Enhancing Oil Price Forecasting Through an Intelligent Hybridized Approach

Hicham BOUSSATTA[1], Marouane CHIHAB[2], Younes CHIHAB[3], Mohammed CHINY[4]
Laboratory of Computer Sciences, Faculty of Sciences Ibn Tofail University, Kenitra, Morocco[1, 2, 3]
Faculty of Sciences Ibn Tofail University, Kenitra, Morocco[4]

***Abstract*—The oil market has long experienced price fluctuations driven by diverse factors. These shifts in crude oil prices wield substantial influence over the costs of various goods and services. Moreover, the price per barrel is intricately intertwined with global economic activities, themselves influenced by the trajectory of oil prices. Analyzing oil behavior stands as a pivotal means for tracking the evolution of barrel prices and predicting future oil costs. This analytical approach significantly contributes to the field of crude oil price forecasting. Researchers and scientists alike prioritize accurate crude oil price forecasting. Yet, such endeavors are often challenged by the intricate nature of oil price behavior. Recent times have witnessed the effective employment of various approaches, including Hybrid and Machine Learning techniques to address similarly complex tasks, though they often yield elevated error rates, as observed in financial markets. In this study, the goal is to enhance the predictive precision of several weak supervised learning predictors by harnessing hybridization, particularly within the context of the crude oil market's multifaceted variations. The focus extends to a vast dataset encompassing CPSE Stock ETF prices over a period of 23 years. Ten distinct models, namely SVM, XGBoost, Random Forest, KNN, Gradient Boosting, Decision Tree, Ridge, Lasso, Elastic Net, and Neural Network, were employed to derive elemental predictions. These predictions were subsequently amalgamated via Linear Regression, yielding heightened performance. The investigation underscores the efficacy of hybridization as a strategy. Ultimately, the proposed approach's performance is juxtaposed against its individual weak predictors, with experiment results validating the findings.**

Keywords—Oil market; prediction; crude oil; hybrid approach; CPSE stock ETF price; machine learning; stock markets

I. INTRODUCTION

Hybridization approach refers to the combination of two different approaches or models to improve the accuracy of exchange rate forecasting in time series analysis. [1], given the significance of time-series prediction in many real-world situations, it is important to carefully select an appropriate model. For this reason, numerous performance measures have been proposed in the literature [1-7] to assess forecast accuracy and compare different models. These are known as performance metrics [6]. The goal of time series models is to gain an understanding of the underlying factors and structure that shaped the observed data, fit a model, and uses it for forecasting. These models have a wide range of applications in the daily operations of electric utilities, such as energy generation planning, energy purchasing, load switching, and contract evaluation [8], The purpose of forecasting is to make and improve decisions, increase profits, and in the case of forecasting oil prices, better decisions largely depend on the accurate prediction of trends, actual prices, and expected prices $x(t)$ and $\tilde{x}(t)$. The ability to predict movement can be measured statistically (R2) [9]. Similarly, the significance of forecasting lies in reducing the risk or uncertainty involved in short-term decision making and planning for long-term growth. Forecasting the demand and sales of a company's products usually begins with a macroeconomic forecast of the overall level of economic activity, such as Gross National Product (GNP). Companies use macroeconomic forecasts of general economic activity as inputs for their microeconomic estimates of the demand and sales for the industry and the firm. The demand and sales for a business are typically estimated based on its historical market share and planned marketing strategy (e.g, forecasting by product line and region). Companies use long-term forecasts for the industry and the economy to determine the necessary investment in plant and equipment to achieve their long-term growth objective. The focus of this study is on multi-step ahead prediction of crude oil prices. This involves extrapolating the crude oil price series by predicting multiple time-steps into the future without access to future outputs. Despite the influence of many complex factors, oil prices exhibit highly non-linear behavior, making it challenging to predict future oil prices, especially when looking several steps ahead (Fan et al., 2008) [10]. The unpredictability of crude oil prices is due to their sensitivity to fluctuations in both global demand and supply. The world economy was destabilized for a decade when oil supplies were disrupted 40 years ago. The formation of the OPEC cartel and the nationalization of the oil industries in the Middle East led to a quadrupling of world oil prices and caused steep recessions in the mid-1970s. The 1979 overthrow of the Shah of Iran by Muslim clerics disrupted Iran's oil supplies, leading to another round of even deeper recessions. The productivity of the future oil market and the expected accuracy of future prices are evaluated. The precision of forecasts using futures prices is compared to that of other methods, including time series and econometric models, as well as key forecasts. The predictive power of futures prices is further investigated by comparing the forecasting accuracy of end-of-month prices with weekly and monthly averages, using different weighting systems [11] Previous studies have shown that the behavior of oil prices is non-linear and traditional econometric and statistical methods struggle to provide accurate predictions in these cases. To address this issue, newer techniques like genetic algorithms, artificial neural networks, and support vector machines have emerged [13]. Alizadeh and Mafinezhad [13] used a General

Regression Neural Network (GRNN) model to forecast Brent oil prices by incorporating seven types of variables as inputs. The authors claimed that this model performed well under various conditions and provided a high level of accuracy. Previous studies have shown that oil price behavior is non-linear, and traditional econometric and statistical models may not be sufficient for analyzing this behavior [12]. To address this, new techniques such as genetic algorithms, artificial neural networks, and support vector machine have emerged [13]. Alizadeh and Mafinezhad [13] used a General Regression Neural Network (GRNN) model to forecast Brent oil price, incorporating seven types of structures as inputs. They found that their model provided a high level of accuracy under challenging conditions. Predicting oil prices is a challenging task due to its significant impact on various economic and non-economic aspects. There is currently a lack of consensus among experts on the most effective methods and models for forecasting oil prices. To address this issue, a hybridization approach that combines multiple models can be used to increase forecasting accuracy [14]. In this study, the aim is to enhance the accuracy of crude oil price predictions by combining weak predictors through hybridization. The dataset comprises daily CPSE Stock ETF prices spanning 23 years. Ten different machine learning models are utilized. (SVM, Random Forest, Gradient Boosting, Neural Network, XGBoost, Decision Tree, Ridge, Lasso, Elastic Net and KNN) to make individual predictions, and then combine these predictions using a Linear Regression model to achieve enhanced performance. The results clearly illustrate the advantages of employing the hybridization approach. After testing the 10 models, The SMRM approach yielded the most accurate results, as it converges towards the minimum of the empirical response and minimizes information loss, outperforming the other models. The study's findings suggest that the Hybrid Proposed System (SMRM) stands out as the most efficient option, outperforming all other individual models. The SMRM achieved an average negative MAPE of -0.023, which was the highest among all models and sets. The proposed SMRM hybrid system offers a promising solution for predicting crude oil prices, leveraging the power of machine learning, and combining multiple models to better capture the complex relationships between different factors. The experiments reveal that SMRM excels over existing models in both accuracy and stability, making it a valuable tool for investors, traders, and other stakeholders in the energy sector. The system can also be continually refined and improved by incorporating irregular factors like political risks and extreme weather events, which can help to better predict changes in crude oil prices. With further development, this approach could have important implications for supporting decision-making and risk management in the energy sector, enabling stakeholders to make more informed and effective decisions in the dynamic and complex world of stock market trading. By providing more accurate and reliable predictions of crude oil prices, the SMRM hybrid system has the potential to revolutionize how we approach predicting crude oil prices, providing valuable insights that can help optimize decision-making and drive greater value in the energy sector. This passage highlights that the main research contribution is the development of the SMRM hybrid system, which combines machine learning models to improve the accuracy and stability of crude oil price predictions. It also emphasizes the system's potential to enhance decision-making and risk management in the energy sector.

