Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 8412895, 9 pages
https://doi.org/10.1155/2022/8412895

Research Article

Optimized Multivariate Adaptive Regression Splines for Predicting Crude Oil Demand in Saudi Arabia

Eman H. Alkhammash,[1] **Abdelmonaim Fakhry Kamel**,[2] **Saud M. Al-Fattah**,[3] **and Ahmed M. Elshewey**[4]

[1]*Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia*
[2]*Faculty of Graduate Environmental Studies, Ain Shams University, Cairo, Egypt*
[3]*Saudi Aramco, Dhahran, Saudi Arabia*
[4]*Faculty of Computers and Information, Computer Science Department, Suez University, Suez, Egypt*

Correspondence should be addressed to Eman H. Alkhammash; hms_1406@hotmail.com and Ahmed M. Elshewey; elshewy86@gmail.com

Received 4 November 2021; Revised 16 December 2021; Accepted 24 December 2021; Published 10 January 2022

Academic Editor: Jorge E. Macias-Diaz

This paper presents optimized linear regression with multivariate adaptive regression splines (LR-MARS) for predicting crude oil demand in Saudi Arabia based on social spider optimization (SSO) algorithm. The SSO algorithm is applied to optimize LR-MARS performance by fine-tuning its hyperparameters. The proposed prediction model was trained and tested using historical oil data gathered from different sources. The results suggest that the demand for crude oil in Saudi Arabia will continue to increase during the forecast period (1980–2015). A number of predicting accuracy metrics including Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) were used to examine and verify the predicting performance for various models. Analysis of variance (ANOVA) was also applied to reveal the predicting result of the crude oil demand in Saudi Arabia and also to compare the actual test data and predict results between different predicting models. The experimental results show that optimized LR-MARS model performs better than other models in predicting the crude oil demand.

1. Introduction

The development of prediction techniques and machine learning models is a critical task for crude oil demand [1]. The prediction techniques can predict different features in oil [2] including oil price, oil demand, oil viscosity, etc. Prediction models and techniques can present many advantages in energy sector such as energy planning, strategy formulation, and energy advancement. The design of prediction models and techniques is a complex task which has huge impacts for the economic trajectories of countries, energy companies, and other industrial sectors [3]. According to the International Energy Agency (IEA), the global demand for crude oil accounted for about 41% of the total fuel share in 2016. According to the Organization of the Petroleum Exporting Countries (OPEC), Saudi Arabia is one of the world's largest oil consumers, ranking fifth after Russia with a 3.4% share of global oil consumption in 2016.

There are numerous models that support the crude oil demand prediction, including autoregressive conditional heteroscedasticity (ARCH) model [4], other time series models, artificial neural networks [5], and fuzzy theory predictions [6, 7].

Machine learning models play an important role in the evaluation and prediction tasks. The features included in the dataset can be used to perform predictions. Machine learning models can also perform future predictions based on the available in the dataset [8]. Regression analysis is a

statistical process used to assess the relationship between various variables. In the field of machine learning, regression analysis models are widely used for predictions. The idea of the regression analysis is to show how the dependent variable (predicted variable) changes when one of the independent variables changes and other independent variables remain constant [9]. When the independent variables are restricted, regression analysis is used to obtain the average value of the dependent variable. There are three main processes for regression analysis which are (1) determining the strength of the predictors, (2) predicting an effect, and (3) trend prediction [10]. Many techniques have been presented in the field of regression analysis, which can be divided into parametric method and nonparametric method. In parametric method, the parameters contain all information about the data. The parameters contain all of the information required to predict the value of future data from the model. For example, in linear regression with a single variable, two parameters (intercept and coefficient) must be known in order to predict a new value. In nonparametric method, because more information is available, the ability to predict new values is more flexible because the parameters in the nonparametric method have infinite dimensions, and the data characteristics are superior to parametric models.

The purpose of this paper is to propose the LR-MARS model for predicting the demand for crude oil in Saudi Arabia. To improve the accuracy of the MARS model, social spider optimization is applied to improve the hyperparameters of the MARS model.

2. Related Work

This section outlines relevant studies in regard to Artificial Intelligent (AI) models for predicting the demand for crude oil. In [11], the authors proposed wavelet method to predict oil price in the long term. The proposed model can forecast the Brent oil price one year ahead. Several time series prediction approaches were compared to [11] model such as ARIMA, GARCH, and Holt-Winters. Result has shown that [11] model provides better prediction models than the other models. China's crude oil demand was predicted using soft and hard computing [12]. In [13], three estimated models for the price of petroleum called theories model, simulation model, and informal model were used. The informal estimate model performs better results than the other two models. The authors in [14] make use of eight artificial neural networks (ANN) and fuzzy regression (FR) for oil price prediction. The analysis of variance (ANOVA) and Duncan's multiple range test (DMRT) are then used to test the forecast produced by ANN and FR. The mean absolute percentage error (MAPE) was calculated for ANN models and the results have shown that ANN models outperform the FR models. For verification and validation purposes, the author have applied Spearman correlation test. The authors in [15] studied the factors that play a role in affecting the demand for oil in thirty developed countries using cointegration functions model. The variables used in the study were energy prices and national income. The result has shown strong relationship between income and the demand of energy and oil. In [16], distinct nine oil models were studied and compared. Oil price, gross domestic product (GDP), and time trend for improvement were considered among the most influencing factors of the models. A method that estimates coefficients was used in the comparison of econometric response of these models [16]. Another study focused on the markets of global crude oil and natural gas in the period 1918–1999 [17]. This study predicted price and income elasticities for crude oil, demand models, and natural gas supply [17].

