Knowledge-Based Machine Learning Approaches to Predict Oil Production Rate in the Oil Reservoir

Ayman Mutahar AlRassas[1,2], Chinedu Ejike[3](✉), Salman Deumah[4],
Wahib Ali Yahya[5], Anas A. Ahmed[4,6], Sultan Abdulkareem Darwish[7],
Asare Kingsley[4], and Sun Renyuan[1]

[1] School of Petroleum Engineering, China University of Petroleum (East China),
Qingdao 266580, China

[2] Institute of Subsurface Energy Systems, Clausthal University of Technology, 38678
Clausthal-Zellerfeld, Germany

[3] Mewbourne School of Petroleum and Geological Engineering, University of Oklahoma,
Norman, OK 73019, USA
ejikechinedu@gmail.com

[4] College of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249,
China

[5] Department of Petroleum Engineering, China University of Geosciences (Wuhan),
Wuhan 430074, China

[6] Department of Petroleum and Natural Gas Engineering, University of Khartoum,
Khartoum 11111, Sudan

[7] College of Petroleum Engineering, Xi'an Shiyou University, Xi'an 710065, China

Abstract. Predicting the oil production rate is a crucial means to improve the operation of hydrocarbon reservoirs and manage the economic plans for oil companies. However, developing a reliable model to predict oil production rate using traditional numerical frameworks are challenging and requires too much time to attain a single model. Thus, in this paper, Machine Learning (ML) techniques are presented as a robust and intelligent framework to predict oil production rates accurately and timely. The ML techniques include Multiple Linear Regression (MLR), Random Forest (RF), Decision Tree (DT), and K-nearest neighbor (KNN). These

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 282–304, 2024.
https://doi.org/10.1007/978-981-97-0268-8_24

four techniques were engaged to predict the oil production rate of real oilfield data of 11 wells. The 11 oil wells were considered as datasets to achieve a precise prediction of the oil production rate. The available datasets were split into two subsets of training and testing data sets. Furthermore, the Root Mean Squared Errors (RMSE) and determination-coefficient values (R^2) regression metrics were employed to evaluate the model performance. Hence, the comparative analysis of the proposed models was presented for all 11 selected production wells. The analysis of results showed that RF can be considered the best predictive ML model for predicting oil production rate with the lowest RMSE and the highest R^2 scores in all 11 production wells. In the KT911H well, the RF model achieved the most accurate results with RMSE and R^2 0.868 and 0.9993 respectively. In addition, the study analysis illustrated that the RF can be considered to be the best predictive model, which was also applied to indicate the relationship between the input parameters and the oil production rate. A sensitivity analysis of the RF model indicated that the liquid volume, water cut, and gas pressure are the most important input parameters affecting the oil production rate performance in all 11 production wells. This paper therefore presents a pragmatic approach for predicting the oil production rate of a typical oilfield and the parameters with the most effects on the prediction based on machine learning techniques.

Keywords: Predictions · Oil Production Rate · Machine Learning · Oilfield · Model Performance Evaluation

Nomenclature

DCA	Decline Curve Analysis
RNS	Reservoir Numerical Simulation
LSTM	Long Short-Term Memory
LSSVM	Least Square Support Vector Machine
SVM	Support Vector Machine
ML	Machine Learning
AI	Artificial Intelligence
KNN	K-nearest neighbors
DT	Decision Tree
RF	Random Forest
Y	Response Variable
MLR	Machine Learning Techniques
RMSE	Root Mean Square Errors
R^2	Determination Co-efficient Value
βo	Y Intercept
$\beta_1 \& \beta_2$	Regression Coefficients first and second independent variables
$x_1 \& x_2$	First and Second Independent Variables
β_p	Regression Coefficient of the Last explanatory variable x_p
ε	residuals
GOR	Gas Oil Ratio

1 Introduction

In the last decades, the oil industry has gained a lot of improvement experiences which have had a huge positive impact on the global and regional businesses and economies associated with this kind of industry and other energy sectors [1–5]. One of the main keys that made this industry so impactful for decision-making is how the well's performance and production forecasting [6–9].

Generally, the performance of an oil well is characterized by many factors such as rate transient analysis, microseismical data, and well completion configurations. In the field, the well's performance can be achieved by implementing a refinement of the statistical interpretations of the well's data by normalizing either one or two key parameters that embody those factors, such as tonnage and lateral lengths[10]. Those wells are then grouped into certain categories for various reasons such as identical completion design, thus reducing the sample sizes without influencing the well performance.

1.1 Well Performance and Decline Curve Analysis Optimization

There are two basic methods for analyzing and predicting of oil production rate, the decline curve analysis (DCA) and reservoir numerical simulation (RNS). The DCA method uses the production data to make predictions of the well's problems and performance. DCA has many advantages: the required data is easily obtainable, it can be illustrated easily with graphs, it shows results on a timely basis, and its analysis is easily handable [10]. However, for DCA, production predictions are made using ideal models, resulting in real production curves that may not exhibit the same level of smoothness as their output.

RNS has proven to be a highly effective tool in the examination of intricate reservoir problems [11]. Nonetheless, this approach follows a bottom-up methodology that entails a laborious and time-intensive process, involving the development of a geological model, a numerical model, and the execution of history matching. Each of these individual models has to be near-perfect for an accurate prediction of the production rate.

The utilization of big data mining, machine learning, and artificial intelligence has emerged as the primary technological advancements employed for intelligent investigation, advancement, and production inside the upstream petroleum sector. Reservoir engineers are responsible for the storage and management of substantial quantities of oilfield data. Additionally, they engage in comprehensive mining and examination of this data to enhance and optimize oilfield development strategies. The employing of artificial intelligence technology enables a thorough analysis and interpretation of oil reservoir data, hence significantly enhancing the efficiency and economic advantages associated with oilfield development [12–15].

The work in this article was therefore carried out by using artificial intelligence and machine learning techniques from a decision tree, a multiple linear regression, and a random forest technique. Those machine learning applications have major aspects in the oil industry in general and especially for the prediction of the well's production performance which is one of the most emerging sectors being explored and developed nowadays. The focus of this article is to highlight the relevance of the chosen machine learning techniques employed to forecast and optimize the performance of the wells.

Many works have been carried out using machine learning techniques for production performance optimization. For instance, Huang et. al [16] developed a Long Short-Term Memory (LSTM) neural network model to analyze the impact of gas injection on the prediction of production performance in a carbonate reservoir. According to the authors' statement, the LSTM method exhibited an average error that was 43.75% lower compared to the traditional RNS approach. Furthermore, the LSTM approach exhibited a total CPU time and comprehensive computing power consumption that constituted just 10.43% and 36.46% of the RNS's, respectively. [16].

Panja et al. utilized a Least Square Support Vector Machine (LSSVM) method to achieve precise predictions of oil recovery [17]. Wang et al. devised a data mining approach for well production performance They employed four distinct supervised learning techniques, including Random Forest, adaptive boost, support vector machine (SVM), and neural network, to forecast the first-year oil production in unconventional reservoirs. [18]. Osarogiagbon et al. provide a comprehensive analysis of the existing literature pertaining to hazardous events and supervised machine learning algorithms. The review offers a concise overview of the methodologies, achievements, and constraints associated with these algorithms [19] (Table 1).

Table 1 Summarized the differences between DCA, RNS, and ML techniques for production performance

Decline Curve Analysis (DCA)	A direct method to forecast the production is not appropriate for reservoirs that possess complex geological characteristics
Reservoir Numerical Simulation (RNS)	A mature used reservoir development and prediction solution with a lack of perfection in the geological uncertainty and long-running worktime
Machine Learning Techniques (ML)	High speed and accurate data foundation especially with high-quality data, but it requires a mature reservoir development solution to be applicable

2 Evolutions of Analytics

Analytics can be classified into four distinct types. The automation of descriptive analytics has been feasible since the inception of computing. However, advancements in machine learning and artificial intelligence (AI) have now made it possible for enterprises to automate the processes of issue identification, outcome prediction, and action prescription.

- Descriptive Analytics (What occurred?): demonstrating what is truly occurring based on the provided data, typically through dashboards and reporting tools.
- Diagnostic Analytics (Why did it occur?): Examining previous performance to achieve not only what occurred, but also why it occurred.

- Prescriptive Analytics (What should we do?): Making recommendations on what should be done and why.
- Predictive Analytics (What could occur?): Explains what possibilities are likely to happen, usually in the form of a predictive forecast.

