Research Article

GAZİ
JOURNAL OF ENGINEERING SCIENCES

# Explainable and Robust Oil Price Forecasting Using Hybrid Deep Learning and SHAP Interpretability: A Long-Term Analysis of WTI Prices from 1986 to 2025

**Onur Polat[a], Neslihan Parlak[b], Esra Söğüt[*c], Eda Fendoğlu[d], Muammer Türkoğlu[e]**

**Submitted:** 22.04.2026 **Revised :** 09.06.2026 **Accepted:** 17.06.2026 doi:10.30855/gmbd.070526A05

[a] Bingöl University, Dept. of Computer Engineering - Bingöl, Türkiye, Orcid: 0000-0001-9313-4910
[b] Samsun University, Dept. of Software Engineering - Samsun, Türkiye, Orcid: 0009-0006-4248-9504
[c,*] Gazi University, Fac. of Technology, Dept. of Computer Engineering - Ankara, Türkiye, Orcid: 0000-0002-0051-2271
[d] Malatya Turgut Özal University, Social and Human Sciences Faculty, Dept. of International Business Management – Malatya, Türkiye, Orcid: 0000-0003-4092-7137
[e] Samsun University, Dept. of Software Engineering - Samsun, Türkiye, Orcid: 0000-0002-2377-4979
[*] *Corresponding author: esrasogut@gazi.edu.tr*

**Keywords:** Crude oil price forecasting, deep learning, time series analysis, SHAP analysis, energy markets

## ABSTRACT

Crude oil is an essential input for strategic sectors such as petrochemicals, industrial production, transportation, defense, and manufacturing. Fluctuations in crude oil prices affect not only energy markets but also global financial stability and policymaking. Therefore, accurate crude oil price forecasting is crucial for maintaining energy market stability, supporting sustainable energy policies, and improving economic planning. This study proposes a hybrid deep learning model that combines Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Multi-Head Attention mechanisms for crude oil price forecasting. The proposed architecture uses multi-scale convolutional layers to capture temporal patterns at different scales. BiLSTM layers learn sequential dependencies, while the attention mechanism highlights the most informative features. The dataset includes West Texas Intermediate crude oil spot prices and macroeconomic and energy market indicators from January 1986 to March 2025. This period covers approximately 39 years of market history. The proposed model achieved an RMSE of 0.1994, an MAE of 0.1641, and an $R^2$ score of 0.9224 on the test set. Model robustness was further evaluated under different market regimes, including the COVID-19 shock, the recovery period, and the post-pandemic market environment. SHapley Additive Explanations (SHAP) analysis was also performed to identify the most influential predictors. The findings indicate that the Consumer Price Index, jet fuel prices, gasoline prices, diesel prices, and other energy-related variables are among the most important predictors of crude oil prices. Overall, the proposed framework effectively captures the nonlinear behavior of crude oil markets while providing transparent and interpretable predictions.

## 1. Introduction

Energy is one of the cornerstones of the economic and social infrastructure of modern societies. Demand for energy in industry, transportation, and many areas of daily life is growing rapidly, especially in developing countries. A large portion of global energy demand is still met by non-renewable energy sources. Crude oil, which stands out among these sources, is of critical importance both for energy production and for the petrochemical industry [1].

**To cite this article**: O. Polat, N. Parlak, E. Söğüt, E. Fendoğlu, and M. Türkoğlu, "Explainable and Robust Oil Price Forecasting Using Hybrid Deep Learning and SHAP Interpretability: A Long-Term Analysis of WTI Prices from 1986 to 2025", *Gazi Journal of Engineering Sciences,* vol.12, no.2, pp. 201-220, 2026. doi:10.30855/gmbd.070526A05

Crude oil is a strategic commodity that drives economic growth on a global scale. Changes in oil prices directly affect not only energy markets but also exchange rates, inflation rates, stock markets, and overall macroeconomic balances [2]. Therefore, accurately predicting the future course of oil prices has become a critical necessity for investors, policymakers, and the energy sector. However, since crude oil prices fluctuate based on numerous dynamic and unpredictable factors, this forecasting process is highly complex and fraught with uncertainty [3].

In the literature, crude oil price forecasting has generally been addressed through time series modeling, econometric methods, and, more recently, artificial intelligence-based approaches. While traditional econometric models perform well under certain assumptions, they fall short in fully capturing the nonlinear behavior and irregular events characteristic of oil prices [4]. For example, statistical methods such as GARCH and ARIMA are useful for short-term fluctuations but are weak in the face of unpredictable shocks such as sudden crises, geopolitical tensions, or natural disasters [5].

In this context, machine learning and deep learning techniques have gained significant attention in recent years and have yielded promising results in oil price forecasting. Thanks to their ability to learn non-linear patterns and handle large data sets, these models can produce more accurate predictions than traditional models. In particular, algorithms such as support vector machines (SVM) and long short-term memory (LSTM) demonstrate strong performance in both short-term and medium-term predictions [5].

Predicting crude oil prices using deep learning-based methods and artificial intelligence algorithms is of great importance in today's energy economics research. In this context, the aim of this study is to forecast crude oil prices by developing convolutional neural networks (CNN)-based model to overcome the limitations of traditional methods defined in the literature and to make predictions with higher accuracy. Within this scope, factors that could be effective in predicting oil prices were analyzed, taking into account historical price data, macroeconomic indicators, and important global developments. Unlike many previous studies that rely on relatively short observation periods, this study utilizes a comprehensive monthly dataset covering the period from January 1986 to March 2025, corresponding to approximately 39 years of market history. This period includes several major structural breaks and market disruptions, such as the 2008 Global Financial Crisis, the 2014 oil price collapse, the COVID-19 pandemic, and the subsequent energy market recovery period. Therefore, the proposed framework is evaluated under multiple market regimes, providing a more comprehensive assessment of forecasting capability and robustness. The performance of the developed model was evaluated comparatively with different machine learning algorithms, and the results were interpreted in a way that would contribute to both decision-makers and academic circles.

The main contributions of this study stand out in several respects. First, a novel hybrid deep learning model combining CNN, BiLSTM, and the Attention mechanism has been proposed. This architecture captures both short-term patterns and long-term dependencies as well as contextual relationships in time series simultaneously, providing higher accuracy compared to traditional econometric and machine learning methods. Second, the study utilizes a long-term dataset covering the period from 1986 to 2025, enabling the proposed model to be evaluated across multiple market regimes and structural breaks. Third, the model's performance has been compared with different machine learning algorithms under a strict chronological leakage-safe evaluation protocol, providing a more realistic assessment of forecasting performance. Fourth, additional robustness analyses were conducted on the COVID-19 shock period and the subsequent recovery period to evaluate the model's generalization capability under extreme market conditions. Fifth, the model's prediction process has been made transparent through SHapley Additive Explanations (SHAP) analysis, and the relative effects of the variables used on the predictions have been quantitatively explained. This determined that indicators such as jet fuel, gasoline prices, industrial production index, and exchange rates play a critical role in predicting oil prices. Finally, the developed model not only contributes to the academic literature in terms of methodological diversity but also produces practical value by providing more reliable and applicable forecasts to decision-makers operating in the energy market.

## 2. Related Works

Predicting crude oil prices is of great importance for global economic stability, financial markets, and the sustainable planning of energy policies. Therefore, numerous studies have been conducted in the literature on forecasting crude oil prices. These studies are generally grouped into three main categories: traditional econometric models, machine learning methods, and deep learning-based hybrid approaches.

Classic econometric models such as ARIMA, VAR, and GARCH can produce successful forecasts during stable periods under linear assumptions, but they cannot adequately represent the nonlinear and sudden change-sensitive dynamics inherent in oil prices. This shortcoming has led researchers to more flexible, adaptable, and data-driven artificial intelligence-based models. Today, machine learning and deep learning-based methods can provide more accurate predictions than traditional models thanks to their ability to learn complex patterns and handle large data sets.

One of the early examples of this trend, Li et al. (2019) proposed an innovative deep learning approach that combines sentiment information extracted from online news texts with financial indicators to predict crude oil prices [2]. The study used 6,756 news headlines collected from Investing.com's "Crude Oil News" section, along with daily West Texas Intermediate (WTI) spot price, futures price, dollar index, and Dow Jones index data. When qualitative information obtained through CNN-based feature extraction and LDA topic modeling was combined with financial data, a significant increase in prediction accuracy was observed. In particular, Random Forest (RF) and Support Vector Regression (SVR) models demonstrated higher performance compared to predictions based solely on financial data. This study shows that qualitative data sources can be a powerful complementary element in oil price forecasting [2]. An et al. (2019) developed a hybrid wavelet decomposition model based on CEEMDAN–LSTM to forecast Brent and WTI oil prices [5]. The model aims to capture the nonlinear and multiscale nature of the series by combining GARCH and wavelet components. The findings revealed that the proposed hybrid structure outperformed other methods, with Mean Absolute Error (MAE)=1.631 and Root Mean Squared Error (RMSE)=1.912 for WTI, and MAE=1.587 and RMSE=1.833 for Brent. These results demonstrate that deep learning models based on wavelet decomposition are effective in modeling complex market dynamics [5]. Shambulingappa (2020) compared different deep learning architectures for crude oil price forecasting [6]. In the study, which evaluated CNN–LSTM and CNN–gated recurrent unit (GRU) models, it was determined that the CNN–GRU architecture yielded the lowest error rate with RMSE ≈ 0.28 and MAE ≈ 0.19. This result shows that integrating CNN-based feature extraction into time series models increases prediction accuracy [6]. To overcome these shortcomings, Obite et al. (2021) compared classical and machine learning models to estimate oil production volumes in Nigeria [7]. In the study, which tested ARIMA, Artificial Neural Network (ANN), and RF models, the ANN model provided the highest accuracy with values of RMSE=0.010, Mean Squared Error (MSE)=0.0001, and R-squared ($R^2$)=0.994. The findings revealed that ANNs outperformed traditional statistical methods in production processes dominated by nonlinear relationships [7].