Panel quantification analysis techniques were used to estimate long-term income and price elasticities in crude oil demand in the Middle East [18]. Data employed in the study covered the period 1971–2002. The result has shown high price inelasticity and slight income elasticity [18]. A prediction model for crude demand based on cointegration and a vector error correction model (VEC) is introduced [19]. Four main factors that affect the crude oil demand were considered: GDP, population growth, oil price, and the share of industrial sector in GDP. Both error correction model (ECM) and Johansen cointegration test were applied for the estimation of elasticities.

In [20], the International Energy Agency (IEA) proposed the scenarios for future oil demand for China in 2006 World Energy Outlook. The study concluded that the minimum statistical (lower bound) annual oil consumption in developed countries is 11 barrels per capita. [21]. Another study in [21] developed crude oil demand models that combines variance analysis and a flexible fuzzy regression model. The results demonstrated the superiority of fuzzy regression over the conventional model. The data used covered the period 1990–2005 for different countries: Japan, Canada, Australia, and United States [21]. In [22], the authors used data that covers the period 1981–2005. Input variables include population, GDP, oil imports, and export of oil. The study demonstrated the benefits of the optimization of particulate swarm (PSO) versus GA in estimating and predicting Iran's crude oil demand. In the domain of energy consumption prediction, another study [23] compared the performance of energy consumption prediction using conventional econometric and artificial intelligence-based models. The result reflected that AI-based models are robust and scalable for prediction. The results also showed that, in national level, the prediction of yearly energy consumption is preferred using conventional models. Moreover, nonlinear regression models obtain the lowest average MAPE (1.79%) for long-term prediction.

SSO has been successfully used to solve the continuous optimization problems [24]. In [24] the researchers adopted SSO and support vector regression as short-term electric load forecasting model. Results showed that SSO helps to achieve good results [24]. Another study in [25] used SSO algorithm to search for optimal cluster centers in fuzzy c-means clustering algorithm. The results showed that SSO improved the performance of fuzzy c-means clustering algorithm among other optimization algorithms [25]. Another study in [26] used SSO algorithm to solve discrete optimization problems. SSO was used for the problem of traveling salesman [26]. SSO was compared to eighteen

algorithms and the experimental results revealed that the performance of SSO algorithm in solving discrete problems was very useful for both low and middle-scale TSP datasets [26].

3. Materials and Methods

3.1. Linear Regression Model. On real-world data, linear regression model works perfectly. There are numerous advantages to using linear regression, such as the fact that the linear regression model in training is faster than many predictive models [27]. Linear regression is used to compute the strength of the relationship between the dependent variable and the independent variables, as well as to determine which independent variables have no relationship with the dependent variable and which independent variables contain redundant information about the dependent variable. Furthermore, linear regression models are simple to implement and use a small amount of memory [28]. If there is only one independent variable in a linear regression model, the regression function is a straight line; if there are two independent variables, the regression function is plane; and if there are n independent variables, the regression function is hyperplane with $n-$ dimensions [10]. If the actual values and predicted values are fitted, then the actual values will be similar to the predicted values. However, if there is a difference between the actual and predicted values, this difference is referred to as a cost, loss, or error. The regression function $\widehat{y}$ dependent on n independent (predictor) variables x_1,x_2, …, x_n is calculated using the following equation:

$$\widehat{y} = w_0x_0 + w_1x_1 + \cdots + w_nx_n + b. \tag{1}$$

Equation (1) represents how the value of $\widehat{y}$ varies with the independent x_1,x_2, …, x_n. w_0, w_1, …, w_n, where x_1, x_2, …, x_n. w_0, w_1, …, w_n are known as feature weights (model coefficients) and b is called a constant bias term (intercept).

3.2. Ridge Regression Model. Ridge regression is a model for multiple regression in order to perform data analysis. In ridge regression, the independent variables are highly correlated. Ridge regression model is used to avoid overfitting and to reduce the complexity of the model. New values that are predicted by ridge regression model give better results when the predictor variables are correlated [10]. Ridge regression model learns two parameters w, b by using the same standard of the least squares with adding a penalty term to make an appropriate variation for the parameter w. The penalty term in ridge regression is known as regularization in order to perform restriction to the model and reduce the overfitting, and also the coefficients of the regression are controlled using the regularization methods; this will reduce the sampling error and minimize the variance [29]. Also, L2 regularization is used for ridge regression model to minimize the residual sum of square (RRS) of the coefficients [29]. RSS for ridge regression can be expressed as in the following equation:

$$\mathrm{RSS}(w,b) = \sum_{i=1}^{N} \left(y_i - (wx_i + b)\right)^2 + \alpha \sum_{j=1}^{p} w_j^2, \tag{2}$$

where α is the penalty term. When the value of α is high, this means that the model is simple and more regularization. The penalty term α adjusts the parameters when the values of the parameters are high, so ridge regression minimizes the parameters to make the model simple and reduce the complexity of the model.

