mining

MDPI

Article

The Role of EDA in Developing Robust Machine Learning Models for Lithology and Penetration Rate Prediction from MWD Data

Jesse Addy [1,*], Ishmael Anafo [2] and Erik Westman [1,*]

[1] Department of Mining & Minerals Engineering, Virginia Tech, Blacksburg, VA 24061, USA
[2] Department of Mining Engineering, University of Utah, Salt Lake City, UT 84112, USA; ishmaelbanafo@gmail.com
* Correspondence: addyjesse18@vt.edu (J.A.); ewestman@vt.edu (E.W.)

Abstract

Measure-While-Drilling (MWD) data provide real-time insight into subsurface conditions and drilling performance, yet their complexity and operational noise often hinder reliable modeling. This study demonstrates the role of Exploratory Data Analysis (EDA) in developing robust machine learning (ML) models for lithology classification and penetration rate (PR) prediction in mining operations. A structured EDA workflow—comprising data integrity assessment, feature distribution analysis, correlation mapping, and depth-wise parameter profiling—was implemented to identify redundant attributes, isolate non-productive intervals, and enhance dataset consistency. Through EDA-informed normalization and feature selection, data consistency and model performance were significantly improved. Machine learning algorithms, including Decision Tree, Random Forest, and Multi-Layer Perceptron, were trained on the refined dataset. The Random Forest Classifier achieved 98.45% accuracy in lithology prediction, while the Random Forest Regressor produced the most accurate PR estimation (R^2 = 0.83, RMSE = 0.52). These results highlight EDA as a critical foundation for constructing physics-informed, data-driven models that enhance predictive reliability and operational efficiency in mining environments.

Keywords: exploratory data analysis; machine learning; measure while drilling; lithology; penetration rate

1. Introduction

Exploratory Data Analysis represents a critical stage in the data science workflow, serving as the bridge between raw data acquisition and model development. First articulated by Tukey [1], EDA emphasizes understanding data through visualization, pattern recognition, and iterative inspection rather than relying solely on statistical inference. Subsequent works by Church [2] and Komorowski [3] reinforced EDA as a process of "quantitative detective work," where analysts uncover relationships, inconsistencies, and hidden structure that inform subsequent modeling. Modern perspectives position EDA as an adaptive, feedback-driven approach that enhances transparency and reliability in machine learning pipelines [4–7]. Despite this, EDA is often treated as a preliminary or optional step, rather than a methodological core of robust predictive modeling.

The increasing volume and complexity of geotechnical and drilling data have heightened the importance of systematic exploratory analysis. Measurement-While-Drilling systems continuously capture drilling parameters such as penetration rate (PR), weight

Academic Editor: Roohollah Shirani Faradonbeh

Received: 25 December 2025
Revised: 27 January 2026
Accepted: 9 February 2026
Published: 4 March 2026

on bit, torque, rotational speed, and standpipe pressure. These data encapsulate valuable information about rock lithology and subsurface conditions, yet their potential remains underexploited in many industrial and research contexts. MWD signals are typically noisy, context-dependent, and sensitive to operational factors, leading to uncertainty in downstream modeling if not adequately explored and preprocessed [8–10]. Guidelines such as the Florida Method of Test for MWD [8] and early technical reviews stress that the reliability of predictive analytics depends as much on data quality assurance and calibration as on algorithm selection. Effective exploratory analysis enables the identification of anomalous readings, scaling issues, and sensor biases that could otherwise propagate through to model outputs.

Recent advances in machine learning have shown promising results in predicting lithology [11–19] and estimating penetration rate [20,21] from MWD data. However, the literature reveals that these efforts have primarily focused on optimizing algorithmic architectures, often overlooking the fundamental role that EDA plays in shaping input features and ensuring generalization. Muraina et al. [5] emphasized that robust modeling requires an in-depth understanding of data distributions, correlations, and multicollinearity before training, while the editorial The Artificial Intelligence Advantage [6] noted that AI and ML methods can augment, rather than replace, exploratory reasoning. Automation frameworks such as SmartEDA [7] now provide systematic tools for data visualization and variable profiling, offering scalable solutions for large-volume industrial datasets. Nonetheless, in drilling analytics—where geological heterogeneity, sensor drift, and operational variability intersect—EDA remains underutilized as a design principle guiding feature engineering and model validation.

Integrating EDA into the machine learning workflow can substantially improve interpretability and performance in predictive drilling models. By examining variable interactions through correlation matrices, scatterplots, and principal-component projections, practitioners can detect redundancies and nonlinear dependencies that affect model bias. Visualization of class distributions and parameter trends enables a more grounded understanding of lithological transitions, while clustering and anomaly detection help isolate noise and outliers before training. These exploratory insights support model design choices such as hyperparameter tuning, feature weighting, and data augmentation strategies. Moreover, EDA-informed preprocessing ensures that derived features—such as normalized torque or pressure ratios—better represent the underlying physical processes captured by MWD instrumentation.

This study focuses on the role of EDA in constructing robust ML models for predicting lithology and penetration rate from MWD data. The research demonstrates how detailed exploratory analysis of drilling parameters enhances both the accuracy and interpretability of machine learning predictions. By systematically quantifying data quality, identifying feature relevance, and revealing patterns of geological variability, EDA is shown to function not merely as a preliminary step but as a methodological foundation for reliable model development. The study thereby bridges a persistent gap in data-driven geotechnical research: the disconnect between exploratory understanding and predictive performance. Ultimately, this integration promotes a transparent, reproducible workflow that improves both model generalization and confidence in decision-support systems for drilling and mineral exploration.

2. Materials and Methods

2.1. Data Source and Description

This study applied an EDA-driven machine learning workflow to predict lithology and penetration rate using MWD data collected from an iron-ore mine in the United States. The

mine forms part of a major iron-bearing province characterized by banded iron formations composed of alternating hematite-rich, magnetite-bearing, and cherty layers.

Drilling was conducted using semi-automated rigs equipped with real-time monitoring systems, which continuously recorded mechanical and hydraulic parameters at approximately 2-inch (5 cm) depth intervals. The raw dataset contained approximately 235,501 valid measurements acquired from 1436 drill holes, with each record representing a sampling interval of about 2 inches. Each observation captured both spatial coordinates (X, Y, Z) and drilling parameters describing the mechanical and hydraulic response of the rock during drilling. In total, 14 recorded variables were available, including penetration rate, feed pressure, flushing pressure, rotation pressure, weight on bit, rotation torque, engine rotation speed, feed force, and water injection volume. These variables collectively characterize the operational dynamics of the drilling process and form the foundation for subsequent predictive modeling. A concise description of the recorded parameters and their functions is presented in Table 1.

Table 1. Recorded MWD and spatial parameters and their descriptions.

