Leveraging Machine Learning for Predicting Marginal Field Oil Well Production Using Electrical Submersible Pump Operations Data

Mohamed Zafir Bin Mohamad Gazali[1], Rabeea Jaffari[2]*, Areej Fatemah Meghji[3],
Tuan Mohammad Yusoff Shah[4], Syed Muhammad Noaman Ahmed Shah[5], Muhammad Hammad Rasool[6]
Department of Mechanical Engineering, Universiti Teknologi Petronas, Sri Iskander, Malaysia[1, 4]
Department of Petroleum & Geoscience, Universiti Teknologi Petronas, Sri Iskander, Malaysia[6]
Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan[2, 3]
Westminster International University, Tashkent, Uzbekistan[5]

Abstract—**Electrical Submersible Pumps (ESPs) are critical components of oil well production that can be evaluated for their longevity through daily performance predictions. Traditional predictive methodologies have failed to handle the complexity of dynamic multi-well systems, creating a demand for automated Machine Learning (ML) pipelines based on operational data. In this study, we develop a multivariate time series ESP dataset based on factors impacting the total oil production of four wells (A1-A4) in the Jay field, Malaysia. A naive baseline pipeline used a 95%/5% chronological train-test split across pure regression models and a multivariate Prophet setup. However, due to the presence of data leakage, flat-trend autocorrelation, and the inability to process mechanical well shut-ins, the result of such an approach was unsatisfactory, producing extremely poor R^2 scores. In order to combat poor results and to find a more optimal solution, we have created a synchronized-lag multivariate pipeline evaluated across a 5-Fold Walk-Forward Cross-Validation framework. Noise has been eliminated by cutting leading zeros from the dataset pre-production, while stationarity has been provided through first-order differencing of target values for tracking daily changes (Δ y). To ensure temporal synchronization, a strict one-day chronological lag protocol has been used across all sixteen operational engineering features, splitting them up into a historical baseline (t-1) and an operational delta change feature (t-2 to t-1). The resulting input has allowed us to turn pure regression models into time series models forced to predict changes in targets based on purely historical pump parameters. Results show that non-linear ensemble models perform significantly better than Prophet under rolling validation. The Light Gradient Boosting Machine (LGBM) proved to be the champion model, achieving a statistically significant R^2 score of 0.31. Feature importance analysis showed that the operational adjustments of localized wells (specifically associated Gas_A2_delta and lag_oil_1) drive downstream fluid changes the most. Against the platform capacity of 9,500 barrels per day, LGBM's MAE translates to only 4.1% daily variation forecast.**

Keywords—Oil production prediction; Electrical Submersible Pump (ESP); machine learning; forecasting

I. Introduction

Artificial Lift (AL) is an essential method of combating low production difficulties that most operators of marginal and mature fields are subjected to most of the time. Since its emergence, AL technology has greatly enhanced the rate of production, and it has optimized costs by coming up with innovative solutions. One of the most popular and common forms of AL is the Electrical Submersible Pump (ESP), which is employed to increase the production rate of a low-pressure well. ESPs are essential in the oil and gas industry as a tool for improving oil production in reservoirs. ESPs are also part of the efficient functioning of an oil production well since they are very important in aiding in lifting and transportation of the well production to the surface. Nevertheless, the performance and design of an ESP are highly dependent on factors that do not apply in one well to another, such as the fluid properties, flow conditions, and well completion design [1]. They should be given due consideration during the design of ESP installations to ensure that the ESP system is well adapted to the special conditions of the well. The ESP systems have several important parts, as they are a downhole pump, motor, seal section, and power cable. These components are important in designing and structuring their performance in an ESP.

To maximize and achieve maximum efficiency, the ESP system needs to be customized and tailored to the conditions of the well. This can possibly be in terms of modifications in materials or coating on parts, design modifications in the pumping mechanism, and modifications in the operating parameters of the mechanism. The majority of unplanned Operating Expenditure (OPEX) increments are the result of equipment failures that contribute to production downtime, which results in the state of deferred oil production and will mean a lot of financial implications to the operators [2]. The operators have begun to focus on optimizing production as well as keeping the critical equipment in good condition. This is a big challenge if production reduces following a few years of production plateau, whereby the operators should contemplate the ideal kind of AL, bearing in mind that the properties of the well vary over time.

Sarvestani and Hadipour considered the different AL techniques to mature oilfields and that the data-driven analysis should be provided to optimise the equipment operation and prevent premature failure without investing in the physical changes or workovers [3]. Much of the research has been

*Corresponding author.

carried out in order to enhance production in unconventional fields that use AL. Zalavadia et al. provided optimization of the artificial lift timing and selection (ALTS) through a hybrid data-driven and physics-based workflow and optimal pump unit selection and speed to get the desired production rate [4]. According to Harryanto and Yersaiyn, it is frequently a challenge in calculating which well producers will need AL in the short and medium term, and this depends upon dynamic modelling studies [5]. Nevertheless, the wells will automatically fall out of natural flow, and predicting production with the aid of numerical modelling is challenging because of the constraints in the characterization and the representation of the reservoir heterogeneity.

Although ALS and ESPs are widely used in the production of oil, the existing predictive mechanisms for production optimization are still difficult to adapt and inefficient. Conventional production forecasting tools like Decline Curve Analysis (DCA) [6], reservoir simulation [7], and analytical modeling tools [8] involve a large amount of manual work and have inherent assumptions that lack sufficient expression regarding the dynamics of well conditions. Not only can the combination of Machine Learning (ML) and ESP operations data serve as a strong alternative due to its ability to provide real-time flexibility and flexible capabilities to capture nonlinear, complex relationships in the dynamics of oil wells, but it can also contribute to the significant improvement of the effectiveness of oil well production in marginal fields [9, 10]. Using enhanced ML and analytics, the operators have the chance to enhance the characterization of a reservoir, anticipate failures, as well as enhance the performance of pumps and optimize schedules [11]. Another way in which the integration of ML with numerical modelling can help provide accurate information on which wells will need AL later is through the incorporation of the technology into numerical modelling [12]. This method has the potential to optimize production strategies and reduce operational risks of reservoir heterogeneity.

