A novel ensemble method for the accurate prediction of the major oil prices in Tanzania

Godfrey Joseph Saqware[1,3] and Ismail, B[2].

[1]Deparment of Statistics, Mangalore University,
Karnataka, INDIA.
Email : [1,3]godjose70@yahoo.com
[2]Department of Statistics, Yenepoya (Deemed to be University),
Karnataka, INDIA
Email:[2]prof.ismailb@gmail.com
[3]Deparment of Statistics, University of Dar es Salaam,
Dar es Salaam, Tanzania.

Abstract: Global development relies much on oil to run different types of machines. Using oil to power many types of equipment is very important to world economic growth. The analysis of oil prices is crucial for the country's long-term stability. However, global monopoly producers, wars, and pandemics have contributed to the volatility of crude oil prices. As a result, the optimal prediction model for oil prices becomes crucial. The performance of several ensemble strategies on single traditional and machine learning models was examined in this study. We found that the weighted ensemble technique outperformed other ensemble and single models in predicting petrol and diesel prices in Tanzania based on four performance metrics. Furthermore, a spike in global oil prices necessitates global economic and political stability for non-oil-producing nations to avoid suffering the consequences. Finally, other ensemble approaches may be used and compared to predict the oil prices.

***Keywords*: Autoregressive Integrated Moving Average (ARIMA), Oil prices, Ensemble methods, Machine learning.**

(Article history: Received: 20th April 2022 and accepted 13th June2023)

1. Introduction

Higher oil prices are beneficial to alternative energy companies, but little is known about how sensitive their financial performance is to changes in oil prices. Concerns about energy security and growing environmental awareness are contributing factors to an increase in concern about the future of the oil industry[1]. Several variables, including several years of unexpectedly high unconventional oil production, declining global demand, a dramatic change in OPEC policy, the reduction of some geopolitical threats, and an increase in the dollar, have contributed to the fluctuation of the oil prices. However, it is impossible to determine the relative relevance of each issue and the rapid increase in oil supply from unconventional sources [2]. This section explains the global and Tanzania oil prices perspectives.

1.1 The global oil prices perspectives

The Organization of Petroleum Exporting Countries (OPEC) influences oil demand and supply. The interaction of invariant and variable components determines oil price volatility. Some invariants that determine oil prices are feedstock prices, exploration costs, drilling expenses, oil chemical composition, production costs, distribution costs, marketing costs, and packaging and storage costs. Global economic activity, production, consumption levels, the value of the US dollar ($), current supply and demand, geopolitical factors, weather-related shifts, and political events are all variables [3]. However, the total connection of the entire system changes considerably, with a high degree of integration in the global crude oil markets over time. Each recognized price contributes uniquely to the global crude oil market's information flow of return and volatility networks[4]. During crises, higher levels of volatility are associated with disruptions in oil supply and demand and higher volatility persistence during financial/economic crises, suggesting that volatility persistence is a critical element in global economic and financial instability [5]. Through the supply and demand framework and cointegration theory, the cointegrating relationship between crude oil prices and global economic activity demonstrates that fluctuations in the Kilian economic index significantly impact crude oil prices through long-run and short-run equilibrium [6]. Liquidity has exponentially increased in recent years, resulting in a statistically significant increase in world oil prices[7]. The long-standing dynamic relationship between global economic activity and crude oil prices has been trending upwards in recent years. The impact of various oil price indices and confounding variables on oil market investors, such as geopolitical risk, armed conflicts, economic policy uncertainty, and equity market uncertainty, has proven to be significant [8]. The impact of oil prices on global economic transformation can be better understood by looking at products and services through global value chains (GVC). Furthermore, transportation has an important structural role in shaping GVCs, illustrating the long-term impact of energy costs on the global economy's structure and interconnection. Understanding the role of oil in a globalized economy is critical for decoupling economic growth from energy expansion and moving toward a low-carbon economy [9]. Real spot gasoline and heating oil prices were revealed to be responsive to oil-specific demand shocks rather than structural oil supply shocks. It's assumed that oil specific demand shocks, like Granger caused spot gasoline and heating oil prices, are the main drivers of spot oil prices [10].

1.2 The Tanzania oil price perspectives

Rising oil prices, power shortages, and political unrest are major causes of GDP drop in emerging nations. Therefore, it causes irregular and unbalanced in the developing countries' Gross Domestic Product (GDP). Tanzania's GDP as a developing country is influenced by consumption based on government and household final expenditures and exports[11]. Oil plays an important role in Tanzania's day-to-day economic activities. Oil imports have been accompanied by imported inflation and a persistent trade imbalance.

