IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 9 February 2024, accepted 22 February 2024, date of publication 26 February 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370440

RESEARCH ARTICLE

A Novel Approach for Reconstruction of IMFs of Decomposition and Ensemble Model for Forecasting of Crude Oil Prices

MUHAMMAD NAEEM[1], MUHAMMAD AAMIR[1], JIAN YU[2], AND OLAYAN ALBALAWI[3]
[1]Department of Statistics, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan
[2]School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 96006, New Zealand
[3]Department of Statistics, Faculty of Science, University of Tabuk, Tabuk 71491, Saudi Arabia

Corresponding author: Muhammad Aamir (aamirkhan@awkum.edu.pk)

ABSTRACT In recent eras, the complexity and fluctuations of the global crude oil prices have affected the economic progress of society. It is therefore, the oil price prediction has hauled the attention of scholars and policymakers. Driven by this critical concern for forecasting of crude oil prices, we introduces a novel hybrid model keeping in mind the primary objective of enhancing prediction accuracy while considering the specific characteristics as inherent in the data. To achieve this achievement, the trend is eliminated, allowing the scrutiny of whether the residual component validates the assurance of a series ran by stochastic trends. Following the removal of the trend, the residual component undergoes rigorous evaluation through autoregressive model following the decomposition model. Then we got support from the support vector machine, autoregressive integrated moving average and long-short term memory. The predictions accuracy can be evaluated by using the various performance metrics. The proposed hybrid model's robustness and forecasting performance are rigorously evaluated through Diebold-Mariano test in comparison to competing models. Furthermore, the forecasting ability is evaluated via directional forecast. Ultimately, the empirical findings explicitly determine the superior predictive capabilities of the proposed hybrid model over alternative approaches.

INDEX TERMS Crude oil prices, decomposition and ensemble model, forecasting, reconstructions of IMFs.

I. INTRODUCTION

Due to the globalization of economic and finance sector, the crude oil has gained significant prominence in the social and economic progress. Oil, currently, stands as the foremost and crucial energy source. The crude oil market serves as the foundation of the petroleum industry, exhibits a significantly larger trading volume compared to other oil product markets. During the previous two decades, it has garnered considerable attention due to its strong association with corporate strategic planning, risk management, household expenditures, and various other factors. The volatility of the crude oil prices has influenced not only the economic growth of countries on a large scale and the financial investment decisions of related industries, but it has also intertwined the everyday life intricately. The present global economy has been undergoing through a period of transformation due to the insignificant policies of the government, energy crises market, cultural and marketplace psychology factors. While underlying the economic indicators and market need serve as fundamental drivers of energy price fluctuations, the shortage and non-renewable quality of energy resources, combined with inequitable sharing and expenditure of resources across countries, contribute to the highly unpredictable nature of energy price developments. In particular, the outbreak of the novel coronavirus pandemic in 2020 triggered the short-term economic stagnation, bringing the transportation and manufacturing industries to close. This led to a sharp contraction in the crude oil and energy resource requirements, causing the world economy to experience a "dark moment."

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães.

Market panic ensued, reaching a point where the May futures cost of Brent crude oil turned negative. Then the year 2021, the successful rollout of COVID-19 vaccinations and robust government economic stimulus measures led to a gradual increase in energy demand and a subsequent rise in global energy prices. However, in the latter half of the year, recurring outbreaks hindered the recovery of the energy supply chain, resulting in depleted natural gas inventories in Europe and the United States. This, in turn, led to a staggering surge in the price of UK natural gas on the IPE exchange in 2021. Given that the Europe heavily relies on gas for electricity generation, the soaring natural gas prices caused electricity prices to skyrocket, resulting in significant production reductions or temporary factory closures for numerous industrial enterprises in Europe. Moreover, energy expenses in different regions worldwide experienced notable increases, with Asia facing coal shortages. Additionally, an international oil prices remained consistently high throughout the year, with a cumulative increase of over 60% in 2021. Presently, the global energy conflicts and crises continue to unfold.

Among the various energy needs, crude oil holds a vital position as a cornerstone of the industry and a crucial component of the transportation sector. It serves as a fundamental energy source, a raw material for the chemical industry, and a strategic resource necessary for socio-economic betterment. Crude oil plays a pivotal role in economic progress and remains the leading global fossil energy source, accounting for around 33% of total energy consumption. Adverse changes in crude oil prices can significantly impact a country's economic development, social stability, and national security. As a significant energy commodity, crude oil possesses not only inherent physical aspects, both financial and political characteristics in nature. Fluctuations in crude oil prices exert a substantial influence with regard to the international economy, financial markets, and local well-being. Moreover, the complex nature of crude petroleum assets categorizes the global crude oil market is an example of multi-dimensional nonlinear system. The variations in crude oil prices result from a myriad of interconnected fundamental and secondary risk factors, encompassing supply and demand dynamics in the crude oil market, changes in the value of the US currency, high-risk trading, geopolitical tensions, and natural calamities, and various other elements. Over the last 20-years, worldwide crude oil rates have exhibited significant volatility, characterized by shorter cycles of large surges and declines, and reduced periods of smooth transitions. Examples include the 2001 economic recession triggered by the bursting of the internet bubble, causing a 40% decrease in oil rates; the 2008 worldwide economic crisis, leading to a 75% decrease in oil prices; the period of May to June 2009, characterized by global economic recovery, weak dollar due to the Central bank's unconventional monetary policy, low OPEC production, and continuous flow of funds into the energy sector, resulting in a 36.5% price increase within two months. The period from 2014 to 2016 witnessed the shale oil revolution, leading to a 76% drop in oil prices. In 2018, the U.S.-China trade dispute, U.S.-Iran conflict, and increased OPEC production caused a 42% decline in oil prices. International COVID-19 crisis in 2020 had a severe impact, with WTI crude oil rates dropping sharply from \$63.27/barrel to -\$37.63/barrel, marking a staggering 159% decline. Given the high market risks associated with crude oil futures, it becomes crucial to develop robust methodologies for accurately forecasting the direction and scale of crude oil price variations. These forecasts aid microeconomic agents in identifying and hedging price risks in the crude oil market, while also uncovering potential profit opportunities.

Nevertheless, predicting the crude oil prices presents inherent challenges due to the multitude of factors that influence them. In addition to the forces of supply and demand [1], political events, the impact of alternative energy sources, stock market performance, exchange rates of major oil-importing or exporting nations, and various other elements hold significant sway over crude oil prices [2]. The interplay of these factors contributes to the intricate and unpredictable nature of the crude oil market.

Owing to the intricate interplay trying together the crude oil market as well as the factors mentioned above, the temporal data of crude oil prices exhibits non-linear and non-stationary characteristics [3]. Consequently, accurately forecasting crude oil prices proves to be a highly challenging endeavour. Extensive research has been conducted to address this issue and consequently, researchers and analysts continuously strive to develop more effective forecasting methods to enhance accuracy, which is also the main focus of our study.

During the initial stages of research, numerous conventional econometric models were developed. These models has three categories [4]: Simple models, models employing exponential smoothing techniques, and AR models, including the autoregressive integrated moving average (ARIMA) model [5], ARCH model [6], and generalized autoregressive conditional heteroskedasticity (GARCH) model [7]. The models within the AR family have gained popularity in effectively modelling and predicting the price fluctuations' volatility in crude oil, thanks to their impressive performance [4]. These traditional econometric models have established a strong foundation for the advancement of forecasting the crude oil prices. Though, it is widely recognized that conventional models struggle to accurately capture the non-linear components present in the temporal data of crude oil prices [4].

With the advancement of computational technologies, a range of modern methods have emerged, characterized by the integration of artificial intelligence approaches. These methods incorporate artificial neural networks (ANN), support vector machine (SVM) [8], [9], visual graph [10], or network prediction techniques [11]. Leveraging these modern methods, they exhibit the capability to handle the time-varying nature of crude oil price and deliver encouraging results [12]. Considering the nature of model input, we categorize these state-of-the-art models into three groups:

TABLE 1. Categorization of the modern approaches.

	Model	Characteristic
Modern models	Time Series	Use only the input of the crude oil price time series for analysis.
	Structural	Incorporate additional market data as inputs for analysis.
	Signal decomposition	Use the individual components obtained after decomposing the crude oil price time series as inputs for further analysis.

sequential models, architectural models, and signal-analysis models. These models are shown in following table 1.

The sequential model primarily relies on the crude oil price temporal data as its input. For instance, in current studies, the authors Khan et al. [13] proposed a predictive model that combines the algorithm for optimizing using chicken swarms with an artificial neural network, resulting in improved performance compared to other prediction models inspired by biological neural networks. Similarly, the authors Bristone et al. [14] employed a visibility graph to represent temporal data as a grid and utilized K-core network centrality to capture the nonlinear characteristics of the crude oil price dataset. These are just a few examples of such approaches. However, due to the limited scope of input variables, the time-series model faces challenges in obtaining sufficient information for accurate predictions.

The structural model provides a solution to address the aforementioned issue. By incorporating additional influences and variables, the structural framework demonstrates excellent performance. For instance, in study [15], a decision tree model incorporating crude oil supply and demand, monthly gross domestic product, and consumer price index as factors yielded superior results. Another paper [16] highlighted the significance of news texts in crude oil price forecasting, while [17] even utilized Twitter public sentiment regarding US foreign rule as an explanatory variable. These are just a few examples showcasing the effectiveness of the structural model. However, it is crucial to note that the performance of the structural model heavily relies on the quality of the selected factors. Selecting suitable variables poses a challenge, as it necessitates specialized financial data and thorough analysis and research.

Moreover, the realm of academic research pertaining to the crude oil market is extensive, mainly because of its significance and intricate connections with various financial markets, such as gold, stocks, and exchange rates [18], [19], [20]. Amidst this vast body of literature, the predicting crude oil prices has emerged as a pressing and practical concern. Nevertheless, accurately predicting the crude oil prices are still poses integral challenges due to numerous factors come into play, including the forces of demand and supply [21], and other factor like the substantial impact of political events, the use of alternative energy sources, stock market indices, exchange rates of key the countries that import or export oil, and many other variables [22]. The combination of these factors contributes to the complexity and unpredictability of the crude oil market.