Similarly, He et al. (2022) proposed a hybrid classification approach based on multi-feature fusion (MDF) to predict crude oil price direction [3]. The study compared MDF–SVM–TS, MDF–MLP–TS, and MDF–MLR–TS models, and the MDF–SVM–TS model yielded the best results with 36.4% accuracy, Macro-F1=0.341, and Macro-Recall ≈ 0.35. This research has contributed a new perspective to the literature by focusing on price direction classification beyond the price level [3]. Guliyev and Mustafayev (2022) evaluated Logistic Regression (LR), Decision Tree (DT), RF, AdaBoost, and XGBoost algorithms to classify the direction of WTI prices during the 1991–2021 period [1]. According to the results of the study, the XGBoost model showed the best performance with Sensitivity=0.8750, Specificity=0.9388, Accuracy=0.9178, and AUC=0.9694. Furthermore, according to SHAP analysis, the most influential variables include RefinCost, WTIFutP, USDX, Unemployment Rate, and IPCPN [1]. Jahandoost et al. (2023) developed hybrid attention-based deep neural network models considering 39 different features at daily, weekly, and monthly intervals. The study determined that the proposed model had a lower error rate on the test set compared to the best single-period model and that the multi-scale learning approach more effectively captured the complex dependencies in the time series [8]. Jha et al. (2024) compared the performance of different regression models using eight macroeconomic variables (oil supply, demand, natural gas price, exchange rates, gold and silver prices, etc.) in a dataset consisting of daily data from 2009 to 2021 [4]. The study evaluated support vector regression (SVR), Ridge, Lasso, and multivariate regression methods. According to the results, the SVR model (RBF kernel) showed the highest performance with $R^2$ = 0.9951, RMSE = 0.0766, MAPE = 4.25, and MSE = 0.005. These findings reveal that SVR-based models incorporating macroeconomic indicators provide high accuracy in oil price forecasting and achieve lower error rates compared to classical regression methods [4].

In the study by Foroutan and Lahmiri (2024), various deep learning and machine learning models were compared to predict crude oil and precious metal prices [9]. Analyses conducted on WTI, Brent, gold, and silver price data from January 4, 2000, to March 25, 2022, revealed that long short-term memory (LSTM), bidirectional LSTM (BiLSTM), gated recurrent unit (GRU), BiGRU, T2V-BiLSTM, T2V-BiGRU, convolutional neural networks (CNN), CNN-BiLSTM, CNN-BiGRU, temporal convolutional network (TCN), TCN-BiLSTM, and TCN-BiGRU. The findings report that the TCN model showed the best performance for WTI, Brent, and silver prices, while LightGBM yielded comparable results to TCN [9]. Building on this line of work, Bi et al. (2024) proposed the VMD-BiTCN-BiGRU-Attention hybrid model based on variational mode decomposition [10]. Analyses conducted on a dataset covering the years 1990–2024 determined that the proposed model provides a distinct advantage in processing non-stationary and

non-linear time series and outperforms many methods in the literature in terms of prediction accuracy [10]. Finally, Zeng et al. (2025) developed the hybrid GARCH–BiLSTM–KAN (GBK-Net) model using 39 years (1986–2025) of daily WTI price data [11]. Comparative analyses reported that the GBK-Net model minimized prediction error and provided a statistically significant advantage over GARCH, EGARCH, LSTM, and CNN–LSTM–KAN models [11].

A review of the literature generally shows that deep learning and hybrid models offer high accuracy in oil price forecasting. However, a significant portion of existing studies focus on a single model architecture, do not sufficiently address explainability components (e.g., SHAP, LIME), and often evaluate forecasting performance using relatively short observation periods. Furthermore, although recent studies have demonstrated promising results using Transformer-based architectures and graph neural network frameworks in complex forecasting tasks, comprehensive comparisons involving these approaches remain limited in the context of long-term crude oil price forecasting. In this context, the proposed multi-scale CNN-BiLSTM-Attention-based hybrid model effectively models nonlinear dynamics by capturing both short- and long-term dependencies while enhancing the transparency of predictions through SHAP-based explainability analysis. In addition, the use of a long-term dataset spanning 1986–2025 and a strict chronological leakage-safe evaluation protocol enables a more realistic assessment of forecasting performance under different market regimes and structural breaks. Thus, the study fills methodological gaps in the literature and offers a reliable, interpretable, and highly accurate forecasting approach for decision support processes in energy markets. Future studies may investigate whether Transformer-based models and graph neural network architectures can further improve forecasting performance by capturing long-range temporal dependencies and complex inter-variable relationships.

## 3. Proposed Methodology

In this study, a hybrid deep learning model combining CNN, BiLSTM, and multi-head attention mechanisms is proposed for the purpose of predicting oil prices. The general schematic representation of the model is shown in Figure 1. This structure is designed to extract local features from time series data, learn long-term dependencies, and focus on critical time steps.

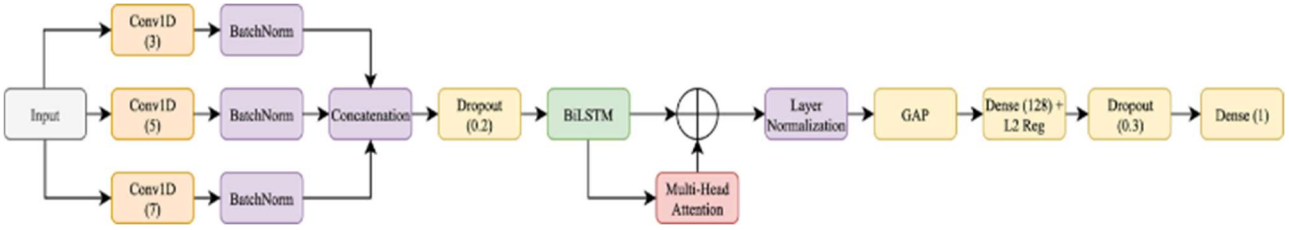



Figure 1. The schema of proposed regression model

The general schematic representation of the model proposed in this study is shown in Figure 1. The general process steps of the proposed system given in Figure 1 are as follows:

- Data Preparation and Normalization: Raw data is normalized before starting model training. This step prevents variables with different scales from negatively affecting model learning and speeds up the training process.
- Feature Extraction (Multi-Scale Convolutional Layers): The input time series is passed through three parallel 1D Convolutional layers with kernel sizes of 3, 5, and 7. These layers' capture short-, medium-, and long-term local patterns and extract meaningful features from the data. The outputs obtained after convolution are combined to enable the model to learn information at different scales simultaneously.
- Regularization (Dropout 0.2): Dropout is applied after feature combination to prevent overfitting. This method supports the model to learn in a more generalizable way.
- Learning Time Dependencies (BiLSTM Layer): Feature vectors are transferred to a BiLSTM layer. This structure captures both forward and backward dependencies in the time series, enabling the learning of long-term relationships.
- Attention Mechanism (Multi-Head Attention + Residual Connection): BiLSTM outputs are fed into a multi-head attention mechanism, placing greater weight on critical time steps. The BiLSTM and attention outputs are combined via a residual connection, followed by layer normalization. This step enables the model to learn contextual information more effectively.

- Feature Vector Compression (Global Average Pooling - GAP): The time dimension is reduced to a fixed-length vector using GAP. This reduces the number of parameters while preserving essential information.
- Dense Layers (Dense + L2 Reg.): The GAP output is fed into a dense layer with 128 neurons. In this layer, ReLU activation and L2 regularization are applied to increase the model's learning capacity and prevent overfitting. Then, Dropout (0.3) is applied to strengthen the regularization.
- Output Layer (Dense 1): In the final stage, regression estimation is performed using a single-neuron linear output layer. This value is obtained as the predicted value of oil prices.

These steps are designed to improve the accuracy of the model and create a more effective regression model. Details and theoretical background of these steps are provided in the subheadings.

### 3.1. Normalization

Prior to training the model, all variables were normalized to prevent scale differences from negatively affecting the learning process and to ensure faster and more stable convergence of the model. In this study, Z-score normalization, which is widely used and provides statistical stability, especially in machine learning models, was preferred [12]. This normalization method is defined as follows:

$$x' = \frac{x-\mu}{\sigma} \quad (1)$$

Here, x represents the raw data point, μ represents the mean of the data set, δ represents the standard deviation, and $x^1$ represents the normalized value. This allows the data to be scaled so that its mean is 0 and its standard deviation is 1.

