An Investigation into the Optimization of Electrically Submersible Pumps through the Application of Machine Learning

Anthony Kerunwa, Chimuanya Winifred Ibewuike, Lilian Enyang Ndoma-Egba, Ndubuisi Okereke, Chukwuebuka Francis Dike*, Federal University of Technology Owerri, Owerri, Nigeria

Abstract

An Electrical Submersible Pump, commonly abbreviated as ESP, is a type of dynamic displacement pump that is specifically designed and employed to extract crude oil from the wellbore when the natural pressure differential is insufficient to facilitate this process effectively. The ESP operates on the principle of generating momentum to lift fluids, transferring them from the inlet to the outlet of the system. In its pursuit to enhance oil production, the ESP system encounters a variety of constraints that can be broadly categorized into pressure-related, pump-specific, fluid-related, and motor-related issues. To ensure the system's optimal performance and design, it is crucial to conduct a comprehensive analysis of these constraints.

In this study, the Random Forest (RF) algorithm was employed as a powerful analytical tool to investigate the aforementioned constraints and to assess their influence on the rate of oil production. To develop the model, a dataset comprising 36 instances from an oilfield located in the Niger-Delta region was utilized. Following the development of the model, a thorough statistical evaluation and validation process was conducted to ensure the reliability and accuracy of the findings.

The results of the study were quite promising, with the Random Forest algorithm demonstrating a regression score of 99.66%, indicating an exceptionally high level of accuracy in predicting the oil production rate. Additionally, the Mean Absolute Error (MAE) was recorded at 0.00866, the Mean Squared Error (MSE) at 0.0001019, and the Root Mean Squared Error (RMSE) at 0.0101. These statistical metrics collectively suggest that the Random Forest algorithm provided predictions that were very close to the actual values of the oil production rate.

Upon validation, the Random Forest algorithm was found to yield values that were in close proximity to the real oil production rates, further underscoring its effectiveness as a predictive tool in the context of oilfield operations. This study not only highlights the potential of machine learning algorithms in addressing complex engineering challenges but also provides valuable insights that could inform the design and operation of ESP systems to maximize their efficiency and productivity.

Introduction

Fossil fuels contribute about 85% of the energy demand globally with 87MMstb/day and 37MMMstb/year (Sheng 2011; Kerunwa et al. 2024a), and this has created the need for continuous production despite several associated hurdles. Conventional oil well undergoes natural supplemental, secondary and tertiary recovery phases in their production life (Avwioroko et al. 2014). At the natural supplemental phase, the reservoir relies on her natural energy for crude oil recovery. (Akpoturi and Ofesi 2017), but when these energy drops, secondary recovery approach consisting of water or natural gas flooding is injected to maintain pressure and recover oil (Austad et al.

Improved Oil and Gas Recovery

DOI: 10.14800/IOGR.1353

Received January 05, 2025; revised March 18, 2025; accepted May 27, 2025.

*Corresponding author: dike.chukwuebuka@futo.edu.ng

2010). When the secondary recovery approach becomes ineffective due to capillary and viscous forces, enhanced oil recovery (EOR) techniques comprising of injection of chemicals and methods excluding natural gas and oil are deployed (Kerunwa et al. 2024b; Dike et al. 2024). The reservoir presents various challenges that necessitate the application of multiple engineering techniques to facilitate the flow of oil from the wellbore. However, the fluids encounter an additional obstacle at the wellbore. In some instances, crude oil ascends to the surface with ease, a phenomenon that is contingent upon the intersection of the inflow performance relationship (IPR) and the vertical lift performance (VLP) (Okologume and Ofesi 2016). Conversely, crude oil that has entered the wellbore may struggle to reach the surface due to constraints such as the presence of water, a limited amount of entrapped natural gas, and a minimal differential pressure, which can be attributed to increased bottomhole pressure and decreased reservoir pressure (Elshan 2013). Consequently, to enable the extraction of crude oil trapped in the wellbore to the surface, a technique known as Artificial Lift is required (Bellarby 2009). Close to 50% of oilfield globally utilizes artificial lift to enhance crude production (Guo et al. 2007). The artificial lift system can be categorized into pump and gas lift system. The pump system makes use of positive or dynamic displacement mechanism in creating pressure differential needed to lift the crude from the wellbore to the surface (Bellarby 2009).