This study proposes the application of ML to explore the ESP performance and oil well behaviors. The objective of the research is to forecast the future production of oil wells based on the actual performance of the well and the available ESP operation data. The data used in the analysis are obtained from an available offshore facility in Jay in a time series format. Time series is a characteristic of the data, which is essential in comprehending the changing nature of oil production and performance over the years [13, 14]. Wells in this Jay field are also ESP-equipped and are in operation. Two types of ML models, 1) Time series forecasting and 2) Regression models, are explored in terms of performance on the obtained ESP dataset of the Jay field. These models are compared to the prediction results to determine the performance of the most effective method to predict oil production on the dataset. The study also offers in-depth insights into the performance of ESPs in oil wells during oil production and the reasons behind their effectiveness and reliability.

The rest of this study is structured in the following way. Section II of this study briefly discusses the existing ML strategies to forecast the oil production performance. Section III offers the methodology of predicting oil production using an ESP time series dataset. The experimental results and discussions are given in Section IV with the detailed comparative analysis of different types of ML models that were applied in this study. Lastly, in Section V, the conclusion and future work are given.

II. Literature Review

Here, different traditional and ML methods of oil production forecasting have been reviewed.

A. Traditional Techniques of Oil Production Forecasting

1) Decline Curve Analysis (DCA) is a conventional oil and gas well production prediction technique. It approximates historical data to a model such as exponential, hyperbolic, or harmonic decline, which assumes that the output decreases with time [6]. Nevertheless, it presupposes constant decline rates, which frequently cannot capture real effects. It proves to have little success in unconventional reservoirs with non-linear and complex behaviors, such as those found in shales. Furthermore, DCA ignores changing dynamics of reservoirs and changes in operations with time that may make long-term forecasts inaccurate.

2) Reservoir simulation is a model that is used to develop a detailed production forecast of the underground fluid movement with the use of geological, fluid, and operational data. It requires high computing capabilities, skills, and precise input information. Nevertheless, its shortcomings consist of requiring a lot of resources and expertise, geological information, which is usually uncertain, and it is time-consuming and expensive to develop and execute the models [7].

3) Analytical models: simplify the behaviour of reservoirs with equations solvable to predict quick reserves and production. They are based on assumptions such as constant flow, in which accuracy is restricted in complicated situations. The major drawbacks are simplification of actual dynamics and inability to adapt to complex production patterns [8].

Besides the above method-specific constraints, there are some more general constraints with the traditional ways of producing forecasts, which may affect the reliability of the forecasts. These consist of bad quality and access to data, especially at early well stages; changing reservoir and well behaviors such as water breakthrough or formation damage; unpredictable human factors such as decisions of operators, and poor calibration of the models to past data [15].

B. ML-Based Oil Production Forecasting

Given the inefficiency of the conventional approach to predicting the future of the oil industry, operators are dedicating themselves to the implementation of newer technologies based on ML to attain the targeted results [16]. ML is finding use in a range of predictive maintenance and monitoring tasks in the oil and gas industry. Recent studies have also explored AI-driven approaches for evaluating carbon capture, utilization, and storage (CCUS) potential in Kazakhstan's oil fields [17] as well as developing advanced predictive correlations for wax precipitation behavior in crude oil [18]. These studies highlight the growing adoption of intelligent data-driven techniques for improving operational efficiency and sustainability in the sector. This includes process monitoring and fault diagnosis of mooring lines in subsea oil exploration [19], corrosion detection and life expectancy prediction in oil and gas pipelines [20, 21], efficient pressure zone detection [22], and drilling operations [23], and so on. ML technologies could reveal some sophisticated trends in massive data as well as evolve with evolving reservoir and operational conditions. These techniques are more effective in non-linear and high-dimensional data compared to traditional ones, and so they are more likely to be more accurate in changing environments [24].

ML-based Artificial Neural Networks (ANNs) have been employed to optimize various aspects of oil and gas operations. Study [25] demonstrated that ANN models can accurately predict offshore pipeline conditions using historical inspection data, enabling more effective inspection and maintenance planning. E. Khamehchi *et al.* developed an ANN model to estimate key gas lift parameters, including production rates and pressures, using field data, providing a data-driven alternative to traditional empirical correlations that often require difficult-to-obtain inputs [26]. Similarly, another study applied ANNs to predict the remaining life of flexible risers by integrating degradation, operational, and advanced inspection data [27]. However, these approaches do not specifically address oil production forecasting using time-series modeling.

Hosseini and Akilan [14] researched the application of deep learning (DL) and high-performance computing in the oil industry. The study focuses on the oil and gas sector, but specifically on the use of intelligent assistive tools (IATs) in production forecasting. The study under consideration deals with oil production prediction and predictive methods via Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) as accurate, generalizable DL models based on time-series data of the Volve oil field. An autoregressive model and an LSTM are also used by Singh and Pateriya [28] to predict future oil production through an ESP time series dataset. Although these DL methods are very accurate with large and rich datasets, they need large amounts of computational resources. Moreover, these DL models are not very interpretable and thus not suitable for marginal fields with limited data and greater operational variability.