The rest of the paper is organized as follows: Section II reviews the current literature on forecasting crude oil prices. Following that, Section III details methodology, and Section IV covers the proposed approach. The empirical results are presented in section V.

II. Related Work

Crude oil prices are difficult to predict due to a complex pricing system with insufficient information, numerous variables, and inaccurate elements [15]. Despite this challenge, researchers are actively exploring methods to accurately forecast crude oil prices and manage related risks. While traditional methods like Arima and Arma have been used, they often fall short in the face of complex data and asymmetric effects. The growth of AI and text mining technologies provide new opportunities to predict crude oil prices and measure investment risk. One such successful machine learning model [16] uses artificial intelligence to predict crude oil prices, including the use of Decision Trees (DTs), a commonly used technique in crude oil modeling. To further improve accuracy, [17] incorporates technical trading indicators like RSI and Stochastic Oscillator into the Random Forest model for minimizing investment risk. These advances in AI technology have the potential to significantly improve the accuracy of crude oil price forecasts. Two PCA-KNN models, which combined PCA for information reduction and KNN for oil price forecasting, were tested on historical EUR/USD exchange rate data sets over a 10-year period. These models achieved the highest success rate of 77.58%. To improve the success rate, [18] presents a novel approach to financial time series prediction by combining K-Nearest Neighbor (KNN) Regression with Principal Component Analysis (PCA). The authors aim to enhance the accuracy of financial time series forecasting, a critical task in the finance domain, [19] proposed a PANK model, which involves three components: (1) Principal Component Analysis to minimize redundant information, (2) Affinity Propagation Clustering for feature extraction through example generation and corresponding cluster formation, and (3) nearest k-neighbour regression reformulated through nested regression for prediction modeling. The PANK model was tested on a 15-year historical data set of the Chinese stock market index, yielding a success rate of 80%. Previous studies have defined the behavior of oil prices as a statistical system, but these methods only provide logical results for linear behavior. The study in [20] used the SVM model to estimate oil prices, but these methods are inadequate for highly complex and non-linear data. The study in [21] recently compared different forecasting models, including ARIMA, FNN, ARFIMA, MS-ARFIMA, and the RW model, and found that the SVM model performed best and is a strong candidate for crude oil price forecasting in one or more stages. Machine learning models often require hand-crafted features, which can make them challenging to implement in real-world situations, especially with large amounts of data. A recent and successful approach comes from the subfield of machine learning, deep learning [22]. The accuracy of forecasting can be improved through hybridization by combining simple forecasts from multiple weaker predictors [23]. A hybrid system combining Artificial Neural Networks (ANN) and Recurrent Neural Networks (RES) based on text mining was proposed to improve prediction performance [24]. Another study proposed a Multi-Intelligent Bat-Neural Network Multi-Agent System (BNNMAS) for predicting the price of oil-linked stocks, comparing it to genetic algorithm neural network (GANN) and generalized neural network regression (GRNN) [24]. However, both systems are subject to performance problems due to the identification of tuning parameters. A recent study [25] developed an intelligent system for forecasting oil prices using time series models, but the system is not suitable for predicting long-term trends. In reference to the study of oil prices, various methods have been used to make predictions and analyze the factors affecting its fluctuation. The research in [26] utilized the Complementary Empirical Ensemble Mode Decomposition (CEEMD) to break down the barrel price into its components and identify the impact of extreme events on crude oil prices. The researcher combined the Iterative Cumulative Sum of Squares (ICSS) test and Chow's test to detect structural breaks, then used ARIMA and SVM models to forecast oil prices. The results showed that the SVM-CEEMD-ARIMA model with structural breaks was the best performing model compared to SVM and ARIMA models alone. During the COVID-19 pandemic, [27] attempted to predict oil price movements using ANN and SVM models. The results showed that his model outperformed other ANN and SVM models, with an RMSE value of 0.6018 and a MAE value of 0.5295. The study in [28] used a Convolutional Neural Network (CNN) to extract features from online news texts, divided into categories such as oil price, oil production, overall oil consumption, and oil stocks. Other models such as MLR, BPNM, SVM, RNN, and LSTM were used for prediction and the results showed that social media factors play a role in oil price prediction. During the Russian-Ukrainian conflict, [29] forecasted oil prices by using exotic options such as Asian Options, Barrier Options, and Gap Options. The GARCH model and Monte Carlo simulation were used to study the options and the results showed that considering the overall performance of all exotic options was better. In [30], the authors analyzed the relationship between oil prices and various wars, including the first and second Gulf Wars and the Russian-Ukrainian War. The results showed that the relationship between real GDP growth and oil prices differed between periods and that it was possible to predict oil prices during the Russian-Ukrainian War. The ongoing conflict between Russia and Ukraine continues to have a significant impact on the financial market and oil prices. The research in [31] used the TVP-VAR technique to identify the sources of oil market volatility and the interconnections between gold, crude oil, and the stock market on February 24, 2022. The results showed that the conflict between Ukraine and Russia affects the interdependence of the markets analysed, both in stable and war situations. In [32], the authors compared the performance of support vector machines (SVM), K-nearest neighbors (KNN), and random forest (RF) models in predicting crude oil prices. The authors found that the SVM model outperformed the KNN and RF models in terms of both in-sample and out-of-sample prediction accuracy. Guliyev and Mustafayev in [33] to compare their predictive accuracy and identify the most effective model for crude oil price forecasting. The three machine learning models used in the study are Linear Regression, Support Vector Regression (SVR), and Random Forest Regression. These models are trained using a range of explanatory variables, such as supply and demand factors, macroeconomic indicators, geopolitical risks, and oil market-specific variables, such as oil inventories. The study found that all three models can effectively predict the WTI crude oil price changes with reasonable accuracy. However, the Random Forest model produced the most accurate forecasts compared to the other two models. In addition, the study found that geopolitical risks, such as tensions between OPEC members and potential supply disruptions, have the most significant impact on WTI crude oil prices. The study's results suggest that machine learning models can be a useful tool for crude oil price forecasting. The findings could be useful for traders, investors, and policymakers who need to make informed decisions based on the expected future price dynamics of crude oil. The study in [34] is to evaluate the effectiveness of different machine learning models in predicting the closing prices of stocks. The study employs three machine learning algorithms: Random Forest (RF), Support Vector Regression (SVR), and Multilayer Perceptron (MLP) for predicting the closing prices of stocks using a range of input features such as volume, moving averages, and technical indicators. The study finds that all three machine learning models are effective in predicting the closing prices of stocks, with Random Forest performing the best, followed by Support Vector Regression and Multilayer Perceptron. The study also found that technical indicators, such as Relative Strength Index (RSI) and Moving Average Convergence Divergence (MACD), were the most effective input features for the prediction models. Overall, the study suggests that machine learning models can be useful for predicting the closing prices of stocks and can help traders and investors make informed decisions based on expected future prices. [35] is to develop a hybrid artificial intelligence model to predict the uniaxial compressive strength of oil palm shell concrete. The study uses a combination of machine learning algorithms, including Artificial Neural Networks (ANN), Genetic Programming (GP), and Support Vector Regression (SVR), to develop a hybrid model for predicting the uniaxial compressive strength of oil palm shell concrete. The model uses input features such as the water-binder ratio, curing time, and oil palm shell content. The study finds that the hybrid model outperforms individual machine learning models in predicting the uniaxial compressive strength of oil palm shell concrete. The hybrid model also achieves a high prediction accuracy with a coefficient of determination (R-squared) value of 0.965. The results of the study suggest that the hybrid model can be a useful tool for predicting the uniaxial compressive strength of oil palm shell concrete, which is important for the design and construction of sustainable building materials. The hybrid model can also be extended to predict the properties of other types of concrete by modifying the input features. The research in [36] is to propose a novel hybrid model for forecasting crude oil prices based on time series decomposition. The study combines two machine learning algorithms, Support Vector Regression (SVR) and Artificial Neural Networks (ANN), with a time series decomposition method called Seasonal and Trend decomposition using Loess (STL), to create a hybrid model for crude oil price forecasting. The study finds that the proposed hybrid model outperforms individual machine learning models and traditional time series models in forecasting crude oil prices. The hybrid model achieves a high prediction accuracy with a Mean Absolute Percentage Error (MAPE) value of 3.22%. The results of the study suggest that the proposed hybrid model can be a useful tool for predicting crude oil prices, which is important for making informed decisions in the oil and gas industry. The study also highlights the importance of combining different machine learning algorithms and time series decomposition methods for improving the accuracy of crude oil price forecasting. The study in [37] is to propose a novel approach for predicting crude oil prices by combining complex network analysis and deep learning algorithms. The study uses a complex network analysis to identify the relationships and dependencies between different economic variables, such as exchange rates and stock prices, and crude oil prices. The identified network is then used as input for deep learning algorithms, specifically a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN), to predict crude oil prices. The study finds that the proposed approach outperforms traditional machine learning models and time series models in predicting crude oil prices. The CNN and RNN models achieve high prediction accuracy with a Mean Absolute Error (MAE) value of 0.55 and 0.51, respectively. The results of the study suggest that the proposed approach can be a useful tool for predicting crude oil prices, which is important for making informed decisions in the oil and gas industry. The study also highlights the importance of considering the complex relationships and dependencies between different economic variables in crude oil price prediction. In [38], Abdollahi and Ebrahimi propose a new hybrid model for predicting the Brent crude oil price. The proposed model combines two different techniques: an ensemble of Extreme Learning Machines (ELMs) and a wavelet transform. The study first applies a wavelet transform to decompose the time series data into different frequency bands, which allows the model to capture different patterns and trends in the data. The decomposed signals are then used as input for the ELM ensemble, which is a machine learning technique that combines multiple ELM models to improve prediction accuracy. The study finds that the proposed hybrid model outperforms several benchmark models, including traditional statistical models, machine learning models, and other hybrid models. The proposed model achieves a high prediction accuracy, with a Mean Absolute Percentage Error (MAPE) value of 1.76% for one-day ahead forecasting and 3.31% for five-day ahead forecasting. The results of the study suggest that the proposed hybrid model can be an effective tool for predicting the Brent crude oil price, which is important for making informed decisions in the oil and gas industry. The study also highlights the importance of combining different techniques to improve prediction accuracy and to capture different patterns and trends in the data. The research in [39] is to propose a weighted hybrid data-driven model for forecasting daily natural gas prices. The proposed model combines two different techniques: an Empirical Mode Decomposition (EMD) method and a Support Vector Machine (SVM) method. The study first applies the EMD method to decompose the original time series data into several Intrinsic Mode Functions (IMFs). These IMFs capture the different temporal scales of the natural gas prices and are used as input variables for the SVM method. The SVM method is a popular machine learning algorithm that can be used for regression and classification tasks. To further improve the performance of the model, the study introduces a weight-based approach that assigns different weights to the historical data based on their importance. The weights are calculated using a genetic algorithm that searches for the optimal weight values. The study finds that the proposed weighted hybrid data-driven model outperforms several benchmark models, including traditional statistical models, machine learning models, and other hybrid models. The proposed model achieves a high prediction accuracy, with a Mean Absolute Percentage Error (MAPE) value of 1.87% for one-day ahead forecasting and 2.55% for five-day ahead forecasting. The results of the study suggest that the proposed weighted hybrid data-driven model can be an effective tool for forecasting daily natural gas prices, which is important for making informed decisions in the energy industry. The study also highlights the importance of combining different techniques and introducing a weight-based approach to improve prediction accuracy.