As a result, Tanzania must choose biofuels and speed up the discovery and exploitation of recognized oil and gas reserves [12]. There is no uniform rule for net oil importers in Sub-Saharan Africa, including Botswana, Kenya, and Tanzania. The Botswanan pula strengthens versus the US dollar, Kenyan shilling, and Tanzanian shilling after an oil price peak[13]. Tanzanian enterprises have a long way to go before competing with demanding oil and gas customers. Government regulations and local content legislative requirements imposed on international oil companies (IOCs) impact local participation. It also depends on local firms' willingness and capacity to upgrade to international standards in the petroleum sector[14]. Tanzania's major gas discoveries have significantly influenced the political economy. International policy debate during the previous decade risks skewing our understanding of these fundamentally political processes. Examining the methods and focusing on the aspects that impact contract negotiations could help to improve Tanzania's energy sector[15]. Tanzania's offshore natural gas deposits have been identified. As a result, the government established local content policies (LCPs). The study identified widespread skepticism, which impacted their suggestions for policy to include in the LCP. One argument is that the administration dominated policymaking and did not engage in public dialogues [16].

This paper is divided into five sections. Section two discusses related literature on ensemble models for predicting oil prices; Section three describes the study's material, methods, and ensemble approaches. Section four discusses the obtained results. Finally, section five will provide a conclusion and recommendations for further studies.

2. Review of the related literature

Time series forecasting accuracy is a prominent issue in academics, having applications in various industries. In recent years, machine learning and soft computing communities have shown interest in ensemble techniques. Ensemble-based algorithms such as Bagging, Adaboost, Negative Correlation, and combination rules have shown better results [17]. A large body of research suggests that combining predictions improves forecasting accuracy and robustness. Experiments show that mode ensembles may automate neural network models and overcome data sampling uncertainty, neural network training stochasticity, and prediction distribution [18]. Ensembles are commonly regarded as one of the most successful methods for problem prediction. According to earlier theoretical studies of ensembles, ensemble members' variance is one of the key causes of this performance [19]. Combining time-series prediction models from numerous models is better than relying on a single model. A combined prediction enhances total accuracy greatly and is frequently better than the forecast of each component model[20]. In recent years, Deep Extreme Learning Machine (DELM) models such as Hierarchical ELM (H-ELM) and deep representations learning using Extreme Learning Machine (ELM) have been proposed for usage in a variety of machine learning applications. Combining a deep learning scheme with an ensemble pruning paradigm and a unique double-deep ELMs ensemble system (DD-ELMs-ES) outperforms single models tackling time series forecasting issues [21]. The Long Short-Term Model (LSTM) ensemble technique has shown to be superior to other ensemble strategies. Furthermore, several LSTM models' weights can be dynamically changed and upgraded using a novel approach to create a composite prediction output that captures nonlinear statistical aspects in time series, enhancing accuracy [22]. A layered ensemble architecture (LEA) for time series forecasting issues consists of two layers, each incorporating a multilayer perceptron (MLP) network that outperforms other ensemble and nonensemble techniques in terms of prediction accuracy [23]. An ensemble of three approaches was proposed: decision trees, gradient boosted trees, and random forest. The ensemble's weights were calculated using the weighted least squares technique. Both dynamic and static ensembles outperformed individual models [24]. A boosting deep learning technique based on the extended AdaBoost algorithm is given for producing various fundamental predictors. A novel dynamic error correction strategy improves fundamental predictors even further. The essential predictors are combined using a stacking ensemble approach using kernel ridge regression as the meta predictor to get the final predicting results. The suggested multi-population non-dominated sorting genetic algorithm-II is for ensemble pruning, which improved forecasting accuracy and stability [25]. Bayesian optimization dynamic ensemble (BODE) overcomes the limits of single model-based methods. It produces a dynamic ensemble prediction combination for time series with time-varying underlying patterns. The BODE approach included ten alternative model options, including statistical methodologies and deep neural networks. The weights of the different combinations were altered based on the Bayesian optimization algorithm (BOA) and hyperparameter (HP) tuning, which gave the prediction performance greater flexibility and generality [26].

According to the examined literature, the ensemble forecast techniques are currently understudied, particularly machine learning models that have lately been developed. Therefore, the proposed study combines the Autoregressive Integrated Moving Average (ARIMA) model with machine learning algorithms to assess prediction performance. The ensemble of the single models will be done using mean, median, and weighted ensemble techniques.

3. Methodology

The data, predictive models, and ensemble methods for the chosen models are all described in this section.

3.1 Data

This paper considers monthly recorded data of the oil prices collected from the Energy and Water Utilities Regulatory Authority (EWURA), the government of the United Republic of Tanzania. The data covers the time between January 2005 and December 2021. The analysis of the two datasets is important due to climate change, war, and the global coronavirus pandemic, which ultimately affects the worldwide demand and supply of oil. The data was divided into 80% and 20% training and testing sets. Fig. 1 shows the data partition, the blue colour indicates the training sets, and the red line indicates the testing set.