Saghir et al. used streaming time-series analytics to track and measure the performance of progressive cavity pumps (PCPs) in 448 coal seam gas wells [13]. The researchers employed clustering methods to automatically identify the operational states of the pumps and anomalies such as gas locking, solids formation, and torque problems. The strategy assisted engineers in finding out the problems in pumps in real time so that failures can be prevented at the early stages. Although effective in identifying the onset of pump problems, the model fails to anticipate failures before the occurrence of symptoms, it only shows them when abnormal behavior has already been initiated. Also, the clustering task is based on the expert labeling of domains, which is not scalable or automatable. The model is also field-specific (PCPs in coal seam gas wells), and thus, it is not generalized to other types of lift or environments. Even though the research employs a time series analytics technique in the management of natural gas wells, it identifies abnormal performance of the AL systems in real-time, but not the forecasting of oil production of the oil.

Devshali et al. experimented with several models and discovered that an LSTM autoencoder was most effective in identifying anomalies that resulted in failures [29]. Their findings indicated that the model would give as much as 60 days of early warning so that the operators can plan proactive maintenance and minimize downtime. The dataset was also quite limited, with 47 failure events, which is not very large to use with DL. Model performance metrics such as false alarms or prediction confidence were also not very transparent. The experiment is exclusively on the detection of failure and built-in real-time optimization of operations. Also, the LSTM is a black box, which provides no or minimal explainability and information on why it fails.

Agwu et al. apply neural networks in predicting the oil flow rates in ESPs. The model proved to be highly accurate, memory consuming low and could be expressed in an explicit mathematical expression [30]. It was much better than traditional empirical correlations and could be used to estimate ESP rates in real-time situations, as well as in the field. The model was trained and tested on 275 data points gathered in the Middle East ESP-lifted wells and therefore cannot be accurately generalized to other fields or operating conditions. It is also specialized in steady-state behavior, that is, it might not be effective in transient flow regimes or in dynamic processes such as the start-up or shutdown of pumps. Moreover, it fails to take into consideration operational risks such as future failures and thus is more of a performance estimator than a diagnostic or a predictive instrument. The experiment is not developed as a time series.

Although current research has investigated the application of ML techniques to production forecasting, most of the current methods are based on a set of deterministic models with minimal flexibility in modeling intricate and non-linear production processes. There is still a big gap in the capability to come up with a powerful ML-driven system that is not only capable of predicting oil production but also prevents the occurrence of signs of ESP performance degradation beforehand. The reservoir data and patterns can be investigated with the help of ML algorithms over large volumes to enhance the accuracy of reservoir characterization. Using ML in numerical modelling, it could result in an increase in forecasting capabilities and more accurate information on the understanding of which wells will need AL in the future.

The study fills this gap by developing a full ESP data set to be used in oil production prediction via ML and by applying various ML models (time-series prediction and regression models) methodically to compare them in terms of predictive accuracy and reliability. The novelty of this work is the strategy to combine ML techniques with adaptability based on real time, the increased accuracy of predicting the production trend, and the possibility to introduce historical data on the processes of organizational work to plan proactive maintenance. In this study, the proposed research hypothesis is that a data-driven prediction framework will enable more precise and timely insights regarding the optimization of production, which will result in a decrease in the number of unplanned downtime and the optimization of resource allocation, as well as an increase in the ESP operating time.

III. Methodology

This study is executed in three stages, as shown in Fig. 1. The stages are described in the subsequent sections.

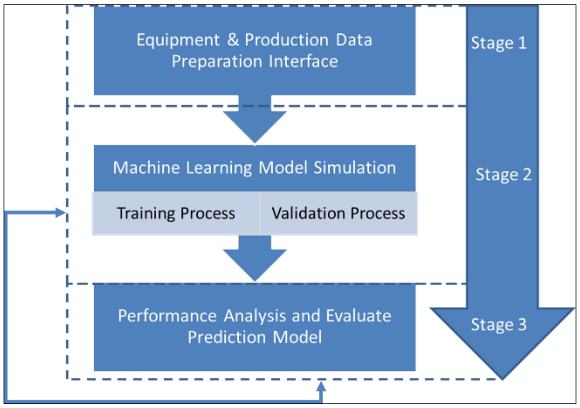

Fig. 1. Research methodology for oil production forecasting using ML models.

A. *Equipment & Production Data Preparation*

This research utilizes real operations data from the Jay field in offshore Peninsular Malaysia. Specific parameters are identified before acquiring the field data. The field data contains daily production data from four oil-producing wells (A1-A4) starting from 8 January 2020 to 26 October 2022. Fig. 2 shows the total field production data.

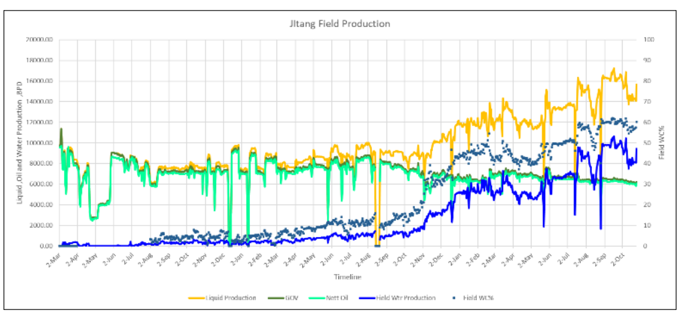

Fig. 2. Jay field production data from Jan 2020 to Oct 2022.

The chart presented in Fig. 2 offers an in-depth analysis of various production parameters over time, which is essential for gaining insight into the oil field's operational performance. The x-axis indicates a timeline from Q1 2020 to Q3 2022, while the two y-axes are used for different parameters. The left y-axis is used to show the production parameters in barrels per day (BPD), while the right y-axis is used for water cut percentage (WC%).