2.1 Descriptive Analytics

Descriptive analytics, regarded as the foundational kind of analytics, refers to the analysis of historical data in order to determine what events or phenomena occurred within a specific dataset. This form of analytics encompasses several arithmetic operations, including mean, median, maximum value, percentage, and other similar calculations, is used by almost every organization. It provides organizations with useful insights into past actions; however, descriptive analytics may not provide the causes of the problem. That is why data consultants do not advise organizations to only use descriptive analytics.

2.2 Diagnostic Analytics

It clarifies the causal factors that contributed to an event occurring in the past. Organizations that utilize diagnostic analytics are more inclined to comprehend the causal connections between actions so they acquire a profound understanding of the underlying factors contributing to events. The research utilizes numerous techniques including data discovery, data mining, and correlation methods. Additionally, statistical expressions such as probability, likelihood, and distribution of outcomes are employed in the analysis.

2.3 Prescriptive Analytics

Analysts attempt to determine the appropriate course of action or strategies for maximizing forthcoming prospects based on a provided dataset. Despite utilizing advanced machine learning algorithms and adhering to established business guidelines, prescriptive analytics is not devoid of limitations. The efficacy of prescriptive analytics is contingent upon the extent to which the model integrates external data alongside the internal data set of the organization.

2.4 Predictive Analytics

Predictive analytics is a field of study that focuses on the extraction and analysis of data with the objective of making predictions about an event of interest, often pertaining to future events. The utilization of data-driven learning technologies for the purpose of predicting these unidentified occurrences has the potential to enhance decision-making processes. Predictive models have the capability to identify patterns and correlations within datasets, enabling organizations to make informed predictions on future outcomes, grounded in empirical evidence rather than mere conjecture. Consequently, the primary objective of predictive analytics is to improve the process of human decision-making by supplementing or surpassing dependence on human knowledge, personal experience, and subjective intuition. The incorporation of predictive modeling within this objective has the potential to yield advantageous outcomes in mitigating both implicit and explicit biases. The predictive analytics process involves a series of consecutive steps:

2.4.1 Identification

A critical first step is identifying the problem and determining the outcome and objectives. The capacity to determine the purpose of the problem will assist in the selection of suitable data to be utilized for the model.

2.4.2 Data Collection

Data mining techniques are employed to facilitate the preparation of data for analysis through the storage and manipulation of data obtained from multiple sources. A notable characteristic of data mining is its comprehensive documentation of all linkages or correlations that can potentially be identified within the dataset, irrespective of their source. The utilization of statistical or machine learning algorithms is capable of detecting patterns, establishing correlations within data, and making predictions for novel data, constituting an integral component of the predictive analytics process. Data mining can be used to gather knowledge about data relationships, which can then be applied in predictive modeling.

2.4.3 Data Analysis

Data analysis is the systematic process of examining, cleaning, and modeling data with the aim of uncovering significant insights and information. Statistics play a crucial role in the process of validating assumptions and testing hypotheses during the examination of data. Statistics enables the examination of purposeful and targeted associations within data through the utilization of sophisticated statistical methodologies, including multivariate analytic approaches such as advanced regression or time series models.

Regression models are one of the most widely used predictive analytics techniques. These models provide a mathematical representation of the relationship between the predictor variable and the outcome variable. Machine learning techniques have their origins in several academic disciplines, such as artificial intelligence, wherein their initial purpose was to facilitate the development of computer learning capabilities. In contrast to conventional statistical approaches, which often necessitate certain data qualities and generally rely on a limited set of essential variables to generate outcomes, machine learning models utilize a computer-based methodology that incorporates numerous elements to identify similarities and patterns within the data. In general, these models tend to prioritize enhanced predictive accuracy by leveraging a wider array of unstructured data, such as text and images, at the expense of interpretability.

2.4.4 Modeling

Identifies patterns and relationships in data and predicts future outcomes based on those patterns and relationships. A predictive model's main assumption is that a future event will occur in the same way that previous events have. Some researchers contend that this assumption is flawed because past behavior does not always predict future behavior.

2.4.5 Model Development and Monitoring

These are the last steps in the predictive analytics process. The process of model development involves integrating the findings of analytics into the decision-making process. For instance, using a predictive tool to discern a discernible pattern that portrays the probability that a well to produce oil. Once this established pattern has been identified, the model should be utilized to predict the future risk associated with production. Model monitoring is a crucial practice employed to effectively oversee and evaluate the performance of a model, with the primary objective of verifying its appropriate functioning. The premise that past production events may accurately predict future production events is inherently erroneous. While certain reservoirs may exhibit homogeneity in their behavior, it is important to note that these homogeneities are not necessarily absolute. Even though some reservoirs are homogenous in their routines, these homogeneities are not absolute, and reservoir changes can occur, invalidating the model used to predict the well. Model deployment and monitoring may have an impact on decision-making; therefore, ensuring an accurate, valid model is critical. It is important to acknowledge that the utilization of models should not be confined to decision-making in isolation, but rather as a complementary tool to inform and support decision-making processes.

3 Methodology

In this study, we used some supervised machine learning techniques, which are Decision Tree (DT), Random Forest (RF), Multiple Linear Regression (MLR) and K-nearest neighbors (KNN). All these models were developed in Python.

3.1 Decision Tree

The decision tree is a supervised learning method that predicts values of responses by learning decision rules derived from data The decision tree constructing algorithm works top–down at every node, by choosing the best variable that best splits the current training subset according to the homogeneity of the target variable within the subsets. [19]. In general, it is a nonparametric method that is used for both classification and regression problems. Also, the decision tree is non-parametric since it does not suppose any distributional properties about the data [20]. The decision tree consists of decision nodes and terminal leaves where each node *n* implements a function, that gives discrete outcomes with branches [20]. It has two types which are classification and regression trees. The classification tree applies to problems where the output data is discreet whereas RT applies to problems where the output is continuous [13, 20]. As a result, the total DT branch of this work can be depicted in Fig. 1.

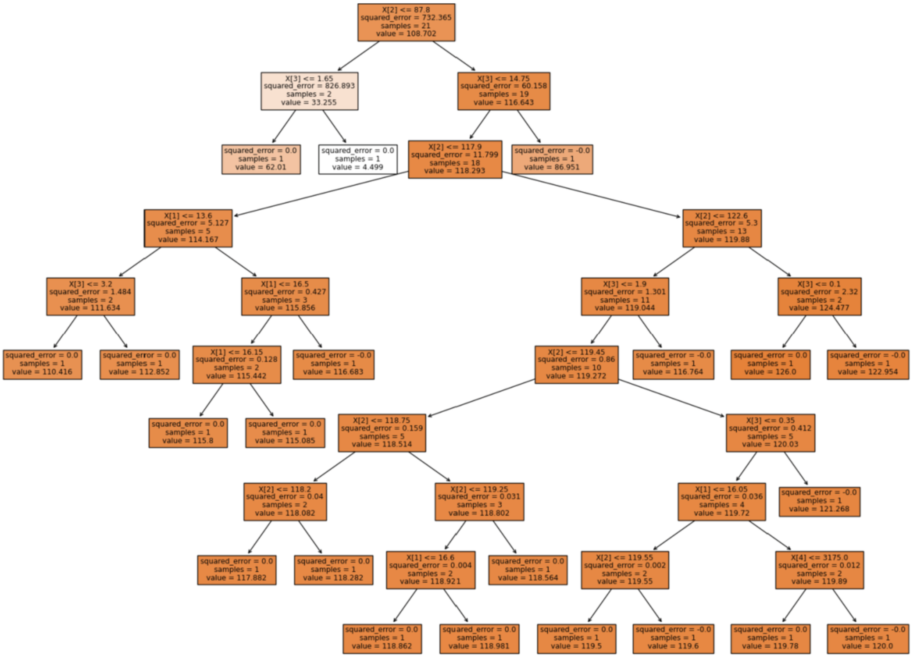
X[2] <= 87.8 squared_error = 732.365 samples = 21 value = 108.702
X[3] <= 1.65 squared_error = 826.893 samples = 2 value = 33.255
X[3] <= 14.75 squared_error = 60.158 samples = 19 value = 116.643
squared_error = 0.0 samples = 1 value = 62.01
squared_error = 0.0 samples = 1 value = 4.499
X[2] <= 117.9 squared_error = 11.799 samples = 18 value = 118.293
squared_error = -0.0 samples = 1 value = 86.951
X[1] <= 13.6 squared_error = 5.127 samples = 5 value = 114.167
X[2] <= 122.6 squared_error = 5.3 samples = 13 value = 119.88
X[3] <= 3.2 squared_error = 1.484 samples = 2 value = 111.634
X[1] <= 16.5 squared_error = 0.427 samples = 3 value = 115.856
X[3] <= 1.9 squared_error = 1.301 samples = 11 value = 119.044
X[3] <= 0.1 squared_error = 2.32 samples = 2 value = 124.477
squared_error = 0.0 samples = 1 value = 110.416
squared_error = 0.0 samples = 1 value = 112.852
X[1] <= 16.15 squared_error = 0.128 samples = 2 value = 115.442
squared_error = -0.0 samples = 1 value = 116.683
X[2] <= 119.45 squared_error = 0.86 samples = 10 value = 119.272
squared_error = -0.0 samples = 1 value = 116.764
squared_error = 0.0 samples = 1 value = 126.0
squared_error = -0.0 samples = 1 value = 122.954
squared_error = 0.0 samples = 1 value = 115.8
squared_error = 0.0 samples = 1 value = 115.085
X[2] <= 118.75 squared_error = 0.159 samples = 5 value = 118.514
X[3] <= 0.35 squared_error = 0.412 samples = 5 value = 120.03
X[2] <= 118.2 squared_error = 0.04 samples = 2 value = 118.082
X[2] <= 119.25 squared_error = 0.031 samples = 3 value = 118.802
X[1] <= 16.05 squared_error = 0.036 samples = 4 value = 119.72
squared_error = -0.0 samples = 1 value = 121.268
squared_error = 0.0 samples = 1 value = 117.882
squared_error = 0.0 samples = 1 value = 118.282
X[1] <= 16.6 squared_error = 0.004 samples = 2 value = 118.921
squared_error = 0.0 samples = 1 value = 118.564
X[2] <= 119.55 squared_error = 0.002 samples = 2 value = 119.55
X[4] <= 3175.0 squared_error = 0.012 samples = 2 value = 119.89
squared_error = 0.0 samples = 1 value = 118.862
squared_error = 0.0 samples = 1 value = 118.981
squared_error = 0.0 samples = 1 value = 119.5
squared_error = -0.0 samples = 1 value = 119.6
squared_error = 0.0 samples = 1 value = 119.78
squared_error = -0.0 samples = 1 value = 120.0

