energies

Article

An Optimized Gradient Boosting Model by Genetic Algorithm for Forecasting Crude Oil Production

Eman H. Alkhammash

Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; eman.kms@tu.edu.sa

Abstract: The forecasting of crude oil production is essential to economic plans and decision-making in the oil and gas industry. Several techniques have been applied to forecast crude oil production. Artificial Intelligence (AI)-based techniques are promising that have been applied successfully to several sectors and are capable of being applied to different stages of oil exploration and production. However, there is still more work to be done in the oil sector. This paper proposes an optimized gradient boosting (GB) model by genetic algorithm (GA) called GA-GB for forecasting crude oil production. The proposed optimized model was applied to forecast crude oil in several countries, including the top producers and others with less production. The GA-GB model of crude oil forecasting was successfully developed, trained, and tested to provide excellent forecasting of crude oil production. The proposed GA-GB model has been applied to forecast crude oil production and has also been applied to oil price and oil demand, and the experiment of the proposed optimized model shows good results. In the experiment, three different actual datasets are used: crude oil production (OProd), crude oil price (OPrice), and oil demand (OD) acquired from various sources. The GA-GB model outperforms five regression models, including the Bagging regressor, KNN regressor, MLP regressor, RF regressor, and Lasso regressor.

Keywords: gradient boosting model; genetic algorithm; oil production; oil price; oil demand

check for updates

Citation: Alkhammash, E.H. An Optimized Gradient Boosting Model by Genetic Algorithm for Forecasting Crude Oil Production. *Energies* **2022**, *15*, 6416. https://doi.org/10.3390/en15176416

Academic Editors: Xiangchun Wang, Gang Yao and Sanyi Yuan

Received: 19 August 2022
Accepted: 30 August 2022
Published: 2 September 2022

1. Introduction

The growth of AI-based research has demonstrated that it has the potential to be a future path for all disciplines. In many sectors of the economy (including communication, e-commerce, etc.), AI is already widely employed, but there is still more work to be done in the oil sector. One of the big challenges is the vast amounts of data, data from several sources in complicated formats, data with high dimensions and coupling, unstructured data, etc. [1].

Machine learning and deep learning techniques have the potential for application in the oil and gas sector and can cover several aspects of the oilfield, such as oil price and oil demand [2].

The traditional approach for forecasting oil production is the Numerical reservoir simulation (NRS), which is based on a numerical model and produces good results [3]. However, the NRS models have significant disadvantages, such as being time-consuming and cumbersome and demanding the construction of a precise static model, as well as various dynamic model parameters [4,5].

The successive geometric transformations model (SGTM) is a model that can be used to estimate electric power consumption in combined-type industrial zones. It is a neural-like model developed by Kachenko and Izonin [6]. Studies show that SGTM is effective in comparison to statistical regression analysis. The general regression neural network with SGTM was found to increase the predictive accuracy based on the recovery of missing data [7].

Analytical approaches are used to compute several types of adjustments of wellbore flow rate. Some hypotheses are necessary for establishing the analytical solution depending on the complexity of the well structure, boundary conditions, and reservoir heterogeneity [4,8].

Additionally, the conventional decline curve analysis (DCA) technique is used to forecast the production rate [9]. The DCA technique uses empirical equations, such as harmonic, hyperbolic, and exponential models. Because the harmonic and exponential curves can be obtained from the hyperbolic decline curve, it may be seen as a generalized model [10].

These models, however, cannot take into account the actual formation variables. As a result, using DCA to ensure correct performance is difficult.

Deep learning (DL) and machine learning (ML) applications in the oil sector cover different applications, including forecasting oil production [11,12] forecasting pressure volume temperature (PVT) [4,13,14] forecasting oil demands [15,16] and detecting oil spills [17].

In this paper, a novel regression model has been developed. The model is evaluated using a variety of performance measures, and the correctness of the proposed model and other models is compared. The main contribution of this paper is:

- A novel optimized model using the GB algorithm called the GA-GB model is proposed for forecasting crude oil production.
- The GA algorithm is employed to optimize GB parameters providing better performance for the GB model
- Extensive comparisons with five models (Bagging regressor, K-nearest neighbors (KNN) regressor, MLP regressor, RF regressor, and Lasso regressor) were performed to validate the performance of the GA-GB model using three real-world datasets.
- The proposed GA-GB model has been successfully used for three distinct datasets obtained from various resources for oil production (OilProd), oil price (OilPrice), and oil demand (OD). These datasets go through data normalization and data imputation during the preparation step.
- The results obtained by computing several performance measures such as MAE, MSE, MedAE, RMSE, and R^2 to predict oil production (OilProd), oil price (OilPrice), and oil demand (OD) using GA-GB demonstrate lower error than other traditional methods. It reveals that Bayesian optimization employing a genetic algorithm is more effective than other traditional methods.

The rest of this paper is organized as follows: Section 2 outlines several studies for crude oil forecasting. Section 3 describes the OProd, OPrice, and OD datasets. Section 4 describes the methodology. Section 5 contains the experimental results, and the conclusion is given in Section 6.

2. Related Works

Several machine learning and deep learning models have been applied in the field of oil production. This section describes studies that use machine learning and deep learning techniques in oil production.

Cheng, Y., and Yang, Y. use the long-short term memory (LSTM) and gated recurrent method for oil production, taking into account the time series. The dataset was collected from China and India. The experiment shows that LSTM and GRU are effective models for the dynamic prediction of oil well production [18]. The proposed models focus on the prediction of oil production and are applied only to two oil wells, one in northwestern China and the second Campbell Basin in India.

AlRassas, A. M. et al. proposed a hybrid model called AO-ANFIS, which is a modified ANFIS Adaptive Neuro-Fuzzy Inference System model and optimization algorithm Aquilla Optimizer (AO). The proposed model has been applied to forecast two different oil fields in China and Yemen. Comparisons were applied to the traditional ANFIS model and other models employing various optimization techniques. Results show that AO has significantly improved the accuracy of the prediction [3]. The proposed model does not use a mutation approach might further improve the AO algorithm's search process, increasing the accuracy of ANFIS. Moreover, the proposed model has been applied only to forecast oil in China and Yemen.

