble Model (Training and Testing).

Parameters	Train MSE	Train RMSE	Train MAE	Test MSE	Test RMSE	Test MAE
ARIMAX: 0.1 LSTM: 0.9	0.00481	0.06938	0.05294	0.00659	0.08117	0.06921

Fig. 3. Actual vs Predicted WTI Crude Oil Price from Ensemble Model.

power of combining ARIMAX and LSTM models using an ensemble method to get more accurate and reliable predictions.

The ensemble augments predictive capacity and strength by combining the strengths of these two models. Such studies illustrate the value of ensemble learning in improving prediction quality.

3.3. Robustness testing

3.3.1. Cross validation

We assess our models using a five-stride cross-validation process, a widely accepted technique for evaluating machine learning models' performance. The dataset is split into training and testing partitions (the model is trained on one and tested on the other) to ensure an unbiased evaluation.

For cross-validation, the dataset is divided into five subsets or folds. Train on four folds and operate on the residual fold, repeating this entire process five times, whereby each fold is a test set once. We can use this iteration average as an overall performance measure that balances risk and reduces bias [77]. It is, instead, for comparing models' out-of-sample performances [78] with the actual versus the predicted from the test data. We did this for ARIMAX, LSTM, and ensemble models; it exposes the pluses and minuses. Cross-validation overcomes the model bias (overfitting) and improves model selection efficiency [79,80]. It aims to choose the best model that performs adequately and generally across a task's different subsamples or parts.

3.3.2. Sensitivity analysis

We perform a sensitivity analysis on ensemble weights for multiple ARIMAX and LSTM models to examine their time-varying sensitivity in predictive robustness. Stein et al. [81] showed that when their ensemble base models are weighted, they remove every best, thus doing a far better job. We assess the contribution of approach features to overfitting, underfitting, and validating the model by tuning its hyperparameters [82,83]. This technique is unsupervised and is also utilized for making the model more robust [84]. For the ARIMAX model, we conduct a sensitivity analysis by examining how variations in the autoregressive, differencing, and moving average orders affect both positive and negative testing performance. The concern is with further adjusting these orders to enhance out-of-sample forecasting. Through this process, the most appropriate parameters are selected to maximize accuracy and minimize outliers from the model [82].

We also examine the impact of different LSTM units and dropout rates on model performance. Findings indicate that some sets of hyperparameters are more stable and generalize better than others. Results also show that an appropriate hyperparameter configuration reduces the risk of overfitting and improves the model's generalization for prediction tasks.

4. Results

4.1. Cross-validation results

The results section details all the findings from model development and reinforcement (performance measures on ARIMAX, LSTM,

and Ensemble models) and takes care of performance metrics. It further provides the models' performance on the train and test datasets with cross-validation and sensitivity analysis results, as shown in Table 12.

The ARIMAX model's cross-validation performance is somewhat deranged. Specifically, the cross-validation results MSE are 0.06250 (train) and 0.03608 (test), while actuals are 0.07901 (train) and 0.55603 (test). It indicates worse performance in real-world testing and suggests potential improvements in robustness.

For RMSE, cross-validation yields 0.24164 (train) and 0.07170 (test), whereas the actual values are 0.28109 (train) and 0.74568 (test). It highlights a performance degradation from cross-validation to real-world data, with error rates higher in the actual test data.

Additionally, there is a slight increase in MAE from cross-validation to actual performance. Cross-validation MAE values are 0.20517 (train) and 0.06434 (test), while actual values rise to 0.20818 (train) and 0.59872 (test). The slight increase in train MAE and significant rise in test MAE suggest that the model performs similarly during training but significantly worse in actual testing conditions.

The LSTM model (Table 13) shows mixed performance, with better results during training but poorer results during testing. The cross-validation MSE values are 0.01106 (train) and 0.00289 (test), while the actual MSE values are 0.005603 (train) and 0.006808 (test). The lower training MSE suggests strong performance on the training data, but the higher test MSE indicates poorer performance on unseen data.

For RMSE, the cross-validation values are 0.10445 (train) and 0.01389 (test), compared to actual values of 0.074855 (train) and 0.082509 (test). The lower training RMSE suggests better performance, while the higher test RMSE points to difficulties in generalizing new data.

The MAE follows a similar pattern. The cross-validation values are 0.08439 (train) and 0.01391 (test), while the actual values are 0.057068 (train) and 0.069445 (test). Although the training MAE is lower, the higher test MAE indicates reduced prediction accuracy on new data.

The Ensemble model (Table 14) exhibits a similar trend, performing better during training but worse during testing. The cross-validation MSE values are 0.01159 (train) and 0.00356 (test), while the actual MSE values are 0.00481 (train) and 0.00659 (test). The lower training MSE suggests good performance, but the higher test MSE indicates challenges in generalization.

For RMSE, the cross-validation values are 0.10668 (train) and 0.01629 (test), while the actual values are 0.06938 (train) and 0.08117 (test). The lower training RMSE suggests better performance, but the higher test RMSE shows poorer performance on new data.

Similarly, the MAE results follow the same pattern. The cross-validation values are 0.08675 (train) and 0.01716 (test), compared to actual values of 0.05294 (train) and 0.06921 (test). The lower training MAE indicates higher accuracy, while the higher test MAE indicates decreased reliability on unseen data. Cross-validation shows that all three models perform better during training than testing. The ARIMAX model experiences the most remarkable performance drop, suggesting a need for improved robustness. The LSTM and Ensemble models perform well in training but struggle to maintain accuracy during testing, indicating potential overfitting and challenges with generalization.

4.2. Sensitivity analysis results

4.2.1. ARIMAX sensitivity analysis results

The ARIMAX model's sensitivity analysis, in Table 15, examines the impact of different parameter settings on training and testing performance. Parameter configurations such as (1, 1, 1) and (2, 1, 1) result in high test MSE values (156.351 and 159.647), indicating poor performance. These settings also show elevated RMSE and MAE during testing, suggesting unstable predictions.

In contrast, parameter sets like (1, 1, 3), (2, 1, 3), and (3, 1, 3) yield much lower test MSE values (around 0.72041–0.73850), with corresponding reductions in RMSE and MAE, indicating more robust performance. These configurations effectively capture the dynamics of WTI crude oil prices.

Interestingly, the (4, 1, 4) setup achieves the lowest test MSE (0.55603) and the smallest RMSE (0.74568), with MAE dropping to 0.59872. While this reduces error, the higher train MSE suggests potential overfitting due to increased complexity. It highlights that higher values of p and q in the ARIMAX model can improve MSE but may lead to overfitting. The analysis underscores the importance of carefully calibrating parameters to optimize accuracy and robustness. Some parameter sets (Fig. 4), such as (1, 1, 1) and (2, 1, 1), result in higher test MSE and worse predictions.

