FUDMA Journal of Engineering and Technology
ISSN: 3092-9385
https://fjet.fudutsinma.edu.ng/index.php/fjet/
Volume 2, Issue 1, 2026, 367-380

Machine Learning Models for Predicting Flow Rate for Niger Delta Oil Wells

Olusiji A. ADEYANJU[1*], Joseph O. OLAIDE[2]

[1*, 2]Department of Petroleum and Gas Engineering, University of Lagos, Lagos, Nigeria

[1*]oadeyanju@unilag.edu.ng, [2]josepholaide10@gmail.com

Abstract

Most choke correlations used to determine oil production rates through chokes are invalid for most fields. Due to the complexity of multiphase flow behaviour, varying process conditions of oil wells, and limited data used to develop these correlations, they do not accurately predict oil flow rates in many of the Nigerian oil wells. This study presents an analytical method for improved oil flow rate estimation in Nigerian oil wells employing machine learning using the typical Gilbert choke input parameters (flowing tubing head pressure, choke size, and gas-liquid ratio) with the addition of flowing well temperature and basic sediment and water content. Six non-linear machine learning models were developed to estimate oil flow rate. These are: CatBoost, TabNet, Random Forest, XGBoost, Support vector machine with radial basis function kernel, and Gaussian process regression with radial basis function kernel. All six models outperformed existing choke correlations using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE) evaluation metrics, with CatBoost and Random Forest returning the best performance. The CatBoost model achieved an R^2 of 97%, an RMSE of 353 STBD, and a MAPE of 7.81%, while the Random Forest model achieved an R^2 of 97%, an RMSE of 369 STBD, and a MAPE of 8.55%. A parameter importance and sensitivity analyses showed that basic sediment and water content and choke size have the highest impact on the oil production rate determination. A consistent negative trend was observed in the sensitivity analysis for the basic sediment and water parameters, an indication of the need to minimize basic sediment and water levels for optimal oil production estimation. The developed model will be of significant assistance to petroleum industry operators in the Niger Delta region of Nigeria for quick effective estimates of oil flow rates.

Keywords: *Choke correlations, Multiphase flow, Machine learning, Gilbert choke input, Non-linear models, Niger Delta region.*

1.0 Introduction

Multiphase (oil, water, and gas) flow measurement in petroleum production systems is of great importance. Although significant advancements have been made in multiphase flow measurement, there is still much to be done in accuracy and robustness. The use of choke correlations in flow estimation as well as control development is a well-established method. The production choke valve as an instrument within the petroleum production system used to control multiphase flow is, commonly referred to as the wellhead choke. The wellhead choke assists in controlling the fluid flow, thereby lowering the possibility of sand transport, as well as managing the pressure difference responsible for both water coning and gas coning (Sanni et al., 2020). As a result of the significant effect of chokes on petroleum production system performance, multiphase flow through chokes needs to be studied for performance modeling of oil and gas production.

This formed the basis of this study's investigation. Several authors have studied the multiphase flow phenomenon and have developed choke correlations using wellhead choke size and fluid properties such as wellhead pressure, gas-to-liquid ratio, specific oil gravities, specific gas gravities, and basic sediment & water (Gilbert et al., 1954; Beiranvand et al., 2012; Okon et al., 2015; and Ghorbani et al., 2018). The essence of these choke correlations is to determine the oil production rate at specific fluid properties and choke size set points. This helps in oil production optimization and oil well performance analyses. However, most choke correlations created are not robust and are field-specific and perform poorly for fields with process conditions different from those for which they were developed.

Prior to oil production rate determination, it is important to identify the flow regime behaviour of the production fluids. Typically, two flow regimes are exhibited by the production fluid: critical (sonic) or sub-critical (subsonic) flow. Sonic flow occurs when a multiphase fluid flows through a choke and attains a Mach number of 1 (i.e. the fluid velocity is equal to the speed of sound in the medium) in the throat of the choke, while subsonic flow occurs when fluid flows through a choke at a velocity below the sonic velocity ($M < 1$). Under sonic conditions, pressure waves downstream are unable to move upstream, and the flow rate becomes independent of changes in downstream pressure (Guo et al., 2007). When the downstream-to-upstream pressure ratio is around 0.55–0.6, flow is critical; otherwise, it is subsonic. For subsonic flow, the flow rate is a function of pressure drop

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

across the system, while for sonic flow, it is constant regardless of downstream disturbances. Therefore, in modeling flowrate through a choke, it is required to ascertain whether the flow is sonic or subsonic.
As seen in Figure 1, it has been determined that if the ratio of after-choke pressure P2 to before-choke pressure P1 is about 0.55–0.6, it will be a critical flow, and if this ratio is greater, the flow will be sub-critical (Guo et al., 2007). In the subsonic flow case, the fluid flow rate is determined by the pressure drop in the production system. Sonic flow is considered more favorable since it allows no upstream disturbances in the choke that may lead to the compromise of the production system's flow. Therefore, in estimating the connection between flow rate and pressure drop for multiphase fluids via a choke, we must ascertain the fluid flow condition, whether the flow is subsonic or sonic.

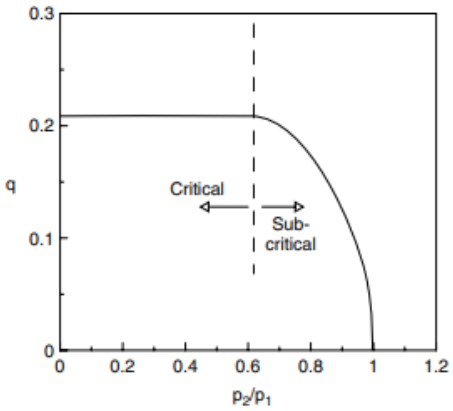

Figure 1: A diagram showing the relationship flow rate and the ratio of upstream and downstream pressure in chokes

Most developed choke correlations are only valid when fluid flows at critical flow. The Gilbert (1954) choke correlation is the most extensively utilized choke correlation for critical flow. This correlation was developed by Gilbert using over 260 well test data points from an oil field in California, USA. The wellhead pressure (sometimes called the upstream pressure), the gas-to-liquid ratio (or gas-to-oil ratio), and the wellhead choke size (or choke diameter) are the only variables that affect this correlation.
The general Gilbert-Choke correlation and its exponent is given by equation 1.

$$Q = A\frac{P_{wh}S^B}{GLR^C} \quad (1)$$

where,
P_{wh} = flowing wellhead pressure in psig,
GLR = Gas-to-liquid ratio in SCF/STB,
Q = Gross liquid rate in STB/D,
S = Choke size in (*/64 inch)
A, B, and C are empirical constants with values of 0.1, 1.89, and 0.546 respectively.