Tadjer, A., et al. [19] adopted two models named DeepAR and Prophet time series analysis were used to predict oil production. The models were applied to selected wells of the Midland fields in the USA. Results show that DeepAR and Prophet analysis are useful for better understanding how oil wells behave and can reduce over/underestimations caused by forecasting with a single decline curve model [19]. The proposed models can solve non-linear short-term oil production forecasting and need improvement to increase predicting performance over extended time horizons.

Makhotin, I., et al. [20] compare machine learning models' rankings of waterflooding efficiency against expert rankings. In particular, Linear regression (LR) models along with neural networks (NN) and GB on decision trees. According to the findings, machine learning models reduce computing complexity and are useful for rating reservoirs. It should be mentioned. Nevertheless, that historical information from Texas waterflood projects was used to perform this study. It was constrained by a certain set of criteria in the database as well as the geological aspects of the area.

Al-qaness, M.A., et al. have adopted a modified Aquila Optimizer (AO) with the Opposition-Based Learning (OBL) technique to optimize ANFIS parameters in a model called AOOBL-ANFIS. The proposed model has been applied to several real-world oil production datasets. Different comparisons between AOOBL-ANFIS and other models such as Autoregressive Integrated Moving Average (ARIMA), LSTM, and classical ANFIS model are applied. The results show that AOOBL-ANFIS outperformed the compared models [21]. Despite the fact that AOOBL-ANFIS outcomes are high. However, the AOOBL-ANFIS mode has limitations affecting its performance. For example, selecting the ratio of solutions that will be updated using the OBL is a significant parameter that causes the time complexity of the proposed model.

Werneck, R., et al. [22] introduced a new setup called N-th Day for forecasting multiple outputs with machine-learning algorithms and assessing a number of learning models. Four deep-learning models are adopted for forecasting oil production. The obtained results indicate that specific architectures are critical for forecasting oil and gas production [6]. There was no data leakage during the training phase. The proposed method, nevertheless, was centered on oil production.

Duan, Y., et al. [23] have combined the autoregressive integral moving average (ARIMA) model with RTS (Rauch Tung Striebel) smoothing for forecasting oil production. The ARIMA-RTS model has a greater prediction accuracy than the ARIMA-Kalman model in predicting the same gas well. The ARIMA-RTS model can aid in improving the prediction of fuels and gas well production with stimulation. However, the ARIMA-RTS model is validated using one actual well production data, and the study was centered on forecasting oil production.

Alkhammash, E. H., et al. [15] proposed combined LR and multivariate adaptive regression splines for predicting crude oil demand in Saudi Arabia. The social spider optimization (SSO) algorithm is used to optimize LR-MARS parameters. The findings indicate that Saudi Arabia's demand for crude oil will continue to rise over the predicted period (1980–2015). The proposed model focused on the prediction of oil demand and was applied only to oil production in Saudi Arabia country.

Unlike other studies, this paper proposed a new model called GA-GB on three different datasets, which are OPrice, OilProd, and OD. Results show that the proposed model is useful for forecasting oil price, oil production, and oil demand. To the best of our knowledge, there is no study that applies one model to successfully forecast oil price, oil production, and oil demand. In addition, unlike other approaches, the proposed GA-GB model has been applied to forecast crude oil in several countries, including the top producers and others with less production.

3. Dataset Description

The experiment uses three different datasets. The first dataset reflects the Country Yearly Oil Production (Barrels per day) and covers the period from 1960 to 2020. The second dataset describes the yearly spot crude oil price from 1983 to 2020. Both datasets are gathered from (https://asb.opec.org/data/, accessed on 1 March 2022). The third dataset is crude oil demand gathered from different sources. The gross domestic product (GDP) feature is gathered from the sources: OPEC, IEA, International Monetary Fund (IMF), Saudi Statistics Authority, and World Bank and covers the period 1980–2015. The crude oil demand features are year, oil demand, GDP, population, Brent prices, Light-Duty Vehicles (LDV), and Heavy-Duty Vehicles (HDV) [16]. Table 1 describes selected spot crude oil prices ($/b) Yearly. Table 2. describes a sample of the world crude oil production by country (1000 b/d)

Table 1. Selected spot crude oil prices ($/b) Yearly.

City	Year	Price
Saudi Arabia-Arab Heavy	2020	41.45
OPEC–ORB	2020	41.47
Nigeria–Forcados	2020	41.56
United Kingdom–BrentDated	2020	41.67
Algeria–Zarzaitine	2020	41.72
Russia–Urals	2020	41.83
Angola–Cabinda	2020	42.29
United Arab Emirates–Dubai	2020	42.31
Norway–Ekofisk	2020	42.33

Table 2. Sample of the world crude oil production by country (1000 b/d).

City	Year	Production
Saudi Arabia	2020	9213.2
Sudans	2020	230.4
Syrian Arab Rep.	2020	22.4
Thailand	2020	117.0
United Arab Emirates	2020	2778.6
United Kingdom	2020	930.5
United States	2020	11,283.0
Venezuela	2020	568.6
Vietnam	2020	193.7
Yemen	2020	42.0

Table 3 describes a number of statistical metrics such as mean, standard error, median, standard deviation, etc., of the three datasets oil price, oil production, and oil demand. For instance, the standard deviation of oil price is 31.27554387, oil production is 8949.166932, and oil demand is 774.0563839.

Table 3. The statistical data analysis of the three datasets Oil price, Oil production, and Oil demand.

	Oil Price	Oil Production	Oil Demand
Mean	41.9362201	3165.570318	1539.087014
Standard Error	1.085145875	141.0411122	129.0093973
Median	28.185	460.1	1230.609065
Standard Deviation	31.37554387	8949.166932	774.0563839
Sample Variance	984.4247534	80087588.78	599163.2855
Kurtosis	−0.344726451	28.64603253	−0.247425069
Skewness	0.876459222	5.038060934	0.962853111
Confidence Level	2.129934178	276.5186522	261.9030003

4. Methodology

In order to forecast crude oil output, this paper proposes since an optimized prediction model based on GA and GB. Figure 1 illustrates the four key steps of the proposed GA-GB model development: (1) data preprocessing stage, (2) GA optimization, (3) GB-GA model, and (4) performance evaluation.