4.2.2. LSTM sensitivity analysis results

The sensitivity analysis for the LSTM model, in Table 16, varying the number of units and dropout rates, reveals that the best test performance occurs with 50 units and a 0.4 dropout rate, achieving a test MSE of 0.006808, RMSE of 0.082509, and MAE of 0.069445. This configuration effectively prevents overfitting by balancing a moderate number of units with a high dropout rate.

Table 12
Cross-Validation vs. Actual Performance Metrics for the ARIMAX Model.

Metric	CV Train	CV Test	Actual Train	Actual Test	Train Explanation	Test Explanation
MSE	0.06250	0.03608	0.07901	0.55603	Higher than average (Worse)	Higher than average (Worse)
RMSE	0.24164	0.07170	0.28109	0.74568	Higher than average (Worse)	Higher than average (Worse)
MAE	0.20517	0.06434	0.20818	0.59872	Slightly higher than average (Slightly Worse)	Higher than average (Worse)

Table 13
Cross-Validation vs. Actual Performance Metrics for the LSTM Model.

Metric	CV Train	CV Test	Actual Train	Actual Test	Train Explanation	Test Explanation
MSE	0.01106	0.00289	0.005603	0.006808	Lower than average (Better)	Higher than average (Worse)
RMSE	0.10445	0.01389	0.074855	0.082509	Lower than average (Better)	Higher than average (Worse)
MAE	0.08439	0.01391	0.057068	0.069445	Lower than average (Better)	Higher than average (Worse)

Table 14
Cross-Validation vs. Actual Performance Metrics for the Ensemble.

Metric	CV Train	CV Test	Actual Train	Actual Test	Train Explanation	Test Explanation
MSE	0.01159	0.00356	0.00481	0.00659	Lower than average (Better)	Higher than average (Worse)
RMSE	0.10668	0.01629	0.06938	0.08117	Lower than average (Better)	Higher than average (Worse)
MAE	0.08675	0.01716	0.05294	0.06921	Lower than average (Better)	Higher than average (Worse)

Table 15
ARIMAX Model Sensitivity Analysis Results Across Parameter Configurations.

Parameters	Train MSE	Train RMSE	Train MAE	Test MSE	Test RMSE	Test MAE
(1, 1, 1)	0.07925	0.28151	0.21371	156.351	12.5040	101.727
(1, 1, 2)	0.07841	0.28002	0.21346	148.052	121.677	0.97895
(1, 1, 3)	0.07672	0.27698	0.20880	0.72041	0.84877	0.62230
(1, 1, 4)	0.07679	0.27711	0.20882	0.72299	0.85029	0.62388
(2, 1, 1)	0.07887	0.28084	0.21414	159.647	126.351	10.2760
(2, 1, 2)	0.07701	0.27749	0.20933	0.73850	0.85936	0.63371
(2, 1, 3)	0.07688	0.27727	0.20949	0.72942	0.85406	0.62823
(2, 1, 4)	0.07673	0.27701	0.20878	0.72051	0.84883	0.62237
(3, 1, 1)	0.07678	0.27710	0.20923	0.72566	0.85186	0.62580
(3, 1, 2)	0.07674	0.27701	0.20861	0.71603	0.84619	0.61985
(3, 1, 3)	0.07670	0.27695	0.20874	0.72412	0.85096	0.62480
(3, 1, 4)	0.07647	0.27654	0.21008	0.63270	0.79543	0.57848
(4, 1, 1)	0.07675	0.27704	0.20910	0.72231	0.84989	0.62367
(4, 1, 2)	0.07614	0.27593	0.20995	0.63127	0.79453	0.58044
(4, 1, 3)	0.07681	0.27715	0.20862	0.72076	0.84898	0.62284
(4, 1, 4)	0.07901	0.28109	0.20818	0.55603	0.74568	0.59872

LSTM models with 150 units consistently produce slightly higher test errors than those with 50 or 100 units, demonstrating that increased complexity does not always enhance performance and may result in overfitting.

Moreover, lower dropout rates (Fig. 5), such as 0.2, typically result in higher test errors, suggesting insufficient regularization. Therefore, the optimal LSTM setup requires a balanced combination of units and dropout rate.

4.2.3. Ensemble sensitivity analysis results

The ensemble model's sensitivity analysis, in Table 17, focused on varying weight values to assess their impact on performance. The optimal weight combination of 10 % ARIMAX and 90 % LSTM achieved the best results, with an MSE of 0.006589, RMSE of 0.081170, and MAE of 0.069214, demonstrating that emphasizing LSTM improves predictive accuracy.

As ARIMAX's weight increased, in Fig. 6, test errors worsened, with the maximum weight of 0.95 yielding the poorest performance (MSE = 0.019205, RMSE = 0.138581, MAE = 0.109796). It suggests that overemphasizing ARIMAX in the ensemble weakens its predictive power, highlighting the need for careful weight balancing.

The analysis underscores the importance of tuning model parameters. The ARIMAX configuration (4, 1, 4) achieved a test MSE of 0.55603, while the LSTM setup with 50 units and a 0.4 dropout rate resulted in a test MSE of 0.006808. However, the 10 % ARIMAX and 90 % LSTM ensemble model outperformed individual models, achieving the lowest test MSE of 0.006589. It illustrates that optimal weight configuration in an ensemble model enhances predictive performance beyond specific models.

Although the complexity of the models used in this study introduces a theoretical risk of overfitting, several precautions were taken to minimize this concern. Feature scaling, dropout regularization, extensive grid search, and cross-validation were all applied to improve the models' ability to generalize. The ensemble structure itself, which combines the strengths of ARIMAX and LSTM, was deliberately designed to mitigate model-specific biases and prevent over-reliance on any single architecture. Nonetheless, as with many deep learning approaches, the LSTM components may still capture patterns specific to the training period, particularly in the absence of a fully independent validation set. While the reported test performance reflects strong predictive accuracy, future research may benefit from stricter validation protocols or adaptive regularization techniques to further mitigate the risk of overfitting under volatile market conditions.

