Pak. J. Statist.
2024 Vol. 40(4), 407-428

FORECASTING OF CRUDE OIL PRICES USING HYBRID TIME SERIES AND MACHINE LEARNING MODELS

Badr Alnssyan[1,§], Dost Muhammad Khan[2,§]
Muhammad Ali[2] and **Ahmed Z. Afify[3]**

[1] Department of Management Information Systems
College of Business and Economics, Qassim University
Buraydah 51452, Saudi Arabia. Email: b.alnssyan@qu.edu.sa

[2] Department of Statistics, Abdul Wali Khan University Mardan
Mardan, KP, Pakistan. Email: muhammad.ali@awkum.edu.pk

[3] Department of Statistics, Mathematics and Insurance
Benha University, Benha 13511, Egypt.
Email: ahmed.afify@fcom.bu.edu.eg

[§] Corresponding authors

ABSTRACT

The global economy is strongly influenced by the production of crude oil, a major nonrenewable energy source. Businesses and economies all over the world are challenged by a greater degree of unpredictability due to the volatility and dynamics of crude oil prices. Several decomposition techniques, including empirical mode decomposition (EMD) and its different variants, are seamlessly incorporated. These decomposition techniques are also integrated with various machine-learning algorithms, which include the support vector machine (SVM), random forest, decision tree and artificial neural network (ANN), to build the hybrid model for crude oil prediction with intrinsic mode functions (IMF's) and residue generated from the actual West Texas Intermediate (WTI). Since the proposed hybrid model is based on the data-decomposition and supervised machine-learning algorithm, therefore the IMFs and residue component extracted from the daily closing prices of WTI are given as input features to these supervised learning techniques. Three important statistical metrics including the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) are utilized to check the prediction performance of the proposed model. The results such as the RMSE, MAE, and MAPE values of 1.446, 1.259, and 2.194 confirm that the complementary ensemble empirical mode decomposition with adaptive noise (CEEMDAN) integrated with SVM technique as a dependable and effective crude oil price forecast tool and demonstrate its improved precision. The results ensure profitability in an unpredictable economy and fosters commodity stability in prices, both of which assist firms decrease their risks.

KEYWORDS

ANN, SVM, Random Forest, Decision Tree, Empirical Mode Decomposition, Machine Learning Models, Neural Networks, WTI, Economic Independence, Economic and Social Policies, Market Access, Soft Skills, SDGs.

1. INTRODUCTION

One of the important commodity that shapes the globe's structure is petroleum, or crude oil. The production of goods and commodities as well as the status of the global economy generally can be impacted by fluctuations in the market value of crude oil (Hu, 2021). It is frequently challenging to foresee crude oil prices having the lowest possible prediction error. Several industries eagerly await the falling of crude oil prices, which is particularly true for the countries that import and export this essential good (Gupta and Pandey, 2018).

Numerous distinct variables influence the complicated and unexpected trend of crude oil prices (Herawati and Djunaidy, 2020). The energy sector is constantly shifting, both internally and externally, as a result of such intricate and multifaceted parameters. It is widely accepted that such factors are getting more and more complicated, taking into account both theoretical and practical elements. Forecasting oil prices effectively is consequently very difficult due to the extremely volatile nature of the oil market (Lu et al., 2021).

It is noteworthy to mention here that both the reliable prediction techniques as well as the authenticity of the crude oil price time series data are two major segments that will lead to precise projection and will directly influence the global economy (Busari and Lim, 2021).

Experts in the field of financial time series utilized different univariate time series as well as econometric models to forecast the crude oil prices precisely. The most common of these methods are generalized autoregressive conditional heteroscedasticity (GARCH) (Hou and Suardi, 2012), error correction models (ECM) (He, Wang and Lai, 2010), random walks (RW) (Chikobvu and Chinhamu, 2013), vector autoregressive regressive models (VAR) (Zhou et al., 2023), and autoregressive integrated moving averages (ARIMA) (Zhao and Wang, 2014).

These traditional time series and econometric models undoubtedly produce reliable forecast, but their practicality is usually limited because of their dependence on certain parameters that is why failed to capture these changes and therefore, produces poor prediction results (Wu, Wu and Zhu, 2019).

In order to address these issues in the classical methods and enhance the prediction accuracy, experts in the financial time series focused on the hybrid models, which are based on different data decomposition (Looney and Mandic, 2008) and machine learning techniques.

Several researchers have shown in their studies that among other machine learning techniques the most widely used are SVM and ANN, which produces promising results as compared to classical methods that requires some assumptions [Deng, Ma and Zeng (2021), Srijiranon, Lertratanakham and Tanantong (2022) and Qiu, Suganthan and Amaratunga (2018)].

Over time, ANN have gained popularity as a potent computational framework that resembles the way the human brain works (Hamdi and Aloui, 2015). Complex time series data with a nonlinear nature can be efficiently modeled and predicted using ANN techniques. These techniques work well when dealing with noisy, dynamic data sets in

enormous numbers (Shabri and Samsudin, 2014). One benefit of this method is that it produces useful data, which are comparable to real-world scenarios while also decreasing inaccuracies. It is a rapidly expanding machine learning method for making predictions about data (Kareem, Hamad and Askar, 2021).

In order to forecast the WTI crude oil prices, a unique hybrid approach is suggested that depends on the data decomposition method called CEEMDAN which is combined with SVM. The main goal of the CEEMDAN technique is to separate the time series into discrete monotonous residues and IMFs. This is accomplished by means of an iterative sifting procedure that separates the prominent oscillatory modes until a monotone residue is found, meticulously dissecting the fluctuations present in the data (Liu et al. (2012) and Ali et al. (2023)).

The main objective of this research work is to propose an improved hybrid model that is based on the data decomposition and supervised machine learning technique to forecast the daily closing prices of WTI. This research work utilized not only the benchmark mark EMD to decompose the data but also used its different variants such as ensemble empirical mode decomposition (EEMD), and CEEMDAN. The most popular supervised machine learning techniques including the SVM, decision tree, random forest, and ANN are trained by providing the extracted IMFs and residue as input features.

The remainder of this paper is organized as follows. The related work is presented in section 2. Methods and materials are given in section 3. The proposed hybrid model is studied in section 4. The accuracy measures are presented in section 5. Section 6 provides the results and discussion. Finally, conclusion of the paper in section 7.

2. RELATED WORK

Recently, financial time series experts and analysts have found forecasting of crude oil prices to be an intriguing and appealing area of study. The daily price of crude oil is predicted using a variety of methods, including hybrid models, machine learning approaches (especially SVM and ANN), and traditional time-series forecasting methods. This part will address research articles that predominantly use these hybrid versions of soft computing models, which are widely used to forecast crude oil prices. Nevertheless, ANN and its hybrid forms have been used in a large number of researches, a few of which are covered here, in the literature.