Prediction accuracy, with a Mean Absolute Percentage Error (MAPE) value of 1.87% is for one-day ahead forecasting and 2.55% is for five-day ahead forecasting. The results of the study suggest that the proposed weighted hybrid data-driven model can be an effective tool for forecasting daily natural gas prices, which is important for making informed decisions in the energy industry. The study also highlights the importance of combining differeProposed System

In this article, the aim is to enhance prediction accuracy by combining several weak predictors through hybridization to address the varying degrees of variability in the oil market. Ten models are used (SVM, XGBoost, Random Forest, KNN, Gradient Boosting, Decision Tree, Ridge, Lasso, Elastic Net and Neural Network) to obtain basic predictions and then integrate them in a Linear Regression for a better outcome. The previous section discussed different types of algorithms including Text Mining algorithms, genetic algorithms, and deep learning algorithms. Although advancements have been made in dynamic system modeling and analysis over the past 23 years to minimize prediction error, the uncertainty of learning models remains limited. However, combining diverse predictive models can prove to be effective in improving precision accuracy. The study centers on the integration of various regression methods (SVM, Random Forest, Gradient Boosting, Neural Network, XGBoost, Decision Tree, Ridge, Lasso, Elastic Net and KNN) to make predictions and produce a final prediction through stacking. The idea is to construct a predictive model by combining various models, as shown in the diagram:

A. Units

The original training dataset, (X), consists of m observations and n features, resulting in an (m×n) matrix. Multiple models M are trained on X using a training method, such as cross-validation. These models make predictions for the result, (y), which are then consolidated into a second training dataset, (X^I2), with a shape of (m ×M). These (M) predictions serve as features for this second-level dataset. The goal of creating a second-level model is to produce the final prediction by utilizing the combination of these different models.