Fig. 1 Decision tree regression model of prediction oil production rate

3.2 K-Nearest Neighbors

K-nearest neighbors KNN algorithm is the simplest and non-parametric supervised learning approach that can be utilized for both classification and predictions. In both regression and classification, the input parameters are composed of the positive integer k closest training datasets within a feature space [13]. Also, KNN determines the class of a new object based on majority votes among k number of neighbors of the objects [21]. The privileges of the KNN method are the simplicity of its applications when solving complex tasks, its ability to reverse calibration, and the need not to re-estimate the model when there are additional new objects to the training data [13].

3.3 Multiple Linear Regressions

Multi Linear regression is a statistical technique that evaluates the linear connection between two or more variables. The predicted outcome is called the dependent variable and the variables on which the outcome is based are called the independent variables. Ordinary linear regression and multiple linear regression are the two main types of linear regression. The difference between the two is related to the number of explanatory variables used to predict the outcome variable.

Ordinary linear regression models the linear relationship between two variables (a dependent variable and an independent variable) by fitting a line of best fit that closely approximates the data points. The best-fit line is determined by squaring the vertical distance between each potential line and the closest point (least squares method).

The best-fit line is the line that reduces the sum of the squared intervals [22]. Therefore, ordinary linear regression predicts the outcome variable with a single explanatory variable.

On the other hand, multiple linear regressions rely on multiple explanatory variables to quantify the dependent variable. Establishes the linear correlation between many independent variables and an outcome variable. Instead of a line of best fit, several regression algorithms use least squares to determine the plane (p-dimensional plane) that best fits the scatterplot [22]. To perform p-dimensional plane fitting,"the sum of squares of the deviations of the points from the plane is reduced" [22]. The optimized line for each explanatory variable is determined by calculating the regression coefficient that produces the smallest residual and the t-statistic of the algorithm and the associated p-value [23].

This makes multiple linear regression more suitable for real-world scenarios because it can be used to model the effects of many independent parameters on a given outcome while quantifying the relative contribution of each individual explanatory variable in the data set [24].

Multiple linear regression algorithms can be summarized by the equation below;

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_P + \varepsilon \quad (1)$$

where the terms:

Y = response variable.

c = y-intercept; the predicted outcome when all other variables equal zero.

$\beta_1 \& \beta_2 =$ regression coefficients of the first, second, and third independent variables $x_1 \& x_2$ respectively.

$\beta_P =$ regression coefficient of the last explanatory variable x_P

ε = residuals.

Assumptions for Multiple Linear Regressions

a) The correlation between the dependent variable and each individual term of the independent variables must be linear.
b) The statistical approach for sampling the observations must be independent and devoid of obvious linear correlation among the independent variables in the dataset.
c) The residuals must follow a normal distribution.
d) The variance must be constant and the size of the residual error must be similar (homoscedasticity).

3.4 Random Forest

"Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [25]. As a supervised learning algorithm, random forest can handle classification and regression problems. It operates on a concept known as ensemble learning, which is basically a technique that integrates several classifiers to address complicated problems ("Random Forest," n.d.). Random forest performs similarly to boosting on a wide range of problems. In addition, they are easy to train and refine. Numerous decision trees make up a random forest algorithm. The generated cluster of trees that constitute the algorithm is trained via bagging or bootstrap aggregating [26]. To issue the final outcome, the algorithm depends on each individual tree class prediction. The final prediction is the average or majority by votes of the trees. For a more accurate prediction of final outcomes, the number of uncorrelated trees in the algorithm should be increased since as the population of trees increases, the likelihood of predicting incorrectly dramatically decreases.

Because random forest makes predictions based on the mean or majority by votes of several decision trees, it is possible some trees may predict wrongly while others may predict correctly. Therefore, in order for random forest to perform better by predicting accurate results rather than an assumed outcome, there has to be some real values or signals included in the feature variable of the dataset and each tree's prediction must have extremely low correlations ("Random Forest," n.d.).

4 Results and Discussions

In this research, we introduced four powerful and dependable soft computing algorithms for predicting the oil production rate of a real oilfield dataset of 11 wells. The suitable selection of input parameters in the machine-learning model is a crucial task. Therefore, in accordance with the majority of published machine-learning-based research, the following input parameters were selected: production time, casing pressure, liquid volume, water cut, gas pressure, and gas-oil ratio (GOR), while the oil production rate is defined as the output parameter.

The framework of this research was developed using four different datasets from the eleven wells. We performed eleven groups of experiments on eleven wells with the training and testing of the ML algorithms. In this study, the multiple linear regressions (MLR), RF, DT, and KNN are the selected machine-learning models. To assess and compare the performance of these developed models, two statistical metrics were perfectly applied for comparing the training and testing dataset stages of all predictive models. More so, to evaluate discrete models for overfitting, the developed MLR, RF, DT, and KNN models' performances for predicting oil production rate for all wells' datasets include Well-S95, Well-KT905H, Well-KT906H, Well-KT907H, Well-KT908DH, Well-KT909H, Well-KT910H, Well-KT911H, T912CH, TK918, and TK919H. Overall, the proposed ML models were evaluated for each well dataset using the R^2 and RMSE metrics. Table 2 highlights the statistical metrics outcomes for training and testing.

Consequently, it is reported that the RF model is tuned by a kernel function, which has a direct impact on results accuracy in all eleven wells: for Well-KT911H, the RMSE of RF is the lowest (0.868) compared to that of DT and KNN of (1.866 and 5.599), respectively. Furthermore, its determination-coefficient score R^2 is very close to the standard precise value of 1 (0.9993) and much higher than that of DT, LR, and KNN (0.9969, 0. 9684, and 0. 9720), respectively. The RF and DT models for Well- KT908DH have the highest R^2 values (0.9991, 0.9977) and the lowest RMSE (1.560, 2.572), respectively. The RF and DT models for Well-S95 have the highest R^2 values (0.999 and 0.997) and the lowest RMSE (0.902, 1.338) respectively. These three wells are ordered respectively in term of their results of high accuracy. Thus, the evaluation results indicated that the proposed RF model performs the optimal accuracy, followed by the DT model.