3.3. Multivariate Adaptive Regression Splines Model. MARS model is a nonlinear and nonparametric regression approach that uses piecewise linear splines to simulate the nonlinear relationship between the dependent and independent variables [30]. The MARS model is built as a linear combination of the following basis functions BFs showed in the following equation:

$$f(x) = \beta_0 + \sum_{i=1}^{m} \beta_i \mathrm{BF}_i, \tag{3}$$

where β_i, $i = 1, 2, \ldots..m$ are unknown coefficients that can be estimated using the least square method and m is the number of terms found in the final model using a forward backward stepwise process. BF_i is the $i-th$ basis function defined from piecewise linear basis functions and based on knot C. BF_i is calculated from the following set functions that is showed in the following equation:

$$\mathrm{BF}_i = \left\{ \left|x - C_i\right|^+, \ \left|C_i - x\right|^+ \right\}, \tag{4}$$

where $|x - C_i|^+$ and $|C_i - x|^+$ are given by

$$\left|x - C_i\right|^+ = \max\left(0, x - C_i\right), \tag{5}$$

$$\left|C_i - x\right|^+ = \max\left(0, C_i - x\right). \tag{6}$$

Finally, the predicted model is built with m numbers of BFs to provide the lowest generalized cross validation (GCV) value that is calculated by the following equation:

$$\mathrm{GCV} = \frac{\mathrm{SSE}i}{(1 - vmi/n)^2}, \tag{7}$$

where SSEi is the sum of square error, where SSE$i = \sum (O - f(x))^2$ and v is the smoothing parameter.

3.4. Analysis of Variance (ANOVA). ANOVA is a statistical analysis technique which is developed by R.A. Fisher in the 1920s. ANOVA can be used for many purposes such as comparing group mean. Two hypotheses are applied to determine the output of the comparison, namely, null hypothesis and alternative hypothesis. ANOVA is also known as analysis of an analysis of variance because it compares two variance estimations, namely, variation within groups and variation between groups. In this paper we perform a one-way ANOVA. The purpose of a one-way between-groups ANOVA is to show if there are any differences among the means of two or more groups/models. When at least two of

the groups/models have means that are significantly different from each other, the ANOVA test is significant in this case. However, it does not reveal which of the groups/models are different.

3.5. Social Spider Optimization Algorithm. The social spider optimization (SSO) is swarm intelligence-based metaheuristic algorithm [31]. SSO is chosen in this study because it is a new heuristic algorithm that solves difficult optimization problems. It is a vital model to search for the global optimum through performing a simulation to the social spider behavior. SSO mimics the behaviors of spiders. Spiders identify the position of prey via the vibration that occurred on the spider web. Any unusual vibration is a sign for the social spider to search for food and move into the source of vibration. The search area of SSO uses chain-like social spider structure. The direction of the food is determined by insects through signals generated through vibrations from the spider web. Equations (8) and (9) define the SSO operation.

The vibration intensity [32] at position x is calculated by the following equation:

$$I(x, x, \text{iter}) = \log\left(\frac{1}{F(x) - C} + 1\right), \tag{8}$$

where $F(x)$ denotes the cost function and C denotes a constant number.

The iteration attenuation is given by the following equation:

$$I(x_1, x_2, \text{iter}) = I(x_1, x_2, \text{iter}) \times \exp\left(-\frac{D(x_1, x_2)}{\sigma \times r_a}\right), \tag{9}$$

where $D(x_1, x_2)$ indicates the distance between x_1 and x_2. The standard deviation of all members along one searched dimension is indicated by σ. The free parameter is r_a.

3.6. The Proposed Prediction Model. This paper combines both LR model and MARS model based on SSO to develop an optimized LR-MARS prediction model that predicts crude oil demand. The proposed LR-MARS model is developed based on five main stages as demonstrated in Figure 1. There are five stages used to develop the LR-MARS model which are (1) data collection and data preprocessing stage, (2) determining training and testing sets, (3) LR model and MARS model, (4) using SSO, and (5) performance evaluation.

3.6.1. Data Collection and Preprocessing. The process of data collection starts with collecting different features for crude oil demand from different sources. Data are tracked and verified for any externality or inconsistency. For example, the gross domestic product (GDP) feature is gathered from the sources: OPEC, IEA, International Monetary Fund (IMF), Saudi Statistics Authority, and World Bank. The data used in this article come from various sources and cover the period 1980–2015 [3]. Features such as year, oil demand, GDP, population, Brent prices, Light-Duty Vehicles (LDV), and Heavy-Duty Vehicles (HDV) are shown in Table 1.