II. LITERATURE REVIEW

Oil price forecasting has enticed substantial consideration from researchers, primary to the proposal of various methods and models aimed at enhancing prediction accuracy. An alignment with this, the author in [23] conducted new method in this field by employing a concise data-driven model for oil price forecasting, however, the authors only consider the crude oil prices for Nigeria and also not considered the other factors like economic growth, financial market and dollar exchange rate. Furthermore, the authors Tang and Hammoudeh [24] utilized a nonlinear regression model specifically for forecasting oil prices within the oil-producing nations' group (OPEC) using GA along with SVR to enhance to prediction accuracy. Besides its increase in the prediction accuracy the model has some labour costs as of tuning parameters. Similarly, the authors [25] using random forest regression to predict oil futures prices during the COVID period. However, the authors only finds the impact of outbreak on the crude oil prices. In the same way, the authors [26] introduced a two- layer non-nagative matrix factorization model to achieve more precise oil prices. Still, it only find the risk of crude oil prices in oil market. The authors in [27] utilized an ANFIS model for forecasting monthly data of crude oil, though it only focus on the sentiment analysis to forecast the crude oil prices. In the same way, the authors in [28] forecasted oil prices using econometric model but it is only for the year 1986 to 1991. The authors in [29] used the target zone model to investigate the nature of oil price globally. Nevertheless, the authors only used the monthly data in the study. The authors in [30] further incorporated the ARIMA and GARCH model to improve the accuracy of oil price forecasting. However, the data was used from 1983 to 2003.

As the oil prices are unstable and non-linear in nature. Therefore, no such models can fully predict it. For this, various methods are introduced to the address this issue. One of the approaches is to decompose the whole data and then used the models that in such a way that the ensemble models are integrated at the end for the prediction. But still, the researchers are working to enhance it more and more and reduce the models complexities. The authors Zhang et al. in [31] constructed a hybrid model based on EEMD and applied AI techniques for crude oil price forecasting.

Several recent research studies have employed advanced methodologies to enhance the price of oil forecasting. The authors in [32] presented an iteration-based hybrid model that incorporated macroeconomic indicators, technical analysis signals, and demonstrated its superior performance compared to conventional models. In contrast the authors only used the data from 1980 to 1998 and only work on the three possible scenarios of the prices: linear, chaotic, and probabilistic. Similarly, the authors [33] proposed a non-iterative ensemble learning model utilizing random vector functional link (RVFL) approach for predicting oil prices, which exhibited improved forecasting accuracy when compared to support vector machine. In their study the authors only used the two methods for the prediction of crude oil prices.

The authors in [12] introduced a GARCH model for the characterization of crude oil prices, thought it used a general additive function of previous observations. Despite that, the authors only compared the parametric and non-parametric GARCH models. Gregor et al. [38] proposed a multi-layered model incorporating web search data and oil-related factors as predictors, demonstrating the superior performance of their model compared to other approaches in an empirical study. Although, the authors used the forecast combination method within range 1 to 24 months and used the mid of month within average. However, it improved in a case when the discrete wavelet transform used as an input for ANN. Additionally, [34] improved the validity of predicting oil prices by employing predictive modelling using regression with robust error metrics and regularization restrictions. The authors used the simple average method to ensemble the models with the help of different wavelet families. Despite that, the authors only used the short term prediction. In the same way, [35] proposed a hybrid model integrating wavelet de-noising, EMD, ARIMA, and fractionally integrated GARCH models to forecast crude oil data, and their model outperformed traditional methods. Moreover, leveraging modern methods like SVM, ANN and LSTM etc, they exhibit the capability to handle the non-stationary nature of crude oil price and deliver promising results [12].

The ARIMA model, vector autoregression (VAR), and GARCH have been developed as traditional statistical models for crude oil price forecasting. These models are constructed on complete statistical theories, allowing for the testing of model performance and parameters based on classical statistical principles. However, the performance of traditional statistical models may not always meet expectations due to the inherent instability and nonlinearity of crude oil price time series. Nevertheless, the study bound restricts their practical use in economic and management problems, as certain statistical hypotheses need to be satisfied, such as data stability testing.

To fulfill the need for precise and adaptable prediction models, artificial intelligence methods provide an option to conventional statistical models. Unlike their counterparts, AI models don't depend on meeting statistical hypotheses or adhering to a specific formula. As an alternative, they leverage the advanced algorithms and machine learning to analyze the data patterns and generate the accurate forecasts.

In more recent times, the adoption of Long Short-Term Memory technique for time-series modelling has gained a great attention. This is primarily attributed to its ability to facilitate end-to-end modelling, seamless integration of independent variables, and self-feature extraction capabilities. However, the authors used the forecast returns of oil prices [36]. LSTM has demonstrated remarkable success in various sequence prediction tasks, spanning domains such as voice identification [36], AI-driven translation [37], image/video classification [38], melody composition [39], and more [40], [41], [42], [43], [44], [45]. These studies highlight the effectiveness of LSTM in capturing the intricate dependencies within complex nonlinear time series systems.

The challenge is the selection and tuning of appropriate model architectures and hyper-parameters as well as may struggle with capturing long-term dependencies and understanding the underlying dynamics of time series data [46]. Hybrid models, which combine the strengths of both classical and machine learning approaches, offer several advantages over individual models in time series forecasting [47].

Since the oil prices are erratic and exhibit non-linear patterns. The effectiveness of the structural model heavily relies on the quality of the selected factors, complete prediction remains poses a challenging task and it requires an extensive research and expertise in temporal data. To tackle this, different methods have been introduced. One approach involves decomposing the overall data and utilizing models in such a manner where the ensemble models are integrated for prediction. Nonetheless, the researchers continue to strive for improvement, aiming to reduce the model complexities. To address the challenge of selecting factors amid complexity and ensuring an ample supply of input information, a signal-decomposition model is introduced. Although it solely employs the crude oil prices data as regressor, the model's actual input is multifaceted. By applying signal splitting techniques, the time series is broken down into several sub-series, each capturing different frequency horizons. Some sub-series reflect short-term fluctuations, while others depict the long-term trends inherent in the original series [48]. Various decomposition methods can be employed to categorize the signal-decomposition model, including the wavelet method [49], empirical mode decomposition [26], ensemble empirical mode decomposition [26], [50], seasonal adjustment methods [51] and many various decomposition methods [48], [52]. The authors in [53] used the type of autocorrelation integrated with Q-learning Swarm Optimization for finding the possible interaction amid of irrigation points with the credentials of high-impact irrigation zones. Furthermore, for taking the oscillation between dry and rainy seasons, the intrinsic mode decomposition is employed with zero-crossing method. Then the authors used the GRU approach for the extraction of IMFs information. However, in his approach the authors used the irrigation data. Moreover, the authors in [54] developed the EMD in combination with spectral analysis for the decomposition of suspended sediment concentrations temporal data. The authors also incorporates the Incosh function in the ridge regression for handling the outliers in the data. To handle the outliers in the data, the authors [55] developed the adoptive trimmed mean. In their studies, the proposed method is better than the stable regulation parameters counterparts. Moreover, the authors [56] developed the robust adaptive rescaled lncosh neural network regression model for the time series data that handles the outliers and noise problems. However, the wind speed data is used.

The EEMD is proposed to address the issue of scale separation without relying on an intermittency perception test. This new method, known as NAD: noise-assisted data analysis, offers a solution to the problem at hand. It has many advantages like robustness, adaptive nature, ensemble techniques, preservation of signal features, compared to other methods for IMF reconstruction [57].

The main goal of this research is to forecast the Brent crude oil price using a hybrid model that considers the specific properties of the oil price data. To achieve this, the Brent oil price data is decomposed, resulting in separate subseries with unique nonlinear and volatile features. The proposed model employs various models to predict each subseries based on its individual characteristics. By combining these forecasts, more accurate predictions are obtained compared to other approaches. This methodology ensures a robust and precise forecasting outcome.

While AI and machine learning models show the significant promise in time series forecasting, they also encounter challenges. A primary issue is the need for extensive amounts of high-quality training data. Time series data typically display intricate patterns, trends, and seasonality, demanding large and representative datasets for successful model training. We, also know that the individual model can't be efficient in order to boost the prediction accuracy as shown in the results and in the existence literature. In alignment with the existing body of literature, the main objective of this study is to predict the Brent crude oil data using a hybrid model that effectively captures the unique characteristics of the oil price time series data and to enhance the prediction accuracy. To achieve this, the Brent oil prices data is decomposed, resulting in separate subseries with unique chaotic and unstable features. Then, the trend is eliminated, allowing the scrutiny of whether the residual component validates the assurance of a series ran by stochastic trends. Following the removal of the trend, the residual component undergoes rigorous evaluation through autoregressive model following the decomposition model. The proposed model adopts various techniques tailored to forecast each subseries based on its inherent nature. This methodology ensures a robust and precise forecasting outcome, as the linear combination of these forecasts produces outcomes with greater accuracy when compared to alternative methods.

III. METHODOLOGY

In the upcoming section, we will delve into the two methods employed in the proposed model. Firstly, in Section III, a brief introduction to EEMD will be presented and the reconstruction of IMFs. Following that, Section III-A, III-B and III-C will cover the fundamental concepts of the ARIMA, SVM and LSTM network. All the work are carried out in Python and R version 3.11.2 and 4.4.3 respectively.

The EEMD which stands for Ensemble Empirical Mode Decomposition, is a powerful algorithm for partitioning data that operates on the time series. This method is an extension of the EMD introduced by the authors in [57]. The EEMD is a flexible technique that allows for the decomposition of a data into multiple Intrinsic Mode Functions (IMFs). If the data $D(t)$ is decomposed into $s(t)$ the true signal and $n(t)$ the noise, then it can be represented in (1).

$$D(t) = s(t) + n(t) \tag{1}$$

The authors in [58] an effort to mitigate the expansion of moving the low frequency mode into the disturbed region, Huang et al. introduced a small amount of noise to earthquake data. However, they did not fully consider the consequences of this noise addition in the EMD method. To address a particular challenge in the original EMD method, the authors Flandrin et al. in [59] incorporated additional noise as a solution. Although they defined the true answer as the mean breakdown of the Dirac function,

$$x[n] = lim_{\varepsilon\to 0^+} E\{x[n] + \varepsilon r\delta[n]\} \tag{2}$$

The notation $[n]$ denotes the *nth* data point, while $x[n]$ represents the Dirac function. Additionally, , $\varepsilon r\delta[n]$ corresponds to a random number , ε symbolizes the infinitesimal parameter, and $E\{\}$ represents the expected value.