The progressive cavity pump (PCP) and sucker rod pump (SRP) fall under positive displacement pumps while electrical submersible pump (ESP) and hydraulic pump (HP) are described as dynamic pumps. The gas-lift system utilizes the differential head by reducing the density of crude oil, which in turn reduces drawdown and well's inflow rate. Among the pumps, ESP have recorded widespread acceptance due to its performance and durability. ESP functions by lifting fluid from inlet to outlet under a momentum it creates. Its mode of operation can be defined as dynamic and centrifugal displacement (Ikekeazu and Anerobic 2020). The pump generates hydraulic power as result of the action of electric motors and injects energy needed to produce fluid for oil recovery. The fluid flows into the system through the impeller to the diffuser. The impeller provides the wellbore heads pressure through high-speed rotation, while the exerted kinetic energy by the impeller in turn is transformed to kinetic energy by the diffuser (Zhu and Zhang 2018). In achieving its goal of improved oil production, ESP systems face several constraints which could be pressure-based, pump-based, fluid-based and motor-based. This is needed for this constraint to be effectively managed and/or overcome for optimal production.

Studies on oil production by artificial lift using ESP design have been carried out by several authors. Gomaa et al. (2020) carried out electrical submersible pump (ESP) design on a vertical well. From the result of their experimental design, a pump with 106 stages, with required horsepower of 571hp and temperature of 255 ºF recorded 64% efficiency and recorded 13479.2rb/day. Kerunwa et al. (2022) carried out a study of ESP on field production network optimization in an oilfield in the Niger-Delta. From the result of their study, a 1.16% and 2.66% oil rate increase was achieved without optimization, while with optimization, 2.66% oil rate was achieved. Several studies have been carried out on oil production by ESP design, but these studies have not considered the impact of these ESP design parameters in oil rate production. In this study Random Forest Machine Learning (RF-ML) was utilized to optimize the parameters that the highest design parameters required to attain the best oil rate.

Methodology

The materials employed in this study encompass the Scikit-Learn package, a Python program, and a well dataset. The Scikit-Learn package is a globally recognized machine-learning program that possesses predictive and statistical functionalities. The dataset utilized in this study comprises 36 entries, featuring independent variables such as operating frequency, water-cut, gas separation, intake pressure, discharge pressure, generated pump head, pump power, and pump efficiency, with motor efficiency and motor speed serving as dependent variables (**Table 1**).

Table 1—Data utilized for the study.