Field Name	Description
Hole ID	Unique identifier for each drilling hole or borehole.
X	The horizontal coordinate (easting) of a point in the drilling operation.
Y	The horizontal coordinate (northing) of a point in the drilling operation.
Z	The vertical coordinate (depth) of a point in the drilling operation.
Lithology	Description or classification of the geological formation encountered during drilling.
Engine RPM	Rotations per minute of the drilling engine.
Rotation Speed	Speed at which the drill bit rotates.
Feed Pressure	Pressure applied to feed the drill bit into the formation.
Flushing Pressure	Pressure of the flushing fluid is used to remove cuttings from the borehole.
Penetration Rate	Rate at which the drill bit advances into the formation.
Rotation Pressure	Pressure applied to the drill bit during rotation.
Feed Force	Force applied to feed the drill bit into the formation.
Weight On Bit	Downward force exerted on the drill bit by the drilling rig.
Rotation Torque	Torque applied to the drill bit during rotation.
Water Injection Volume	Volume of water injected into the borehole during drilling.

Ten lithological categories were identified from geological mapping and exploration logs and encoded as integer labels (2–13) to maintain consistency with the MWD dataset. These classes represent distinct rock units observed in the deposit and are broadly interpreted as follows:

- 2—Magnetite-bearing ore;
- 3—Claystone or minor alteration zone;
- 4—Waste rock or breccia lens;
- 7—Thin shale band or unclassified material;
- 8—Taconite or intermediate banded iron formation;
- 9—Hematite-rich ore;
- 10—Banded low-grade taconite;
- 11—Chert or siliceous layer;
- 12—Hematitic shale;
- 13—Transitional lithology or boundary rock.

2.2. Data Cleaning and Preprocessing

Prior to analysis, the raw MWD dataset underwent a systematic cleaning and preprocessing workflow to ensure consistency, remove redundancies, and improve data quality for subsequent exploratory and machine learning analyses. All procedures were implemented

in Python 3.11.5 using the pandas, numpy, and seaborn libraries. Initial inspection of the dataset revealed the presence of four redundant lithology-related columns: Lithology.1, yfit21, yfit23, and yfit25. Pairwise comparisons between these columns and the primary Lithology field showed less than 0.35% variation across all entries, confirming that they represented near-identical values. Consequently, these columns were removed to eliminate redundancy and prevent multicollinearity in subsequent analyses.

2.3. Exploratory Data Analysis

Exploratory Data Analysis was central to this study, serving as both a diagnostic and interpretive framework for understanding the relationships among drilling parameters, spatial context, and lithology. The EDA process guided all subsequent cleaning, feature selection, and modeling decisions.

2.3.1. Spatial Distribution and Geological Structure

The first stage of exploration examined the spatial organization of the dataset to assess borehole coverage, depth distribution, and lithological continuity. Three-dimensional and two-dimensional visualizations of the drilling coordinates were generated to evaluate sampling density and stratigraphic layering across the mining area.

Figure 1a–d illustrate the spatial distribution of boreholes across the study area. The 2D and 3D visualizations reveal two primary drilling clusters separated along the *X*-axis, along with vertically stratified lithological layers consistent with the geological framework of the deposit. A focused view of the main drilling region (X > 39,000) highlights denser sampling and lateral lithological continuity, supporting the inclusion of spatial coordinates as predictive features in subsequent analyses.

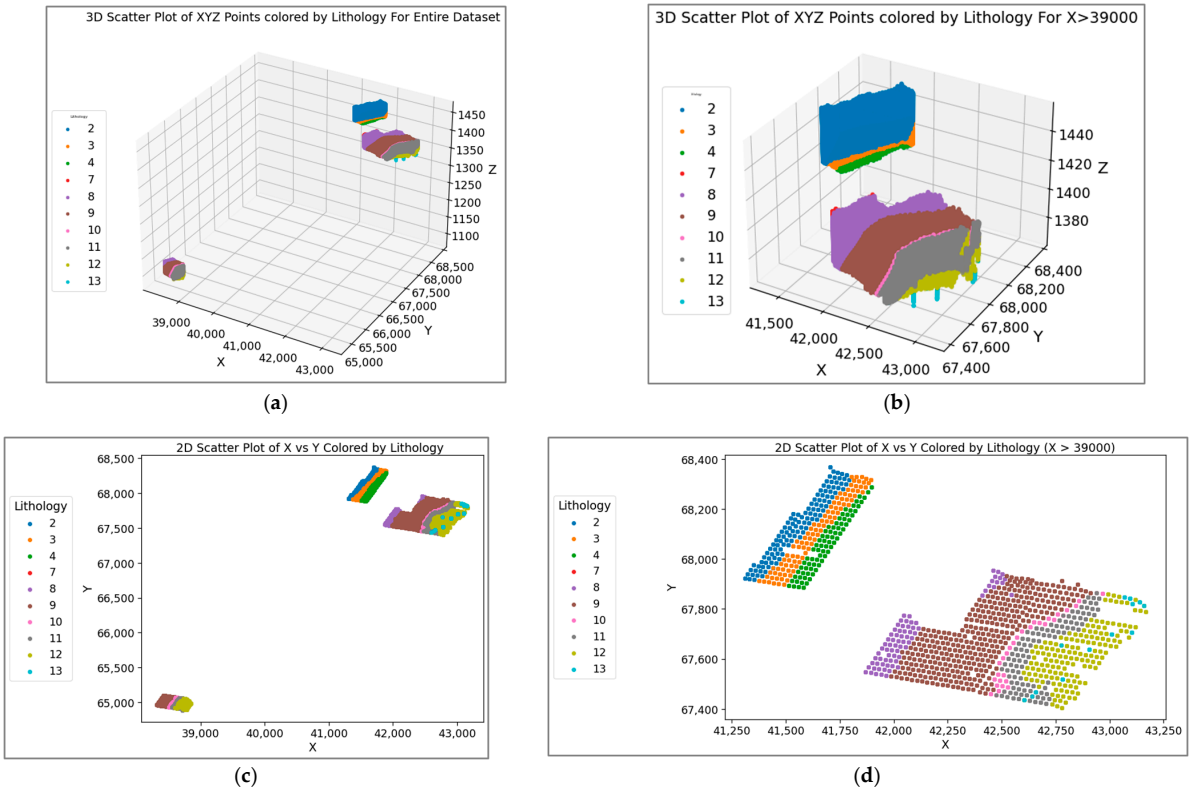


Figure 1. Spatial distribution of boreholes and lithological layering. (**a**) 3D scatter plot of X–Y–Z coordinates across all boreholes. (**b**) 3D subset (X > 39,000) highlighting the main drilling zone and lithological layering. (**c**) 2D scatter plot of X–Y coordinates colored by lithology for the entire dataset. (**d**) 2D projection of X–Y coordinates for X > 39,000, showing lateral lithological continuity.