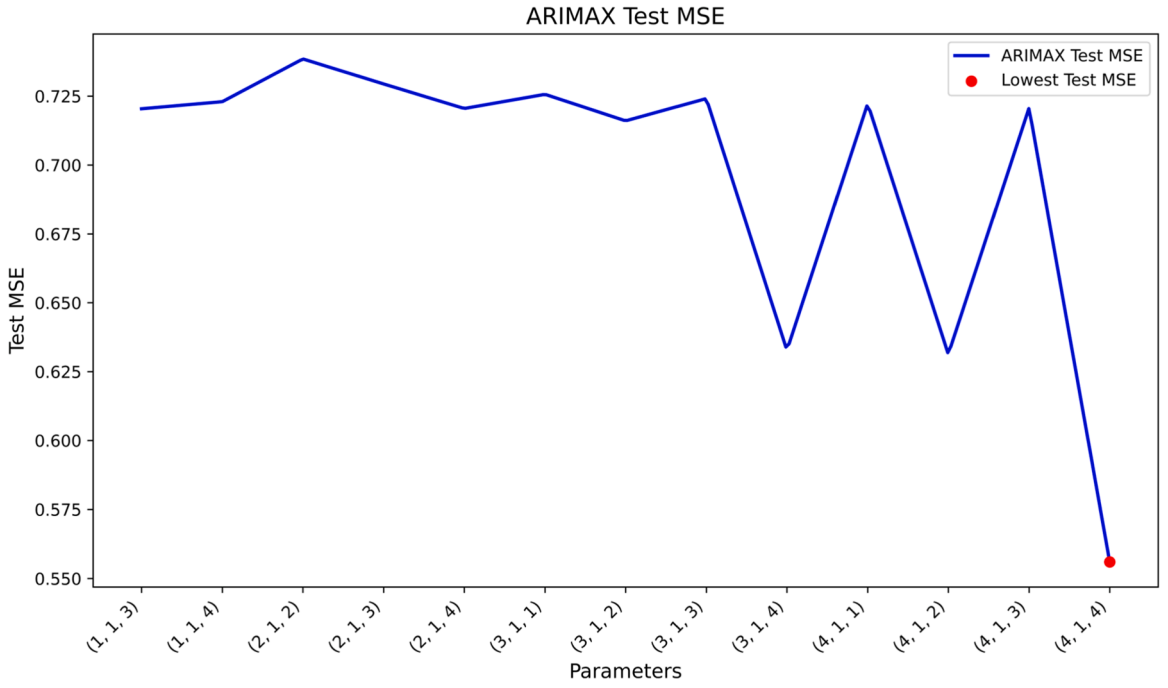


Fig. 4. Sensitivity Analysis of ARIMAX Model Test MSE Across Parameter Configurations.

Table 16
LSTM Model Sensitivity Analysis Results Across Parameter Configurations.

Units	Dropout	Train MSE	Train RMSE	Train MAE	Test MSE	Test RMSE	Test MAE
150	0.4	0.005710	0.075565	0.057617	0.006949	0.083361	0.069751
150	0.3	0.005611	0.074907	0.057181	0.006915	0.083158	0.068908
150	0.2	0.005487	0.074074	0.056944	0.007000	0.083665	0.069507
50	0.4	0.005603	0.074855	0.057068	0.006808	0.082509	0.069445
50	0.3	0.005693	0.075451	0.057379	0.006859	0.082818	0.069330
50	0.2	0.005811	0.076231	0.057733	0.007138	0.084485	0.071353
100	0.4	0.005687	0.075409	0.057882	0.007058	0.084013	0.070341
100	0.3	0.005712	0.075579	0.057680	0.006918	0.083172	0.069337
100	0.2	0.005534	0.074389	0.057017	0.006859	0.082816	0.069033

4.3. *Forecast-error comparison via the Diebold–Mariano test*

To formally test whether the observed difference in forecast accuracy between the ensemble and the stand-alone LSTM model was statistically significant, we applied the Diebold–Mariano (DM) test. The test was conducted on the one-step-ahead forecast errors from a common evaluation period consisting of $n = 37$ monthly observations.

Under a squared-error loss function, the loss differential series is defined as $d_t = e_{1,t}^2 - e_{2,t}^2$, where $e_{1,t}$ and $e_{2,t}$ denote ensemble and LSTM errors, respectively. The DM statistics:

$$DM = \frac{\overline{d}}{\sqrt{s_d^2/n}} \tag{13}$$

follows a two-sided t-distribution with $n - 1$ degrees of freedom [85].

The Diebold–Mariano tests demonstrate a clear preference for the ensemble. For both loss functions, the DM statistics are large and positive (6.076 for MSE, 7.283 for MAE) with values < 0.0001, so the null hypothesis of equal predictive accuracy is rejected. In other words, the ensemble's forecast errors are not just marginally but significantly smaller than those of the stand-alone LSTM, confirming the ensemble's statistical superiority in this sample.

4.4. *Benchmarks*

The benchmarks section details all the findings from model development in comparison to base models, as summarized in Table 18.

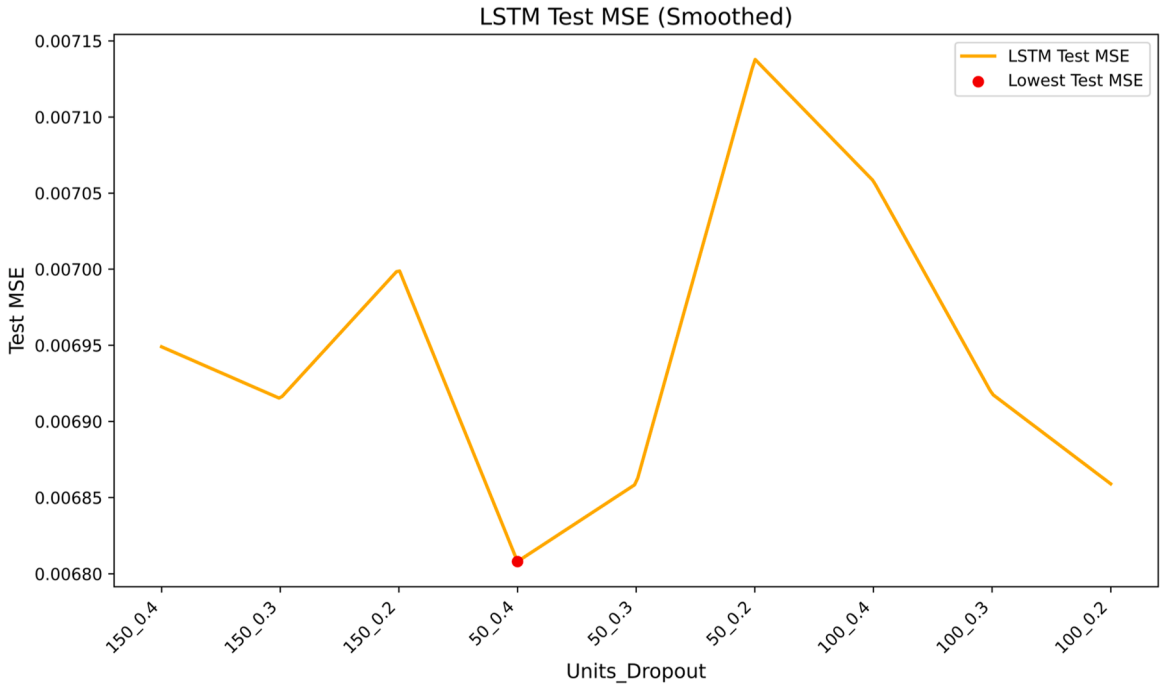


Fig. 5. Sensitivity Analysis of LSTM Model Test MSE Across Units and Dropout Configurations.

