A New Drilling Rate of Penetration Prediction Model by Particle Swarm Optimization and Gradient Boosting Regression

Faris Aiman Jamaluddin[1], Marina Yusoff[2], Diva Kurnianingtyas[3], Mohamad Taufik Mohd Salledud-din[4]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI),
Universiti Teknologi MARA (UiTM), Shah Alam, 40450, Selangor, Malaysia[1]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI),
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam, 40450, Selangor, Malaysia[2]
Faculty of Computer Science, Universitas Brawijaya, Malang, 65145, Indonesia[2,3]
PETRONAS Research Sdn Bhd, Kawasan Institusi Bangi, Bandar Baru Bangi, Selangor[4]

***Abstract*—The oil and gas industry continuously evolves to enhance operational efficiency and productivity while minimizing costs and environmental impact. Among the critical aspects of oil and gas operations, drilling efficiency is a key factor in accessing underground hydrocarbon reservoirs. Traditional machine learning models and current regression models have shown limitations in accurately modelling the rate of penetration due to the high nonlinearity of data. This project focuses on the rate of penetration prediction for drilling optimization. This study proposed a new drilling rate of penetration prediction model with the embedding of particle swarm optimization in a gradient boosting regression method. A solution representation of the particle is introduced as a hyperparameter strategy to explore the optimal parameter for predicting drilling datasets. Extensive experiments were carried out using two splitting strategies, One for All and All for One, across the drilling wells. The proposed GBR+PSO hybrid model achieved a mean absolute error of 1.205, representing a reduction of approximately 89.68% compared to the best-performing baseline model, K-Nearest Neighbor with the One for All splitting strategy, which achieved a mean absolute error of 11.68. The hybrid solution could enhance drilling ROP predictions, advancing the drilling rate of penetration strategy. It has the potential to support the development of autonomous drilling optimization, thus contributing to more efficient, reliable, and cost-effective drilling rate of penetration strategies in future drilling operations.**

***Keywords*—Drilling; gradient boosting regression; machine learning; rate of penetration; particle swarm optimization**

I. Introduction

The oil and gas industry, along with its related sectors, is continually seeking innovative and competitive ways to enhance operational efficiency and productivity [1] [2]. The objective is primarily to minimize costs and reduce the environmental impact. Reaching this objective requires improving processes in oil and gas operations; one of them is the drilling process to estimate and adjust the penetration rate (ROP) [3] [4]. In recent years, demands on the oil and gas sector have continued to increase. However, the cost of offshore exploration and development is very high, especially for hydrocarbon production [5]. The ROP is a a key drilling performance indicator through subsurface formations, regardless of whether the drilling is targeting oil, gas, or any other substance. ROP plays a vital role, serving as a fundamental metric in drilling operations in various industries, primarily in oil and gas exploration [6], geothermal energy production [7], mineral exploration [8], and environmental monitoring [9]. In spite of the critical role of ROP prediction in drilling optimization, understanding how drilling variables influence ROP remains an open question in drilling [10][11][12]. ROP presents a highly nonlinear problem. A lot of efforts had been established as early as a traditional models as such Bourgoyne and Young model in 1974 [13] and until an advanced AI methods [14][17][18].

Despite these significant advancements, existing studies still require further improvement in optimization efficiency, adaptive feature engineering, computational complexity reduction, explainability, robustness against geological variability, and real-time adaptive learning capability for intelligent drilling applications. Although ML has become more popular in drilling engineering, there is still a lack of thorough studies that concentrate on ML and optimization methods for ROP prediction. Therefore, the main aim of this research is to improve the ROP prediction by using a hybrid approach of ML and optimization methods while utilizing the University of Stavanger Rate of Penetration (USROP) dataset as a main resource [19]. Key questions addressed include the necessity of incorporating ROP variables into ML models, the comparative effectiveness of the hybrid approach versus traditional models for ROP estimation, and the utilization of predictive models for optimization purposes. The research questions are:

- How to construct a new proposed model compared to existing machine learning methods in terms of MAE for predicting ROP in drilling operations?
- What is the effect of different training, testing, and validation splitting scenarios on the predictive performance of the proposed model across varying drilling well conditions?
- To what extent does the improved prediction accuracy of the proposed model translate into measurable gains in drilling efficiency, intelligent drilling optimization, and real-time operational decision-making?

This study proposes a hybrid data-driven method to predict ROP by integrating GBR and PSO. Our contributions in hybrid ROP prediction are:

- The study introduces a new GBR+PSO hybrid model for ROP prediction, demonstrating enhanced predictive performance with significantly lower MAE values than existing machine learning methods.
- A thorough training, testing, and validation incorporating MAE analysis and benchmarking, splitting scenario of the drilling well.
- The approach improves drilling efficiency by increasing a significant prediction accuracy across different conditions in intelligent drilling optimization and operational decision-making strategies.

The organization of the study is structured as follows: Section II presents the related work. Section III is dedicated to the preliminaries, which consists ROP optimization and prediction, score metrics and a proposed scenario. In Section IV, the explanation of the material and methods includes the proposed hybrid GBR+PSO method and performance measure. The results and performance evaluations are discussed in Section V and Section VI, respectively, and Section VII concludes the study.

II. Related Work

Recent advancements in ROP prediction have increasingly adopted artificial intelligence (AI), machine learning (ML), and deep learning (DL) approaches to improve drilling efficiency and predictive reliability [20]. Accurate ROP prediction remains challenging due to nonlinear drilling behavior, noisy sensor signals, heterogeneous geological formations, concept drift, and high-dimensional drilling parameters [21][22]. Inappropriate feature selection may introduce redundant variables, resulting in reduced prediction accuracy and poor model generalization [20]. Consequently, recent studies have emphasized intelligent feature selection, adaptive learning, and optimization strategies to enhance prediction performance and real-time drilling capability [21][22]. For instance, Attention Mechanism-enhanced Bidirectional Long Short-Term Memory (AM-BiLSTM) model integrated with Kalman filtering and mutual information analysis for feature selection, where dominant drilling parameters such as weight on bit (WOB), total depth, weight on hook, and flow pumps were identified as significant predictors [20]. Even though AM-BiLSTM model achieved significant improvements in Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) compared to Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and BiLSTM-based models, further enhancement is still required to strengthen feature dependency learning capability. In particular, improving adaptive feature representation and model generalization across varying drilling environments remains a critical challenge.