2.3.2. Statistical Overview of Drilling Parameters

Descriptive statistics were computed for the major drilling parameters in the raw MWD dataset to establish baseline variability and operational behavior before any preprocessing. Table 2 summarizes these statistics, including measures of central tendency, dispersion, and percentile ranges.

Table 2. Descriptive statistics for primary MWD parameters in the raw dataset.

Category	Engine RPM	Feed Pressure	Flushing Pressure	Penetration Rate	Rotation Pressure	Feed Force	Rotation Torque	Weight on Bit
count	235,501.00	235,501.00	235,501.00	235,501.00	235,501.00	235,501.00	235,501.00	235,501.00
mean	1799.48	2557.54	49.65	2.12	1853.05	66,286.50	2767.32	66,331.62
std	2.44	512.03	6.87	1.41	298.58	12,143.80	565.85	12,151.97
min	1652.00	1.45	0.00	0.07	732.45	15,987.96	550.00	15,998.95
10%	1798.00	2001.54	40.61	0.79	1493.90	52,905.24	2062.50	52,939.60
50%	1800.00	2686.12	49.31	1.74	1841.99	69,976.22	2750.00	70,023.77
90%	1801.00	3021.16	58.02	3.84	2210.39	75,948.33	3437.50	75,999.96
max	1816.00	3125.59	118.93	9.97	4013.23	77,719.04	6875.00	77,771.92

The raw dataset exhibited substantial variability across drilling parameters. Variables such as feed pressure, flushing pressure, and rotation torque displayed wide ranges and high standard deviations, reflecting fluctuations due to changing lithological hardness and bit–rock interaction. In contrast, engine RPM and rotation speed remained nearly constant, indicating stable operational control throughout drilling. The raw distributions of these parameters are illustrated in Figure 2, which visualizes the variability and distribution shape of both spatial and mechanical variables.

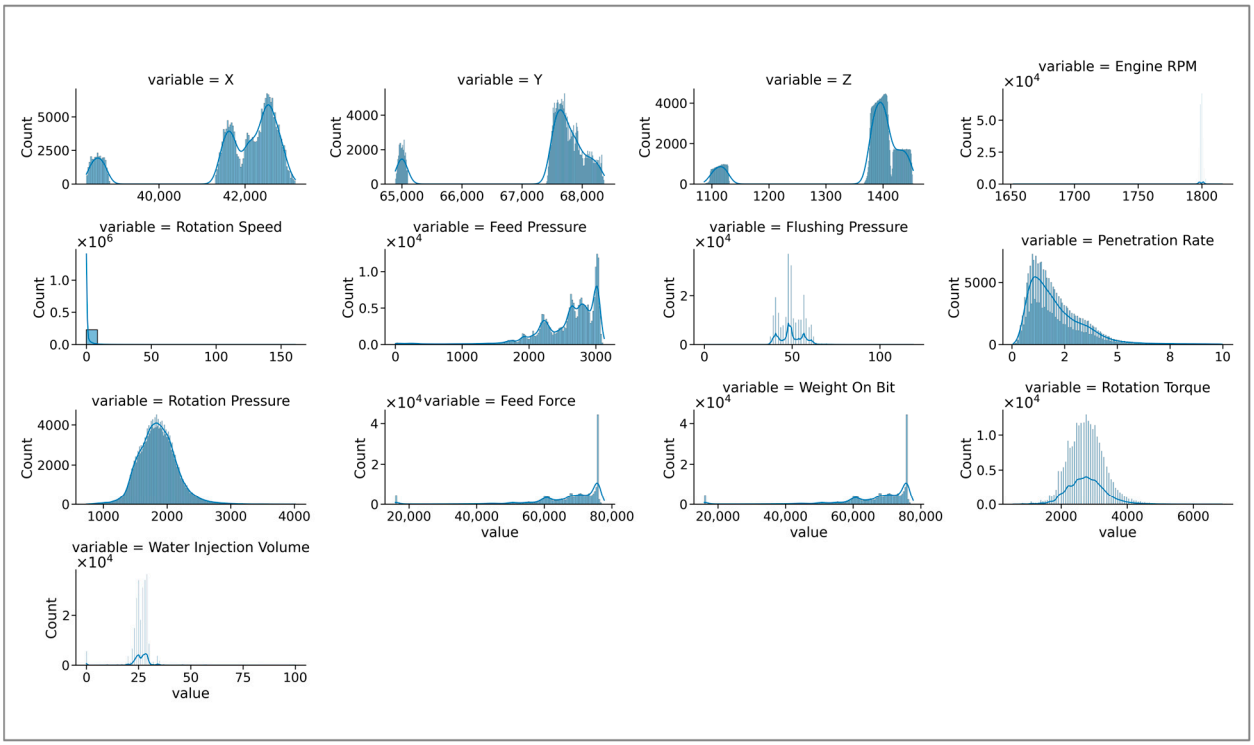


Figure 2. Histograms showing univariate distributions of the principal MWD parameters and spatial coordinates in the raw dataset. The bars represent the frequency of observations, and the blue line shows the smoothed distribution trend for each variable.

Figure 2 shows that X and Y coordinates exhibit multimodal distributions corresponding to distinct blast areas, while Z reveals discrete layering consistent with vertical stratification in the deposit. Mechanical response variables such as penetration rate, feed force, and rotation torque display positively skewed distributions, reflecting geological heterogeneity and transitions between soft and hard rock formations. Conversely, control variables (engine RPM, rotation speed) show narrow, centered distributions indicative of consistent drilling operations. However, further statistical examination revealed that rotation speed values were predominantly zero across all observations, suggesting a recording or sensor anomaly. Consequently, this feature was excluded from subsequent analyses, as it provided no meaningful variation.

To complement these numerical trends, the categorical distribution of lithology was analyzed using a Pareto chart (Figure 3). This analysis revealed that lithologies 9, 8, and 2 together accounted for approximately 80% of all recorded intervals, while the remaining categories appeared infrequently. The dominance of these three lithologies underscores the stratified nature of the orebody and highlights the inherent class imbalance within the dataset. Understanding this imbalance was critical for guiding preprocessing, particularly in balancing class representation during modeling.