Gilbert's correlation has been modified by several other researchers using adjustments to the constants and exponents through regression parameters based on data from various oil fields. Most empirical correlations like the one listed below apply only to critical flow conditions. Some notable researchers who have modified this correlation as shown in Table 1 include: Baxendell (Baxendell, 1957), Ros (Ros, 1960), Achong (Achong, 1961), Pilehavari (Pilehavari, 1981), Owolabi (Owolabi et al., 1991), Beiranvand (Beiranvand et al., 2012) and Okon (Okon et al., 2015).

Table 1: Empirical constants for different Gilbert-type choke correlations

Correlations	A	B	C
Gilbert	0.1	1.89	0.546
Ros	0.574	2	0.5
Baxendell	0.1046	1.93	0.546
Achlong	0.2618	1.88	0.65
Pilehavari	0.0214	2.11	0.313
Owolabi et al.	0.028	1.83	0.289
Beiranvand et al.	0.0328	2.275	0.589

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

Correlations	A	B	C
Okon et al.	0.1943	1.71	0.505

Other researchers have modified the Gilbert choke correlations, Table 2 by the addition of basic sediment and water, flowing well temperature, and temperature at standard conditions. The general form for this modified correlation is given as equation 2.

$$Q = A\frac{P_{wh}{}^{F} S^{B} (1-\frac{BS\&W}{100})^{D} (\frac{T}{T_{SC}})^{E}}{GLR^{C}} \quad (2)$$

where,
P_w = Flowing wellhead pressure in psig
GLR = gas-to-liquid ratio in SCF/STB
Q = Gross Liquid rate in STB/D
S = Choke size in (1/64 inch)
BS&W = Basic sediment and water (%)
T = Flowing well temperature (°F)
T_{sc} = Standard Temperature (60 °F)
A, B, C, D, E, and F = Empirical constants

Table 2: Empirical constants for the modified Gilbert-type choke correlation with BS&W and T inclusion

Correlations	A	B	C	D	E	F
Beiranvand et al. (2012)	1	1.5	0.1	1	-0.8	0.5
Okon et al. (2015)	0.0509	1.8134	0.6749	0.02235	0.000029	1.321
Ghorbani et al. (2018)	0.0575	2.3	0.709	0.00001	0	1

Multiphase flow is the most common fluid flow condition for most oil wells including those in Niger Delta region of Nigeria. However, only a few researchers have investigated multiphase flow behaviour in the Niger Delta region of Nigeria. Okon et al. (2015) studied the relationship between wellhead pressure and production rate using the Microsoft Excel Solver optimization method, General Reduced Gradient (GRG), for Niger Delta wells. Sanni et al., (2020) investigated the further expansion of the Choubineh et al., (2017) choke model also with the use of GRG to develop an oil flow rate model for the Nigeria oil fields. However, these researches mainly used optimization techniques and standard models, which may not provide the best, thorough representation, especially the nonlinear characteristics of multiphase flow. The linearization that is necessary when applying the GRG approach could be a source of errors because the nature of relationships between the variables might be more complicated than linear transformations. Moreover, it should be admitted that the GRG method, being sensitive to initial guesses and local optima, may pose some concerns as to the model parameters' accuracy. Other issues include implementation concerns like the inability of Excel Solver for large sets of data and multiple models. Furthermore, there is a notable gap in research employing advanced computational methods, such as machine learning, to model multiphase flow rates. Machine learning techniques have the potential to better capture the non-linear and complex relationships inherent in multiphase flow. Dabiri et al. (2024) used only Gilbert-type parameters in their study of nonlinear machine learning techniques for the estimation of the wellhead choke flow rate. Their solution was for non-Nigerian oil wells; hence, the applicability of that solution in the case of models for flowing wells in Nigeria is limited. For this study, several nonlinear machine learning models will be developed using the typical Gilbert choke input parameters with the addition of flowing well temperature and basic sediment and water content to better estimate the oil flow rate in Nigerian fields.

2.0 Materials and Methods

2.1 Data Collation and Preparation

This study analysed 1,431 production flow test records from 21 selected wells producing in the south-eastern region of Nigeria, one of Nigeria's most productive oil regions. These records consist of the following variables: choke size (x/64" in), flowing wellhead pressure (psi), flowline temperature (°F), gross liquid rate (BPD), gas-liquid ratio, and BS &W (%). The output data is the measured gross liquid rate.

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

Prior to model implementation, the dataset underwent preprocessing, including data cleaning, outlier detection and removal, and dataset partitioning. Data cleaning involved correcting numerical inconsistencies, and addressing missing or erroneous values across all features. The 'Choke X/64"' column was processed to extract and convert fractions into numerical values, while non-numeric representations, such as "Full Open," were assigned equivalent numerical value of "64 in". Erroneous symbols, multiple decimal points, and placeholders were removed to ensure data consistency. Outliers were identified using box plots and subsequently removed using the interquartile range (IQR) method. The lower and upper bounds were set using the 1.5*IQR rule, and extreme values in key variables, including flowing wellhead pressure, flowline temperature, oil production rate, and gas-oil ratio, excluded. Additionally, records with zero or negative oil production were removed to maintain data integrity. Following preprocessing, the dataset was randomly partitioned into training (75%) and testing (25%) subsets to facilitate model training and evaluation (Figure 2). Summary statistics of the cleaned dataset are presented in Tables 3 and 4.

