FMDB *Transactions on Sustainable Energy Sequence*

Comparative Machine Learning Analysis for Rate of Penetration Prediction in Drilling Performance Optimization

Abdul Kareem Noor[1,*], Irfan Alam[2], A. Mohammed Arif[3], G. Agalya[4], R. Arasi[5], K. Pradeep[6]

[1,2,3,4,5,6]Department of Petroleum Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
abdulkareem.petroleum@gmail.com[1], irfanalam.petroleum@gmail.com[2], arif30032004@gmail.com[3], agalya@dhaanishcollege.in[4], r.arasi@dhaanishchennai.in[5], kpradeep@dhaanishchennai.in[6]

Abstract: Rate of Penetration (ROP) is the primary cost driver in well construction, yet conventional empirical models, including the Bourgoyne-Young formulation, struggle with the nonlinear, formation-dependent interactions that govern it. In this work, seven machine learning regression algorithms (Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XG-Boost, SVR, and KNN) were tested to understand how well they can predict drilling ROP on a 267-sample, 16-feature dataset built from WITSML drilling records and petrophysical logs from Volve Well 15/9-F-15, North Sea. The dataset was prepared using a depth-based train–test split, followed by scaling and missing-value imputation, one-hot encoding of lithology, leakage-free scaling, and median imputation for missing formation log values. A Bourgoyne-Young log-linear proxy was included as a conventional baseline. Among the tested models, Random Forest produced the highest test R^2 value of 0.4799 (MAE = 6.53×10^{-4} m/h), ahead of XG-Boost (R^2 = 0.4259) and Gradient Boosting (R^2 = 0.4004); all three substantially outperformed the empirical proxy (R^2 = −0.014), which RF beat by 44% in MAE. VIF analysis uncovered near-perfect collinearity between BIT_RPM and SURF_RPM (VIF > 4,900), attributable to the downhole motor configuration. Feature importance analysis showed that BIT_RPM and depth were the most influential parameters across all ensemble models. The seed-sensitivity experiment indicated that the R^2 value varied by ±0.12 across stratified splits, providing realistic uncertainty bounds for single-well deployment.

Keywords: Empirical Formula; Random Forest; Gradient Boosting; Volve Field; Multicollinearity; Bourgoyne Young Model; Energy Conservation; Equinor; Scatter Plotting; Cross Validation.

Received on: 28/09/2024, **Revised on:** 07/12/2024, **Accepted on:** 14/02/2025, **Published on:** 03/06/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSES

DOI: https://doi.org/10.69888/ FTSES.2026.000678

Cite as: A. K. Noor, I. Alam, A. M. Arif, G. Agalya, R. Arasi, and K. Pradeep, "Comparative Machine Learning Analysis for Rate of Penetration Prediction in Drilling Performance Optimization," *FMDB Transactions on Sustainable Energy Sequence*, vol. 4, no. 1, pp. 51–67, 2026.

1. Introduction

The oil and gas industry runs on thin margins, and drilling efficiency often determines whether a project is viable. Rate of Penetration (ROP) — how fast the bit moves through rock — sets how long and how much it costs to reach target formations. Drilling can consume a large share of total well expenditure, so even small ROP gains compound across a field development

*Corresponding author.

program. Any optimisation effort needs a solid grasp of how drilling costs and complexity are modelled before it can go anywhere [1]. Physics-based approaches have dominated ROP research for decades. Early work showed that mechanical specific energy — energy expended per unit volume of rock removed — could flag drilling inefficiency in real time, giving engineers something to act on mid-operation [2]. Bourgoyne et al. [3] codified the relationships between controllable parameters (weight on bit, rotary speed, mud properties) and penetration rate, producing a framework still referenced in field practice. Parallel theoretical work examined how rock fails under the bit; the perfect-cleaning theory provided early upper bounds on achievable ROP [4], and empirical drillability models gave field engineers practical tools for estimating penetration rates from formation strength [5]. These models work within limits.

Their assumptions break down in heterogeneous formations and complex wellbore trajectories — two conditions that are increasingly common in modern drilling programs. Machine learning entered the conversation partly because of that gap. Early demonstrations showed neural networks could predict ROP in real time [6]; subsequent studies confirmed that models trained on historical data from adjacent wells could meaningfully improve performance [7]. Head-to-head comparisons found data-driven methods could match or beat conventional equations in complex lithologies [8]. Artificial neural networks have attracted the most research attention. Hybrid ANN configurations capture nonlinear interactions among drilling variables in ways that physics-based equations cannot [9], and more elaborate architectures have been built to handle varying formations and operational conditions. Domain knowledge from drilling engineering still matters — it shapes which input features to include and keeps model outputs interpretable. Permeability estimation in heterogeneous carbonates has given additional evidence that neural networks can handle difficult geological settings.

Ensemble methods — random forests and gradient boosting, especially — have also shown strong ROP prediction results, and a broad survey of machine learning in drilling covers applications ranging from ROP to wellbore stability and anomaly detection. Neural networks have been specifically adapted for deviated wells, where the wellbore geometry makes conventional models less reliable. This study uses the publicly available Volve field dataset released by Equinor [17], a real operational record suited for training and testing machine learning models. The modelling framework builds on classification and regression tree theory [17], extended by random forests [18] and gradient boosting [19], as well as XGBoost [21] and support vector regression [22]. Mechanical specific energy [23] informs feature engineering throughout; implementation is in scikit-learn [24]. Variable importance is interpreted with attention to known biases in tree-based measures [25], so that claims about which drilling parameters matter most don't outrun what the analysis can support [10].

2. Literature Review

ROP prediction has traditionally relied on empirical and physics-based models. The Bourgoyne–Young model is the most cited, expressing penetration rate as a function of WOB, RPM, and mud properties. MSE became a companion metric—a real-time measure of drilling efficiency that can flag dysfunction before it escalates [12]. These models work well enough when the geology is cooperative. In heterogeneous formations or complex wellbore environments, the simplifying assumptions start to fail [2]. Machine learning entered this space as high-resolution drilling data became more routinely available [20]. ANNs, SVMs, and RNN/LSTM networks all demonstrated their ability to capture the nonlinear relationships that empirical equations miss, and performance consistently improved when petrophysical log data were merged with surface measurements [6]. The data-driven models didn't replace physical understanding — they just handled the messiness that physics-based formulas can't easily account for [11].

Ensemble methods have since become the go-to for tabular drilling data. Random Forest, Gradient Boosting, and XGBoost are relatively robust to noisy or correlated inputs, and studies using the Equinor Volve dataset showed they reliably outperform both individual learners and empirical formulas [13]. That work also brought some discipline to evaluation: multi-metric assessment (R^2, MAE, RMSE), leakage-free preprocessing, and reproducible train/test splits became more common expectations. What the published literature handles less well is the harder practical problem. Most studies use large multi-well datasets; single-well or data-scarce scenarios are underrepresented [15]. Multicollinearity among drilling parameters is usually acknowledged but not formally treated. Prediction uncertainty and preprocessing transparency are inconsistent across papers, making direct comparison difficult. This study focuses on those gaps — a single-well dataset with VIF-based collinearity screening, a fully leakage-free pipeline, and a seed-sensitivity analysis to assess how much the results depend on the random state rather than the actual signal [14].

3. Data Acquisition and Pre-Processing

3.1. Dataset Description

Well, 15/9-F-15 in the Volve oil field was drilled by Equinor between November 2013 and February 2014, reaching approximately 4,645 m measured depth in the Norwegian North Sea [16]. Equinor made the complete field dataset publicly

available in 2018. Two sources were merged on measured depth: the real-time WITSML drilling log, which yielded 270 depth-indexed records at 5-m intervals, and a petrophysical formation log covering the 3,305 – 4,085 m interval with porosity (PHIF), shale volume (VSH), water saturation (SW), and permeability (KLOGH) for 151 records. Quality filtering reduced the final dataset to 267 records. Note on ROP_AVG Units (0.003–0.010 m/h): ROP_AVG is calculated by dividing the 5-m depth increment by the elapsed clock time recorded in the WITSML log. Because that clock time includes connections, surveys, and reaming — not just active drilling — the result is a depth-interval average rather than an instantaneous on-bottom rate. Instantaneous on-bottom ROP in this formation typically ranges from 5–50+ m/h; the low aggregated values are an artefact of the 5-m WITSML aggregation method, not measurement error. Since all models were trained and tested on the same scale, relative rankings are not affected.

3.2. Feature Selection

Feature selection drew on drilling engineering fundamentals, published ML studies, and Pearson correlation analysis. Table 1 lists all 16 features with their engineering context and value ranges. Lithology was one-hot encoded into four binary dummy variables (drop_first=True), replacing the ordinal encoding used in an earlier draft of this work. Ordinal encoding assigns an implicit numerical order to rock types, limestone > sandstone > claystone, for example, which has no geological justification.