The Liquid Production, represented by the yellow line, refers to the total volume of liquids produced, combining both oil and water production. This line is observed to be at a higher level compared to other statistics, implying a significant level of fluid production throughout the period. The GOV, or Gross Oil Volume, represented by the green line, refers to the overall rate of oil production without factoring in the content of water within the oil, whereas the Net Oil, represented by the aqua line, refers to net oil production after removing impurities or water content from oil. Field Water Production, represented by the dark blue line, refers to the water produced during oil production, showing a steady increase throughout the period, emphasizing its importance as a byproduct of oil production. The chart includes several key production metrics.

One of the most informative aspects of the chart is the "Field WC%" line, which is shown by a dashed black line on the y-axis located on the right side of the chart. "Field WC%" stands for the percentage of water within the total production mix. This line gives insight into the fluid mixture that is being produced from the wells. Particularly, the "Field WC%" line rises over time, especially in the latter period of the chart. This rise indicates that the mixture of oil and water is increasing over time, which could be a sign that the oil well is becoming old or that water is encroaching on the oil wells.

1) Dataset overview: To develop the required time series dataset, the numeric values for the input feature variables (Full Tubing Head Pressure (FTHP), ESP speed, Gas factor, and Water factor for the four wells (A1, A2, A3, and A4)), and the target output total oil production, Nett_oil, are retrieved from the daily production report. The input feature variables are defined as follows:

a) Full Tubing Head Pressure (FTHP): Full Tubing Head Pressure, as the name suggests, refers to the pressure at the top of the production tubing within a well. This is an important parameter when measuring the performance of a well, as it shows how much pressure is being exerted by fluids within the well. Full Tubing Head Pressure is affected by a number of factors, including fluid density, temperature, and fluid flow. When it comes to oil production prediction, Full Tubing Head Pressure is used as an indication of how well a well is doing, and it is an important parameter when trying to calculate oil flow rates and how pressure within a reservoir may change.

b) Electrical Submersible Pump (ESP) speed: This refers to the speed of the Electrical Submersible Pump, which is utilized to pump fluids from oil wells, especially in low-pressure environments. ESP speed is crucial in determining the flow rate of the oil and water being pumped out of the oil well. By varying the ESP speed, the flow rate of the oil and water can be optimized. In oil production prediction, ESP speed is essential in determining the number of hydrocarbons that can be pumped out of the oil well.

c) Gas factor: Gas factor, which is normally given in standard cubic feet of gas per barrel of liquid produced (scf/bbl), refers to the volume of gas present in the oil. Gas factor is one of the important parameters used to understand the interaction of gas and liquid in the oil field. Gas factor also helps in the oil recovery process. This is because if the volume of gas increases, it may indicate a variation in the pressure of the oil field. Gas factor also helps in the estimation of the total production capacity of the oil well.

d) Water factor: Water factor is defined as the volume of water produced per barrel of oil (bbl/bbl), and it is a vital parameter for evaluating the efficiency of oil recovery processes. Water factor provides a means of gaining insight into the water-to-oil ratio, which can be a vital parameter for evaluating the health of the oil reservoir and the efficiency of enhanced oil recovery processes. An increase in the water factor may indicate problems such as water encroachment and inefficient oil recovery. In oil production prediction, the water factor provides a means of evaluating the efficiency of oil production wells and making decisions regarding the management of water production.

A summary of the features used in this research is provided in Table I.

TABLE I. ESP DATA FOR OIL PRODUCTION PREDICTION USING ML MODELS

S.#	Variable	Category	Description	Unit
1	Nett_oil	Target	Total net oil production from all the wells	Barrels of oil per day (bopd)
2	ESP_A1	Feature	Pump speed for Well A1	Hertz (Hz)
3	FTHP_A1		Full Tubing Head Pressure for Well A1	Pounds per square inch (psi)
4	ESP_A2		Pump speed for Well A2	Hertz (Hz)
5	FTHP_A2		Full Tubing Head Pressure for Well A2	Pounds per square inch (psi)
6	ESP_A3		Pump speed for Well A3	Hertz (Hz)
7	FTHP_A3		Full Tubing Head Pressure for Well A3	Pounds per square inch (psi)
8	ESP_A4		Pump speed for Well A4	Hertz (Hz)
9	FTHP_A4		Full Tubing Head Pressure for Well A4	Pounds per square inch (psi)
10	Gas_A1		Gas produced by well A1	Million standard cubic feet per day (MMscf/d)
11	Water_A1		Water produced by well A1	Barrels of water per day (bwpd)
12	Gas_A2		Gas produced by well A2	Million standard cubic feet per day (MMscf/d)
13	Water_A2		Water produced by well A2	Barrels of water per day (bwpd)
14	Gas_A3		Gas produced by well A3	Million standard cubic feet per day (MMscf/d)
15	Water_A3		Water produced by well A3	Barrels of water per day (bwpd)
16	Gas_A4		Gas produced by well A4	Million standard cubic feet per day (MMscf/d)
17	Water_A4		Water produced by well A4	Barrels of water per day (bwpd)

The developed time series dataset with 1,023 records and 17 columns is depicted in Fig. 3 via a dataset snapshot. As shown in Fig. 3, the chronological nature of the dataset, based on the defined column for dates, provides a means of evaluating the impact of operations and environmental factors on the production output over the period of interest, thereby creating a holistic approach for forecasting the production dynamics in the future. The remaining 16 columns are related to the different parameters for the four wells: A1-A4, as defined in Table I.