Ding (2018) addressed the challenge of forecasting international crude oil prices through a hybrid modeling approach that combines different methods to enhance prediction accuracy. The suggested technique is referred to as EEMD-ANN-ADD, which combines the ANN along with EEMD by adding a decompose-ensemble component to a single AI model. Using Akaike's information criterion (AIC) to select a model, the EEMD for data breakdown, ANN for individual forecasting, and adding ensemble technique for ensemble predictions are the four steps in the methodology. The "decomposed-ensemble" algorithms outperform conventional composite algorithms in terms of forecasting precision, according to the Diebold-Mariano test, which has been adjusted to account for both level and directional measurements.

Song et al. (2021) utilized three different decomposition methods such as time varying filtering based empirical mode decomposition (TV-EMD), wavelet transform (WT), and complementary empirical mode decomposition (CEEMD). After decomposing, the series the Elman neural network (ENN) were used to build the model. It is evident from this study the TVF-EMD-ENN model outperforms the other models in terms of prediction accuracy.

Shambora and Rossiter (2007) used a model based on ANN using moving average crossover as inputs. These forecasts serve as the foundation for the financial indicators that are produced for both purchases and sales. According to the ANN model, the crude oil futures market is currently exhibiting remarkable profits, highlighting a question about the system effectiveness.

A novel method known as the EEMD-SBL-ADD was proposed by Li et al. (2018) for predicting nonlinear and nonstationary crude oil prices. The individual forecasts are combined by adding them in the last stage. Several assessment metrics are taken into account, such as the runtime duration, Diebold-Marino (DM) test, RMSE, Dstat, MAPE, and model confidence set (MCS) test, all of which show greater accuracy than the current forecasting techniques.

As part of metrological research, Ruiz-Aguilar et al. (2021) suggested a hybrid model, which is capable of forecasting wind speed. The investigation analyzes data gathered from the Bay of Algeciras, Bay of Algeciras, and Spain, using ANNs and ensemble learning techniques. There are several hourly prediction horizons, each with its own set of benefits and drawbacks. In terms of short-term (1h) and medium-term (24h) predictions with good correlation coefficients, the EMD-PE-ANN technique surpasses individual ANN algorithms.

A novel method called the WANN is proposed by Shabri and Samsudin (2014) to forecast crude oil prices on a daily basis in which ANN is used in conjunction with discrete wavelet transforms. The authors has demonstrated that when a single day lead time is given for the forecast, the WANN model outperforms a normal ANN model in predicting crude oil prices.

Shabri and Samsudin (2014) proposed the use of hybrid models in which one is based on EMD and the other on ANNs. In contrast to SVR and standalone ANN models without decomposition, short-term Nifty stock index forecasts produced by this hybrid EMD-ANN model are significantly better.

Although a considerable amount of research has been done on the use of several models, including machine learning methods, to forecast crude oil prices, there is a clear lack of research on the CEEMDAN-SVM model in particular. Although the CEEMDAN-SVM model is known to be beneficial in managing nonlinear and nonstationary data, there is not much research that thoroughly assesses its efficacy, resilience, and applicability in various market scenarios and time periods. Moreover, earlier research did not provided a comparison analysis of the CEEMDAN-SVM model and other hybrid methods. It is critical to fill this research gap in order to assess the reliability and effectiveness of the CEEMDAN-SVM model as a crude oil price prediction technique.

3. METHODS AND MATERIAL

The segmentation techniques employed in this research, such as EMD and CEEMDAN, which successfully simplify the complex structure of the data to independent IMFs and a single monotone residue is briefly reviewed in this part of the manuscript. We also discuss how SVM is used as a key component of the suggested model. Thus, we examine the SVM approach after explaining the decomposition strategies.

3.1 Empirical Mode Decomposition (EMD)

As a signal preprocessing procedure, Huang et al. (1998) first presented the EMD approach in 1998. This method is commonly used for decomposing dynamic signals, enabling the division of time series into several IMFs and a monotonic residue by applying the Hilbert-Huang transform (HHT) technique [Kong and Zhang (2020), Agana and Homaifar (2018) and Yu, Dai and Tang (2016)]. The complex signal is deconstructed into discrete oscillatory segments with varying frequencies, resulting in the isolation of a monotonic residue. The determination of the IMF components is predicated on the satisfaction of two conditions: (i) The extrema and zero-crossing points are either the same or differ by only one point, and (ii) Both the mean value of the upper envelope and the mean value of the lower envelope must be zero at any given point. The step-by-step procedure of the EMD technique is outlined as follows:

i. Identify all the data $\{y_i\ (t)\}$ localized extrema, or localized maxima and minima.

ii. Determine the data upper and lower envelope denoted by $\{U(t)\}$ and $\{L(t)\}$.

iii. Connect each of the minima and maxima using the cubic spline interpolation method to get the average of the upper and lower envelope, or $M(t)$:

$$Mean(t) = \frac{U(t) + L(t)}{2} \tag{1}$$

iv. Subtract the mean envelope calculated in Step 3 from the original signal to obtain the first component, i.e.

$$k_1(t) = y(t) - Mean(t) \tag{2}$$

The first IMF can be regarded as $k_1(t)$ if it satisfies the two requirements for the IMF as stated above; if not, Steps 1 through 4 will be repeated, treating $k_1(t)$ as a new data.

v. To obtain $r_1(t)$, the initial IMF identified in Step 4, is removed from the actual data $y(t)$, i.e.

$$r_1(t) = y(t) - k_1(t) \tag{3}$$

vi. Here, $r_1(t)$ is viewed as a fresh signal, and the filtering process from Step 1 is repeated. The real signal, $y(t)$, will be decomposed as follows after the last EMD phase, and the overall signal trend will be a smooth monotonic residue:

$$y(t) = \sum_{i=1}^{n} k_i(t) + r_n \tag{4}$$

It is important to note that r_n is the residue and $k_1(t), k_2(t), \ldots, k_n(t)$ are all different IMFs with varying frequencies ranging from high to low. Where r_n is the residue and $k_1(t), k_2(t), \ldots, k_n(t)$ are different IMFs with different frequencies that vary from high to low. The detailed decomposition is presented in the following flowchart given in Figure 1.