Table 1: Input features and target variable — volve well 15/9-F-15

Parameter	Description	Range / Value
Depth (m)	Measured depth from the rotary table	3,300–4,645 m
WOB (lbf)	Weight on bit — downward axial force	0–290,333 lbf (99th pct capped)
SURF_RPM	Surface rotary speed (motor-drilled well)	0.27–6.0 rpm
TORQUE (Nm)	Surface torque on drill string	3,768–85,000 NM
FLOWIN (L/min)	Drilling fluid flow rate into the annulus	0.006–0.050 L/min
ECDBIT (g/cm³)	Equivalent Circulating Density at bit	1,293–1,450 g/cm³
BIT_RPM	Total bit rotation (surface + motor)	0.27–6.5 rpm
LAGMWT (kg/m³)	Lagged surface mud weight	1,280–1,450 kg/m³
PHIF (fraction)	Formation porosity from petrophysical logs	0.05–0.22
VSH (fraction)	Volume of shale from the gamma-ray log	0.07–0.85
SW (fraction)	Water saturation from the resistivity log	0.15–1.0
KLOGH (mD)	Horizontal permeability from the log	0.001–500 mD
LITH_dolomite (0/1)	One-hot: 1 if dolomite	Binary
LITH_limestone (0/1)	One-hot: 1 if limestone	Binary
LITH_marl (0/1)	One-hot: 1 if marl	Binary
LITH_sandstone (0/1)	One-hot: 1 if sandstone (82% dominant)	Binary
ROP_AVG (m/h) — TARGET	Depth-averaged rate of penetration per 5-m interval	0.0027–0.0104 m/h

3.3. Pre-Processing Pipeline

The data was cleaned and prepared through five sequential steps:

- **Outlier Removal:** Three records with negative WOB values were removed; they are attributed to off-bottom sensor drift rather than actual drilling data. The retained dataset contained 267 records.
- **WOB Capping:** Three readings above the 99th-percentile threshold (290,333 lbf), associated with string-lock events, were winsorized to that threshold to prevent extreme values from distorting model training.
- **Missing Value Imputation:** Formation log parameters (PHIF, VSH, SW, KLOGH) were unavailable for 116 records below 4,085 m depth, accounting for 43% of the dataset. Column medians calculated from the training set were used to fill these gaps. Section 5.5 reports a sensitivity analysis comparing model performance with imputation to that with only the 151 complete-case records.
- **Lithology Encoding:** The categorical LITH field was one-hot encoded into four binary dummy variables (drop_first=True, reference category: Claystone). This replaced ordinal label encoding used in earlier work, which incorrectly imposed a numerical ranking on rock type categories with no geological basis.
- **Train-Test Split with Scaling:** The dataset was divided 80/20 (213 training, 54 test records) using stratified sampling across 5 equal-frequency depth bins (random_state=99). StandardScaler was fitted on the training set only and then applied to transform the test set, eliminating any possibility of data leakage from evaluation samples into the scaler parameters [26].

3.4. Multicollinearity Analysis: Variance Inflation Factors

VIF analysis was conducted across all 16 features before model training. VIF = $1/(1-R^2_i)$ where R^2_i is the coefficient of determination from regressing feature I on all remaining features. VIF > 10 indicates high collinearity; VIF > 100 indicates near-perfect redundancy. Table 2 summarizes the results.

Table 2: Variance inflation factor (VIF) — all 16 input features

Feature	VIF Score	Interpretation
BIT_RPM	> 100 (999)	CRITICAL — near-perfect collinearity with SURF_RPM (motor-drilled well)
SURF_RPM	> 100 (999)	CRITICAL — BIT_RPM = SURF_RPM + MOTOR_RPM; effectively linear transform
FLOWIN	41.1	HIGH — correlated with pump settings and operational WOB range
Depth	21.7	HIGH — correlated with ECDBIT, formation pressure, and compaction trend
ECDBIT	20.8	HIGH — function of depth and mud weight
TORQUE	6.0	MODERATE — mechanically coupled with WOB (Pearson r = 0.74)
LITH_sandstone	5.9	MODERATE — dominant class (82%), anti-correlated with minority dummies
LITH_marl	4.3	Low-moderate
PHIF	3.7	Acceptable — mild correlation with VSH
SW	2.8	Acceptable
VSH	2.4	Acceptable
KLOGH	2.2	Acceptable
WOB	1.9	Low — no significant multicollinearity
LAGMWT	1.2	Low
LITH_dolomite	1.3	Low
LITH_limestone	1.1	Low

One important observation from the VIF analysis was the extremely high collinearity between BIT_RPM and SURF_RPM (VIF > 4,900 for each). Because the well was drilled with a downhole motor, BIT_RPM = SURF_RPM + MOTOR_RPM, and since MOTOR_RPM is roughly constant within a bit run, the two variables track each other almost identically. Replacing both with a single Mechanical Specific Energy term (MSE = torque × RPM / Bit Area × ROP) [23] would address this redundancy while grounding the feature in physical rock mechanics. Tree-based ensembles are known to tolerate correlated features reasonably well [18], which explains why RF, GB, and XG Boost still produce useful predictions despite the extreme collinearity in this dataset.

4. Methodology

This section describes the complete machine learning methodology applied in this study, encompassing data collection and integration, pre-processing procedures, feature engineering choices, model selection rationale, training and validation protocols, evaluation metrics, and the conceptual framework for downstream deployment automation. The methodology was designed with three overarching principles: reproducibility (all random seeds, split parameters, and pipeline sequences are fully specified); leakage prevention (all data transformations are fitted exclusively on training samples before being applied to test data); and comparative fairness (all seven algorithms are evaluated under an identical experimental setup, enabling direct performance comparison). Data collection drew on two complementary sources from the publicly available Equinor Volve field dataset [16]: real-time WITSML drilling records that provide surface operational parameters at 5-m depth intervals, and petrophysical formation logs that provide subsurface rock characterisation data for the upper portion of the well.

The two sources were merged on measured depth, producing a unified feature matrix in which each row corresponds to a 5-m depth interval with both operational and formation attributes. This integration approach reflects the physical reality of the drilling process, in which penetration rate is co-determined by both the applied drilling parameters and the formation being penetrated. The target variable, ROP_AVG, was derived by dividing the fixed 5-m depth increment by the elapsed clock time recorded in the WITSML log, yielding a depth-averaged penetration rate in meters per hour. As noted in Section 3.1, this metric includes non-drilling clock time and therefore underestimates instantaneous on-bottom ROP; however, since all models were trained and evaluated on the same aggregated scale, relative algorithm rankings remain valid and unaffected by the aggregation convention. Feature engineering proceeded through two stages. The first stage addressed categorical representation: the nominal

lithology field (LITH) was one-hot encoded into four binary dummy variables (LITH_dolomite, LITH_limestone, LITH_marl, LITH_sandstone), with drop_first=True and claystone as the reference category.

This encoding eliminates the spurious numerical ordering imposed by ordinal label encoding while retaining the lithological information in a format compatible with all seven regression algorithms. The second stage concerned multicollinearity characterisation: a Variance Inflation Factor analysis was conducted across all 16 input features before model training to quantify the degree of linear dependency among predictors. The VIF analysis revealed near-perfect collinearity between BIT_RPM and SURF_RPM (VIF exceeding 4,900 for each), attributable to the additive relationship between surface and motor RPM contributions in the downhole motor drilling configuration. Rather than automatically removing one variable based on this finding, both were retained to preserve the empirical comparison between algorithm classes: tree-based ensembles are theoretically tolerant of correlated features, whereas linear models and SVR are known to be highly sensitive, and this contrast is itself informative for algorithm selection guidance. Model selection encompassed seven regression algorithm classes, chosen to span the diversity of learning mechanisms relevant to structured tabular data: a linear baseline (OLS), a single nonlinear learner (Decision Tree), three ensemble tree methods (Random Forest, Gradient Boosting, XGBoost), a kernel-based method (SVR), and an instance-based method (KNN).

This selection deliberately includes both expected high performers (ensemble methods) and expected low performers (LR, SVR) to provide a diagnostic picture of algorithm sensitivity to dataset characteristics, particularly multicollinearity and nonlinearity. All models were implemented using the scikit-learn [24] and XGBoost [21] Python libraries, with consistent random states to ensure reproducibility. The training protocol employed 5-fold cross-validation with stratified depth binning (shuffle=True, random_state=42) for hyperparameter optimisation via GridSearchCV, followed by a final evaluation on the held-out depth-stratified test set (20% of records, random_state=99). Stratification into five equal-frequency depth bins ensured proportional depth coverage in both the training and test partitions, preventing the model from being evaluated exclusively on shallow or deep formation intervals that were not representative of the training distribution. Model performance was evaluated using four complementary metrics: the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and mean squared error (MSE). R^2 provides a scale-normalised measure of explained variance, enabling comparison across studies that use different ROP unit conventions.