Table 17
Ensemble Model Sensitivity Analysis Across Weight Configurations.

Weight	MSE	RMSE	MAE
0.05	0.006677	0.081716	0.069123
0.10	0.006589	0.081170	0.069214
0.15	0.006592	0.081192	0.069305
0.20	0.006688	0.081780	0.069395
0.25	0.006876	0.082922	0.069486
0.30	0.007157	0.084597	0.069577
0.35	0.007529	0.086772	0.069667
0.40	0.007995	0.089413	0.069828
0.45	0.008552	0.092477	0.071013
0.50	0.009202	0.095926	0.072741
0.55	0.009944	0.099720	0.075524
0.60	0.010778	0.103819	0.079108
0.65	0.011705	0.108191	0.083216
0.70	0.012724	0.112802	0.087447
0.75	0.013836	0.117626	0.091743
0.80	0.015040	0.122636	0.096214
0.85	0.016336	0.127811	0.100685
0.90	0.017724	0.133132	0.105211
0.95	0.019205	0.138581	0.109796

The seasonal naïve forecast remains a practical yardstick in crude-oil analytics, yet empirical evidence shows it is generally surpassed by structured alternatives. Using monthly Kenyan crude-oil prices, Lumumba et al. [86] reported that ARIMA and ETS specifications achieved lower information-criteria scores and forecast errors than the naïve comparator, even though the latter remained valuable for quick reference and transparency.

A broader comparison of univariate models by Tularam and Saeed [87] reached the same conclusion, finding that exponential-smoothing and ARIMA variants consistently outperformed the naïve baseline across multiple accuracy measures when applied to international oil-price series. Retaining the seasonal naïve benchmark, therefore, enables policymakers and analysts to quantify the incremental benefits conferred by advanced hybrids such as the ARIMAX–LSTM fusion proposed in the present study.

Moving-average models have consistently served as credible baselines for evaluating more sophisticated forecasting techniques: they outperformed a tuned ARIMA specification in daily stock-price prediction [88], delivered accurate short-term projections of tourist arrivals [89], and, when the window length is optimally chosen, matched or exceeded other naïve benchmarks in formal time-series experiments [90].

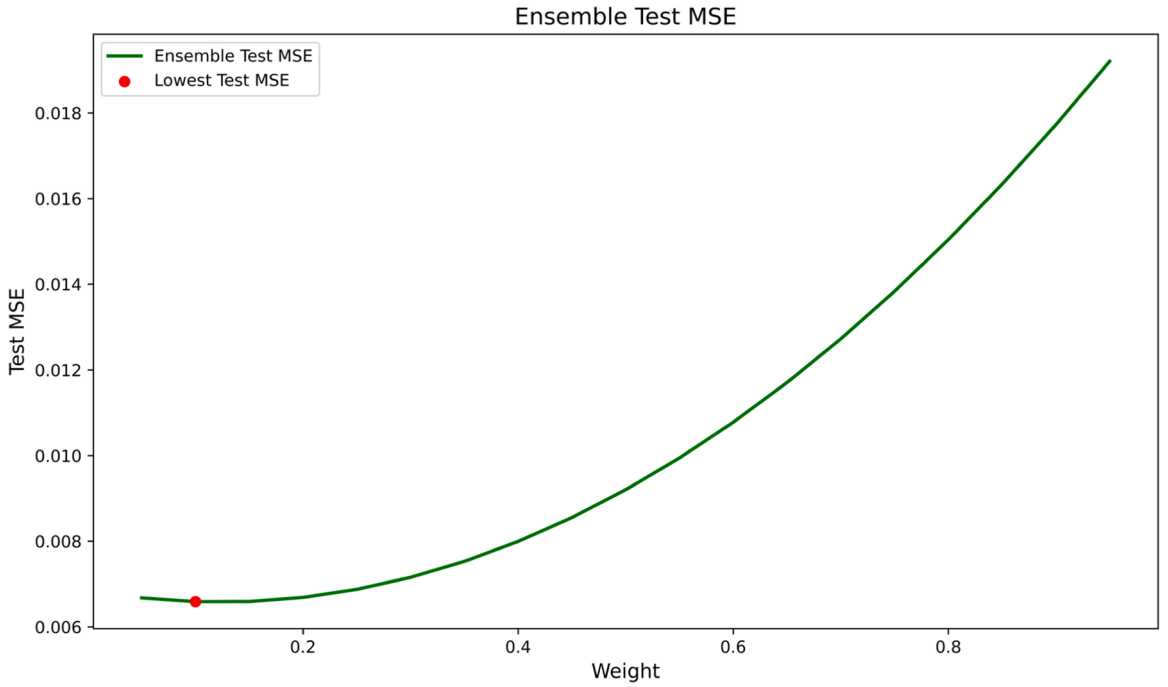


Fig. 6. Sensitivity Analysis of Ensemble Model Test MSE Across Weight Configurations.

Table 18
RMSE of Forecast Benchmarks.

Model / Benchmark	MSE (USD^2/bbl^2)	RMSE (USD/bbl)	MAE (USD/bbl)
ARIMAX	≈ 266.6	≈ 16.3	≈ 13.1
LSTM	≈ 0.991	≈ 9.95	≈ 8.37
Ensemble	≈ 0.960	≈ 9.79	≈ 8.34
12-Month Moving Avg	214.96	14.66	11.32
Seasonal Naïve	433.63	20.82	16.28

Across all error metrics, the neural-network approaches outperform the traditional statistical baselines. The ensemble registers the lowest RMSE (≈ 9.79 USD bbl^{-1}) and MAE (≈ 8.34 USD bbl^{-1}), representing about a 33 % reduction in RMSE relative to the 12-month moving-average benchmark and more than a 50 % drop versus the seasonal-naïve alternative. As illustrated by the results reported in Section 4.3, the ensemble's numerical advantage over the stand-alone LSTM is not only economically meaningful but also statistically significant; this indicates that the improvement stems from the effective combination of linear and nonlinear components rather than random variation. By contrast, the ARIMAX specification not only lags behind the neural models but even trails the simple moving average, suggesting that linear dynamics alone are insufficient for the market's pronounced nonlinear behavior.