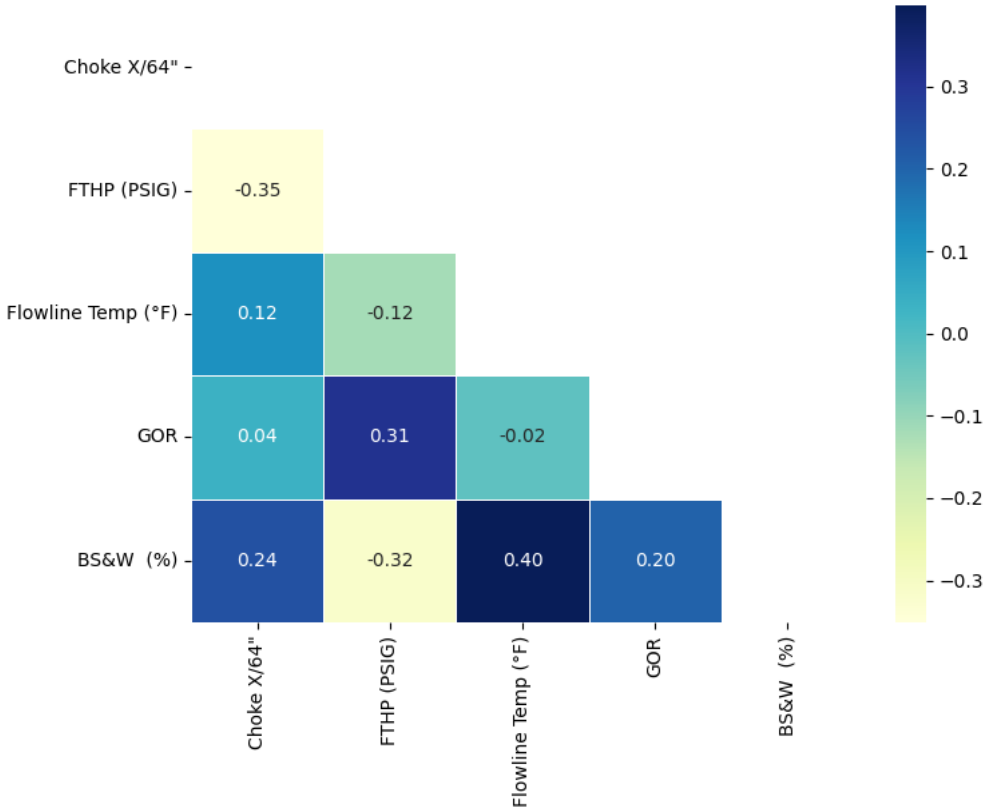

Figure 2: Data Training Analysis

Table 3: The train data summary statistics

Features	Minimum	Maximum	Mean	Median	Variance	Standard deviation
Choke (X/64")	1	192	78.63	70	1339.56	36.6
FTHP (PSIG)	105	438	254.7	260	4625.36	68.01
Flowline Temp (°F)	93	124	109.8	110	32.83	5.73
QL (BPD)	489	12271.7	4456.51	4084.1	4149531.96	2037.04
GLR (SCF/STB)	0	467.58	137.72	130.04	9623.61	98.1
BS&W (%)	0	95	54.1	62.5	819.10	28.62

Table 4: *The test data summary statistics*

Features	Minimum	Maximum	Mean	Median	Variance	Standard Deviation
Choke (X/64")	1	192	78.2	70	1553.15	39.41
FTHP (PSIG)	103	443.6	254.31	259.65	4752.72	68.94

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

Features	Minimum	Maximum	Mean	Median	Variance	Standard Deviation
Flowline Temp (°F)	93.4	122	109.35	110	28.62	5.35
QL (BPD)	238.3	10984	4299.46	3976.5	4144725.94	2035.86
GLR	0	464.57	143.77	130.97	10569.90	102.81
BS&W (%)	0	96.5	53.23	60.5	916.27	30.27

2.2 Machine Learning

To estimate the gross liquid rate, six non-linear machine learning algorithms were employed: random forest, categorical boosting (catboost), extreme gradient boosting (xgboost), attentive interpretable tabular learning (tabnet), support vector machine with a radial basis function (rbf) kernel, and gaussian process regression with an rbf kernel. these algorithms were selected for their ability to capture the complex, non-linear relationships in well test data.

Bayesian optimization was used to fine-tune the hyperparameters of each algorithm, ensuring optimal performance. to improve model generalization and prevent overfitting, k-fold cross-validation was applied. this technique divides the dataset into k subsets, with the model trained and validated k times across different partitions. in this study, 10-fold cross-validation was implemented, meaning the training data was randomly divided into 10 subsets, with each subset serving as a validation set once while the model was trained on the remaining data.

2.2.1 Attentive Interpretable Tabular Learning (tabnet)

This is a powerful technique for interpretable tabular data deep learning developed by arik and pfister (2019). here, the algorithm works by training a model via gradient-descent-based optimization; it takes raw tabular data and does not need any preprocessing of the input, hence readily integrable into end-to-end learning. in the present study, a tabnet algorithm is adopted for predicting the gross liquid rate, and the hyperparameter used for training is shown in Table 5.

Table 5: The hyperparameter used for training

Parameter	Description	Value
n_d	Dimensionality of the decision network (decision-making layer size).	9
n_a	Dimensionality of the attention mechanism.	49
n_steps	Number of rounds of learning in the model.	8
gamma	Regularization parameter controlling sparsity.	1.0697
lambda_sparse	Penalty to enforce sparsity in learned features.	2.2657e-06
optimizer_fn	Optimizer used for training.	AdamW
optimizer_lr	Learning rate for the optimizer.	0.01129
mask_type	Type of mask used for attention (Entmax encourages sparsity).	entmax
seed	Random seed for reproducibility of the results.	42

2.2.2 Gradient Boosting

Gradient boosting, first proposed by Friedman (2001), is an ensemble learning algorithm where each successive model corrects the errors of the previous one, thereby minimizing the total error. Several extensions of gradient boosting exist, including Categorical Boosting (CatBoost) and Extreme Gradient Boosting (XGBoost), both of which were used in this study. CatBoost by Prokhorenkova et al., (2018) used a version of "ordered boosting" for avoiding overfitting problems and improving generalization. Here, permutations are done instead of using the actual data so that target leakage cannot happen. XGBoost uses added parameters to gradient boosting introduced by (Chen and Guestrin, 2016) to avoid overfitting by penalizing the model for its complexity. This will involve tree pruning and shrinking in order to optimize the performance of the model. The hyperparameters used in training the catboost and xgboost model are shown in Table 6.

Table 6: The hyperparameters used in training the catboost and xgboost model

	iterations	learning_rate	max_depth	random_state
XGBoost	1079	0.2363	5	42
CatBoost	1300	0.1104	10	42

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

2.2.3 Random Forest

This is an ensemble method based on decision trees, introduced by (Breiman, 2001), which works by training 'n' decision trees and then takes the mean or mode of the output from these 'n' trees. The hyperparameters used in training the random forest model are shown in Table 7.

Table 7: The hyperparameters used in training the random forest model

	iterations	max_depth	random_state
Random Forest	**400**	**12**	**42**

Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel: Probably one of the most powerful machine learning algorithms in the domain of regression analysis, Support Vector Machine (SVM) was developed by Vladimir Vapnik along with his colleagues working at AT&T Bell Laboratories. The concept of SVMs involves finding a function that would represent the target variable within some level of acceptable error (Cortes & Vapnik, 1995). The SVM allows the use of the radial basis function for its kernel, which allows it to map inputs to higher dimensions and hence model nonlinear relationships in data points. In this work, the gross liquid rate has been predicted using an SVM algorithm with an RBF kernel. The hyperparameters used in training the random forest model is shown in Table 8.