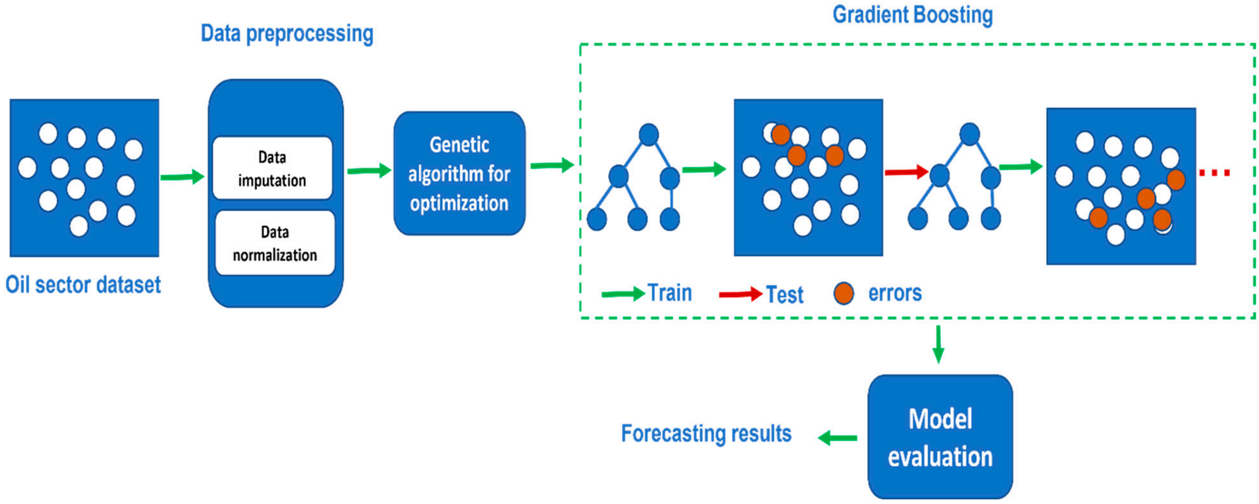


Figure 1. The stages of GA-GB forecasting model.

4.1. Dataset Preprocessing

4.1.1. Data Imputation

When dealing with real-world problems such as crude oil production, missing values are common when data is gathered over long time periods from disparate sources. Imputation is a process in the data preprocessing stage that is used to replace missing values with substituted data using basic statistical parameters such as median and mode.

In this study, we used mean imputation. The mean imputation is a simple method that replaces all missing values. The variance of the imputed variables is reduced using mean imputation. The standard errors are also reduced using mean imputation.

4.1.2. Data Normalization

The quality of the data can have a direct impact on the models' ability to learn; consequently, it is vital that we preprocess our data before utilizing it as inputs into the suggested model. Normalization is used for preprocessing in this paper. Normalization can be used to scale input values with various scales if the data includes various scales. Making variables similar to one another is what normalization does. In normalization, each variable receives equal weight, ensuring that no one variable dominates the model's output. Rescaling (min-max normalization), mean normalization, and Z-score normalization or standardization are examples of techniques used for data normalization.

Normalization adjusts each input value independently by subtracting the mean (centering) and dividing by the standard deviation to modify the mean and standard deviation of the distribution to zero and one [24]. The following equation is used to determine normalization.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x represents the input value, μ the mean value, and σ the standard deviation value.

The following equation is used to compute standard deviation, where x_i denotes the input:

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

4.2. Genetic Algorithm

The genetic algorithm (GA) is an optimization technique developed by Holland in 1992 [25]. This technique was greatly influenced by biological species evolution and the natural selection mechanism. GA is a probabilistic approach and doesn't require specific data to lead a search [24,26,27].

Individuals in populations are traditionally referred to be candidate solutions, which gradually replace the preferable solutions over time. The digits 0 and 1 indicate chromosomes that form a linear string, which provides a solution for each candidate. Generation is the overall population size created by optimizing each iteration [26]. The generation in GA is produced via three fundamental genetic operators, namely reproduction, cross-over, and mutation. A reproduction operator is described as a process of choosing the best chromosomes based on their scaled values while taking into account the given criterion of fitness. The second operator is the cross-over, which creates new individuals by fusing particular parts of existing individuals of parents. Recombination can be used in a variety of ways, including single-point cross-over and two-point crossover. Despite this, the process of crossover chooses two parents and a random cross-over point. The first offspring is produced by combining the left side of the first parent's gene with the opposite side of the second parent's gene, and the second offspring is produced by repeating the first procedure in the opposite direction [26,28]. Elements of a chromosome are randomly substituted in the mutation operator. 4.5.

4.3. Gradient Boosting Algorithm

The Boosting Technique is a supervised machine learning algorithm that was developed in the last two decades. The boosting method is an ensemble technique that employs numerous weak learners that focus on the errors that occur at each step until a strong model for regression and classification is produced. In this paper, we use the bosting approach for regression purposes. Gradient boosting is made up of three primary components: The loss function identifies the difference between actual and predicted values. A weak learner is a decision tree to make a prediction. An additive model to minimize the loss function. Each weak learner model attempts to fix errors introduced by previous weak learner models to improve the model's prediction and reduce its prediction error [29,30].

Consider a set of random input variables $x = \{x_1,\ x_2, \ldots\ x_n\}$ together with a response variable z.

$$\widetilde{F}(x) = \underset{F(x)}{\text{argmin}}\, L_{z,x}(z,\ F(x)) \tag{3}$$

The loss function is used a squared error function to estimate the approximation function:

$$Loss(z,\ F(x)) = (z - F(x))^2 \tag{4}$$

The following equation may be used to get the gradient of the loss function $Loss(z,\ F(x))$ [29]:

$$\widetilde{z_i} = \left[\frac{\partial\, Loss(z_i,\ F(x_i))}{\partial\, F(x_i\ )}\right]_{F(x)=F_{m-1}(x)} \tag{5}$$

When we utilize regression trees $h(x_i; b)$ with parameter b as weak learners, it might generalize the computation range of the gradient. Typically, it is a parameterized function of the input variables x with parameter b [29]. The following equation may be solved to get the tree [29]:

$$b_m = \arg\min_{b,\beta} \sum_{i=1}^{N} [\widetilde{z_i} - \beta h(x_i,\ b)]^2 \tag{6}$$

where b_m is the weight value, commonly known as the expansion coefficient of each weak learner and $m,\ \beta$ is the parameters collected at iteration m.