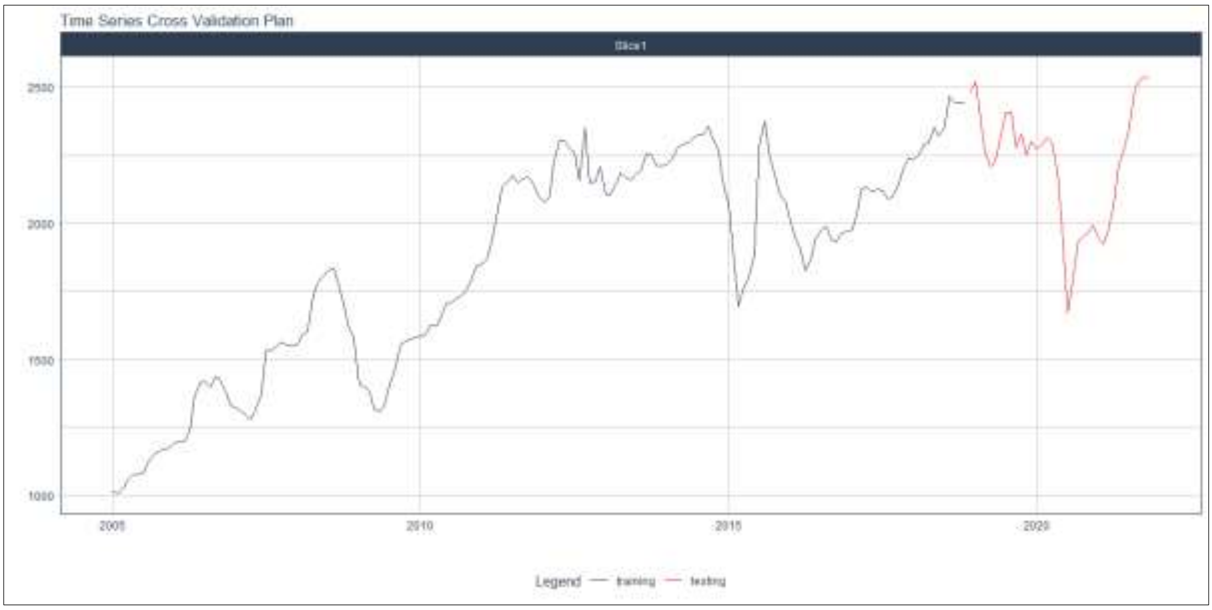

Figure 1: Data partitioning

3.2 Predictive models

Effective predictive models unquestionably aid in timely and proper decision-making [27]. Therefore, an effective predictive model that can handle oil prices is required. Given the abundance of predictive models now in use, it is necessary to evaluate their effectiveness to choose the most suitable and practical model for use in a given real-world scenario. This sub-section explains individual predictive models and the proposed ensemble techniques for the selected models.

3.2.1 Autoregressive Integrated Moving Average Model (ARIMA) Model

The Autoregressive Moving Average (ARMA) model is the combination of the Autoregressive (AR) and Moving Average (MA) Models. The autoregressive model predicts the current value based on the previous observations $w_{t-1}, w_{t-2}, ..., w_{t-p}$, while the Moving Average model predicts the current value using the residuals of the previous values $\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-p}$.

The AR model of order p is given by:

$$w_t = \psi_1 w_{t-1} + \psi_2 w_{t-2} + ... + \psi_p w_{t-p} + \varepsilon_t \tag{1}$$

The model in Eqn. (1) can be re-written as:

$$w_t - \psi_1 w_{t-1} - \psi_2 w_{t-2} - ... - \psi_p w_{t-p} = \varepsilon_t \tag{2}$$

By applying the backward shift operator $B^j w_t = w_{t-j} \{ j = 1, 2, 3, ..., p \}$, the model in Eqn. (2) can be re-written as:

$$w_t - \psi_1 B w_t - \psi_2 B^2 w_t - \ldots - \psi_p B^2 w_t = \varepsilon_t$$
$$(1 - \psi_1 B - \psi_2 B^2 - \ldots - \psi_p B^2) w_t = \varepsilon_t \tag{3}$$
$$\psi(B) w_t = \varepsilon_t$$

The MA model of order q is given by:

$$w_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q} \tag{4}$$

By applying the backward shift operator $B^j \varepsilon_t = \varepsilon_{t-j} \{ j = 1, 2, 3, \ldots, q \}$, the model in Eqn. (4) can be re-written as: -

$$w_t = \varepsilon_t + \phi_1 B \varepsilon_t + \phi_2 B^2 \varepsilon_t + \ldots + \phi_q B^q \varepsilon_t$$
$$w_t = (1 + \phi_1 B + \phi_2 B^2 + \ldots + \phi_q B^q) \varepsilon_t \tag{5}$$
$$w_t = \phi(B) \varepsilon_t$$

The $ARMA(p,q)$ model is used for forecasting a stationary time series, and it's given in Eqn. (6) below.