Oil Flow Rate	Variable Data			Pressure Data		Pump Data				
Oil Rate (STB/d)	Ope. Freq. (Hz)	Water Cut (%)	Gas Sep. Eff. (%)	Intake Pressure (psi)	Discharge Pressure (psi)	Pump Head Gen. (FT)	Pump Power (HP)	Pump Eff. (%)	Motor Eff. (%)	Motor Speed (rpm)
1533.3	50	45	50	2547.79	2721.92	499.378	16.74	54.3392	63.5558	2941.56
1509.8	50	45	55	2549.79	2726.15	505.44	16.6948	54.3108	63.4929	2941.6
1487.7	50	45	60	2551.69	2730.14	511.116	16.6535	54.2775	63.432	2941.64
1465.9	50	45	65	2553.55	2732.02	516.612	16.6117	54.238	63.3706	2941.68
0	50	50	50	0	0	0	0	0	0	0
0	50	50	55	0	0	0	0	0	0	0
0	50	50	60	0	0	0	0	0	0	0
0	50	50	65	0	0	0	0	0	0	0
0	50	55	50	0	0	0	0	0	0	0
0	50	55	55	0	0	0	0	0	0	0
0	50	55	60	0	0	0	0	0	0	0
0	50	55	65	0	0	0	0	0	0	0
2100	60	45	50	2499.21	2752.89	726.532	30.2447	60.4534	71.7619	3497.54
2086.8	60	45	55	2500.36	2757.98	737.212	30.6753	61.2217	72.2969	3499.55
2073.7	60	45	60	2501.5	2763.05	747.871	31.1049	61.988	72.8306	3501.55
2060.7	60	45	65	2502.65	2768.07	758.395	31.5287	62.7437	73.357	3503.51
1319.4	60	50	50	2553.37	2852.04	840.438	28.0311	54.0685	70.9984	3514.55
1290.4	60	50	55	2556.09	2856.24	844.204	27.8095	53.5598	70.8398	3515.08
1263.4	60	50	60	2558.62	2860.14	847.692	27.6022	53.0847	70.6916	3515.57
1236.7	60	50	65	2561.13	2864	851.112	27.3957	52.6128	70.5443	3516.06
0	60	55	50	0	0	0	0	0	0	0
0	60	55	55	0	0	0	0	0	0	0
0	60	55	60	0	0	0	0	0	0	0
0	60	55	65	0	0	0	0	0	0	0
2444.7	70	45	50	2469.03	2792.11	924.87	48.5702	56.2451	77.3177	4063.68
2429.5	70	45	55	2470.36	2797.55	935.764	48.6507	56.5599	77.333	4063.5
2414	70	45	60	2471.72	2802.97	946.555	48.7254	56.8629	77.3471	4063.33
2398.5	70	45	65	2473.08	2808.32	957.117	48.7947	57.1502	77.36	4063.18
1960.8	70	50	50	2493	2882.37	1094.44	49.6665	60.1615	77.5133	4061.23
1946.5	70	50	55	2494.37	2887.23	1103.44	49.7044	60.3437	77.5198	4061.15
1930.7	70	50	60	2495.9	2892.49	1113.2	49.7388	60.5361	77.5256	4061.07
1916.3	70	50	65	2497.27	2897.25	1121.93	49.7691	60.7009	77.5307	4061.01
1105.7	70	55	50	2560.46	2993.12	1195.07	41.6485	48.9004	75.1619	4079.5
1074.9	70	55	55	2563.67	2997.28	1197.27	41.1825	48.2276	75.0305	4080.56
1047.5	70	55	60	2566.52	3000.98	1199.21	40.7685	47.6298	74.9139	4081.51
1021.3	70	55	65	2569.24	3004.51	1201.06	40.3723	47.0583	74.8026	4082.5

Data Process. The dataset was inspected to confirm the unavailability of irregularities and errors such as missing values and duplicates. Using datafram.isna () key within the python program it was observed that the entire data were present in the entry. The entries check confirmed that there were no missing values or duplicate values in the data.

Data Visualization. Data visualization was carried out to see the relation between the various variables present within the dataset. In attaining this, the multivariate visualization technique using Pearson correlation heatmap was utilized. The choice of correlation heatmap is tied to its ability to provide comprehensive overview of relationship between the variables within the dataset.

Feature Engineering. Feature Engineering was also conducted to identify and select those variables from the dataset that significantly contributes to the target variable oil rate. Using standard scale options, data normalization was carried out before the logarithmic transformation of target variables was carried out. This transformation is critical for variance stabilization and distribution normalized aimed at achieving linear correlation and appropriate model. The transformed target variables a-times yields better model performance, based on regression metrics. This procedure is vital especially when original data exhibits extreme values, heteroscedasticity (non-constant variance) or skewness.

Model Selection. In this study, Random Forest (RF) algorithm from Scikit-Learning Package of Python Software was utilized. Scikit-Learning Package of Python Software comprises several algorithms such as Linear Regression (LR) algorithm, Neural Network and Support Vector (SV) Algorithm. The choice of RF machine learning (RF-ML) is tied to its bagging technique setup which aids it to utilize the ensemble learning method for regressor in machine learning, and its potential to capture non-linear correlation and interaction features in very complex manner. There is variation in the chosen requirement for grouping the number of branches and nodes. The number of trees present in RF is one of the most vital hyper-parameters of RF which is used to determine the performance of the model.

Data Splitting. In the data splitting stage, the dataset was split into the training set (80%) and the testing set (20%). This was carried out to ensure that ML model is sufficiently trained on a large data (training data) while retaining a separate portion for analysis to confirm how well the model will perform with the unseen data (testing data).

Model Training. The RF algorithm utilized a diverse combination of decision trees with branches and nodes for grouping and regression. The root node (at the treetop) was divided to yield two branches which are defined by satisfactory observations. For the regression, partitions are selected to lower the variations of samples labels. While the training was going on, the hyper properties of the model were tuned using grid search optimization approach to derive the optimal input properties of the models. The optimization methodology was utilized to update the input features of the model. The expected primary outcome to be derived from the learning model is to derive the optimal oil rate.