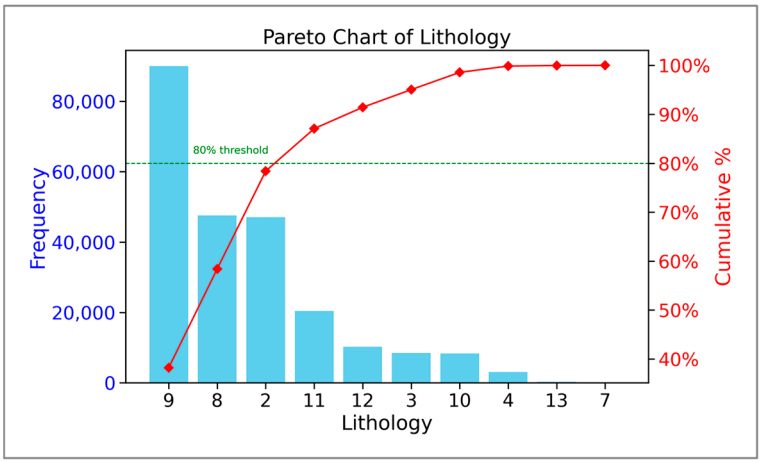
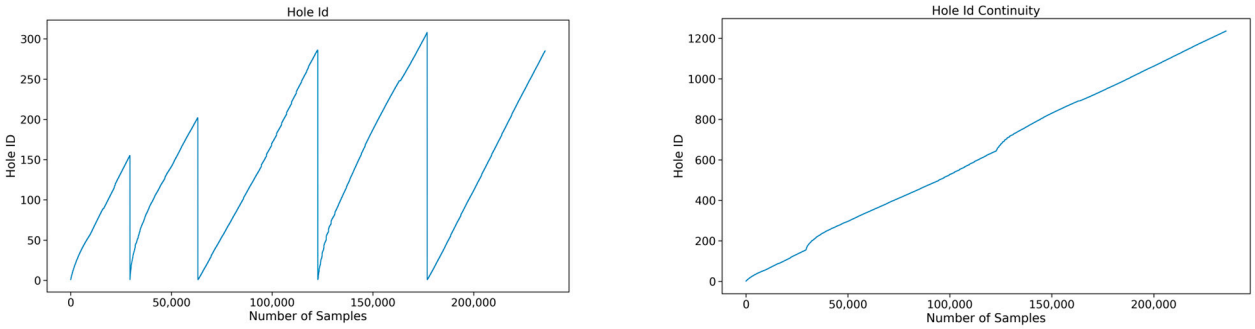


Figure 3. Pareto chart of lithology frequency. Blue bars represent the frequency of each lithology class, the red line indicates the cumulative percentage, and the green dashed line marks the 80% threshold. Three dominant lithologies (9, 8, and 2) collectively account for approximately 80% of all samples.

2.3.3. Hole ID Continuity

Before conducting multivariate analyses, the continuity of hole identifiers was examined to ensure consistent indexing across the drilling campaigns. Each blast pattern in the raw dataset originally numbered holes independently (e.g., restarting from Hole ID = 1 for each pattern). This discontinuity risked disrupting sequential analyses and spatial referencing across multiple patterns.

A line plot of Hole ID versus sample index revealed several discontinuous jumps (Figure 4a), confirming that hole numbering restarted within each blast. To maintain a unified spatial index and preserve sequential integrity, a correction was implemented to render all Hole IDs continuous across the entire dataset. After adjustment, the updated Hole ID sequence (Figure 4b) showed a smooth, cumulative progression consistent with a single continuous dataset. This modification ensured that subsequent analyses—particularly spatial and depth-based modeling—treated all boreholes as part of a single integrated drilling operation rather than disjointed subsets.



(**a**) (**b**)

Figure 4. Correction of Hole ID sequencing across blast patterns in the MWD dataset. (**a**) Discontinuous Hole ID numbering observed in the raw data across multiple blast patterns. (**b**) Continuous Hole ID sequence after renumbering, ensuring consistent spatial indexing.

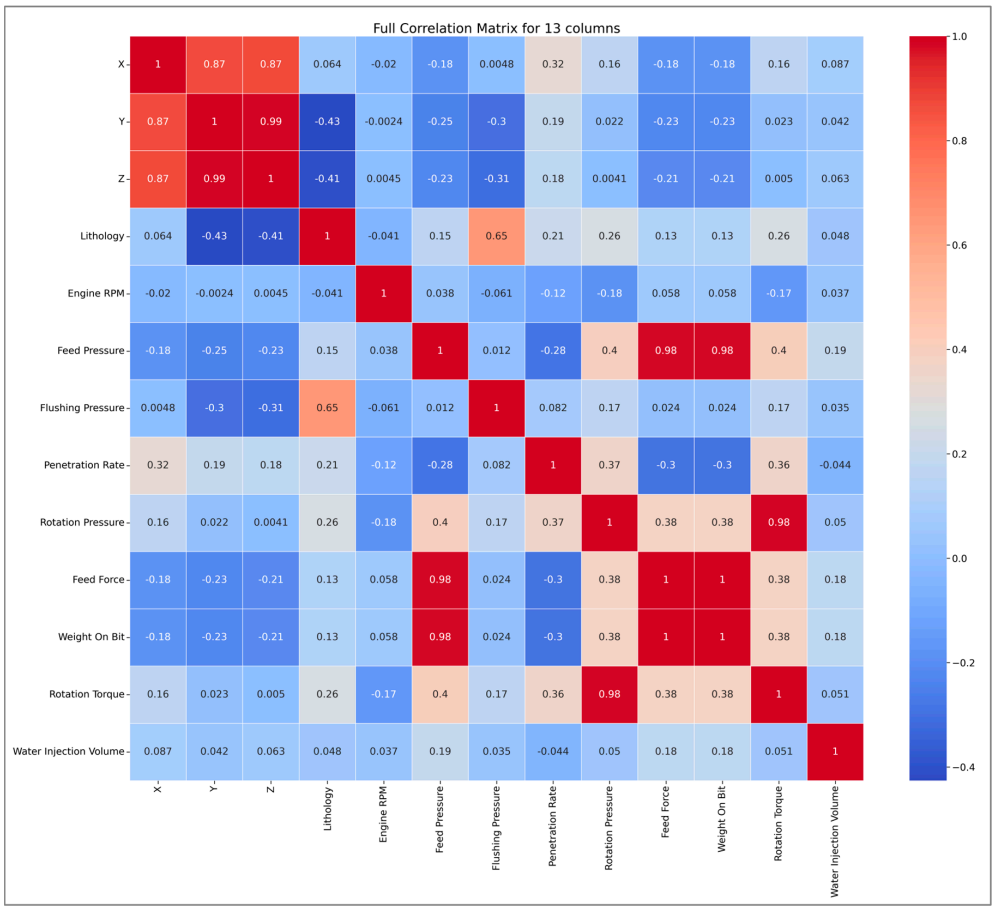
2.3.4. Correlation and Feature Interaction Analysis

Following univariate exploration, multivariate analysis was performed to examine relationships among all mechanical, hydraulic, and spatial parameters. A Pearson correlation matrix (Figure 5a) was generated using 13 variables to identify redundant or weakly associated features.