$$w_t = \psi_1 w_{t-1} + \psi_2 w_{t-2} + \ldots + \psi_p w_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q} \tag{6}$$

Now, supposed $w_t \sim I(d)$ is the stationary series after d^{th} the order of differencing, we say $z_t = (1-B)^d w_t$, and $ARIMA(p,d,q)$ can be written as:

$$(1-B)^d w_t = z_t = \psi_1 z_{t-1} + \psi_2 z_{t-2} + \ldots + \psi_p z_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q} \tag{7}$$

We are introducing the backward shift operator as in Eqn. (3) and (6), Eqn. (8) can be written as:

$$[1 - \psi_1 B - \psi_2 B^2 - \ \ldots \ - \psi_p B^p][1-B]^d w_t = [1 + \phi_1 B + \phi_2 B^2 + \ \ldots \ + \phi_q B^q] \varepsilon_t$$
$$\psi(B)[1-B]^d w_t = \phi_1(B) \varepsilon_t \tag{8}$$

The above-postulated model in Eqn. (9) works only when there is no seasonal variation. If the variable contains a seasonal component, then it is called Seasonal ARIMA or SARIMA model. The $SARIMA(p,d,q)*(P,D,Q)_S$ incorporates both non-seasonal and seasonal components in a multiplicative model as given in Eqn. (9) below.

$$\psi(B) \Psi(B^S)(1-B)^d (1-B^S)^D w_t = \phi(B) \Phi(B^S) \varepsilon_t \tag{9}$$

Where:

Ψ and ψ are the AR parameters for the seasonal and non-seasonal components, respectively, while Φ are ϕ the MA parameters for the seasonal and non-seasonal components respectively, B is the backward operator d, D and are the differencing terms.

3.2.2 Facebook FB Prophet Model

The FB Prophet is a developed time series prediction model that is both simple and practical without much data pre-processing[28]. The FB Prophet model can fit quickly and automatically fill in the missing values. Furthermore, the FB Prophet model allows for flexible adjustment of periodicity, making it suited for various applications. The FB Prophet time series forecasting model comprises trend, cyclic, festival, and error items. The model's composition is as follows:

$$y(t) = g(t) + s(t) + h(t) + e(t) \tag{10}$$

Where, $g(t)$ is the trend function, $s(t)$ is the periodic term function, $h(t)$ is the impacts from the holidays, and the $e(t)$ is the error term.

The above can be expanded as:

i) The `$g(t)$ can either take the form of the logistic growth model or the piece-wise regression model as follows: -

a) For the logistics growth model, $g(t) = \frac{c}{1+e^{(-k(t-m))}}$, C is the saturation value, k is the growth rate, and m is the bias parameter.

b) For the piece-wise regression, $y = \begin{cases} \beta_0 + \beta_1 x & x \le c \\ \beta_0 - \beta_2 c + (\beta_1 + \beta_2)x & x > c \end{cases}$, where c is the trend change point.

ii) The periodic term function $s(t)$ is based on the Fourier series, which results in periodic effects on model flexibility (Harvey & Shephard, 1993). With a conventional Fourier series, the approximate arbitrary smooth seasonal effects is as follows:

$$s(t) = \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right],$$

t represent periods and P is the regular periodic length of the time series; it can be daily, monthly, or yearly.

iii) The impacts from the holidays, $h(t) = Z(t)\mathrm{K}_i;\ \ \mathrm{K}_i \sim N(0,\sigma)$,

$Z(t) = \left[\mathrm{I}(t \subset D_1), \ldots, \mathrm{I}(t \subset D_i), \ldots., \mathrm{I}(t \subset D_L)\right]$, i represents holidays; D represents the collection of past and future holidays; K represents the impact of each holiday on the forecast.

3.2.3 Extreme Gradient Boosting Regression

Extreme Gradient Boosting (XGBOOST) is a model that has been consistently refined and enhanced by several scientists' research[29]. The model is based on Boosting Tree models and is a learning framework. The classic Boosting Tree models use only the first derivative information.

Because the residual of the previous $n-1$ trees is used while training the n^{th} tree, distributed training is challenging to execute. The XGBOOST conducts a second-order Taylor expansion on the loss function and may automatically employ the CPU's multithreading for parallel computation.

Integrate the tree model with the addition method, assuming a total of K trees, and use F to represent the basic tree model, then:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{11}$$

The objective function is given in Eqn. (12).

$$L = \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{12}$$

Where the loss function L denotes the difference between the predicted and actual values; Ω is the regularisation function that prevents overfitting.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{13}$$

Where T and w denote the number of leaves per tree and the weight of each tree's leaves, respectively. Finally, after each split following the second-order Taylor expansion of the objective function, the derived information gain of the objective function is provided in Eqn. (14).

$$Gain = \left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda} \right] - \gamma \tag{14}$$

A splitting threshold γ in (14) prevents overfitting and inhibits the model's tree development. The leaf node can only split if the information benefit is larger than γ. This is the same as pre-pricing the tree while optimizing the objective function.