4.4. Performance Evaluation

Five different measures of errors are introduced to evaluate the proposed models in order to validate the performance and effectiveness of the prediction models, which are Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-squared are used to assess the performance of each model (R^2) as shown in Equations (7)–(11), where y_{real_i} denotes actual values, y_{pred_i} denotes predicted [31].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| y_{real_i} - y_{pred_i} \right| \tag{7}$$

$$MedAE = median\left(\left| y_{real_i} - y_{pred_i} \right|, \ldots, \left| y_{real_n} - y_{pred_n} \right| \right) \tag{8}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(y_{real_i} - y_{pred_i} \right)^2 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_{real_i} - y_{pred_i} \right)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_{real_i} - y_{pred_i} \right)^2}{\sum_{i=1}^{N} \left(y_{real_i} - \overline{y} \right)^2} \tag{11}$$

5. Results and Discussion

In order to achieve the predictions accurately for the three datasets, a novel model called genetic algorithm-gradient boosting (GA-GB) is constructed. Also, other machine learning regression models are constructed in this paper to be compared with the GA-GB model. In addition, data cleaning and data preprocessing were performed on the three datasets. Five evaluation metrics, namely, mean absolute error (MAE), median absolute error (MedAE), mean square error (MSE), coefficient of determination (R2), and root mean squared error (RMSE), were computed to evaluate the performance of the models. The results demonstrated that the GA-GB model achieved the best results.

The experimental results are executed using the Jupyter notebook version (6.4.6). Jupyter notebook helps in executing and writing python codes easily. Jupyter notebook is widely utilized as an open source for implementing and executing machine learning regression models. The results of four models, namely, bagging regressor, k-nearest neighbor (KNN) regressor, multi-layer perceptron (MLP), random forest (RF) model, and lasso regressor, are used to be compared with the results of the GA-GB model. In Table 4, the best parameters of the GB model are obtained by GA. The GA is used to select the best parameters for the models and minimize the loss function that gives the best results. It is a random-based optimization technique where random changes are made to the current solutions in order to produce new ones. These changes are made slowly to obtain solutions until finding the best solution.

Table 4. The best parameters for the gradient boosting model using a genetic algorithm.

Models	Tuning Parameters	Best Parameters
GB	n_estimators = [50,100,150,200,250] learning_rate = [0.1, 0.01, 0.001, 0.0001]	n_estimators = 150 learning_rate = 0.001

For the bagging regressor model, the parameters used for the model are presented in Table 5. For the KNN regressor model, the parameters used for the model are presented in Table 6.

Table 5. Parameters used for bagging regressor model.

Model	Parameters
Bagging regressor	n_estimators = 100, max_samples = 5

Table 6. Parameters used for KNN regressor model.

Model	Parameters
KNN regressor	n_neighbors = 5, weights = distance.

In MLP, we have used three layers. The first layer consists of 64 hidden units/neurons with the ReLU activation function. The second layer consists of 32 hidden units with the ReLU activation function. The final layer is the output layer which consists of one unit with a linear activation function. The optimizer used is the Adam optimizer, and the number of epochs is 100. The parameters used for the MLP model are presented in Table 7.

Table 7. Parameters for MLP model.

Batch Size	Learning Rate	Epochs	Optimizer	Activation Function Used in Output	Activation Function Used in Hidden
32	0.0001	100	Adam	Linear	Relu

For RF regressor model, the parameters used for the model are presented in Table 8. For Lasso regressor model, the parameters used for the model are presented in Table 9.

Table 8. Parameters used for RF regressor model.

Model	Parameters
RF regressor	max_depth = 15, n_estimators = 150

Table 9. Parameters used for lasso model.

Model	Parameters
Lasso model	alpha = 1, fit_intercept = true

The performance of the regression models for **oil production (OProd)** dataset are demonstrated in Table 10.

Table 10. Performance of the regression models for oil production dataset.

Models	MAE	MSE	MedAE	RMSE	R^2
GA-GB	0.002	3.8×10^{-2}	0.0008	0.001	99.8%
Bagging regressor	0.006	0.0003	0.0015	0.004	99%
KNN regressor	0.009	0.0005	0.007	0.008	98.2%
MLP regressor	0.004	0.0002	0.0013	0.003	99.1%
RF regressor	0.008	0.0004	0.003	0.006	98.74%
Lasso	0.06	0.009	0.07	0.06	95.4%

From Table 10, the GA-GB model gives the best results as can be seen in bold. The lasso model gives the least results.

The performance of the regression models for the **oil price (OPrice)** dataset is demonstrated in Table 11.

Table 11. Performance of the regression models for oil price dataset.

Models	MAE	MSE	MedAE	RMSE	R^2
GA-GB	0.0002	1.1×10^{-7}	0.0001	0.0011	99.99%
Bagging regressor	0.0020	6.48×10^{-4}	0.006	0.0076	99.8%
KNN regressor	0.0006	5.43×10^{-7}	0.0005	0.0023	99.96%
MLP regressor	0.0007	1.57×10^{-6}	0.0007	0.0037	99.95%
RF regressor	0.0060	7.36×10^{-4}	0.009	0.0092	99.2%
Lasso	0.04	0.005	0.04	0.03	96.1%

From Table 11, the GA-GB model gives the best results (the bold). The lasso regressor model gives the least results.

The performance of the regression models for the **oil demand (OD)** dataset is demonstrated in Table 12.

Table 12. Performance of the regression models for oil demand dataset.