Table 8: The hyperparameters used in training the random forest model

	C (Regularization Parameter)	Gamma (Kernel Coefficient)	Kernel
SVM	31.8281	0.000441	"rbf"

2.2.4 Gaussian Process Regression (GPR) with a Radial Basis Function (RBF) kernel:

GPR-RBF is a non-parametric-based probabilistic model applied in regression analysis. It is not a mere forecast, but an estimate of the measure of confidence in each forecast produced along with it (Kim & Lee, 2021). It further assumes the normality over functions and uses the assumed normality to make a forecast of the target variable. In this research, the GPR-RBF algorithm is adopted for gross liquid rate prediction. To optimize the performance of the model, specific hyperparameters were selected. The RBF kernel was scaled by a factor of 2, with a length scale of 400.37, which controls the smoothness of the function approximation. A larger length scale ensures smoother predictions, assuming long-range dependencies in the data. Additionally, the noise level parameter, alpha = 0.0251, was used to account for observational uncertainty, ensuring robust predictions. To maintain reproducibility, the random state was set to 42. The parameters used for the model are shown in Table 9.

Table 9: Parameters used in the Model

Hyperparameter	Description	Value
Kernel	Defines the covariance function of the model	2 * RBF(length_scale)
Length Scale	Controls the smoothness of the function	400.3666
Alpha	Added noise variance to improve stability	0.025108
Random State	Ensures reproducibility	42

2.3 Model Evaluation

The predictions of the developed gross liquid rate machine learning models will be compared with the actual gross liquid rate values using the metrics below. This comparison will assess the performance of the developed models. The choice of metrics was carefully selected to evaluate how well the individual models understand the input features and the extent of the deviation between the model predictions and the actual gross liquid rate.

1. Coefficient of Determination (R-squared)

$$R^2 = 1 - \frac{RSS}{TSS} \quad (3)$$

RSS stands for "residual sum of squares"
TSS stands for "total sum of squares"

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (4)$$

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

$$TSS = \sum_{i=1}^{n}(y_i - \underline{y})^2 \quad (5)$$

Where,

y_i = true value
$\hat{y}$ = predicted value
n = number of data points
$\underline{y}$ = mean value of a sample

2. Root Mean Squared Deviation (RMSD)

$$RMSD = \sqrt{\frac{\Sigma^n{}_{n=1}(y_i - \hat{y})^2}{n}} \quad (6)$$

3. Mean Squared Deviation (MSD)

$$MSD = \frac{\Sigma^n{}_{n=1}(y_i - \hat{y})^2}{n} \quad (7)$$

4. Mean Absolute Deviation (MAD)

$$MAD = \frac{\Sigma^n{}_{n=1}|y_i - \hat{y}|}{n} \quad (8)$$

5. Mean Absolute Percentage Deviation (MAPD)

$$MAPD = |\frac{y_i - \hat{y}}{y_i}| \, X \, 100 \quad (9)$$

3.0 Results and Discussion

The CatBoost, TabNet, Random Forest, XGBoost, Random Support Vector Machine, and Gaussian Process models used for estimating the gross liquid rate were assessed using the R^2, root-mean-square deviation (RMSD), mean absolute deviation (MAD), and mean absolute percentage deviation (MAPD) metrics. The models were applied to both the training and test datasets. The reason for evaluating both sets is to ensure the models are not overfitting on either set. From Table 5 and Table 6, the individual model results are not far apart for both train and test data sets. The CatBoost model had the best results across all metrics on both the training and test sets, with an RMSD of 54.3 STBD and R^2 of 1.00 for the training set, and 353.8 STBD and R^2 of 0.97 for the test set. The Random Forest model also proved to be a good one coming in second to the CatBoost model overall. The Random Forest model was able to provide an RMSD of 206.3 STBD and R^2 of 0. 99 for the training set, and 369. 6 STBD and R^2 of 0.97 for the test set. The rest of the models including XGBoost, TabNet, Gaussian Process-RBF, and SVM-RBF also had good performances with R^2 values ranging from (0.98 - 0.99) and RMSD between (232 - 281 STBD) for the training sets, and R^2 between (0.94 - 0.96) and RMSD ranging from (381 - 513 STBD) for the test sets.

3.1 Lower GLR Section: GLR < 300 SCF/STB

Shown in Tables 10, 11, and 12 are the statistical measures of the R^2, RMSE, MAE, and MAPE values for the eight Gilbert-type choke models used, three modified Gilbert-type models employed, and six predictive models built. According to the findings shown above, it can be stated that the six predictive models proposed perform better than the literature models in all the metrics used. Particularly, CatBoost and Random Forest outperform other models, yielding a high R^2 score of 97% with an RMSE of 360 STBD, and an R^2 score of 97% with an RMSE of 377 STBD, respectively.

Table 10: Statistical performance of the 6 predictive models for GLR < 300 SCF/STB

Model	R-Squared	RMSE	MAE	MAPE
TabNet	0.96	416.39	244.95	9.45
RandomForest	0.97	377.62	241.2	8.63
CatBoost	0.97	360.8	217.54	7.8
XGBoost	0.96	392.29	243.7	8.93
GaussianProcess	0.96	391.13	241.1	7.7
SVM	0.94	516.14	291.94	8.78

Table 11: Statistical performance of the 8 Gilbert-type choke models for GLR < 300 SCF/STB

Model	R-Squared	RMSE	MAE	MAPE
Gilbert	-150.27	25273.64	7735.29	159.04
Ros	-19535.63	287216.4	126606.27	2691.65
Baxendell	-258.63	33110.23	10667.75	219.13
Achlong	-950.53	63386.37	16632.97	338.98
Pilehavari	-125.11	23075.66	12901.89	266.84
Owolabi	-7.3	5919.1	3122.1	67.36
Beiranvand	-950.16	63374.17	16894.99	317.05
Okon	-88	19385.14	7182.26	157.24

Table 12: Statistical performance of the 3 modified Gilbert-type models for GLR < 300 SCF/STB

Model	R-Squared	RMSE	MAE	MAPE
Modified Beiranvand	-1.76	3411.6	2685.78	61.38
Ghorbani	-3798.26	126658.3	25543.16	457.41
Modified Okon	-156.76	25810.01	6891.24	152.84

3.2 Upper GLR Section: (≥300 SCF/STB)

Tables 13, 14, and 15 contain the statistical measures of the R^2, RMSE, MAE, and MAPE values for the eight Gilbert-type choke models used, the three modified Gilbert-type models employed, and six predictive models built. Based on the results presented, the six machine learning models outperform the literature models in all metrics. Specifically, the XGBoost, CatBoost, and Random Forest models demonstrate superior performance, achieving remarkable R^2 scores of 96%, 95%, and 95%, respectively, with corresponding RMSE values of 231 STBD, 265 STBD, and 269 STBD.