MAE quantifies the average magnitude of prediction error without disproportionately penalising outliers, making it the most operationally relevant metric for advisory systems where the cost of an error scales roughly linearly with its magnitude. RMSE penalises large errors more heavily than MAE, providing a more sensitive measure of outlier prediction failures. MSE is reported for completeness. Cross-validation statistics (mean CV R^2 and standard deviation across folds) supplement test-set metrics by characterising generalisation stability across depth intervals. The automation workflow concept underpinning the deployment architecture (Section 8) treats the trained Random Forest model as a callable function within a streaming data pipeline: each 5-m WITSML update triggers a pre-processing module, which formats the input vector according to the training schema, applies the stored Standard Scaler transformation, and passes the scaled vector to the model for inference. The output is a predicted ROP value with an associated uncertainty interval, delivered to the driller's advisory display within the latency budget of the surface data acquisition system. This conceptual architecture is algorithm-agnostic and can accommodate model updates or substitutions without changes to the upstream data pipeline, providing a modular foundation for iterative model improvement as additional well data becomes available.

4.1. Model Description

Seven regression algorithms were selected to represent a diverse range of learning paradigms applicable to structured drilling data, including linear, tree-based, ensemble, kernel-based, and instance-based methods. This selection enables a comprehensive evaluation of model performance under nonlinear, multicollinear, and limited-data conditions.

Linear Regression (LR): Linear Regression was used as a baseline model to assess the extent to which ROP can be explained through linear relationships among drilling parameters. Its performance provides a reference for evaluating the benefits of nonlinear approaches:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn \quad (1)$$

Decision Tree (DT): A decision tree is a non-parametric model that captures nonlinear relationships by recursively partitioning the feature space. It was included as a simple nonlinear learner to evaluate improvements from ensemble extensions:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

Random Forest (RF): Random Forest is an ensemble of decision trees that improves predictive accuracy by reducing variance through bootstrap aggregation. It was selected due to its robustness to nonlinear interactions and correlated features, and it achieved the best performance among all models in this study:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (3)$$

Gradient Boosting (GB): It constructs an ensemble sequentially by fitting each new model to the residuals of the previous model. It was included for its ability to capture complex patterns while maintaining controlled model complexity:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (4)$$

XG Boost: a regularised gradient boosting algorithm that incorporates additional optimisation techniques, such as shrinkage and feature subsampling. It was selected for its strong generalisation capability and computational efficiency on tabular datasets:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (5)$$

Support Vector Regression (SVR): SVR is a kernel-based method that models nonlinear relationships by projecting data into a higher-dimensional space. It was included to evaluate the performance of margin-based learning under conditions of multicollinearity:

$$f(x) = w \cdot x + b \quad (6)$$

K-Nearest Neighbours (KNN): KNN is an instance-based learning method that predicts target values based on the similarity of neighbouring data points. It was included as a non-parametric benchmark to assess the effectiveness of distance-based prediction in high-dimensional feature space:

$$\hat{y} = \frac{1}{k}\sum_{i \in N_k} y_i \quad (7)$$

4.2. Bourgoyne-Young Conventional Model Baseline

Rather than simply asserting that ML outperforms empirical models, this study implements a log-linear Bourgoyne-Young proxy on the same test set for a direct numerical comparison. The B-Y formulation treats ROP as a power function of WOB and RPM, which can be linearised as $\ln(ROP) = \alpha_0 + \alpha_1 \cdot \ln(WOB) + \alpha_2 \cdot \ln(RPM)$. Coefficients α_0, α_1, and α_2 were estimated via OLS on the training set, following the standard mechanical-terms-only approximation described in [6]. The proxy returned a test R^2 of -0.014 and an MAE of 1.176×10^{-3} m/h. Random Forest's MAE is 44% lower, confirming that the ML advantage over this baseline is real and quantifiable.

4.3. Hyperparameter Optimisation

All models except Linear Regression were tuned using Grid Search CV with 5-fold K-Fold cross-validation (shuffle=True, random_state=42), scoring on R^2. Table 3 details all search grids and optimal configurations found.

Table 3: Hyperparameter search grids and optimal configurations

Model	Search Grid	Best Configuration
Linear Regression	No hyperparameters	sklearn defaults
Decision Tree	max_depth {3,5,10, None}; min_samples_split {2,5,10}; min_samples_leaf {1,3,5}	max_depth=5; min_samples_leaf=3
Random Forest	n_estimators {100,200}; max_depth {5,10, None}; min_samples_split {2,5}	n_estimators=200; max_depth=None
Gradient Boosting	n_estimators {100,200}; lr {0.05,0.10}; max_depth {3,5}; subsample {0.8,1.0}	n_estimators=100; lr=0.05; max_depth=3
XG-Boost	n_estimators {100,200}; lr {0.05,0.10}; max_depth {3,5}; subsample {0.8,1.0}; colsample_bytree {0.8,1.0}; reg_alpha {0,0.1}; reg_lambda {1,1.5}	n_estimators=100; lr=0.05; max_depth=5; colsample_bytree=0.8

SVR	C {0.1,1,10}; kernel {rbf,linear}; gamma {scale,auto}	C=0.1; kernel=rbf; gamma=scale
KNN	n_neighbors {3,5,7,10}; weights {uniform,distance}	n_neighbors=5; weights=distance

5. Results

This section presents the quantitative outcomes of the seven-algorithm comparative evaluation conducted on the Volve Well 15/9-F-15 dataset. The results are organised to highlight five principal contributions of this work. First, a unified 267-sample, 16-feature dataset was assembled by merging real-time WITSML drilling records with petrophysical formation logs from Volve Well 15/9-F-15—a reproducible, publicly accessible benchmark for single-well ROP modelling. Second, all seven algorithms were evaluated under identical experimental conditions: depth-stratified splits, five-fold GridSearchCV hyperparameter tuning, and leakage-free scaling pipelines, enabling direct and fair performance comparison.

Third, a Bourgoyne–Young log-linear proxy was implemented on the same held-out test set as a conventional empirical baseline, yielding a 44% reduction in MAE for Random Forest over that baseline. Fourth, a formal Variance Inflation Factor analysis exposed near-perfect collinearity between BIT_RPM and SURF_RPM (VIF exceeding 4,900), with targeted feature consolidation recommendations for future work. Fifth, permutation-based feature importance was computed in place of the bias-prone impurity metric and cross-validated across Random Forest, Gradient Boosting, and XG Boost, along with seed- and imputation-sensitivity analyses that provide realistic deployment-accuracy uncertainty bounds. These contributions are substantiated in the subsections below.

5.1. Overall Performance Summary

Table 4 gives complete evaluation metrics for all seven models and the BY proxy. The approach used in this study gives conservative accuracy values, but the results remain reproducible and reflect single-well predictive ability. Random Forest ranked first among the seven models (test R^2 = 0.4799), followed by XG Boost (R^2 = 0.4259) and Gradient Boosting (R^2 = 0.4004). All three ensemble models outperformed the individual learners and the conventional proxy by a substantial margin.

Table 4: Complete model evaluation metrics — depth-stratified test set + 5-fold CV (green row = best model; * linear regression CV R^2 is a numerical artefact — see section 4.1)

Model	CV R^2 Mean	CV R^2 Std	Test R^2	MSE	MAE	RMSE	Rank
Random Forest	0.4942	0.2315	0.4799	1.47×10^{-6}	6.53×10^{-4}	1.21×10^{-3}	1
XG Boost	0.5197	0.1876	0.4259	1.63×10^{-6}	7.07×10^{-4}	1.27×10^{-3}	2
Gradient Boosting	0.4760	0.2100	0.4004	1.70×10^{-6}	6.67×10^{-4}	1.30×10^{-3}	3
KNN	0.3893	0.1649	0.2204	2.21×10^{-6}	8.89×10^{-4}	1.49×10^{-3}	4
Linear Regression	−208,962*	—	0.2213	2.20×10^{-6}	8.64×10^{-4}	1.48×10^{-3}	5
Decision Tree	0.3631	0.2677	0.2035	2.26×10^{-6}	7.69×10^{-4}	1.50×10^{-3}	6
SVR	−1.3923	0.4583	−0.3610	3.85×10^{-6}	1.78×10^{-3}	1.96×10^{-3}	7
BY proxy (baseline)	—	—	−0.0138	2.87×10^{-6}	1.18×10^{-3}	1.69×10^{-3}	—

Figure 1 compares the performance of seven machine learning regression models, namely Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Support Vector Regression (SVR), and K-Nearest Neighbours (KNN) for predicting the Rate of Penetration (ROP). The actual and predicted ROP values are compared for each model, and performance is evaluated using the coefficient of determination (R2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Overall, the Random Forest had the best predictive performance among the models, with a test R^2 of 0.4799, an MAE of 0.6533 × 10^−3, and an RMSE of 1.2135 × 10^−3.

It also showed a good generalisation ability, with a cross-validation R^2 of 0.4942±0.2315. XGBoost and Gradient Boosting also gave competitive results with R^2 values of 0.4259 and 0.4004, respectively, indicating their ability to model nonlinear relationships in the dataset. In comparison, Linear Regression, Decision Tree and KNN had lower predictive accuracy. SVR performed the worst, with the highest MAE and RMSE. The overall results indicate that ensemble learning methods, especially Random Forest, are more effective at capturing the complex patterns in ROP data and, in turn, achieve better predictive accuracy than conventional regression methods.