Models	MAE	MSE	MedAE	RMSE	R^2
GA-GB	0.001	0.00010	0.004	0.002	98.6%
Bagging regressor	0.004	0.00018	0.006	0.004	98.2%
KNN regressor	0.003	0.00017	0.006	0.004	98.2%
MLP regressor	0.007	0.00027	0.009	0.008	98%
RF regressor	0.006	0.00022	0.008	0.007	98.1%
Lasso	0.08	0.005	0.09	0.09	94.9%

From Table 12, the GA-GB model gives the best results. The lasso regressor model gives the least results. Figures 2–4 compare between the models in the term of coefficient of determination (R^2) for the three datasets, OProd, Oprice, and OD, respectively.

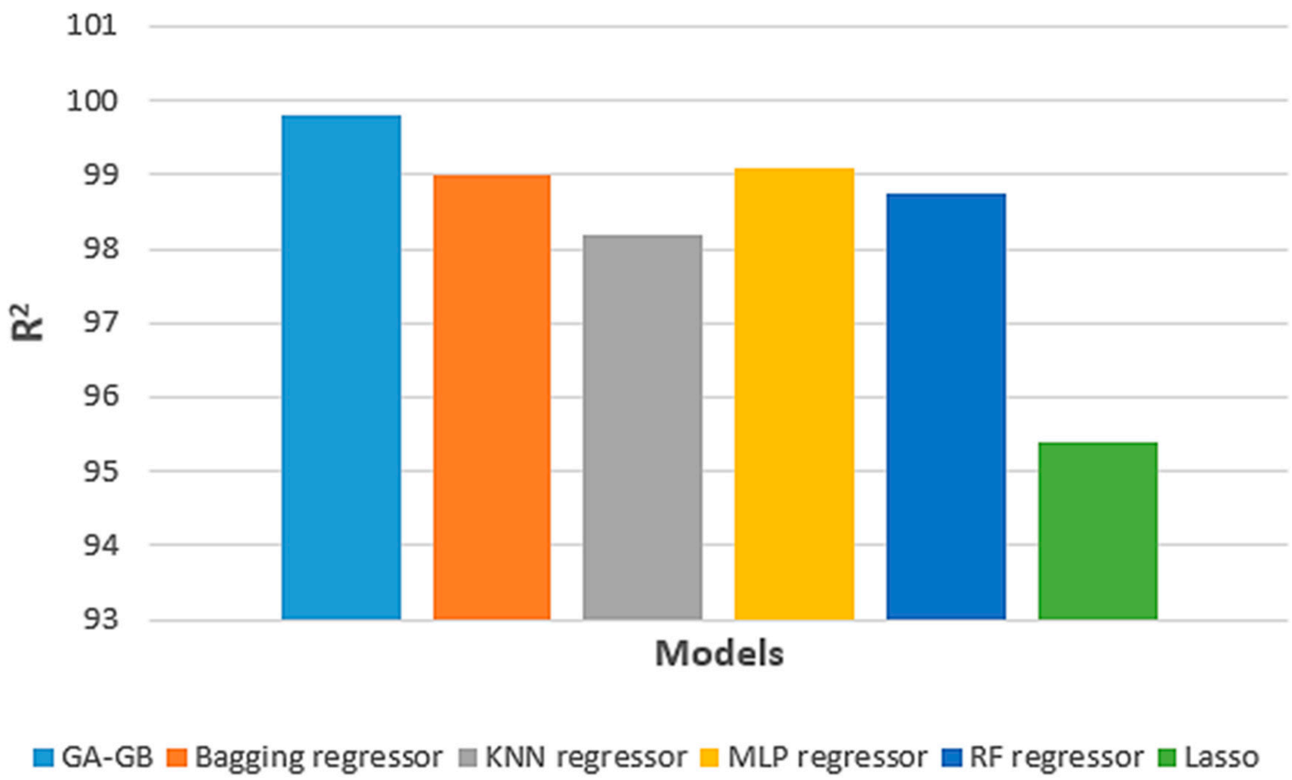


Figure 2. Comparison between the models used in the term of coefficient of determination (R^2) using OProd dataset.