5. Discussion and conclusion

Crude oil, WTI, is one of the most significant global commodities in determining economic stability and policymaking. The volatility of crude-oil prices (driven mainly by geopolitical risks, structural economic changes, and environmental policies) makes this prediction important in supporting accurate decisions. More sophisticated models, such as classical econometric methods and recent state-of-the-art machine-learning techniques, are needed to capture complex and poorly characterized dynamics in crude-oil prices. This research is motivated by the necessity of combining linear and nonlinear modeling in a solid forecasting framework that can help obtain better accuracy and generalization.

This study was guided by three central research questions, and the results provide distinct answers to each. First, regarding the comparative performance of machine-learning and econometric models, our findings indicate that the LSTM model is substantially more accurate and reliable than the traditional ARIMAX model for forecasting WTI crude-oil prices. This is evidenced by the LSTM model's significantly lower test-error metrics (e.g., Test MSE = 0.0068) compared to the ARIMAX model (Test MSE = 0.556), demonstrating its superior ability to capture the market's complex nonlinear dynamics. In response to the second question—whether combining these approaches yields a superior forecast—our investigation yielded a critical insight. While our proposed ARIMAX–LSTM ensemble demonstrated the highest accuracy on the single chronological test set (Test MSE = 0.0066), a more rigorous 5-

fold cross-validation revealed the standalone LSTM model to be the most robust performer on average. This discrepancy is a key finding of our study. It suggests that while the ensemble's advantage in the final chronological test set is statistically significant (DM = 6.076, $p < 0.001$), this superiority may be period-specific. The superior average performance of the standalone LSTM in cross-validation indicates that relying solely on one test set could be misleading. Finally, concerning the third question on the impact of parameter sensitivity, the analysis confirms that model accuracy is highly dependent on hyper-parameter tuning. The sensitivity analysis identified optimal configurations for each model—such as (4, 1, 4) for ARIMAX and 50 units with a 0.4 dropout rate for LSTM—and demonstrated that sub-optimal choices could drastically degrade performance. This underscores the critical need for rigorous parameter optimization to ensure model robustness.

Building on previous hybrid forecasting models, this study introduces a structured ARIMAX–LSTM ensemble optimized through grid search and validated via robustness testing. Unlike purely black-box approaches, it integrates an interpretable econometric component while addressing both linear and nonlinear price dynamics, thereby offering a practical tool for forecasting under volatile market conditions.

The findings of this research correspond with prior investigations that suggest a role for mixed methods. Studies by Wang and Wu [28] and Aamir et al. [31] showed the importance of using an ensemble of statistical and machine-learning models for better forecasting. Decomposition techniques are discussed by Lin and Sun [34] and Abdollahi [43] from the perspective of neural networks. This paper builds on these results by introducing a weighted ensemble model synthesizing deep learning and econometric approaches that lays a fuller foundation for robustness and sensitivity analysis.

Before outlining future directions, it is worth discussing the broader implications of our findings and limitations. Although the ensemble model exhibited statistically significant improvements in predictive accuracy over the standalone LSTM on the final chronological test set, its usefulness for policy applications must still be evaluated with caution. Even with this period-specific advantage, policymakers should augment ensemble forecasts with supplementary validation or scenario-based stress testing to guard against unforeseen market shifts. Nevertheless, the structured integration of ARIMAX and LSTM provides an adaptable framework that can serve as a preliminary forecasting tool for energy-policy planning, hedging strategies, and macro-economic simulations. Importantly, the sensitivity of model performance to hyper-parameter choices highlights a limitation of generalizability, particularly in the face of abrupt market-regime shifts or exogenous shocks. Future implementations in policy contexts should therefore consider integrating real-time model-adaptation mechanisms or diversifying ensemble components to enhance stability.

Research directions for future studies can evolve the model by including additional exogenous variables (geopolitical or environmental signals, etc.) that increase flexibility. Supplementary work into more advanced machine-learning architectures and real-time stress testing would provide the model with further reliability. This improvement would make the model more viable in facing the increasingly sophisticated new crude-oil-market forecasting challenges.

Data availability

Data will be made available on request.

References

[1] L. Kilian, Exogenous oil supply shocks: how big are they and how much do they matter for the U.S. Economy? Rev. Econ. Stat. 90 (2008) 216–240, https://doi.org/10.1162/rest.90.2.216.

[2] H.M. Ertuğrul, M.T. Kartal, S.K. Depren, U. Soytaş, Determinants of electricity prices in Turkey: an application of machine learning and time series models, Energ (Basel) 15 (2022) 7512, https://doi.org/10.3390/en15207512.

[3] S. Kılıç Depren, M.T. Kartal, H.M. Ertuğrul, Ö. Depren, The role of data frequency and method selection in electricity price estimation: comparative evidence from Turkey in pre-pandemic and pandemic periods, Renew. Energy 186 (2022) 217–225, https://doi.org/10.1016/j.renene.2021.12.136.

[4] I. Energy Agency, World Energy Outlook 2024 (2024). www.iea.org/terms.

[5] https://www.eia.gov, U.S. Energy Information Administration (n.d.).

[6] S. Mukhtarov, M. Azizov, M.T. Kartal, H. Eynalov, Catalyzing green transformation: mitigating oil price impact on CO2 emissions in Saudi Arabia via renewable energy transition, Environ. Econ. Policy Stud. (2024), https://doi.org/10.1007/s10018-024-00416-1.

[7] J. Alvarez-Ramirez, J. Alvarez, E. Rodriguez, Short-term predictability of crude oil markets: a detrended fluctuation analysis approach, Energy Econ. 30 (2008) 2645–2656, https://doi.org/10.1016/j.eneco.2008.05.006.

[8] W.E. Shambora, R. Rossiter, Are there exploitable inefficiencies in the futures market for oil? Energy Econ. 29 (2007) 18–27, https://doi.org/10.1016/j.eneco.2005.09.004.

[9] S. Aydın, K. Altun, Early prediction of fabric quality using machine learning to reduce rework in manufacturing processes, Int. J. Optim. Control Theor. Appl. (IJOCTA) 14 (2024) 308–321, https://doi.org/10.11121/ijocta.1462.

[10] H.H. Yildirim, M. Yavuz, Evaluation of wind energy investment with artificial neural networks, Int. J. Optim. Control: Theor. Appl. (IJOCTA) 9 (2019) 142–147, https://doi.org/10.11121/ijocta.01.2019.00780.