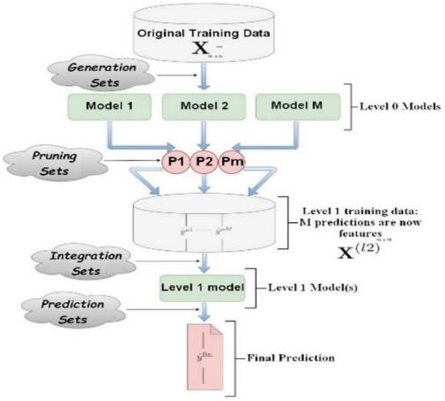

Fig. 1. Stacking Multi-Regression Models (SMRM).

B. Purning Sets

Evaluation criteria such as MAPE, MAE, RMSE and R-Squared R2 play a crucial role in selecting the best models based on ranking. These criteria are presented in Table I.

TABLE I. META-FEATURES USED

Features	Description
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
R2	R-Squared

C. Integration Sets

This section explains the behavior of the selected models as shown in Table II in their combinations for predicting barrel price. Stacking models is a technique used to create a secondary dataset for cross-validation using k-fold to address two key issues:

To make off-sample predictions, the stacking process uses predictions f1,...,fm to determine the generator biases for the learning set in different regions, where each model is most effective. The right linear combination with the weights vector is then learned by the meta-learner ai,...,(i=1,...m):

$$f_{stacking}{}^{(x)} = \sum_{i=1}^{m} a_i f_i(\boldsymbol{x}) \qquad (1)$$

D. Prediction Sets

The final predictions are generated from the training data X or from the second-level learner's model(s). The stacking model is used to select various sub-learners and to study how to collect and combine sub-models and their predictions. A meta-model is employed to merge the best predictions from all models. Each model provides predictions for the outcome (y), which are integrated into a second training dataset (X^I2), resulting in (m ×M) predictions. These second-level data possess M new characteristics. A second-level model is created to generate the results used for the final prediction. Fig. 1 illustrates the overall design of the results after applying this approach. Three models were generated and the basic model type at level 0, as well as the differences between the ten models, is explained as follows:

1) The first model used in the base layer is Random Forest
2) The second model used in the base layer is KNN
3) The third model used in the base layer is SVM
4) The fourth model used in the base layer is Gradient Boosting
5) The fifth model used in the base layer is Decision Tree
6) The sixth model used in the base layer is Ridge.
7) The seventh model used in the base layer is Lasso
8) The eighth model used in the base layer is Elastic Net
9) The ninth model used in the base layer is XGBoost
10) The tenth model used in the base layer is Neural Network

TABLE II. DATA MINING ALGORITHMS

Algorithm	Description
KNN	K-Nearest Neighbour
Decision Tree	Decision Tree
SVM	RSuport Vector Machine
Neural Network	Neural Network
Ridge	Ridge
Lasso	Lasso
Elastic Net	Elastic Net
XGBoost	XGBoost
Random Forest	Random Forest
Gradient Boosting	Gradient Boosting

III. STUDY OF THE PROPOSED APPROACH

The prediction capacity is tested against some reference models. First, the data description will be presented in Section (A). Second, all measures for evaluating prediction performance and the statistical tests that compare and adjust predictive accuracy will be discussed in Section (B). Finally, the stacking learning sets algorithm will be explained in Section (C).

A. Dataset

The ETF Prices data was used as a reference point and was uploaded to the ETF prices website [40]. This data represents the global oil price and is daily in nature, covering the period from 2000 to 2023 with 5751 observations. The data consists of important factors that impact supply and demand and the dependent variable of oil consumption. These variables were selected to model the barrel price series for the following reasons: Firstly, they are closely linked to oil prices and represent various drivers of the end price. Secondly, the relationship between these factors and the oil price series is noisy, non-linear, and volatile, but one of them is likely to provide valuable information on oil price schedules at any given time. Thirdly, more insights can be gained by including as many variables as possible. Finally, the system that contains all these models, namely (SVM, Random Forest, Gradient Boosting, XGBoost, Neural Network, Decision Tree, Ridge, Lasso, Elastic Net and KNN) is mainly powerful in modeling high-dimensional data using all these variables. The data is divided into two parts, with the training samples consisting of the first 80% of observations of all series and the rest used as test data, as shown in Fig. 2. All data is obtained from websites, including the Energy Information Administration (EIA), Exchange Traded Funds (ETF), and Yahoo Finance (36). The visualization of the entire actual time series (annual and monthly) is shown in Fig. 3 and Fig. 4. For model formation, oil prices and exogenous variables are pre-processed using first differences and standardization. The use of first differences helps remove zero values, and the use of standardized variables avoids estimation problems such as an explosion of parameters. The methods are designed to model price series rather than oil performance series. The proposed system SMRM is developed with the aim of identifying and validating the factors that contribute to oil price variations. To achieve this objective, a hybrid system SMRM is implemented and utilized to understand the fluctuation of the barrel price while utilizing information from Yahoo Finance. Table IV demonstrates the storage and testing of the data based on its quantitative characteristics. The system reveals 12 key performance indicators which were used as explanatory variables in the model with the expected next-day oil price as the dependent and output variable. These key performance indicators are listed in Table III.

TABLE III. KEY PERFORMANCE INDICATORS AFFECTING CRUDE OIL PRICES

Algorithm	Description
S_3	Moving average for past 3 days
S_9	Moving average for past 9 days
RSI_3	Moving average of Relative Strength Index for past 3 days
RSI_9	Moving average of Relative Strength Index for past 9 days
MME_26	Exponential Moving Average for past 26 days
MME_12	Exponential Moving Average for past 12 days
%K	Stochastic Oscillator
RVI	Relative Vigo Index
MOM	Momentum
MACD	Mobile Average Convergence Divergence
%D_3	Moving Average of %L for past 3 days
%D_9	Moving Average of %L for past 9 days

TABLE IV. OIL MARKET EXPLANATORY VALUES

Features	Description
Close	Reference to the end of a trading
Open	Reference to the starting period of trading
Hight	Reference to the involving large amounts of price
Low	Reference to the reaching of the price

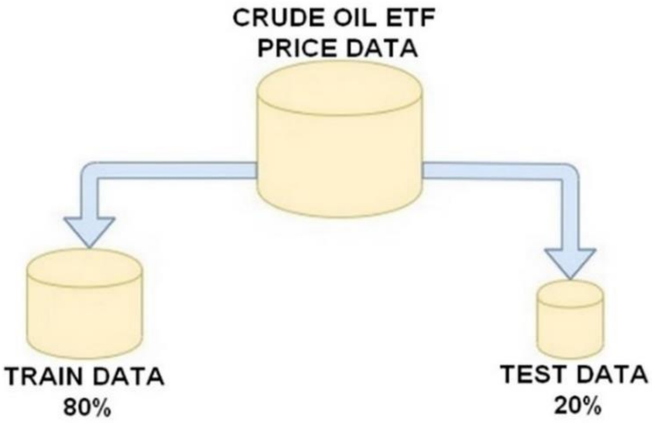

Fig. 2. Splitting the dataset.