Model Validation. After the training using RF-model, the model was validated using indices to determine its performance in comparison with actual dataset. The metrics used for the regression task included mean square error (MSE), mean absolute error (MAE), regression (R^2) and root mean square error (RMSE)

Comparison with Existing Model. The model was compared with actual field data, and Kerunwa et al. (2022) approach in **Eq. (1)** through **(5)**.

$$h = \frac{p_{discharge} - p_{sunction}}{0.433} \quad \text{,......} (1)$$

$$p_{discharge} = p_{wf} - 0.433Y_LD \qquad (2)$$

where h represents the pumping head, ft; $p_{discharge}$ signifies the discharge pressure, psi; $p_{sunction}$ denotes the suction pressure, psi; Y_L and D correspond to the specific gravity of the production fluid and the depth of the production interval in feet, respectively.

$$D_{pump} = D - \frac{p_{wf} - p_{sunction}}{0.433Y_L} \qquad (3)$$

where D_{pump} is the minimum pump depth, ft.

The quantity of stages in ESP is specified as,

$$n_s = \frac{Z}{L_s} \qquad (4)$$

where n_s denotes the number of design stages; Z represents the total dynamic head, measured in feet; and L_s signifies the lift per stage.

The equation for ESP motor horsepower calculation is given as,

$$p_{hm} = p_{hs}n_s\rho_f \qquad (5)$$

where p_{hm} represents the motor horsepower; p_{hs} denotes the horsepower per stage; n_s signifies the number of stages; and ρ_f corresponds to the specific gravity of the fluid.

Results and Discussion

Multivariate Analysis. Figure 1 illustrates the Pearson correlation heatmap for all variables within the dataset. It illustrates a robust positive correlation between the oil production rate and several operational parameters, including operating frequency, pump power, pump efficiency, motor efficiency, and motor speed. This suggests that a reduction in these variables would have a substantial effect on the oil production rate from the well. Conversely, the oil production rate exhibited a moderate positive correlation with discharge pressure and pump head generation, indicating that decreases in these factors do not significantly influence the oil production rate. Furthermore, the oil production rate demonstrated a strong inverse correlation with water-cut and intake pressure, and a moderate inverse correlation with gas separator efficiency. The strong inverse correlation indicates that increases in water-cut and intake pressure result in a decrease in the oil production rate, whereas the moderate inverse correlation suggests that enhancements in gas separator efficiency led to a marginal reduction in the oil production rate.

Model Performance. **Table 2** presents the goodness-of-fit metrics for the model generated by the machine learning algorithm. It is evident that the logistic regression model achieved a regression accuracy of 99.66% in proximity to the actual dataset, signifying its efficacy as a model. The mean squared error (MSE) of 0.0001019 indicates a minimal average squared discrepancy between the predicted and actual values. Furthermore, the root mean squared error (RMSE) of 0.0101 suggests a minor deviation between the actual and predicted values. This is corroborated by the mean absolute error (MAE) values.

Table 2—Model evaluation result.

Model	MAE	MSE	RMSE	R^2 (%)
Random Forest Regression	0.00866	0.0001019	0.0101	99.66