The scatter plot of the predicted gross liquid rate against the actual gross liquid rate for the 6 predictive models applied to the test data set is shown in Figure 3.

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

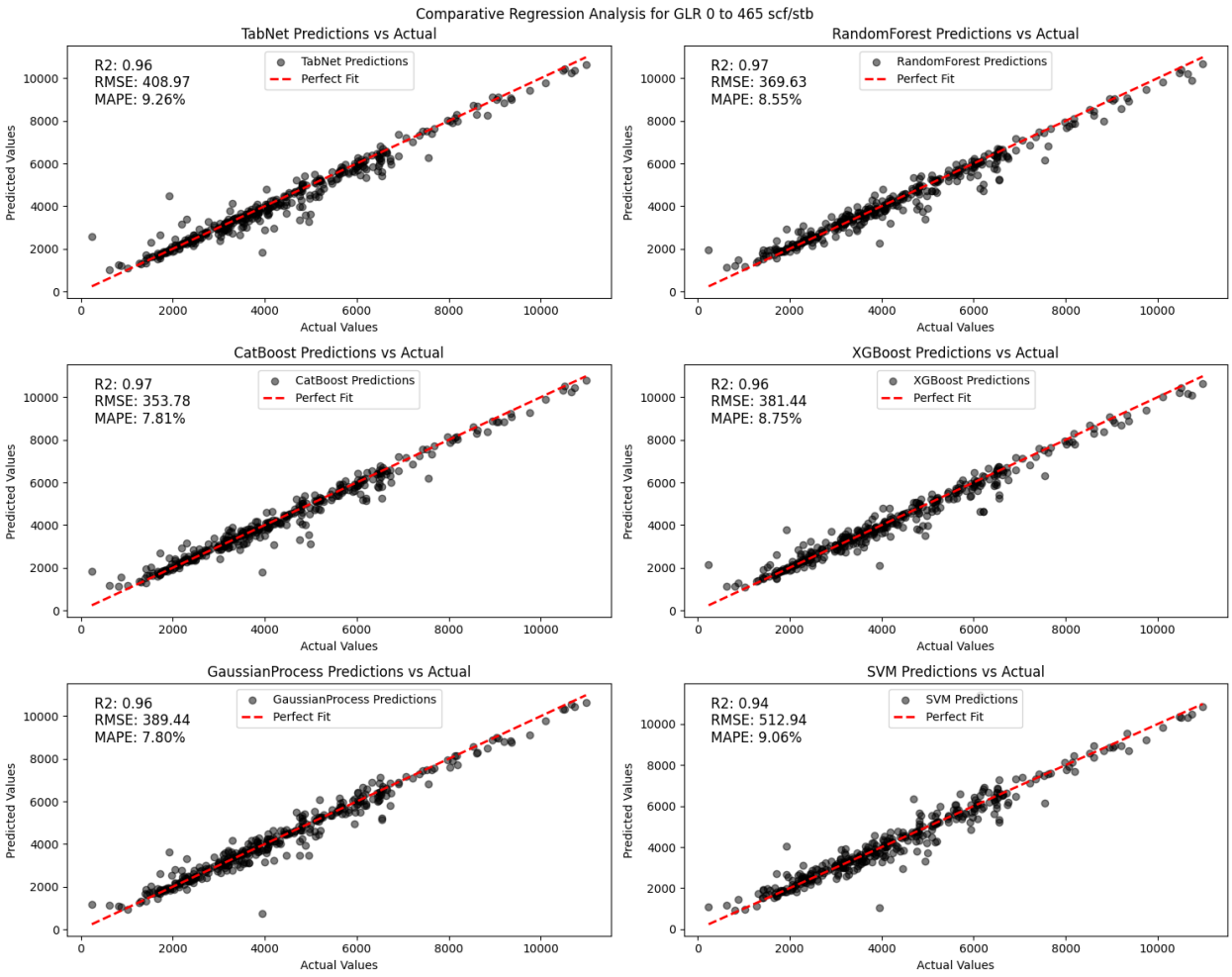

Figure 3: A scatter plot of the predicted gross liquid rate against the actual gross liquid rate for the 6 predictive models applied to the test data set

The statistical performances of the developed models on the train data are shown in Table 13 and 14.

Table 13: Statistical performance of the developed models on the train data

Model	R-Squared	RMSD	MAD	MAPD
TabNet	0.98	280.82	183.04	4.60
RandomForest	0.99	206.30	132.18	3.58
CatBoost	1.00	54.30	37.43	1.10
XGBoost	0.98	260.36	163.28	4.18
GaussianProcess-RBF	0.98	289.18	189.85	5.06
SVM-RBF	0.99	231.55	167.77	4.48

Table 14: Statistical performance of the developed models on the test data

Model	R-Squared	RMSD	MAD	MAPD (%)
TabNet	0.96	408.97	242.70	9.26
RandomForest	0.97	369.63	236.96	8.55
CatBoost	0.97	353.78	215.49	7.81
XGBoost	0.96	381.44	238.58	8.75
GaussianProcess-RBF	0.96	389.44	243.97	7.80
SVM-RBF	0.94	512.94	296.34	9.06

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

3.2.1 Comparative Analysis Between the Developed Machine Learning Models and Empirical Choke Correlations

A comparative analysis was conducted using popular and commonly used Gilbert-type choke models (Gilbert, Ros, Achlong, Baxendell, Pilehavari, Owolabi et al., Beiranvand et al., Okon et al.), modified Gilbert-type models with new parameter inclusions (Beiranvand et al., Okon et al., Ghorbani et al.), and newly developed predictive models (CatBoost, TabNet, Random Forest, XGBoost, SVM, and Gaussian Process models). To conduct this comparative study, the data was divided into two sections: the lower GLR (0 - 199 SCF/STB) section and the upper GLR (300 - 465 SCF/STB) section. The literature models exhibited very poor performance in both sections, as anticipated, due to their development using reservoir and flow conditions that are atypical of Niger Delta fields. For lower GLR section: GLR< 300 SCF/STB case, Tables 15, 16 and 17 showed the statistical measures of the R^2, RMSD, MAD, and MAPD values for the eight Gilbert-type choke models used, three modified Gilbert-type models employed, and six predictive models built. According to the findings shown below, it can be stated that the six predictive models proposed perform better than the literature models in all the metrics used. Particularly, CatBoost and Random Forest outperform other models, yielding a high R^2 score of 97% with an RMSD of 360 STBD, and an R^2 score of 97% with an RMSD of 377 STBD, respectively.