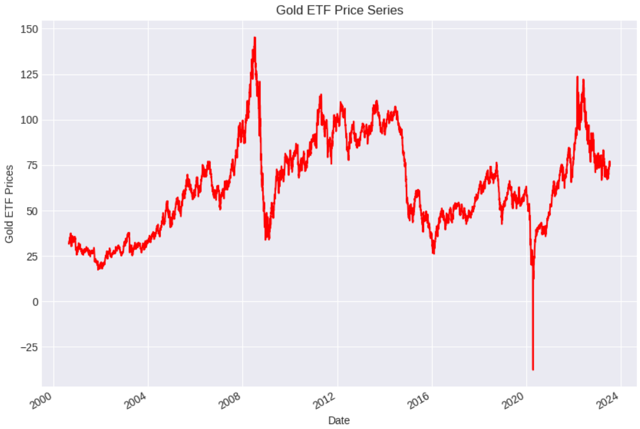

Fig. 3. Annual crude oil price.

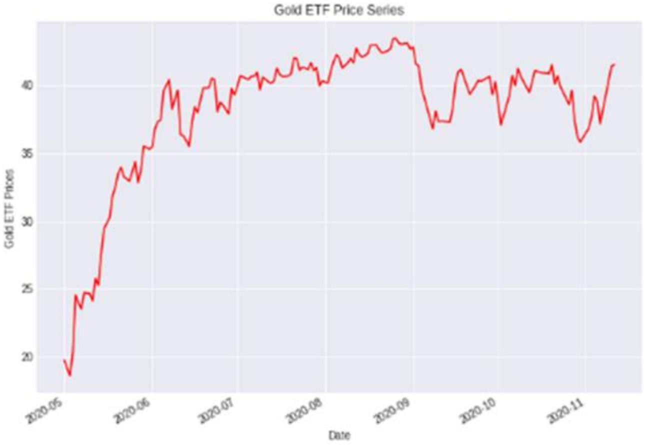

Fig. 4. Monthly crude oil price.

B. Performance evaluation criteria and statistical test

In evaluating the performance of the prediction models, four important indicators were used. These include the Mean Absolute Percentage Error MAPE calculated using Eq. (2), Mean Absolute Error MAE calculated using Eq. (3), Root Mean Squared Error (RMSE) calculated using Eq. (4) and R-Squared R2 calculated using Eq. (5). These indicators play a crucial role in estimating the performance of prediction models across various aspects.

$$MAPE = \frac{1}{N}\sum_{t=1}^{N} \frac{y(t) - \hat{y}(t)}{\hat{y}(t)} \qquad (2)$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}(y - \hat{y})^2 \qquad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{y}(t))^2} \qquad (4)$$

$$R2 = 1 - \frac{\sum_{t=1}^{N}(y - \hat{y})^2}{\sum_{t=1}^{N}(y - \hat{y})^2} \qquad (5)$$

Or y(t) and y ̂(t) represent, respectively, the actual value and the predicted value, a(t)=1 if (y(t+1)- y(t))(y ̂(t+1)- y(t))≥ or a(t)=0, and N is the size of the predictions.

C. *The Algorithm for Staking Learning Sets*

This section outlines the process of the proposed system, which is based on a typical sequence of the dataset for improved prediction. The general design of the method is expressed using pseudo code.

```
Input:  Dataset {D=(x_1,y_1 ),(x_2,y_2 ),…,(x_m,y_m )}
First-level learning algorithms  L_1,L_2,…,L_n
          Second-level learning algorithm L
Process:
% Have a training of first-level individual learner h_t by applying the first-level learning algorithm L_t to the original dataset D
     for t = 1,…, T:
         h_t=L_t (D)
     end
% Generation of a new dataset
    D^'=∅
    for t = 1,…,m
           for  t = 1,…,T
                  z_it=h_i (x_i) % Used h_t to predict training dataset x_i
           end
       D^'=D^'∪ {((z_I1  ,z_i2,…,z_T ),y_i )}
   end
% Have a training of the second h^' learner by applying the second level, learning the L algorithm to the new dataset set D^'
```

IV. Results

A. *Benchmarking and Comparison of the Predictive Modelling*

The results of testing 10 models (SVM, XGBoost, Random Forest, KNN, Gradient Boosting, Decision Tree, Ridge, Lasso, Elastic Net and Neural Network) show that the use of the Random Forest algorithm is more effective and closer to reality. This is because the Random Forest algorithm minimizes information loss and converges towards the minimum of empirical response, as shown in Fig. 6. On the other hand, the other models appear to be less effective and less modulable, with more variability present in the data, as seen in Fig. 5 and 7 in which, the case of the two SVM and KNN models is considered.

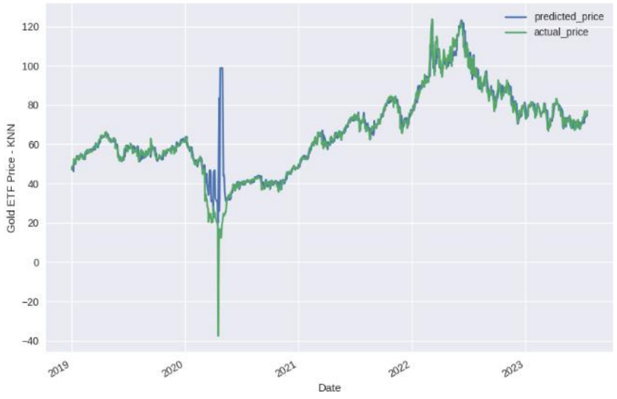

Fig. 5. Predicted and actual ETF Price via the KNN model.

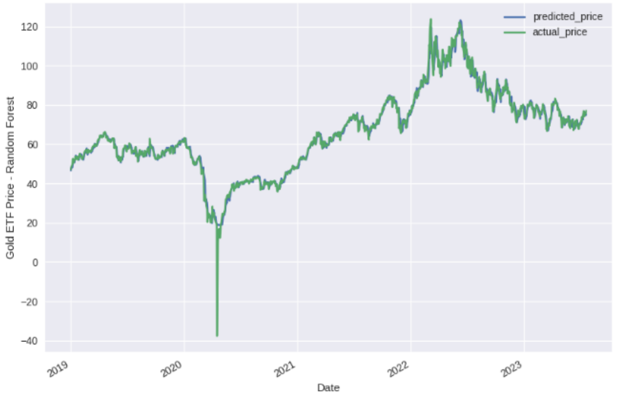

Fig. 6. Predicted and actual ETF price via the Random Forest regressor model.

B. *Examination of Algorithms*

First, the ETF Price data was utilized as a reference dataset, and various machine learning models were examined on the dataset. The assessment will include the following 10 algorithms.