In another research, a hybrid framework combining Fuzzy C-Means, Dynamic Time Warping, Genetic Algorithm, and Extreme Learning Machine, where correlation analysis was employed for feature selection and GA was used to optimize the ELM model under different drilling conditions [12]. The GA-optimized ELM achieved the best prediction accuracy among comparative methods, highlighting the importance of drilling condition identification and optimization strategies for accurate ROP prediction. Furthermore, a hybrid online ROP prediction framework incorporating formation drillability (FD) sensing, Savitzky-Golay filtering, and incremental learning strategies, where the inclusion of FD as an additional predictive feature improved prediction accuracy by at least 19% compared to six existing methods [22]. The similar performance on the optimized LSTM achieved the best overall performance with prediction precision greater than 96.87%, minimum Average Relative Percentage Error (ARPE), MAPE, RMSE, and bias, as well as maximum R^2 performance [10].

III. Preliminaries

A. ROP Optimization Process and Prediction

ROP optimization is possible in numerous ways. Fig. 1 illustrates two general methods for this optimization process on ROP. The ROP model is utilized to determine the optimal drilling parameters for a specific optimization problem, such as minimizing drilling time, drilling MSE, or other objectives [19]. This is then used to establish standard drilling parameters as such of weight on bit or drill bit in rotary speed (RPM). Fig. 1(a) and Fig. 1(b) indicate the difference between continuous learning [20][19] and a reference well method [19]. As seen in Fig. 1(a), the model is able to be constructed in real-time using continuous learning. This necessitates training the model while drilling, which is more challenging to accomplish due to the computational requirements of the software and hardware. The first dataset is extremely small and skeletal at the beginning, which is another disadvantage that slows down the machine learning training process. There is a warm-up phase in which no model is used or a temporary model is employed. A reference well, as demonstrated in Fig. 1(b) is analogous to a drilled well. This enables users to construct models offline that need to be done only once, and is generally more easier compared than the continuous learning methods. In both methodologies, the optimum drilling parameters are computed using the ROP model and then applied in the drilling controls for parameters such as bit weight, RPM and mud flow [13]. The method would be based on local constraints and data availability. Best practices do not yet exist however, a hybrid approach is also feasible.

Effective ROP optimization depends on the development of a reliable ROP prediction model. Traditionally, both analytical and data-driven models were initiative and improvised to achieve this target [15]. In order to train the machine learning model or estimate model constants, both methods require reference datasets. Reality and the constructed models should have a smaller predicted error that would be references for the future drilling. This implies that the precision of a model can be utilized only for similar drilling such as tools, methods and depth. In modern drilling operations, continuous learning techniques have gained prominence [16][20]. Information to be referred, or trained, can be sourced from the local wells at a given time in a continuous learning process in parallel of the model creations.

B. Score Metrics

MAE is suggested as the key metric for ROP prediction.

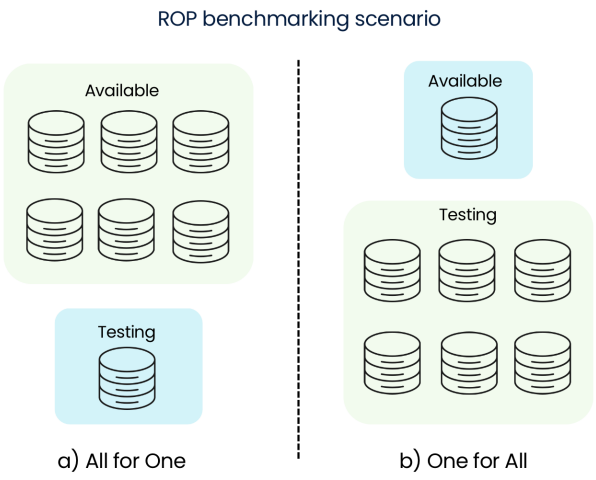

Fig. 1. ROP performance evaluation strategy.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{1}$$

where, n = number of samples, t = sample number consecutively, At = actual value at sample t, and Ft = forecasted value at sample t. The reason for this choice is that ROP modelling is mostly done for drilling time optimization, where the interest is in the cost per meter drilled. A specific error value, e.g., 10 m/h, will be of roughly the same significance to an operator whether the true value is 30 m/h or 80 m/h [19]. While MAE is widely used, other studies may employ alternative performance metrics such as the MAPE and the Weighted Mean Absolute Percentage Error (WMAPE). MAPE, defined as a percentage of the actual values used to estimate error, can be useful for relative comparisons of performance across ROP ranges. However, it becomes problematic when actual values approach zero. WMAPE, on the other hand, handles sample imbalance by weighting errors by the true values, making it a more robust alternative in highly heterogeneous datasets. The evaluation measure must be properly selected based on the specific purpose and data characteristics of the study in ROP prediction.

C. Proposed Scenario

In ROP prediction, three performance evaluation strategies are adapted from the previous work of Tunkiel et al. [19] to evaluate ROP prediction models. The scenarios are All for One (AFO) and One for All (OFA). AFO, as illustrated in Fig. 1(a), in which one complete well is set aside for testing and all but one well are available for training and validation. This necessitates a final MAE score for all wells combined and enables seven different runs, with varying wells set aside for testing as cross-validation. Fig. 1(b) shows the One for All (OFA) situation, where only one well is used to estimate, and the rest are estimated by this model, as suggested by the final situation. This creates seven train/test iterations as cross-validation, just like the second scenario. Regarding iteration independence, in both the AFO and OFA strategies, each iteration is treated as an independent entity. This demands that the algorithms designed must be able to work with data from a single iteration without being influenced by preceding iterations. This habit would make it possible to guarantee that the models remain objective and do not inadvertently incorporate irrelevant information that could undermine their precision.

IV. Materials and Techniques

This section elaborates on the materials, data sources, and research methodologies described. The proposed approach captures the steps to see the performance of the enhancement of GBR+PSO models on ROP prediction optimization. The approach includes data preparation, pre-processing, construction of the proposed methods, and evaluation. In addition, we introduced a particle representation or solution mapping for the PSO. Fig. 2 demonstrates the flow of the proposed methodology. This process begins with the data extraction and pre-processing stage, where raw data is cleaned of outliers, missing values are imputed, overlapping data are removed, and the data are sorted. The pre-processing stage is rigorously carried out to ensure the equality and consistency of datasets, thereby ensuring that high-quality data are gathered and creating a strong model. The subsequent feature selection yields an optimal dataset containing the most descriptive predictors of the ROP model. The essence of the methodology lies in the predictive modelling achieved using machine learning techniques.

Three regressors namely, AdaBoost, Gradient Boosting, and K-Nearest Neighbours are assessed for performance across multiple algorithmic paradigms. Hyperparameter optimization made possible through PSO is applied to optimize the performance of selected models. PSO is a population-based stochastic optimization procedure well suited to navigating complicated, high-dimensional search spaces, making it suitable to determine optimal parameter configurations.

Model validation involves using performance measures such as MAE and R^2, which provide numerical measures of predictive accuracy and model fitness. The convergence analysis of the PSO algorithm is also conducted to ensure the stability and reliability of hyperparameter tuning. The design system consists of two approaches, namely AFO and OFA, enabling the comparison of model performance across different training and test paradigms.