The author Gledhill [60] introduced a unique approach in data analysis by incorporating noise to assess the resilience of the EMD algorithm. Drawing from his findings regarding the impact of noise on the EMD algorithm, the author's assumed that the clean data, unaffected by noise, provided the true reference answer. Consequently, he introduced the discrepancy, Δ, to quantify the differences between the outcomes produced by the EMD algorithm and the reference data.

$$\Delta = \sum\nolimits_{k=1}^{m} \left(\sum\nolimits_{t} (Ir_k(t) - kn_k(t))^2\right)^{\frac{1}{2}} \tag{3}$$

In his wide-ranging investigation into the intricate distribution of the "discrepancy" induced by noise, he considered *Irk* and *Ink* as the *kth* components of the IMF obtained from the data without and with the addition of noise, respectively. The variable m denotes the sum of total of IMFs obtained from the data. Through this thorough analysis, Gledhill aimed to gain a comprehensive understanding of how the noise-induced deviations manifest within the dataset. After careful examination, he reached the conclusion that the EMD algorithm exhibits a satisfactory level of stability when subjected to minor disturbances or perturbations. Furthermore, in the EMD methodology, the dataset $D(t)$ is decomposed into intrinsic mode functions (IMFs), represented as I_k, which capture the underlying components or patterns within the data.

$$D(t) = \sum\nolimits_{k=1}^{m} I_k + r_n \tag{4}$$

The residue of the data $D(t)$ after extracting n number of IMFs is denoted as rn. It represents the remaining component that cannot be further decomposed into additional IMFs.

As indicated in Eq. 1, the data consists of a combination of signal and noise components. In order to enhance the data accuracy, employing the ensemble mean method proves to be effective. This technique involves gathering data from multiple independent observations, each of which incorporates distinct sources of noise.

In order to simulate the random noise that can occur during the measurement process, white noise is intentionally added. This additional white noise is considered representative of the potential random noise encountered during measurements. In this context, each "artificial" observation, denoted by the index i, is created under these conditions.

$$D_i(t) = D(t) + w_i(t) \tag{5}$$

In scenarios where only one observation is present, the generation of multiple-observation ensembles involves incorporating diverse instances of white noise, denoted as $wi(t)$, into the single observation. This approach, outlined in (5), aims to replicate the effect of having separate observations with varying sources of random noise.

The step of EEMD are as follows:

1. Introduce a series of white noise to the target data.
2. Perform a decomposition of the data, including the added white noise, into individual Intrinsic Mode Functions (IMFs).
3. Repeat the process by incorporating various white noise series each time and decomposing the data into IMFs again.
4. Calculate the average (ensemble means) of the associated IMFs obtained from the decompositions as the final outcome.

According to a well-established statistical principle, the influence of the added white noise is expected to diminish.

$$\alpha_n = \frac{\alpha}{\sqrt{n}} \tag{6}$$

$$In\alpha_n + \frac{\alpha}{2} In(n) = 0 \tag{7}$$

In this context, n represents the count of ensemble components, α denotes the magnitude of the added noise, and α_n signifies the ultimate standard deviation of the error, described as the disparity pertaining to both the input signal and the associated IMF(s).

The ultimate truth determined by EEMD is attained as the ensemble size approaches infinity, implying that the number becomes infinitely large.

$$I_k(t) = \frac{1}{n} \sum_{j=1}^{n} \{I_k(t) + \omega r_j(t)\} \tag{8}$$

$$I_k(t) + \omega r_j(t) \tag{9}$$

In the equation, the jth trial of the kth IMF represents the noisy signal, and the magnitude of the integrated noise, ω is not necessarily small. However, it is crucial to have a large number of trials in the ensemble, denoted by n. The disparity between the true value and the ensemble result follows a widely recognized statistical principle: it diminishes proportionally to the inverse square root of n, as illustrated in (6).

Given the established definition of the truth, the discrepancy (Δ) should be considered instead of the one specified in (3). This revised measure accounts for the discrepancy between the truth and the obtained results, incorporating the necessary adjustments to ensure accuracy and reliability.

$$\Delta = \sum_{k=1}^{m} \left(\sum_{t} E\{(Ir_k(t)\} - kn_k(t))^2 \right)^{\frac{1}{2}} \tag{10}$$

where E{} denotes the expected value, as expressed in (8). This statistical concept encapsulates the anticipated outcome or average result, providing a measure of central tendency in relation to the given equation.

In this section, we propose a new method of forecasting approach that combines IMFs reconstruction with an optimized combined forecasting model. The aim is to improve the accuracy of the forecasting process. This approach involves reconstructing the IMFs and utilizing an optimal combined forecasting model.

To improve the precision of crude oil price forecasting, an approach which combines IMFs reconstruction, and an optimized combined forecasting model is introduced. The goal is to address the issue of certain decomposed IMFs having similar impacts on the original series in terms of trend and accuracy. To tackle this, the data is portioned using the difference method. Once the trend is removed, then the objective is to scrutinize whether the residual component shows the characteristics of a series with a stochastic trend and the residual component is subjected to estimate using the autoregressive model and then examine the significance of the coefficient being estimated [61] method is utilized to reconstruct the IMFs based on their comprehensive contribution index. The series is represented as the combination of a non-random trend, random walk, and stationary variation components as given in the following in (11).

$$y_t = \tau_t + \vartheta_t + \epsilon_t \tag{11}$$

The LM test is conducted to examine the hypothesis that the random walk component has zero variance. The asymptotic distribution of the test statistic is derived under both the null hypothesis and the alternative hypothesis, which assumes the series is difference stationary. The overall framework of this novel forecasting approach is depicted in figure 2.

A. ARIMA MODEL

Integrated Moving Average, known as Autoregressive Integrated Moving Average [62], is formed by integration the autoregressive (AR) model with the moving average (MA) model after differencing the time series data. The necessary equations are written below.

$$y_t = C + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + e_t + \varepsilon_1 e_{t-1} + \varepsilon_2 e_{t-2} + \ldots + \varepsilon_q e_{t-q} \tag{12}$$

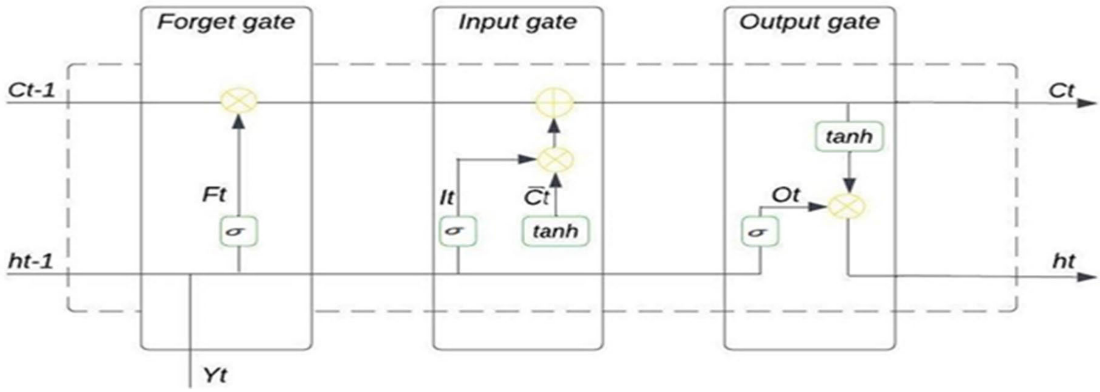


FIGURE 1. Structure of LSTM model.

The AR component of the ARIMA model indicates that the time series is regressed on its own past data, capturing the relationship between current and previous values. The MA component indicates that the forecast error is a linear combination of past errors, considering the residual errors in the model. The component signifies that differencing has been applied to the data to achieve stationarity, a prerequisite for the ARIMA model.

The above in (12) represents the predictors used in the ARIMA (p,d,q) model, where the autoregressive (AR) part consists of lagged p data points and the moving average (MA) part includes lagged q errors, all of were modified by differenced. The forecast is based on the differenced value of *yt* in the *dth* order. In the ARIMA model, the coefficients β and ε are estimated through methods like maximum likelihood estimation, as the model develops an understanding from the training data of oil prices. Selecting the appropriate values for p, d, and q can be challenging, but by trying out different combinations and evaluating the model's performance, one can determine the optimal configuration. In the current study, an auto ARIMA is used using 80% training and 20% testing parts of the daily crude oil prices.

B. SVM MODEL

The support vector machine was first presented by Vapnik [63] as a machine learning model that works on the minimum structural risk criterion. It excels at modelling nonlinear relationships, effectively mitigating the complexities associated with solving problems in high-dimensional spaces. This model has extended and further helped in a new angle on nonlinear combination forecasting, important to its extensive adoption by researchers in various disciplines for addressing predictive challenges. To demonstrate, SVMs have originate utility in fields beyond prediction, such as evaluating the performance of spark ignition systems [64] in areas like automotive engineering and combustion research and identification of parameters [65]. In the same way, the authors Qi and Zhang [66] introduced SVM for crude oil prices and in the same the authors in [67] used SVM with optimization techniques for the six months of crude oil prices data which gave better results from convolutional models. The author in Vapnik [63] employ ε-Robust loss function for the purpose of regression analysis and temporal forecasting.

$$C(f(x)-y)=\begin{cases}|f(x)-y|-\varepsilon & for \quad |f(x)-y|\geq\varepsilon\\ 0 & otherwise\end{cases} \tag{13}$$

C. LSTM MODEL

The Long Short-Term Memory [68] (LSTM) neural network is a modified version of the Recurrent Neural Network (RNN). LSTM incorporates three gates, namely the entry gate, memory gate, and output gate, into its unit. These gates play a crucial role in updating the data stored in the memory unit. By doing so, LSTM effectively manages the balance between retaining and discarding historical data. Additionally, LSTM resolves the challenges of gradient vanishing and exploding, which are common issues in traditional RNNs. When upgrading the cell state of the LSTM unit, each gate exerts control in the following manner:

The LSTM neural network incorporates three gates within its unit to control the flow of information:

i Input gate: This gate selectively determines the data stored in the cell state utilizing the current input and the earlier hidden state.

ii Forget gate: The forget gate selectively determines which information from the previous cell state should be discarded, considering the current data input and the earlier hidden state.

iii Output gate: Using the current input and the updated cell state, the output gate conditionally determines the information to be output from the LSTM unit.