Table 2 Statistical metrics results of ML models performance for training and testing sets

Statistical Metrics		RMSE R^2		RMSE R^2	
Well	Model	Training set		Testing set	
Well-S95	MLR DTM RFM KNN	3.184	0.982	3.48	0.978
		0	1	1.338	0.997
		0.32	1	0.902	0.999
		0	1	4.162	0.969
Well-KT905H	MLR DTM RFM KNN	4.006	0.935	4.439	0.927
		0	1	1.2	0.995
		0.291	1	1.164	0.995
		0	1	3.875	0.945
Well-KT906H	MLR DTM RFM KNN	4.419	0.965	5.036	0.957
		0	1	1.95	0.993
		0.266	1	1.089	0.998
		0	1	4.423	0.967
Well-KT907H	MLR DTM RFM KNN	6.111	0.942	6.57	0.932
		0	1	1.536	0.996
		0.323	1	1.222	0.998
		0	1	6.747	0.928
Well-KT908DH	MLR	9.903	0.9610	10.278	0.9625
	DTM	0.000	1.0000	2.572	0.9977
	RFM	0.637	0.9998	1.560	0.9991
	KNN	0.000	1.0000	9.742	0.9663
Well- KT909H	MLR	5.472	0.871	5.346	0.8742
	DTM	0.000	1.000	0.834	0.9969
	RFM	0.529	0.999	0.581	0.9985
	KNN	0.000	1.000	4.492	0.9111
Well- KT910H	MLR	4.896	0.943	4.578	0.9477
	DTM	0.000	1.000	1.355	0.9954
	RFM	0.480	0.999	0.812	0.9984
	KNN	0.000	1.000	3.003	0.9775
Well-KT911H	MLR	5.872	0.972	5.948	0.9684
	DTM	0.000	1.000	1.866	0.9969
	RFM	0.437	1.000	0.868	0.9993
	KNN	0.000	1.000	5.599	0.9720

(*continued*)

Table 2 (*continued*)

Statistical Metrics		RMSE R^2		RMSE R^2	
Well	Model	Training set		Testing set	
Well-T912CH	MLR	4.180	0.877	4.213	0.8714
	DTM	0.000	1.000	1.162	0.9902
	RFM	0.250	1.000	0.860	0.9946
	KNN	0.000	1.000	1.590	0.9817
Well-TK918	MLR	3.778	0.856	4.031	0.8740
	DTM	0.000	1.000	0.871	0.9941
	RFM	0.337	0.999	0.678	0.9964
	KNN	0.000	1.000	1.604	0.9801
Well-TK919H	MLR	2.403	0.895	2.214	0.8956
	DTM	0.000	1.000	0.724	0.9888
	RFM	0.275	0.999	0.509	0.9945
	KNN	0.000	1.000	1.966	0.9177

Furthermore, the prediction visual plots by all proposed models were presented, and the comparative configuration plot of real oil production rate and predicted values are visualized for the datasets of each well as illustrated in Fig. 2. The oil production rate values of test and train data for each well are highlighted graphically in the subplots of Fig. 2. Obviously, the RF model proved its high precision in Fig. 2 by accumulating the data points that are identically close to the 45° slop line. The results indicated that the DT, KNN, and MLR also predict oil production rate with high accuracy due to the larger quantity of data points presented in the Well-S95 and Well-KT211H, but by comparison, the RF is the proposed method achieved the most optimized predictive performance by taking into account variant tendency and related data of oil production rate.

The proposed predictive RF method also provides a perfect oil production rate prediction for the other wells, particularly in Well-KT908DH and Well-S95. This study demonstrated the superiority of the RF model in predicting the oil production rate for the eleven wells. However, in the case of testing conditions shown in Fig. 3 (a), and (b), the performance reference models (DT, MLR, and KNN) improved slightly. The RF provided an optimal accurate prediction by taking the highest place of the R^2 and the lowest place of the RMSE. As demonstrated in Fig. 3, the RF and DT models outperformed the other ML models.

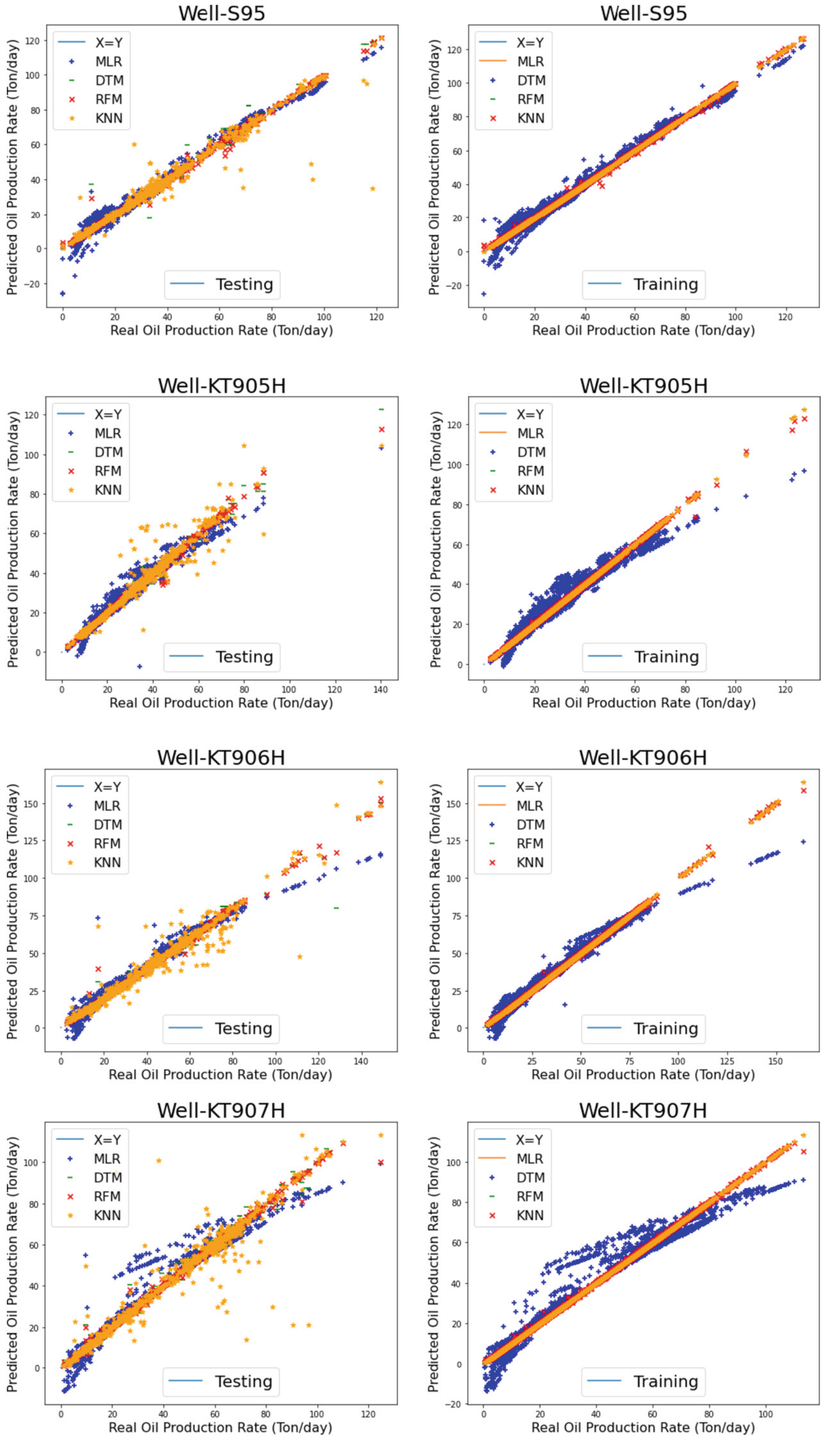

Fig. 2 A visual plot of predicted versus real oil production rate values for testing and training datasets of all wells