The availability in both the training and the testing phases leads to a comprehensive analysis of the models. The training process involves tuning the selected machine learning models to the preprocessed data, while the testing process assesses the generalization performance of the models on unseen data. The integration of a combination of robust preprocessing, advanced optimization techniques, and rigorous validation presents an all-encompassing and methodological solution to ROP prediction, which could be beneficial in generating useful knowledge for enhancing drilling efficiency and optimizing resource recovery in shale reservoirs. Strategic comparison between AFO and OFA methods enables the diversification of model training and testing performance under varying operational conditions. Detailed steps are elaborated in the following sub-sections.

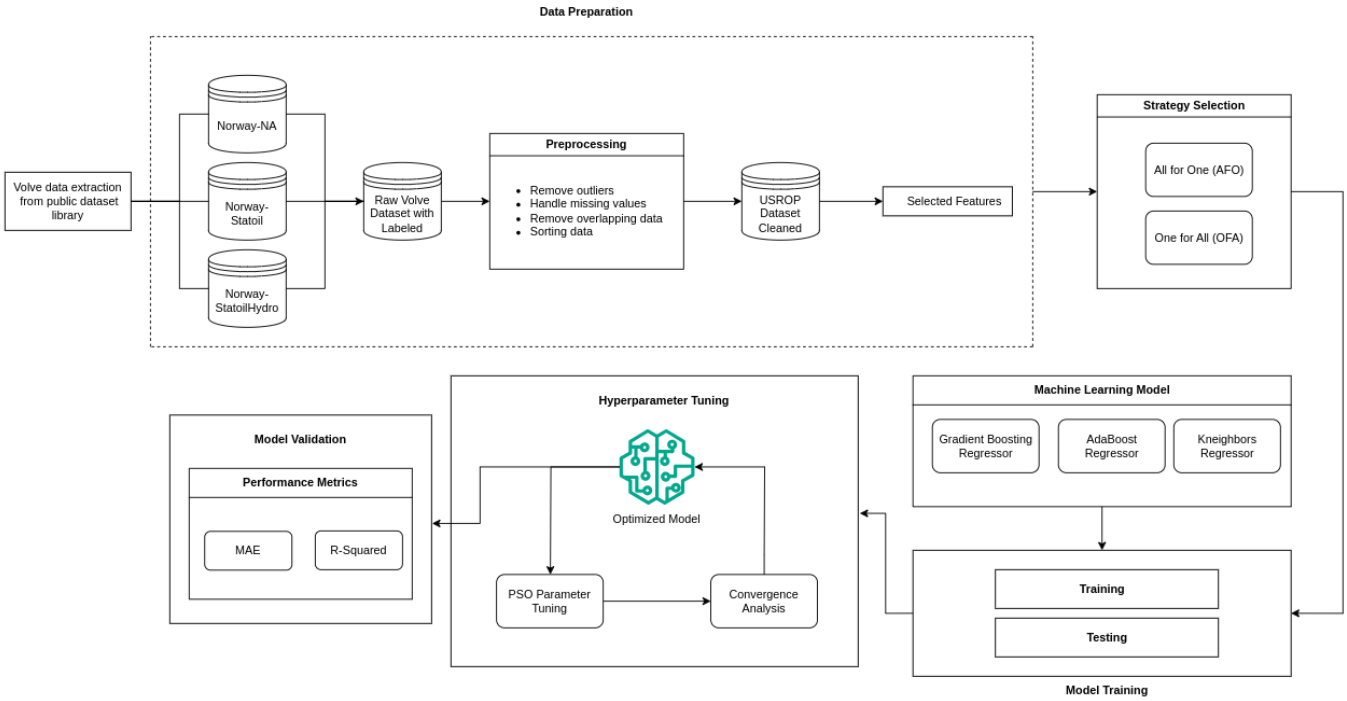

Fig. 2. Flow of methodology.

A. Data Preparation

This study utilized the same precise data preparation as that of the USROP reference paper. Experiments in numerous drilling areas are facilitated by the Volve database's large-scale real-time drilling data. Meanwhile, the logs differ in features, contain missing values, and vary in quality. A good subset of Volve must address this problem. Seven wells in total were selected based on the logged attributes. Data quality and the logged common attributes were the selection criteria. The datasets were acquired from the USROP dataset, focusing on the Volve field. [19]. The attributes are selected: Measured Depth [m], Weight on Bit [kkgf], Average Standpipe Pressure [kPa], Average Surface Torque [kN.m], Rate of Penetration [m/h], Average Rotary Speed [rpm], Mud Flow In [L/min], Mud Density In [g/cm3], Diameter [mm], Average Hook-load [kg], Hole Depth (TVD) [m], USROP Gamma [gAPI]. Minimal processing was done to the attributes to preserve the original data. This is necessary as the drilling logs often contain erroneous, non-physical values. There may be sentinel values indicating no reading (typically -999), corrupted values from the mud-pulsing system, and others. Samples containing weights on bit values below 0 and above 35 were truncated. The same way rows with mud density in, mud flow in, and average surface torque values below zero were removed, as well as with ROP values above 100 and average standpipe pressure above 25,000. Meanwhile, the diameter refers to the nominal wellbore diameter. Forward and backward filling were used to fill small gaps in the data resulting from uneven logging frequency across different equipment.

Distribution of ROP values by frequency across five ranges shows a pronounced leftward skew. The concentration of data in the 1-20 m/h and 21-40 m/h ranges suggests that the majority of drilling operations experienced quite low penetration rates. This may indicate natural limitations imposed by the penetrated geological rocks, such as high rock strength or unfavourable drilling conditions. Alternatively, it can indicate suboptimal drilling operations or equipment inefficiencies that prevent higher ROP. The declining rate observed with increasing ROP intervals also indicates the challenge of achieving high penetration rates. The explanation for this trend could be rising depth, which typically leads to higher compressive stresses and poorer rock drillability. The diminishing returns of increasing weight on bit (WOB) or rotary speed at high ROP values could also be a factor for this trend. A detailed investigation of geological conditions, drilling parameters, and operating practices is critical to elucidate the underlying causes of this skewed ROP distribution and to identify potential means of enhancing drilling performance.

B. Feature Correlation Coefficient

This correlation heatmap illustrates the interrelations among various drilling parameters and their impact on ROP (Fig. 3). Colour intensity and direction (red positive, blue negative) indicate the strength and direction of the linear relationships among variables. Surprisingly, ROP also shows a moderate negative correlation with Measured Depth (-0.39) and Hole Depth (TVD) (-0.42), indicating a decrease in ROP with increasing depth, perhaps due to factors such as increasing formation compaction and pressure. Conversely, there is a moderate positive correlation between ROP and Mud Flow In (0.25), suggesting the potential for greater ROP at higher mud flow rates, possibly due to more effective cuttings transport and hole cleaning.