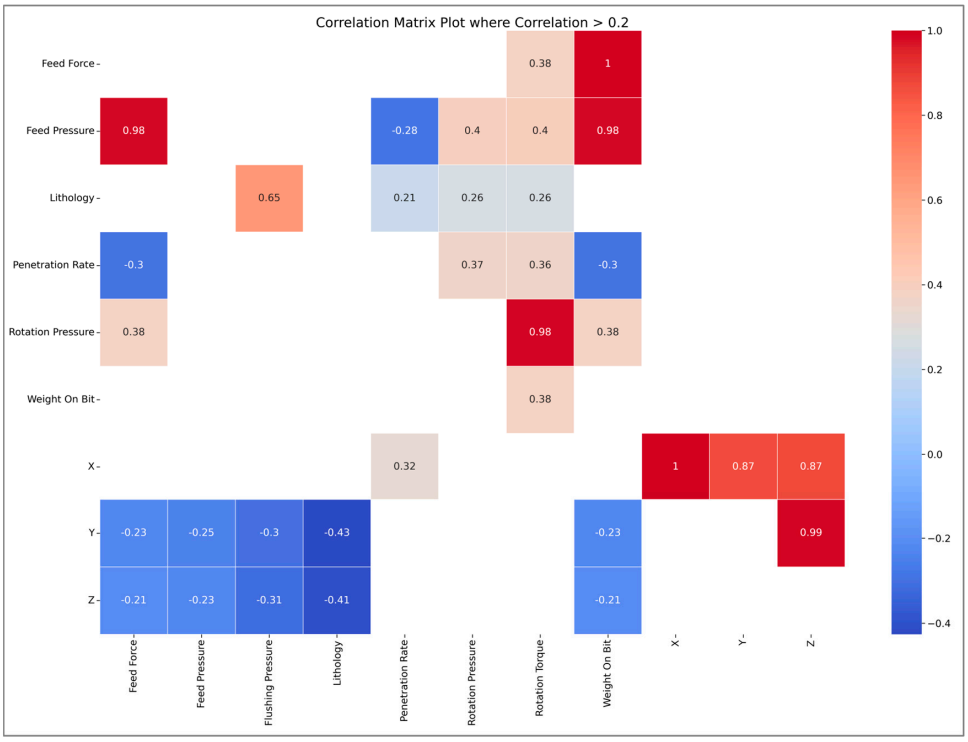
The full matrix revealed several strong linear associations. Feed Force and Weight on Bit exhibited the highest correlation ($r \approx 0.98$), indicating that both describe nearly identical load responses during drilling. Feed Pressure and Rotation Pressure also showed a strong positive relationship ($r \approx 0.83$), reflecting coupled mechanical–hydraulic effects within the drilling system. Moderate positive correlations ($r \approx 0.3$–0.4) were observed between Penetration Rate and Feed Pressure, and between Rotation Torque and Rotation Pressure, suggesting interdependent influences of feed energy and bit–rock resistance.

In contrast, parameters such as Water Injection Volume and Flushing Pressure displayed negligible correlations ($|r| < 0.2$) with all other variables, indicating limited predictive or diagnostic relevance. The removal of Water Injection Volume and Flushing Pressure was a heuristic decision informed by expert knowledge of drilling operations at this specific site. These variables are frequently adjusted at the driller's discretion and are not directly governed by the physical response of the geological formation. A correlation threshold of $|r| = 0.2$ was therefore adopted as a conservative exploratory screening criterion to identify variables with weak linear association. This threshold was not treated as a strict physical cutoff but rather applied in conjunction with visual inspection of correlation matrices and physical interpretation of drilling mechanics. The low correlation values of these variables reinforced the hypothesis that they primarily represent operational variability rather than lithological drivers. The strong performance of the nonlinear models indicates that the retained feature set was sufficient to capture the dominant nonlinear relationships governing drilling response. Based on this criterion, variables with overall correlations below 0.2 were excluded to streamline the feature set and reduce noise. The filtered correlation matrix (Figure 5b) shows the remaining meaningful relationships among 11 retained parameters.

These results confirmed that key operational features—Feed Pressure, Feed Force, Weight On Bit, Rotation Pressure, and Rotation Torque are highly interdependent, while Water Injection Volume and Flushing Pressure displayed negligible linear correlations with other variables, suggesting that their effects on drilling performance may be independent or nonlinear. The resulting reduced feature set served as the foundation for subsequent modeling and interpretation.



(**a**)



(**b**)

Figure 5. Pearson correlation matrices for MWD parameters showing linear relationships among spatial and operational variables. (**a**) Full correlation matrix illustrating all pairwise relationships. (**b**) Filtered correlation matrix displaying only correlations with $|r| > 0.2$.

2.3.5. Drilling Parameter Behavior and Operational Anomalies

The earlier correlation analysis from Figure 5 revealed that Penetration Rate appeared negatively (though weakly) correlated with Engine RPM, Feed Pressure, Feed Force, and Weight on Bit. This observation was physically implausible—under typical drilling conditions, higher mechanical input should generally correspond to faster penetration. To investigate this inconsistency, a depth-based analysis was performed to examine how drilling parameters evolved through individual boreholes.

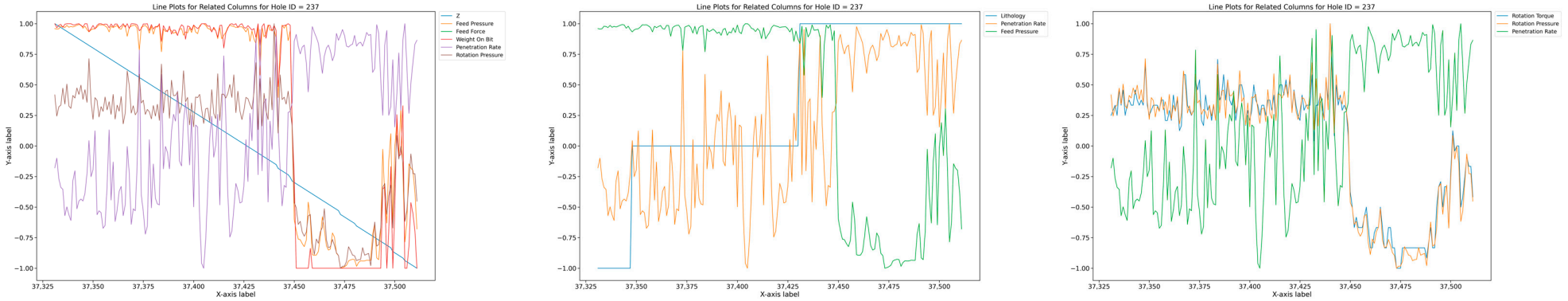
The drilling data were recorded continuously from the collar to the toe of each hole, making each borehole an independent depth sequence. Analyses were therefore conducted one hole at a time to preserve depth continuity. Hole ID 237, which intersected the greatest number of lithological transitions, was selected for detailed evaluation and depth-wise line plots were generated for selected parameters (Figure 6a–c).