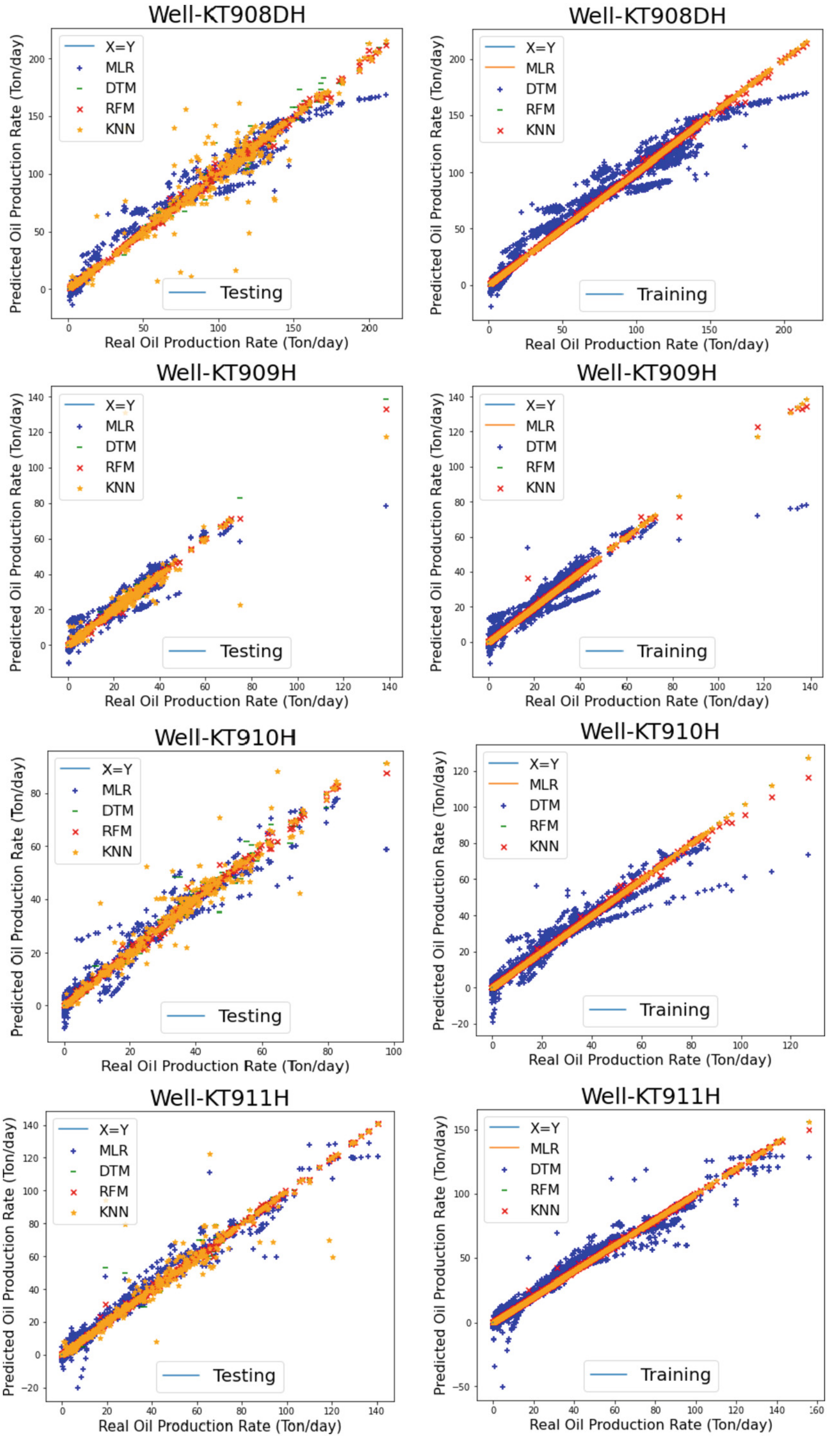

Fig. 2 (*continued*)

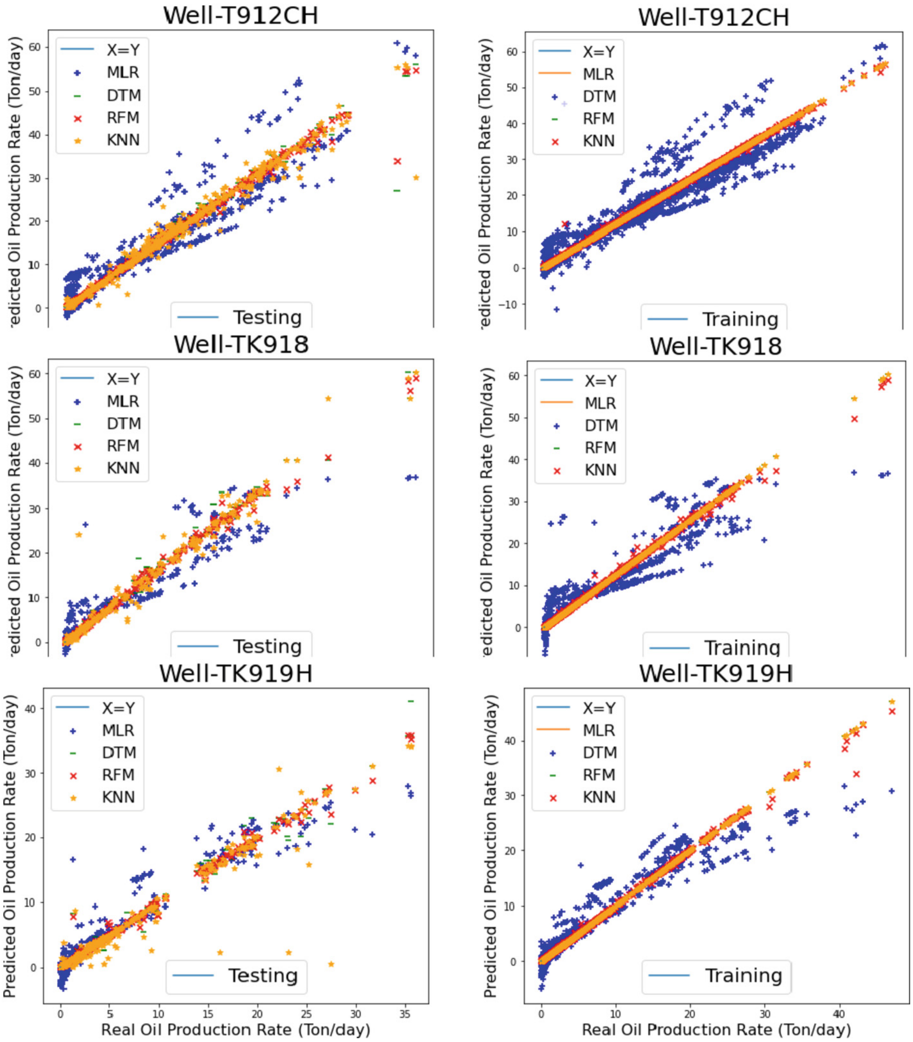

Fig. 2 (*continued*)

Figure 3(a) revealed that the RF model demonstrated the best accurate results based on R^2 error in Well-KT911H, Well-KT908DH, Well-S95, and Well-KT909H, respectively. Figure 3(b) further showed that the RF model proved the validated results based on RMSE in the Well-TK919H, Well-KT909H, Well-TK918, and Well-T912CH, respectively. In addition, the decision tree model (DT) is validated as the second-best algorithm for predicting production performance, which demonstrated better results based on R^2 error for Well-KT908DH, Well-KT909H, Well-KT911H, and Well-S95, as represented in Fig. 3(a). According to Fig. 3(b), the performance path of the DT model based on RMSE is different which provided accurate results in the Well-TK919H, Well-KT909H, Well-TK918, and Well-KT912CH.

A possible reason for the weak predictive performance of the other ML models might be because they depend on the effect underlying the data quantity, where the ML technique discovers the relations through data mining.

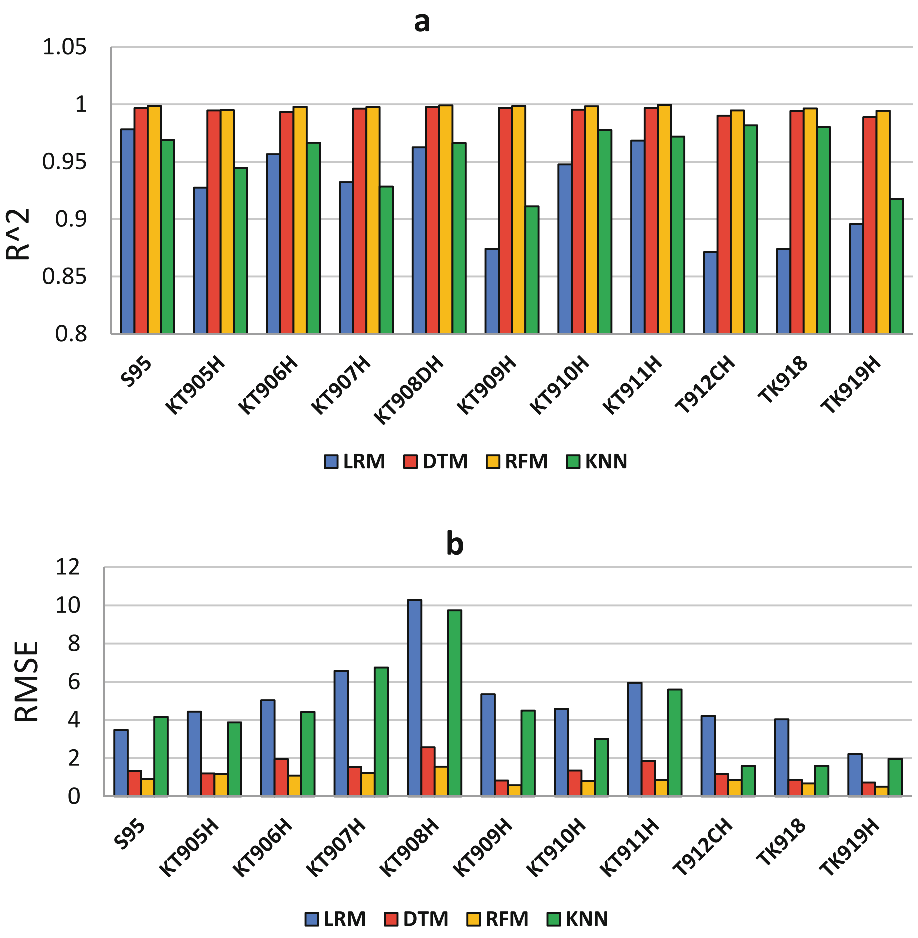

Fig. 3 Comparison performance results of the (a) R^2, and (b) RMSE results of proposed ML models for the testing dataset of all wells