3.2.4 The Elastic Net Regression model

Elastic Net is a regularised regression model that includes L_1 (lasso) and L_2 (ridge) regression penalties. The discovered constraint and limitation of lasso regression is the inability to choose the number of predictors. When the Elastic Net is employed alone, it becomes the ridge regression since it contains the lasso regression penalty. Thus, the ridge regression coefficient is first determined in the technique of regularisation with an elastic net. Then, the lasso technique is applied to the ridge regression to shrink the coefficient.

$$L_{enet} = \frac{\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}\right)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}\left|\hat{\beta}_j\right|\right) \tag{15}$$

In addition, Elastic Net allows setting and picking a lambda value. If we set $\alpha = 0$ or $\alpha = 1$, the model corresponds to ridge and lasso, respectively. Furthermore, if alpha we set $\alpha = 0$ and $\alpha = 1$, the penalty function reduces to L_1 (ridge) and L_2 (lasso) terms, respectively. The optimization of the elastic net regression model requires alpha values between 0 and 1.

3.2.5 Neural Network Autoregression (NNAR) model

The Neural Network Autoregression $NNAR(p,k)$ Model comprises p components that denote the numbers of lagged values and hidden nodes, respectively. If the dataset contains seasonality, the model becomes $NNAR(p,P,k)$, where P indicates the number of seasonal lags. The value p is chosen based on the Akaike Information Criterion (AIC). Since the Neural nets have an inherent random component, the model should be run several times (20 is the minimum requirement). The $NNAR(p,P,k)$ model is given in Eqn. (16) below.

$$y(t) = f\left(y_{t-1} + y_{t-2} + ... + y_{t-d}\right) + e_t \tag{16}$$

3.3 Ensemble methods for the selected models

Suppose M_1, M_2, M_3, M_4 and M_5 represent ARIMA, FB Prophet, XGBOOST, Elastic Net, and NNAR models. The three ensemble methods will be defined as follows: -

(a) Simple average ensemble method

The simple average takes the prediction from the five models M_1, M_2, M_3, M_4 and M_5. Then Eqn. (17) shows the simple average ensemble method.

$$f_c = (f_{M_1} + f_{M_2} + f_{M_3} + f_{M_4} + f_{M_5}) / 5 \tag{17}$$

(b) Median ensemble method

Since there is an odd number of models, the formula for the median ensemble method is given in Eqn. (18).

$$f_c = f\left(\frac{P}{2} + 0.5\right) \tag{18}$$

(c) Weighted ensemble method

The weighted ensemble approach uses the average of the predicted technique or provides varying weights to distinct predictions based on their performance. A superior single prediction model receives a higher loading, Eqn. (19) gives the weighted ensemble formula.

$$f_c = \sum_{j=1}^{5} W_j M_j, \; j = 1,2,3,4,5 \tag{19}$$

The loading weight W_j is assigned to the predicted model M_j and f_c is the weighted ensemble prediction model.

3.4 Performance Metrics

The metrics help to identify the best statistical-performing model. In this case, we use the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scale Error (MASE), and Root Mean Square Error (RMSE) criterion to select the best single and ensemble models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_t - \hat{y}| \tag{20}$$

$$MASE = \frac{|y_t - \hat{y}|}{\frac{1}{n}\sum_{i=1}^{n}|y_t - \hat{y}|} \tag{21}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_t - \hat{y}}{\hat{y}}\right| *100 \tag{22}$$

$$RMSE = \frac{1}{n}\sqrt{\sum_{1=1}^{n}(y_t - \hat{y})^2} \tag{23}$$

4. Results and discussion

This section analyses oil prices and compares the individual or single models with the proposed ensemble model. This shows whether combining various models enhances their capacity for prediction.

4.1 A plot of the oil prices

Since 2005, the price of gasoline and diesel has been fluctuating. Oil prices fell due to the 2008 economic downturn and the coronavirus, but prices rose again in 2021. Many sectors in Tanzania are influenced by oil prices, particularly gasoline and diesel. The agriculture and transportation sectors rely heavily on pricing for their everyday operations than the biofuel or production costs. The oil prices influence food pricing in developing nations due to transportation expenses [30]. The dynamic relationship between global crude oil prices and food price indices is examined using a nonlinear approach. The underlying regime depends on the fundamental split in global food and oil prices co-movement. Furthermore, when a threshold is reached, the adjustment process of the food price indices towards equilibrium is extremely durable and rises quicker than the oil price, as per the data[31]. Fig. (2) below shows the plot of the major oil prices in Tanzania from 2005 to 2021.