Table 1 describes a number of statistical metrics such as mean, standard error, median, standard deviation, etc., of the features of the dataset which are oil demand, GDP, population, LDV, and HDV. For instance, the maximum value of the oil demand is 3318.656317, the minimum value is 602, and the standard deviation is 774.0563839.

In statistics, the correlation matrix shows the correlation coefficients between variables. The correlation matric of the features of Saudi Arabia oil demand dataset is shown in Table 2. Each cell represents the correlation value between two variables. As can be seen in Table 2, the correlation coefficient of the features is closer to 1 which means that we have strong positive correlation between each two features in the dataset.

Data preprocessing stage is an essential step in machine learning [33]. The quality of the data can directly affect the ability of the models to learn; thus, it is critical that we preprocess our data before using data as inputs into the proposed model. In this paper, preprocessing is done using normalization. If the data contains input values with varying scales, normalization can be used to scale these values. Normalization scales each input value separately through subtracting the mean (centering) and dividing by the standard deviation in order to change the distribution's mean to zero and standard deviation to one [33]. Normalization is calculated using the following equation:

$$z = \frac{x - \mu}{\sigma}, \tag{10}$$

where x is the input value, μ is the mean value, and σ is the standard deviation value. Mean value (μ) is calculated using the following equation:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{11}$$

Standard deviation (σ) is calculated using the following equation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}. \tag{12}$$

3.6.2. Training and Testing Sets. The crude oil demand dataset is split into train data (90%) and test data (10%). Following that, the train data is split further into training set (50% of train data) and validation set (50% of train data).

3.6.3. LR Model and MARS Model. The training set (50% of train data) is trained by LR model and the validation set (50% of train data) is used as an input to the LR model to make predictions through LR model. LR model provides two predictions (validation prediction set and test prediction set). Finally, the validation prediction set will be trained with MARS model to create LR-MARS model. This LR-MARS is

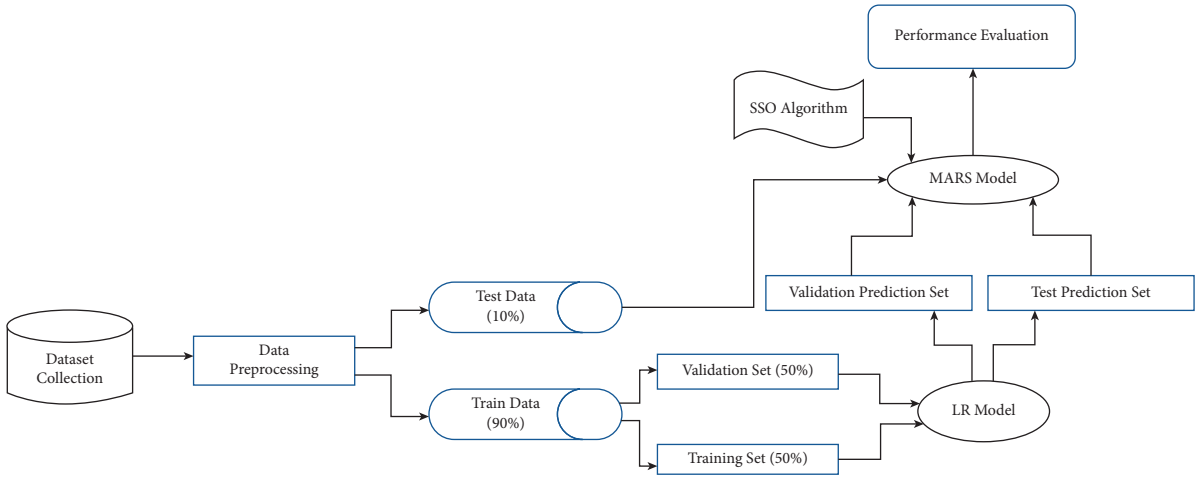

Figure 1: The stages for the proposed LR-MARS model.

Table 1: Statistical data analysis of features for Saudi Arabia oil demand.

Statistic	Oil demand (MBD)	GDP (bill SAR)	Population (MM)	Brent price ($/Bbl)	LDV (M)	HDV (M)
Mean	1539.087014	1481.705056	19.57855192	41.68838889	5390.042144	3587.345317
Standard error	129.0093973	79.40556539	1.079092403	5.189740084	578.5509528	288.0411458
Median	1230.609065	1359.74	19.2705	28.5755	4214.092	3085.9235
Standard deviation	774.0563839	476.4333924	6.474554418	31.13844051	3471.305717	1728.246875
Sample variance	599163.2855	226988.7774	41.91985491	969.6024771	12049963.38	2986837.26
Kurtosis	−0.247425069	−0.292910639	−1.010727991	0.214674645	−0.121742024	−0.158338512
Skewness	0.962853111	0.655422825	0.157041289	1.232271204	0.950014993	0.857854418
Minimum	602	778.227	9.32	12.713	1268.38	1199.523
Maximum	3318.656317	2545.24	31.016	111.62	13749.2784	7713.0792
Confidence level (95.0%)	261.9030003	161.2018679	2.190674043	10.53573249	1174.520876	584.7546137

Table 2: Correlation matrix of features for Saudi Arabia oil demand dataset.