### 3.2. Multi-scale convolutional feature extraction

To capture both the short-term fluctuations and longer-term structural patterns of crude oil price time series, the model employs three parallel one-dimensional (1D) convolutional layers. These layers were configured with kernel sizes of 3, 5, and 7, respectively, to extract information from different time windows simultaneously. Each convolutional block contained 128 filters and applied the ReLU activation function to ensure nonlinear transformation. Convolution is generally expressed in the following form [13-15]:

$$y_i = \sum_{k=1}^{K} x_{i+k-1}.w_k + b \quad (2)$$

Here, x represents the input time series, $w_k$ represents the filter weights, b represents the bias, and $y_i$ represents the convolution output. The feature maps obtained from each of the three convolutional layers are concatenated together to combine multi-scale information into a single representation. Dropout is applied to this combined feature vector at a rate of 0.2 to reduce overfitting and increase the model's generalization capacity.

### 3.3. Bidirectional long short-term memory

The multi-scale features extracted by the convolutional layers are fed into the BiLSTM layer to more effectively model the sequential dependencies of the time series. The BiLSTM structure creates a richer representation by learning the relationships of the time series both from the past to the future and from the future to the past simultaneously.

The basic updates in an LSTM cell are defined as follows [16-19]:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_t - 1, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_t - 1, x_t] + b_i) \\ \tilde{C}_t &= \tan h\,(W_c \cdot [h_t - 1, x_t] + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_t - 1, x_t] + b_o) \\ h_t &= o_t \odot \tan h\,(C_t) \end{aligned} \quad (3)$$

Here, $f_t$ is the forget gate, $i_t$ is the input gate, $o_t$ is the output gate, $C_t$ is the cell state, and $h_t$ is the hidden state. These gates enable the network to regulate long-term information flow and reduce problems such as gradient vanishing.

In the BiLSTM architecture, the time series is processed both forward and backward. The hidden state ($h_t$) -> obtained in the forward direction and the hidden state ($h_t$) <- obtained in the backward direction are combined to form the final representation:

$$h_t = Concat\left(\overrightarrow{h_t}, \overleftarrow{h_t}\right) \quad (4)$$

Thanks to this structure, the model can access more comprehensive contextual information compared to unidirectional LSTMs and demonstrates stronger performance, particularly in data structures with complex dependencies such as economic time series.

### 3.4. Multi-head attention

The sequential feature representations obtained from the BiLSTM layer are fed into the Multi-Head Attention mechanism to learn the relative importance of different time steps in the time series [16,20,21]. This mechanism enables the model to assign higher weights to time steps carrying critical information while reducing the impact of less important steps, thereby improving the learning process.

The basic attention function is defined as follows [20-22]:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Here, the (query), K(key), and V(value) matrices are derived from BiLSTM outputs. The scaled dot-product attention function is preferred because it provides computational efficiency and stability in high-dimensional representations.

The Multi-Head Attention mechanism is achieved by running multiple independent attention heads in parallel and is expressed as follows:

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O \quad (6)$$

Here, each $(head)_i$ enables the model to create a richer and more versatile contextual representation by learning the input in a different projection space. After the outputs of the heads are combined, a residual projection is applied using an output weight matrix $W^O$.

Following the attention mechanism, residual connections and layer normalization are now used to prevent problems such as gradient vanishing and learning breakdown while increasing the depth of the model:

$$\widehat{H} = LayerNorm(H + MultiHead(H,H,H)) \quad (7)$$

This structure prevents information loss by combining the attention output with the original representation, while also increasing the stability of the learning process [20-23]. Thus, the model is able to capture the contextual relationships of the time series more consistently and improve its overall performance.

### 3.5. Global average pooling and dense layers

The time series representation obtained from the attention mechanism was subjected to GAP to reduce the number of model parameters and fix the time dimension [23]. GAP produces a more compact and generalizable representation by

taking the average of each feature across all time steps. The operation is defined as follows:

$$z = \frac{1}{T}\sum_{t=1}^{T} h_t \quad (8)$$

Here, $h_t$, represents the output of the attention mechanism at the $t_{th}$ time step; T denotes the total number of time steps. Thanks to GAP, the model minimizes information loss while reducing the variability on the time axis to a single vector. This vector is passed through a dense layer with 128 neurons, a ReLU activation function, and L2 regularization to increase learning capacity [23]:

$$h = \mathrm{ReLU}(Wz + b) \quad (9)$$

Here, W represents the weight matrix and b represents the bias term. L2 regularization prevents the weights from growing excessively, making the model more resistant to overfitting. Following this layer, Dropout was applied at a rate of 0.3 to reduce overfitting and improve the model's overall performance. In the final stage, a single-neuron linear output layer, suitable for regression problems, was used to obtain the final prediction value:

$$\hat{y} = Wh + b \quad (10)$$

This structure ensures that the model has both strong representational power and produces statistically more stable predictions.

### 3.6. Implementation details

The proposed model was trained using an end-to-end Adam optimization algorithm. The learning rate was set to $1\times10^{-3}$; MSE was chosen as the loss function, and MAE was selected as the evaluation metric. This choice ensures both the effective minimization of error magnitudes and the statistically reliable assessment of the model's prediction accuracy.

The developed hybrid architecture consists of a combination of multi-scale convolutional layers, BiLSTM, and multi-head attention mechanisms. The model's input data is a single-channel time series with a length of 100. In the first stage, short, medium, and long-term patterns were extracted using three parallel Conv1D layers with kernel sizes of 3, 5, and 7. Batch Normalization was applied after each convolutional operation, and the outputs obtained from these three channels were combined to form a 384-dimensional feature map.

Dropout was applied to this combined representation to reduce the risk of overfitting. Subsequently, a 64-unit BiLSTM layer was used to learn sequential temporal dependencies more deeply and bidirectionally. To capture time-series contextual relationships more effectively, a 4-headed, 32-key-dimensional Multi-Head Attention layer was added; this layer was supported by residual connections and Layer Normalization to enhance the model's stability and learning performance.

In the final stage, the generated sequence was converted into a fixed-size vector using the GAP operation; then, a Dense layer with 128 neurons, ReLU activation, and L2 regularization was applied. After this layer, another dropout step was added. The final output was generated using a single-neuron Linear Dense layer, suitable for the regression task.

The model architecture and the number of parameters for each layer are provided in Table 1. As a result, the model has a total of 316,673 parameters. Of these parameters, 315,905 are trainable, while 768 are non-trainable. This architecture has been optimized with both computational efficiency and high prediction accuracy in mind. To ensure a realistic forecasting scenario and eliminate temporal information leakage, the proposed model was evaluated using a strict chronological train-test split strategy. The earliest 70% of observations (January 1986–May 2013) were used for model training, while the most recent 30% (June 2013–March 2025) were reserved exclusively for testing. Hyperparameter optimization, model fitting, and data preprocessing procedures were performed using only the training data. Furthermore, the parameters of the StandardScaler were estimated exclusively from the training set and subsequently

applied to the test set. This evaluation protocol ensures that no future information is available during model development and better reflects real-world forecasting conditions.

Table 1. Proposed model layer structure and parameter counts

| Layer (Type) | Output Shape | Number of Parameters |
|---|---|---|
| Input Layer (Input) | (100, 1) | 0 |
| Conv1D (kernel=3, 128) | (100, 128) | 512 |
| Conv1D (kernel=5, 128) | (100, 128) | 768 |
| Conv1D (kernel=7, 128) | (100, 128) | 1.024 |
| Batch Normalization ×3 | (100, 128) | 1.536 |
| Concatenate | (100, 384) | 0 |
| Dropout | (100, 384) | 0 |
| BiLSTM (64×2) | (100, 128) | 229.888 |
| Multi-Head Attention | (100, 128) | 66.048 |
| Residual + Add | (100, 128) | 0 |
| Layer Normalization | (100, 128) | 256 |
| Global Average Pooling | (128) | 0 |
| Dense (128, ReLU) | (128) | 16.512 |
| Dropout | (128) | 0 |
| Dense (1, Linear) | (1) | 129 |
| Total | - | 316.673 |

## 4. Experimental Results

In the experimental studies, we used the Python software platform, and a high-performance laptop equipped with an Intel Core i7-13700HX processor and RTX 4080 Mobile GPU running on Ubuntu 22.04.5 LTS operating system. This robust hardware configuration enabled us to carry out our applications and computations smoothly with high efficiency.

### 4.1. Dataset

This study has created a comprehensive dataset for predicting oil prices. The WTI crude oil spot price, which most accurately reflects price movements in global energy markets, was used as the dependent variable. Various macroeconomic and energy market indicators were selected as independent variables in order to include economic and sectoral factors affecting oil prices in the model.
The basic variables used in the data set are as follows:

- US Dollar Index (DXY): Reflects the impact of exchange rate fluctuations on oil prices.
- Consumer Price Index (CPI): Shows the relationship between inflation and energy consumption.
- INDPRO (Industrial Production Index): Represents the impact of economic activity on oil demand.
- US Crude Oil Production: Explains supply-side dynamics.
- XAU/USD (Gold Price): Affects volatility in energy markets based on risk perception.
- IGREA (Global Real Economic Activity): Represents global demand conditions.
- PPI Energy Sub-Items: Gasoline (WPU0571), jet fuel (WPU057203), diesel (WPU057303), natural gas (WPU0531), propane (WPU05320104), butane (WPU05320105), and electricity (WPU0543).