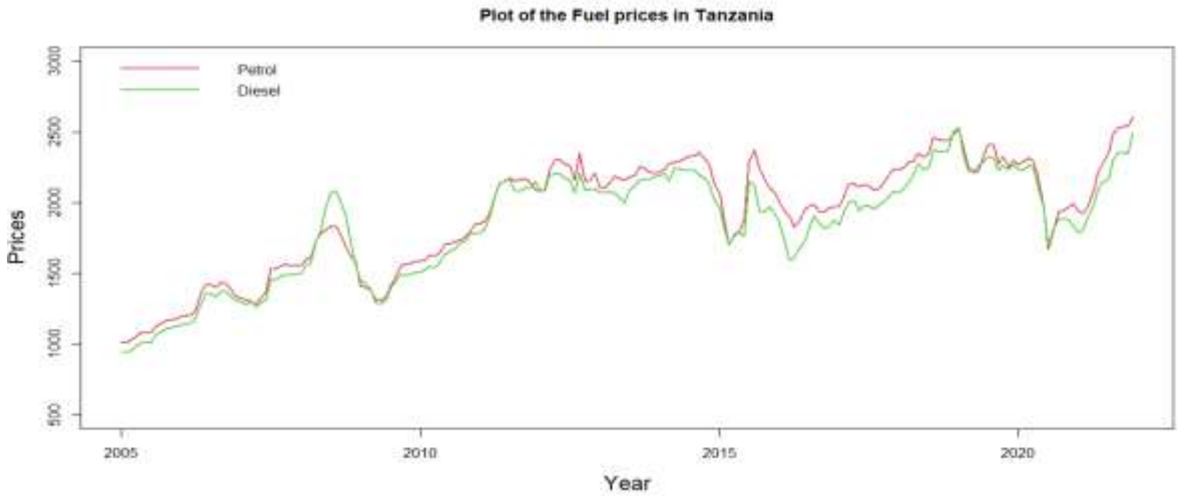

Figure 2: Plot of the oil prices in Tanzania

4.2 Single and ensemble models for petrol prices

In Table 1, based on the MAE, MAPE, MASE, and RMSE, the FB Prophet model with regressors outperformed other models in predicting the petrol prices in Tanzania. In Table 2, we carried out ensemble mean, median, and weighted methods. The result shows that the weighted ensemble method has improved the performance compared to the other ensemble methods. Furthermore, the weighted ensemble methods have also outperformed the best single model in predicting petrol prices in Tanzania.

Table 1:Result of the single models for the petrol prices

Model	Data set	MAE	MAPE	MASE	RMSE
ARIMA (1,1,0)	Test	247.17	12.2	3.36	314
ELASTIC NET	Test	340.29	16.62	4.63	415.62
XGBOOST	Test	219.41	10.85	2.98	285.27
NNAR (4,1,10) [12]	Test	228.16	10.16	3.1	267.4
PROPHET W/ REGRESSORS	Test	**185.59**	**8.77**	**2.52**	**229.07**

Table 2: Results of the ensemble models for the petrol prices.

Ensemble Model	Data set	MAE	MAPE	MASE	RMSE
ENSEMBLE (MEAN): 5 MODELS	Test	175.04	8.72	2.38	246.53
ENSEMBLE (MEDIAN): 5 MODELS	Test	219.25	10.84	2.98	284.99
ENSEMBLE (WEIGHTED): 5 MODELS	Test	**167.89**	**8.16**	**2.28**	**219.27**

4.3 Single and ensemble models for the prediction of the diesel prices

In Table 3, based on the MAE, MAPE, MASE, and RMSE, the NNAR (4,1,10) [12] model outperformed other models to predict diesel prices in Tanzania. In Table 4, we carried out ensemble mean, median, and weighed methods for the five models. The result shows that the weighted ensemble method has performed well compared to the other ensemble methods. The weighted ensemble methods have also outperformed the best single model in predicting the diesel prices in Tanzania.

Table 3: Result of the single model for the diesel prices

Model	Data set	MAE	MAPE	MASE	RMSE
ARIMA (1,1,0)	Test	249.27	12.69	4.05	317.08
ELASTIC NET	Test	281.31	14.25	4.57	360.51
XGBOOST	Test	190.17	9.8	3.09	265.23
NNAR (4,1,10) [12]	Test	**154.51**	**7.04**	**2.51**	**198.32**
PROPHET W/ REGRESSORS	Test	217.68	10.18	3.54	244.51

Table 4: Results of the ensemble model for the diesel prices.