Taking advantage of these three gates, the LSTM model effectively manages the information stored in its memory unit. By utilizing the input gate, LSTM can selectively determine which past data to retain and flexibly modify the cell state with new information. This capability allows LSTM to forget irrelevant information and retain relevant information, thereby enhancing its ability to capture long-term interdependencies in the data. Through the learning process, LSTM receives input information and utilizes its unique structure to process and store the data. The architecture of an LSTM

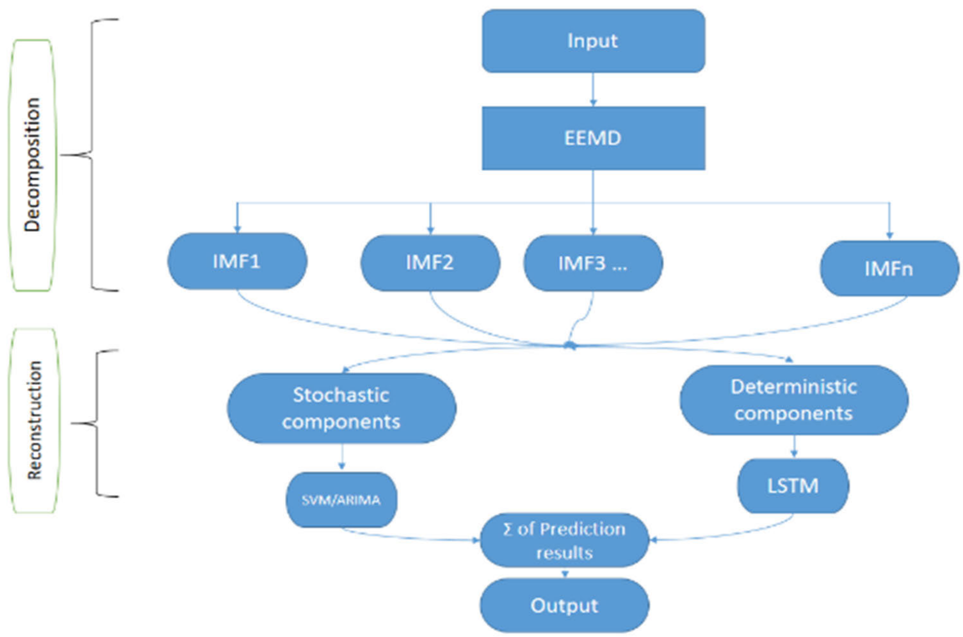


FIGURE 2. **Flowchart of the proposed method.**

segment is depicted in Figure 1. The calculation method of an LSTM unit can be summarized as follows:

1. At each time step t, the values of the candidate memory cell C_t, input gate I_t, and forget gate F_t are computed using the following equations:

$$\tilde{C}_{t\,t} = \tanh\left(w_c\left[h_{t-1}, y_t\right] + B_c\right) \quad (14)$$
$$I_t = \delta\left(w_i\left[h_{t-1}, y_t\right] + B_i\right) \quad (15)$$
$$f_t = \delta\left(w_f\left[h_{t-1}, y_t\right] + B_f\right) \quad (16)$$

where w_c, w_i and w_f denote the resultant weight matrices, B_c, B_i and B_i denote the resultant biases, $ht - 1$ denotes the output value of the LSTM unit at the previous time step, the input value y_t at time t, where tanh denotes the hyperbolic tangent activation function within interval (-1, 1), and δ is the sigmoid activation function within the range (0, 1).

2. The value of the memory cell C_t at time t is calculated using the following equation:

$$C_t = f_t C_{t-1} + I_t \tilde{C}_t \quad (17)$$

The value of the memory cell $_t$ at time t, which depends on the last memory cell state $_{t-1}$, is calculated using the weight matrix w_0 and bias B_0 of the output gate:

3. The value of the output gate denoted as O_t and the corresponding output value of the LSTM unit h_t at time t are calculated using the following formulas:

$$o_t = \sigma\left(w_0\left[h_{t-1}, y_t\right] + B_0\right) \quad (18)$$
$$h_t = o_t \tanh\left(C_t\right) \quad (19)$$

By incorporating the architecture of three control gates and a memory cell, LSTM is designed to effectively capture and update long-term information, enabling it to learn and model the long-term dependencies in time series data.

TABLE 2. **Descriptive statistics of daily crude oil prices.**

Description	Values
Minimum	9.12
1st Quarter	55.24
Median	73.26
Mean	77.34
3rd Quarter	105.99
Maximum	133.18

This capability allows LSTM to retain and utilize crucial information over extended periods, enhancing its performance in tasks involving long-term dependencies. In this study, we used 64 hidden units in the first LSTM layers followed by 32 hidden units in the second LSTM layers. Furthermore, there are 16 hidden units in third LSTM layers with 0.2 dropout and finally the single unit is in the dense layer for the output. In the model, the number of epochs were 50, with min_delta 1e^-04 with patience set to 20. We have used early stopping function to avoid over and under-fitting during the training process. The general structure of LSTM is shown in the following figure 1.

To increase the validity of crude oil price data forecasting, a method which pools IMFs rebuild, and an enhanced combine forecasting model is introduced. The objective is to address the problem of certain decomposed IMFs having

FIGURE 3. The Brent daily crude oil prices.

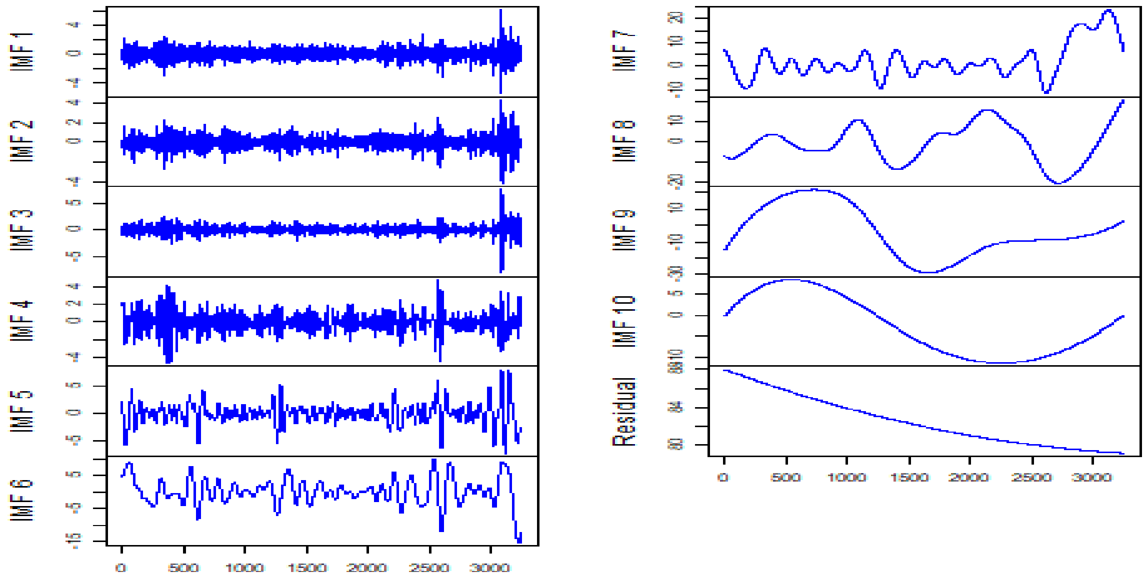


FIGURE 4. Number of IMFs using EEMD.

TABLE 3. Forecasting evaluation criterions.

Criterion	Formula
Mean square error	$MSE = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \widetilde{Y}_i)^2$
Mean absolute error	$MAE = \frac{1}{m}\sum_{i=1}^{m}\lvert Y_i - \widetilde{Y}_i\rvert$
Mean absolute percentage error	$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left\lvert\frac{Y_i-\widetilde{Y}_i}{Y_i}\right\rvert * 100$
Mean Absolute Scaled error	$MASE = \frac{\frac{1}{m}\sum_{i=1}^{m}(\lvert Y_i-\widetilde{Y}_i\rvert)}{\frac{1}{m-1}\sum_{i=1}^{m}(\lvert Y_i-\widetilde{Y}_i-1\rvert)}$
Median percentage error	$MDPE = \text{median}\frac{(Y_i-\widetilde{Y}_i)}{Y_i} * 100$
Standard deviation percentage error	$SDPE = \frac{1}{m}\sqrt{\frac{[\sum(\lvert Y_i-\widetilde{Y}_i\rvert)]}{Y_i - mean\ PE)^2}} * 100$
Root mean squared scale error	$RMSSE = \sqrt{\frac{(Y_i-\widetilde{Y}_i)^2}{\frac{(Y_i-Y_t)^2}{m}}}$

similar effects on the original series in terms of trend and accuracy. The trend is eliminated, allowing the scrutiny of whether the residual component validates the assurance of a series ran by stochastic trends. Following the removal of the trend, the residual component undergoes rigorous evaluation through autoregressive model following the decomposition model. The following steps are used in the proposed method.

Step 1: Decomposition of data

TABLE 4. (a). Performance metrics of the models. (b). Performance metrics of the models.