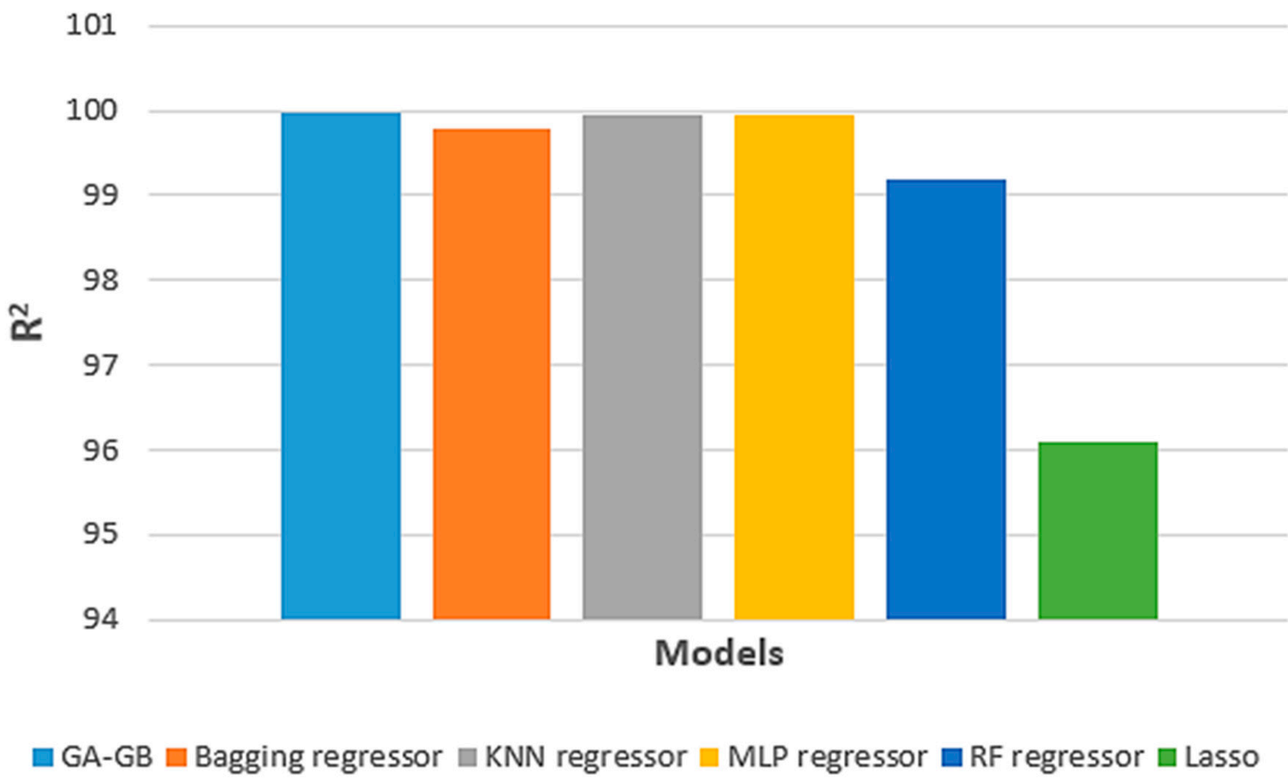


Figure 3. Comparison between the models used in the term of coefficient of determination (R^2) using OPrice dataset.

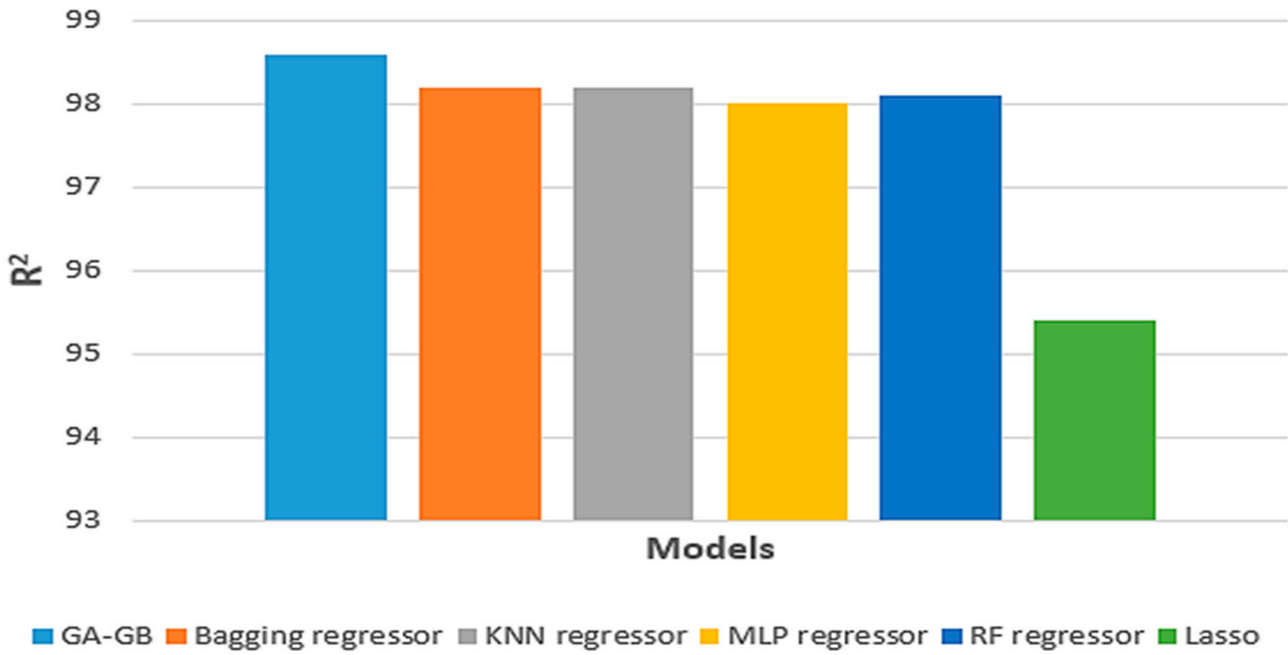
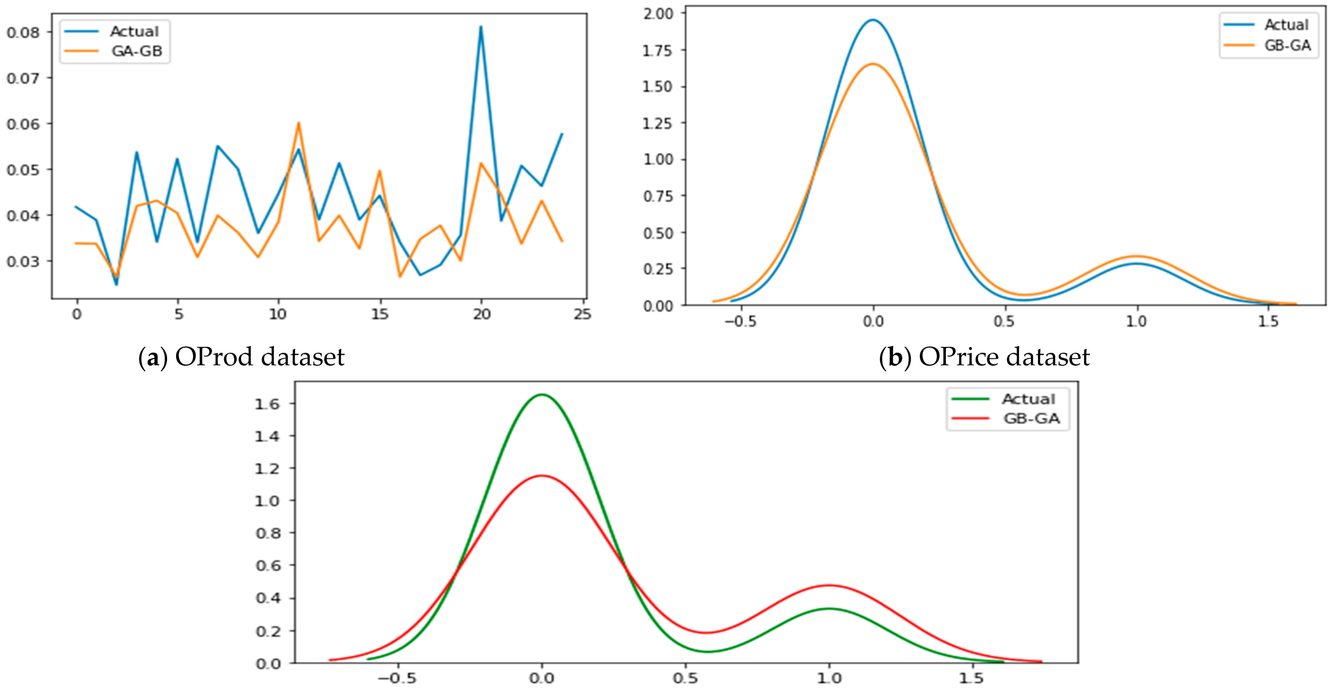


Figure 4. Comparison between the models used in the term of coefficient of determination (R^2) using OD dataset.