1) *KNN*
2) *Random* Forest
3) *SVM*
4) *XGBoost*
5) *Gradient* Boosting
6) *Neural* Network
7) *Decision* Tree
8) *Ridge*
9) *Lasso*
10) *Elastic* Net

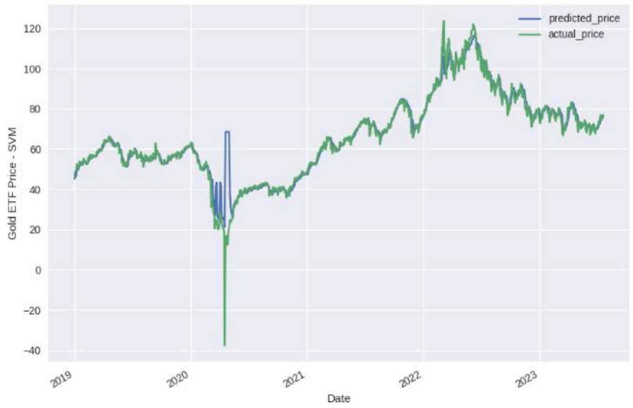

Fig. 7. Predicted and actual ETF Price via the SVM model.

All the algorithms will be evaluated, and their average performance will be compared by describing the distribution of accuracy scores for each. The models will undergo evaluation based on the MAPE. Due to the stochastic nature of the algorithm and numerical accuracy differences, the results may vary. In this context, it was found that the Random Forest algorithm gave the best result with a negative MAPE of approximately -0.021, as shown in Table VI, and with a best R2 of 95.40% as shown in Fig. 8.

C. The Combination of Models

The approach defines the Stacking Multi-Regression Models (SMRM) by initially presenting a list of tuples for the 10 basic models and subsequently defining the Linear Regression, which acts as a meta-model, to combine the predictions of the basic models and learn how to best combine the outputs of each of the 10 separate models (see Fig. 1). This implementation allows us to assess the performance of each model and the findings indicate that SMRM is the most efficient when compared to the other models, with a negative MAPE of approximately -0.023. The average and median scores for the SMRM are the highest in comparison to the other individual models, as seen in Table V. However, A stacking set can be chosen as the final model, fine-tune it, and use it to make predictions on novel datasets using the linear model created from all the training data. The prediction method estimates the ETF Price (y) for the explanatory variable X as shown in Fig. 10. The results show that the R2 has a score of 97% as shown in Fig. 9. A score close to 100% indicates that the model effectively explains the ETF prices for crude oil. The cumulative returns, represented as a purchase signal, are shown in Fig. 11, where a "1" value indicates that the expected price for the next day is higher than the current day's price, while no position is taken otherwise.

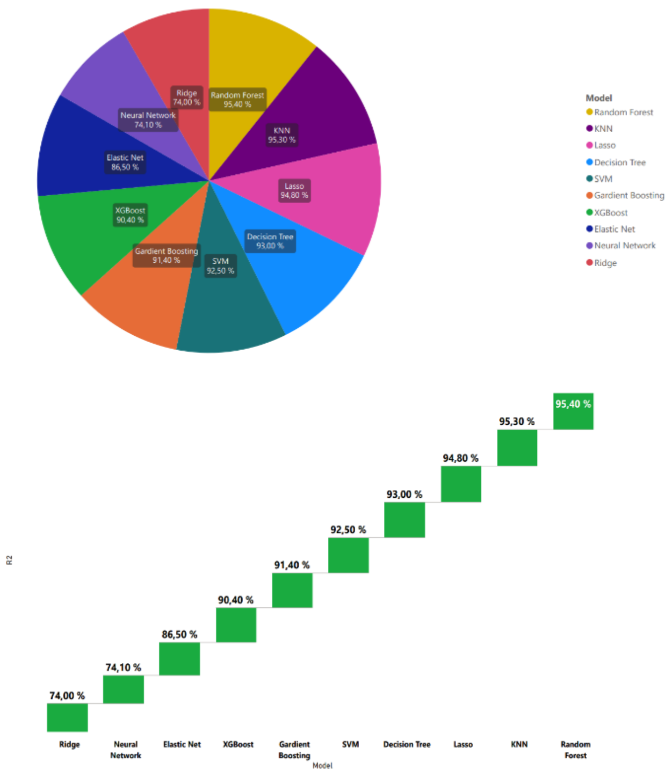

Fig. 8. Average performance of algorithms.

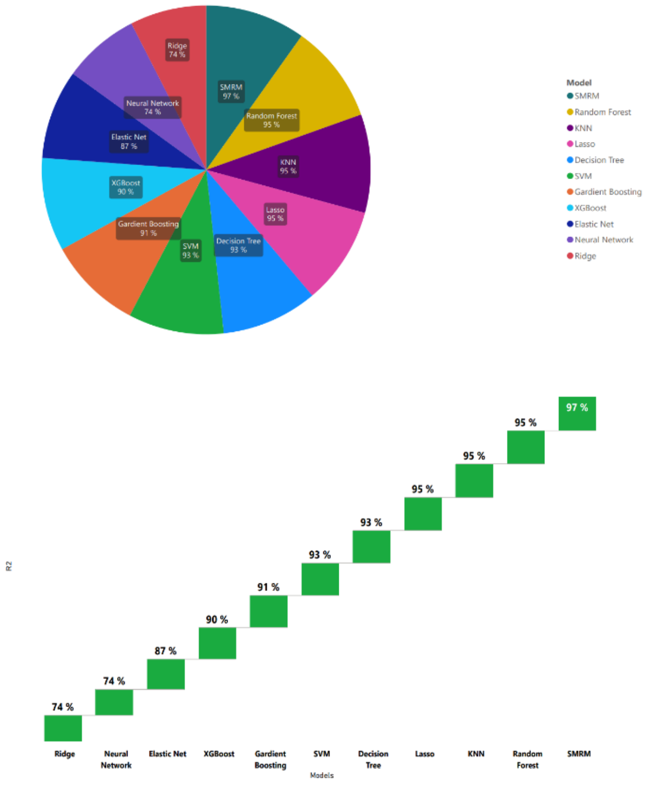

Fig. 9. Average algorithm performance and SMRM.

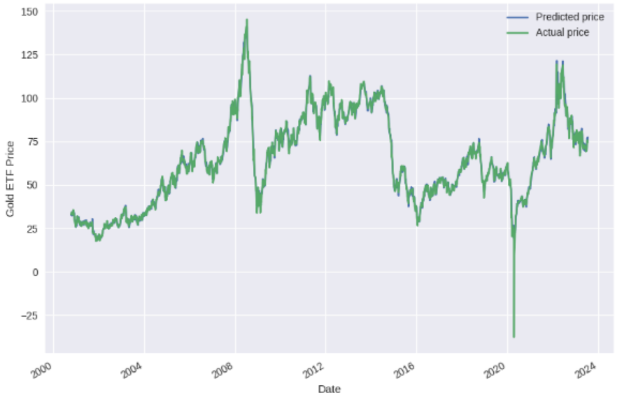

Fig. 10. Predicted price and ETF price.