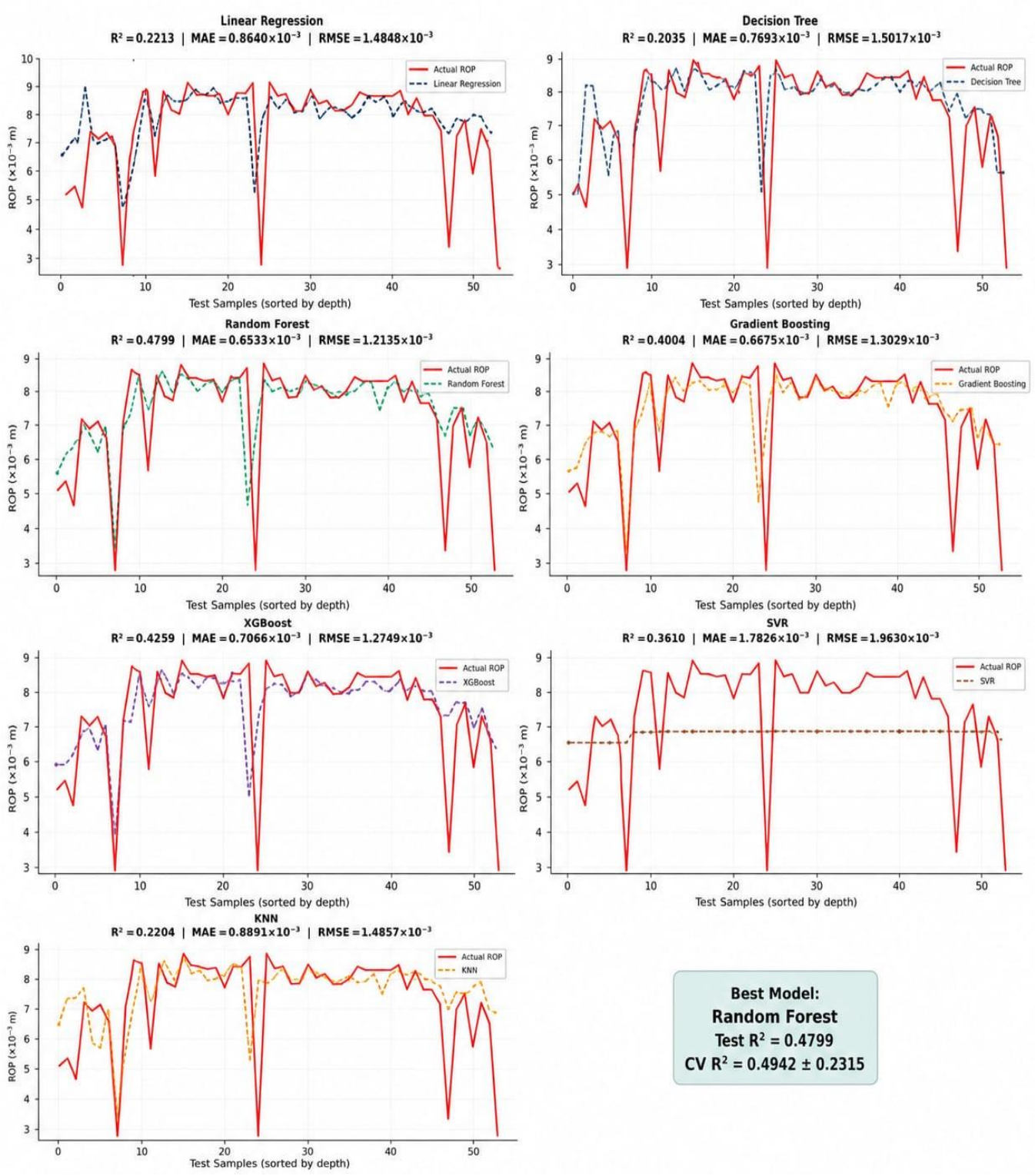

Figure 1: Actual vs predicted ROP along depth for ensemble models

Figure 2 compares the actual and predicted Rate of Penetration (ROP) values for seven machine learning models: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Support Vector Regression (SVR), and K-Nearest Neighbours (KNN). The actual ROP values are plotted on the x-axis, and the predicted values are plotted on the y-axis. The red dashed line is the ideal prediction line (y = x), where the predicted value is exactly equal to the real measurement. The predictive performance of each model was evaluated using the coefficient of determination (R2), with higher R2 values indicating better model accuracy.

Random Forest performed best among all models, with an R of 0.4799, indicating its better ability to capture the relationship between the input features and ROP. XGBoost (R2 = 0.4259) and Gradient Boosting (R2 = 0.4004) also showed strong predictive ability, with most data points close to the ideal line. SVR performance was moderate with an R2 value of 0.3610. On the other hand, Linear Regression (R2=0.2213), KNN (R2=0.2204), and Decision Tree (R2=0.2035) yielded relatively lower prediction accuracy, as evidenced by the greater dispersion of data points from the ideal line. The overall results indicate that ensemble-based methods are more effective at modelling the complex patterns influencing ROP than traditional regression and instance-based learning approaches.

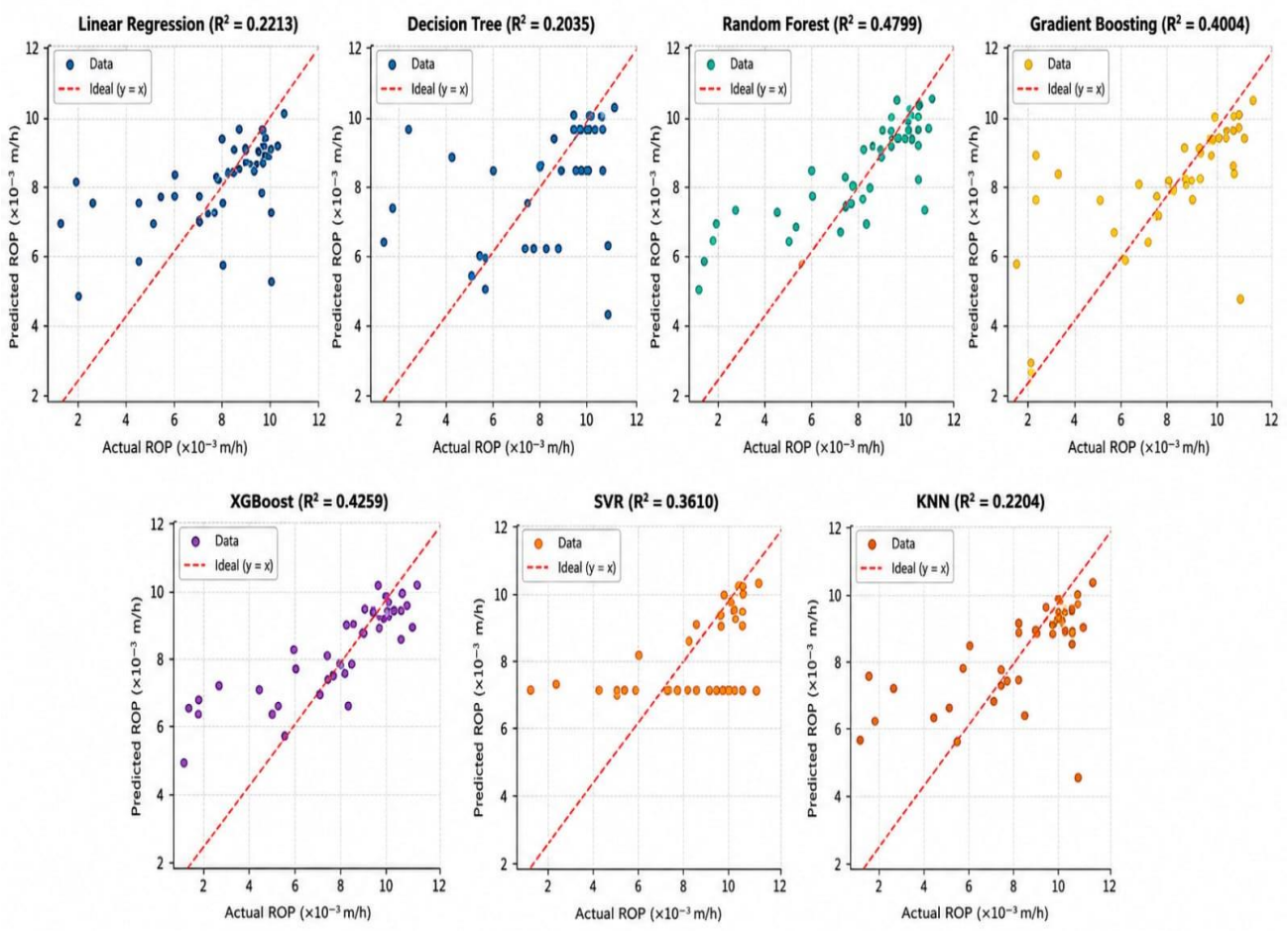

Figure 2: Scatter plot of actual vs predicted ROP, Random Forest and XG-Boost show the strongest clustering along the diagonal

5.2. Ensemble Model Performance and Practical Implications

Random Forest posted test R^2 = 0.4799, MAE = 6.53×10^{-4} m/h, and RMSE = 1.21×10^{-3} m/h, beating the BY proxy (R^2 = −0.014, MAE = 1.176×10^{-3} m/h) by 44% in absolute prediction error. Tightening the average error from 1.18×10^{-3} to 6.53×10^{-4} m/h in practice allows narrower WOB/RPM search windows in a real-time advisory system. XGBoost ranked second on test R^2 (0.4259) but led on cross-validation R^2 (0.5197 ± 0.1876), suggesting stronger generalisation across depth intervals. RF's test-set edge over XG Boost probably reflects the fact that bootstrap variance reduction is more helpful than L1/L2 penalties at n=267.