The data has been obtained from reliable sources such as EIA, FRED, and Investing.com and covers the period from January 1986 to March 2025. The resulting dataset spans approximately 39 years and includes multiple market regimes characterized by both stable and highly volatile periods. Notable events represented within the dataset include the 2008 Global Financial Crisis, the 2014 oil price collapse, the COVID-19 oil market shock, and the subsequent post-pandemic recovery period. The inclusion of these events enables a comprehensive evaluation of model robustness under different market conditions. The US economy's dominant position in global oil markets is the key factor supporting the data

sources' focus on this region. Sample rows related to the structure of the data set are presented in Table 2. This table provides an example view representing the format and data structure of the variables used in the model.

Table 2. Sample data from the dataset

| Date | WTI | DXY | CPI USA | INDPRO | Oil Production | Gold Price | IGREA | WPU057303 | WPU0543 | WPU0531 | WPU0571 | WPU057203 | WPU0532 0104 | WPU053 20105 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jan-1986 | 22.93 | 123.71 | 109.900 | 55.6453 | 283248 | 326.5 | -37.82508200 | 81.9 | 114.2 | 97.5 | 79.6 | 79.9 | 105.4 | 105.1 |
| Feb-1986 | 15.46 | 121.78 | 109.700 | 55.2634 | 256855 | 347.3 | -51.77545200 | 63.7 | 115 | 95.3 | 70.7 | 77.2 | 99 | 89.3 |
| Mar-1986 | 12.61 | 114.94 | 109.100 | 54.8728 | 279413 | 335.6 | -55.41517800 | 49.5 | 114.4 | 100.5 | 55.5 | 70.2 | 78.5 | 66.6 |
| Apr-1986 | 12.84 | 117.78 | 108.700 | 54.9350 | 265917 | 333.1 | -61.57598000 | 49.5 | 113.7 | 96.7 | 50.9 | 59.2 | 68.2 | 49.8 |
| May-1986 | 15.38 | 112.28 | 109.000 | 55.0399 | 273964 | 345.3 | -66.41025400 | 45.6 | 114.1 | 95.9 | 56.6 | 52.9 | 64.6 | 45.5 |

### 4.2. Performance metrics

In this study, four common performance metrics are used to evaluate the performance of the proposed model: MAE, RMSE, MSE and $R^2$. The mathematical expressions of these metrics are given below (Equations 11-14):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (11)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (12)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (14)$$

Here, $y_i$, $\hat{y}_i$, $\bar{y}$, and n represent the actual values, predicted values, average of actual values, and total number of data, respectively. These metrics were used to comprehensively measure the prediction accuracy and generalization ability of the model.

### 4.3. Results

The performance of the proposed CNN–BiLSTM–Attention-based model was evaluated using a strict chronological train-test split strategy. The earliest 70% of observations (January 1986–May 2013) were used for training, while the remaining 30% (June 2013–March 2025) were reserved exclusively for testing. This evaluation protocol prevents temporal information leakage and provides a more realistic assessment of forecasting performance under real-world conditions. The model was trained using the Adam optimization algorithm (learning rate = 0.001); MSE was chosen as the loss function, and MAE was chosen as the evaluation metric.

The proposed architecture consists of a multi-scale CNN module comprising three parallel Conv1D layers with kernel sizes of 3, 5, and 7 to capture multi-scale temporal patterns. The extracted features are subsequently fed into a 64-unit BiLSTM layer to model long-term temporal dependencies, followed by a four-head attention mechanism that selectively emphasizes the most informative time steps. Dropout layers are incorporated throughout the network to reduce the risk of overfitting. The final regression prediction was obtained through GAP and dense layers. Experimental performance results are presented in Table 3.

Table 3. The performance results (score) for proposed regression model

| MAE | MSE | RMSE | R-Squared |
|---|---|---|---|
| 0.1641 | 0.0398 | 0.1994 | 0.9224 |

The results presented in Table 3 demonstrate that the proposed model achieved strong forecasting performance under

the adopted chronological evaluation protocol. The obtained MAE, RMSE, and $R^2$ values indicate that the model successfully captures the nonlinear dynamics of crude oil prices while maintaining robust out-of-sample prediction capability. Since the test set consists exclusively of future observations that were not available during training, these results provide a realistic estimate of the model's forecasting performance under real-world market conditions.

In addition to the numerical results, a visual assessment of the model's prediction performance is presented in Figure 2. Figure 2(a) illustrates the temporal relationship between the actual and predicted crude oil prices over the test period. The close alignment between the two curves indicates that the proposed model successfully captures both long-term trends and short-term fluctuations in oil prices, even during periods of increased market volatility. Figure 2(b) presents the relationship between actual and predicted values in the form of a scatter plot. The concentration of observations around the perfect-fit line demonstrates a strong agreement between predicted and actual values, indicating that the model achieves reliable forecasting performance across different price levels.

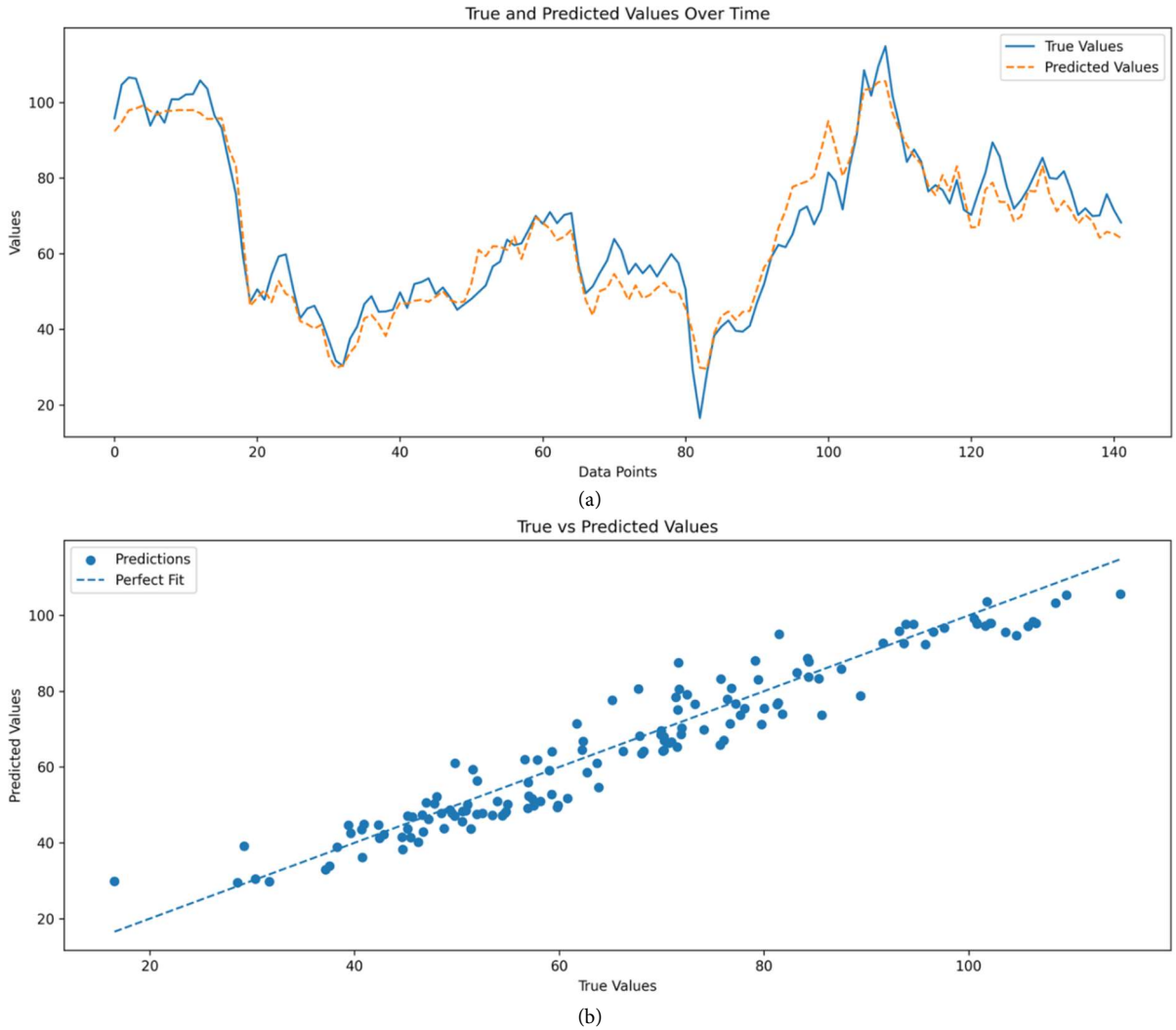



Figure 2. Prediction performance of the proposed model: (a) Actual and Predicted Oil Prices Over Time, (b) True vs Predicted Scatter Plot.