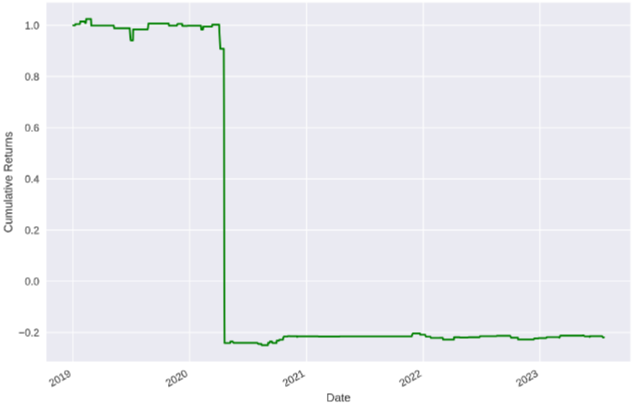

Fig. 11. Cumulative returns signals crude oil price.

V. Discussion

In this section, a stacking set can be chosen as the final model. Both individual and overall predictors have been applied to address the problem of estimating the expected price of the next day, providing a more comprehensive evaluation of the stability of the SMRM hybrid system. All models were examined in their predictions. It is evident that the proposed SMRM hybrid system was the best for forecasting oil prices, as shown in Fig 10, compared to the other models studied. In all models, the Random Forest Regressor model based on the SMRM hybrid system not only achieved the highest accuracy in estimation, as measured by MAPE, but also achieved the highest success rate in R2, as measured by R2. Additionally, among the models studied, the KNN performed the worst in growth forecasts. This model not only had the lowest MAPE, but also achieved the worst score in terms of R-Squared, as measured by R2. This could be since SVM was a linear regressor, which couldn't capture non-linear models. In addition to the Random Forest, KNN, and the other models based on the SMRM hybrid system, which produced the best and worst results, respectively, all models studied produced encouraging mixed results, which were analyzed using four evaluation criteria (i.e., MAPE, MAE, RMSE, and R2. First, in terms of level accuracy, the results of the MAPE measurement showed that the Random Forest based on SMRM achieved the best results, followed by the other models that are based on SMRM was the weakest, as shown in Table VI. Second, high accuracy does not necessarily imply a high success rate in predicting the R2. It is crucial that the R2 is correct for the policy maker to make an investment plan in oil-related processes (production, price, and demand). Thus, comparison of the R2 is essential. Similar conclusions can be drawn from Table VII regarding the R2. The SMRM hybrid system achieved significantly higher results and closer to 100 than all others, followed by the other 10 overall models (i.e., SVM, XGBoost, Random Forest, KNN, Gradient Boosting, Decision Tree, Ridge, Lasso, Elastic Net and Neural Network). The 10 overall methods typically outperformed individual forecasting models, and among the overall methods, the Random Forest model based on the SMRM hybrid system produced the best results, while the other models based on SMRM, The Ridge model exhibited the lowest R-Squared at 74%, as shown in Table VII. The Random Forest Regressor model within the SMRM hybrid system demonstrated the ability to adapt to the data, meaning that the difference between the predicted and observed values is important (RMSE and MAE), as shown in Tables VIII and IX.

TABLE V. MAPE BY MULTIPLE REGRESSION MODELS

Algorithm	MAPE
KNN	-0,023
Random Forest	-0,021
SVM	-0,031
XGBoost	-0,022
Gradient Boosting	-0,022
Neural Network	-0,035
Decision Tree	-0,03
Ridge	-0,035
Lasso	-0,028
Elastic Net	-0,032
SMRM	-0,023

TABLE VI. MAPE BY INDIVIDUAL REGRESSION MODELS

Algorithm	MAPE
KNN	-0,023
Random Forest	-0,021
SVM	-0,031
XGBoost	-0,022
Gradient Boosting	-0,022
Neural Network	-0,035
Decision Tree	-0,03
Ridge	-0,035
Lasso	-0,028
Elastic Net	-0,032

TABLE VII. R2 BY MULTIPLE REGRESSION MODELS

Algorithm	R2
KNN	99,30%
Random Forest	99,40%
SVM	98,50%
XGBoost	99,40%
Gradient Boosting	99,40%
Neural Network	74,10%
Decision Tree	99%
Ridge	74%
Lasso	94,80%
Elastic Net	86,50%
SMRM	98,90%

TABLE VIII. MAPE BY MULTIPLE REGRESSION MODELS

Algorithm	RMSE
KNN	-2,043
Random Forest	-1,997
SVM	-3,162
XGBoost	-2,028
Gradient Boosting	-1,996
Neural Network	-7,721
Decision Tree	-2,538
Ridge	-7,725
Lasso	-4,351
Elastic Net	-6,071

TABLE IX. MAPE BY INDIVIDUAL REGRESSION MODELS

Algorithm	MAE
KNN	-1,303
Random Forest	-1,25
SVM	-1,77
XGBoost	-1,258
Gradient Boosting	-1,249
Neural Network	-1,428
Decision Tree	-1,68
Ridge	-1,42
Lasso	-1,39
Elastic Net	-1,43

VI. CONCLUSION AND FUTURE WORK

Predicting crude oil prices is a difficult task that requires a nuanced understanding of a wide range of economic and political factors. The SMRM hybrid system proposed in this paper is an innovative approach that leverages machine learning to better capture the complex relationships between these factors and crude oil prices. By combining multiple models, SMRM can generate more accurate and reliable predictions, which can help investors and traders make more informed decisions. The experiment results illustrate that SMRM surpasses existing prediction models in terms of both accuracy and stability, making it a valuable tool for anyone interested in predicting crude oil prices. While there is still much work to be done to fully understand and predict stock market trends, SMRM represents a major step forward in this field, and holds the potential to revolutionize the approach to crude oil price prediction in the coming years. Predicting crude oil prices is a complex and challenging task, requiring a nuanced understanding of economic and political factors. The proposed SMRM hybrid system represents a significant improvement over existing approaches, as it can leverage both quantitative and qualitative factors to generate more accurate predictions. By learning from past data, the system can continually improve its forecasts, making it a robust and flexible forecasting tool that can support decision-making for a range of stakeholders, including investors and policymakers. Experiments demonstrate that SMRM outperforms existing models in terms of accuracy and stability, highlighting its potential as a powerful tool for predicting crude oil prices in the years to come. However, the proposed SMRM hybrid system offers a robust tool for predicting crude oil prices with heightened accuracy and reliability. However, there are still challenges to address, such as quantifying the impact of irregular factors like political risks and extreme weather events on crude oil prices. To address these challenges, future research will aim to incorporate these factors into the SMRM hybrid system and quantify their impact, leading to even more accurate predictions. With continued research and development, SMRM has the potential to revolutionize crude oil price prediction and help stakeholders make more informed decisions in the dynamic and complex world of stock market trading. In conclusion, the proposed SMRM hybrid system offers a promising solution for predicting crude oil prices, leveraging the power of machine learning, and combining multiple models to better capture the complex relationships between different factors. The experiments reveal that SMRM excels over existing models in both accuracy and stability, making it a valuable tool for investors, traders, and other stakeholders in the energy sector. The system can also be continually refined and improved by incorporating irregular factors like political risks and extreme weather events, which can help to better predict changes in crude oil prices. With further development, this approach could have important implications for supporting decision-making and risk management in the energy sector, enabling stakeholders to make more informed and effective decisions in the dynamic and complex world of stock market trading. By providing more accurate and reliable predictions of crude oil prices, the SMRM hybrid system has the potential to revolutionize how approach to predicting crude oil prices, providing valuable insights that can help to optimize decision-making and drive greater value in the energy sector.