	Oil demand (MBD)	GDP at 2010	Population	Brent prices	LDV	HDV
Oil demand (MBD)	1					
GDP at 2010	0.945006302	1				
Population pop	0.952639467	0.924158747	1			
Brent prices	0.84962017	0.834478322	0.747342982	1		
LDV	0.996837432	0.954042182	0.962077275	0.828231293	1	
HDV	0.994288957	0.944879459	0.970942636	0.807996732	0.998513235	1

used to make final predictions on the test prediction set to obtain the final predicted output that is in turn compared with the actual test data.

3.7. Performance Metrics. To validate the performance and effectiveness of the prediction models proposed, five error analysis criteria are introduced to evaluate the proposed models, as shown in equations (13)–(17), where y_{real_i} is the actual values, y_{pred_i} is the predicted values, and $\overline{y}$ is the mean value of actual values [24, 34]. For each model, the performance is evaluated using the Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-squared (R^2).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_{\text{real}_i} - y_{\text{pred}_i}\right|, \tag{13}$$

$$\text{MedAE} = \text{median}\left(\left|y_{\text{real}_1} - y_{\text{pred}_1}\right|, \ldots\ldots, \left|y_{\text{real}_N} - y_{\text{pred}_N}\right|\right), \tag{14}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{\text{real}_i} - y_{\text{pred}_i}\right)^2, \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\text{real}_i} - y_{\text{pred}_i}\right)^2}, \tag{16}$$

Table 3: Comparison of prediction performances using machine learning models.

Models	LR-MARS	LR	Ridge regression
MAE	0.024	0.042	0.055
MedAE	0.023	0.047	0.054
MSE	0.0007	0.0026	0.0036
RMSE	0.02	0.05	0.06
R^2	99.9%	99.6%	99.4%

Table 4: Comparison of prediction performances using LR-MARS model with different cases.

Models	Case 1	Case 2	Case 3
MAE	0.024	0.034	0.046
MedAE	0.023	0.041	0.048
MSE	0.0007	0.0038	0.0041
RMSE	0.02	0.031	0.036
R^2	99.9%	99.3%	99.1%

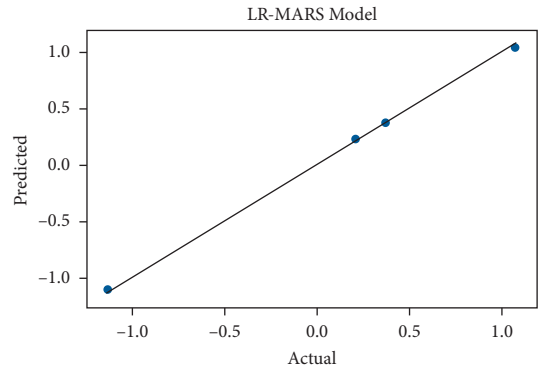

Figure 2: A cross-plot of the actual and predicted crude oil demand using LR-MARS model.

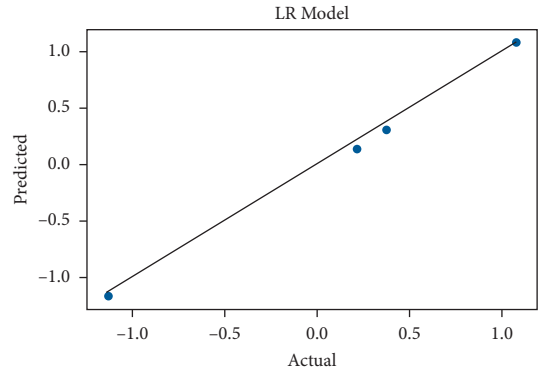

Figure 3: A cross-plot of the actual and predicted crude oil demand using LR model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{\text{real}_i} - y_{\text{pred}_i}\right)^2}{\sum_{i=1}^{N}\left(y_{\text{real}_i} - \overline{y}\right)^2}. \tag{17}$$

4. Results and Discussion

The implementation of the models is done using Google Colab notebook. Google Colab notebook helps to write and execute python in the browser, where it is an open source and widely used for the implementation of machine learning algorithms such as regression, classification, and clustering. To evaluate the performance of the optimized LR-MARS model in crude oil demand prediction more effectively, other models are chosen for comparison. Furthermore, the models commonly used in machine learning are chosen. SSO has been used to perform tuning to the two hyperparameters (penalty term and maximum number of basis functions (BFs)). The population of SSO metaheuristic algorithm consists of 30 members and the

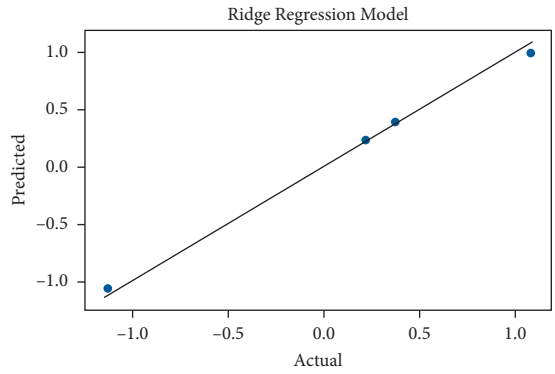

Figure 4: A cross-plot of the actual and predicted crude oil demand using ridge regression model.