When Figure 2(a) and Figure 2(b) are evaluated together, it can be observed that the proposed model accurately follows the overall movement of crude oil prices while maintaining a strong correspondence between predicted and actual values. These findings confirm that the proposed CNN-BiLSTM-Attention-based architecture provides robust and reliable forecasting performance under previously unseen market conditions.

### 4.4 Robustness analysis across market regimes

To further assess the robustness and generalization capability of the proposed CNN-BiLSTM-Attention architecture, an additional analysis was conducted across different market regimes. Particular attention was given to the COVID-19 pandemic period because it represents one of the most disruptive events in modern energy markets, characterized by unprecedented demand shocks, supply chain disruptions, extreme price volatility, and rapid recovery dynamics. Since the adopted chronological train-test split places the COVID-19 and post-pandemic periods entirely within the test set, these observations provide an opportunity to evaluate the model under previously unseen and highly challenging market conditions.

The forecasting performance of the proposed model and several benchmark machine learning methods was evaluated separately for the Pre-COVID, COVID Shock, COVID Recovery, and Post-COVID periods. The coefficient of determination ($R^2$) values obtained for each period are presented in Table 4.

Table 4. Comparison of forecasting performance across different market regimes ($R^2$)

| Method | Full Test | Pre-COVID | COVID Shock | COVID+Recovery | Post-COVID |
|---|---|---|---|---|---|
| Proposed Model | 0.9224 | 0.4501 | 0.6538 | 0.7752 | 0.7953 |
| ExtraTrees | 0.8443 | -0.2289 | 0.3095 | 0.7262 | 0.5938 |
| Random Forest | 0.7543 | -0.3664 | 0.2319 | 0.6108 | 0.0366 |
| Decision Tree | 0.7498 | 0.3316 | 0.1005 | 0.5631 | 0.0486 |
| Gradient Boosting | 0.6951 | -0.6107 | -0.0435 | 0.3926 | -0.1425 |
| Lasso | 0.5409 | 0.6339 | 0.6059 | 0.4431 | -2.9829 |
| Ridge | 0.5046 | 0.0779 | 0.4672 | 0.4028 | -3.3097 |
| AdaBoost | 0.6342 | -0.5906 | 0.0930 | 0.2605 | -0.7992 |
| MLP | 0.5231 | -1.4753 | -0.0426 | 0.1019 | -2.1212 |
| Linear Regression | 0.1779 | 0.0703 | 0.1680 | -0.1336 | -6.3854 |
| SVR | -0.3310 | -3.4537 | -3.8628 | -0.3267 | -6.1481 |
| KNN | -1.0149 | -20.6822 | -24.1610 | -4.5567 | -0.6316 |
| Gaussian Process | -1.3102 | -6.0137 | -0.0012 | -0.6721 | -13.5496 |

As shown in Table 4, the proposed model consistently achieved the highest $R^2$ values across all evaluated market regimes. On the complete test set, the proposed architecture obtained an $R^2$ value of 0.9224, outperforming the strongest conventional machine learning baseline, Extra Trees ($R^2$ = 0.8443). Similar superiority was observed during the COVID Shock period, where the proposed model achieved an $R^2$ value of 0.6538, indicating a stronger ability to capture abrupt market movements and structural changes.

The advantage of the proposed architecture becomes even more evident during the COVID Recovery and Post-COVID periods. While most machine learning methods experienced substantial performance degradation or even negative $R^2$ values, the proposed model maintained stable forecasting capability, achieving $R^2$ values of 0.7752 and 0.7953, respectively. These results indicate that the combination of multi-scale convolutional feature extraction, bidirectional temporal modeling, and attention mechanisms provides superior adaptability to changing market conditions.

In addition to the $R^2$-based comparison, the forecasting behavior of the proposed model throughout the entire test period is illustrated in Figure 3. The highlighted region corresponds to the COVID-19 shock and recovery period.

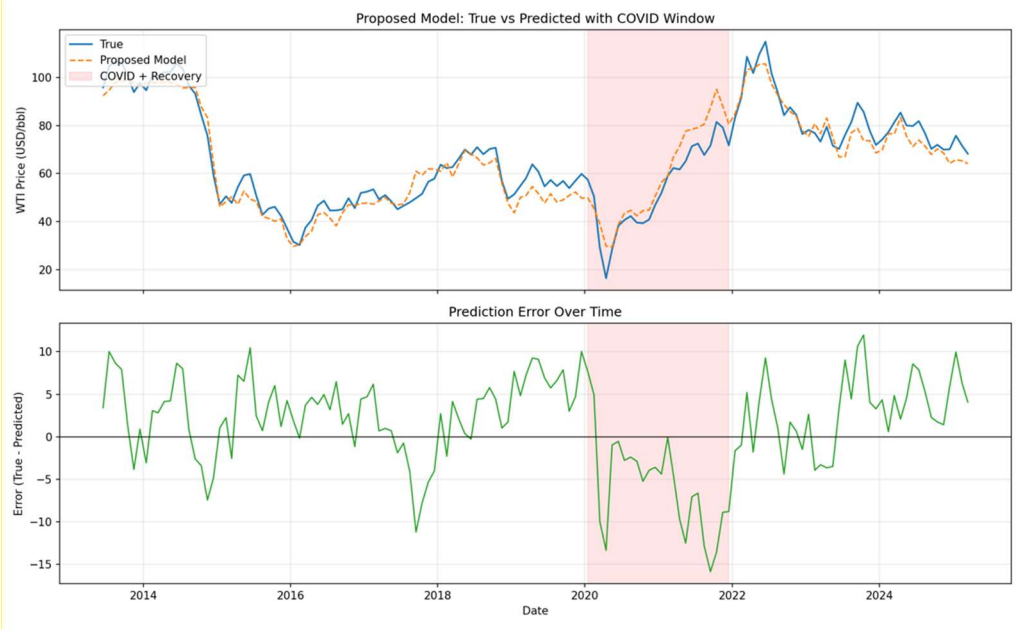



Figure 3. Actual and predicted crude oil prices during different market regimes, including the COVID-19 shock and recovery period.

As illustrated in Figure 3, the proposed model successfully follows the overall trajectory of crude oil prices across different market regimes. Although prediction deviations increase during periods of extreme volatility, particularly during the COVID-19 recovery phase, the model continues to capture the major turning points and long-term trends. The close alignment between actual and predicted values throughout the test period further supports the robustness of the proposed forecasting framework.

To provide a complementary error-based assessment, the RMSE and MAE values obtained by the proposed model for each market regime are reported in Table 5.

Table 5. Error metrics of the proposed model across different market regimes (USD/bbl)

| Period | RMSE | MAE |
|---|---|---|
| Full Test | 5.77 | 4.75 |
| Pre-COVID | 5.44 | 4.58 |
| COVID Shock | 6.05 | 4.85 |
| COVID + Recovery | 8.10 | 6.78 |
| Post-COVID | 5.23 | 4.36 |

The results presented in Table 5 confirm that forecasting difficulty increases during periods characterized by extraordinary market behavior. The highest error values were observed during the COVID Recovery period (RMSE = 8.10, MAE = 6.78), reflecting the exceptional volatility and rapid market adjustments experienced during this phase. Nevertheless, the model maintained acceptable forecasting accuracy and successfully tracked overall market movements. Lower error values were observed during the Pre-COVID and Post-COVID periods, with the best performance achieved during the Post-COVID period (RMSE = 5.23, MAE = 4.36).

Overall, the results demonstrate that the proposed CNN-BiLSTM-Attention architecture remains effective across diverse market conditions and retains strong forecasting capability even during periods of severe economic uncertainty and structural disruptions. These findings provide additional evidence regarding the practical applicability and robustness of the proposed framework for long-term crude oil price forecasting.

### 4.5. Ablation study

This ablation study was conducted to systematically evaluate the contribution of the proposed model's components to its prediction performance. In this context, the Multi-Scale CNN, BiLSTM, Attention, and Residual blocks were disabled one by one, and the performance of each variant was compared using the RMSE, MAE, and $R^2$ metrics. The results obtained under the adopted chronological evaluation protocol are presented in Table 6.

Table 6. Ablation study results obtained under the chronological evaluation protocol

| Variant | Multi-Scale CNN | BiLSTM | Attention | Residual | RMSE | MAE | $R^2$ |
|---|---|---|---|---|---|---|---|
| FULL | ✓ | ✓ | ✓ | ✓ | 0.1994 | 0.1641 | 0.9224 |
| – Residual | ✓ | ✓ | ✓ | × | 0.2131 | 0.1607 | 0.9113 |
| – Attention | ✓ | ✓ | × | × | 0.2487 | 0.1953 | 0.8793 |
| –Multi-Scale CNN | × | ✓ | ✓ | ✓ | 0.2521 | 0.1921 | 0.8760 |
| – BiLSTM | ✓ | × | ✓ | ✓ | 0.3722 | 0.3131 | 0.7296 |

Table 6 shows that the full model, with all components active, achieved the best overall performance (RMSE = 0.1994, MAE = 0.1641, $R^2$ = 0.9224). These results indicate that the combined use of multi-scale feature extraction, bidirectional temporal dependency modeling, attention mechanisms, and residual connections provides the most effective architecture for crude oil price forecasting. Removing the Residual component resulted in a moderate decrease in performance (RMSE = 0.2131, $R^2$ = 0.9113), indicating that residual connections contribute to learning stability and facilitate information flow throughout the network. When the Attention mechanism was removed, forecasting performance deteriorated further (RMSE = 0.2487, $R^2$ = 0.8793). This finding confirms that the attention mechanism helps the model focus on the most informative temporal patterns and improves prediction accuracy. A similar performance decline was observed when the Multi-Scale CNN block was excluded (RMSE = 0.2521, $R^2$ = 0.8760). Since this component extracts information from different temporal scales through multiple convolution kernels, its removal reduces the model's ability to capture diverse market dynamics. The largest performance degradation occurred when the BiLSTM layer was removed (RMSE = 0.3722, MAE = 0.3131, $R^2$ = 0.7296). This substantial decrease demonstrates the critical role of bidirectional temporal modeling in learning long-term dependencies and complex sequential relationships within crude oil price series.