As shown in Figure 6a–c, the parameters maintained consistent relationships with depth until approximately sample 37,450, where a distinct anomaly emerged. At this point, Penetration Rate spiked sharply while Feed Pressure, Feed Force, and Weight on Bit dropped to their lowest values. This inverse behavior occurred immediately after a lithological change, suggesting that it reflected operational rather than geological or mechanical conditions.

In Figure 6a, Feed Pressure, Feed Force, Weight on Bit, and Rotation Pressure remain steady and well-correlated through most of the borehole. Near depth 37,450, however, all three load-related parameters collapse while Penetration Rate rises abruptly—a pattern inconsistent with genuine rock cutting, where reduced load should result in slower advancement.

Figure 6b situates this anomaly within its geological context. A lithological transition precedes the irregular response, after which Feed Pressure declines and Penetration Rate fluctuate erratically. This indicates the bit likely entered a problematic layer—possibly fractured or unstable—that disrupted normal drilling equilibrium.

In Figure 6c, Rotation Torque and Rotation Pressure remain nearly constant across the same interval, while Penetration Rate diverges sharply. The absence of a torque increase confirms that the high penetration readings were not due to added mechanical energy but instead to non-productive string motion, such as hole-cleaning operations used to restore circulation.



(**a**) (**b**) (**c**)

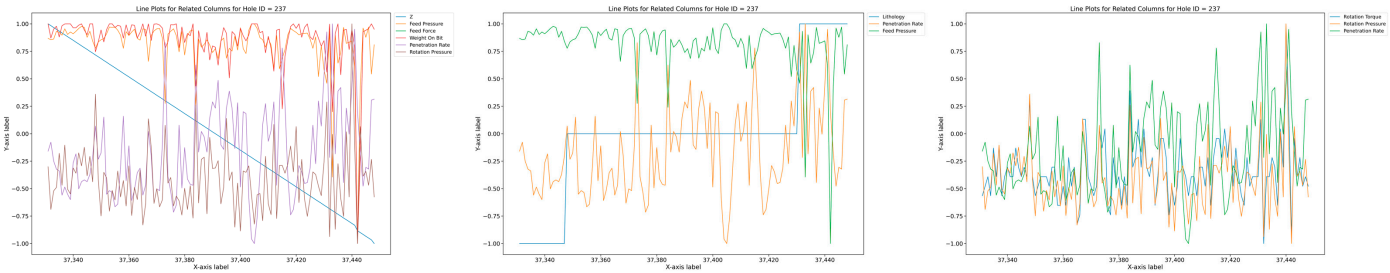
Figure 6. Depth-wise variation in key MWD parameters for Hole ID 237, illustrating the observed penetration rate anomaly near sample 37,450. (**a**) Feed-, weight-, and pressure-related parameters. (**b**) Lithology, penetration rate, and feed pressure. (**c**) Torque- and rotation-related parameters.

2.3.6. Data Filtering and Its Effect on MWD Parameters

To ensure data quality and restrict the modeling dataset to periods of active drilling, a quantitative filter was applied. Rows with Weight-on-Bit values below the 20th percentile were removed, under the assumption that active drilling occurs primarily at higher bit loads. This filtering step effectively excluded intervals associated with cleaning or non-productive operations.

The selection of the 20th percentile Weight-on-Bit (WOB) threshold was based on an EDA-driven, empirically motivated assessment of the physical and statistical relationship between Weight on Bit (WOB) and Penetration Rate (PR), rather than formal numerical optimization. Exploratory sensitivity checks were performed by examining the WOB–PR relationship at multiple percentile levels (5th, 10th, 15th, 20th, and 25th). Below the 20th percentile, the relationship between WOB and PR was weakly negative, which is physically counter-intuitive for active drilling and characteristic of non-productive operations such as collaring or hole-cleaning, where the drill bit is not fully engaged with the formation. At the 20th percentile, this relationship transitioned to a weakly positive correlation, indicating the onset of a stable drilling regime in which increased bit load contributes meaningfully to penetration. Based on this observed transition, filtering the bottom quintile of WOB values allowed the dataset to better represent intervals of active rock cutting, ensuring that subsequent analyses and model training focused on lithological resistance rather than operational artifacts.

Following this refinement, the same parameter relationships were re-examined for Hole ID 237. As shown in Figure 7a–c, the anomalous behavior observed earlier near sample 37,450 disappeared, and the drilling variables now exhibit physically consistent relationships with depth. Feed Pressure, Feed Force, and Weight on Bit vary proportionally with Penetration Rate, reflecting realistic drilling response under stable operating conditions.



(**a**) (**b**) (**c**)

Figure 7. Depth-wise variation in key MWD parameters for Hole ID 237, after filtering out low Weight-on-Bit intervals (<20th percentile). (**a**) Feed-, weight-, and pressure-related parameters. (**b**) Lithology, penetration rate, and feed pressure. (**c**) Torque- and rotation-related parameters.

Similarly, Rotation Torque and Rotation Pressure now show synchronized variation with Penetration Rate, confirming proper mechanical coupling between rotation and penetration. Lithological transitions also align more smoothly with parameter fluctuations, indicating that the filtered dataset primarily represents intervals of active cutting rather than non-productive movements or hole-cleaning events.

To quantify the effect of filtration, descriptive statistics were calculated for key MWD parameters before and after filtering the data. Table 3 summarizes representative metrics (mean, standard deviation, and selected percentiles).

Table 3. Comparison of key MWD parameter statistics before and after filtering low Weight-on-Bit values (<20th percentile).

Parameter	Dataset	Mean	Std	10th %	50th %	90th %	Max
Feed Pressure (kPa)	Unfiltered	2557.54	512.03	2001.54	2686.12	3021.16	3125.59
	Filtered	2751.62	235.05	2380.09	2780.4	3026.96	3125.59
Penetration Rate (m/min)	Unfiltered	2.12	1.41	0.79	1.74	3.84	9.97
	Filtered	1.98	1.25	0.79	1.71	3.51	9.97
Rotation Pressure (kPa)	Unfiltered	1853.05	298.58	1493.9	1841.99	2210.39	4013.23
	Filtered	1901.47	260.62	1579.47	1886.96	2224.9	3607.12
Feed Force (N)	Unfiltered	66,286.5	12,143.8	52,905.24	69,976.22	75,948.33	77,719.04
	Filtered	70,995.02	4739.35	63,263.03	71,965.67	75,948.33	77,719.04
Rotation Torque (Nm)	Unfiltered	2767.32	565.85	2062.5	2750	3437.5	6875
	Filtered	2859.72	494.28	2268.75	2818.75	3506.25	6462.5
Weight on Bit (N)	Unfiltered	66,331.62	12,151.97	52,939.6	70,023.77	75,999.96	77,771.92
	Filtered	71,043.31	4742.55	63,306.29	72,014.55	75,999.96	77,771.92

Filtering produced noticeable effects across all major drilling parameters. The mean and lower percentile values of Feed Pressure, Feed Force, and Rotation Torque increased, reflecting the exclusion of intervals with abnormally low bit load and energy input. This shift confirms that the removed records corresponded to non-drilling activities such as hole cleaning.