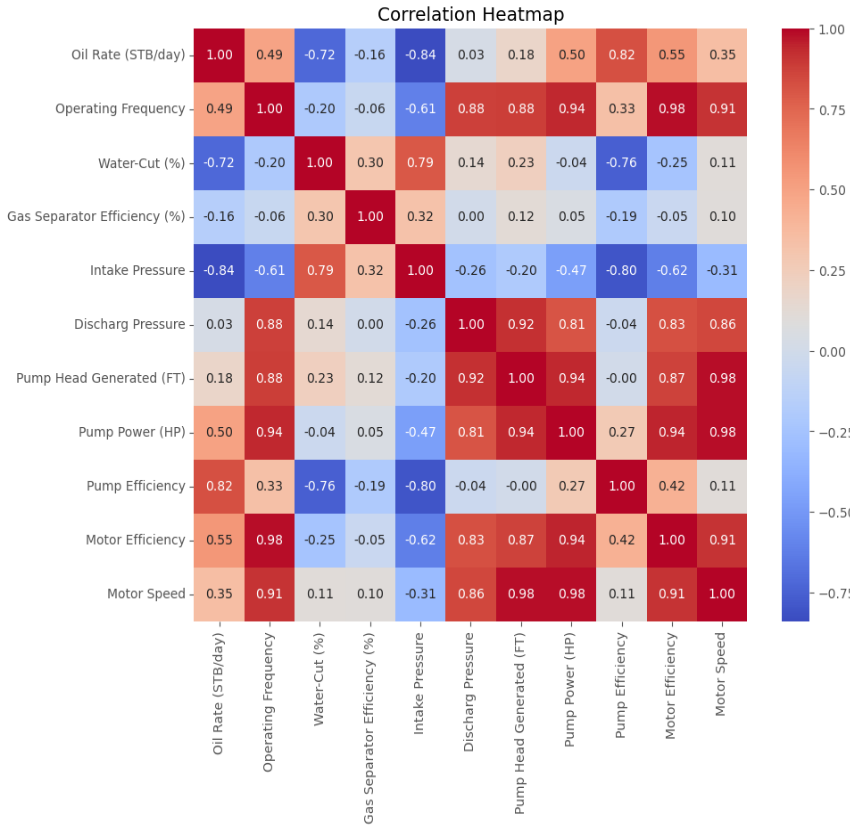

Figure 1—Correlation heat-map.

Upon a thorough examination of **Figure 2**, it becomes quite apparent that the various data points are indeed distributed in a manner that closely aligns with the trend-line. This observation serves to strongly corroborate the reliability and accuracy of the developed model. Consequently, this lends significant support to the notion that the model is well-suited for use in subsequent research endeavors. Furthermore, when the model was applied to the actual oil rate dataset, the results were nothing short of impressive, as evidenced by the near-perfect fit that is clearly illustrated in **Figure 3**. This remarkable alignment between the model's predictions and the actual data points further underscores the model's efficacy and potential utility in practical applications.

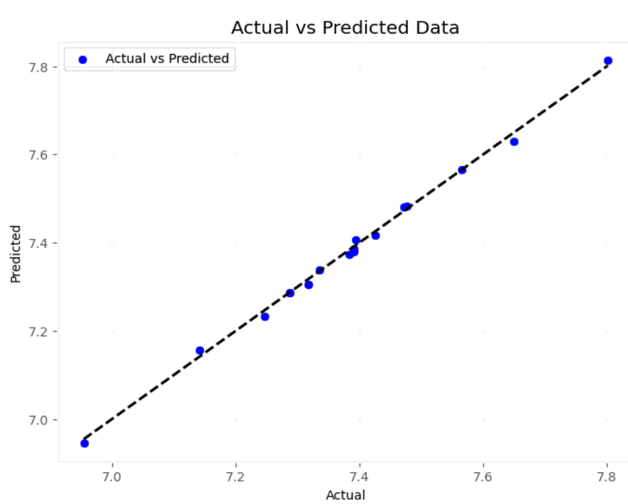

Figure 2—Scatterplot with trend line.

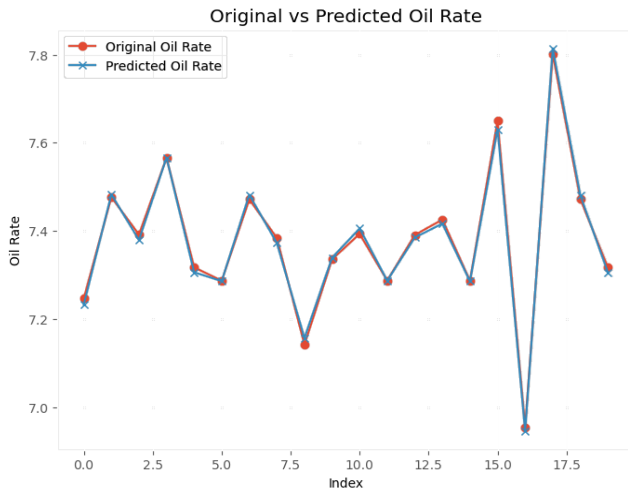

Figure 3—Trend analysis.