Overall, the ablation results indicate that all components contribute positively to forecasting performance. Among them, the BiLSTM layer appears to be the most influential component, while the Multi-Scale CNN and Attention mechanisms provide substantial complementary benefits. The best performance is achieved when all components operate together within the proposed hybrid architecture.

### 4.6. Dependent parameter analysis

The SHAP method was used to explain the model's prediction mechanism and evaluate the relative effects of the variables. SHAP transparently reveals the decision structure of the model by measuring the marginal contribution of each feature to the prediction in complex deep learning models. This increases the model's explainability and allows for a more systematic evaluation of the functions of economic variables in the prediction process.

In this context, Table 7 presents the average contributions of the energy price indices, macroeconomic indicators, and production-based parameters in the dataset to the model output. The results indicate that the PPI components associated with fuel and energy products make the greatest contribution to the model, highlighting their importance in the prediction process.

According to Table 7, CPI is the most influential variable in the proposed model, contributing 19.53% of the total importance. This finding suggests that inflation-related economic conditions play a major role in long-term crude oil price forecasting. Among the energy-related indicators, jet fuel (WPU057203), diesel fuel (WPU057303), and gasoline price indices (WPU0571) are the most influential variables, collectively accounting for more than 40% of the overall

importance. This result confirms that refined petroleum product prices provide strong signals regarding crude oil market dynamics. Gold prices (10.38%) and the Industrial Electric Power Index (WPU0543) (8.80%) also contribute substantially to the model, indicating that broader economic activity and energy consumption patterns contain valuable predictive information. Industrial production (INDPRO) remains an important explanatory factor with a contribution of 6.45%, reflecting the close relationship between economic activity and energy demand. The remaining variables, including propane prices, butane prices, natural gas prices, oil production, IGREA, and DXY, exhibit relatively smaller contributions. Although their individual impacts are limited, they provide complementary information that improves the model's ability to capture different aspects of the crude oil market.

Table 7. SHAP analysis results based on the proposed model

| Index | Description | Degree of Importance |
|---|---|---|
| CPI | Consumer Price Index (USA) | 19.53% |
| WPU057203 | Jet Fuel Price Index | 14.33% |
| WPU057303 | Diesel Fuel Price Index | 14.13% |
| WPU0571 | Gasoline Price Index | 13.29% |
| XAU/USD | Gold Spot Price/ons | 10.38% |
| WPU0543 | Industrial Electric Power Index | 8.80% |
| INDPRO | Industrial Production Index (USA) | 6.45% |
| WPU05320104 | Propane Price Index | 5.21% |
| WPU05320105 | Butane Price Index | 3.98% |
| WPU0531 | Natural Gas Price Index | 1.46% |
| Oil Production | US Crude Oil Production | 0.95% |
| IGREA | Global Real Economic Activity Index | 0.92% |
| DXY | US Dollar Index | 0.59% |

The contribution ratios presented in Table 7 clearly show which variables are more decisive in the model's predictions and increase the model's transparency. In addition, contribution distributions are presented in Figure 4 so that the effects of the features can be examined visually.

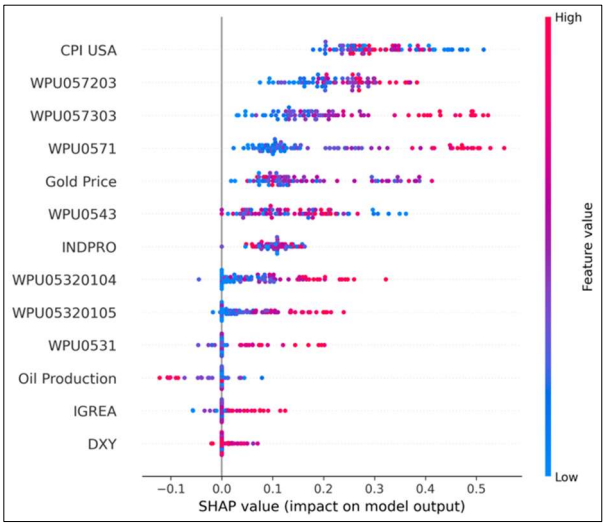



Figure 4. SHAP Summary Graph: Impact of Features on Model Predictions

Figure 4 visually displays the distribution of SHAP values for the variables and allows for the interpretation of both the magnitude and direction of the effect. The color scale indicates the change in variable values from low (blue) to high (red).

Consistent with the results presented in Table 5, Figure 4 confirms that CPI USA, jet fuel prices, diesel fuel prices, and gasoline prices exert the strongest influence on the model output. The concentration of high feature values (red points) in regions associated with positive SHAP values indicates that increases in these variables generally contribute to higher predicted crude oil prices.

Positive and significant effects are also observed for Gold Price, WPU0543, and INDPRO. These variables contribute to the model by reflecting broader macroeconomic activity, industrial demand, and energy consumption patterns. Their SHAP distributions indicate that the model incorporates not only direct energy market indicators but also wider economic signals when generating forecasts. On the other hand, the SHAP distributions of variables such as IGREA, Oil Production, Natural Gas Prices, and DXY are concentrated within a relatively narrow range. This indicates that their direct influence on model predictions is limited compared to the dominant variables. Nevertheless, these indicators provide complementary information that helps the model better represent broader market conditions.

This SHAP analysis reveals that the proposed model is primarily driven by a combination of macroeconomic conditions and refined petroleum product prices. The dominant influence of CPI, jet fuel, diesel fuel, and gasoline price indices demonstrates that both inflationary dynamics and energy market conditions play critical roles in crude oil price forecasting. These findings indicate that the proposed CNN–BiLSTM–Attention architecture captures not only short-term market movements but also broader economic relationships, providing predictions that are both statistically accurate and economically meaningful.

These results confirm that the proposed CNN-BiLSTM-Attention-based structure possesses a prediction mechanism that is not only statistically but also economically consistent, and scientifically clarifies which variables the model prioritizes in decision-making processes.

## 5. Discussion

In this section, the findings of the study are discussed in comparison with the existing literature and its practical contributions and limitations are discussed. After detailing the original contributions of the study to the literature, suggestions for future research are provided.

### 5.1. Comparison with machine learning methods

The comparison of machine learning methods with the deep learning-based proposed model highlights the advantages of the method over traditional algorithms. In this context, linear methods (Linear Regression, Ridge, Lasso), ensemble approaches (RF, Gradient Boosting, Extra Trees, AdaBoost) and ANN-based MLP regressors were evaluated. The results of all methods are presented comparatively in Table 8 in terms of RMSE, MAE, and $R^2$ metrics.

Table 8. Comparison of proposed model with other machine learning methods

| Method | RMSE | MAE | $R^2$ |
|---|---|---|---|
| Linear Regression | 0.6490 | 0.4504 | 0.1779 |
| Ridge | 0.5038 | 0.3660 | 0.5046 |
| Lasso | 0.4850 | 0.3351 | 0.5409 |
| Random Forest | 0.3548 | 0.3186 | 0.7543 |
| Gradient Boosting | 0.3952 | 0.3584 | 0.6951 |
| Decision Tree | 0.3580 | 0.2915 | 0.7498 |
| ExtraTrees | 0.2825 | 0.2445 | 0.8443 |
| AdaBoost | 0.4329 | 0.3728 | 0.6342 |
| MLP | 0.4943 | 0.4181 | 0.5231 |
| Proposed Model | 0.1994 | 0.1641 | 0.9224 |

When examining the results presented in Table 8, it can be observed that the proposed CNN-BiLSTM-Attention model achieves the best overall forecasting performance among all evaluated methods. The proposed model obtained the lowest error values (RMSE = 0.1994, MAE = 0.1641) and the highest coefficient of determination ($R^2$ = 0.9224), indicating a strong ability to capture the nonlinear dynamics of crude oil prices. Among the conventional machine learning methods, Extra Trees achieved the closest performance ($R^2$ = 0.8443), followed by RF ($R^2$ = 0.7543) and Decision Tree ($R^2$ = 0.7498). However, all benchmark methods remained noticeably below the performance level of the proposed deep learning architecture.