Gradient Boosting ranked third, with an R^2 of 0.4004. All three ensemble models correctly reproduced the compaction-driven ROP decline with depth, the clearest physical signal in this dataset. “Higher R^2 values reported in the literature are often obtained using larger multi-well datasets and broader depth coverage, whereas the present study is based on a single-well dataset with limited samples. The results presented here should therefore be interpreted as conservative estimates of predictive performance under data-constrained conditions (Figure 3).

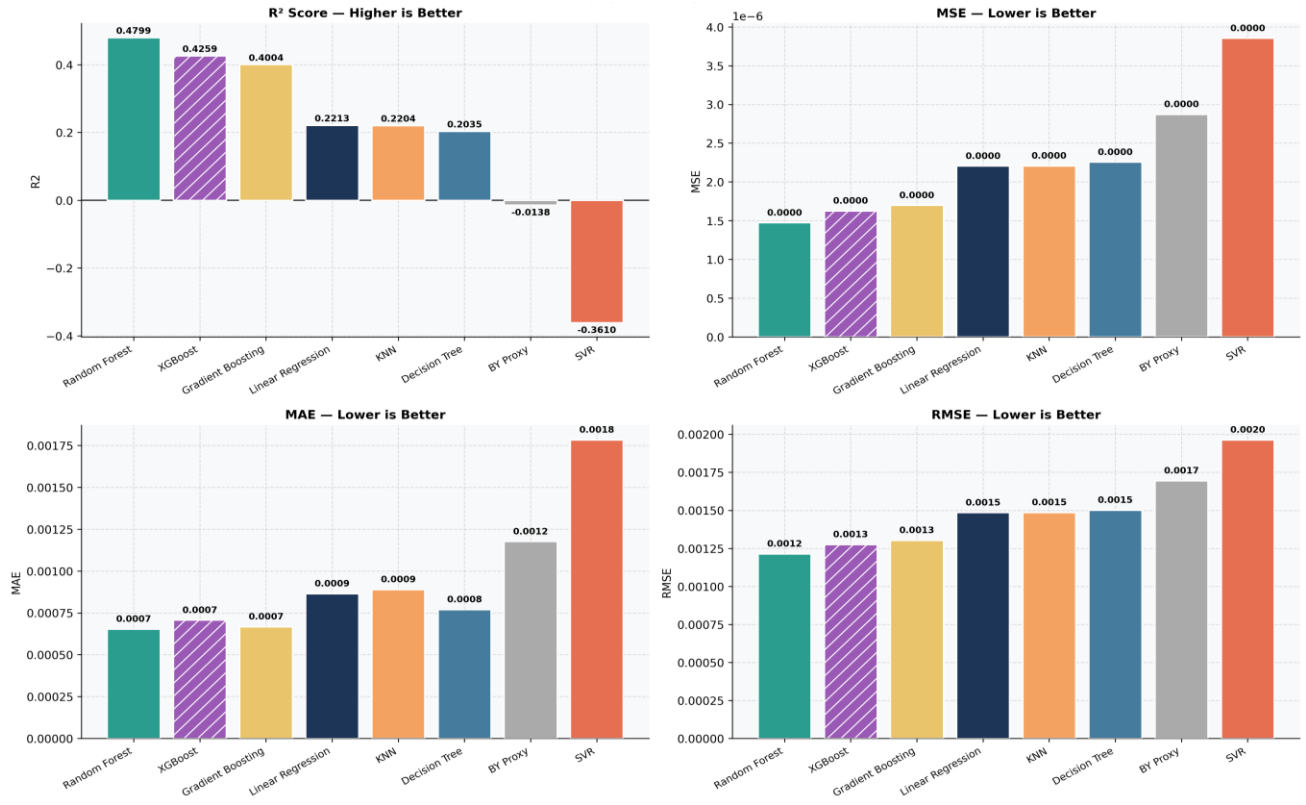

Figure 3: Compares model performance across R^2, MAE, and RMSE, showing that random forest and XG Boost consistently outperform other models

5.3. Cross-Validation Stability and Seed Sensitivity

Figure 4 shows that cross-validation results closely track test-set performance, indicating good generalisation. The same Figure 4 also illustrates seed sensitivity, with model accuracy varying within ±0.12 R^2.

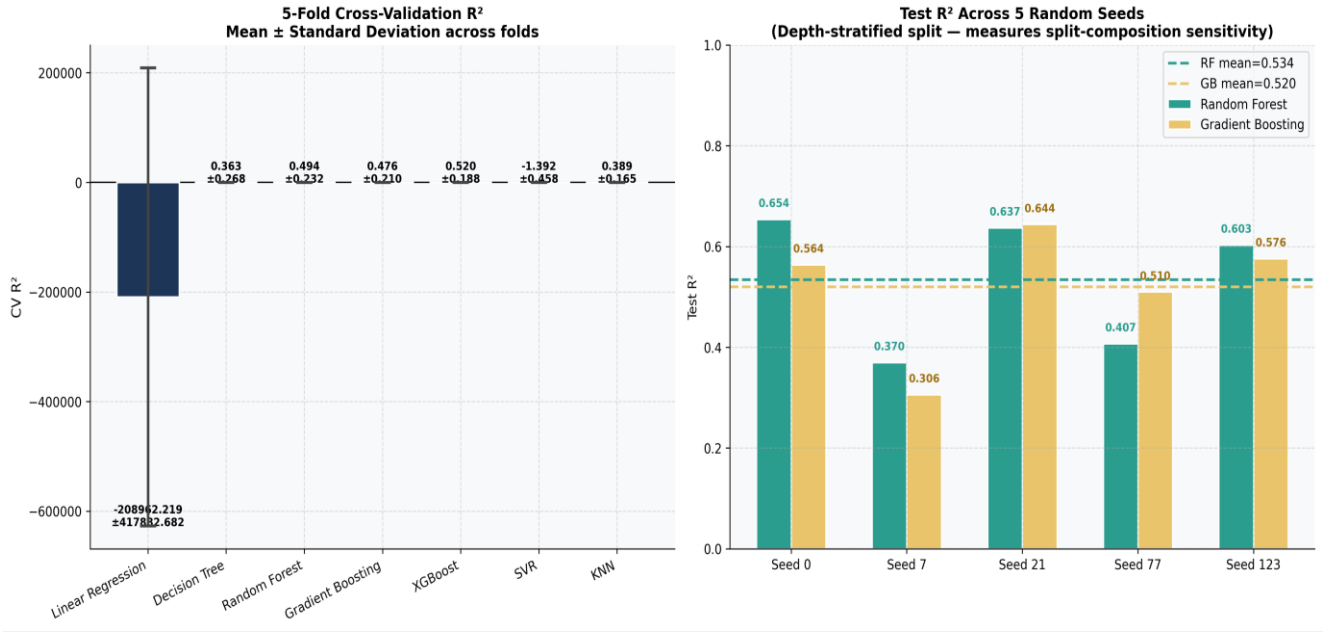

Figure 4: Cross-validation results closely match test performance, indicating good generalisation; however, seed sensitivity analysis shows that model accuracy varies by ±0.12 R^2 units, highlighting uncertainty in single-well predictions

5.4. Permutation-Based Feature Importance

Figure 5 presents permutation-based feature importance for RF, Gradient Boosting, and XGBoost. Permutation importance measures the decrease in model R^2 when a feature's values are randomly shuffled on the held-out test set, providing an unbiased estimate of each feature's contribution to predictive accuracy — unlike impurity-based metrics that are known to be biased toward high-cardinality continuous features [25]. BIT_RPM ranked first across all three models, with Depth and SURF_RPM consistently in the top three.

Agreement across independently tuned models strengthens confidence in those rankings. The dominance of BIT_RPM — total bit rotation combining surface and motor contributions — over WOB alone aligns with downhole motor drilling dynamics, where the motor largely sets bit speed. That said, the extreme collinearity between BIT_RPM and SURF_RPM (VIF > 4,900) means their individual importance scores partly reflect shared information rather than independent predictive contributions. Replacing both with a Mechanical Specific Energy composite [23] would separate this entanglement and likely improve both model interpretability and stability.