Flushing Pressure and Penetration Rate both showed reduced variability after filtration, suggesting improved consistency in hydraulic conditions and more stable rock–bit interaction. The Rotation Pressure parameter also narrowed in range, indicating the removal of extreme operational events, such as freeing a stuck bit.

Overall, the refined dataset exhibited tighter distributions and reduced noise, making it more representative of true drilling performance and suitable for subsequent modeling analyses.

2.4. Modeling Framework and Approach

The cleaned and filtered dataset (Section 2.3.6) was used to develop machine learning models for two primary objectives: lithology classification and penetration rate prediction. The modeling framework integrated exploratory data analysis, data normalization, and multiple supervised learning algorithms to capture both linear and nonlinear relationships between the MWD parameters and their target outputs. All analyses were conducted in Python 3.11.5 using the scikit-learn machine learning library, supplemented with visualization and evaluation tools.

2.4.1. Data Setup and Preprocessing

The refined dataset was partitioned into training (75%) and testing (25%) subsets using a heuristic split designed to balance learning and generalization rather than a borehole-by-borehole split. While MWD data are sequential and spatially correlated along individual boreholes, this strategy was selected to preserve statistical distribution parity between the training and testing subsets. In geological datasets, strict borehole-wise splitting can introduce distribution shifts in which test boreholes contain lithological units or drilling conditions not represented in the training data, thereby confounding model evaluation. Randomized shuffling ensures that both subsets sample the same population of drilling conditions across the site.

To prevent bias, the training and testing subsets were verified to be qualitatively identical, meaning they exhibited similar statistical distributions across all variables. This

verification was performed using Ci_tools, a specialized Python 3.11.5 library developed by Dr. Rajive Ganguli (University of Utah). Ci_tools provides utilities for dataset normalization and distributional comparison by analyzing percentile statistics between subsets [22]. The data was iteratively shuffled until the statistical characteristics of the training and testing sets were consistent, ensuring the model was tested on data drawn from the same population as the training set. After validation, the dataset was divided into input and target subsets for each task. Subsequent modeling and evaluation were performed using standard scikit-learn pipelines, ensuring reproducibility and modular implementation.

2.4.2. Feature Configuration and Normalization

The input feature space comprised 11 predictive variables, including both spatial coordinates and MWD parameters: Engine RPM, Feed Pressure, Flushing Pressure, Rotation Pressure, Weight on Bit, Rotation Torque, and Feed Force. Lithology, encoded as categorical integers, served as the target variable for classification, while Penetration Rate was modeled as a continuous output for regression. All continuous features were normalized using z-score normalization to standardize scale and enhance numerical stability during training.

2.4.3. Model Selection and Implementation

To evaluate both interpretability and predictive capability, a combination of linear, tree-based, and neural network algorithms was applied to each prediction task:

1. Penetration Rate Regression:
    - Linear Regression (LR);
    - Decision Tree Regression (DTR);
    - Random Forest Regression (RFR);
    - MLP Regressor.
2. Lithology Classification:
    - Decision Tree (DT);
    - Random Forest (RF);
    - Multi-Layer Perceptron (MLP) Classifier.

For the ensemble models, the Random Forest was trained with 50 estimators, while the MLP utilized the ReLU activation function and Adam optimizer with hidden layer configurations (100, 50). Decision tree models were constrained using maximum depth derived from data size ($\log_2$(n/min_samples)) to prevent overfitting.

2.4.4. Model Evaluation and Performance Metrics

Model performance was assessed using both quantitative metrics and diagnostic visualization tools. Although the lithology distribution is imbalanced, ensemble-based classifiers such as Random Forest inherently reduce bias toward dominant classes through bootstrap sampling and recursive partitioning. Model performance was therefore evaluated using precision, recall, and F1-score in addition to accuracy, with spatial prediction maps used to assess behavior in both dominant and rare lithologies. For classification, performance was evaluated using Accuracy, Precision, Recall and F1-score. For regression, the following metrics were computed: Coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

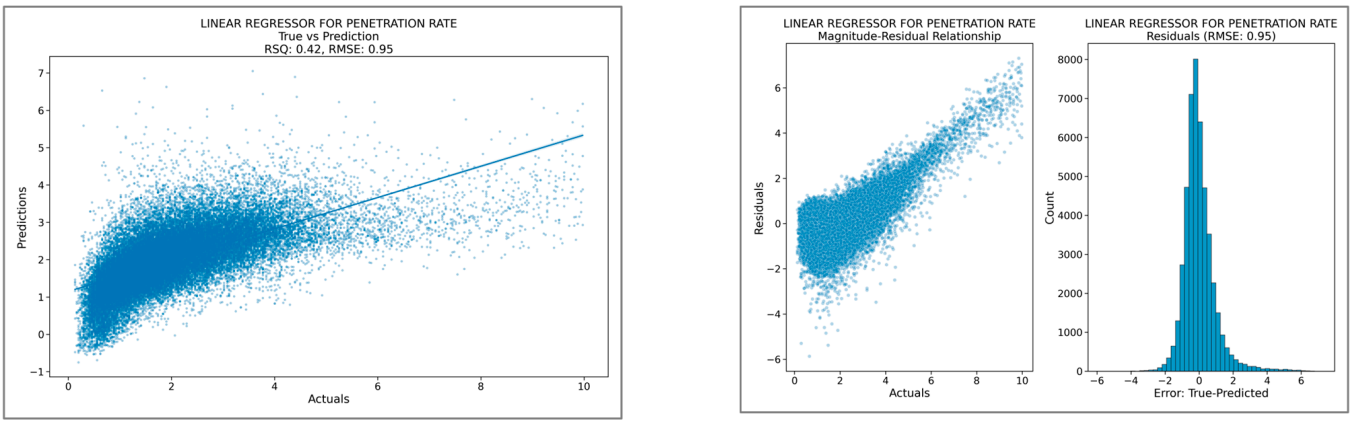
Evaluation and visualization were conducted to enable direct comparison across model types. In addition to numerical results, regression models were examined using true vs. predicted plots and residual distributions, while classification results were analyzed using normalized confusion matrices and 2D spatial prediction maps (X–Y plane). The spatial plots compared predicted versus actual lithologies, highlighting zones of agreement and misclassification.

3. Results

This section presents the performance outcomes of the machine learning models developed for Penetration Rate Regression and Lithology Classification. Each model's results are evaluated using quantitative metrics and diagnostic visualizations, highlighting their predictive accuracy, consistency, and interpretability.