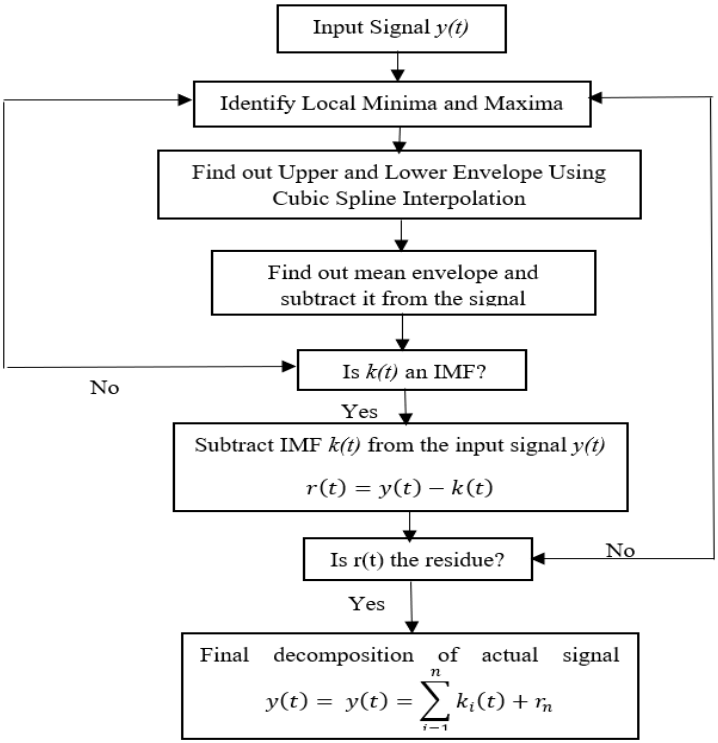

Figure 1: Schematic View of the EMD Algorithm

3.2 Complementary Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

By adding Gaussian white noise to the signal, the mode-mixing issue in the EMD method can be resolved. This new variant of EMD is known as EEMD (Wu and Huang, 2009). Reconstruction errors may result from the EEMD technique inability to eliminate Gaussian white noise following signal reconstruction. The full ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was introduced by Torres et al. (2011) as a solution to this problem.

It effectively resolves the mode mixing issue, drastically reduces computation costs, and makes reconstruction mistakes insignificant. Give an account of the function $E_j(\cdot)$, which allows the jth mode to be reached by EMD and let $w_j(\cdot)$ be the standard normal distribution white noise. The subsequent steps outline the CEEMDAN technique:

i. By applying the EMD approach, the Gaussian white noise added signal $y_i(t) = y(t) + \gamma_0 w_i(t)$ (where γ_0 is a noise coefficient, $i = 1,2,\dots,L$) can be decomposed to produce the very first IMF. Next, we define the first mode as follows:

$$\overline{IMF_1} = \frac{1}{L}\sum_{i=1}^{L} IMF_{i1} \tag{5}$$

ii. Calculate the first residue

$$r_1(t) = y(t) - \overline{IMF_1} \tag{6}$$

iii. Decompose residue $r_1(t) + \gamma_1 E(w_i(t))$ to obtain the 2nd mode as

$$\overline{IMF_2} = \frac{1}{L}\sum_{i=1}^{L} E_1\left[r_1(t) + \gamma_1 E_1\left(w_i(t)\right)\right] \tag{7}$$

iv. The residue that results can be acquired by mathematically replicating a similar procedure for every IMF.

$$R_m(t) = y(t) - \sum_{j=1}^{m} \overline{IMF_j} \tag{8}$$

The symbol *"m"* stands for the total number of IMFs. At different intervals, the IMFs collectively deconstruct the original signal attributes. The residue effectively reduces the predicted error and captures the flatter trend of the actual data.

3.3 Support Vector Machine (SVM)

Vapnik (1995) introduced the SVM method, which is currently the most used supervised machine learning approach that utilizes structured risk reduction criterion and statistical theory. The SVM approach can be applied to regression as well as classification tasks in real-world scenarios. It is usually applied to categorization difficulties, though. Unlike previous machine learning algorithms as BPN, which operate on the premise of minimizing the empirical error, the fundamental idea behind SVM is to contract the upper bound of generalization error (Lin et al., 2008). SVM in general is formulated as a minimization problem, mathematically:

$$min\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i \text{ Subject to } y_i\left((\omega.\phi(x_i)) + b\right) \geq 1 - \xi_i \text{ with } \xi_i \geq 0 \tag{9}$$

The parameter *C* is a regularization factor and it is one of the SVM hyperparameters: this constant must be set before solving the minimization problem, $\phi(x_i)$ is a nonlinear transformation that takes the data into a high dimensional space, which is called reproducing kernel Hilbert space also known as the feature space. Fortunately, $\phi(x_i)$ does not need to be computed explicitly. The dual formulation of the SVM minimization does not need $\phi(x_i)$ but only the inner product of the transformation of two data points $\phi(x_i)^T\phi(x_j)$. The dual form of the SVM minimization is:

$$\min \frac{1}{2}\sum_{ij} y_i\alpha_i y_j\alpha_j K(x_i, x_j) - \sum_i \alpha_i$$
$$\text{subject to } \sum_i y_i\alpha_i = 0 \text{ with } 0 \le \alpha_i \le C \tag{10}$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is a kernel function that represents the ϕ mapping.

The SVM kernel is a function that changes non-separable problems into separate ones by taking a low dimensional input space and transforming it into a higher dimensional space. It is most helpful in cases with non-linear separation. In nonlinear separation problems, a kernel function $K(x_i, x_j)$ is utilized to carry out the mapping and construct a hyperplane. There is a trade-off between the misclassification error and maximizing the margin that can be controlled by the parameter 'C', technically known as the regularization parameter. The most well-known and widely used kernels for SVM are linear, polynomial, and radial basis. The mathematical structure of these kernels are defined in the following equations (11-13).

$$\text{Linear Kernel: } K(x_i, x_j) = a + \sum_{i=1}^{n} b_i \,.\, (x, x_i) \tag{11}$$

$$\text{RBF Kernel: } K(x_i, x_j) = exp\left(-\sigma\left\|x_i - x_j\right\|^2\right) \tag{12}$$

$$\text{Polynomial kernel: } K(x_i, x_j) = (\,\sigma x_i \cdot x_j + C)^d, \tag{13}$$

Just "C" which is a needs to be optimized as a hyperparameter for the linear kernel. However, for the polynomial kernel, there are three parameters to tune: C, d, and σ. In contrast, for the RBF kernel, both the regularization parameter C and the kernel parameter σ need to be optimized simultaneously with the use of the grid search method.

4. THE PROPOSED HYBRID MODEL

The proposed methodology unfolds in two pivotal stages. Initially, leveraging advanced techniques like EMD, EEMD, SEMD, and CEMDAN, it adeptly decomposes the nonlinear and nonstationary time series data of oil prices into a set of distinct IMF components alongside a singular monotone residue. Following this, in the subsequent stage, these precisely derived IMF components and the residue are harnessed as input features for the development of hybrid models, encompassing ANN, SVM, Random Forest, and Decision Tree. The culmination involves a rigorous comparison of these hybrid models through accuracy metrics such as RMSE, MAPE, and MAE. The subsequent paragraphs provide an overview of every step details.

Step 1: The historical data for the price of crude oil was meticulously collected from the Yahoo Finance website. The obtained unprocessed data was then thoroughly preprocessed to satisfy the requirements needed for the efficient application of EMD and its variations. Furthermore, dividing the data into an 80% training set and a 20% testing set was an essential step. To evaluate the expected accuracy of the proposed model, this was done purposefully.

Step 2: Several distinct decomposition approaches, including EMD, EEMD, CEEMDAN, and SEMD, were needed for this comprehensive analysis of crude oil price data. First, the data was separated into separate components, like IMFs, which are characterized by a range of frequencies ranging from low to high. One monotone residue was identified in order to improve our understanding of the fundamental patterns in the historical data on the price of crude oil.