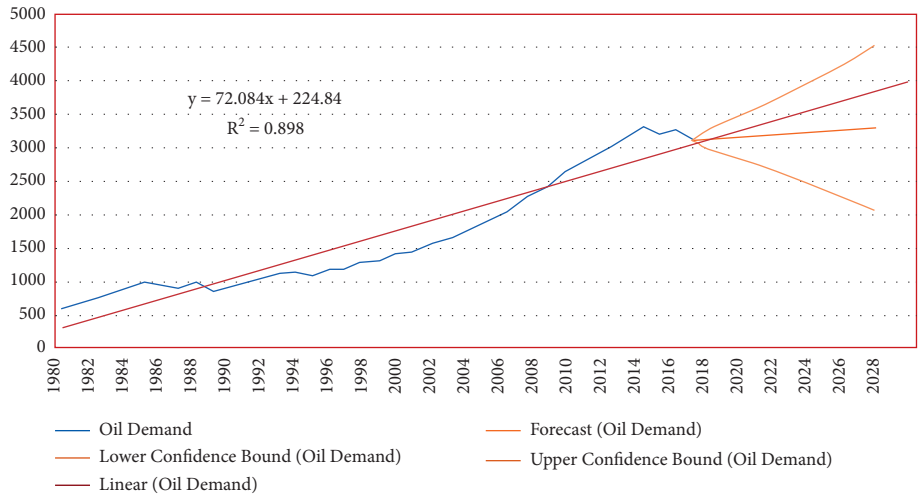

Figure 5: A cross-plot of the actual and predicted crude oil demand using ANOVA model.

Table 5: Comparison of prediction performance for LR-MARS optimized model and ANOVA model, respectively.

Models	LR-MARS	ANOVA
MSE	0.0007	0.0493
RMSE	0.02	0.2374
R^2 (%)	99.9%	89%

Table 6: Analysis of the source of variation.

Source of variation	Sum of square SS	Degree of freedom DF	Mean of square MS	F-value	P value	$F_{critical}$
Between groups	0.001588403	3	0.000529468	10.00246025	0.00982156	3.490294819
Within groups	2.582510317	12	0.215209193			
Total	2.58409872	15				

maximum number of generations is 100. The output of the optimization process is that the maximum number of basis functions (BFs) is 42 and the penalty term is 1.46. The prediction model proposed in this paper, which combines linear regression model with multivariate adaptive regression splines model, has shown high prediction accuracy when predicting crude oil demand in Saudi Arabia. To effectively evaluate the performance of LR-MARS in crude oil demand prediction, traditional prediction models of machine learning are used in this paper as comparative experiments. During the experiment, LR model and ridge regression model are used for crude oil demand prediction as comparative tests. To objectively evaluate and describe the performance of the three prediction models, the prediction error values of each model are calculated according to equations (13)–(17). The experimental results of MAE, MedAE, MSE, RMSE, and R^2 of the test data are shown in Table 3.

Among all the experimental models in Table 1, ridge regression model has the largest error, and its MAE, MedAE, MSE, RMSE, and R^2 are 0.055, 0.054, 0.0036, 0.06, and 99.4%, respectively. The MAE, MedAE, MSE, RMSE, and R^2 of LR model are 0.042, 0.047, 0.0026, 0.05, and 99.6%, respectively. The error of LR-MARS model with optimizing the two hyperparameters (penalty term and maximum number of basis functions (BFs)) using SSO algorithm is the smallest; its MAE, MedAE, MSE, RMSE, and R^2 are 0.024, 0.023, 0.0007, 0.02 and 99.9%, respectively, which is significantly lower than the other two models. It can be seen from Table 3 that LR-MARS model with optimizing the two hyperparameters (penalty term and maximum number of basis functions (BFs)) using SSO algorithm has a high accuracy in predicting crude oil demand and is more effective than the other models. Table 4 demonstrates a comparison of LR-MARS model with different cases: case 1, optimizing the two hyperparameters (penalty term and maximum number of basis functions (BFs)) using SSO algorithm, case 2, optimizing the one hyperparameter (penalty term) using SSO algorithm, and third, without optimizing any hyperparameter.

Figures 2–4 show a cross-plot of the actual and predicted crude oil demand using LR-MARS model, LR model and ridge regression model, respectively.

4.1. Analysis of Variance (ANOVA). In this section, we use ANOVA for two purposes. The first purpose is to predict the crude oil demand in Saudi Arabia. The second purpose is using ANOVA to compare the actual test data and the predicted data results between LR-MARS, LR, and ridge regression model, respectively. R^2 which is also known as coefficient of determination, is used to calculate how close the data are to the fitted regression line. The value ($R^2 = 0.898$) indicates a better fit for the model as shown in Figure 5.