Table 15: Statistical performance of the 6 predictive models for GLR < 300 SCF/STB

Model	R-Squared	RMSD	MAD	MAPD
TabNet	0.96	416.39	244.95	9.45
RandomForest	0.97	377.62	241.2	8.63
CatBoost	0.97	360.8	217.54	7.8
XGBoost	0.96	392.29	243.7	8.93
GaussianProcess	0.96	391.13	241.1	7.7
SVM	0.94	516.14	291.94	8.78

Table 16: Statistical performance of the 8 Gilbert-type choke models for GLR < 300 SCF/STB

Model	R-Squared	RMSD	MAD	MAPD
Gilbert	-150.27	25273.64	7735.29	159.04
Ros	-19535.63	287216.4	126606.27	2691.65
Baxendell	-258.63	33110.23	10667.75	219.13
Achlong	-950.53	63386.37	16632.97	338.98
Pilehavari	-125.11	23075.66	12901.89	266.84
Owolabi	-7.3	5919.1	3122.1	67.36
Beiranvand	-950.16	63374.17	16894.99	317.05
Okon	-88	19385.14	7182.26	157.24

Table 17: Statistical performance of the 3 modified Gilbert-type models for GLR < 300 SCF/STB

Model	R-Squared	RMSD	MAD	MAPD
Modified Beiranvand	-1.76	3411.6	2685.78	61.38
Ghorbani	-3798.26	126658.3	25543.16	457.41
Modified Okon	-156.76	25810.01	6891.24	152.84

For upper GLR section: GLR ≥300 SCF/STB case, Tables 18, 19, and 20 showed the statistical measures of the R^2, RMSD, MAD, and MAPD values for the eight Gilbert-type choke models used, the three modified Gilbert-type models employed, and six predictive models built. Based on the results presented, the six machine learning models outperform the literature models in all metrics. Specifically, the XGBoost, CatBoost, and Random Forest models demonstrate superior performance, achieving remarkable R^2 scores of 96%, 95%, and 95%, respectively, with corresponding RMSD values of 231 STBD, 265 STBD, and 269 STBD.

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

Table 18: Statistical performance of the 6 predictive models for GLR ≥ 300 SCF/STB

Model	R-Squared	RMSD	MAD	MAPD
TabNet	0.93	314.09	215.78	7.2
RandomForest	0.95	268.56	193.32	7.87
CatBoost	0.95	264.91	193.61	7.99
XGBoost	0.96	231.16	179.07	6.64
GaussianProcess	0.9	378.19	283.97	9.45
SVM	0.83	480.6	351.18	12.63

Table 19: Statistical performance of the 8 Gilbert-type choke models for GLR ≥ 300 SCF/STB

Model	R-Squared	RMSD	MAD	MAPD
Gilbert	-0.95	1635.04	963.37	30.52
Ros	-1353.19	43044.48	33386.52	1081.38
Baxendell	-2.79	2275.93	1075.73	33.54
Achlong	-3.85	2576.9	1276.56	40.75
Pilehavari	-23.22	5756.33	3430.41	108.98
Owolabi	-0.62	1486.71	921.01	29.62
Beiranvand	-6.6	3223.96	1369.17	39.68
Okon	-1.12	1704.35	872.3	28.73

Table 20: Statistical performance of the 3 modified Gilbert-type models for GLR ≥ 300 SCF/STB

Model	R-Squared	RMSD	MAD	MAPD
Modified Beiranvand	-2.25	2110.82	1848.3	61.39
Ghorbani	-6.52	3212.25	1363.27	39.01
Modified Okon	-0.07	1212.81	978.24	33.8

3.2.2 Sensitivity Analysis

A feature importance analysis was conducted on the six different machine learning models. From Figure 4, it is evident that basic sediment and water content were the most crucial features for modeling the gross liquid rate, while the gas-liquid ratio and flowing temperature were identified as the least important features.

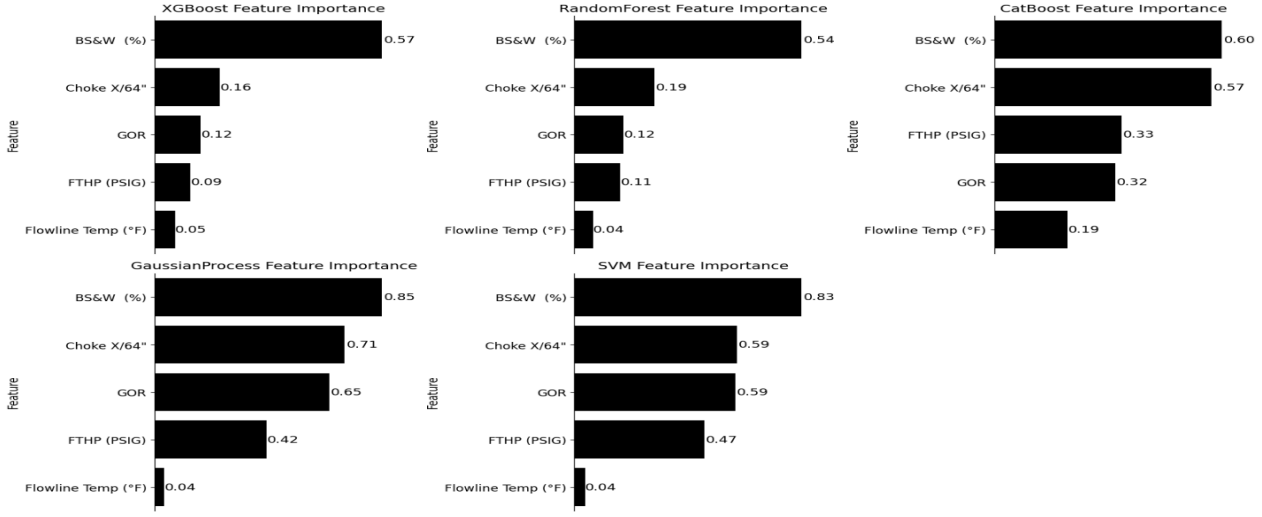

Figure 4: A feature importance plot for the 6 predictive models applied to the test data set

A sensitivity analysis was performed on the best-performing model (CatBoost) to assess the magnitude and direction of each feature. According to Figure 5, the following trends were observed:

- There is an increasing trend in choke size up to a point (150/64“), beyond which further increases have negligible effects.