Linear models such as Linear Regression, Ridge, and Lasso exhibited considerably lower forecasting performance, suggesting that the relationship between crude oil prices and the selected explanatory variables cannot be adequately represented through linear assumptions alone. Ensemble-based approaches, including RF, Extra Trees, Gradient Boosting, and AdaBoost, demonstrated improved performance by capturing nonlinear relationships, yet they were unable to model temporal dependencies explicitly. The superior performance of the proposed model indicates that the integration of multi-scale convolutional feature extraction, bidirectional temporal learning, and attention mechanisms provides a significant advantage for forecasting complex and highly volatile crude oil price series.

Overall, the results demonstrate that although conventional machine learning methods provide reasonable forecasting capability, the proposed CNN-BiLSTM-Attention architecture consistently outperforms all benchmark models across all evaluation metrics. These findings suggest that deep learning architectures specifically designed to capture both local patterns and long-term temporal dependencies are more suitable for modeling the complex behavior of crude oil prices under a realistic chronological forecasting setting.

### 5.2. Comparison with existing literature

When examining the literature on oil price forecasting, it is observed that the majority of studies rely on traditional econometric models (ARIMA, GARCH, VAR, etc.), but these models are limited in adequately representing the nonlinear structures and sudden market shocks in oil prices. In recent years, machine learning and deep learning-based methods have gained prominence in the literature because they can model non-linear dynamics more effectively. As part of this transformation, the proposed CNN-BiLSTM-Attention-based model offers an architecture that comprehensively addresses the multi-scale and sequential dependencies of complex time series. In this context, the prominent studies and performance results in the literature are summarized in Table 9.

It should be noted that direct comparisons with previous studies should be interpreted with caution because datasets, forecasting horizons, evaluation protocols, and preprocessing procedures differ considerably across the literature. Unlike many previous studies that rely on random train-test partitioning or conventional cross-validation approaches, the present study employs a strict chronological train-test split strategy to prevent temporal information leakage. Therefore, the reported performance values represent a more realistic out-of-sample forecasting scenario and provide a more rigorous assessment of forecasting capability under previously unseen market conditions.

The comparisons in Table 9 demonstrate the variety of methods used in the literature and their performance levels. For example, although Zeng et al. (2025) GARCH–BiLSTM–KAN structure can model mixed volatility dynamics, the 2.49 RMSE and 0.981 $R^2$ values indicate that errors cannot be reduced below a certain level despite the complex hybrid structure [11]. Although Bi and colleagues' (2024) combination of VMD-based component decomposition with the BiTCN–BiGRU–Attention structure offers high accuracy, the complexity of the model components and their dependence on preprocessing increase the application cost [10]. SVR-based models examined by Jha et al. (2024) achieved lower error levels than classical regression methods, highlighting the importance of nonlinear relationships; however, these methods fail to capture sequential dependencies and multi-scale patterns, often being limited to single-step prediction performance [4]. Furthermore, many of the existing studies have been limited to relatively short forecasting horizons or specific market periods. In contrast, the proposed framework is evaluated using a dataset spanning approximately 39 years (1986–2025), encompassing major market disruptions, including the Global Financial Crisis, the 2014 oil price collapse, and the COVID-19 pandemic.

Table 9. Comparison of methods and performance metrics for crude oil price forecasting

| Study (Year) | Data / Target | Inputs | Model(s) | RMSE | MAE | $R^2$ |
|---|---|---|---|---|---|---|
| Zeng et al. (GBK-Net, 2025) [11] | WTI daily prices (1986–2025) | Returns + GARCH volatility | GARCH–BiLSTM–KAN | 2.49 | – | 0.981 |
| Bi et al. (2024) [10] | International crude oil prices | VMD decomposition components | VMD + BiTCN + BiGRU + Attention | 1.4417 | 0.7973 | 0.9953 |
| Jha et al. (2024) [4] | WTI & Brent spot prices | 8 macroeconomic variables (gold, natural gas, FX, supply–demand indicators, etc.) | SVR (RBF), Ridge, Lasso, Multiple Regression | 0.0766 | – | 0.9951 |
| Li et al. (2019) [2] | Crude oil prices (Brent & WTI) | Financial indicators + news texts (LDA + sentiment features) | SVR (RBF), RF, Linear Regression | 0.0040 | 0.0030 | – |
| Jahandoost et al. (2023) [8] | WTI prices | Daily, weekly, monthly multi-scale features | CNN–GRU–Attention (proposed), CuDNNLSTM-1dCNN, LSTM NN | – | 1.5495 | – |
| An et al. (2019) [5] | Brent & WTI prices | GARCH + Wavelet components | CEEMDAN–LSTM hybrid | 1.912 | 1.631 | – |
| Shambulingappa et al. (2020) [6] | Crude oil prices | Time-series components | CNN–LSTM / CNN–GRU + skip connections | 0.28 | 0.19 | – |
| Obite et al. (2021) [7] | Nigerian crude oil production | Production time series | ARIMA, ANN, RF | 0.010 | – | 0.994 |
| Foroutan & Lahmiri (2024) [9] | WTI, Brent, Gold, Silver (2000–2022) | Price series (+ Time2Vector embedding) | LSTM / BiLSTM / GRU / BiGRU / CNN / TCN / T2V-BiLSTM / T2V-BiGRU / RF / LightGBM / SVR / KNN | 1.894 | 1.444 | – |
| Proposed study (2026) | WTI spot prices (1986–2025) | USD index, industrial production, CPI/PPI, global economic activity, jet fuel, gasoline, propane, diesel indicators | Multi-scale CNN, BiLSTM, Multi-Head Attention | 0.1994 | 0.1641 | 0.9224 |

Similar diversity is also observed in deep learning-based studies. Although hybrid models such as CNN–GRU, CNN–LSTM, or CEEMDAN–LSTM have produced strong results in capturing non-linear structures, limitations exist in modeling sequential dependencies together with multi-scale convolutional feature extraction. For example, Shambulingappa (2020) achieved an RMSE of approximately 0.28 with the CNN–GRU model, but the power of long-term dependencies remained limited, especially during volatile periods [6]. Similarly, An et al.'s (2019) model, which combines CEEMDAN decomposition with LSTM, offers a component-based approach but does not allow for the direct learning of multiscale features within the architecture. In terms of addressing temporal depth and multiscale dynamics together, the structure proposed in this study differs from these models in the literature [5].

Although many studies in the literature employ deep learning and machine learning methods for crude oil price forecasting, the transparency of model decision-making processes and detailed analyses of variable contributions have often been overlooked. Furthermore, a considerable number of existing approaches focus on a single modeling component (e.g. CNN, LSTM, GRU, or decomposition techniques) and do not simultaneously address multi-scale feature extraction and long-term temporal dependencies.

This study addresses these limitations through a hybrid architecture that combines multi-scale convolutional layers, BiLSTM, and attention mechanisms, together with SHAP-based interpretability analysis and robustness evaluation across different market regimes. The proposed framework offers a more balanced and holistic approach by simultaneously addressing forecasting accuracy, model interpretability, and generalization under challenging market conditions, including the COVID-19 shock and recovery periods. Consequently, the study contributes not only to predictive performance but also to the practical reliability and transparency of crude oil price forecasting models.

### 5.3. Practical contributions and limitations of the study

The practical contributions of this study enhance the accuracy, interpretability, and applicability of crude oil price forecasting, providing valuable insights for the energy sector and policymakers:

- Development of the CNN-BiLSTM-Attention Model: This study proposes a hybrid deep learning

architecture that combines convolutional neural networks (CNN), bidirectional long-short term memory (BiLSTM), and multi-head attention mechanisms. This model captures local features, sequential dependencies, and long-term contextual relationships simultaneously, providing higher prediction accuracy and generalization capacity compared to traditional econometric methods and classical machine learning approaches.

- Capturing Nonlinear Dynamics: Crude oil prices are influenced by complex and nonlinear relationships such as supply-demand imbalances, macroeconomic fluctuations, and market uncertainties. CNN-based feature extraction, combined with BiLSTM's sequential learning capability and the contextual power of its attention mechanism, has enabled these nonlinear structures to be modeled more effectively. As a result, the model produces more reliable and realistic results in market predictions.
- Increasing Model Explainability with SHAP Analysis: One of the study's significant contributions is making the model's decision-making process transparent using SHAP analysis. This analysis reveals the relative effects of variables such as jet fuel price, gasoline price, US Dollar Index (DXY), and industrial production index on the predictions. This increases the interpretability of the model outputs, providing more reliable and actionable information for energy sector managers and policymakers.
- Applicability in the Energy Sector: The proposed model can be adapted not only to crude oil price forecasting but also to different energy forecasting problems such as natural gas demand, electricity consumption, or renewable energy integration. This flexibility provides significant contributions to energy management, supply chain planning, and the support of sustainable energy policies.

These contributions demonstrate their importance in terms of increasing the accuracy of the study's predictions, enhancing transparency, and expanding its applicability in different energy contexts.

However, the study has some limitations. First, the dataset used is largely based on indicators specific to the US market. This may limit the model's ability to generalize to other geographic regions or global market conditions. Second, although the CNN-BiLSTM-Attention architecture demonstrated strong forecasting performance, no direct comparison was conducted with more recent architectures such as Transformer-based models or Graph Neural Networks (GNNs), which have recently shown promising results in complex forecasting tasks. Therefore, the relative advantages and limitations of the proposed architecture compared with these emerging approaches remain an open research question. Finally, external factors such as geopolitical risks, extreme weather conditions, or sudden macroeconomic shocks have not been included in the model, even though these factors can significantly influence short-term price fluctuations.