In this study, the input parameters selection and importance rely on the best-proposed predictive model. Thus, a fitting model should be used to identify input parameters with the greatest impact. However, the RF method was used to compute the input parameters' importance ranking and permutation importance for only testing datasets shown in Fig. 4. From the subplots of Fig. 4, the first and second most effective input parameters for oil production rate prediction in Well-S95 are liquid volume and water cut, followed by gas pressure and casing pressure. The water cut is the most influential parameter affecting oil production rate in all other wells, followed by liquid volume, casing pressure, gas pressure, and GOR parameter except Well-KT908DH and Well-TK918. In Well-KT908DH, we can see that the water cut is the most influencing parameter, followed by casing pressure. In Well-TK918H, we can see that the water cut is the most influencing parameter, followed by gas pressure. Therefore, as shown in Fig. 4, the production time parameter has no effect on the oil production rate in all wells followed by GOR and casing pressure. Overall, the other input parameters have non-relative importance in the RF model for predicting oil production rate in all wells.

Importantly, by estimating the drop in the model score, the permutation parameter importance can evaluate input parameters and oil production rate. Therefore, the average permutation importance score of parameters is listed in Table 3. The greater effect of the input parameter on the model can be established when the model score displayed a dramatic dropping and a higher ranking is assigned, as shown in subplots of Fig. 4. Accordingly, the liquid volume, water cut, and gas pressure parameters in the RF model for each well have high permutation importance scores, implying that they have the

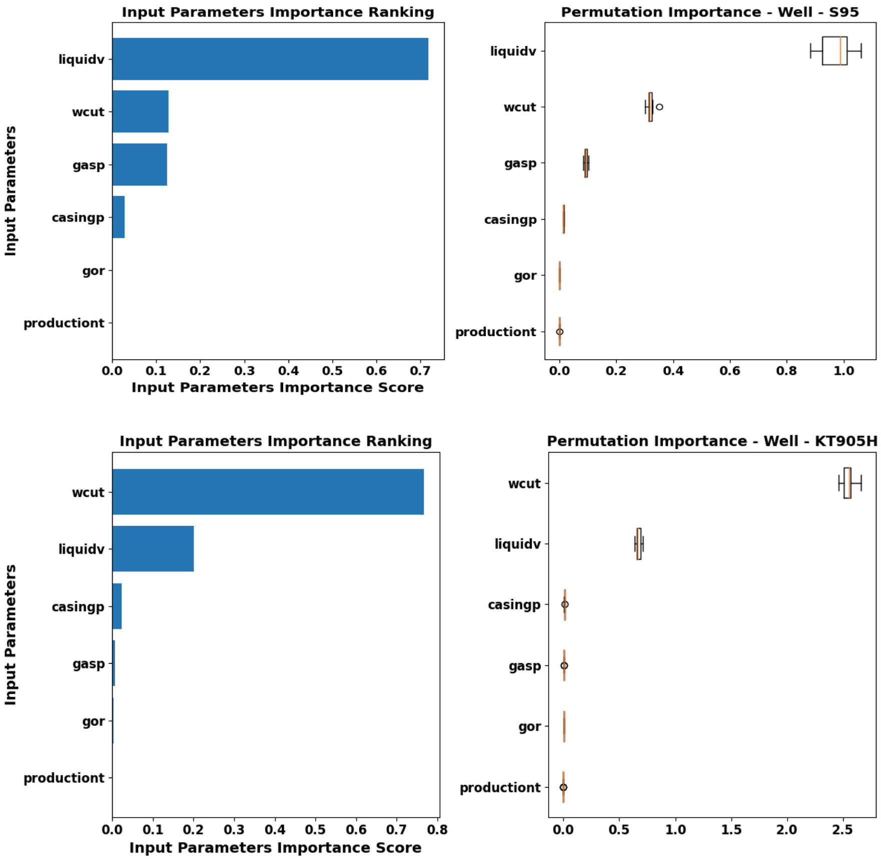

Fig. 4 Importance of input parameters for the prediction of oil production rate in each well using random forest model

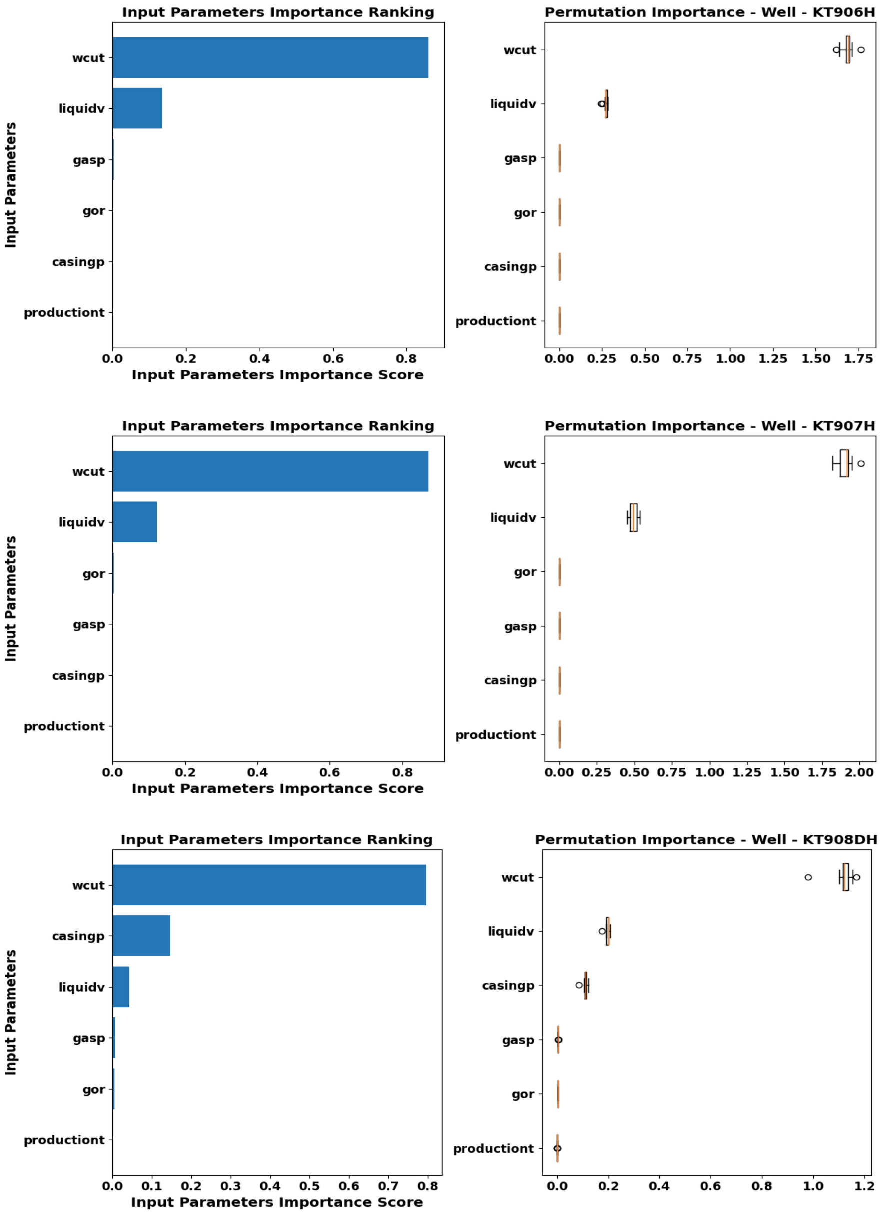

Fig. 4 (*continued*)