Date	Nett_Oil	ESP_A1	FTHP_A1	ESP_A2	FTHP_A2	ESP_A3	FTHP_A3	ESP_A4	FTHP_A4
2020-01-08	0	0	0	0	0.0	0	0.0	0	0.0
2020-01-09	0	35	808	35	847.0	35	795.0	35	866.0
2020-01-10	0	35	766	35	820.0	35	913.0	35	866.0
2020-01-11	0	35	934	35	753.0	35	910.0	35	866.0
2020-01-12	0	35	910	35	758.0	35	901.0	35	867.0
...	...	...	...	...	...	...	...	...	...
2022-10-22	6035	46	665	44	707.0	44	651.0	44	691.0
2022-10-23	6000	46	669	44	702.0	44	650.0	44	691.0
2022-10-24	5914	46	670	44	709.0	44	652.0	44	689.0
2022-10-25	5812	46	668	44	707.0	44	651.0	44	687.0
2022-10-26	6077	48	699	44	709.0	45	669.0	44	687.0

Fig. 3. Formulated ESP time series dataset snapshot.

It is also important to note from Fig. 3 that the initial data points for the oil production target variable 'Nett_Oil' are observed to have zero or near-zero values. This could mean that the initial period of operation was perhaps the start of operations or perhaps a period of downtime. As we move along the dates, especially in 2022, it is clear that there is a marked increase in these production values. The 'ESP' and 'FTHP' columns also show variations that could be related to the operations or perhaps the productivity of the wells or variations in the pumping and tubing equipment. These are critical for diagnostic purposes to determine whether the ESPs were running efficiently or perhaps facing pressure-related problems.

B. Application of ML Models for Oil Production Prediction

Two categories of ML models are applied to predict total oil production: time series forecasting and regression models. The details of these ML categories are as follows:

1) Time series forecasting: This is performed by utilizing the Prophet model [31], where the model has been optimized for forecasting, interpretability, and simplicity, allowing it to handle outlier and missing data challenges robustly. The time series is decomposed by Prophet into three components: Trend, which shows directional changes in production over time; Seasonality, which shows periodic changes in production; and Noise, which shows random variations not explained by trends or seasonality. This decomposition provides a rich understanding of production, which could be used for forecasting. The uncertainty level of Prophet forecasts is set at 0.95, allowing for possible future trend changes and providing a range of predictions. ESP speed, FTHP, gas, and water production are used as features, and oil production, or Nett_oil, is used as the target variable. Prophet has been optimized for its ability to work with complex time series data with seasonal components, making it suitable for use in a dynamic environment like oil production.

2) Regression models: In addition to time series forecasting, six regression models are applied to the dataset to predict Nett_oil as a numeric target. The regression models include Linear Regression [24], Bayesian Linear Regression

[24], Neural Network Regression [32], Decision Tree Regression [33], Boosted Decision Forest Regression [34], and Light Gradient Boosted Machine (LGBM) [35]. Unlike the time series approach, these models do not incorporate the date as an input feature but instead treat the prediction problem as a regression task with the other operational features as predictors. Each regression model is trained using the hold-out method, where 95% of the data (972 records) is used for training and 5% (51 records) for testing. This methodology ensures that the models were evaluated based on their ability to generalize to unseen data. Table II depicts the description and the hyperparameters configuration for the mentioned regression models.

TABLE II. ML REGRESSION MODELS HYPERPARAMETERS CONFIGURATION

S.#	ML Model	Description	Hyperparameters Configuration
1	Linear Regression [24]	Assumes a linear relationship between the independent input/ feature variables and the dependent output/target variable	Single parameter, learning rate 0.1, number of epochs 10, L2 regularization weight 0.001
2	Bayesian Linear Regression [24]	Uses linear regression supplemented by additional information in the form of a prior probability distribution. Prior information about the parameters is combined with a likelihood function to generate estimates for the parameters.	Regularization weight: 1
3	Neural Network Regression [32]	Uses adaptive weights and can approximate non-linear functions of the input features for output predictions.	Single parameter, fully connected, 100 hidden nodes, learning rate 0.005, number of epochs 100, momentum 0, normalizer min-max normalizer
4	Decision Forest Regression [33]	An ensemble of decision trees where each tree outputs a Gaussian distribution as a prediction. An aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees.	Single parameter, Resampling method Bagging, Number of decision trees 8, Maximum depth of the trees 32, Number of random splits per node 128, Minimum number of samples per leaf node 1.
5	Boosted Decision Tree Regression [34]	An ensemble of regression trees using boosting. Boosting means that each tree is dependent on prior trees. Learns by fitting the residuals of the trees that preceded it.	Single parameter, maximum number of leaves per tree 20, minimum number of samples per leaf node 10, learning rate 0.2, total number of trees constructed 100
6	Light Gradient Boosting Machine (LGBM) [35]	An ensemble of regression trees using boosting based on the histogram-based learning technique, allowing the trees to grow using a leaf-wise strategy instead of the level-wise strategy. Reduces training time and resource usage.	Number of leaves 31, maximum depth 1, learning rate 0.1, total number of trees constructed 100

3) ML Model evaluation and benchmarking: To assess the performance of the ML models, several evaluation metrics were employed: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) as described in (1)-(3) respectively.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(Predicted_i - Actual_i)^2}{N}} \tag{1}$$

$$MAE = \sum_{i=1}^{N} \frac{|Actual_i - Predicted_i|}{N} \tag{2}$$

$$R^2 = \frac{\sum_i (Actual_i - Predicted_i)^2}{\sum_i (Actual_i - Meanof Actual_i)^2} \tag{3}$$

The value of RMSE helps to estimate the magnitude of the difference between predicted values and real ones, where bigger differences are punished more, thanks to squared differences. The smaller the RMSE, the better the fit. MAE helps to evaluate the average error between the predicted and actual values in an easy-to-understand way, where all differences are treated equally irrespective of the size of the error. This metric is helpful to estimate the accuracy of the prediction. Lastly, R^2 helps to evaluate the share of variance of the target variable (oil production) that can be explained by the model, and values close to one mean that the model explains much of the data. These metrics were calculated for each model to understand the relative efficiency of the models.