(a). Performance metrics of the models

Model	MSE	MAE	MAPE	MASE	MDPE
Proposed method	**1.13**	**0.874**	**1.37**	**0.572**	-0.990
No-change Forecast	5.13	1.52	2.08	0.998	**-0.760**
LSTM	8.05	1.97	2.64	1.29	-0.993
ARIMA	3469.1	57.92	97.40	37.90	82.69
LSTM (Individual IMFs)	59.49	6.15	9.03	4.03	-7.49
ARIMA (Individual IMFs)	1892.47	36.02	74.21	23.61	46.00

(b). Performance metrics of the models

Model	MPE	SDPE	RMSSE	SMAPE	WS
Proposed method	-0.736	**1.55**	**0.470**	**1.37**	**44.04**
No-change Forecast	**-0.443**	2.88	0.999	1.53	74.05
LSTM	-0.725	3.40	1.25	2.65	85.05
ARIMA	97.40	62.45	25.96	61.17	1957.49
LSTM (Individual IMFs)	-7.85	7.89	3.39	9.67	309.48
ARIMA (Individual IMFs)	40.98	84.27	19.17	44.45	1422.63

TABLE 5. Directional forecast values of the models.

Models	Directional Forecast Accuracy (%)
Proposed method	76.41
LSTM	53.44
ARIMA	58.80
LSTM (Individual IMFs)	55.30
ARIMA (Individual IMFs)	61.08
No-change forecast	74.39

In the first step, the primary data D_t was decomposed into k IMFs using EEMD.

Step 2: Reconstructions of IMFs

The k number of IMFs are further reconstructed into stochastic and deterministic components. For this, the data is portioned using the difference method. Once the trend is removed, then the objective is to scrutinize whether the residual component shows the characteristics of a series with a stochastic trend and the residual component is subjected to estimate using the autoregressive model and then examine the significance of the coefficient being estimated.

Step 3: Ensemble forecast outcomes

In the third step, the forecasted results which are obtained from stochastic and deterministic components are aggregated using simple addition method.

$$PD = \sum\nolimits_{j=1}^{k} St + \sum\nolimits_{m=1}^{n} Dt \quad (20)$$

where PD is the predicted data obtained from stochastic St and deterministic Dt components respectively.

Step 4: Final output

The final output is then compared with using ARIMA, LSTM and No-change-forecast method.

The EEMD method aims to break down the initial series into IMFs and a residue, where each IMF operates that fulfils two criteria:

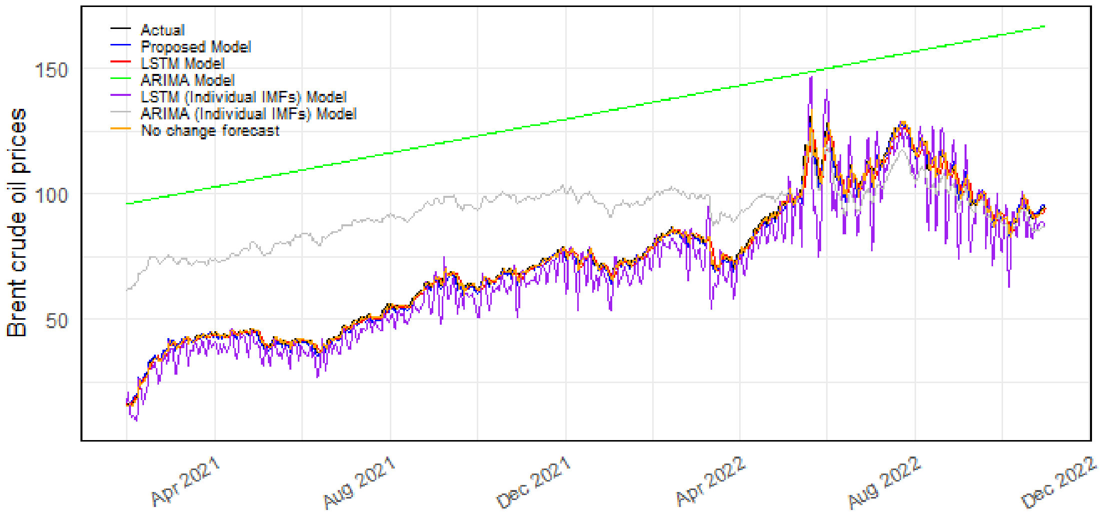
i. In the complete dataset, the count of zero crossings and extreme crossings should be either the same or show a difference by a maximum of one.
ii. At any given point, the mean value of the wrapper formed by the local maxima and local minima is zero.

The impact of the added white noise can be computed as given in (7). The pseudo code of the Algorithm is shown below.

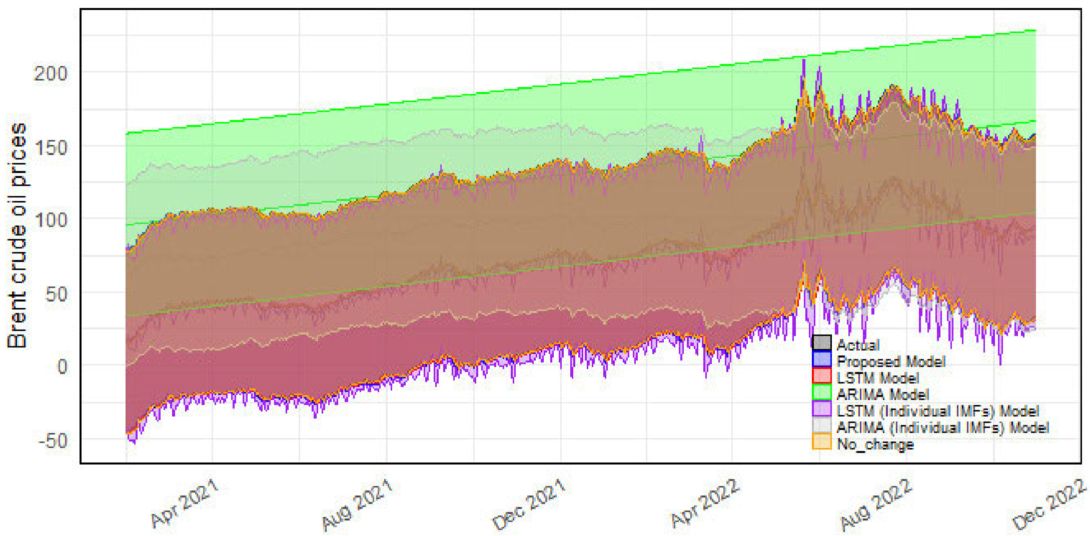
In this study, we have taken daily prices of crude oil data from Jan-04-2010 to Oct-31-2022. There are total 3248 daily prices of crude oil series. Dividing the data into training and testing, we followed the 80/20 rule, with 2598 prices allocated for training and 650 prices for testing which is shown in the figure 3.

TABLE 6. Results of the Diebold-Mariano (DM).

Models	DM values	P-values
Proposed with LSTM	2.253	<0.02
Proposed with ARIMA	-16.693	<0.01
Proposed with LSTM(Individual IMFs)	8.012	<0.01
Proposed with ARIMA (Individual IMFs)	-16.056	<0.01
Proposed with No-change forecast	5.058	<0.01
LSTM with ARIMA	-4.7143	<0.01
LSTM with LSTM (Individual IMFs)	-8.0041	<0.01
LSTM with ARIMA (Individual IMFs)	-6.6975	<0.01
ARIMA with LSTM (Individual IMFs)	-72.737	<0.01
ARIMA with ARIMA (Individual IMFs)	7.0972	<0.01
No-change forecast with LSTM(whole IMFs)	4.00	<0.01
No-change forecast with ARIMA(whole IMFs)	112.30	<0.01
No-change forecast with LSTM(Individual IMFs)	16.59	<0.01
No-change forecast with ARIMA(Individual IMFs)	-35.88	<0.01



(a) Actual and predicted values of the models



(b) Actual and predicted values with 95% CI of the models

FIGURE 5. **(a). Actual and predicted values of the models. (b). Actual and predicted values with 95% CI of the models.**

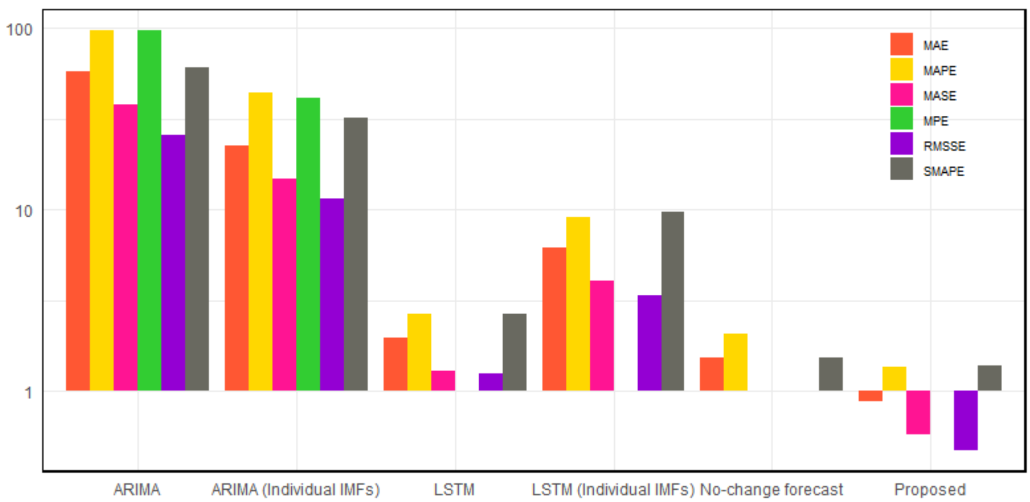


FIGURE 6. **Multiple bar charts of the models using performance metrics.**

Algorithm for the Proposed of Reconstruction of IMFs

initialize $\boldsymbol{Dt}$ ($\boldsymbol{t} = 1, 2, 3, \ldots, \boldsymbol{T}$)
$\boldsymbol{Dit} \leftarrow \boldsymbol{IMF1}, \boldsymbol{IMF2}, \ldots, \boldsymbol{IMFk}$ by using the equation 1
for k = 1 to m do by equation 12
if is_$\boldsymbol{St}$ do
    $\boldsymbol{St}$.append(IMF)
else
    $\boldsymbol{Dt}$.append(IMF)
end if
end for
procedure for $\boldsymbol{PD}$ given in equation 19
for j = 1 to k & m = 1 to n
if $\boldsymbol{St}$ in St(i) do
    $\boldsymbol{ft_1}$ = apply_arima_model($\boldsymbol{St}$)
    arima_ft_1.append($\boldsymbol{ft_1}$)
else
    $\boldsymbol{Dt}$ in $\boldsymbol{Dt}$(i)
    $\boldsymbol{ft_2}$ = apply_lstm_model($\boldsymbol{Dt}$)
    lstm_ft_2.append(ft_2)
end if
end for

The Brent daily data are splits into various IMFs using the EEMD. The number of IMFs are shown in the figure 4.