Step 3: In this phase, the suggested approach is built with an emphasis on achieving the greatest efficiency with lowest values for RMSE, MAE, and MAPE. Since different supervised learning algorithms are used to build the proposed hybrid models. Therefore, the IMFs and residue components separated from the actual WTI crude oil prices in step 2 is used as input features to build the architecture of different hybrid models.

Step 4: Forecasts were made using a number of models, notably the CEEMDAN-ANN, SEMD-ANN, EEMD-ANN, and EMD-ANN. The experimental results unmistakably demonstrate that the CEEMDAN-SVM model performs better than the ensemble models by reaching minimal values in statistical indicators like RMSE, MAPE, and MAE. This demonstrates how the accuracy and prediction capabilities of the CEEMDAN-SVM model surpass those of its ensemble counterparts.

Step 5: In this stage, a thorough evaluation of the CEEMDAN-SVM hybrid model was conducted in conjunction with other hybrid machine-learning models. Amongst the performance measures that formed the basis of the assessment were RMSE, MAPE, and MAE. This comparative study sheds light on how accurate and effective the suggested CEEMDAN-SVM model is in comparison to other hybrid models.

The schematic view of the proposed model is presented in Figure 2.

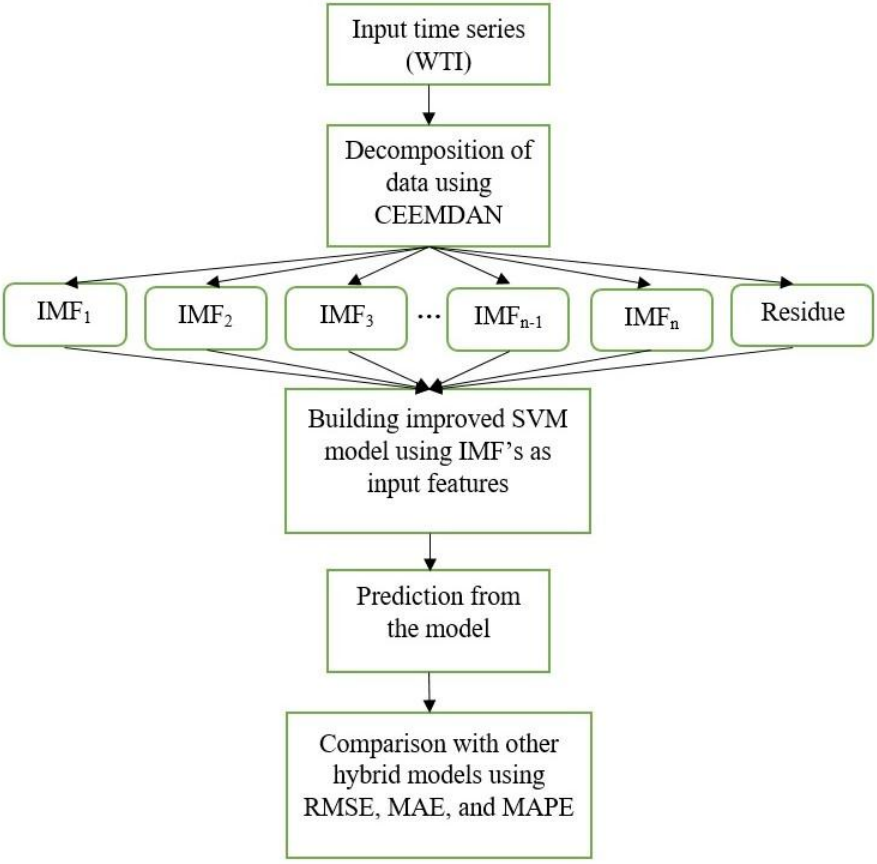

Figure 2: Schematic View of the Proposed CEEMDAN-SVM Model

5. ACCURACY MEASURES

We employ three main performance metrics called the RMSE, MAPE, and MAE. These metrics are listed numerically below, to assess the suggested models accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{A}\sum_{a=1}^{A}(y_a - \acute{y_a})^2} \tag{14}$$

$$\text{MAPE} = \sum_{a=1}^{A}\frac{|y_a - \acute{y_a}|}{|y_a|} \times \frac{100}{A} \tag{15}$$

$$\text{MAE} = \frac{1}{A}\sum_{a=1}^{A}|y_a - \acute{y_a}| \tag{16}$$

Across this study, it is imperative to clarify that the numeral y_a represents the quantities that actually occurred at a given time, while $\acute{y}_a$ represents the numbers that were projected over the same temporal span a. The total number of observations, appropriately denoted by A, is crucial to the evaluation. It is crucial to emphasize that a model's effectiveness is based on how low these indicators are.

6. RESULT AND DISCUSSION

6.1 Preprocessing of the Data

The investigation is based on the price data pertaining to West Texas Intermediate (WTI) crude oil, encompassing the period from March 1, 2003 to October 31, 2023 and comprising a total of 5364 observations. The data was accurately sourced from the Yahoo Finance website (https://finance.yahoo.com). To ensure the reliability of subsequent predictions, a crucial preprocessing phase was undertaken on the collected data before it was employed for model training. The behavior of the WTI crude oil closing prices in the time interval is shown in Figure 2.

Addressing the challenge of missing values, which can significantly affect the model's predictive capabilities, a strategic approach was adopted for imputation. Recognizing the influence of past moments on crude oil prices, a method was employed that leveraged the mean of the two preceding values and the value of the subsequent moment to fill in the gaps within the data. Furthermore, a separation of the data into 80% training and 20% for testing was implemented to facilitate a robust evaluation of the proposed model.

Furthermore, different libraries such as ggplot2, readxl, DescTools, EMD, Rlibeemd, reshape2, tidyr, e1071, caret, quantmod, randomForest, rpart, and neuralnet are utilized in RStudio for data analysis and building the proposed hybrid model.

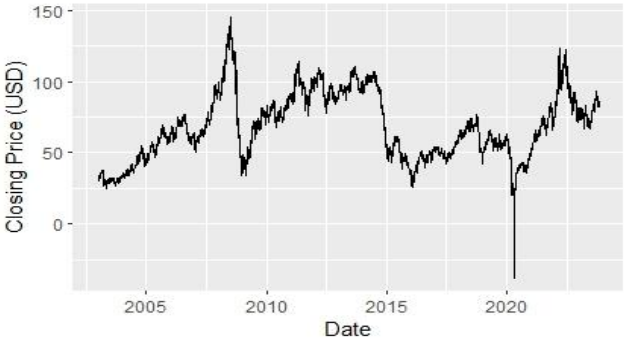

Figure 3: WTI Crude Oil Closing Prices in the Time Interval January 1, 2003 to October 31, 2023

Table 1
Descriptive Statistics of WTI Daily Closing Prices

Count	Min	Max	Mean	Median	Mode	SD
5364	-37.63	145.29	68.19	65.66	44.66	23.43

Table 1 presents a comprehensive overview of the descriptive statistics pertaining to the daily closing prices of West Texas Intermediate (WTI) crude oil. With a total of 5364 observations, this dataset captures a diverse range of price fluctuations in the market. The minimum recorded value is -37.63 USD, indicating an anomaly that can be attributed to the unprecedented events surrounding the COVID-19 pandemic.