All ML models were developed and tested with the help of the Python programming environment Google Colab [36]. The analysis of the metrics of models helped to compare different models and to optimize the training strategy accordingly, as depicted via a feedback loop in Fig. 1.

IV. RESULTS AND DISCUSSION

The results for training all the ML models on the ESP oil production dataset are presented in Table III, with best values highlighted in bold.

TABLE III. TOTAL OIL PRODUCTION PREDICTION RESULTS FROM VARIOUS ML MODELS

S.#	ML Model	Category	RMSE	MAE	R^2
1	Linear Regression [24]	Regression	1893.58	1889.7	-227.36
2	Bayesian Linear Regression [24]		809.02	686.92	-40.68
3	Neural Network Regression [32]		720.44	650.87	-32.05
4	**Decision Forest Regression [33]**		**260.30**	**217.67**	**-3.31**
5	Boosted Decision Tree Regression [34]		773.06	535.72	-37.06
6	Light Gradient Boosting Machine (LGBM) [35]		382.33	279.25	-8.31
7	**Prophet [31]**	**Time Series Forecasting**	**259.09**	**212.47**	**-3.27**

Table III depicts that the multivariate configuration of the Prophet time-series forecasting model originally appeared to perform the best, yielding the lowest absolute error values (RMSE = 259.09, MAE = 212.47) and a top baseline R^2 value of -3.27. Within the pure ML regression category, the Decision

Forest model emerged as the second-best architecture, followed closely by LGBM with an RMSE of 382.33, an MAE of 279.25, and an R^2 of -8.31. Conversely, the plain gradient boosting method utilized in the standard Boosted Decision Tree model scored the worst among the ensemble techniques, proving that standard boosting routines were unviable compared to the bagging and light gradient boosting mechanisms utilized in the Decision Forest and LGBM models. Additionally, the linear frameworks (Linear and Bayesian Linear Regression) performed the worst overall, with R^2 values of -227.36 and -40.68, and RMSE values of 1893.58 and 686.92, respectively. This statistical collapse confirms that linear models lack the representational capacity required to capture the complex, non-linear boundaries governing fluid dynamics and multi-well decline curves.

In particular, the structural issue resided in the way that the baseline was executed. All models (1-6) were implemented purely as cross-sectional regression algorithms, which were fed solely with static feature vectors and the target variable, devoid of any time series-related metrics, therefore making them completely blind to time series patterns in the data. The only model that considered the nature of the problem was the Prophet model. Moreover, all the baseline configurations yielded negative R^2 values, showing a common overfitting constraint of using a standalone validation window. Running models on one 51-day segment (5% test data), being the absolute end of the timeline, hid this structural issue behind deceptively low baseline absolute errors, but made the fit statistically non-viable when encountering any variance.

To reduce variance and create a statistically solid basis for forecasts, a stationary and synchronized in-time pipeline was developed in a 5-Fold Walk-Forward Cross-Validation setting for the top three models (Prophet, Decision Tree, and LGBM). First, data noise was removed by safely cutting off the flat production records before the beginning of oil production, thus bringing the starting point of the timeline to the day of first oil production. Second, the target was transformed into first-order differences, and the model predicted Δy.

To bridge the gap identified in the baseline phase, the pure regression architectures (Decision Forest and LGBM) were fully converted into time-series-aware models. A strict one-day chronological lag protocol was applied globally across all sixteen input feature columns. Every engineering metric was split into a historical performance baseline (day t-1) and an operational delta column tracking mechanical changes (between day t-2 and day t-1). This structured input space forced the tree models to forecast tomorrow's volume adjustments using exclusively known, historical pump parameters and rolling target lags (lag_oil_1). Following model execution, the predicted change vectors were dynamically added back to the historical baseline (y_{t-1}) to calculate final error minimization performance metrics directly on a raw barrel scale as shown in Table IV.

Table IV shows that the synchronized lag alignment allowed the LGBM to firmly cement its lead, securing a positive, valid R^2 score of 0.308, closely followed by the Decision Forest architecture at 0.291. This performance shift highlights that tree-based gradient boosters and parallel ensemble arrays excel at utilizing differenced engineering deltas to resolve sharp day-to-day fluid variations once they are armed with true sequential historical features. While this rigorous cross-validation naturally raised LGBM's absolute MAE to 391.52 barrels, it successfully corrected the model's structural bias, effectively trading a fragile, localized baseline for a robust, mathematically sound, and generalized forecasting framework.

TABLE IV. TOTAL OIL PRODUCTION PREDICTION RESULTS FROM THE TOP PERFORMING ML MODELS UNDER OPTIMIZED TIMELINE CONFIGURATIONS

S.#	ML Model	RMSE	MAE	R^2
1	Decision Forest Regression [33]	679.50	378.71	0.29
2	LGBM [35]	671.14	391.52	0.30
3	Prophet [31]	1929.73	1505.40	-4.40

Concurrently, the performance of the Prophet architecture significantly declined under the cross-validation framework, with its error envelope widening to an MAE of 1505.41 barrels and an R^2 of -4.401. Because Prophet treats external regressors as static linear additive components to its core trend line, its collapse under a rolling validation loop exposes a fundamental flaw in how its equations process physical asset data over time, where relationships between line pressures (FTHP), pump speeds (ESP), and net oil volumes continuously evolve. When forced to evaluate on the newly engineered historical lags and delta changes across expanding rolling horizons, Prophet's linear coefficients became highly unstable and failed to generalize to the subsequent test folds. Furthermore, as an additive curve-fitting tool built on continuous Fourier series, Prophet cannot process the sudden mechanical zero-production records caused by operational well shut-ins, resulting in massive overpredictions when production abruptly resumes.