For the assessment of the current study, we have used different performance metrics as shown in the Table 3. In the same way, we have also used the Diebold-Mariano (DM) test that the test the similarity between the two predictions under the null hypothesis $H\_0 : E(d) = 0$. The results indicate that there is difference d among all the methods in prediction as of P-values.

IV. RESULT AND DISCUSSION

In table 4 (a) and (b), we have compared the proposed method with original data and the individual IMFs using ARIMA, LSTM and No-change-forecast. The results are shown in the following Table 4(a). The value of mean square error for the proposed method, No-change forecast, LSTM, ARIMA, LSTM (Individual IMFs) and ARIMA (Individual IMFs) is 1.13, 5.13, 8.05, 3459.1, 59.49 and 1892.47 respectively. In the same way, the mean absolute error (MAE), mean percentage error, mean absolute percentage error (MAPE) and mean absolute scaled error (MASE) of the proposed method are smaller than the other methods. The median percentage error (MDPE) is a robust statistic that measured the deviation within the predicted and actual observations, thus the MDPE values of the No-change forecast is better than the other methods which indicates that it is more robust to the outliers observations. In addition to, the mean percentage error of No-change forecast is better than the rest of the methods. In the same way, the standard deviation of percentage error (SDPE) calculates the dispersion or variation of percentage errors prediction of time series models. The SDPE value of the proposed method, No-change forecast, LSTM, ARIMA, LSTM (Individual IMFs) and ARIMA (Individual IMFs) is

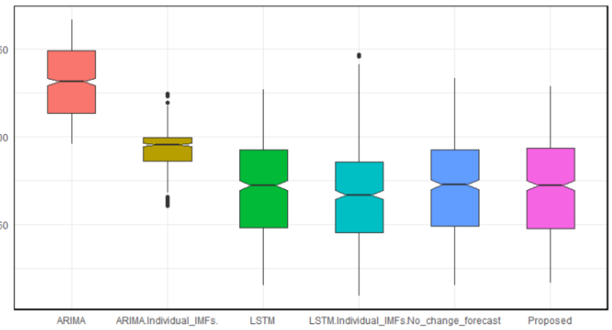


FIGURE 7. Boxplots of the models.

1.55, 2.88, 3.40, 62.45, 7.89 and 84.27 respectively. The value of SDPE shows that the proposed method is more consistent and stable in-terms of performance. Moreover, root mean squared scaled error (RMSSE) of the proposed method, No- change forecast, LSTM and ARIMA is 0.470, 0.999, 1.25 and 25.96 while for LSTM (Individual IMFs) and ARIMA (Individual IMFs) is 3.39 and 19.17 respectively. The value of RMSSE indicates that the proposed method have better performance than the other methods. In additions to the values of symmetric mean absolute percentage error (SMAPE) of the proposed method and No-change forecast is 1.37 and 1.53 whereas of LSTM, ARIMA, LSTM (Individual IMFs) and ARIMA (Individual IMFs) is 2.65, 61.17, 9.67 and 44.45 respectively. The SMAPE equally weights to the underestimation and overestimation of the predicted model and the value proved that the proposed technique prevails over the other models. The WINKLER SCORE (WS) measures both the absolute and relative error of between actual and predicted values. The WS value of the proposed method is 44.04, No-change forecast is 74.05, LSTM is 85.05, ARIMA is 1957.49, while for LSTM (Individual IMFs) and ARIMA (Individual IMFs) is 309.48 and 1422.63 respectively. The WS of the proposed is smaller than the other methods and it indicates that the proposed method is better than the other models in terms of forecast accuracy.

In the following Table 5, we find the directional forecast (DF) accuracy for the models. The DF of the proposed, LSTM and ARIMA models is 76.41, 53.44 and 58.80 while of LSTM (Individual IMFs), ARIMA (Individual IMFs) and No-change forecast is 55.30, 61.08 and 74.39 respectively. The DF values revealed that the proposed method is better than the other models in terms of assessing the ability of a forecasting.

We have find the Diebold-Mariano (DM) tests and it assume that there is no significant difference between the forecasting accuracy of the models in the null hypothesis. In the Table 6 below, the DM values shows that there is difference among forecasting accuracy of the proposed, LSTM, LSTM, ARIMA, LSTM (Individual IMFs) and ARIMA (Individual IMFs) respectively.

The forecasted outcomes of the proposed model in contrast to the factual values and the other methods are plotted in the

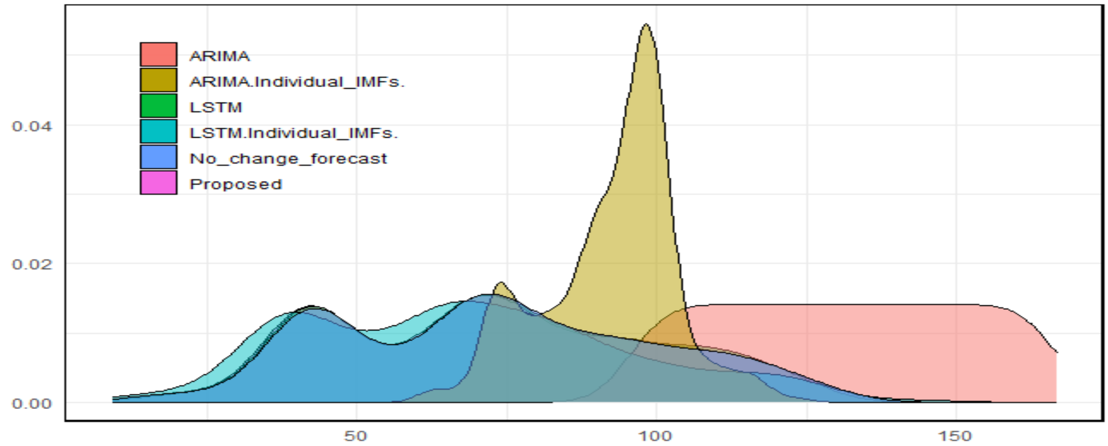


FIGURE 8. Density plots of the predicted models.

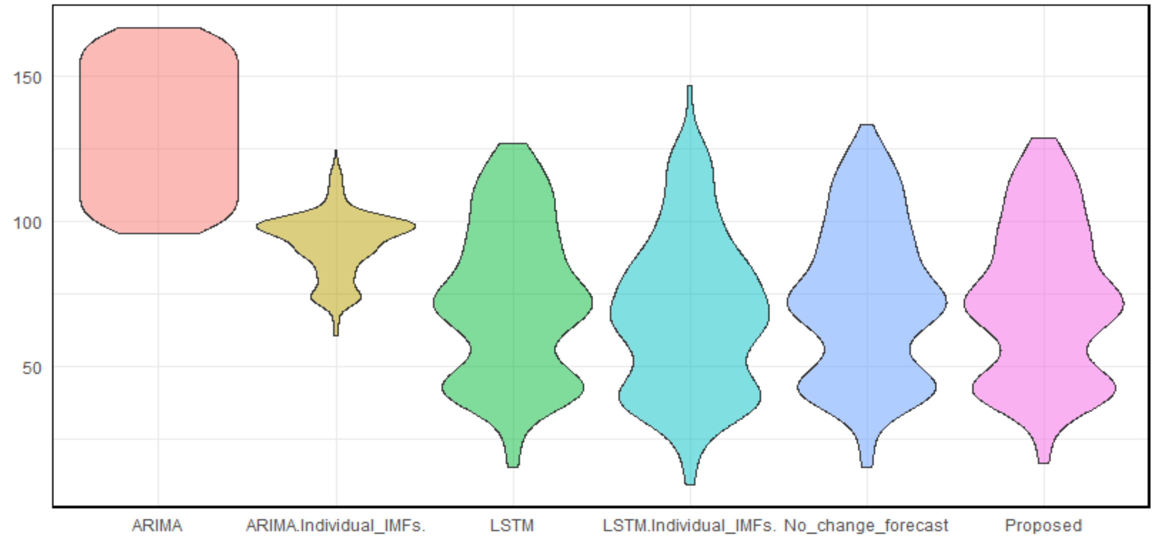


FIGURE 9. Violin plots of the predicted models.

following figure 5 (a). Moreover, we've constructed the 95% confidence interval for each of the predicted models as shown in the figure 5 (b). The ARIMA model predicted the values with constant increase across time. In the same way, ARIMA (Individual IMFs) first away from the original line and then followed the trend of the actual values. There are various peaks and hikes in the LSTM (Individual IMFs) across time with the original values. Moreover, LSTM model followed the pattern of the original values with fluctuations, while the proposed model followed the trend of the actual values.

The figure 6 presents the multiple bar charts of the models. We take Log for on the metrics values for better comparison as some of values of ARIMA model were very unusual with other models. The bar of the proposed model shows better performance among the models following the no-change forecast method. The poor efficiency is of ARIMA model, this is because of probably due the taking of the differences of data to become it stationary.

To find the overall performance of the models, box plot is used. The figure 7 shows the boxplots for each model. In the following figure, the proposed model shows more stability than the other models. Then the no-change forecast method show the stability as compared to the other methods. The LSTM (Individual IMFs) and ARIMA (Individual IMFs) models shows the outliers in the predicted values while the rest of the models including the proposed model didn't show any unusual observations in the predicted values. The ARIMA (Individual IMFs) indicates that it negatively skewed. In the same way, the LSTM, LSTM (Individual IMFs) is to some extent skewed negatively while the proposed model is less skewed negatively as compared to other models except ARIMA model.

The figure 8 shows the Density or kernel density plots. The density plot exploring and provide insights into data characteristics and model performance as well as capture the complex data patterns. The figures of Density plots indicates that ARIMA is more widely distributed. There is highly peak in the ARIMA (Individual IMFs) model and less dense due to the constant increase in the predicted values. The proposed method is equally spread on the predicted data as compared to the other models which shows that it align with the actual data following the no- change forecast method.