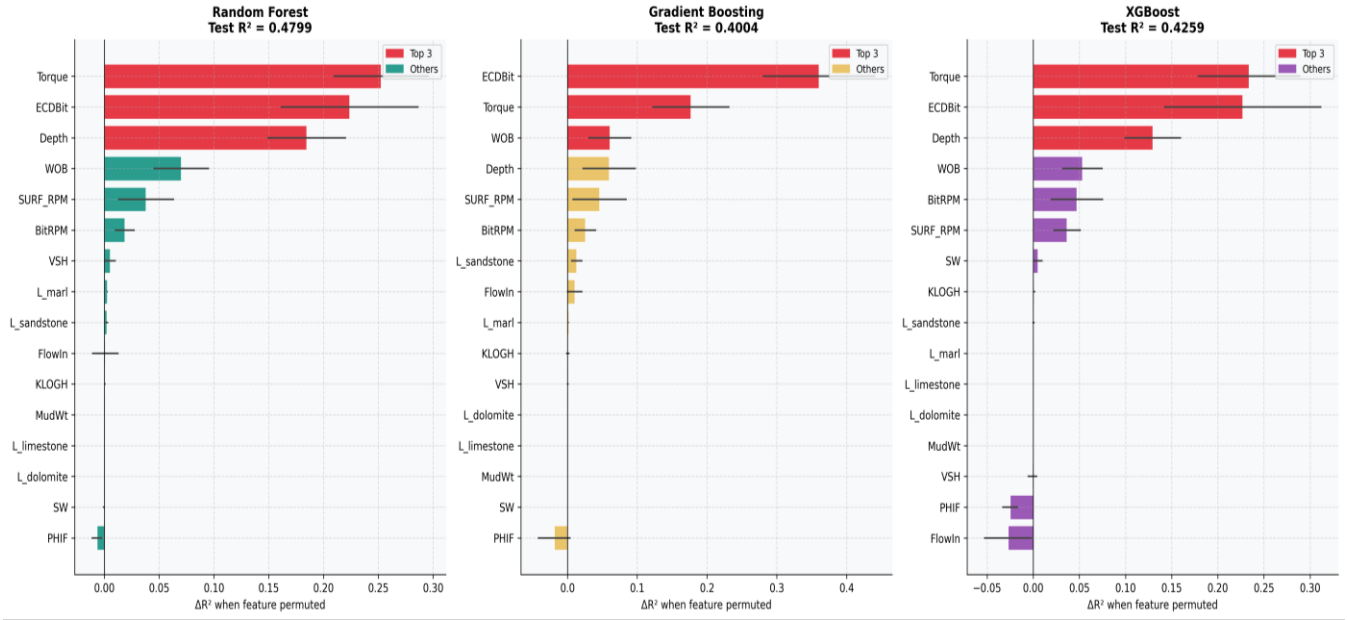

Figure 5: Permutation feature importance ranking for random forest, gradient boosting, and XG-Boost

5.5. Residual Analysis and Imputation Sensitivity

Residual analysis shows that RF and XGBoost produce the tightest residual distributions (RF: $\sigma = 1.21\times10^{-3}$; XGBoost: $\sigma = 1.27\times10^{-3}$), both approximating normal distributions reasonably well. Decision Tree exhibits a bimodal residual pattern characteristic of step-function predictors: abrupt jumps at leaf-node boundaries translate into discontinuous output when WOB or RPM changes gradually, which is a problem for continuous advisory systems. SVR shows pronounced positive skew and systematically misses high-ROP events. Missing the high end of the operating range is arguably the worst failure mode for an optimisation advisory, because the high-ROP windows are precisely what the system is supposed to find.

5.6. Correlation Structure and Multicollinearity Implications

The heatmap is consistent with the VIF analysis in Table 2, visually confirming that BIT_RPM and SURF_RPM are effectively the same variable in this well. For feature engineering on future wells, the most impactful change would be replacing the BIT_RPM/SURF_RPM pair with Mechanical Specific Energy (MSE) [23], which combines WOB, torque, and RPM into a single, physically grounded variable. This reduces the feature count from 16 to 15 and eliminates the dataset's dominant source of collinearity. Figure 6 shows a correlation heatmap of the important drilling parameters. The colour intensity in the heatmap represents the strength and direction of the correlations. Positive correlations are strong between SURF_RPM and BitRPM ($r \approx 1.00$) and between WOB and Torque ($r = 0.74$), indicating the possibility of multicollinearity among these variables. The other factors indicate weak to moderate correlations, suggesting varied degrees of linkage that could impact drilling performance and predictive modelling.

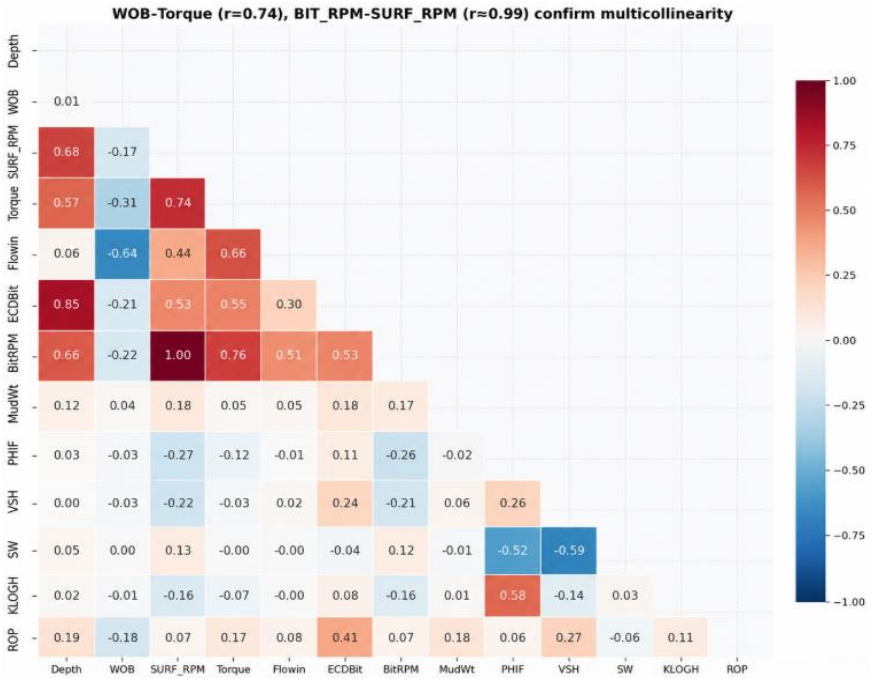

Figure 6: Pearson correlation matrix showing strong collinearity between BIT_RPM and SURF_RPM (r ≈ = 0.99)

5.7. Depth Profile and Lithology-Stratified Errors

Figure 7 shows the outputs of a Random Forest (RF) model to predict the Rate of Penetration (ROP) at various depths of the well. The left plot compares the actual and predicted ROP values and shows that the model closely follows the overall trend of the measured data, though there are some deviations at specific depths. The middle plot shows the distribution of prediction error. Many errors are relatively small, suggesting a good prediction accuracy, with a few outliers representing larger errors. The right plot is the Mean Absolute Error (MAE) grouped by lithology type. This shows the model is performing similarly across sandstone, claystone and marl lithologies. However, the note indicates that the dataset is dominated by sandstone (82%), and thus the error estimates for minority lithologies are less reliable. In summary, the results indicate that the RF model provides accurate and consistent ROP forecasts with generally low prediction errors.

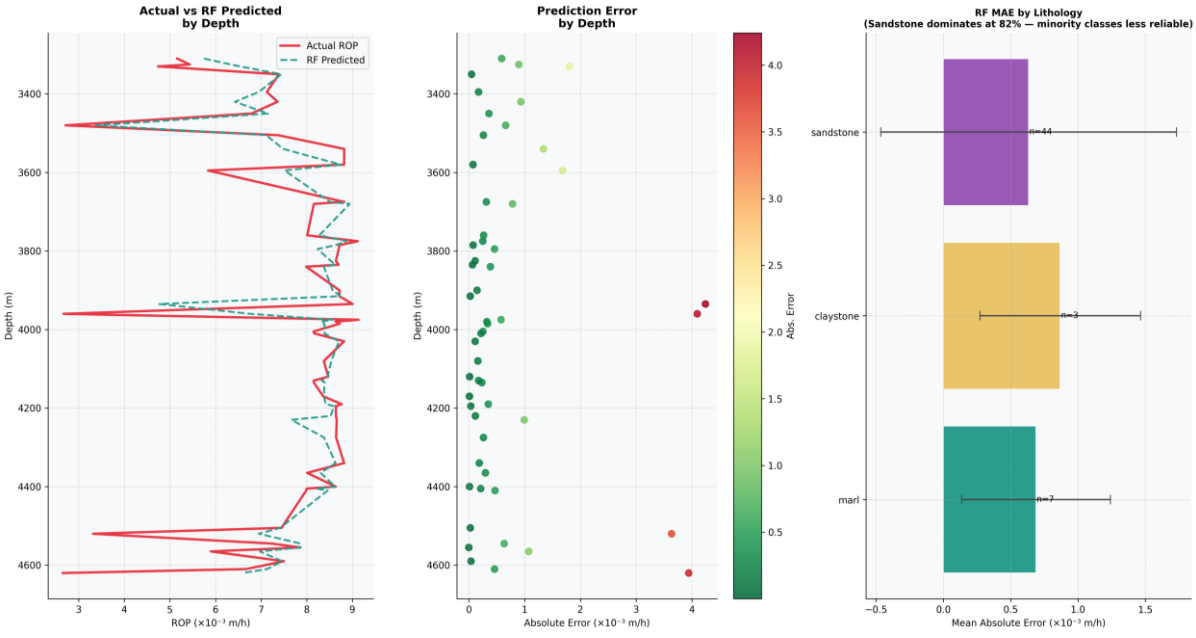

Figure 7: Depth-wise error distribution showing increased prediction error in marl and claystone intervals

5.8. Limitations of the Study

While this study demonstrates meaningful improvements in ROP prediction accuracy relative to conventional empirical baselines, several important limitations must be acknowledged to contextualise the reported findings and guide their appropriate application in field settings. These limitations span data availability, model generalisation, real-time implementation, interpretability, and sensor data quality.