Even though this heatmap is useful as a first-level view

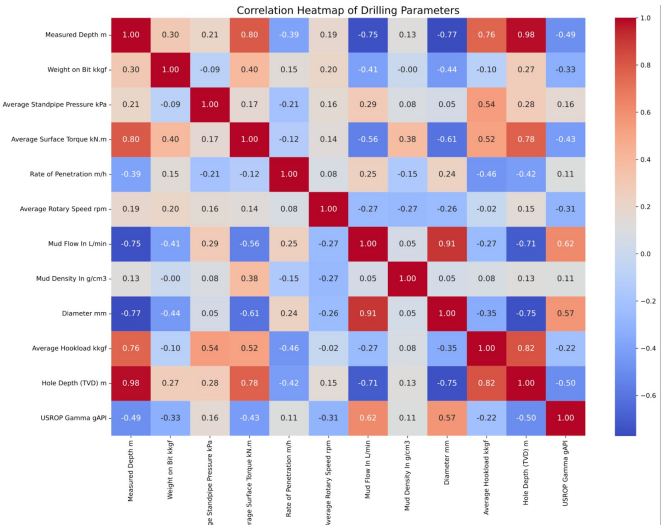

Fig. 3. Features correlation

TABLE I. PARTICLE REPRESENTATION FOR GBR TUNING

Variable	Hyperparameter range
X_1	`n_estimators`: 50–500
X_2	`learning_rate`: 0.01–0.3
X_3	`max_depth`: 3–10
X_4	`min_samples_split`: 2–20
X_5	`min_samples_leaf`: 1–20
X_6	`subsample`: 0.5–1.0
X_7	`max_features`: 0.1–1.0

of drilling parameter correlations with ROP, it has its own limitations. First, the correlations presented are specific to the dataset employed and may not extrapolate to other geological structures and drilling conditions. Second, the interpretation is restricted to linear relationships, whereas the actual interaction of parameters and ROP might be more complex and non-linear. Therefore, further study using new analysis techniques, such as non-linear regression modelling and machine learning algorithms, is needed to develop a larger, predictive ROP estimation model. Further consideration of business sense and operational feasibility will become crucial to demonstrate the actual, real-world practicability of the prediction model when implemented.

C. Proposed Solution

1) Solution representation: Following is the solution mapping or particle representation (see Table I):

whereas,

X1 to X7 are labels used to denote specific GBR parameters in a structured format.

R1 to R7 are shorthand notations corresponding to the ranges of parameters X1 to X7.

X1, R1: n_estimators: Range from 50 to 500.

X2, R2: learning_rate - Range from 0.01 to 0.3.

X3, R3: max_depth - Range from 3 to 10.

X4, R4: min_samples_split - Range from 2 to 20.

X5, R5: min_samples_leaf - Range from 1 to 20.

X6, R6: subsample - Range from 0.5 to 1.0.

X7, R7: max_features - Range from 0.1 to 1.0.

2) Fitness function: The fitness function was calculated using test data from previous references (Tunkiel et al., 2021a). The equation given below describes the fitness function used to calculate MAE, which is applied within the PSO algorithm.

$$\min\left(\frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|\right) \tag{2}$$

whereas,

leftmargin=1.15em

- n: Total number of data points (hours) for drilling speed.
- t: Counter that iterates through each data point, going from 1 to n. Keeps track of the specific hour analysed within the calculation.
- A_t: Actual drilling speed measured at a specific hour (t). A_1 would be the speed for the first hour, A_2 for the second, and so on.
- F_t: Predicted drilling speed for the same hour (t).
- $|A_t - F_t|$: Difference between the actual speed (A_t) and the predicted speed (F_t). It ignores the direction of the error (positive or negative).
- $\sum$: Sum of all the absolute differences between actual and predicted speeds for all the hours ($t = 1$ to $t = n$).
- $(1/n)$: Represents the average.

In this setup, MAE serves as the objective function which PSO aims to minimize. MAE is mainly used to measure the accuracy of the GBR model's predictions. Each particle in the PSO algorithm represents a potential set of GBR hyperparameters. The fitness of each particle is evaluated by training the GBR model with its corresponding hyperparameters and calculating the MAE on a validation dataset.

3) A proposed new hybrid GBR+PSO: In this study, a new hybrid GBR+PSO algorithm is proposed (see Algorithm 1). The following is a step-by-step of a new hybrid GBR+PSO algorithm. The PSO algorithm iteratively updates particles' positions and velocities based on their own best-known positions, their neighbours' best-known positions, and the global best-known position. This update process continues until the algorithm converges on the optimal hyperparameters that yield the lowest MAE, ensuring the most accurate ROP predictions. This dynamic interplay between MAE and PSO effectively guides the search for the optimal hyperparameters by utilizing PSO's exploration and exploitation capabilities to fine-tune the GBR model. Based on Barbosa et al. [23] has been shown few past studies have used the MAE as an objective function for ML optimization. In this PSO setup for GBR hyperparameter tuning, the particles' parameters (hyperparameters for the GBR model) and the objective function (MAE) are tightly linked:

- Particles' parameters (GBR hyperparensuree and train the GBR model.
- The MAE is computed based on the predictions of the GBR model with tcouldhose hyperparameters and serves as the objective function that PSO optimizes.
- The goal of PSO in this setup is to minimize the prediction error of the GBR model. MAE is a straightforward measure of average error and fits well as the objective function for this minimization problem.

Algorithm 1 A New Hybrid GBR+PSO

- Begin.
- Initialize the PSO algorithm for GBR hyperparameter tuning.
- Define the particle representation.
- Initialize the swarm with a population of particles, each representing a candidate set of GBR hyperparameters.
- Set the fitness function as MAE, calculated using Equation (2), where: leftmargin=1.5em,noitemsep,topsep=1pt
  - n: Total number of data points (hours) for drilling speed
  - t: Iteration counter
  - A_t: Actual drilling speed at time t
  - F_t: Predicted drilling speed at time t
- Evaluate the initial fitness of each particle by training the GBR model using its hyperparameters and calculating the MAE on a validation dataset.
- Update the velocity and position of each particle based on: leftmargin=1.5em,noitemsep,topsep=1pt
  - The particle's personal best position (pBest)
  - The global best position found by the swarm (gBest)
- Iterate until the stopping criterion is met (e.g., a fixed number of iterations for convergence to the lowest MAE).
- Select the optimal GBR hyperparameters corresponding to the particle with the lowest MAE.
- Train the final GBR model using the optimized hyperparameters.
- Measure the model training time.
- Predict the test dataset using the trained GBR model.
- Evaluate the final GBR model's performance using MAE.
- End.