4.2. ANOVA Predicting Result. ANOVA is used as a prediction model. Table 5 provides a comparison of ANOVA prediction model and the proposed LR-MARS optimized model. The results show that LR-MARS optimized model gives a high performance comparing to ANOVA model.

In Table 6, the analysis of the source of variation is carried out in two ways: between groups and within groups. Between-groups analysis determines the source of variance of LR-MARS, LR, and ridge regression models, respectively. Within-groups analysis identifies the experimental error between the group and itself. From the ANOVA results in Table 6, SS = 2.582510317, while Mean Square MS = 0.215209193. Therefore, we can conclude that the null hypothesis was rejected because $F_{\text{critical}} = 3.490294819$ and $F = 10.00246025$, where $F_{\text{critical}} < F$. Moreover, since the P value is less than 0.05 (i.e., $0.00982 < 0.05$), this is another indication of the significant differences in the attribute (crude oil demand) between LR-MARS, LR, and ridge regression models, respectively, and therefore is another evidence to reject the null hypothesis.

5. Conclusion

In this paper, a hybrid model called LR-MARS is developed for predicting the crude oil demand in Saudi Arabia. This paper used historical data of one of the world's largest oil producers (Saudi Arabia) to demonstrate the applicability and effectiveness of the proposed LR-MARS model. The dataset used in the LR-MARS consists of seven features: time, oil demand, GDP, population, Brent crude prices, LDV, and HDV. The LR-MARS model is a combination of linear regression model and multivariate adaptive regression splines (MARS) model. We also used SSO algorithm for optimizing two hyperparameters, namely, penalty term and maximum number of basis functions (BFs) for the MARS model. To evaluate the performance of LR-MARS optimized model, we used MAE, MedAE, MSE, RMSE, and R^2 to examine and test the predictions performance for the LR-MARS model that are 0.024, 0.023, 0.0007, 0.02, and 99.9%, respectively. We have also compared LR-MARS optimized model to other machine learning prediction models. The optimized LR-MARS model is more accurate in predicting crude oil demand in Saudi Arabia than other models. Moreover, we have used ANOVA as prediction model to predict the crude oil demand in Saudi Arabia and also to compare the actual test set and predicted results between LR-MARS, LR, and ridge regression models. This paper will be useful for oil demand planning, setting strategies, and future oil investments. Due to the limitation in obtaining some features and the inconsistency of scaling some data, these limitations of features will lead to a certain range of errors in data-processing process and prediction process. Therefore, other possible influencing features can be considered as input variable. As a direction of future work, as splines can be modelled by adding more knots, this will help in increasing the model flexibility. Moreover, cubic spline model and natural cubic spline model can be used to enhance the results.

Data Availability

The data used in this paper were obtained from different sources (OPEC, IEA, International Monetary Fund (IMF), Saudi Statistics Authority, and World Bank) and cover the period 1980 to 2015 [3].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge Taif University Researchers Supporting Project number (TURSP-2020/292), Taif University, Taif, Saudi Arabia, for funding this research.

References

[1] H. Duan, G. R. Lei, and K. Shao, "Forecasting crude oil consumption in China using a grey prediction model with an optimal fractional-order accumulating operator," *Complexity*, vol. 2018, Article ID 3869619, 12 pages, 2018.

[2] 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.

[3] S. M. Al-Fattah, "Application of the artificial intelligence GANNATS model in forecasting crude oil demand for Saudi Arabia and China," *Journal of Petroleum Science and Engineering*, vol. 200, Article ID 108368, 2021.

[4] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.

[5] S. Aiguo and L. Jiren, "Evolving Gaussian RBF network for nonlinear time series modelling and prediction," *Electronics Letters*, vol. 34, no. 12, pp. 1241–1243, 1998.

[6] A. Hatami-Marbini and F. Kangi, "An extension of fuzzy TOPSIS for a group decision making with an application to Tehran stock exchange," *Applied Soft Computing*, vol. 52, pp. 1084–1097, 2017.

[7] F. Gaxiola, P. Melin, F. Valdez, and O. Castillo, "Interval type-2 fuzzy weight adjustment for backpropagation neural networks with application in time series prediction," *Information Sciences*, vol. 260, pp. 1–14, 2014 Mar 1.

[8] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," *Cortex*, vol. 55, pp. 122–129, 2014.

[9] P. Aleksandar, P. Silvana, and Z. P. Valentina, "Multiple linear regression model for predicting bidding price," *Technics Technologies Education Management*, vol. 10, no. 3, pp. 386–393, 2015.

[10] A. M. Elshewey, "Machine learning regression techniques to predict burned area of forest fires," *International Journal of Soft Computing*, vol. 16, no. 1, pp. 1–8, 2021.

[11] Q. Liang, Y. Fan, and Y. M. Wei, "A long-term trend forecasting approach for oil price based on wavelet analysis," *Chinese Journal of Management Science*, vol. 13, no. 1, pp. 30–36, 2005.