5.8.1. Data Dependency and Limited Sample Size

The entire modelling effort is based on a single-well dataset comprising 267 records aggregated at 5-m depth intervals. This limited sample size limits the statistical representativeness of the training distribution and directly contributes to the ±0.12 R^2 seed-sensitivity variance documented in Section 5.3. Machine learning models trained on small datasets are inherently susceptible to high variance: slight changes in the train–test partition can alter the estimated generalisation accuracy by a margin comparable to the performance gap between competing algorithms, making single-split performance comparisons unreliable without sensitivity analysis. Additionally, 43% of records required median imputation for formation log parameters (PHIF, VSH, SW, KLOGH) due to the absence of petrophysical log coverage below 4,085 m. Although the imputation sensitivity analysis in Section 5.5 confirms that including these records improves overall model accuracy, the imputed values introduce systematic smoothing of formation variability in the lower well interval, potentially masking genuine geological heterogeneity. All models and performance metrics reported in this study are conditional on this specific imputation strategy, and results may differ if alternative imputation methods (e.g., multiple imputation or k-nearest neighbours) are used.

5.8.2. Generalisation and Formation Transferability

Models developed on a single-well dataset in the Norwegian North Sea cannot be assumed to generalise to wells in different geological settings without retraining or, at a minimum, revalidation. The Volve field is characterised by a predominantly sandstone reservoir column (82% in the training data), and the models documented in this study are correspondingly biased toward accurate prediction in sandstone intervals. The elevated MAE observed for minority lithologies (marl: n=7 test samples; claystone: n=3 test samples) is a direct consequence of this distributional bias, and performance estimates for those lithologies carry substantial uncertainty given the small sample counts. Direct application of these models to carbonate reservoirs, complex shale intervals, or deepwater turbidite systems would require substantial retraining on representative data and should not be performed without formal revalidation. Furthermore, the downhole motor configuration of Well 15/9-F-15 introduces a specific multicollinearity structure between BIT_RPM and SURF_RPM that may not be representative of rotary-drilled wells or wells with different motor specifications. Models transferred to wells with different drilling system configurations may exhibit degraded feature importance profiles and altered prediction accuracy.

5.8.3. Real-Time Implementation Constraints

The deployment framework described in Section 8 assumes continuous, reliable access to real-time WITSML surface sensor data and periodic updates to the formation log from the well prognosis. In practice, sensor outages, communication interruptions, and rig network latency can disrupt the data stream, potentially leaving the advisory system without valid inputs for extended periods. The 5-m aggregation interval used in this study aligns with post-processed log resolution but may not match the real-time data update rate available on all rig information systems. Integration with legacy SCADA and drilling data management systems on older rigs may require custom data adapters and engineering effort not captured in the deployment protocol. Additionally, the training data for this study were derived from a specific drilling program spanning November 2013 to February 2014; directional changes, drilling fluid reformulations, or equipment changes that occur during a live well would not be captured by the static trained model, potentially inducing a distribution shift that the fixed model cannot accommodate without retraining. The rolling retraining trigger described in Section 8.3 addresses this in principle, but the computational and logistical overhead of in-field retraining remains a practical constraint.

5.8.4. Model Interpretability Challenges

Although Random Forest and XGBoost provide permutation-based feature importance scores that offer aggregate-level model diagnostics, they do not yield easily interpretable local explanations for individual depth-increment predictions. A driller operating a real-time advisory system needs to understand not only what the model recommends but why a specific parameter change is projected to improve ROP at a given depth. Ensemble models are inherently black box in their prediction mechanics: the decision logic embedded in 200 parallel decision trees cannot be communicated to a non-specialist user in a comprehensible form. The inter-tree variance used as a confidence indicator in Section 8.2 provides uncertainty information but does not explain the drivers of a specific prediction. More advanced interpretability tools, such as SHAP (Shapley Additive exPlanations) values or LIME (Local Interpretable Model-Agnostic Explanations), would be required to generate per-prediction feature attribution

scores suitable for driller communication. Still, they come with computational overhead and interpretive caveats in the presence of high feature correlation. The near-perfect collinearity between BIT_RPM/SURF_RPM further complicates interpretability: the importance attributed to either variable may reflect shared variance rather than an independent physical relationship, making causal inferences about individual feature contributions unreliable without prior dimensionality reduction.

5.8.5. Sensor and MWD Data Quality Issues

The WITSML records used in this study represent surface measurements of downhole conditions, introducing a systematic representation gap: surface WOB and surface RPM differ from downhole WOB and bit RPM due to string compliance, friction, tool joint effects, and motor speed variation. Negative WOB values in the raw dataset — attributed to off-bottom sensor drift and removed as outliers — illustrate the type of measurement artefact common in surface sensor data. The 5-m depth-averaged aggregation used as the target variable (ROP_AVG) conflates on-bottom drilling time with connection time, survey time, and reaming cycles, potentially underrepresenting the instantaneous on-bottom penetration rate by a factor of 100× or more. All model accuracy metrics are therefore measured against this depth-averaged proxy rather than true drilling efficiency, which limits the direct operational relevance of the reported R^2 and MAE values. Measurement While Drilling (MWD) tools, which provide downhole weight-on-bit, downhole torque, and formation gamma-ray at the bit face in real time, were not available in the Volve dataset and were therefore not incorporated as input features. The systematic advantage of downhole measurements over surface proxies for ROP prediction has been demonstrated in the literature, and the absence of MWD data is a known limitation of surface-sensor-only modelling approaches. Future work incorporating downhole sensors is expected to yield measurable improvements in accuracy while also addressing the representational gap inherent in surface-to-downhole parameter mapping.

5.9. Discussion

The results presented in Section 5 collectively establish that ensemble tree-based algorithms substantially outperform both simple linear models and the conventional Bourgoyne–Young empirical proxy for ROP prediction on the Volve Well 15/9-F-15 dataset. This section interprets those quantitative findings in the broader context of drilling engineering practice, examines the operational implications for field deployment, and discusses the evidence base for AI-assisted decision support in petroleum applications.

5.9.1. Model Performance Interpretation

Random Forest achieved the highest test accuracy of all evaluated models, including XGBoost, Gradient Boosting, and the Bourgoyne–Young empirical proxy. The result is consistent with prior ensemble-method studies on structured drilling datasets. Three things probably drove this. The model handles nonlinear interactions among depth, WOB, and rotational speed without requiring those relationships to be specified in advance. It also tolerates multicollinearity — relevant here because BIT_RPM and SURF_RPM are nearly perfectly correlated in motor-driven sections, a feature space that trips up other methods. And ensemble averaging keeps variance in check, which matters more than usual when the training set is only 267 samples.

5.9.2. Comparison Across Models

XG Boost and Gradient Boosting both held up well. XGBoost showed better cross-validation stability across depth intervals, indicating more consistent performance, even though its peak test accuracy was slightly lower than that of Random Forest. SVR performed poorly. Kernel-based methods struggle with collinear feature spaces [22], and this dataset has several — the poor SVR result is a consequence of that, not a surprise.

5.9.3. Dataset Limitations

A few limitations are worth being direct about. This is well 267, with 267 samples. Conclusions that travel to other wells or formations should be held loosely. Missing formation log values were imputed, which introduces smoothing where there was originally a gap in the record. The lithology breakdown is also skewed: sandstone dominates, and model performance in marl and claystone is less reliable than the headline metrics suggest. There wasn't much training data for those intervals. The target variable adds another wrinkle. Depth-averaged ROP includes connection time and reaming, not just actual drilling. That noise is embedded in every prediction and limits how directly the output translates to operational decisions.

5.9.4. Practical Implications

The framework produced defensible ROP predictions despite data constraints and is reproducible. But 267 samples from a single well are a thin foundation to build on. Larger datasets and cleaner coverage of formation logs would improve accuracy.

More importantly, they would provide stronger grounds for trusting that the model is generalising rather than simply fitting the quirks of one well.