The maximum closing price observed is 145.29 USD, showcasing the upper bounds of price levels within the examined period. The mean closing price stands at 68.19 USD, providing a measure of central tendency that represents the average value over the dataset. Complementing the mean, the median, which is 65.66 USD, offers additional insights into the distribution of prices, especially in the context of potential outliers.

Interestingly, the mode of 44.66 USD signifies a recurring value, shedding light on a specific price point that frequently occurred. This could be indicative of market conditions or influential factors during certain periods. Adding a layer of understanding to the dataset, the standard deviation (SD) is calculated at 23.43 USD, representing the degree of dispersion of individual data points from the mean.

It is crucial to acknowledge that the occurrence of an outlier, such as the significantly smaller can be directly linked to the global financial disruptions triggered by the COVID-19 pandemic. This anomaly serves as a distressing reminder of the extraordinary economic impacts witnessed during this period, contributing to the broader narrative reflected in the WTI crude oil prices.

6.2 Decomposition of Data with EMD, EEMD, SEMD, and CEEMDAN

Various techniques, including empirical mode decomposition (EMD), ensemble EMD (EEMD), statistical EMD (SEMD), and complete ensemble EMD with adaptive noise (CEEMDAN), were employed to analyze the closing prices of crude oil. The primary objective was to decompose these values into monotone residues and multiple intrinsic mode functions (IMFs). Figure 3 visually represents the results of this decomposition using EMD technique. Similarly, IMF's constructed by EEMD, CEEMDAN and SEMD are presented in Figure 4, Figure 5 and Figure 6, respectively.

In the graphical representation, it is observed that the IMFs extracted using EMD, EEMD, and CEEMDAN exhibit similarities. However, the SEMD method generates fewer IMFs compared to EMD, EEMD, and CEEMDAN, which shows that replacing the cubic spline interpolation with smoothing never improves the decomposition results. This distinction in the number of IMFs highlights the unique characteristics and outcomes of each decomposition technique in capturing the underlying patterns in crude oil closing prices.

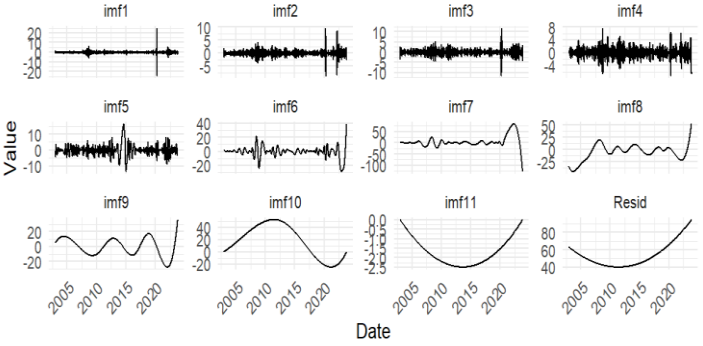

Figure 4: Decomposition of WTI Daily Closing Price with the Method of EMD

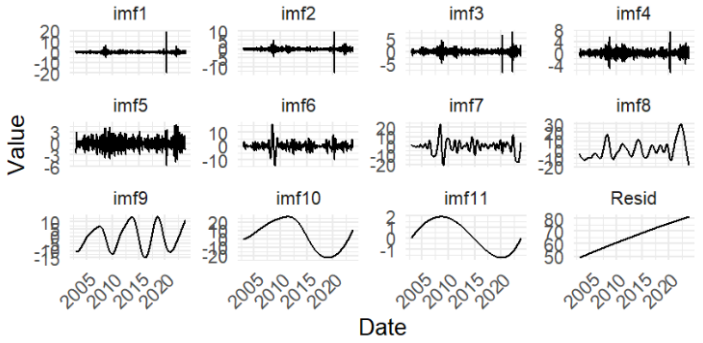

Figure 5: Decomposition of WTI Daily Closing Price | with the Method of EEMD

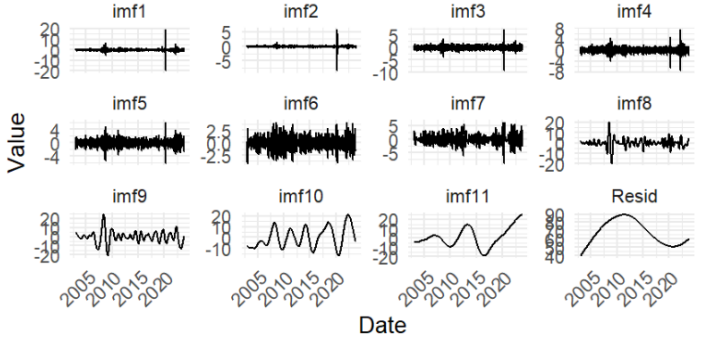

Figure 6: Decomposition of WTI Daily Closing Price with the Method of CEEMDAN

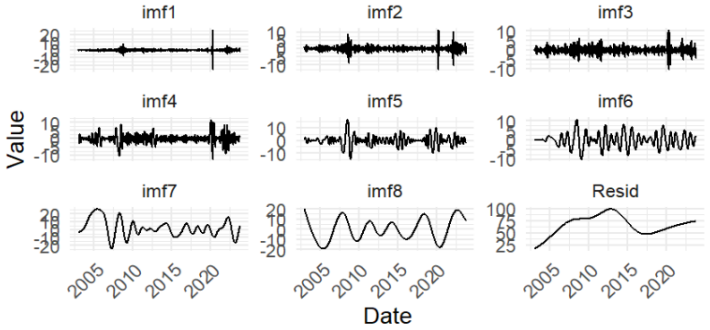

Figure 7: Decomposition of WTI Daily Closing Price with the Method of SEMD

6.3 Intrinsic Mode Function (IMF)

The comprehensive descriptive statistics, encompassing key metrics like mean, median, standard deviation, minimum, and maximum values, have been precisely computed for each intrinsic mode function (IMF) as well as the actual crude oil prices. The insights derived from this analysis are presented in Tables 2-5, offering a detailed breakdown of the decomposition process.

Exploring the results from Table 2, the raw data unfolds with a spectrum of characteristics. From a minimum value of -37.694 to a maximum of 145.290, the data showcases a diverse range. The entire distribution is reflected in the median of 64.900 and mean of 68.070, which denote the center inclinations. In the meantime, the data standard deviation of 23.694, which captures sensitivities and variations, indicates a significant level of volatility.

As we focus on each particular IMF, interesting trends start to show up. The amplitudes of IMF1 and IMF2 are comparatively small, ranging from -24.334 to 24.497. This implies that there may be a lot of noise or high-frequency oscillations among these components. Having readings ranging from -11.265 to 11.244, IMF3 and IMF5 show heterogeneous trends, a combination of positive and negative deviations. IMF4 is particularly notable since it has a wider range, ranging from -15.298 to 10.279. Its positive bias, which is represented by a mean of 0.103 and suggests that the data may contain an ongoing pattern, is prominent. From IMF6-IMF11, there is a range of different wavelengths and amplitudes. Comparing these components with each other, they show larger ranges and higher standard deviations, which illustrate the fine features that the decomposition method captures.