[11] E.N. Cici Karaboğa, G. Şekeroğlu, E. Kızıloğlu, K. Karaboğa, A.M. Acılar, Price prediction of dual-listed stocks with RF and LSTM algorithms: NYSE and BIST comparison, Math. Model Numer. Simul. Appl. 4 (2024) 207–230, https://doi.org/10.53391/mmnsa.1577228.

[12] H. Hüseyin Yildirim, A. Akusta, Key drivers of volatility in BIST100 firms using machine learning segmentation, Int. J. Optim. Control: Theor. Appl. (IJOCTA) 15 (2025) 180, https://doi.org/10.36922/ijocta.1707.

[13] O.O. Okundalaye, N. Özdemir, R.O. Awonusika, Early breast cancer prediction using optimized machine learning and tumor-immune modeling, J. Comput. Appl. Math. 473 (2026) 116875, https://doi.org/10.1016/j.cam.2025.116875.

[14] O.O. Okundalaye, N. Özdemir, F. Evirgen, Leveraging machine learning for early and accurate anaemia diagnosis: a comparative study of classification algorithms, in: 2025: pp. 42–52. https://doi.org/10.1007/978-3-031-90914-6_3.

[15] N. Tekbıyık-Ersoy, Modeling the renewable energy development in Türkiye with optimization, Int. J. Optim. Control: Theor. Appl. (IJOCTA) 15 (2025) 135, https://doi.org/10.36922/ijocta.1664.

[16] H.H. Yıldırım, Ö.F. Rençber, C. Yüksel Yıldırım, Ranking the determinants of financial performance using machine learning methods: an application to BIST energy companies, Math. Model Numer. Simul. Appl. 4 (2024) 165–186, https://doi.org/10.53391/mmnsa.1594426.