GBR was selected for advanced hyperparameter tuning using PSO, with validation performed using the AFO splitting strategy . Unlike the initial phase, this phase utilized all relevant GBR parameters to achieve optimal model performance. The PSO setup, as outlined in Table III, was configured with 100 iterations and 100 particles, default values to ensure robust exploration of the hyperparameter space. The inertia weight was set to 0.5, while the cognitive and social coefficients were both set to 2.0. These parameters were chosen to balance exploration of new solutions and exploitation of known good solutions, ensuring the algorithm efficiently converges to the best hyperparameter configuration. The PSO algorithm works by iteratively adjusting the position of each particle based on its personal best solution and the global best solution found by the swarm. The inertia weight controls the influence of a particle's previous velocity, while the cognitive and social coefficients guide the particles toward self-improvement and collaboration, respectively. This approach ensures a comprehensive search of the hyperparameter space, minimizing the MAE and enhancing the predictive accuracy of the GBR model.

D. Performance Evaluation

This study has used two essential empirical measurements to evaluate and compare the effectiveness of GBR+PSO. By running this experiment and applying PSO to the GBR, the main goal of this study is to identify the optimal set of hyperparameters for the GBR, including the learning rate, the number of trees, and tree depth, that minimise prediction error while maximising overall model performance, as measured by the MAE metric. This optimization step is an important part of my experiment, as it is represented to significantly enhance the model's accuracy and robustness. Plus, these improvements will contribute to more precise and reliable ROP predictions and enable more efficient and informed decision-making in drilling operations. A proposed hybrid GBR+PSO was compared with GBR, ADR and KN models. In evaluating the performance of GBR+PSO, the hyperparameter ranges for GBR were carefully selected to ensure meaningful exploration and avoid impractical configurations. The lower and upper bounds for the hyperparameters were set as follows: `n_estimators` (50 to 500), `learning_rate` (0.01 to 0.3), `max_depth` (3 to 10), `min_samples_split` (2 to 20), `min_samples_leaf` (1 to 20), `subsample` (0.5 to 1.0), and `max_features` (0.1 to 1.0). These ranges were selected based on their significant influence on the performance of the GBR model. For instance, `n_estimators` control the number of boosting iterations, with higher values potentially improving accuracy at the cost of increased computational time. The `learning_rate` determines the contribution of each tree to the final model, with smaller values promoting more robust learning but requiring more iterations. The `max_depth` parameter limits the depth of each tree, preventing overfitting by controlling model complexity. Similarly, `min_samples_split` and `min_samples_leaf` balance underfitting and overfitting by regulating the minimum number of samples required to split a node and the minimum number of samples per leaf node, respectively. The `subsample` parameter introduces randomness by using a fraction of samples for fitting each tree, enhancing generalization. Finally, `max_features` controls the proportion of features considered for each split, reducing overfitting while maintaining predictive accuracy.

By leveraging PSO to fine-tune these hyperparameters within the specified bounds, the GBR model achieves an optimal balance between predictive accuracy and computational efficiency. The iterative nature of PSO ensures that the best hyperparameter configuration is selected based on the training data, resulting in a robust, high-performing model. This approach not only enhances the predictive capabilities of the GBR model but also provides a systematic and efficient method for hyperparameter optimization, making it a valuable contribution to the field of machine learning and predictive

modeling. In this implementation, PSO is used to fine-tune the hyperparameters of the GBR. The optimization process involves 100 iterations, especially during which particles (candidate solutions) explore the search space to minimize the MAE. Each particle will then adjust its position based on its previous experience (personal best) and the best-known solution found by the swarm (global best). The algorithm utilizes inertia weight to control how much a particle's previous velocity influences its current movement, balancing exploration and exploitation. These two coefficients guide the particles: the cognitive coefficient, encouraging self-improvement, and the social coefficient by promoting collaboration toward the best global solution. The parameter ranges are constrained by specific lower bounds ([50, 0.01, 3, 2, 1, 0.5, 0.1]) and upper bounds ([500, 0.3, 10, 20, 20, 1.0, 1.0]) to ensure meaningful exploration. These settings allow PSO to iteratively refine hyperparameter combinations, efficiently converging to an optimal GBR configuration.

The GBR is a machine learning algorithm that builds decision trees to minimize prediction error. In this implementation, the hyperparameters optimized by PSO should significantly influence GBR's performance as follows:

leftmargin=1.15em

- `n_estimators`: The number of boosting iterations (trees) in the model, ranging from 50 to 500. More iterations can improve accuracy but may increase computation time.
- `learning_rate`: Determines the contribution of each tree to the final model. A smaller value (e.g., 0.01 to 0.3) results in more robust learning but requires more iterations.
- `max_depth`: Limits the depth of each tree (3 to 10), controlling model complexity and preventing overfitting.
- `min_samples_split`: Minimum number of samples required to split a node (2 to 20), balancing underfitting and overfitting.
- `min_samples_leaf`: Minimum number of samples per leaf node (1 to 20), affecting tree size and generalization.
- `subsample`: Fraction of samples used for fitting each tree (0.5 to 1.0), introducing randomness for better generalization.
- `max_features`: The proportion of features considered for each split (0.1 to 1.0), which reduces overfitting while maintaining predictive accuracy.

By tuning these hyperparameters based on lower bound and upper bound, the GBR model is able to achieve an optimal balance between predictive accuracy and computational efficiency. Next, PSO ensures the selection of the best configuration based on the training data.

V. Computational Results

This section highlights the computational results of GBR, KN and ABR and the proposed GBR+PSO. The experiments were perform based on an appropriate parameters for each of the methods, GBR and ABR, utilizing the different learning rate of 0.1 and 0.01 meanwhile KN based on the number neibours. All experiments tested on AFO and OFA benchmark splitting strategy.

A. Computational Results Using GBR

GBR employed the OFA methodology with a learning rate tinkered to 0.01 and the same 80/20 train-test split. This experiment was therefore meant to investigate the lower learning rate's impact while targeting a particular drilling scenario. Similar to GBR-EXP1,GBR-EXP3, GBR-EXP4 has shown a promising learning curve. With little tweak on the parameter, we can clearly see the gap between the training and test error showing that to some extent the model is generalizing well. The training error goes down smoothly, which means it is learning from the data. However, the train error also goes down in the beginning and then flattens. This would then mean that, in learning, there is an optimum size of training beyond which the generalization performance does not increase significantly. It appears that the optimal training size is around 60,000. Beyond this point, the model starts to overfit, and the test MAE starts to increase. This may be due to various factors involving data complexity, limitations of model architecture, or noise in data. Further analysis, such as hyperparameter tuning and feature engineering, could potentially improve the model's performance and further reduce the test error.