Ensemble Model	Data set	MAE	MAPE	MASE	RMSE
ENSEMBLE (MEAN): 5 MODELS	Test	160.78	8.2	2.61	221.93
ENSEMBLE (MEDIAN): 5 MODELS	Test	193.91	9.95	3.15	267.92
ENSEMBLE (WEIGHTED): 5 MODELS	Test	**147.48**	**7.33**	**2.4**	**190.82**

Involving the weighted ensemble approach in predicting the oil prices has reduced the errors based on performance metrics. The simple addition ensemble technique outperforms benchmark models regarding prediction accuracy and effectiveness. Therefore, it is a viable tool for predicting complicated time series with significant volatility and irregularity, such as oil prices [32]. A unique decompose ensemble prediction approach that combines ensemble empirical mode decomposition (EEMD), and artificial neural network (ANN) has improved oil price prediction accuracy[33]. The research on ensemble neural networks for oil price prediction employing Feedforward, Recurrent, and Radial Basis Function networks using a weighted average ensemble approach outperformed other models[34]. A novel ensemble empirical mode decomposition method with adaptive noise (CEEMDAN) and extreme gradient boosting (XGBOOST) was proposed. The ensemble CEEMDAN-XGBOOST outperformed other single models predicting crude oil prices[35]. Therefore, the studies confirm that ensemble predictions of different models improve performance of the single model.

5. Conclusion

Superior results were obtained by combining machine learning prediction models such as XGBOOST, Elastic Net, NNAR, and FB Prophet with the ARIMA model. Among the three ensemble methods, the weighted ensemble approach outperformed the other two ensemble and single models in Tanzania's predicting petrol and diesel prices. The prediction of Tanzanian oil prices has improved due to adopting the weighted ensemble technique into single models. Other time-series data may be utilized to assess the efficiency of the weighted ensemble model for the five models. This study used only mean, median, and weighted ensemble approaches; however, more ensemble methods might be applied to investigate prediction performance.

References

[1] I. Henriques and P. Sadorsky, “Oil prices and the stock prices of alternative energy companies,” *Energy Economics*, vol. 30, no. 3, pp. 998–1010, 2008.

[2] J. Baffes, M. A. Kose, F. Ohnsorge, and M. Stocker, “The great plunge in oil prices: Causes, consequences, and policy responses,” *Consequences, and Policy Responses (June 2015)*, 2015.

[3] A. Demirbas, B. Omar Al-Sasi, and A.-S. Nizami, “Recent volatility in the price of crude oil,” *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 12, no. 5, pp. 408–414, 2017.

[4] D. Zhang, Q. Ji, and A. M. Kutan, “Dynamic transmission mechanisms in global crude oil prices: estimation and implications,” *Energy*, vol. 175, pp. 1181–1193, 2019.

[5] M. Zavadska, L. Morales, and J. Coughlan, “Brent crude oil prices volatility during major crises,” *Finance Research Letters*, vol. 32, p. 101078, 2020.

[6] Y. He, S. Wang, and K. K. Lai, “Global economic activity and crude oil prices: A cointegration analysis,” *Energy Economics*, vol. 32, no. 4, pp. 868–876, 2010.

[7] R. A. Ratti and J. L. Vespignani, “Why are crude oil prices high when global activity is weak?,” *Economics Letters*, vol. 121, no. 1, pp. 133–136, 2013.

[8] M. Dong, C.-P. Chang, Q. Gong, and Y. Chu, “Revisiting global economic activity and crude oil prices: A wavelet analysis,” *Economic Modelling*, vol. 78, pp. 134–149, 2019.

[9] F. Picciolo, A. Papandreou, K. Hubacek, and F. Ruzzenenti, “How crude oil prices shape the global division of labor,” *Applied Energy*, vol. 189, pp. 753–761, 2017.

[10] L. H. Ederington, C. S. Fernando, T. K. Lee, S. C. Linn, and H. Zhang, “The relation between petroleum product prices and crude oil prices,” *Energy Economics*, vol. 94, p. 105079, 2021.

[11] A. R. Kira, “The factors affecting Gross Domestic Product (GDP) in developing countries: The case of Tanzania,” 2013.

[12] B. B. Stambuli, “Price and income elasticities of oil demand in Tanzania: An autoregressive approach,” *Business Management Dynamics*, vol. 3, no. 1, p. 75, 2013.

[13] V. Pershin, J. C. Molero, and F. P. de Gracia, “Exploring the oil prices and exchange rates nexus in some African economies,” *Journal of Policy Modeling*, vol. 38, no. 1, pp. 166–180, 2016.

[14] W. Anderson, “Local suppliers in Tanzania: Ready for the petroleum sector?,” *The African Review: A Journal of African Politics, Development and International Affairs*, pp. 51–83, 2016.

[15] R. H. Pedersen and P. Bofin, *The politics of gas contract negotiations in Tanzania: a review*. JSTOR, 2015.

[16] A. Kinyondo and E. Villanger, “Local content requirements in the petroleum sector in Tanzania: A thorny road from inception to implementation?,” *The Extractive Industries and Society*, vol. 4, no. 2, pp. 371–384, 2017.

[17] H. Allende and C. Valle, “Ensemble methods for time series forecasting,” in *Claudio moraga: A passion for multi-valued logic and soft computing*, Springer, 2017, pp. 217–232.

[18] N. Kourentzes, D. K. Barrow, and S. F. Crone, “Neural network ensemble operators for time series forecasting,” *Expert Systems with Applications*, vol. 41, no. 9, pp. 4235–4244, 2014.