Lastly, across all models, the absolute regression metrics (RMSE and MAE) appear high because the target net oil production values (Nett_Oil) fluctuate up to 10,000 bopd while input features operate in mere tens. However, LGBM's final MAE translates to a relative daily forecasting variation of only 4.1% against a platform capacity exceeding 9,500 barrels, proving it is highly effective and reliable for navigating unpredictable operational shutdowns.

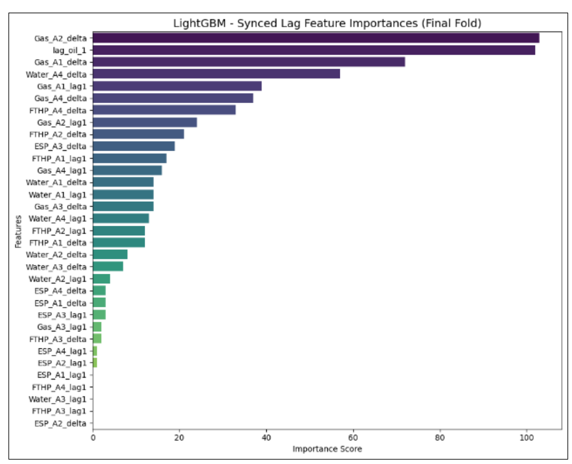

Fig. 4. Feature importance for total oil production using LGBM.

Feature importance diagnostics from the final fold of the champion LGBM model, as depicted in Fig. 4, confirm that the pipeline successfully forced the algorithm to learn from reservoir physics and mechanical transformations. The model assigned its highest predictive weights to Gas_A2_delta (103) and the historical target lag (lag_oil_1) (102), proving that tracking localized directional changes in associated gas rates and immediate production history is crucial for capturing fluid swings. Furthermore, the strong emergence of operational change parameters like Gas_A1_delta (72), Water_A4_delta (57), and FTHP_A4_delta (33) demonstrates that LGBM effectively maps physical disruptions, such as valve manipulations and pressure shifts, directly to downstream oil production fluctuations.

V. Conclusion and Future Work

The purpose of the current research is to expand the use of ML models to increase the accuracy and effectiveness of daily oil production predictions. ML can greatly aid in overcoming numerous limitations that emerge during the application of conventional approaches for prediction and optimization of oil production processes. In this research, an optimized, synchronized-lag multivariate pipeline was engineered to predict daily oil production using real-time ESP data from four active wells (A1-A4). By implementing target differencing to track daily changes and safely trimming pre-production zeroes, the pipeline corrected structural biases and achieved a positive, valid R^2 score of 0.31 using the LGBM model. The results indicate that non-linear ensemble models armed with sequential historical features outperform traditional additive curve-fitting tools like Prophet, which remain structurally incapable of handling sudden mechanical well shut-ins. Feature importance diagnostics confirmed that localized operational changes, specifically associated gas fluctuations (Gas_A2_delta) and immediate historical production lags (lag_oil_1), represent the most dominant parameters driving downstream fluid variations.

For future work, these optimized frameworks can be trained on real-time sensor streams to actively prescribe pump adjustments. Advanced deep learning architectures, such as LSTM networks, can be deployed to dynamically tune pump speeds and flow rates to maximize production while minimizing power consumption. Additionally, ML can be applied to predictive maintenance and early fault detection. By tracking patterns across historical pressure deltas and lift parameters, models can detect anomalies indicating mechanical degradation, alerting operators to potential pump failures before they occur to minimize unplanned downtime.

Acknowledgment

Acknowledgment is extended to the Malaysian marginal field operator who provided the field data used in this study. The organization name and field name cannot be revealed as per the conditions of confidentiality stipulated.

Data Availability

The data that form the basis for this study's findings are not openly accessible. Availability is conditional on request, depending on certain approvals.

References

[1] B. A. Alqahtani, M. A. Alqahtani, and A. M. Alqahtani, "Electric submersible pump setting depth optimization-a field case study," in SPE Middle East Artificial Lift Conference and Exhibition, 2022: SPE, p. D012S009R002.

[2] M. Abdelaziz, R. Lastra, and J. Xiao, "ESP data analytics: Predicting failures for improved production performance," in Abu dhabi international petroleum exhibition and conference, 2017: SPE, p. D031S089R003.

[3] A. Darvish Sarvestani and A. Hadipour, "artificial lift method selection for mature oil fields: A case study," in SPE annual caspian technical conference, 2019: SPE, p. D033S013R014.

[4] H. Zalavadia, P. Singh, U. Sinha., S. Snakaran, "Improving Arificial Lift Timing, Selection, and Operations Strategy Using a Physics Informed Dara-driven Method," in Unconventional Resources Technology Conferemce, Denver, Colorado, June 2023.

[5] E. Haryanto, S. Yersaiyn, A. R. Shah, S. Su, and F. Bouchet, "Management of artificial lift planning & deployment for improved field development plan," in SPE Asia Pacific Oil and Gas Conference and Exhibition, 2020: SPE, p. D021S018R002.

[6] C. Thakuria, M. Al-Ghnemi, A. Bora, and M. Al-Bahar, "Decline Curve Analysis for Production Forecast Assurance-A Novel Approach," in SPE Kuwait Oil and Gas Show and Conference, 2019: SPE, p. D033S015R004.

[7] N. Dousi, F. Syarkawi, and I. Ignatov, "The Application of Reservoir Simulation in Reserves Estimation Methods to Estimate the Range of Technical Uncertainty," in Abu Dhabi International Petroleum Exhibition and Conference, 2020: SPE, p. D021S026R002.

[8] M. Meyet, R. Dutta, and C. Burns, "Comparison of decline curve analysis methods with analytical models in unconventional plays," in SPE Annual Technical Conference and Exhibition, 2013: SPE, p. D031S041R006.