Comparison of Models. **Figure 4** illustrates a detailed comparison between the oil production rates that were forecasted by the current study, the actual production data collected from the field, and the predictions that were previously made by Kerunwa et al. (2022). As can be clearly observed in the figure, the oil production rates that were projected by the methodology employed in this study are found to be in closer agreement and more accurately reflect the real-world production data when compared to the predictions generated by the approach outlined by Kerunwa et al. (2022). This striking similarity between the predicted rates from this study and the actual field data serves to validate and reinforce the effectiveness and reliability of the proposed RF-ML algorithm in accurately forecasting oil production rates. This outcome not only underscores the superiority of the RF-ML algorithm but also highlights its potential as a robust tool for making precise and dependable predictions in the domain of oil production forecasting.

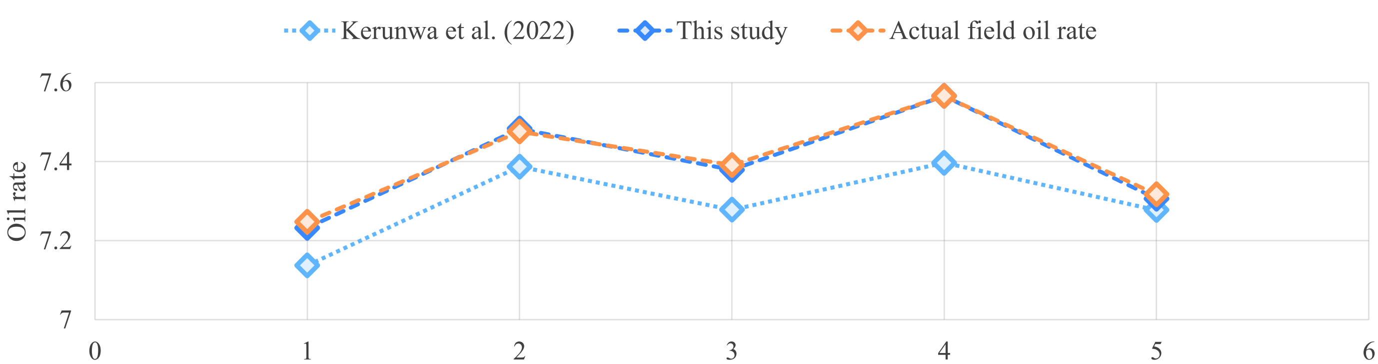

Figure 4—Comparison with previous approach and actual field data.

Conclusion

In the realm of machine learning (ML) research, the application of the Random Forest Regression Algorithm has yielded several significant conclusions. Firstly, the Random Forest regressor demonstrated an impressive 99.6% regression accuracy, which clearly indicates that the model developed through this approach is highly suitable and effective for the intended purpose. This high level of accuracy suggests that the model can make precise predictions and generalizing well from the training data to new, unseen data. Furthermore, when compared to existing models in the field, the Random Forest algorithm exhibited superior performance. This superior performance suggests that the Random Forest model not only meets but exceeds the standards set by previous models, thereby establishing itself as a robust and reliable tool for the task at hand. Specifically, the Random Forest model has shown promise in the domain of Oil Prediction, where accurate and reliable predictions are crucial for various applications, including resource management, market analysis, and strategic planning. The success of the Random Forest regressor in achieving such high regression accuracy and outperforming existing models underscores its potential for widespread adoption in industries that relieve predictive analytics. The algorithm's ability to handle complex datasets and provide accurate forecasts makes it an invaluable asset for researchers and practitioners alike. As a result, the Random Forest Regression Algorithm emerges as a powerful tool in the arsenal of machine learning techniques, particularly for those seeking to enhance their predictive capabilities in the field of Oil Prediction.

Conflicting Interests

The author(s) declare that they have no conflicting interests.

Reference

Avwioroko, J.E., Taiwo, O.A., Mohammed, I.U., et al. 2014. A Laboratory Study of ASP Flooding on Mixed Wettability for Heavy Oil Recovery Using Gum Arabic as a Polymer. Paper presented at the SPE Nigeria Annual International Conference and Exhibition, Lagos, Nigeria, 5-7 August 2014. SPE-172401-MS.

Akpoturi, P. and Ofesi, S.F. 2017. Enhanced Oil Recovery Using Local Alkaline. *Nigerian Jour. of Tech.* **36**(1):515-522.