[17] T. Ulussever, S. Kılıç Depren, M.T. Kartal, Ö. Depren, Estimation performance comparison of machine learning approaches and time series econometric models: evidence from the effect of sector-based energy consumption on CO2 emissions in the USA, Environ Sci. Pollut Res. 30 (2023) 52576–52592, https://doi.org/10.1007/s11356-023-26050-0.
[18] L. Yu, S. Wang, K.K. Lai, Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm, Energy Econ. 30 (2008) 2623–2635, https://doi.org/10.1016/j.eneco.2008.05.003.
[19] A. Ghaffari, S. Zare, A novel algorithm for prediction of crude oil price variation based on soft computing, Energy Econ. 31 (2009) 531–536, https://doi.org/10.1016/j.eneco.2009.01.006.
[20] T. Mingming, Z. Jinliang, A multiple adaptive wavelet recurrent neural network model to analyze crude oil prices, J. Econ. Bus. 64 (2012) 275–286, https://doi.org/10.1016/j.jeconbus.2012.03.002.
[21] T. Xiong, Y. Bao, Z. Hu, Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices, Energy Econ. 40 (2013) 405–415, https://doi.org/10.1016/j.eneco.2013.07.028.
[22] M. Bildirici, Ö. Ersin, Forecasting volatility in oil prices with a class of nonlinear volatility models: smooth transition RBF and MLP neural networks augmented GARCH approach, Pet. Sci. 12 (2015) 534–552, https://doi.org/10.1007/s12182-015-0035-8.
[23] T. Li, M. Zhou, C. Guo, M. Luo, J. Wu, F. Pan, Q. Tao, T. He, Forecasting crude oil price using EEMD and RVM with adaptive PSO-based kernels, Energ (Basel) 9 (2016) 1014, https://doi.org/10.3390/en9121014.
[24] H. Miao, S. Ramchander, T. Wang, D. Yang, Influential factors in crude oil price forecasting, Energy Econ. 68 (2017) 77–88, https://doi.org/10.1016/j.eneco.2017.09.010.
[25] X. Li, J. Ma, S. Wang, X. Zhang, How does Google search affect trader positions and crude oil prices? Econ. Model. 49 (2015) 162–171, https://doi.org/10.1016/j.econmod.2015.04.005.
[26] M. Naderi, E. Khamehchi, B. Karimi, Novel statistical forecasting models for crude oil price, gas price, and interest rate based on meta-heuristic bat algorithm, J. Pet. Sci. Eng. 172 (2019) 13–22, https://doi.org/10.1016/j.petrol.2018.09.031.
[27] M. Bildirici, N. Guler Bayazit, Y. Ucan, Analyzing crude oil prices under the impact of COVID-19 by using LSTARGARCHLSTM, Energ (Basel) 13 (2020) 2980, https://doi.org/10.3390/en13112980.
[28] G. Wang, J. Wu, Crude Oil Price Forecasting Based On the ARIMA and BP Neural Network Combinatorial Algorithm, in: ICLEM 2012, American Society of Civil Engineers, Reston, VA, 2012, pp. 482–487, https://doi.org/10.1061/9780784412602.0075.
[29] H. Lyu, Y. Chang, Research on international crude oil price forecasting model, Int. J. New Dev. Eng. Soc. 1 (2017) 78–81, https://doi.org/10.25236/IJNDES.17325.
[30] R. Alquist, L. Kilian, What do we learn from the price of crude oil futures? J. Appl. Econ. 25 (2010) 539–573, https://doi.org/10.1002/jae.1159.
[31] M. Aamir, A. Shabri, M. Ishaq, Crude oil price forecasting by Ceemdan based hybrid model of Arima and Kalman filter, J. Teknol. 80 (2018), https://doi.org/10.11113/jt.v80.10852.
[32] S. Ramyar, F. Kianfar, Forecasting crude oil prices: a comparison between artificial neural networks and vector autoregressive models, Comput. Econ. 53 (2019) 743–761, https://doi.org/10.1007/s10614-017-9764-7.
[33] L.-T. Zhao, S.-G. Wang, Z.-G. Zhang, Oil price forecasting using a time-varying approach, Energ (Basel) 13 (2020) 1403, https://doi.org/10.3390/en13061403.
[34] H. Lin, Q. Sun, Crude oil prices forecasting: an approach of using CEEMDAN-based multi-layer gated recurrent unit networks, Energ (Basel) 13 (2020) 1543, https://doi.org/10.3390/en13071543.
[35] Q. Qin, K. Xie, H. He, L. Li, X. Chu, Y.-M. Wei, T. Wu, An effective and robust decomposition-ensemble energy price forecasting paradigm with local linear prediction, Energy Econ. 83 (2019) 402–414, https://doi.org/10.1016/j.eneco.2019.07.026.
[36] L. Tang, C. Zhang, L. Li, S. Wang, A multi-scale method for forecasting oil price with multi-factor search engine data, Appl. Energy 257 (2020) 114033, https://doi.org/10.1016/j.apenergy.2019.114033.
[37] Ö. Arslan, A machine learning approach for voice pathology detection using mode decomposition-based acoustic cepstral features, Math. Model Numer. Simul. Appl. 4 (2024) 469–494, https://doi.org/10.53391/mmnsa.1473574.
[38] Y. Zhang, Y. Wei, Y. Zhang, D. Jin, Forecasting oil price volatility: forecast combination versus shrinkage method, Energy Econ. 80 (2019) 423–433, https://doi.org/10.1016/j.eneco.2019.01.010.
[39] F. Villada, D. Arroyave, M. Villada, Pronóstico del Precio del Petróleo mediante Redes Neuronales Artificiales, Información Tecnológica 25 (2014) 145–154, https://doi.org/10.4067/S0718-07642014000300017.
[40] S. Deng, A. Sakurai, Crude oil spot price forecasting based on multiple Crude oil markets and timeframes, Energ (Basel) 7 (2014) 2761–2779, https://doi.org/10.3390/en7052761.
[41] F. Cheng, T. Li, Y. Wei, T. Fan, The VEC-NAR model for short-term forecasting of oil prices, Energy Econ. 78 (2019) 656–667, https://doi.org/10.1016/j.eneco.2017.12.035.
[42] A. Hadjira, H. Salhi, L. Choubar, OPEC basket monthly crude oil price forecasting: comparative study between Prophet Facebook, NNAR, FTS models, Comput. Econ. (2024), https://doi.org/10.1007/s10614-024-10762-7.
[43] H. Abdollahi, A novel hybrid model for forecasting crude oil price based on time series decomposition, Appl. Energy 267 (2020) 115035, https://doi.org/10.1016/j.apenergy.2020.115035.
[44] T. Gunarto, R. Azhar, N. Tresiana, S. Supriyanto, A. Ahadiat, Accurate estimated model of volatility crude oil price, Int. J. Energy Econ. Policy 10 (2020) 228–233, https://doi.org/10.32479/ijeep.9513.
[45] P. Mishra, A probabilistic weighted ensemble algorithm, in: 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE (2018) 1–4, https://doi.org/10.1109/CCAA.2018.8777731.
[46] P.N. Jha, M. Cucculelli, A new model averaging approach in predicting credit risk default, Risks 9 (2021) 114, https://doi.org/10.3390/risks9060114.
[47] C. Ren, Y. Xu, Robustness verification for machine-learning-based power system dynamic security assessment models under adversarial examples, IEEe Trans. Control Netw. Syst. 9 (2022) 1645–1654, https://doi.org/10.1109/TCNS.2022.3145285.
[48] J. Chuah, U. Kruger, G. Wang, P. Yan, J. Hahn, Framework for testing robustness of machine learning-based classifiers, J. Pers. Med. 12 (2022) 1314, https://doi.org/10.3390/jpm12081314.
[49] L. Gosch, D. Sturm, S. Geisler, S. Günnemann, Revisiting robustness in graph machine learning, ICLR (2023). http://arxiv.org/abs/2305.00851.
[50] P. Sohrabi, H. Dehghani, R. Rafie, Forecasting of WTI crude oil using combined ANN-Whale optimization algorithm, Energy Sources B: Econ Plan Policy 17 (2022), https://doi.org/10.1080/15567249.2022.2083728.
[51] T. Yin, Y. Wang, Predicting the price of WTI crude oil using ANN and chaos, Sustainability. 11 (2019) 5980, https://doi.org/10.3390/su11215980.
[52] C. Deng, L. Ma, T. Zeng, Crude oil price forecast based on deep transfer learning: shanghai Crude oil as an example, Sustainability. 13 (2021) 13770, https://doi.org/10.3390/su132413770.
[53] J. Liu, X. Huang, Forecasting crude oil price using event extraction, IEEe Access. 9 (2021) 149067–149076, https://doi.org/10.1109/ACCESS.2021.3124802.
[54] J. Nasir, M. Aamir, Z.U. Haq, S. Khan, M.Y. Amin, M. Naeem, A new approach for forecasting crude oil prices based on stochastic and deterministic influences of LMD using ARIMA and LSTM models, IEEe Access. 11 (2023) 14322–14339, https://doi.org/10.1109/ACCESS.2023.3243232.
[55] B. Jin, X. Xu, Y. Zhang, Peanut oil price change forecasts through the neural network, Foresight. 27 (2025) 595–612, https://doi.org/10.1108/FS-01-2023-0016.
[56] L. Jovanovic, D. Jovanovic, N. Bacanin, A. Jovancai Stakic, M. Antonijevic, H. Magd, R. Thirumalaisamy, M. Zivkovic, Multi-step crude oil price prediction based on LSTM approach tuned by salp swarm algorithm with disputation operator, Sustainability. 14 (2022) 14616, https://doi.org/10.3390/su142114616.
[57] J. Jo, U. Kim, E. Lee, J. Lee, S. Kim, A supply chain-oriented model to predict crude oil import prices in South Korea based on the hybrid approach, Sustainability. 15 (2023) 16725, https://doi.org/10.3390/su152416725.
[58] D.H. Vo, T.N. Vu, A.T. Vo, M. McAleer, Modeling the relationship between crude oil and agricultural commodity prices, Energ (Basel) 12 (2019) 1344, https://doi.org/10.3390/en12071344.