[19] M. Oliveira and L. Torgo, “Ensembles for time series forecasting,” in *Asian Conference on Machine Learning*, 2015, pp. 360–370.

[20] R. Adhikari, “A neural network based linear ensemble framework for time series forecasting,” *Neurocomputing*, vol. 157, pp. 231–242, 2015.

[21] G. Song and Q. Dai, “A novel double deep ELMs ensemble system for time series forecasting,” *Knowledge-Based Systems*, vol. 134, pp. 31–49, 2017.

[22] J. Y. Choi and B. Lee, “Combining LSTM network ensemble via adaptive weighting for improved time series forecasting,” *Mathematical Problems in Engineering*, vol. 2018, 2018.

[23] M. M. Rahman, M. M. Islam, K. Murase, and X. Yao, “Layered ensemble architecture for time series forecasting,” *IEEE transactions on cybernetics*, vol. 46, no. 1, pp. 270–283, 2015.

[24] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska, and F. Martínez-Álvarez, "Multi-step forecasting for big data time series based on ensemble learning," *Knowledge-Based Systems*, vol. 163, pp. 830–841, 2019.

[25] S. Zhang, Y. Chen, W. Zhang, and R. Feng, "A novel ensemble deep learning model with dynamic error correction and multi-objective ensemble pruning for time series forecasting," *Information Sciences*, vol. 544, pp. 427–445, 2021.

[26] L. Du, R. Gao, P. N. Suganthan, and D. Z. Wang, "Bayesian optimization based dynamic ensemble for time series forecasting," *Information Sciences*, vol. 591, pp. 155–175, 2022.

[27] B. P. BV and M. Dakshayini, "Performance analysis of the regression and time series predictive models using parallel implementation for agricultural data," *Procedia computer science*, vol. 132, pp. 198–207, 2018.

[28] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[30] B. M. Dillon and C. B. Barrett, "Global oil prices and local food prices: Evidence from East Africa," *American Journal of Agricultural Economics*, vol. 98, no. 1, pp. 154–171, 2016.

[31] S. Cheng and Y. Cao, "On the relation between global food and crude oil prices: an empirical investigation in a nonlinear framework," *Energy Economics*, vol. 81, pp. 422–432, 2019.

[32] L. Yu, W. Dai, and L. Tang, "A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting," *Engineering Applications of Artificial Intelligence*, vol. 47, pp. 110–121, 2016.

[33] Y. Ding, "A novel decompose-ensemble methodology with AIC-ANN approach for crude oil forecasting," *Energy*, vol. 154, pp. 328–336, 2018.

[34] L. A. Gabralla, H. Mahersia, and A. Abraham, "Ensemble neurocomputing based oil price prediction," in *Afro-European Conference for Industrial Advancement*, 2015, pp. 293–302.

[35] Y. Zhou, T. Li, J. Shi, and Z. Qian, "A CEEMDAN and XGBOOST-based approach to forecast crude oil prices," *Complexity*, vol. 2019, 2019.

Authors Profile

Mr. Godfrey Joseph Saqware, is a Research Scholar at the Department of Statistics, Mangalore University, India under the African Scholarship Scheme (MEA) from the Indian Council of Cultural Relations (ICCR). He completed Masters of Statistics at University of Dar Es Salaam, Tanzania in 2016. He joined the Department of Statistics, University of Dar Es Salaam, Tanzania as an Assistant Lecturer in 2016 to date. Before joining the Department of Statistics, Mr. Saqware worked with Ministry of Finance and Planning, Tanzania as a Statistician from 2014 to 2016. His research areas of interest are Financial Time Series, Econometrics, Bayesian Statistics, Simulation, Artificial Intelligent, Machine Learning, and Deep Learning.

Dr. Ismail, B is a retired Professor of Statistics in the Department of Statistics Mangalore University, Karnataka, India. Currently, Professor Ismail is teaching Biostatistics in the Department of Statistics, at YENEPOYA (Deemed to be University), Derlakatte, Mangalore, Karnataka, India. He completed his Ph.D. (1993) in Statistics at Mangalore University. He did his Post-Doctoral research (Honorary senior research fellow) at the Department of Statistics, University of Glasgow, Scotland-U.K (2000-2001).He got Commonwealth fellow in 2000 awarded by the association of commonwealth universities, U.K., “Franca e Diego de Castro” award for the research paper “Estimation of Error Correlation in Nonparametric Regression” in honor of Diego De Castro, Torino, Italy in 2004.He hold different position such as Chairman of the Department of Statistics, Mangalore University, the Life Member of the Indian Society for Probability and Statistics, Life member of Association of British Scholars, Life member of Indian Society of Geometrics, Life member of Kerala Statistical Association and so on. His areas of specialization are Econometrics, Non-parametric Regression, and Financial Time Series.