Future studies can address these limitations by expanding the dataset with multi-country and global indicators, incorporating geopolitical and environmental variables, and evaluating alternative deep learning architectures such as Transformers and GNNs. In addition, robustness analyses across different market regimes and forecasting horizons may further improve the generalizability and practical applicability of crude oil price forecasting models.

## 6. Conclusion

The objective of this study was to develop an accurate, interpretable, and robust deep learning framework for crude oil price forecasting. To achieve this goal, a hybrid architecture combining multi-scale CNNs, BiLSTM, and multi-head attention mechanisms was proposed. The model was evaluated using a strict chronological train-test split strategy on a dataset spanning approximately 39 years (1986–2025), thereby preventing temporal information leakage and providing a realistic out-of-sample forecasting scenario.

The experimental results demonstrated that the proposed architecture achieved strong forecasting performance and consistently outperformed conventional machine learning approaches. The ablation study further confirmed that the combined use of multi-scale feature extraction, bidirectional temporal modeling, and attention mechanisms plays a critical role in improving forecasting accuracy. In addition, SHAP-based interpretability analysis revealed that CPI, jet fuel prices, diesel fuel prices, and gasoline prices are among the most influential variables affecting crude oil price predictions.

The robustness analyses conducted across different market regimes, including the COVID-19 shock period, recovery phase, and post-COVID period, demonstrated that the proposed framework maintains stable forecasting capability even under highly volatile and previously unseen market conditions. These findings indicate that the model is capable of capturing both short-term market fluctuations and long-term economic dynamics. Despite these promising results, several limitations remain. The dataset is primarily composed of indicators related to the U.S. economy, which may limit the direct applicability of the model to other geographical regions. Furthermore, no direct comparison was performed with emerging architectures such as Transformer-based models or GNNs. Future research may address these limitations by incorporating global indicators, geopolitical and environmental variables, and alternative deep learning architectures.

Overall, the proposed CNN-BiLSTM-Attention framework provides a reliable, interpretable, and practically applicable solution for crude oil price forecasting. The findings contribute to both the academic literature and real-world decision-support processes in energy markets by offering a forecasting approach that balances predictive performance, interpretability, and robustness.

## Conflict of Interest Statement

The authors declare that there is no conflict of interest

## Authorship Contribution Statement

**E.S.** and **O.P:** Conceptualization; **E.S.** and **M.T:** Methodology; **N.P:** Software, **E.S:** Validation; **E.F.** and **M.T: F**ormal Analysis; **N.P.** and **E.F:** Data Curation; **E.F., O.P.** and **M.T:** Writing - Original Draftwriting - review and editing**, N.P.** and **E.F:** visualization.

## References

[1] H. Guliyev and E. Mustafayev, "Predicting the changes in the WTI crude oil price dynamics using machine learning models," *Resources Policy*, vol. 77, 102664, March 2022. doi: 10.1016/j.resourpol.2022.102664

[2] X. Li, W. Shang and S. Wang, "Text-based crude oil price forecasting: A deep learning approach," *International Journal of Forecasting,* vol. 35, no. 4, pp. 1548-1560, September 2019. doi: 10.1016/j.ijforecast.2018.07.006

[3] H. He, M. Sun, X. Li and I. A. Mensah, "A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features," *Energy*, vol. 244, 122706, February 2022. doi: 10.1016/j.energy.2021.122706

[4] N. Jha, H. K. Tanneru, S. Palla and I. H. Mafat, "Multivariate analysis and forecasting of the crude oil prices: Part I–Classical machine learning approaches," *Energy*, vol. 296, 131185, April 2024. doi: 10.1016/j.energy.2024.131185

[5] J. An, A. Mikhaylov and N. Moiseev, "Oil price predictors: Machine learning approach," *International Journal of Energy Economics and Policy*, vol. 9, no. 5, pp. 1-6, June 2019. doi: 10.32479/ijeep.7597

[6] S. H. S, "Crude oil price forecasting using machine learning," *International Journal of Advanced Scientific Innovation*, vol. 1, no. 1, pp. 1-11, December 2020. doi: 10.5281/zenodo.4641697

[7] C. P. Obite, A. Chukwu, D. C. Bartholomew, U. I. Nwosu and G. E. Esiaba, "Classical and machine learning modeling of crude oil production in Nigeria: Identification of an eminent model for application," *Energy Reports*, vol. 7, 3497-3505, November 2021. doi: 10.1016/j.egyr.2021.06.005

[8] A. Jahandoost, M. Houshmand and S. A. Hosseini, "Prediction of west texas intermediate crude-oil price using hybrid attention-based deep neural networks: a comparative study," in Proceedings of the *13th International Conference on Computer and Knowledge Engineering (ICCKE)*, 01-02 November 2023, Mashhad, Iran, Islamic Republic of [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/10326291?casa_token=pgOS94waRCUAAAAA:pv-FU0oRnJFzzZxotL0Dv01ca5ZKMvira7E-VYSnkeX_qHBG-S5GMenzF2wmyMdB86nyY-NTOw. [Accessed: 29 January 2026]

[9] P. Foroutan and S. Lahmiri, "Deep learning systems for forecasting the prices of crude oil and precious metals," *Financial Innovation*, vol. 10, no.1, 111, July 2024. doi: 10.1186/s40854-024-00637-z

[10] M. Bi, Z. Liu and X. Yang, "A Hybrid Prediction Model for International Crude Oil Price Based on Variational Mode Decomposition with BiTCN-BiGRU-Attention Deep Learning Techniques," October 2024. Preprint available at *Research Square*. doi: 10.21203/rs.3.rs-5284460/v1

[11] J. Zeng,W. J. Xie, Z. Y. Wei, Z. L. Wei, H. Y. An, X. W. Jiang, B. Guo and M. Wang, "GBK-Net: A Hybrid GARCH--BiLSTM--KAN Model for Crude Oil Price Forecasting," September 2025. Preprint available at *Preprints*. doi: 10.20944/preprints202509.0646.v1

[12] M. Türkoğlu and D. Hanbay, "Plant recognition system based on extreme learning machine by using shearlet transform and new geometric features," *Journal of the Faculty of Engineering and Architecture of Gazi University*, vol. 34, no. 4, pp. 2097-2112, June 2019. doi: 10.17341/gazimmfd.423674

[13] O. Polat, M. Türkoğlu, H. Polat, S. Oyucu, H. Üzen, F. Yardımcı and A. Aksöz, "Multi-stage learning framework using convolutional neural network and decision tree-based classification for detection of DDoS pandemic attacks in SDN-based SCADA systems," *Sensors*, vol. 24, no. 3, 1040, February 2024. doi: 10.3390/s24031040

[14] A. I. Simsek, E. Koç, B. D. Tasdemir, A. Aksöz, M. Turkoglu, M. and A. Sengur, "Deep Learning Forecasting Model for Market Demand of Electric Vehicles," *Applied Sciences*, vol. 14, no. 23, 10974, November 2024. doi: 10.3390/app142310974

[15] B. Zhao, H. Lu, S. Chen, J. Liu and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162-169, February 2017. doi: 10.21629/JSEE.2017.01.18

[16] E. Soylu, S. Gül, K. A. Koca, M. Türkoğlu, M. Terzi and A. Şengür, "Speech signal-based accurate neurological disorders detection using convolutional neural network and recurrent neural network based deep network," *Engineering Applications of Artificial Intelligence*, vol. 149, 110558, June 2025. doi: 10.1016/j.engappai.2025.110558

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, November 1997. doi: 10.1162/neco.1997.9.8.1735

[18] M. Turkoglu, "Defective egg detection based on deep features and Bidirectional Long-Short-Term-Memory," *Computers and Electronics in Agriculture*, vol. 185, 106152, June 2021. doi: 10.1016/j.compag.2021.106152

[19] I. D. Mienye, T. G. Swart and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," *Information*, vol. 15, no. 9, 517, August 2024. doi: 10.3390/info15090517

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in Neural İnformation Processing Systems*, vol. 30, December 2017. ISBN: 9781510860964

[21] H. Üzen, M. Turkoglu, D. Ozturk and D. Hanbay, "A novel hybrid attention gate based on vision transformer for the detection of surface defects," *Signal, Image and Video Processing*, vol. 18, no. 10, pp. 6835-6851, June 2024. doi: 10.1007/s11760-024-03355-2

[22] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016, Las Vegas, NV, USA [Online]. Available: IEEE, https://www.computer.org/csdl/proceedings/cvpr/2016/12OmNqH9hnp. [Accessed: 12 February 2026]

[23] Ö. Dürrani, S. İşgüzar, A. İmak, T. Ateşşahin, Z. Cömert, S. Z. Dürrani and M. Türkoğlu, "MobileNet-and attention-based MLP-Mixer architecture for geographical-region recognition of a marine fish (Perciformes: Carangidae) using otolith images," *Expert Systems with Applications*, vol. 291, 128585, October 2025. doi: 10.1016/j.eswa.2025.128585