Figure 5 displays a comparison between the actual values and the predicted values for the GA-GB model for the three datasets, OProd, Oprice, and OD, respectively.



(**a**) OProd dataset (**b**) OPrice dataset

(**c**) OD dataset

Figure 5. Displays a comparison between the actual values and the predicted values for the GA-GB model for the three datasets (**a**) Oprice, (**b**) OProd, and (**c**) OD.

The main contribution of this study is to optimize the hyperparameters of different machine learning algorithms based on the Gradient Bayesian (GB) optimizer using a genetic algorithm (GA-GB) to improve the forecasting of crude oil production. Most of the algorithms reflected in the past utilized the traditional methods. However hyperparameters optimization is more capable of improving the forecasting. We computed the performance using standard evaluation metrics such as MAE, MSE, MedAE, RMSE, and R^2. In the past, researchers measured the evaluation performance based on different error metrics [28], as each measure is not suitable for all kinds of problems. So, researchers compute different measures [32] accordingly, as in this study. The minimum value of different measures indicates the more robust algorithm to forecast. In this study, the proposed GA-GB algorithm yielded the highest forecasting prediction performance than the traditional one using all the measures. Moreover, for reflected the graphical presentation of GA-GB only to reflect the visual representation as applied and reflected by the researchers in the past to see the difference between actual and predicted values.

6. Conclusions

This paper proposed an optimized gradient boosting model that employs the GA to tune the parameters of GB. The proposed GA-GB model is used successfully to forecast crude oil production. The price of crude oil and the demand for crude oil are also predicted using the GA-GB model. The experiment results demonstrated a very high performance of the GA-GB. Three different datasets are used in the experiment: OilProd, OilPrice, and OD. The preprocessing stage was performed using data imputation and data normalization. To evaluate the performance of GA-GB optimized model, we utilized MAE, MSE, MedAE, RMSE, and R^2 to evaluate and test the predictions performance for the GA-GB model that are: 0.002, 3.8×10^{-2}, 0.0008, 0.001, 0.001, 99.8%, respectively in OProd dataset, 0.0002, 1.1×10^{-7}, 0.0001, 0.0011, 99.99%, respectively in OPrice dataset, and 0.00010, 0.004, 0.002, and 98.6%, respectively in OD dataset. Five other regression models are compared with GA-GB, including the Bagging regressor, KNN regressor, MLP regressor, RF regressor, and Lasso regressor. The results yielded using the GA-GB model by computing different performance metrics such as MAE, MSE, MedAE, RMSE, and R^2 to predict the oil production, oil price, and oil demand exhibit minimal error than using other traditional methods. It indicates that Bayesian optimization using a genetic algorithm is more powerful than other traditional methods. The proposed model can be best utilized for forecasting oil production, prices, and demand in order to improve the planning of the concerned departments. A limitation of this work is that there are many important factors in AI that was hampered by data unavailability, such as oil imports over consumption. Future work will focus on including the most crucial factors, working at the dataset level, and analyzing the impact of different parameters on oil production.

Funding: This work is supported by Taif University Researchers Supporting Project number (TURSP-2020/292) Taif University, Taif, Saudi Arabia.

Informed Consent Statement: Not applicable.

Data Availability Statement: Oil production and oil price datasets are obtained from: (https://asb.opec.org/data/, accessed on 1 March 2022) whereas oil demand dataset is obtained from [16].