[12] L. Wu, S. Liu, L. Yao, S. Yan, and D. Liu, "Grey system model with the fractional order accumulation," *Communications in Nonlinear Science and Numerical Simulation*, vol. 18, no. 7, pp. 1775–1785, 2013.

[13] J. Tuo and S. Yanbing, "Summary of world oil price forecasting model," in *Proceedings of the 2011 Fourth International Symposium on Knowledge Acquisition and Modeling*, pp. 327–330, IEEE, Sanya, China, October 2011.

[14] A. Azadeh, M. Moghaddam, M. Khakzad, and V. Ebrahimipour, "A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting," *Computers & Industrial Engineering*, vol. 62, no. 2, pp. 421–430, 2012.

[15] I. B. Ibrahim and C. Hurst, "Estimating energy and oil demand functions," *Energy Economics*, vol. 12, no. 2, pp. 93–102, 1990.

[16] H. G. Huntington, "OECD oil demand," *Energy Economics*, vol. 15, no. 1, pp. 49–56, 1993.

[17] N. Krichene, "World crude oil and natural gas: a demand and supply model," *Energy Economics*, vol. 24, no. 6, pp. 557–576, 2002.

[18] P. K. Narayan and R. Smyth, "A panel cointegration analysis of the demand for oil in the Middle East," *Energy Policy*, vol. 35, no. 12, pp. 6258–6265, 2007.

[19] J. Xiong and P. Wu, "An analysis of forecasting model of crude oil demand based on cointegration and vector error correction model (VEC),"in *Proceedings of the 2008 International Seminar on Business and Information Management*, vol. 1, pp. 485–488, IEEE, Wuhan, China, December 2008.

[20] W. P. Nel and C. J. Cooper, "A critical review of IEA's oil demand forecast for China," *Energy Policy*, vol. 36, no. 3, pp. 1096–1106, 2008.

[21] A. Azadeh, M. Khakestani, and M. Saberi, "A flexible fuzzy regression algorithm for forecasting oil consumption estimation," *Energy Policy*, vol. 37, no. 12, pp. 5567–5579, 2009.

[22] E. Assareh, M. A. Behrang, M. R. Assari, and A. Ghanbarzadeh, "Application of PSO (particle swarm optimization) and GA (genetic algorithm) techniques on demand estimation of oil in Iran," *Energy*, vol. 35, no. 12, pp. 5223–5229, 2010.

[23] N. Wei, C. Li, X. Peng, F. Zeng, and X. Lu, "Conventional models and artificial intelligence-based models for energy consumption forecasting: a review," *Journal of Petroleum Science and Engineering*, vol. 181, Article ID 106187, 2019.

[24] A. Sina and D. Kaur, "Short term load forecasting model based on kernel-support vector regression with social spider optimization algorithm," *Journal of Electrical Engineering & Technology*, vol. 15, no. 1, pp. 393–402, 2020.

[25] Q.-T. Bui, Q.-H. Nguyen, V. M. Pham et al., "A novel method for multispectral image classification by using social spider optimization algorithm integrated to fuzzy C-mean clustering," *Canadian Journal of Remote Sensing*, vol. 45, no. 1, pp. 42–53, 2019.

[26] B. A. Emine and E. Ülker, "Dıscrete socıal spıder algorıthm for the travelıng salesman problem," *Artificial Intelligence Review*, vol. 54, no. 2, pp. 1063–1085, 2021.

[27] A. K. Marandi and D. A. Khan, "An impact of linear regression models for improving the software quality with estimated cost," *Procedia Computer Science*, vol. 54, pp. 335–342, 2015.

[28] N. Aghdaei, G. Kokogiannakis, D. Daly, and T. McCarthy, "Linear regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings," *Energy Procedia*, vol. 121, pp. 79–86, 2017.

[29] J. M. Pereira, M. Basto, and A. F. D. Silva, "The logistic lasso and ridge regression in predicting corporate failure," *Procedia Economics and Finance*, vol. 39, pp. 634–641, 2016.

[30] B. Keshtegar, C. Mert, and O. Kisi, "Comparison of four heuristic regression techniques in solar radiation modeling: kriging method vs. RSM, MARS and M5 model tree," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 330–341, 2018.

[31] D. T. Vu, X.-L. Tran, M.-T. Cao, T. C. Tran, and N.-D. Hoang, "Machine learning based soil erosion susceptibility prediction using social spider algorithm optimized multivariate adaptive regression spline," *Measurement*, vol. 164, Article ID 108066, 2020.

[32] J. Q. James and V. O. Li, "A social spider algorithm for global optimization," *Applied Soft Computing*, vol. 30, pp. 614–627, 2015.

[33] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, Article ID 105524, 2020.

[34] M. Y. Shams, O. M. Elzeki, L. M. Abouelmagd, A. E. Hassanien, M. A. Elfattah, and H. Salem, "HANA: a healthy artificial nutrition analysis model during COVID-19 pandemic," *Computers in Biology and Medicine*, vol. 135, Article ID 104606, 2021.