[59] A. Sen, An analysis of crude oil prices in the last decade (2011-2020): with deep learning approach, SSRN Electron J. (2021), https://doi.org/10.2139/ssrn.3906517.
[60] M.T. Kartal, The effect of the COVID-19 pandemic on oil prices: evidence from Turkey, Energy RES LETT 1 (2021), https://doi.org/10.46557/001c.18723.
[61] T.S. Adebayo, M.T. Kartal, Effect of green bonds, oil prices, and COVID-19 on industrial CO2 emissions in the USA: emissions in the USA: evidence from novel wavelet local multiple correlation approach, Energy Environ 35 (2024) 3273–3296, https://doi.org/10.1177/0958305X231167463.
[62] Y. Wang, S. Zhu, C. Li, Research on Multistep time series prediction based on LSTM, in: 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), IEEE (2019) 1155–1159, https://doi.org/10.1109/EITCE47263.2019.9095044.
[63] M. Razaviyayn, T. Huang, S. Lu, M. Nouiehed, M. Sanjabi, M. Hong, Nonconvex min-max optimization: applications, challenges, and recent theoretical advances, IEEe Signal. Process. Mag. 37 (2020) 55–66, https://doi.org/10.1109/MSP.2020.3003851.
[64] L.B.V. de Amorim, G.D.C. Cavalcanti, R.M.O. Cruz, The choice of scaling technique matters for classification performance, Appl. Soft. Comput. 133 (2023) 109924, https://doi.org/10.1016/j.asoc.2022.109924.
[65] L.B.V. de Amorim, G.D.C. Cavalcanti, R.M.O. Cruz, Meta-Scaler: a Meta-learning framework for the selection of scaling techniques, IEEE Trans. Neural Netw. Learn. Syst. 36 (2025) 4805–4819, https://doi.org/10.1109/TNNLS.2024.3366615.
[66] R. ElShawi, Y. Sherif, M. Al-Mallah, S. Sakr, Interpretability in healthcare: a comparative study of local machine learning interpretability techniques, Comput. Intell. 37 (2021) 1633–1650, https://doi.org/10.1111/coin.12410.
[67] G.E.P. Box, G.M. Jenkins, G.C. +Ljung, G.M. Reinsel, Time Series Analysis: Forecasting and Control, Fifth Editiom, John Wiley & Sons, 2016.
[68] M. Yang, J. Xie, P. Mao, C. Wang, Z. Ye, Application of the ARIMAX Model on Forecasting Freeway Traffic Flow, in: CICTP 2017, American Society of Civil Engineers, Reston, VA, 2018, pp. 593–602, https://doi.org/10.1061/9780784480915.061.
[69] A. Kabovic, M. Kabovic, S. Bostjancic Rakas, V. Timcenko, The influence of the input parameters variation of the non-seasonal ARIMAX model on the accuracy of meteorological parameters forecasting, in: 2022 30th Telecommunications Forum (TELFOR), IEEE (2022) 1–4, https://doi.org/10.1109/TELFOR56187.2022.9983670.
[70] W. Wanishsakpong, B.E. Owusu, Optimal time series model for forecasting monthly temperature in the southwestern region of Thailand, Model Earth Syst. Env. 6 (2020) 525–532, https://doi.org/10.1007/s40808-019-00698-5.
[71] Q. Xu, W. Li, D. Kong, X. Zhao, X. Wang, Y. Li, Y. Shen, X. Wang, Z. Zhao, Ultra-short-term wind speed forecast based on WD-ARIMAX-GARCH model, in: 2019 IEEE 2nd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), IEEE (2019) 219–222, https://doi.org/10.1109/AUTEEE48671.2019.9033198.
[72] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[73] A. Ravikumar, H. Sriraman, A deep understanding of long short-term memory for solving vanishing error problem, in: 2023: pp. 74–90. https://doi.org/10.4018/978-1-6684-8531-6.ch004.
[74] Y. Jianying, Z. Xuefei, L. Shiqiang, X. Shaoze, W. Ajun, Z. Yalei, Study on five typical optimization models of LSTM, in: 2023 China Automation Congress (CAC), IEEE (2023) 298–303, https://doi.org/10.1109/CAC59555.2023.10450860.
[75] A.G. Timmermann, Forecast Comb (2005). www.cepr.org.
[76] J. Zhao, A. Takai, E. Kita, Weight-training ensemble model for stock price forecast, in: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2022, pp. 1–6, https://doi.org/10.1109/ICDMW58026.2022.00024.
[77] G. Casalicchio, L. Burk, Evaluation and benchmarking. Applied Machine Learning Using Mlr3 in R, Chapman and Hall/CRC, Boca Raton, 2023, pp. 53–82, https://doi.org/10.1201/9781003402848-3.
[78] K. Arthi, V. Sankaradass, N. Parveen, J. Muralidharan, 7 Methods of cross-validation and bootstrapping, in: toward artificial general intelligence, Gruyter (2023) 145–166, https://doi.org/10.1515/9783111323749-007.
[79] A. Valier, The cross validation in automated valuation models: a proposal for use, in: 2020: pp. 585–596. https://doi.org/10.1007/978-3-030-58814-4_45.
[80] B. Zhu, Y. Liu, General approximate cross validation for model selection, in: Proceedings of the 29th ACM International Conference on Multimedia, ACM, New York, NY, USA, 2021, pp. 5281–5289, https://doi.org/10.1145/3474085.3475649.
[81] B. Van Stein, E. Raponi, Z. Sadeghi, N. Bouman, R.C.H.J. Van Ham, T. Bäck, A comparison of global sensitivity analysis methods for explainable AI with an application in genomic prediction, IEEE Access. 10 (2022) 103364–103381, https://doi.org/10.1109/ACCESS.2022.3210175.
[82] L. Liu, C. Zhou, Y. Gao, Sensitivity Analysis of Microscopic Traffic Simulation Model Parameters, in: CICTP 2020, American Society of Civil Engineers, Reston, VA, 2020, pp. 1579–1589, https://doi.org/10.1061/9780784482933.135.
[83] A.T. Tunkiel, D. Sui, T. Wiktorski, Data-driven sensitivity analysis of complex machine learning models: a case study of directional drilling, J. Pet. Sci. Eng. 195 (2020) 107630, https://doi.org/10.1016/j.petrol.2020.107630.
[84] F. Zhang, K. Luo, W. Zhai, S. Tan, Y. Wang, Non-probabilistic parameter sensitivity analysis for structures based on ellipsoidal model, Adv. Mech. Eng. 10 (2018), https://doi.org/10.1177/1687814018782362.
[85] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 13 (1995) 253–263, https://doi.org/10.1080/07350015.1995.10524599.
[86] V.W. Lumumba, T.W. Mutugi, A. Wagala, Forecasting of monthly crude oil prices in Kenya using comparative time series models, Asian J. Probab. Stat. 26 (2024) 97–109, https://doi.org/10.9734/ajpas/2024/v26i9648.
[87] G.A. Tularam, T. Saeed, Oil-price forecasting based on various univariate time-series models, Am. J. Oper. Res. 06 (2016) 226–235, https://doi.org/10.4236/ajor.2016.63023.
[88] C.P. Tsokos, K-th moving, weighted and exponential moving average for time series forecasting models, Eur. J. Pure Appl. Math. 3 (2010) 406–416.
[89] I. Zoran, M. Ace, N. Zoran, Time series forecasting using a moving average model for extrapolation of number of tourist, UTMS J. Econ. 9 (2018) 121–132.
[90] I. Svetunkov, F. Petropoulos, Old dog, new tricks: a modelling view of simple moving averages, Int. J. Prod. Res. 56 (2018) 6034–6047, https://doi.org/10.1080/00207543.2017.1380326.