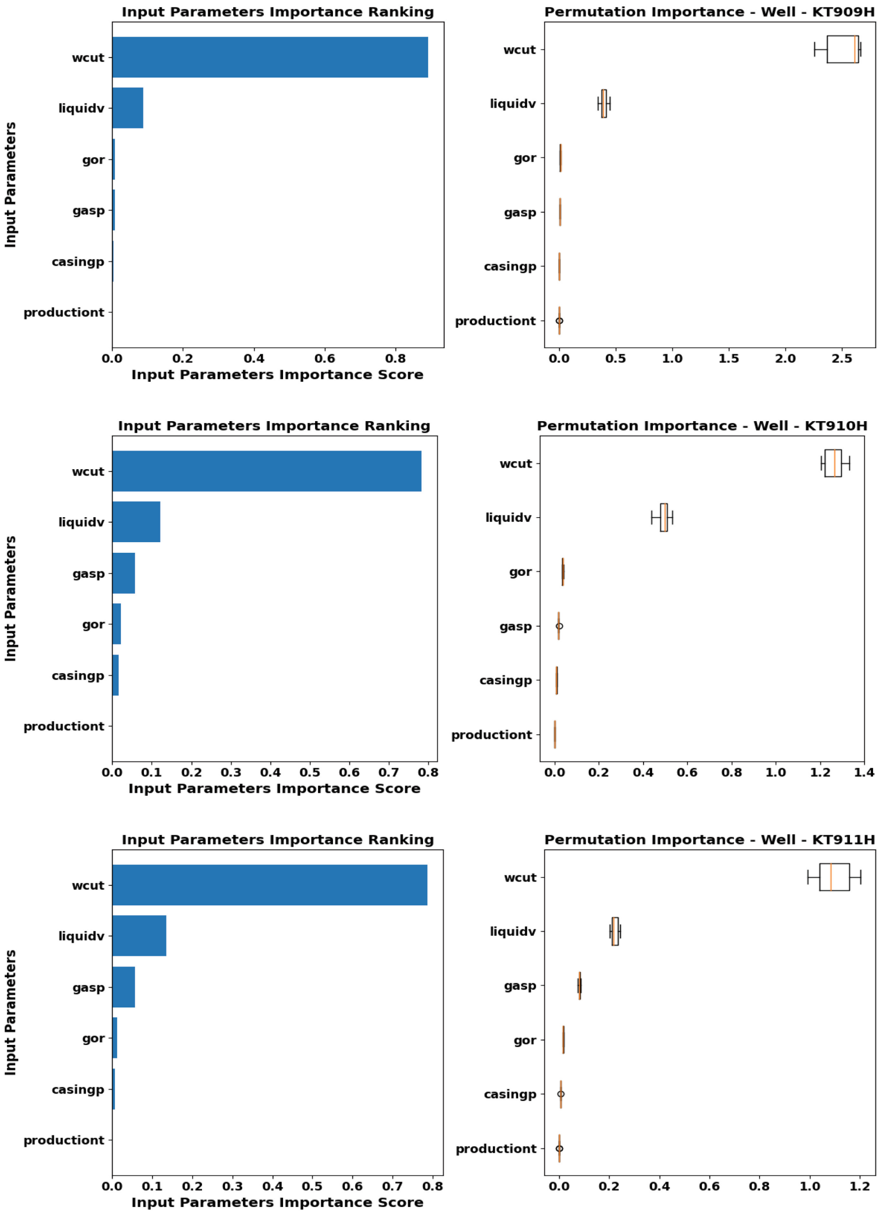

Fig. 4 (*continued*)

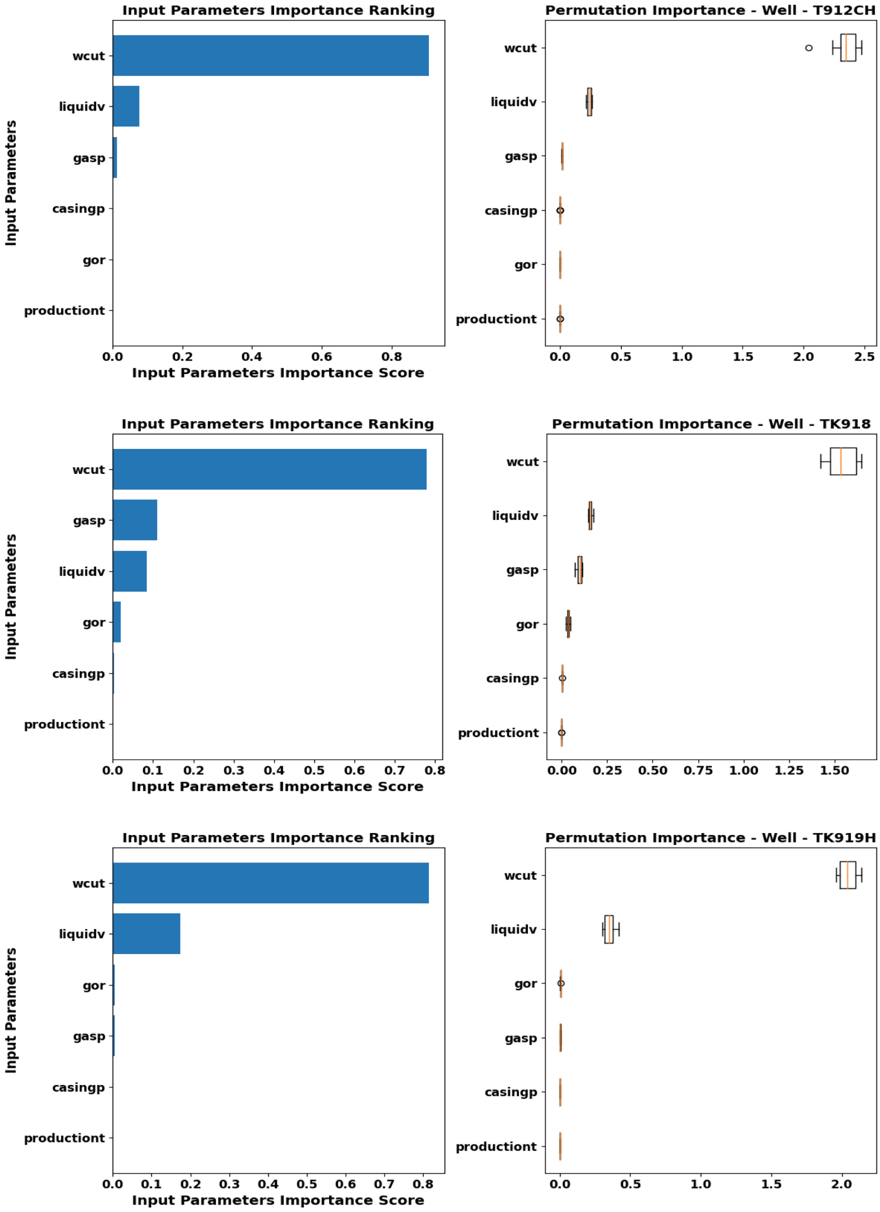

Fig. 4 (*continued*)

greatest impact on the oil production rate. As shown in Table 3 and Fig. 4, production time, GOR, and casing pressure were clearly less important parameters in oil production rate prediction. Furthermore, as shown in Table 3, the production time has a negative impact on the oil production rate in wells KT911H. However, if the production time parameters increase in well KT911H the oil production rate would decrease. Overall, the permutation importance not only demonstrates a parameter's predictive fitness, but it also validates how important a parameter is for a model.

Table 3 Average permutation importance of input parameters on the oil production rate using RF model

Average Permutation Importance Scores						
	GOR	Gas Pressure	Water Cut	Liquid Volume	Casing-Pressure	Production-Time
Well-S95	0.00015	0.09325	0.31995	0.97466	0.01443	0.00005
Well-KT908DH	0.00429	0.00508	1.1153	0.19719	0.11043	0.00000
Well-KT909H	0.01155	0.00894	2.5202	0.39682	0.00369	0.00000
Well-KT911H	0.01517	0.08111	1.0935	0.22028	0.00465	-0.00000

In general, an oil production rate prediction model has been developed for all wells in the area based on the ML models developed, which have shown that the RF model is the most effective for predicting oil production rates in all eleven considered wells. It is interesting to note, however, that Well-S95, which has the highest quantity of data points (4560), exhibited a slightly lower R^2 value (0.999) compared to Well-KT911H, which has 3190 data points (R^2 of 0.9993). This observation suggests that the model may not always benefit from an increase in the number of data points beyond a certain threshold. It may also imply that the different locations and times of the data points could play a role in the model's performance, rather than just the sheer quantity. Thus, each well has its unique features that impact the model performance, and imbalanced data can affect the model's accuracy differently for each well.