Finally, intriguing features are revealed by the residue, which represents the information that remains after breakdown. The residue captures the rest of the details, with a minimum value of 55.95 and a maximum value of 78.86. The residual variability within this component is highlighted by the mean of 77.46 and the standard deviation of 10.88, which provide important insights into the complexities of the decomposed data.

Table 2
Statistical Measures of IMFs and Residue Extracted by Implementing EMD

Description	Min.	Median	Mean	SD	Max.
Actual	-37.63	64.900	68.070	23.694	145.290
IMF1	-24.334	0.024	0.016	1.142	24.497
IMF2	-9.453	-0.001	0.005	1.020	9.653
IMF3	-11.265	-0.004	-0.006	1.335	11.244
IMF4	-15.298	0.046	0.103	1.842	10.279
IMF5	-10.911	0.002	-0.051	2.523	15.359
IMF6	-11.864	0.070	0.295	4.667	19.401
IMF7	-58.215	-0.352	-4.083	15.617	42.787
IMF8	-83.296	-1.219	-6.551	28.006	36.487
IMF9	-20.367	0.881	1.729	12.146	26.061
IMF10	-12.436	-0.233	-0.985	6.374	15.942
IMF11	-0.956	0.136	0.135	0.783	1.226
Residue	55.950	78.860	77.460	10.888	93.410

Table 3
Statistical Measures of IMFs and Residue Extracted by Implementing EEMD

Description	Min.	Median	Mean	SD	Max.
Actual	-37.63	64.900	68.070	23.694	145.290
IMF1	-19.83	0.008	-0.0004	0.856	19.195
IMF2	-13.87	0.002	0.0008	0.680	10.049
IMF3	-8.653	-0.006	0.001	0.837	7.058
IMF4	-6.686	0.011	-0.005	1.025	6.896
IMF5	-5.838	-0.008	0.008	1.520	5.692
IMF6	-14.73	-0.110	-0.098	3.531	16.844
IMF7	-20.96	0.122	0.200	6.761	23.351
IMF8	-12.30	-1.397	0.180	8.641	24.016
IMF9	-15.83	-0.245	-0.636	8.316	12.875
IMF10	-20.67	6.771	5.249	15.918	26.738
IMF11	-0.455	0.332	0.305	0.527	1.022
Residue	44.820	65.410	62.870	7.825	70.930

Table 3 presents a comprehensive statistical analysis that clarifies the unique characteristics of both the residue from the application of EEMD and the IMFs. Examining the IMFs values from IMF1 to IMF11, we find some interesting trends. IMF1 displays a wide range, with a minimum of -19.830 and a maximum of 19.195, suggesting significant variations in the data at hand. IMF11, on the other hand, has a narrower spectrum, ranging from -0.455 at the lowest value to 1.022 at the highest point, indicating lower fluctuation.

The median values of the IMFs represent their central trends. For example, IMF11, with a median of 0.332, provides information on its central trend, whereas IMF1, with a median of 0.008, depicts the middle of its distribution. Mean values provide hidden information that help us comprehend average behavior even further. Measuring the average

characteristics of the data are IMF1, which has a mean of -0.0004, and IMF11, which has a mean of 0.305.

Examining the standard deviations (SD) reveals how the data points are distributed around the mean. IMF5, which has a high standard deviation of 1.520, indicates more variation in its data points. IMF11, on the other hand, shows more clustered data points, indicating a certain level of homogeneity, with a comparatively low SD of 0.527. Now that we are focusing on the residue, this part stands for the unaccounted-for fraction of the signal that the IMFs were unable to catch. As evidenced by its mean value of 62.870 and median value of 65.410, the residue exhibits distinctive features and variability. The uniqueness of residual component is further highlighted by its 7.825 standard deviation, which provides important information about the details of the decomposed signal.

Table 4
Statistical Measures of IMFs and Residue Extracted by Implementing SEMD

Description	Min.	Median	Mean	SD	Max.
Actual	-37.63	64.900	68.070	23.694	145.29
IMF1	-25.84	0.024	0.0023	1.221	25.854
IMF2	-12.829	0.0025	0.010	1.173	11.091
IMF3	-10.425	0.013	0.017	1.591	10.519
IMF4	-12.471	-0.020	-0.0115	1.777	11.565
IMF5	-14.737	0.0362	-0.0221	4.197	16.340
IMF6	-10.129	-0.020	-0.099	4.625	10.671
IMF7	-24.313	0.940	1.931	6.125	25.928
IMF8	-19.214	-0.183	0.044	8.924	19.809
Residue	17.54	67.820	66.300	13.906	99.390

A detailed summary of the wide range of values contained in the residue and intrinsic mode functions (IMFs) can be seen in Table 4. Interesting trends and features embedded in the broken components are revealed by this study. The component with the widest range of values, IMF1, is the most volatile, ranging from -25.840 to 25.854. IMF8, on the other hand, has an even more limited range, ranging from -19.214 to 19.809, indicating a smaller oscillation magnitude. Upon analyzing the IMFs central trend, we discover that each component's median value varies. Having a median value of 0.940, IMF7 leads the group and shows a central tendency towards positive values. IMF8, at the other hand, has the smallest median (-0.183), indicating a propensity for values that are negative. Further information about the average behavior of the IMFs can be gleaned from mean values. IMF6 has the smallest mean of -0.099, suggesting a predisposition towards negative deviations, whereas IMF7 stands out with the highest mean of 1.931, showing an inclination towards positive deviations. The standard deviations, which range from 1.173 to 8.924, highlight the differences across the IMF's dispersion. Smaller numbers for the standard deviation suggest larger clusters of data, whereas larger numbers show greater variability in the distribution of data points. Now that we are looking at the residue, we can see that it ranges from 17.540 to 99.390. With a mean of 0.044 and a median value of 67.820, the residue exhibits distinctive features that set it apart from the IMFs. This

highlights even more how complex the deconstructed signal is and how unique characteristics are captured in each component.

Table 5
Statistical Measures of IMFs and Residue Extracted by Implementing CEEMDAN

Description	Min.	Median	Mean	SD	Max.
Actual	-37.630	64.900	68.070	23.694	145.290
IMF1	-19.855	0.008	-0.0001	0.857	19.235
IMF2	-9.234	0.0002	0.0003	0.270	5.993
IMF3	-9.125	-0.001	0.003	0.664	7.091
IMF4	-7.258	0.002	-0.002	0.833	6.996
IMF5	-6.011	0.003	-0.0003	0.983	5.930
IMF6	-3.975	-0.0004	-0.0003	1.308	4.271
IMF7	-9.728	-0.011	0.006	2.093	6.636
IMF8	-19.903	-0.091	-0.137	4.577	19.732
IMF9	-20.827	0.544	0.662	7.007	23.563
IMF10	-17.183	-3.299	-0.844	8.957	15.324
IMF11	-22.063	0.349	-0.039	10.924	24.090
Residue	36.310	68.980	68.420	14.302	87.670

A thorough examination of the statistical metrics connected to the residue obtained from the CEEMDAN decomposition and intrinsic mode functions (IMFs) is shown in Table 5. Starting with IMF1, we get a broad range of values, indicating notable swings, from -19.855 to 19.235. The median value of 0.008 sheds light on the distinctive nature of IMF1 by indicating a core propensity towards positive deviations.