Austad, T., Alireza, R., and Tina, P. 2010. Chemical Mechanism of Low Salinity Water Flooding in Sandstone Reservoirs. Paper presented at the SPE Improved Oil Recovery Symposium, 24-28 April. SPE-129767-MS.

Bellarby, J. 2009. *Well Completion Design*. Amsterdam, Netherlands: Elsevier.

Dike, C., Izuwa, N. C., Kerunwa, A., et al. 2024. Parametric Evaluation on the Interfacial Tension Response of Agro-Surfactant. *Improved Oil and Gas Recovery* **8**(1):1-12.

Elshan, A. 2013. Development of Expert System for Artificial Lift Selection. Master thesis, Cankaya Ankara, Turkey: Middle East Technical University.

Gomaa, S., Abdelhady, A., Ramzi, H., et al. 2020. Electrical Submersible Pump Design in Vertical Oil Wells. *Pet Petro Chem Eng* J. **4**(5): 000237.

Guo, B., Lyon, W., and Ali, G. 2007. *Petroleum Production Engineering "A Computer-Assisted Approach"*. Amsterdam, Netherlands: Elsevier.

Ikekpeazu, G.E. and Anerobi, J. I. 2020. Oil Recovery by Artificial Lift Systems (ALS): A Review. *Int. Jour. of Innovation. Sci. and Research. Tech*. **5**(1):1-9.

Kerunwa, A., Obibuike, J.U., Okereke, N.U., et al. 2022. Evaluation of Electrical Submersible Pump on Field Production Network Optimization in Niger Delta Oilfield. *Open Journal of Yangtze Gas and Oil* **7**(1): 26-47.

Kerunwa, A., Izuwa, N.C., Dike, C.F., et al. 2024a. Review on the Utilization of Local ASP in the Niger-Delta for Enhanced Oil Recovery. *Petroleum and Coal* **66**(1): 256-275

Kerunwa, A., Dike, C.F., Izuwa, N.C., et al. 2024b. Performance Evaluation of Agro-Materials for Surfactant-Polymer Flooding. *Petroleum and Coal* **66**(1): 308-317.

Okologume, C.W. and Ofesi, S.F. 2016. Comparative Evaluation of Artificial Lift Methods on a Niger-Delta Field. *Academic Research International* **7**(1): 1-16

Sheng, J. 2011. Modern Chemical Enhanced Oil Recovery, Theory and Practice $1^{s}{}_{t}$ ed. Burlington: Gulf Professional Publishin.

Zhu, J. and Zhang, H. 2018. A Review of Experiments and Modeling of Gas Liquid Flow in Electrical Submersible Pumps. *Energies* **11**(1):180-195.

Anthony Kerunwa is an Associate Professor at the Department of Petroleum Engineering, Federal University of Technology Owerri with research interest in drilling, production, reservoir engineering and petroleum economics. Dr. Kerunwa holds a bachelor's degree in Petroleum Engineering from Federal University of Technology Owerri, a Master's degree in Petroleum Engineering from Federal University of Technology Owerri,

and a PhD degree in Petroleum Engineering from Centre for Oilfield Chemicals Research, IPS, University of Port Harcourt.

Chimuanya Winifred Ibewuike is a B.Eng Graduate of Petroleum Engineering with interest in artificial lift design.

Lilian Enyang Ndoma-Egba is a Research Technologist at the Department of Petroleum Engineering, Federal University of Technology Owerri. She has research interest in Drilling Fluids Technology and Cementing Design. Ndoma-Egba holds a Bachelor's Degree in Oil & Gas Engineering, and Master' Degree in Petroleum Engineering.

Ndubuisi Okereke is a Senior Lecturer at the Department of Petroleum Engineering, Federal University of Technology Owerri with research interest in flow assurance, production optimization, drilling, reservoir engineering and procurement. Okereke holds a Bachelor's Degree in Civil Engineering, from Federal University of Technology Owerri, Master's Degree and PhD degree in Subsea Engineering from Cranfield University.

Chukwuebuka Francis Dike is a Research Technologist at the Department of Petroleum Engineering, Federal University of Technology Owerri. He has research interest in drilling, drilling fluids technology, reservoir engineering, enhanced oil recovery and flow assurance. Dike Holds a Bachelor's Degree and Master's Degree in Petroleum Engineering.