In Figure 9, we plotted the Violin plots. It combines both density plots and whisker plots and gives a comprehensive view of data and its nature. The Violin plots are different of each predicted model. All the models are shows bimodal except ARIMA model. The proposed model is less spread as compared to the other models. Following the proposed method, the no-change forecast, and LSTM models show less spread among the other methods.

V. CONCLUSION

Oil price complexity level might impart to the point that the parallel time series contains nonlinear behaviour, shifts in behaviour, and temporal lag. In this paper, we proposes a hybrid model for predicting daily crude oil prices, with an emphasis on apprehending these features to increase the accuracy of forecasts. To achieve this objective, we instigates by decomposing the temporal data into various components using the ensemble empirical mode decomposition. Next, the authors employs various forecasting models tailored to the characteristics of each component to attain accurate predictions.

Furthermore, various performance metrics are conducted for comparison including directional forecast, Diebold-Mariano test, and plots i.e., multiple bar chart, boxplots, density plots and Violin plots for visualizations. On average, the results reveal that solely the suggested hybrid model yields high-quality forecast values.

VI. LIMITATIONS AND FUTURE WORK

Indeed, putting forth a specific model that claims accurate forecasting is a highly challenging endeavor. However, such forecasting efforts are comparable to attempts to shape a future framework, enabling us to formulate effective strategies as necessary based on the prevailing circumstances.

In the current study, we only studied the Brent crude oil data. There can be other financial factors added like dollar exchange, demand, and supply etc. Furthermore, some other temporal data can be used.

The proposed hybrid model might conceivably be enhanced by allowing for the following aspects.

The possibility of decomposition of data with some other approaches. After the decomposition and for the reconstruction of the data, some other method can be employed to enhance the prediction accuracy as compared to the trend is eliminated, the residual component is evaluated through autoregressive model.

The other possibility is of decomposition of data with EEMD and exploring the incorporation of the SVM/ARIMA and LSTM with alternative techniques like the Gated Recurrent Unit (GRUs) and Transformers like Temporal Convolutional Network or genetic algorithm can be considered. Moreover, it is dynamic to test the proposed model for forecasting other energy commodities to assess its stability and applicability. We intends to tackle these aspects in the near future.

VII. LIST OF ABBREVIATION

ARIMA	Autoregressive Integrated Moving Average
DM	Diebold-Mariano
EMD	Empirical Mode Decomposition
EEMD	Ensemble Empirical Mode Decomposition
GRU	Gated Recurrent Unit
IMF	Intrinsic Mode Function
LSTM	Long-short term memory
MSE	Mean Square Error
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MASE	Mean Absolute Scaled error
MDPE	Median percentage error
SDPE	Standard deviation percentage error
RMSSE	Root mean squared scale error
SVM	Support Vector Machine

AUTHORS' CONTRIBUTION

Muhammad Naeem: Conceptualization, Methodology, Software, Writing-original draft, Visualization, and Data creation. Muhammad Aamir: Methodology, Validation, Writing-review & editing, and Supervision. Jian Yu: Writing-review, Visualization, and Supervision. Olayan Albalawi: Visualization, Data creation, Methodology, and Software.

REFERENCES

[1] R. Hagen, "How is the international price of a particular crude determined?" *OPEC Rev.*, vol. 18, no. 1, pp. 127–135, Mar. 1994.

[2] L. Yu, J. Li, L. Tang, and S. Wang, "Linear and nonlinear Granger causality investigation between carbon market and crude oil market: A multi-scale approach," *Energy Econ.*, vol. 51, pp. 300–311, Sep. 2015.

[3] J. Li, S. Zhu, and Q. Wu, "Monthly crude oil spot price forecasting using variational mode decomposition," *Energy Econ.*, vol. 83, pp. 240–253, Sep. 2019.

[4] N. B. Behmiri and J. R. P. Manso, "Crude oil price forecasting techniques: A comprehensive review of literature," 2013. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Crude+Oil+Price+Forecasting+Techniques%3A+A+Comprehensive+Review+of+Literature+7+Jun+2013+Niaz+Bashiri+Behmiri&btnG=

[5] Y. Xiang and X. H. Zhuang, "Application of ARIMA model in short-term prediction of international crude oil price," *Adv. Mater. Res.*, vol. 798, pp. 979–982, Dec. 2013.

[6] C. W. Cheong, "Modeling and forecasting crude oil markets using ARCH-type models," *Energy Policy*, vol. 37, no. 6, pp. 2346–2355, Jun. 2009.

[7] A. Hou and S. Suardi, "A nonparametric GARCH model of crude oil price return volatility," *Energy Econ.*, vol. 34, no. 2, pp. 618–626, Mar. 2012.

[8] L. Yu, X. Zhang, and S. Wang, "Assessing potentiality of support vector machine method in crude oil price forecasting," *EURASIA J. Math., Sci. Technol. Educ.*, vol. 13, no. 12, pp. 7893–7904, Nov. 2017.

[9] W. Xie, L. Yu, S. Xu, and S. Wang, "A new method for crude oil price forecasting based on support vector machines," in *Proc. 6th Int. Conf. ICCS*, Reading, U.K. Berlin, Germany: Springer, May 2006, pp. 444–451. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=W.+Xie%2C+L.+Yu%2C+S.+Xu%2C+and+S.+Wang%2C+%E2%80%98%E2%80%98A+new+method+for+crude+oil+935+price+forecasting+based+on+support+vector+machines%2C%E2%80%99%E2%80%99+in+Computational+936+Science%E2%80%94ICCS%2C+Reading%2C+U.K.%2C+2006%2C+pp.+444%E2%80%93451.&btnG=

[10] J. Zhao, H. Mo, and Y. Deng, "An efficient network method for time series forecasting based on the DC algorithm and visibility relation," *IEEE Access*, vol. 8, pp. 7598–7608, 2020.

[11] F. Liu and Y. Deng, "A fast algorithm for network forecasting time series," *IEEE Access*, vol. 7, pp. 102554–102560, 2019.

[12] R. Bisoi, P. K. Dash, and S. P. Mishra, "Modes decomposition method in fusion with robust random vector functional link network for crude oil price forecasting," *Appl. Soft Comput.*, vol. 80, pp. 475–493, Jul. 2019.

[13] A. Khan, R. Shah, J. Bukhari, N. Akhter, M. Idrees, and H. Ahmad, "A novel chicken swarm neural network model for crude oil price prediction," in *Advances on Computational Intelligence in Energy: The Applications of Nature-Inspired Metaheuristic Algorithms in Energy*, 2019, pp. 39–58. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A.+Khan%2C+R.+Shah%2C+J.+Bukhari%2C+N.+Akhter%2C+M.+Idrees%2C+and+H.+Ahmad%2C+947+%E2%80%98%E2%80%98A+novel+chicken+swarm+neural+network+model+for+crude+oil+price+pre948+diction%2C%E2%80%99%E2%80%99+in+Advances+on+Computational+Intelligence+in+Energy%2C+2019%2C+AQ%3A7+949+pp.+39%E2%80%9358.&btnG=

[14] M. Bristone, R. Prasad, and A. A. Abubakar, "CPPCNDL: Crude oil price prediction using complex network and deep learning algorithms," *Petroleum*, vol. 6, no. 4, pp. 353–361, Dec. 2020.

[15] E. Chen and X. J. He, "Crude oil price prediction with decision tree based regression approach," *J. Int. Technol. Inf. Manag.*, vol. 27, no. 4, pp. 2–16, May 2019.

[16] L.-T. Zhao and G.-R. Zeng, "Analysis of timeliness of oil price news information based on SVM," *Energy Proc.*, vol. 158, pp. 4123–4128, Feb. 2019.

[17] M. Oussalah and A. Zaidi, "Forecasting weekly crude oil using Twitter sentiment of US foreign policy and oil companies data," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 201–208.

[18] L. Yang, Y. Zhu, Y. Wang, and Y. Wang, "Multifractal detrended cross-correlations between crude oil market and Chinese ten sector stock markets," *Phys. A, Stat. Mech. Appl.*, vol. 462, pp. 255–265, Nov. 2016.

[19] X. Liu, H. An, S. Huang, and S. Wen, "The evolution of spillover effects between oil and stock markets across multi-scales using a wavelet-based GARCH–BEKK model," *Phys. A, Stat. Mech. Appl.*, vol. 465, pp. 374–383, Jan. 2017.

[20] B. Mo, H. Nie, and Y. Jiang, "Dynamic linkages among the gold market, US dollar and crude oil market," *Phys. A, Stat. Mech. Appl.*, vol. 491, pp. 984–994, Feb. 2018.

[21] J. Wu, J. Dong, Z. Wang, Y. Hu, and W. Dou, "A novel hybrid model based on deep learning and error correction for crude oil futures prices forecast," *Resour. Policy*, vol. 83, Jun. 2023, Art. no. 103602.

[22] R. Aloui, R. Gupta, and S. M. Miller, "Uncertainty and crude oil returns," *Energy Econ.*, vol. 55, pp. 92–100, Mar. 2016.

[23] A. Amano, "A small forecasting model of the world oil market," *J. Policy Model.*, vol. 9, no. 4, pp. 615–635, Dec. 1987.

[24] L. Tang and S. Hammoudeh, "An empirical exploration of the world oil price under the target zone model," *Energy Econ.*, vol. 24, pp. 577–596, Nov. 2002.

[25] S. Moshiri and F. Foroutan, "Forecasting nonlinear crude oil futures prices," *Energy J.*, vol. 27, no. 4, pp. 81–96, Oct. 2006.

[26] L. Yu, S. Wang, and K. K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Econ.*, vol. 30, no. 5, pp. 2623–2635, Sep. 2008.

[27] F. Gori, D. Ludovisi, and P. Cerritelli, "Forecast of oil price and consumption in the short term under three scenarios: Parabolic, linear and chaotic behaviour," *Energy*, vol. 32, no. 7, pp. 1291–1296, Jul. 2007.

[28] A. Khashman and N. I. Nwulu, "Support vector machines versus back propagation algorithm for oil price prediction," in *Proc. Int. Symp. Neural Netw.*, Guilin, China, Jun. 2011, pp. 530–538.