5 Conclusion

In this work, MLR, KNN, DT, and RF models have been engaged to predict the oil production rate of real oilfield data for 11 wells in the oilfield, each with varying quantities of data points. Overall, the model demonstrated satisfactory performance, with good accuracy for most of the wells studied. These wells are (Well-S95, Well-KT905H, Well-KT906H, Well-KT907H, Well-KT908DH, Well-KT909H, Well-KT910H, Well-KT911H, Well-T912CH, Well-TK918, and Well-TK919H). All the aforementioned models were trained and tested to verify that the models learned the rapport between the input and output data. Thereafter, all the datasets were split in the same distribution as 75% of the data points, which were employed for the training set and the remaining points for testing the developed models. According to the calculated and tabulated results of statistical errors for all models in the study analysis, the built models display their performance precisely.

In general, DT and KNN models equally outperformed the other two models in terms of training performance in all wells, with $\mathbf{R^2}$ and RMSE (1, 0). On the contrary, the RF model performed the best outcome in the testing performance set in all wells as verified with other ML models. Precisely, the performance metrics indicated that the RF model performed the best accurate outcomes in Well-KT911H well with the R^2 and RMSE values of (0.9993, and 0.868), respectively. Additionally, the DT proved its second-order predictive model in the testing performance set in all wells.

In this work, the RF model was used as a selected model to study the relationship between the input parameters and oil production rate. It was deduced that the liquid volume, water cut, and gas pressure are the most important parameters affecting the oil production rate performance in all wells, whereas production time, GOR, and casing pressure are proven less important parameters for the final prediction.

It has been found that the liquid volume parameter has an impact on the accuracy of the model for Well-S95, which has the highest quantity of data points. This well exhibited slightly lower accuracy than Well-KT911H, which has less data points. However, the liquid volume was considered the most influential feature affecting oil production rate performance.

Declaration of Interest. The authors declare no conflict of interest regarding the publication of this paper.

References

1. Alrassas, A.M., Thanh, H.V., Ren, S., Sun, R., Al-areeq, N.M., Kolawole, O., et al.: CO_2 sequestration and enhanced oil recovery via the water alternating gas scheme in a mixed transgressive sandstone- carbonate reservoir : case study of a large middle east oilfield. Energy Fuels **36**, 10299–10314 (2022). https://doi.org/10.1021/acs.energyfuels.2c02185
2. AlRassas, A.M., Vo Thanh, H., Ren, S., Sun, R, Le Nguyen Hai, N., Lee, K.K.: Integrated static modeling and dynamic simulation framework for CO_2 storage capacity in Upper Qishn Clastics, S1A reservoir, Yemen. Geomech. Geophys. Geo-Energy Geo-Res. **8**, 1–23 (2022). https://doi.org/10.1007/s40948-021-00305-x.
3. Al-Mudhafar, W.J., Abbas, M.A., Wood, D.A.: Performance evaluation of boosting machine learning algorithms for lithofacies classification in heterogeneous carbonate reservoirs. Mar. Pet. Geol. **145**, 105886 (2022). https://doi.org/10.1016/j.marpetgeo.2022.105886
4. AlRassas, A.M., Ren, S., Sun, R., Thanh, H.V., Guan, Z.: CO_2 storage capacity estimation under geological uncertainty using 3-D geological modeling of unconventional reservoir rocks in Shahejie formation, block Nv32 China. J. Pet. Explor. Prod. **11**, 2327–2345 (2021). https://doi.org/10.1007/s13202-021-01192-4
5. Al-Qaness, M.A.A., Elaziz, M.A., Ewees, A.A.: Oil consumption forecasting using optimized adaptive neuro-fuzzy inference system based on sine cosine algorithm. IEEE Access **6**, 68394–68402 (2018). https://doi.org/10.1109/ACCESS.2018.2879965
6. Alrassas, A.M., Al-Qaness, M.A.A., Ewees, A.A., Ren, S., Elaziz, M.A., Damaševičius, R., et al.: Optimized anfis model using aquila optimizer for oil production forecasting. Processes **9**, 1–17 (2021). https://doi.org/10.3390/pr9071194
7. Al-qaness, M.A.A., Ewees, A.A., Fan, H., AlRassas, A.M., Abd, E.M.: Modified aquila optimizer for forecasting oil production. Geo-Spatial Inf. Sci. **25**, 1–17 (2022). https://doi.org/10.1080/10095020.2022.2068385
8. AlRassas, A.M., Al-qaness, M.A.A., Ewees, A.A., Ren, S., Sun, R., Pan, L., et al.: Advance artificial time series forecasting model for oil production using neuro fuzzy-based slime mould algorithm. J. Pet. Explor. Prod. Technol. **12**, 383–395 (2022). https://doi.org/10.1007/s13202-021-01405-w

9. Syed, F.I., Alshamsi, M., Dahaghi, A.K., Neghabhan, S.: Artificial lift system optimization using machine learning applications. Petroleum **8**, 219–226 (2022). https://doi.org/10.1016/j.petlm.2020.08.003
10. Bowie, B.: Machine learning applied to optimize duvernay well performance. In: Society of Petroleum Engineers - SPE Canada Unconventional Resources Conference, URC 2018 (2018). https://doi.org/10.2118/189823-ms
11. Ertekin, T.: Heavy Crude Oil Recovery, pp. 379–380 (1984). https://doi.org/10.1007/978-94-009-6140-1
12. Botao, L.: Discussion on current application of artificial intelligence in petroleum industry (2019)
13. Deumah, S.S., Yahya, W.A.: SPE-208667-MS Prediction of Gas Viscosity of Yemeni Gas Fields Using Machine Learning Techniques (2021)
14. Alalimi, A, Pan, L., Al-Qaness, M.A.A., Ewees, A.A., Wang, X., Abd Elaziz, M.: Optimized random vector functional link network to predict oil production from Tahe OIL FIELD in China. Oil Gas Sci. Technol. **76** (2021). https://doi.org/10.2516/ogst/2020081
15. AL-Alimi, D., AlRassas, A.M., Al-qaness, M.A.A., Cai, Z., Aseeri, A.O., Abd Elaziz, M., et al.: TLIA: time-series forecasting model using long short-term memory integrated with artificial neural networks for volatile energy markets. Appl. Energy **343**, 121230 (2023). https://doi.org/10.1016/j.apenergy.2023.121230
16. Huang, R., Wei, C., Wang, B., Yang, J., Xu, X., Wu, S., et al.: Well performance prediction based on Long Short-Term Memory (LSTM) neural network. J. Pet. Sci. Eng. **208**, 109686 (2022). https://doi.org/10.1016/j.petrol.2021.109686
17. Panja, P., Pathak, M., Velasco, R., Deo, M.: Least square support vector machine: an emerging tool for data analysis. In: SPE Rocky Mountain Petroleum Technology Conference Reservoirs Symposium, SPE 2016, p. SPE-180202 (2016)
18. Wang, S., Chen, S.: Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. J. Pet. Sci. Eng. **174**, 682–695 (2019). https://doi.org/10.1016/j.petrol.2018.11.076
19. Osarogiagbon, A.U., Khan, F., Venkatesan, R., Gillard, P.: Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. Process. Saf. Environ. Prot. **147**, 367–384 (2021). https://doi.org/10.1016/j.psep.2020.09.038
20. Oloso, M.A.: Prediction of Reservoir Fluid Properties Using Machine Learning (2018)
21. Balas, V.E.: Intelligent systems reference library 172 recent trends and advances in artificial intelligence and internet of things (n.d.)
22. Uyanık, G.K., Güler, N.: A study on multiple linear regression analysis. Procedia-Soc. Behav. Sci. **106**, 234–240 (2013)
23. Eberly, L.E.: Multiple linear regression. Top. Biostat. 165–187 (2007)
24. Neethu, T.S., Sabu, A.S., Mathew, A., Wakif, A., Areekara, S.: Multiple linear regression on bioconvective MHD hybrid nanofluid flow past an exponential stretching sheet with radiation and dissipation effects. Int. Commun. Heat Mass Transf. **135**, 106115 (2022)
25. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
26. Tony, Y.: Understanding Random forest: How the Algorithm Works and Why It Is So Effective (2019)