Acknowledgments: The author would like to acknowledge Taif University Researchers Supporting Project number (TURSP-2020/292) Taif University, Taif, Saudi Arabia.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Li, H.; Yu, H.; Cao, N.; Tian, H.; Cheng, S. Applications of artificial intelligence in oil and gas development. *Arch. Comput. Methods Eng.* **2021**, *28*, 937–949. [CrossRef]
2. Di, S.; Cheng, S.; Cao, N.; Gao, C.; Miao, L. AI-based geo-engineering integration in unconventional oil and gas. *J. King Saud Univ.-Sci.* **2021**, *33*, 101542. [CrossRef]
3. Mesbah, M.; Vatani, A.; Siavashi, M.; Doranehgard, M.H. Parallel processing of numerical simulation of two-phase flow in fractured reservoirs considering the effect of natural flow barriers using the streamline simulation method. *Int. J. Heat Mass Transf.* **2019**, *131*, 574–583. [CrossRef]
4. AlRassas, A.M.; Al-qaness, M.A.; Ewees, A.A.; Ren, S.; Abd Elaziz, M.; Damaševičius, R.; Krilavičius, T. Optimized ANFIS model using Aquila Optimizer for oil production forecasting. *Processes* **2021**, *9*, 1194. [CrossRef]
5. Nwaobi, U.; Anandarajah, G. Parameter determination for a numerical approach to undeveloped shale gas production estimation: The UK Bowland shale region application. *J. Nat. Gas Sci. Eng.* **2018**, *58*, 80–91. [CrossRef]
6. Tkachenko, R.; Izonin, I.; Kryvinska, N.; Dronyuk, I.; Zub, K. An approach towards increasing prediction accuracy for the recovery of missing IoT data based on the GRNN-SGTM ensemble. *Sensors* **2020**, *20*, 2625. [CrossRef] [PubMed]
7. Tkachenko, R.; Izonin, I. Model and principles for the implementation of neural-like structures based on geometric data transformations. In Proceedings of the International Conference on Computer Science, Engineering and Education Applications, Kiev, Ukraine, 18–20 January 2018; pp. 578–587.
8. Asadi, M.B.; Dejam, M.; Zendehboudi, S. Semi-analytical solution for productivity evaluation of a multi-fractured horizontal well in a bounded dual-porosity reservoir. *J. Hydrol.* **2020**, *581*, 124288. [CrossRef]
9. Wachtmeister, H.; Lund, L.; Aleklett, K.; Höök, M. Production decline curves of tight oil wells in eagle ford shale. *Nat. Resour. Res.* **2017**, *26*, 365–377. [CrossRef]
10. Liang, H.B.; Zhang, L.H.; Zhao, Y.L.; Zhang, B.N.; Chang, C.; Chen, M.; Bai, M.X. Empirical methods of decline-curve analysis for shale gas reservoirs: Review, evaluation, and application. *J. Nat. Gas Sci. Eng.* **2020**, *83*, 103531. [CrossRef]
11. Liu, W.; Liu, W.D.; Gu, J. Forecasting oil production using ensemble empirical model decomposition based Long Short-Term Memory neural network. *J. Pet. Sci. Eng.* **2020**, *189*, 107013. [CrossRef]
12. Song, X.; Liu, Y.; Xue, L.; Wang, J.; Zhang, J.; Wang, J.; Jiang, L.; Cheng, Z. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *J. Pet. Sci. Eng.* **2020**, *186*, 106682. [CrossRef]
13. Liu, J.; Wang, S.; Wei, N.; Chen, X.; Xie, H.; Wang, J. Natural gas consumption forecasting: A discussion on forecasting history and future challenges. *J. Nat. Gas Sci. Eng.* **2021**, *90*, 103930. [CrossRef]
14. Agwu, O.E.; Akpabio, J.U.; Dosunmu, A. Artificial neural network model for predicting the density of oil-based muds in high-temperature, high-pressure wells. *J. Pet. Explor. Prod. Technol.* **2020**, *10*, 1081–1095. [CrossRef]
15. Alkhammash, E.H.; Kamel, A.F.; Al-Fattah, S.M.; Elshewey, A.M. Optimized multivariate adaptive regression splines for predicting crude oil demand in Saudi arabia. *Discret. Dyn. Nat. Soc.* **2022**, *2022*, 8412895. [CrossRef]
16. Al-Fattah, S.M.; Aramco, S. Application of the artificial intelligence GANNATS model in forecasting crude oil demand for Saudi Arabia and China. *J. Pet. Sci. Eng.* **2021**, *200*, 108368. [CrossRef]

17. Capizzi, G.; Sciuto, G.L.; Woźniak, M.; Damaševičius, R. A Clustering Based System for Automated Oil Spill Detection by Satellite Remote Sensing. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2016; pp. 613–623.
18. Cheng, Y.; Yang, Y. Prediction of oil well production based on the time series model of optimized recursive neural network. *Pet. Sci. Technol.* **2021**, *39*, 303–312. [CrossRef]
19. Tadjer, A.; Hong, A.; Bratvold, R.B. Machine learning based decline curve analysis for short-term oil production forecast. *Energy Explor. Exploit.* **2021**, *39*, 1747–1769. [CrossRef]
20. Makhotin, I.; Orlov, D.; Koroteev, D. Machine Learning to Rate and Predict the Efficiency of Waterflooding for Oil Production. *Energies* **2022**, *15*, 1199. [CrossRef]
21. Al-qaness, M.A.; Ewees, A.A.; Fan, H.; AlRassas, A.M.; Abd Elaziz, M. Modified aquila optimizer for forecasting oil production. *Geo-Spat. Inf. Sci.* **2022**, 1–17. [CrossRef]
22. de Oliveira Werneck, R.; Prates, R.; Moura, R.; Gonçalves, M.M.; Castro, M.; Soriano-Vargas, A.; Júnior, P.; Hossain, M.; Hossain, M.; Ferreira, A.; et al. Data-driven deep-learning forecasting for oil production and pressure. *J. Pet. Sci. Eng.* **2022**, *210*, 109937. [CrossRef]
23. Duan, Y.; Wang, H.; Wei, M.; Tan, L.; Yue, T. Application of ARIMA-RTS optimal smoothing algorithm in gas well production prediction. *Petroleum* **2022**, *8*, 270–277. [CrossRef]
24. Mirjalili, S. Genetic algorithm. In *Evolutionary Algorithms and Neural Networks*; Springer: Cham, Switzerland, 2019; pp. 43–55.
25. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; MIT Press: Cambridge, MA, USA, 1992.
26. Khandelwal, M.; Marto, A.; Fatemi, S.A.; Ghoroqi, M.; Armaghani, D.J.; Singh, T.N.; Tabrizi, O. Implementing an ANN model optimized by genetic algorithm for estimating cohesion of limestone samples. *Eng. Comput.* **2018**, *34*, 307–317. [CrossRef]
27. Saemi, M.; Ahmadi, M.; Varjani, A.Y. Design of neural networks using genetic algorithm for the permeability estimation of the reservoir. *J. Pet. Sci. Eng.* **2007**, *59*, 97–105. [CrossRef]
28. Butt, F.M.; Hussain, L.; Jafri, S.H.M.; Alshahrani, H.M.; Al-Wesabi, F.N.; Lone, K.J.; Tag El Din, E.M.; Duhayyim, M.A. Intelligence based Accurate Medium and Long Term Load Forecasting System. *Appl. Artif. Intell.* **2022**, *36*, 2088452. [CrossRef]
29. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
30. Nie, P.; Roccotelli, M.; Fanti, M.P.; Ming, Z.; Li, Z. Prediction of home energy consumption based on gradient boosting regression tree. *Energy Rep.* **2021**, *7*, 1246–1255. [CrossRef]
31. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef]
32. Hussain, L.; Saeed, S.; Idris, A.; Awan, I.A.; Shah, S.A.; Majid, A.; Ahmed, B.; Chaudhary, Q.A. Regression analysis for detecting epileptic seizure with different feature extracting strategies. *Biomed. Eng./Biomed. Tech.* **2019**, *64*, 619–642. [CrossRef]