Regarding IMF3, the median of -0.001229 and the mean value of 0.003 suggest that there is moderate amount of variation within this component. These metrics represent the average and central tendency of IMF3, providing important details regarding its general trend.

IMF8, on the other hand, has unique properties. With a standard deviation of 4.577 and a negative mean of -0.137, IMF8 indicates substantial variation with the possibility of anomalies. The range of values is -19.903 to 19.732. The wide range of numbers emphasizes how dynamic IMF8 is.

When we look at the residue, we see that its mean value is 68.420 and its median is 68.980. These measurements point to a largely consistent structure of the residue part. On the other hand, the value that ranges from 36.310 to 87.670 suggests that there may be unusual or unusual values in the data.

To sum up, the statistical measurements offer significant understanding of the features and differences found in the IMFs and residue obtained through the CEEMDAN breakdown. These findings add to a thorough comprehension of the broken down elements and their distinctive characteristics.

6.4 Comparison of the Proposed Model

The compelling experimental findings presented in Table 6 unequivocally highlight the superior predictive capabilities of the CEEMDAN-SVM model when compared to its counterpart's hybrid models. The exceptional performance of the CEEMDAN-SVM model is evident through remarkably low values of key performance indicators, with an RMSE of 1.446, MAE of 1.259, and MAPE of 2.194, nearly approaching zero. These indicators collectively emphasize the model's potential to deliver highly accurate forecasts.

Table 6
Performance Metrics of Different Hybrid Models Used for WTI Crude Oil Prediction

Model	RMSE	MAE	MAPE
ARIMA	27.105	22.667	34.582
EMD-SVM	24.605	21.324	2.554
EMD-Random Forest	31.113	24.471	2.671
EMD-Decision Tree	41.894	34.023	2.907
EMD-ANN	1.931	1.710	2.404
EEMD-SVM	20.422	17.949	2.464
EEMD-Random Forest	26.535	24.682	3.602
EEMD-Decision Tree	16.569	13.189	14.304
EEMD-ANN	1.534	1.509	2.611
SEMD-SVM	2.530	1.950	2.480
SEMD-Random Forest	16.029	12.669	3.348
SEMD-Decision Tree	15.426	13.607	16.331
SEMD-ANN	1.545	1.987	3.985
CEEMDAN-SVM	1.446	1.259	2.194
CEEMDAN-Random Forest	16.468	13.057	2.955
CEEMDAN-Decision Tree	21.109	17.144	2.462
CEEMDAN-ANN	1.581	1.710	2.363

In a holistic assessment across various scenarios, the CEEMDAN-SVM model consistently outperforms its counterparts, as detailed in Table 6. The CEEMDAN-SVM model exhibits the smallest values for essential statistical measures, including RMSE, MAE, and MAPE. The EEMD-ANN model secures the second-best position with an RMSE of 1.534, followed by the CEEMDAN-ANN model with an RMSE of 1.81, showcasing their respective forecasting capabilities for crude oil prices.

While exact actual and predicted values are not provided here, a comprehensive summary of this research endeavor is effectively conveyed through visual representations in Figure 7. These figures offer a strategic and simplified insight into the alignment between actual and predicted closing prices of crude oil. Moreover, Figure 7 visually indicates a negligible difference between actual and predicted values. However, a closer examination of the accuracy metrics reveals that the CEEMDAN-SVM model attains the lowest values for RMSE, MAE, and MAPE, solidifying its status as the most accurate model for predicting crude oil prices.

Figure 8: Comparison of Actual and Predicted WTI Crude Oil Daily Closing Prices using CEEMDAN-SVM Model

7. CONCLUSION

The main purpose of this research work is to propose a hybrid model that predicts the WTI crude oil daily closing prices accurately. The proposed hybrid model is based on a new variant of EMD, which is known as CEEMDAN and supervised learning algorithm called SVR.

The experimental findings presented in Table 6 clearly highlight the superior predictive performance of the CEEMDAN-SVM model when compared to other hybrid models such as EMD-ANN, EEMD-ANN, SEMD-ANN and others. The exceptional performance of the CEEMDAN-SVM model is evident through remarkably low values of key performance indicators, with a RMSE of 1.446, MAE of 1.259, and MAPE of 2.194, nearly approaching zero. These indicators collectively emphasize the model potential performance to deliver highly accurate forecasts.

In a rigorous assessment across various scenarios, the CEEMDAN-SVM model consistently outperforms its counterparts, as detailed in Table 6. The CEEMDAN-SVM model exhibits the smallest values for essential statistical measures, including RMSE, MAE, and MAPE. The EEMD-ANN model secures the second-best position with an RMSE of 0.319, followed by the EEMD-ANN model with an RMSE of 1.534, showcasing their respective forecasting capabilities for crude oil prices.

While exact actual and predicted values are not provided here, a comprehensive summary of this research endeavor is effectively conveyed through visual representations in Figure 7. This figure offers a strategic and simplified insight into the alignment between actual and predicted closing prices of crude oil. Figure 8 visually indicates a negligible difference between actual and predicted values. However, a closer examination of the accuracy metrics reveals that the CEEMDAN-SVM model attains the lowest values for RMSE, MAE, and MAPE, confirming its status as the most accurate model for predicting crude oil prices.

In a nut shell, the suggested model will help not only the experts in the field of financial time series analysis but the investors as well to make wise decision in investing in the stock prices of WTI.

The main limitation of this research work that all of the supervised machine-learning techniques used in this research work requires the input features, and without suitable input features the hybrid model cannot be trained. Therefore, selecting the suitable input features often remained a challenging task, and wrong selection may distort the prediction accuracy. To overcome this limitation, the author's future work is focused on deep learning technique such as convolution neural network (CNN), long-short term memory (LSTM) network, and recurrent neural network (RNN).

ACKNOWLEDGEMENT

The authors gratefully acknowledge Qassim University, represented by the Deanship of Graduate Studies and Scientific Research, on the financial support for this research under the number (2023-SDG-1-BSRC36941) during the academic year 1445 AH/2023 AD.

Funding:

The authors gratefully acknowledge Qassim University, represented by the Deanship of Graduate Studies and Scientific Research, on the financial support for this research under the number (2023-SDG-1-BSRC36941) during the academic year 1445 AH/2023 AD.

Data Availability:

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Conflicts of Interest:

All authors declare no conflicts of interest.