[9] P. Bangert, "Preventing Slugging by Tuning Choke through Machine Learning," in International Petroleum Technology Conference, 2020: IPTC, p. D023S160R003.

[10] C. Zhan, S. Sankaran, V. LeMoine, J. Graybill, and D.-O. S. Mey, "Application of machine learning for production forecasting for unconventional resources," in SPE/AAPG/SEG Unconventional Resources Technology Conference, 2019: URTeC, p. D023S029R004.

[11] P. Pankaj et al., "Application of data science and machine learning for well completion optimization," in Offshore Technology Conference, 2018: OTC, p. D041S051R005.

[12] S. Unrau, P. Torrione, M. Hibbard, R. Smith, L. Olesen, and J. Watson, "Machine learning algorithms applied to detection of well control events," in SPE Kingdom of Saudi Arabia annual technical symposium and exhibition, 2017: SPE, p. D043S039R001.

[13] F. Saghir, M. E. G. Perdomo, and P. Behrenbruch, "Performance analysis of artificial lift systems deployed in natural gas wells: A time-series analytics approach," Geoenergy Science and Engineering, vol. 230, p. 212238, 2023.

[14] S. Hosseini and T. Akilan, "Advanced deep regression models for forecasting time series oil production," arXiv preprint arXiv:2308.16105, 2023.

[15] I. Yemez, V. Singh, N. H. Orellana, and E. Izaguirre, "Factors contributing to uncertain production forecasts: 3D reservoir modelling perspectives," in International Petroleum Technology Conference, 2014: IPTC, pp. IPTC-18196-MS.

[16] S. Perry et al., "Autoregressive and Machine Learning Driven Production Forecasting–Midland Basin Case Study," in Unconventional Resources Technology Conference, 26–28 July 2021, 2021: GeoScienceWorld, pp. 3104-3117.

[17] B. Khusain et al., "A Python-based Evaluation of Kazakhstan's Fields for Carbon Capture, Utilization, and Storage Projects", Eng. Technol. Appl. Sci. Res., vol. 15, no. 2, pp. 20782–20789, Apr. 2025.

[18] A. Khussainova et al., "Advanced Correlations for Predicting Wax Precipitation in Crude Oil: A Study on Melting Point and Solid-State Transition Temperatures", Eng. Technol. Appl. Sci. Res., vol. 15, no. 2, pp. 21505–21517, Apr. 2025.

[19] R. Jaffari, M. A. Hashmani, H. Taib, N. Abdullah, and S. S. H. Rizvi, "A novel image-based framework for process monitoring and fault diagnosis of mooring lines," Journal of Hunan University Natural Sciences, vol. 47, no. 10, 2020.

[20] H. Yang, L. Lu, and K. Tsai, "Machine learning based predictive models for CO2 corrosion in pipelines with various bending angles," in SPE Annual Technical Conference and Exhibition, 2020: SPE, p. D041S057R002.

[21] M. Sharifi, K.-T. Yao, S. Raghavendra, I. Ershaghi, R. House, and J. Blouin, "Prediction of remaining life in pipes using machine learning from thickness measurements," in SPE Western Regional Meeting, 2015: SPE, pp. SPE-173995-MS.

[22] M. H. Rasool, R. Jaffari, M. Ahmad, N. A. Siddiqui, A. Z. Junejo, and M. A. Abbas, "Novel deep learning framework for efficient pressure zone detection via analysis of pore pressure profiling," Arabian Journal for Science and Engineering, vol. 50, no. 1, pp. 237-248, 2025.

[23] C. I. Noshi and J. J. Schubert, "The role of machine learning in drilling operations; a review," in SPE eastern regional meeting, 2018: SPE, p. D043S005R006.

[24] C. M. Bishop and N. M. Nasrabadi, Pattern recognition and machine learning (no. 4). Springer, 2006.

[25] M. S. El-Abbasy, A. Senouci, T. Zayed, F. Mirahadi, and L. Parvizsedghy, "Artificial neural network models for predicting condition of offshore oil and gas pipelines," Automation in Construction, vol. 45, pp. 50-65, 2014.

[26] E. Khamehchi, F. Rashidi, and H. Rasouli, "Prediction of gas lift parameters using artificial neural network," Enhanced Oil Recovery–Iranian Chemical Engineering Journal (Special Issue), vol. 8, no. 43, 2009.

[27] K. Francis and V. Chaves, "Flexible riser remnant life assessment using artificial neural networks and advanced inspection methods," in Offshore Technology Conference, 2018: OTC, p. D031S030R004.

[28] A. Singh and R. Pateriya, "Artificial oil lift production forecasting and analysis using autoregressive and deep learning models," Available at SSRN 3396282, 2019.

[29] S. Devshali et al., "Predicting ESP failures using artificial intelligence for improved production performance in one of the offshore fields in India," in Abu Dhabi International Petroleum Exhibition and Conference, 2022: SPE, p. D032S183R001.

[30] O. E. Agwu, A. Alkouh, S. Alatefi, R. A. Azim, and R. Ferhadi, "Utilization of machine learning for the estimation of production rates in wells operated by electrical submersible pumps," Journal of Petroleum Exploration and Production Technology, vol. 14, no. 5, pp. 1205-1233, 2024.

[31] Facebook. "Prophet" https://facebook.github.io/prophet/ (accessed Jan 11, 2025).

[32] C. C Aggarwal, "Neural networks and deep learning: a textbook," ed: Springer, 2018.

[33] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, Classification and regression trees. Chapman and Hall/CRC, 2017.

[34] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189-1232, 2001.

[35] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol. 30, 2017.

[36] Google. "Welcome To Colab." https://colab.research.google.com/ (accessed May 2025, 2025).