B. Computational Results Using KNR

In KNN-EXP1, the AFO method has been used with n_neighbors = 3. This setting reduces the neighbourhood size, thus should be allowing the model to capture more local fine details. The experiment uses an 80/20 train-test split to examine the model's behavior when trained on most of the dataset and validated on a smaller portion. Training MAE just remains very low since the model manages to easily fit to the data being trained from. On the other hand, test MAE shows its height relatively throughout the process. This large difference between the training and testing MAE indicates underfitting criteria of the model.

C. Computational Results Using ABR

The learning curves for ABR experiments from ABR-EXP1 to ABR-EXP4, all reflect the overfitting pattern. Whereas the training error decreases rapidly, indicating that the model fits the training data well, either the test error saturates or increases. This also reflects that the model memorizes the training data rather than learning the underlying pattern. This is further supported by the overfitting behaviour evident from the violin plot, since the MAE test values in the various experiments are widely distributed, with noticeably large variability in performance. This could mean that the current ABR model architecture, with its default hyperparameters, is overly simplistic for handling such a large dataset range of 120,000 to 200,000 samples. The model's complexity may not be sufficient to learn the complex relationships in such a large dataset. Hence, it overfits the training data and generalises very poorly with high prediction errors on unseen data.

D. Computational Results Using GBR+PSO

In assessing the impact of particle representation on a new hybrid GBR+PSO performance, the results GBR+PSO obtained MAE: 1.205, which indicate good MAE value and good optimization process and R^2: 0.982 indicates goodness-of-fit measure for linear regression models. From the results, it demonstrates that the hyperparameter tuning procedure, which was driven by PSO, has successfully discovered an optimal set of hyperparameters for the GBR model. This is demonstrated by the algorithm's iterative exploration of various hyperparameter combinations, which resulted in a "best new configuration" with improved MAE values at each step. In the PSO algorithm for hyperparameter tuning, some important parameters guide the search process. This parameter refers to cognitive factor (C1) and social factor (C2). The cognitive factor parameter, which is set to 2.0 by default, gives an idea of how much a particle is influenced by its own pBest to continue the exploration of good areas previously found. On the other hand, the social factor C2 = 2.0 will relate to the degree of attraction by a particle for the gBest discovered by all the swarm. That enhances collaboration or sharing of information between particles. Along with the inertia weight which is set at 0.5 to maintain a balance between exploration and exploitation the described parameters enable this swarm to dive into the space of hyperparameters. The PSO algorithm explored 100 iterations with a swarm of 100 particles evaluating many possible configurations or possible feasible solutions. Table II demonstrates the selected hyperparameter configurations of the Configuration 1 and Configuration 2.

TABLE II. THE SELECTED HYPERPARAMETER CONFIGURATIONS

Parameters	Configuration 1	Configuration 2
n_estimators	95	425
learning_rate	0.166916	0.132280
max_depth	3	6
min_samples_split	17	14
min_samples_leaf	8	9
subsample	0.716358	0.533660

Configuration 1 and Configuration 2 obtains MAE of about 4.178 and 1.205, respectively. The algorithm successfully identified a new best configuration in the second iteration, significantly reducing the MAE from 4.178 to 1.205. This improvement highlights the effectiveness of the PSO algorithm in optimizing hyperparameters, leveraging the interplay between C1, C2, and inertia weight to find a high-performing model. The best fitness (lowest MAE) achieved during the hyperparameter tuning process is 1.205, which corresponds to Configuration 2 in the training output demonstrating the practical benefits of using PSO for hyperparameter tuning.

Fig. 4 shows the "Predicted vs. Actual ROP" plot of GBR+PSO experiment provides a compelling visual representation of the model's performance, corroborating the findings from the learning curve and the hyperparameter tuning process. The strong positive correlation between predicted and actual ROP values, evident in the tight clustering of points around the diagonal line, demonstrates the model's ability to accurately capture the underlying relationships in the data. The low MAE which is 1.205 and of about R^2 of 0.982 values further support this observation. While the plot reveals a strong

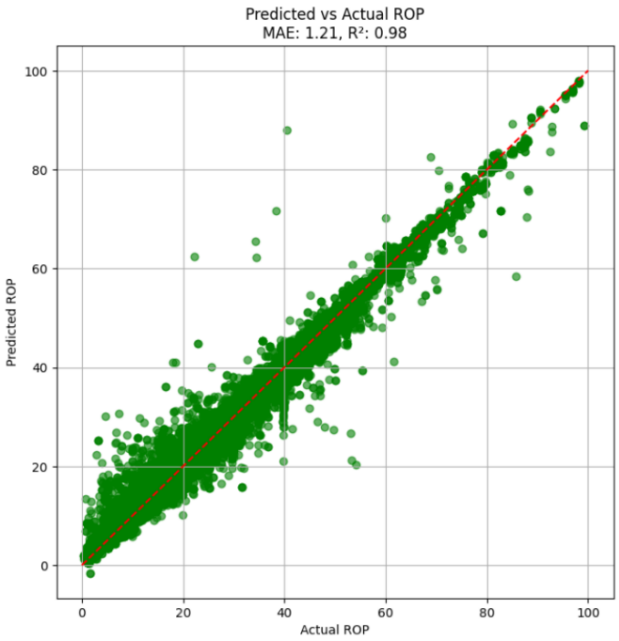

Fig. 4. Predicted versus actual ROP for the best GBR+PSO experiment.

overall correlation, it also provides valuable insights for further refinement. By analyzing the distribution of points and identifying areas where deviations from the diagonal line, we can pinpoint specific scenarios or data regions, where the model's predictions might be less accurate. This plot gives insight to guide further hyperparameter tuning, feature engineering, or data collection efforts to enhance model performance and improve the accuracy and reliability of ROP predictions.

VI. DISCUSSION

The experimental results presented in the tables provide a comprehensive comparison of various machine learning models, including GBR, Adaboost, KNR, and GBR+PSO. The performance of these models is evaluated using the MAE metric, which is particularly suitable for predicting the ROP in drilling operations. MAE measures the average absolute difference between predicted and actual values, providing a clear, interpretable metric of model accuracy. For ROP prediction, where even small deviations can have significant operational implications, MAE is an appropriate metric because it directly quantifies the average error in predicted ROP values, ensuring the model's predictions are reliable and actionable. The performance metrics for different models of the ROP prediction is demonstrated in Table III.