REFERENCES

1. Agana, N.A. and Homaifar, A. (2018). EMD-based predictive deep belief network for time series prediction: An application to drought forecasting. *Hydrology*, 5(1), 18.
2. Ali, M., Khan, D.M., Alshanbari, H.M. and El-Bagoury, A.A.A.H. (2023). Prediction of complex stock market data using an improved hybrid emd-lstm model. *Applied Sciences*, 13(3), 1429.
3. Busari, G.A. and Lim, D.H. (2021). Crude oil price prediction: A comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. *Computers & Chemical Engineering*, 155, 107513.
4. Chen, Y., He, K. and Tso, G.K. (2017). Forecasting crude oil prices: a deep learning based model. *Procedia Computer Science*, 122, 300-307.
5. Chikobvu, D. and Chinhamu, K. (2013). Random walk or mean reversion? Empirical evidence from the crude oil market. *Istatistik Journal of the Turkish Statistical Association*, 6(1), 1-9.
6. Deng, C., Ma, L. and Zeng, T. (2021). Crude oil price forecast based on deep transfer learning: Shanghai crude oil as an example. *Sustainability*, 13(24), 13770.
7. Ding, Y. (2018). A novel decompose-ensemble methodology with AIC-ANN approach for crude oil forecasting. *Energy*, 154, 328-336.

8. Gupta, V. and Pandey, A. (2018). Crude oil price prediction using LSTM networks. *International Journal of Computer and Information Engineering*, 12(3), 226-230.
9. Hamdi, M. and Aloui, C. (2015). Forecasting crude oil price using artificial neural networks: a literature survey. *Econ. Bull.*, 35(2), 1339-1359.
10. He, Y., Wang, S. and Lai, K.K. (2010). Global economic activity and crude oil prices: A cointegration analysis. *Energy Economics*, 32(4), 868-876.
11. Heddam, S., Vishwakarma, D.K., Abed, S.A., Sharma, P., Al-Ansari, N., Alataway, A., Dewidar, A.Z. and Mattar, M.A. (2024). Hybrid river stage forecasting based on machine learning with empirical mode decomposition. *Applied Water Science*, 14(3), 46.
12. Herawati, S. and Djunaidy, A. (2020). Implementing Method of Empirical Mode Decomposition based on Artificial Neural Networks and Genetic Algorithms for Crude Oil Price Forecasting. In *Journal of Physics: Conference Series*, 1569(2), 022075). IOP Publishing.
13. Hou, A. and Suardi, S. (2012). A nonparametric GARCH model of crude oil price return volatility. *Energy Economics*, 34(2), 618-626.
14. Hu, Z. (2021). Crude oil price prediction using CEEMDAN and LSTM-attention with news sentiment index. *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles*, 76, 28.
15. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C. and Liu, H.H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971), 903-995.
16. Jothimani, D., Shankar, R. and Yadav, S.S. (2016). A hybrid EMD-ANN model for stock price prediction. In Swarm, Evolutionary, and Memetic Computing: *6th International Conference, SEMCCO* 2015, Hyderabad, India, December 18-19, 2015, Revised Selected Papers 6 (pp. 60-70). Springer International Publishing.
17. Kareem, S., Hamad, Z.J. and Askar, S. (2021). An evaluation of CNN and ANN in prediction weather forecasting: A review. *Sustainable Engineering and Innovation*, 3(2), 148-159.
18. Kong, X. and Zhang, T. (2020). Improved Generalized Predictive Control for High-Speed Train Network Systems Based on EMD-AQPSO-LS-SVM Time Delay Prediction Model. *Mathematical Problems in Engineering*, 2020(1), 6913579.
19. Li, T., Hu, Z., Jia, Y., Wu, J., & Zhou, Y. (2018). Forecasting crude oil prices using ensemble empirical mode decomposition and sparse Bayesian learning. *Energies*, 11(7), 1882.
20. Lin, S.W., Ying, K.C., Chen, S.C. and Lee, Z.J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, 35(4), 1817-1824.
21. Liu, H., Chen, C., Tian, H.Q. and Li, Y.F. (2012). A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. *Renewable Energy*, 48, 545-556.
22. Looney, D. and Mandic, D.P. (2008). A machine learning enhanced empirical mode decomposition. In 2008 *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1897-1900). IEEE.

23. Lu, Q., Sun, S., Duan, H. and Wang, S. (2021). Analysis and forecasting of crude oil price based on the variable selection-LSTM integrated model. *Energy Informatics*, 4(Suppl 2), 47. https://doi.org/10.1186/s42162-021-00166-4
24. Qiu, X., Suganthan, P.N. and Amaratunga, A.G. (2018). Ensemble incremental random vector functional link network for short-term crude oil price forecasting. In 2018 *IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1758-1763). IEEE.
25. Ruiz-Aguilar, J.J., Turias, I., González-Enrique, J., Urda, D. and Elizondo, D. (2021). A permutation entropy-based EMD–ANN forecasting ensemble approach for wind speed prediction. *Neural Computing and Applications*, 33(7), 2369-2391.
26. Shabri, A. and Samsudin, R. (2014). Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model. *Mathematical Problems in Engineering*, 2014(1), 201402.
27. Shabri, A. and Samsudin, R. (2014). Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model. *Mathematical Problems in Engineering*, 2014(1), 201402.
28. Shambora, W.E. and Rossiter, R. (2007). Are there exploitable inefficiencies in the futures market for oil? *Energy Economics*, 29(1), 18-27.
29. Song, C., Chen, X., Wu, P. and Jin, H. (2021). Combining time varying filtering based empirical mode decomposition and machine learning to predict precipitation from nonlinear series. *Journal of Hydrology*, 603, 126914.
30. Srijiranon, K., Lertratanakham, Y. and Tanantong, T. (2022). A hybrid framework using PCA, EMD and LSTM methods for stock market price prediction with sentiment analysis. *Applied Sciences*, 12(21), 10823.
31. Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In 2011 *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4144-4147). IEEE.
32. Vapnik V.N. (1995). *The nature of statistical learning Theory*, Springer Verlag, New York, Inc.
33. Wu, Y.X., Wu, Q.B. and Zhu, J.Q. (2019). Improved EEMD-based crude oil price forecasting using LSTM networks. *Physica A: Statistical Mechanics and its Applications*, 516, 114-124.
34. Wu, Z. and Huang, N.E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 1(01), 1-41.
35. Yu, L., Dai, W. and Tang, L. (2016). A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. *Engineering Applications of Artificial Intelligence*, 47, 110-121.
36. Zeng, Q., Qu, C., Ng, A.K. and Zhao, X. (2016). A new approach for Baltic Dry Index forecasting based on empirical mode decomposition and neural networks. *Maritime Economics & Logistics*, 18, 192-210.
37. Zhao, C.L. and Wang, B. (2014). Forecasting crude oil price with an autoregressive integrated moving average (ARIMA) model. *Fuzzy Information & Engineering and Operations Research & Management*, 275-286.
38. Zhou, D., Siddik, A.B., Guo, L. and Li, H. (2023). Dynamic relationship among climate policy uncertainty, oil price and renewable energy consumption—findings from TVP-SV-VAR approach. *Renewable Energy*, 204, 722-732.