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

- For flowing tubing head pressure (FTHP), a slight decrease is observed at lower pressures (0-200 psi), followed by an increase.
- There is a consistent decrease in basic sediment and water content, indicating a strong correlation with the target.
- A slight increase in the gas-liquid ratio is observed up to 100 SCF/STB, after which it consistently decreases.
- Flowing well temperature shows a complex relationship, indicating a weak correlation with the target.

The practical significance of these findings is that the initial decline followed by a sharp rise in the gross liquid rate sensitivity factor suggests an optimal choke size that maximizes production without excessive pressure drop or formation damage (Almohammad et al., 2019; Gidado et al., 2023). However, when this choke size exceeds a certain limit, it does not impact the flow rate significantly. This is also echoed by the optimization studies in shale gas wells by (Wu et al., 2022). The degree of gross liquid rate sensitivity rises with FTHP, which reveals at higher pressures, there is an increased oil throughput. The initial increase in the gross liquid rate sensitivity factor with GLR may suggest the use of the gas lift method to improve the gross liquid rate. However, the use of gas lifts must be done carefully as high gas concentrations have detrimental impacts on production efficiency because of operation difficulties and back pressure problems (Dwivedi et al., 2022). The result also revealed a negative relationship between BS&W and gross liquid rate sensitivity factor, which shows the principal need to bring down the levels of BS&W for optimal oil production.

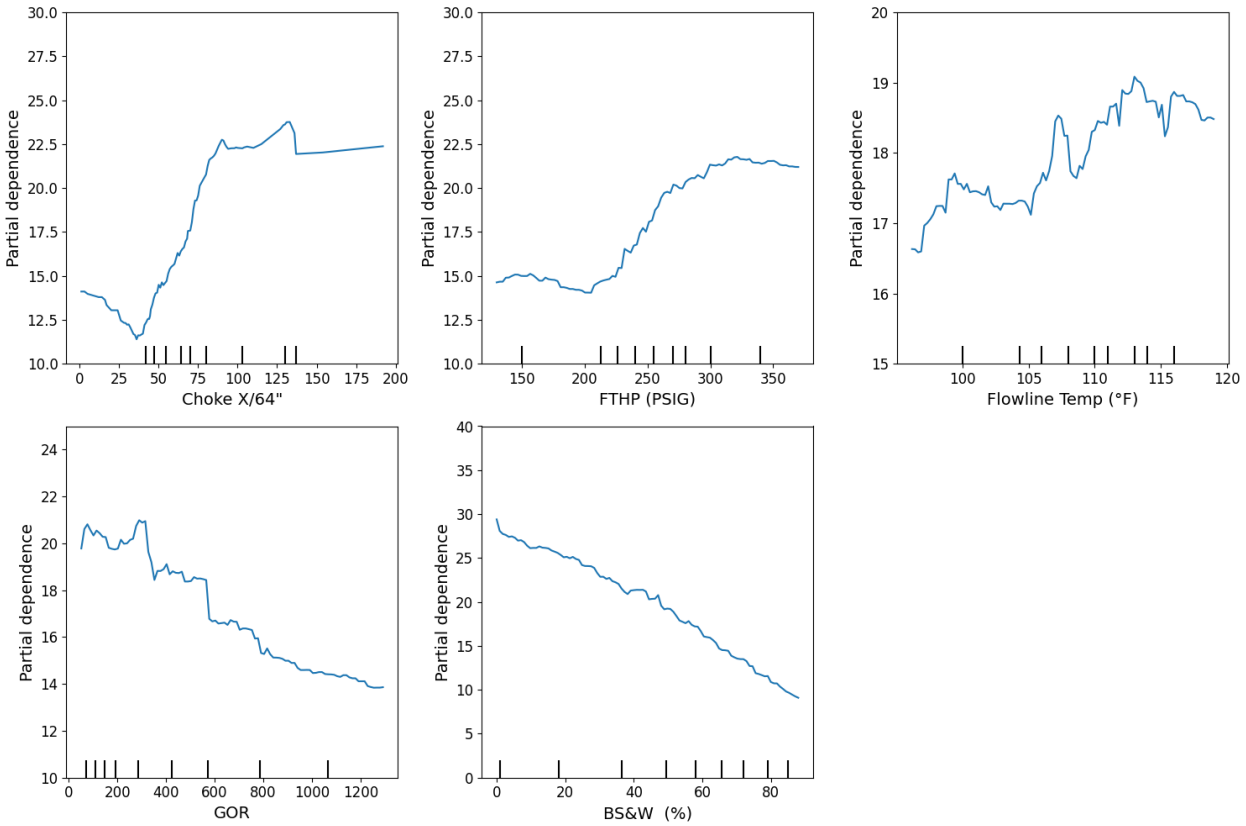

Figure 5: A sensitivity analysis plot for the best-performing model (Catboost) on the train and test data set

4.0 Conclusion

An analytical method was employed using non-linear machine learning to create a robust model useful for estimating the gross oil rate in Nigerian oil-producing wells. The CatBoost, TabNet, Random Forest, XGBoost, Support Vector Machine, and Gaussian Process models were employed for this modelling using 1,431 production flow test records from 21 wells in the Niger Delta region of Nigeria. A comparison was made between the developed models and existing literature models, and the former exhibited significantly superior performance. The following conclusions can be drawn from this research.

1. Machine learning techniques are powerful tools for modeling and predicting the gross oil rate at surface production facilities.
2. All six models investigated in this study outperformed the existing choke correlations using the coefficient of determination (R^2), root mean square deviation (RMSD), and mean absolute percentage deviation (MAPD) evaluation metrics, with CatBoost and Random Forest performing best. The CatBoost model

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026

achieved an R^2 score of 97%, an RMSD of 353 STBD, while the Random Forest model achieved an R^2 score of 97%, and an RMSD of 369 STBD.

3. The predictive models also demonstrated superior performance compared to existing correlations in both regions with low GLR < 300 SCF/STB and regions with high GLR ≥300 SCF/STB. CatBoost and Random Forest performed best amongst all the models, yielding a high R^2 score of 97% with an RMSD of 360 STBD, and an R^2 score of 97% with an RMSD of 377 STBD, respectively in low GLR regions. In high GLR regions, the XGBoost, CatBoost, and Random Forest models achieved remarkable R^2 scores of 96%, 95%, and 95%, respectively, with corresponding RMSD values of 231 STBD, 265 STBD, and 269 STBD.
4. These models can be used to estimate the oil rate downstream of the choke in most Niger Delta fields. They are useful tools for reliable and quick estimates of oil production rates, hence facilitating an informed decision-making process.