REFERENCES

[1] Khashei, M., & Mahdavi Sharif, B. (2021). A Kalman filter-based hybridization model of statistical and intelligent approaches for exchange rate forecasting. Journal of Modelling in Management, 16(2), 579-601.

[2] C. Hamzacebi, "Improving artificial neural networks' performance in seasonal time series forecasting", Information Sciences 178 (2008), pages: 4550-4559.

[3] G.P. Zhang, "A neural network ensemble method with jittered training data for time series forecasting", Information Sciences 177 (2007), pages: 5329–5346.

[4] G.P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing 50 (2003), pages: 159–175.

[5] H. Park, "Forecasting Three-Month Treasury Bills Using ARIMA and GARCH Models", Econ 930, Department of Economics, Kansas State University, 1999.

[6] L.J. Cao and Francis E.H. Tay "Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting", IEEE Transaction on Neural Networks, Vol. 14, No. 6, November 2003, pages: 1506-1518.

[7] R. Lombardo, J. Flaherty, "Modelling Private New Housing Starts In Australia", Pacific-Rim Real Estate Society Conference, University of Technology Sydney (UTS), January 24-27, 2000.

[8] Ahmad, A., Javaid, N., Guizani, M., Alrajeh, N., & Khan, Z. A. (2016). An accurate and fast converging short-term load forecasting model for industrial applications in a smart grid. IEEE Transactions on Industrial Informatics, 13(5), 2587-2596.

[9] Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics, 30(5), 2623-2635.

[10] Xiong, T., Bao, Y., & Hu, Z. (2013). Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices. Energy Economics, 40, 405-415.

[11] Kumar, M. S. (1992). The forecasting accuracy of crude oil futures prices. Staff Papers, 39(2), 432-461.

[12] Liu, Jinlan, Yin Bai, and Bin Li. "A new approach to forecast crude oil price based on fuzzy neural network." Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on. Vol. 3. IEEE, 2007.

[13] Alizadeh, A., and Kh Mafinezhad. "Monthly Brent oil price forecasting using artificial neural networks and a crisis index." Electronics and Information Engineering (ICEIE), 2010 International Conference On. Vol. 2. IEEE, 2010.

[14] Safari, A., & Davallou, M. (2018). Oil price forecasting using a hybrid model. Energy, 148, 49-58.

[15] Yi, Yao, and Ni Qin. "Oil price forecasting based on selforganizing data mining." Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on. IEEE, 2009.

[16] Nwulu, N. I. (2017, September). A decision trees approach to oil price prediction. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-5). IEEE.

[17] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

[18] Tang, L., Pan, H., & Yao, Y. (2018, March). K-Nearest Neighbor Regression with Principal Component Analysis for Financial Time Series Prediction. In Proceedings of the 2018 International Conference on Computing and Artificial Intelligence (pp. 127-131).

[19] Tang, L., Pan, H., & Yao, Y. (2018). PANK-A financial time series prediction model integrating principal component analysis, affinity propagation clustering and nested k-nearest neighbor regression. Journal of Interdisciplinary Mathematics, 21(3), 717-728

[20] Zhang, Y., He, J., & Yin, T. F. (2012). Research on petroleum price prediction based on SVM. Computer Simulation, 29(3), 375.

[21] Yu, L., Zhang, X., & Wang, S. (2017). Assessing potentiality of support vector machine method in crude oil price forecasting. Eurasia Journal of Mathematics, Science and Technology Education, 13(12), 7893-7904.

[22] Kumar, Y. J. N., Preetham, P., Varma, P. K., Rohith, P., & Kumar, P. D. (2020, July). Crude Oil Price Prediction Using Deep Learning. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 118-123). IEEE.

[23] [23] Wang, S., Yu, L., & Lai, K. K. (2004, July). A novel hybrid AI system framework for crude oil price forecasting. In Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management (pp. 233-242). Springer, Berlin, Heidelberg.

[24] [24] Hafezi, R., Shahrabi, J., & Hadavandi, E. (2015). A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. Applied Soft Computing, 29, 196-210.

[25] Nguyen, H. V., Naeem, M. A., Wichitaksorn, N., & Pears, R. (2019). A smart system for short-term price prediction using time series models. Computers & Electrical Engineering, 76, 339-352.

[26] Cheng, Y., Yi, J., Yang, X., Lai, K. K., & Seco, L. (2022). A CEEMD-ARIMA-SVM model with structural breaks to forecast the crude oil prices linked with extreme events. Soft Computing, 26(17), 8537-8551.

[27] Kaymak, Ö. Ö., & Kaymak, Y. (2022). Prediction of crude oil prices in COVID-19 outbreak using real data. Chaos, Solitons & Fractals, 158, 111990.

[28] Wu, B., Wang, L., Wang, S., & Zeng, Y. R. (2021). Forecasting the US oil markets based on social media information during the COVID-19 pandemic. Energy, 226, 120403.

[29] Shen, Z. (2022, July). Optimal Oil-based Exotic Options Strategies Under the Background of War: An Empirical Study in the Context of the Russia-Ukraine Conflict. In 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022) (pp. 954-961). Atlantis Press.

[30] Sun, Y. (2022, July). The Impacts of Wars on Oil Prices. In 2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022) (pp. 167-170). Atlantis Press.

[31] Ha, L. T. (2022). Dynamic interlinkages between the crude oil and gold and stock during Russia-Ukraine War: evidence from an extended TVP-VAR analysis. Environmental Science and Pollution Research, 1-14.

[32] Yuan, X., & Li, X. (2021). Mapping the technology diffusion of battery electric vehicle based on patent analysis: A perspective of global innovation systems. Energy, 222, 119897.

[33] Wang, J., Zhou, H., Hong, T., Li, X., & Wang, S. (2020). A multi-granularity heterogeneous combination approach to crude oil price forecasting. Energy Economics, 91, 104790.

[34] Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia computer science, 167, 599-606.

[35] Zhang, J., Li, D., & Wang, Y. (2020). Predicting uniaxial compressive strength of oil palm shell concrete using a hybrid artificial intelligence model. Journal of Building Engineering, 30, 101282.

[36] Abdollahi, H. (2020). A novel hybrid model for forecasting crude oil price based on time series decomposition. Applied energy, 267, 115035.

[37] Bristone, M., Prasad, R., & Abubakar, A. A. (2020). CPPCNDL: Crude oil price prediction using complex network and deep learning algorithms. Petroleum, 6(4), 353-361.

[38] Abdollahi, H., & Ebrahimi, S. B. (2020). A new hybrid model for forecasting Brent crude oil price. Energy, 200, 117520.

[39] Wang, J., Lei, C., & Guo, M. (2020). Daily natural gas price forecasting by a weighted hybrid data-driven model. Journal of Petroleum Science and Engineering, 192, 107240.

[40] Yahoo Finance, https://finance.yahoo.com/quote/CL%3DF/history?p=CL%3DF