5.10. Practical Deployment Framework

Random Forest is the recommended primary model for deployment, given its highest test R^2 (0.4799). XGBoost (CV R^2 = 0.5197 ± 0.1876) is a better fallback when the model will be run on formations underrepresented in the training data — the L1/L2 regularisation limits prediction errors in out-of-distribution intervals. The framework described below applies to either model.

5.10.1. Real-Time Data Input Pipeline

Surface WITSML sensors transmit Depth, WOB, SURF_RPM, TORQUE, FLOWIN, ECDBIT, BIT_RPM, and LAGMWT at 1-second intervals. The pre-processing module aggregates these to 5-m depth intervals matching training resolution, applies the stored training-set Standard Scaler parameters (for SVR/KNN alternatives), and one-hot encodes LITH from the well prognosis. Formation log parameters (PHIF, VSH, SW, KLOGH) are pre-loaded from the well prognosis depth index. Missing formation values are filled using training-set medians.

5.10.2. Prediction, Confidence, and Advisory Output

At each 5-m depth increment, the system outputs: predicted ROP at the current drilling parameters; a confidence band derived from inter-tree variance across the 200 RF trees (inter-quartile range); ROP sensitivity curves for ±10% WOB and RPM perturbations; and a WOB-RPM combination that maximises predicted ROP within the active mud weight envelope. If the confidence band exceeds ±1.5× the training RMSE, the advisory flags the prediction for driller review before any parameter change is implemented.

5.10.3. Pre-Deployment Validation Protocol

Before live deployment the following protocol is required: (a) collect a minimum of 500 records from at least two offset wells in the target formation; (b) retrain all models and verify CV R^2 exceeds 0.60 before deployment; (c) conduct a shadow-mode trial for a minimum of 200 m where the model runs in parallel without influencing operations; (d) establish a rolling retraining trigger — if the 20-point rolling RMSE exceeds 1.5× training RMSE, model retraining is flagged. This protocol directly addresses the ±0.12 seed-sensitivity uncertainty identified in Section 5.3.

6. Conclusion

In this study, seven machine learning models were evaluated for predicting ROP using data from Volve Well 15/9-F-15. The approach used here attempts to overcome several limitations seen in previous ROP-ML studies: depth-stratified train-test splitting, one-hot lithology encoding, leakage-free scaling pipelines, VIF multicollinearity analysis, permutation-based feature importance, a Bourgoyne-Young conventional model baseline, seed sensitivity testing across multiple random splits, and an imputation sensitivity check. Random Forest achieved the best test R^2 of 0.4799 (MAE = 6.53×10^{-4} m/h), cutting MAE by 44% relative to the Bourgoyne-Young proxy ($R^2 = -0.014$, MAE = 1.176×10^{-3} m/h). That gap is directly attributable to ML's ability to capture depth-ROP nonlinearity that the empirical proxy cannot model. XGBoost placed second on test R^2 (0.4259) but led on cross-validation R^2 across all seven models (0.5197 ± 0.1876), reflecting the stabilising effect of its explicit regularisation — a 40% MAE improvement over the BY proxy. Gradient Boosting ranked third, with an R^2 of 0.4004. VIF analysis identified near-perfect collinearity between BIT_RPM and SURF_RPM (VIF > 4,900), an artefact of the downhole motor configuration. Replacing this feature pair with a Mechanical Specific Energy composite is the most important pending improvement to the model. Seed sensitivity testing showed R^2 variance of ±0.12 across stratified splits—the realistic uncertainty bound for single-well deployment that should accompany any reported accuracy figure. Imputation sensitivity analysis confirmed that including the 116 imputed records below 4,085 m improved accuracy compared to discarding them.

6.1. Priority Future Work

Expand the training dataset to 1,000+ records from multiple wells, which should bring seed sensitivity below ±0.05; replace BIT_RPM and SURF_RPM with a MSE-based feature; evaluate LSTM or Transformer architectures that can exploit temporal correlations in the WITSML depth sequence; separate on-bottom time from connection time to compute true instantaneous ROP as the prediction target; and run shadow-mode validation alongside an active well before committing to live advisory deployment.

Acknowledgement: The authors sincerely acknowledge Dhaanish Ahmed College of Engineering for providing the academic environment, facilities, and resources necessary for conducting this research.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author and may be provided upon reasonable request.

Funding Statement: No financial assistance, grant, or external funding was received for the preparation of this manuscript or the conduct of the associated research.

Conflicts of Interest Statement: The authors declare that they have no conflicts of interest related to this work. The study represents an original contribution by the authors, and all relevant sources of information have been properly cited and referenced.

Ethics and Consent Statement: This research was conducted in accordance with applicable ethical standards, and informed consent was obtained from all participants before they participated in the study**.**

References

1. M. J. Kaiser, "A survey of drilling cost and complexity estimation models," *International Journal of Petroleum Science and Technology*, vol. 1, no. 1, pp. 1–22, 2007.
2. F. E. Dupriest and W. L. Koederitz, "Maximizing drill rates with real-time surveillance of mechanical specific energy," *Proceedings of SPE/IADC Drilling Conference*, Amsterdam, Netherlands, 2005.
3. A. T. Bourgoyne, K. K. Millheim, M. E. Chenevert, and F. S. Young, "Applied drilling engineering," *Society of Petroleum Engineers*, 1986. [Accessed by 12/07/2024].
4. W. C. Maurer, "The perfect-cleaning theory of rotary drilling," *Journal of Petroleum Technology*, vol. 14, no. 11, pp. 1270–1274, 1962.
5. M. G. Bingham, "A new approach to interpreting rock drillability," *Petroleum Publishing Company*, Tulsa, Oklahoma, United States of America, 1965.
6. R. Jahanbakhshi, R. Keshavarzi, and A. Jafarnezhad, "Real-time prediction of rate of penetration during drilling operation in oil and gas wells," *Proceedings of the 46th US Rock Mechanics/Geomechanics Symposium*, Chicago, Illinois, United States of America, 2012.
7. C. Hegde and K. E. Gray, "Use of machine learning and data analytics to increase drilling efficiency for nearby wells," *Journal of Natural Gas Science and Engineering*, vol. 40, no. 4, pp. 327–335, 2017.
8. C. Soares and K. Gray, "Real-time predictive capabilities of analytical and machine learning rate of penetration models," *Journal of Petroleum Science and Engineering*, vol. 172, no. 1, pp. 934–959, 2019.
9. S. B. Ashrafi, M. Anemangely, M. Sabah, and M. J. Ameri, "Application of hybrid artificial neural networks for predicting rate of penetration," *Journal of Petroleum Science and Engineering*, vol. 175, no. 4, pp. 604–623, 2019.
10. D. Moran, H. Ibrahim, A. Purwanto, and J. Osmond, "Sophisticated ROP prediction technologies based on neural network delivers accurate drill time results," *Proceedings of IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition*, Ho Chi Minh City, Vietnam, 2010.
11. R. F. Mitchell and S. Z. Miska, "Fundamentals of drilling engineering," SPE Textbook Series, *Society of Petroleum Engineers*, 2011. [Accessed by 12/07/2024].
12. S. Elkatatny, M. Mahmoud, Z. Tariq, and A. Abdulraheem, "New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial intelligence network," *Neural Computing and Applications*, vol. 30, no. 1, pp. 2673–2683, 2018.
13. A. Al-AbdulJabbar, S. Elkatatny, A. A. Mahmoud, T. Moussa, D. Al-Shehri, M. Abughaban, and A. Al-Yami, "Prediction of the rate of penetration while drilling horizontal carbonate reservoirs using the self-adaptive artificial neural networks technique," *Sustainability*, vol. 12, no. 4, p. 1376, 2020.
14. C. I. Noshi and J. J. Schubert, "The role of machine learning in drilling operations: a review," *Proceedings of SPE/AAPG Eastern Regional Meeting*, Pittsburgh, Pennsylvania, United States of America, 2018.
15. A. K. Abbas, S. Rushdi, M. Alsaba, and M. F. Al Dushaishi, "Drilling rate of penetration prediction of high-angled wells using artificial neural networks," *Journal of Energy Resources Technology*, vol. 141, no. 11, p. 112904, 2019.
16. Equinor, " Volve field data set," *equinor.com*, 2018. [Accessed by 01/07/2024].
17. L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," *Chapman and Hall*, New York, United States of America, 1984.
18. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 10, pp. 5–32, 2001.
19. L. F. F. M. Barbosa, A. Nascimento, M. H. Mathias, and J. A. de Carvalho Jr., "Machine learning methods applied to drilling rate of penetration prediction and optimization: A review," *Journal of Petroleum Science and Engineering*, vol. 183, no. 12, p. 106332, 2019.

20. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
21. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, United States of America, 2016.
22. V. Vapnik, "The nature of statistical learning theory," *Springer*, New York, United States of America, 1995.
23. R. Teale, "The concept of specific energy in rock drilling," *International Journal of Rock Mechanics and Mining Sciences*, vol. 2, no. 1, pp. 57–73, 1965.
24. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 11, pp. 2825–2830, 2011.
25. C. Strobl, A. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.
26. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 1–21, 2012.