Additional well data will increase the strength of this machine learning techniques in estimating the gross liquid rate in Nigerian oil wells. Future research may be conducted using these data-driven techniques as additional well data becomes available leading to the development of a more robust models for oil flow rates estimation.

Nomenclature

A = Proportionality constant
B = Gas-to-liquid ratio constant
C = Choke size constant
D = BS&W constant
E = Flowing temperature term constant
F = Flowing tubing head or wellhead pressure constant
FTHP = Flowing tubing head or wellhead pressure (Psi)
GOR = Gas-to-oil ratio (SCF/STB)
GLR = Gas-to-liquid ratio (SCF/STB)
MSD = Mean square deviation
MAD = Mean absolute deviation
MAPD = Mean absolute percentage deviation
MAPE = Mean absolute percentage error
Q = Gross liquid flowrate (STB/D)
QL = Gross liquid flowrate (STB/D)
P_w = Flowing tubing head or wellhead pressure (psi)
P_{wh} = Flowing tubing head or wellhead pressure (psi)
R^2 = Coefficient of determination or performance
RMSD= Root mean squared deviation
RMSE= Root mean squared error
S = Choke size (*/64th inches)
SCF = Standard cubic feet
STB = Stock tank barrel
STBD = Stock tank barrel per day
T = Flowing temperature (°F)
T_{sc} = Standard condition temperature (°F)

References

Achong, I. B. (1961). Revised bean and performance formula for Lake Maracaibo wells" published by the University of Zulia. *Maracaibo, Venezuela*.

Almohammad, H. (2019). Effective Production Network Debottlenecking Analysis by Finding Optimum Choke Size. In *Abu Dhabi International Petroleum Exhibition and Conference* (p. D012S142R001). SPE.

Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 8, pp. 6679-6687).

Baxendell, P. B. (1957). Bean performance-lake wells. *Shell Internal Rep.*

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chang, W., Wang, X., Yang, J., & Qin, T. (2023). An improved CatBoost-based classification model for ecological suitability of blueberries. *Sensors*, *23*(4), 1811.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Choubineh, A., Ghorbani, H., Wood, D. A., Moosavi, S. R., Khalafi, E., & Sadatshojaei, E. (2017). Improved predictions of wellhead choke liquid critical-flow rates: modelling based on hybrid neural network training learning based optimization. *Fuel*, *207*, 547-560.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Dabiri, M. S., Hadavimoghaddam, F., Ashoorian, S., Schaffie, M., & Hemmati-Sarapardeh, A. (2024). Modeling liquid rate through wellhead chokes using machine learning techniques. *Scientific Reports*, *14*(1), 6945.

Dwivedi, V., Al Zaabi, F., & Jaffres, B. (2022). The Benefits of an Integrated Approach (Development, Completion, Production Operation, and Integrity) for Gas-Lift Design on an UAE Offshore Field. In *Abu Dhabi International Petroleum Exhibition and Conference* (p. D022S169R004). SPE.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Ghorbani, H., Wood, D. A., Moghadasi, J., Choubineh, A., Abdizadeh, P., & Mohamadian, N. (2019). Predicting liquid flow-rate performance through wellhead chokes with genetic and solver optimizers: an oil field case study. *Journal of Petroleum Exploration and Production Technology*, *9*(2), 1355-1373.

Gidado, A. O., Adeniyi, A. T., Olusola, B., & Giwa, A. (2023). Sensitivity Analysis of Choke Size Selections on Reservoir Pressure Drawdown Using Prosper Modelling for Reservoir Management. In *SPE Nigeria Annual International Conference and Exhibition* (p. D031S013R005). SPE.

Gilbert, W. E. (1954). Flowing and gas-lift well performance. In *Drilling and production practice*. OnePetro.

Gou, B., Lyons, W. C., & Ghalambor, A. (2007). Petroleum production engineering.

Khan, K., Ahmad, W., Amin, M. N., Ahmad, A., Nazar, S., & Alabdullah, A. A. (2022). Compressive strength estimation of steel-fiber-reinforced concrete and raw material interactions using advanced algorithms. *Polymers*, *14*(15), 3065.

Kim, J., & Lee, J. (2021). Identifiability of covariance kernels in the Gaussian process regression model. *arXiv preprint arXiv:2108.04715*.

Okon, A. N., Udoh, F. D., & Appah, D. (2015). Empirical wellhead pressure-production rate correlations for Niger Delta oil wells. In *SPE Nigeria Annual International Conference and Exhibition. OnePetro*.

Owolabi, O. O., Dune, K. K., & Ajienka, J. A. (1991, August). Producing the multiphase flow performance through wellhead chokes for the Niger Delta oil wells. In *International Conference of the SPE Nigeria Section Annual Proceedings, August* (pp. 28-30).

Pilehvari, A. A. (1981). *Experimental study of critical two-phase flow through wellhead chokes*. University of Tulsa.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Ros, N. C. J. (1960). An analysis of critical simultaneous gas/liquid flow through a restriction and its application to flowmetering. *Applied Scientific Research*, *9*(1), 374-388.

Safar Beiranvand, M., & Babaei Khorzoughi, M. (2012). Introducing a new correlation for multiphase flow through surface chokes with newly incorporated parameters. *SPE Production & Operations*, *27*(04), 422-428.

Safar Beiranvand, M., Mohammadmoradi, P., Aminshahidy, B., Fazelabdolabadi, B., & Aghahoseini, S. (2012). New multiphase choke correlations for a high flow rate Iranian oil field. *Mechanical Sciences, 3(1)*, 43-47.

Sanni, K., Longe, P., & Okotie, S. (2020). New Production Rate Model of Wellhead Choke for Niger Delta Oil Wells. *Journal of Petroleum Science and Technology*, *10*(4), 41-49.

Singh, H. (2024). Machine Learning Application of Generalized Gaussian Radial Basis Function and Its Reproducing Kernel Theory. *Mathematics*, *12*(6), 829.

Wu, J., Yang, X., Di, Y., Li, P., Zhang, J., & Zhang, D. (2022). Numerical simulation of choke size optimization in a shale gas well. *Geofluids*, *2022*(1), 2197001.

Received: 09/03/2026; Revised: 24/03/2026; Accepted: 26/03/2026; Published: 15/04/2026