The superior performance of the GBR+PSO model, with an MAE of 1.205, highlights the effectiveness of PSO in fine-tuning hyperparameters. PSO's ability to efficiently explore the hyperparameter space and balance exploration with exploitation allows it to identify configurations that minimize prediction errors [24][25]. This is particularly important for ROP prediction, where the relationship between input parameters includes weight on bit, standpipe pressure, and rotary speed and ROP is complex and non-linear. The learning curve analysis for GBR+PSO would likely show a steady decrease in training and validation errors as the number of iterations

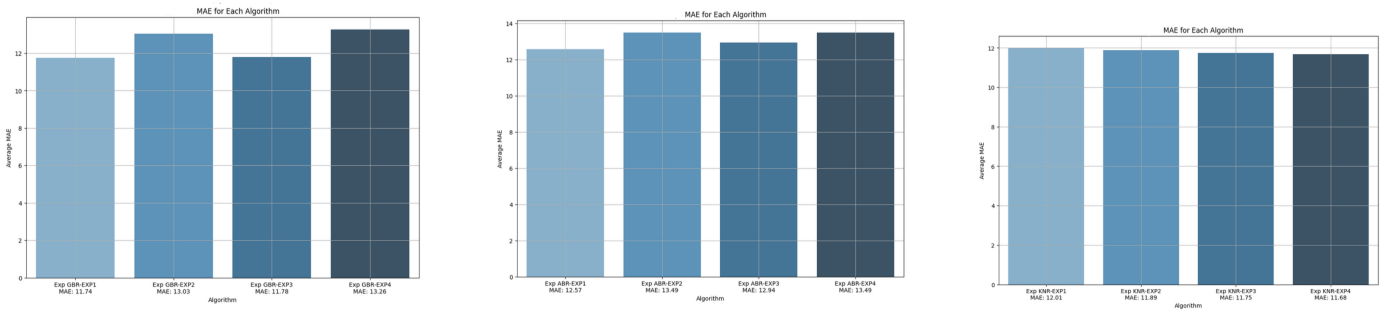

Fig. 5. MAE score for the standalone experiments: (a) GBR, (b) ABR, and (c) KNR.

TABLE III. PERFORMANCE METRICS FOR DIFFERENT MODEL CLASSIFIERS.

Method and Experiment ID	Benchmark Splitting Strategy	Parameter	MAE
GBR-EXP1	AFO	Learning Rate: 0.1	11.74
GBR-EXP2	AFO	Learning Rate: 0.01	13.03
GBR-EXP3	OFA	Learning Rate: 0.1	11.78
GBR-EXP4	OFA	Learning Rate: 0.01	13.26
ABR-EXP1	AFO	Learning Rate: 1.0	12.57
ABR-EXP2	AFO	Learning Rate: 0.01	13.49
ABR-EXP3	OFA	Learning Rate: 1.0	12.94
ABR-EXP4	OFA	Learning Rate: 0.01	13.49
KNR-EXP1	AFO	$n\_neighbors = 3$	12.01
KNR-EXP2	AFO	$n\_neighbors = 5$	11.89
KNR-EXP3	OFA	$n\_neighbors = 3$	11.75
KNR-EXP4	OFA	$n\_neighbors = 5$	11.68
GBR+PSO (Proposed Method)	AFO	Parameter range based on Table II	1.205

increases, indicating that the model is learning effectively without overfitting. This behavior is critical for ensuring that the model generalizes well to unseen data, which is essential for real-world drilling applications.

To further validate the model's performance, predict vs. actual plots are highly useful. These plots provide a visual comparison between the predicted and actual ROP values, allowing for a direct assessment of the model's accuracy. For a well-performing model like GBR+PSO, the points in the predict vs. actual plot should closely align with the 45-degree line, indicating minimal deviation between predictions and ground truth. This visualization confirms the model's accuracy and helps identify any systematic errors or biases in the predictions, which can be addressed in subsequent iterations of model refinement.

In summary, using MAE as a performance measure [20], along with learning curves, predict vs. actual plots, and violin plots, creates a strong way to assess and confirm the reliability of ROP prediction models. The GBR+PSO model's exceptional performance, as evidenced by its low MAE, illustrates the value of advanced optimization techniques and thorough model validation. By leveraging these tools, researchers and practitioners can ensure that their models not only achieve high accuracy but also exhibit consistent and reliable behavior across a wide range of drilling conditions, ultimately leading to more efficient and safer drilling operations (see Fig. 5).

VII. CONCLUSION

The research under this study has framed the potential of hybrid optimization techniques in improving the ROP prediction in drilling operations. This study has created a plan to make ROP predictions more accurate by mixing machine learning algorithms with traditional optimization methods, while also dealing with the challenges of drilling in areas that contain hydrocarbons. The results indicated that the methodologies introduced can deliver substantial improvements in drilling efficiency, enabling reductions in operational costs and increased productivity in the oil and gas industry. This work also underlined the importance of data-driven approaches in modern drilling practices. Given the increasing demands for efficiency and cost-effectiveness within the industry, this capability to predict ROP with good accuracy will become relevant. The findings of the research indicate that the developed hybrid GBR+PSO model would be very useful for drilling engineers in driving data-informed decisions powered by real-time and predictive analytics. This development does not just guarantee the operational success but also contributes to achieving those goals the industry has laid out concerning minimum environmental impact with maximum resource extraction. The objective of the research was to assess the overall effect of the suggested methodologies on drilling performance. The implemented hybrid optimization approach, when executed, resulted in a measurable increase in ROP, thereby increasing the overall productivity of the drilling operation.

This study is quite important because it shows that the practical benefits of adopting advanced predictive models are tangible in the field. The positive results associated with this objective are indicative of further research and development in the field of drilling optimization. Since the industry is in continuous evolution, advanced technologies and methodologies must be integrated not only to stay competitive but also to address the challenges posed by increasingly complex drilling environments. Despite these promising results, the study has certain limitations. The model was developed and validated using historical drilling datasets from a limited number of wells, which may not fully capture the variability of diverse geological formations and operational conditions encountered across the broader industry. Building on these findings, future work should focus on real-time field validation and deployment, testing the GBR+PSO model on live drilling rigs with real-time data streaming to evaluate its practical performance, latency, and reliability under actual field conditions, rather than solely on historical datasets. Further research could also explore integration with other optimization algorithms, hybridizing GBR with other metaheuristic algorithms such as Ant Colony Optimization, Grey Wolf Optimizer, or Whale Optimization Algorithm, to benchmark against PSO and identify the most robust optimization strategy. Additionally, future

studies should explore embedding the hybrid GBR+PSO model into autonomous or semi-autonomous drilling control systems to enable fully automated, closed-loop optimization of drilling parameters.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Faris Aiman b Jamaluddin: Writing – original draft, formal analysis, data curation. Marina Yusoff: Writing – review and editing, visualization, validation, supervision, methodology, and conceptualization. Diva Kurnianingtyas – validation, editing. Mohamad Taufik Mohd Salledud-din – data understanding, validation.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no competing interests that could have appeared to influence the work reported in this study.

ACKNOWLEDGMENT

The authors gratefully acknowledge support from the Universiti Teknologi MARA (UiTM), Malaysia, and the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Petronas Research Sdn.Bhd for knowledge transfer and financial support provided to this research project.