[29] J. Nasir, M. Aamir, Z. U. Haq, S. Khan, M. Y. Amin, and M. Naeem, "A new approach for forecasting crude oil prices based on stochastic and deterministic influences of LMD using ARIMA and LSTM models," *IEEE Access*, vol. 11, pp. 14322–14339, 2023.

[30] K. He, L. Yu, and K. K. Lai, "Crude oil price analysis and forecasting using wavelet decomposed ensemble model," *Energy*, vol. 46, no. 1, pp. 564–574, Oct. 2012.

[31] J.-L. Zhang, Y.-J. Zhang, and L. Zhang, "A novel hybrid method for crude oil price forecasting," *Energy Econ.*, vol. 49, pp. 649–659, May 2015.

[32] Y. Zhang, F. Ma, B. Shi, and D. Huang, "Forecasting the prices of crude oil: An iterated combination approach," *Energy Econ.*, vol. 70, pp. 472–483, Feb. 2018.

[33] L. Tang, Y. Wu, and L. Yu, "A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting," *Appl. Soft Comput.*, vol. 70, pp. 1097–1108, Sep. 2018.

[34] X. Hao, Y. Zhao, and Y. Wang, "Forecasting the real prices of crude oil using robust regression models with regularization constraints," *Energy Econ.*, vol. 86, Feb. 2020, Art. no. 104683.

[35] L. Lin, Y. Jiang, H. Xiao, and Z. Zhou, "Crude oil price forecasting based on a novel hybrid long memory GARCH-M and wavelet analysis model," *Phys. A, Stat. Mech. Appl.*, vol. 543, Apr. 2020, Art. no. 123532.

[36] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[38] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1462–1471.

[39] D. Eck and J. Schmidhuber, "A first look at music composition using LSTM recurrent neural networks," *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, vol. 103, pp. 48–56, 2002.

[40] W. Gao, M. Aamir, A. B. Shabri, R. Dewan, and A. Aslam, "Forecasting crude oil price using Kalman filter based on the reconstruction of modes of decomposition ensemble model," *IEEE Access*, vol. 7, pp. 149908–149925, 2019.

[41] E. Antwi, E. N. Gyamfi, K. A. Kyei, R. Gill, and A. M. Adam, "Modeling and forecasting commodity futures prices: Decomposition approach," *IEEE Access*, vol. 10, pp. 27484–27503, 2022.

[42] J. Liu and X. Huang, "Forecasting crude oil price using event extraction," *IEEE Access*, vol. 9, pp. 149067–149076, 2021.

[43] R. Tao, X. Zhang, and L. Zhao, "Forecasting crude oil prices based on an internet search driven model," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 4156–4161.

[44] X. Liu, "Research on crude oil price forecasting based on MLR-SVR decomposition integrated model," in *Proc. Int. Conf. Inf. Technol., Commun. Ecosystem Manag. (ITCEM)*, Dec. 2022, pp. 56–60.

[45] H. Yang, Y. Zhang, and F. Jiang, "Crude oil prices forecast based on EMD and BP neural network," in *Proc. Chin. Control Conf. (CCC)*, Jul. 2019, pp. 8944–8949.

[46] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *Int. J. Forecasting*, vol. 14, pp. 35–62, Mar. 1998.

[47] C.-J. Lu, T.-S. Lee, and C.-C. Chiu, "Financial time series forecasting using independent component analysis and support vector regression," *Decis. Support Syst.*, vol. 47, no. 2, pp. 115–125, May 2009.

[48] Z. Qian, Y. Pei, H. Zareipour, and N. Chen, "A review and discussion of decomposition-based hybrid models for wind energy forecasting applications," *Appl. Energy*, vol. 235, pp. 939–953, Feb. 2019.

[49] J. C. Reboredo and M. A. Rivera-Castro, "A wavelet decomposition approach to crude oil price and exchange rate dependence," *Econ. Model.*, vol. 32, pp. 42–57, May 2013.

[50] W. Sun and Z. Li, "Hourly $PM_{2.5}$ concentration forecasting based on mode decomposition-recombination technique and ensemble learning approach in severe haze episodes of China," *J. Cleaner Prod.*, vol. 263, Aug. 2020, Art. no. 121442.

[51] S. J. Koopman and P. H. Franses, "Constructing seasonally adjusted data with time-varying confidence intervals," *Oxford Bull. Econ. Statist.*, vol. 64, no. 5, pp. 509–526, Dec. 2002.

[52] E. Jianwei, Y. Bao, and J. Ye, "Crude oil price analysis and forecasting based on variational mode decomposition and independent component analysis," *Phys. A, Stat. Mech. Appl.*, vol. 484, pp. 412–427, Oct. 2017.

[53] S. Zhao, Z. Yang, S. Zhang, J. Wu, Z. Zhao, D.-S. Jeng, and Y.-G. Wang, "Predictions of runoff and sediment discharge at the lower yellow river delta using basin irrigation data," *Ecological Informat.*, vol. 78, Dec. 2023, Art. no. 102385.

[54] S. Zhang, J. Wu, Y.-G. Wang, D.-S. Jeng, and G. Li, "A physics-informed statistical learning framework for forecasting local suspended sediment concentrations in marine environment," *Water Res.*, vol. 218, Jun. 2022, Art. no. 118518.

[55] D. VandenHeuvel, J. Wu, and Y.-G. Wang, "Robust regression for electricity demand forecasting against cyberattacks," *Int. J. Forecasting*, vol. 39, no. 4, pp. 1573–1592, Oct. 2023.

[56] Y. Yang, H. Zhou, J. Wu, Z. Ding, Y.-C. Tian, D. Yue, and Y.-G. Wang, "Robust adaptive rescaled lncosh neural network regression toward time-series forecasting," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 9, pp. 5658–5669, Sep. 2023.
[57] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, Jan. 2009.
[58] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.
[59] P. Flandrin, P. Goncalves, and G. Rilling, "EMD equivalent filter banks, from interpretation to applications," in *Hilbert–Huang Transform and Its Applications*. Singapore: World Scientific, 2005, pp. 57–74.
[60] R. J. Gledhill, "Methods for investigating conformational change in biomolecular simulations," Doctoral dissertation, Univ. Southampton, 2003. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=R.+J.+Gledhill%2C+%E2%80%98%E2%80%98Methods+for+investigating+conformational+change+in+1095+biomolecular+simulations%2C%E2%80%99%E2%80%99+Doctoral+dissertation%2C+Univ.+Southampton%2C+AQ%3A8+1096+2003.&btnG=
[61] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *J. Econ.*, vol. 54, pp. 159–178, Oct. 1992.
[62] M. Aamir, A. Shabri, and M. Ishaq, "Improving forecasting accuracy of crude oil prices using decomposition ensemble model with reconstruction of IMFs based on ARIMA model," *Malaysian J. Fundam. Appl. Sci.*, vol. 14, no. 4, pp. 471–483, Dec. 2018.
[63] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer, 1999.
[64] Q. Zuo, X. Zhu, Z. Liu, J. Zhang, G. Wu, and Y. Li, "Prediction of the performance and emissions of a spark ignition engine fueled with butanol-gasoline blends based on support vector regression," *Environ. Prog. Sustain. Energy*, vol. 38, no. 3, p. e13042, May 2019.
[65] H.-Y. Zuo, Z.-Q. Luo, J.-L. Guan, and Y.-W. Wang, "Identification on rock and soil parameters for vibration drilling rock in metal mine based on fuzzy least square support vector machine," *J. Central South Univ.*, vol. 21, no. 3, pp. 1085–1090, Mar. 2014.
[66] Y.-L. Qi and W.-J. Zhang, "The improved SVM method for forecasting the fluctuation of international crude oil price," in *Proc. Int. Conf. Electron. Commerce Bus. Intell.*, 2009, pp. 269–271.
[67] Z. Xiao-lin and W. Hai-wei, "Crude oil prices predictive model based on support vector machine and particle swarm optimization," in *Software Engineering and Knowledge Engineering: Theory and Practice*, 2012, pp. 645–650.
[68] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

MUHAMMAD NAEEM received the bachelor's degree from the Islamia College, University of Peshawar, in 2014, and the master's degree in statistics from Abdul Wali Khan University Mardan, Pakistan, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Statistics. He is a Lecturer with the Government Degree College Lund Khwar, Mardan, Pakistan. His research interests include time series modeling, linear models, machine learning, applied statistics, and computational statistics.

MUHAMMAD AAMIR received the master's degree in statistics from Quaid-i-Azam University, Islamabad, Pakistan, in 2013, and the Ph.D. degree in mathematics from Universiti Teknologi Malaysia, Malaysia, in 2018. Currently, he is an Assistant Professor with the Department of Statistics, Abdul Wali Khan University Mardan, Pakistan. His current research interests encompass sampling, machine learning for mathematical problem diagnostics, econometrics, and time series analysis. Additionally, he actively contributes to the academic community by serving as a Reviewer for various international journals, including IEEE Access, *PLOS One*, *Resources Policy*, *Energy*, *Journal of Cluster Science*, and *Heliyon*.

JIAN YU received the B.Sc. degree in computer science from Zhejiang University, China, the M.Sc. degree in computer systems from Chinese Academy of Sciences, China, and the Ph.D. degree in computer software and theory from Peking University. He is currently working as an Associate Professor with Auckland University of Technology, New Zealand. He is the Director of the Ubiquitous and Web Computing Research Laboratory (UbiWeb), Department of Computer Science and Software Engineering, ECMS. His current research interests include deep learning for recommender systems, graph neural networks, complex networks, web and ubiquitous computing, and service-oriented computing. He has more than 140 publications. He has served as a PC Member for over 100 international conferences, including CORE Rank A conferences ACM CIKM, ICSOC, IEEE SCC, EAI MobiQuitous, COOPIS, and IEEE TrustCom.

OLAYAN ALBALAWI received the master's degree in mathematical science from Queensland University of Technology, Australia, in 2011, the Master of Statistics degree from The University of New Mexico, USA, in 2016, and the Ph.D. degree in applied statistics from the University of New South Wales, Australia, in 2020. He is currently an Assistant Professor with the Department of Statistics, Faculty of Science, University of Tabuk. He has published numerous articles in reputable journals.