REFERENCES

[1] J. Li, L. Stamford, and A. Gallego-Schmid, "Full environmental life cycle costing analysis of repurposing onshore abandoned oil and gas wells for geothermal power generation," Applied Thermal Engineering, p. 130469, 2026.

[2] X. Yu and G. Ren, "A data-driven approach to achieve low-carbon building energy optimization by using BIM technology," International Journal of Advanced Computer Science and Applications, vol. 16, no. 8, 2025.

[3] Y. Chen, P. Chen, X. Dai, J. Gong, and K. Lu, "Prediction of rate of penetration in oil and gas well drilling: A hybrid data-driven approach with multi-stage data preprocessing," *Engineering Applications of Artificial Intelligence*,vol. 163, p. 113085, 2026.

[4] Y. Song, Z. Song, J. Yang, L. Wei, and J. Tang, "Enhancing energy efficiency and sustainability in offshore drilling through real-time multi-objective optimization: Considering lag effects and formation variability," *Reliability Engineering & System Safety*, vol. 261, p. 111138, 2025.

[5] S. Davoodi, M. Al-Shargabi, D. A. Wood, and M. Mehrad, "Advancement of artificial intelligence applications in hydrocarbon well drilling technology: A review," *Applied Soft Computing*, vol. 176, p. 113129, 2025.

[6] Z. W. Cai, R. L. Zhao, Z. M. Huang, Y. P. Qiao, Q. Y. Yue, C. L. Li, et al., "Research on the development trends of measurement while drilling (MWD) technology in oil and gas drilling," *Petroleum Science*, 2025.

[7] M. Y. Amer, S. K. Salem, M. S. Farahat, and A. M. Salem, "Reducing drilling cost of geothermal wells by optimizing drilling operations: Cost effective study," *Journal of Unconventional Resources*, vol. 7, p. 100196, 2025.

[8] Y. Zhang, L. Yu, L. Yang, Z. Hu, and Y. Liu, "Data-driven framework for predicting rate of penetration in deepwater granitic formations: A marine engineering geology perspective with comprehensive model interpretability," *Journal of Engineering Geology*, vol. 351, p. 108039, 2025.

[9] P. Chen, "Advancements and future outlook of safety monitoring, inspection and assessment technologies for oil and gas pipeline networks," *Journal of Pipeline Science and Engineering*, p. 100267, 2025.

[10] M. Naderi, A. Bagheri, and M. N. Khujin, "Hybrid LSTM-Desirability Concept for Drilling Rate of Penetration (ROP) Prediction and Optimization," Array, p. 100900, 2026.

[11] M. Bizhani and E. Kuru, "Towards drilling rate of penetration prediction: Bayesian neural networks for uncertainty quantification," *Journal of Petroleum Science and Engineering*, vol. 219, p. 111068, 2022.

[12] X. Yang, M. Wu, C. Lu, W. Li, L. Chen, and S. Du, "Prediction of rate of penetration based on drilling conditions identification for drilling process," *Neurocomputing*, vol. 579, p. 127439, 2024.

[13] A. Sauki, P. N. F. M. Khamaruddin, S. Irawan, I. Kinif, S. Ridha, S. A. Ali, and M. A. Ali, "Development of a modified Bourgoyne and Young model for predicting drilling rate," Journal of Petroleum Science and Engineering, vol. 205, p. 108994, 2021.

[14] R. Alharbi, N. Alageel, M. Alsayil, R. Alharbi, and A. A. Alhakamy, "Prediction of oil production through linear regression model and big data tools," International Journal of Advanced Computer Science and Applications, vol. 13, no. 12, 2022.

[15] A. Sharma, T. Burak, R. Nygaard, E. Hoel, T. Kristiansen, and M. Welmer, "Hybrid ROP modeling: Combining analytical and data-driven approaches for drilling," Geoenergy Science and Engineering, vol. 251, p. 213877, 2025.

[16] Yehia, T., Gasser, M., Ebaid, H., Meehan, N., & Okoroafor, E. R. (2024). Comparative analysis of machine learning techniques for predicting drilling rate of penetration (ROP) in geothermal wells: A case study of FORGE site. Geothermics, 121, 103028.

[17] R. Zhang, Z. Zhu, Z. Yan, T. Pan, X. Song, G. Li, et al., "EHTGNN: An Explainable Hybrid Temporal Graph Neural Network for robust rate of penetration prediction in drilling operations," Geoenergy Science and Engineering, vol. 259, p. 214347, 2026.

[18] R. H. Allawi, W. J. Al-Mudhafar, M. A. Abbas, and D. A. Wood, "Leveraging boosting machine learning for drilling rate of penetration (ROP) prediction based on drilling and petrophysical parameters," *Artificial Intelligence in Geosciences*, vol. 6, no. 1, p. 100121, 2025.

[19] A. T. Tunkiel, D. Sui, and T. Wiktorski, "Reference dataset for rate of penetration benchmarking," Journal of Petroleum Science and Engineering, vol. 196, p. 108069, 2021.

[20] Q. Yuan, M. He, Z. Chen, M. Liu, and X. Chen, "A real-time prediction method for rate of penetration sequence in offshore deep wells drilling based on attention mechanism-enhanced BiLSTM model," *Ocean Engineering*, vol. 325, p. 120820, 2025.

[21] D. Rezki, L.-H. Mouss, A. Baaziz, and T. Bentrcia, "Adaptive prediction of Rate of Penetration while oil-well drilling: A Hoeffding tree based approach," Engineering Applications of Artificial Intelligence, vol. 143, p. 111465, 2025.

[22] C. Gan, Y. Wang, W.-H. Cao, K.-Z. Liu, and M. Wu, "Real-time formation drillability sensing-based hybrid online prediction method for the rate of penetration (ROP) and its industrial application for drilling processes," *Control Engineering Practice*, vol. 164, p. 106487, 2025.

[23] L. F. F. M. Barbosa, A. Nascimento, M. H. Mathias, and J. A. de Carvalho, "Machine learning methods applied to drilling rate of penetration prediction and optimization - A review," Journal of Petroleum Science and Engineering, vol. 183, p. 106332, 2019, doi: 10.1016/j.petrol.2019.106332.

[24] M. Yusoff, D. Ehsan, M. Y. Sharif, and M. T. M. Sallehud-din, "Topology approach for crude oil price forecasting of particle swarm optimization and long short-term memory," International Journal of Advanced Computer Science & Applications, vol. 15, no. 1, 2024.

[25] M. Daviran, A. Maghsoudi, and R. Ghezelbash, "Optimized AI-MPM: Application of PSO for tuning the hyperparameters of SVM and RF algorithms," Computers & Geosciences, vol. 195, p. 105785, 2025.