3.1. Penetration Rate Prediction

Four regression models were implemented to predict Penetration Rate (PR): Linear Regression (LR), Decision Tree Regression (DTR), Random Forest Regression (RFR), and Multi-Layer Perceptron (MLP). Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and residual-based diagnostics. Figure 8a–h illustrate the true-predicted relationships and residual behaviors for each model.



(a) **(b)**

Figure 8. Linear Regressor performance for penetration rate prediction. (**a**) True vs. Predicted PR showing overall trend and underestimation of higher values (R^2 = 0.42, RMSE = 0.95) where each dot represents one observation. (**b**) Residual analysis: **left**—magnitude–residual relationship; **right**—residual distribution histogram.

3.1.1. Linear Regression (LR)

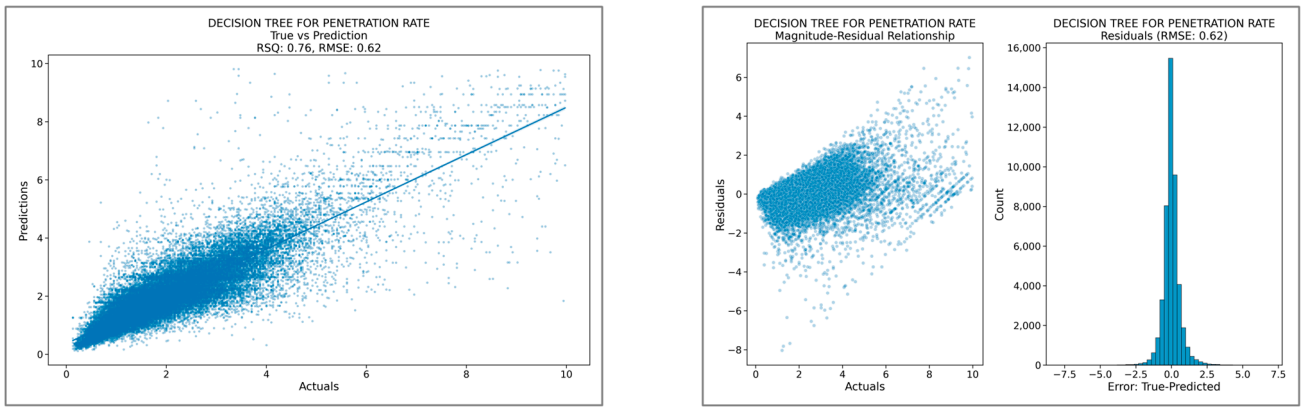
The Linear Regression model served as the baseline. As shown in Figure 8a,b, it captured the general increasing trend between actual and predicted penetration rates but systematically underestimated higher values.

Residuals displayed a widening spread with increasing PR magnitude, indicating heteroscedasticity and unmodeled nonlinear effects. Although residuals were approximately centered on zero, their dispersion confirmed the model's limited capacity to capture the complex, coupled influence of hydraulic and mechanical drilling parameters.

3.1.2. Decision Tree Regression (DTR)

Introducing nonlinear decision boundaries markedly improved prediction fidelity. As shown in Figure 9a,b, residual dispersion narrowed relative to the linear model, and the histogram exhibited a sharp, symmetric distribution.

Predictions were more consistent across the PR range, though slight overfitting was visible for high-rate observations. The discrete step pattern in predictions reflects the tree's piecewise constant structure. Overall, DTR captured dominant nonlinear relationships effectively but remained sensitive to localized noise due to its single-tree formulation.

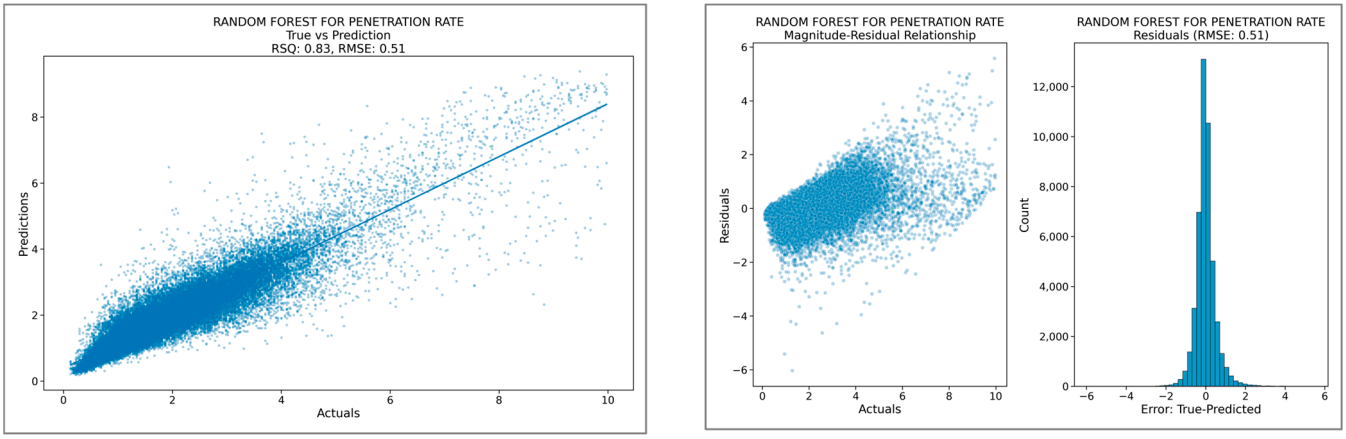


(**a**) (**b**)

Figure 9. Decision Tree Regressor performance for penetration rate prediction. (**a**) True vs. Predicted PR illustrating improved fit across the full range of observations (R^2 = 0.76, RMSE = 0.62) where each dot represents one observation (**b**) Residual analysis: **left**—magnitude–residual relationship; **right**—residual distribution histogram.

3.1.3. Random Forest Regression (RFR)

The Random Forest ensemble delivered the highest overall predictive accuracy and stability. Figure 10a,b show a near-linear alignment between predicted and actual values with minimal deviation from the 1:1 reference line.



(**a**) (**b**)

Figure 10. Random Forest Regressor performance for penetration rate prediction. (**a**) True vs. Predicted PR showing strong alignment and minimal dispersion (R^2 = 0.83, RMSE = 0.51) where each dot represents one observation. (**b**) Residual analysis: **left**—magnitude–residual relationship; **right**—residual distribution histogram.

Residuals were tightly centered and normally distributed, confirming excellent generalization and low variance. Ensemble averaging mitigated local noise and smoothed erratic fluctuations, producing a robust and physically consistent mapping between MWD parameters and penetration rate.

3.1.4. Multi-Layer Perceptron (MLP) Regressor

The Multi-Layer Perceptron (MLP) Regressor, configured with four hidden layers (100–100–100–100 neurons) and ReLU activation, was evaluated on both the training and testing datasets to assess model generalization and convergence behavior.

As shown in Figure 11a–d, both the training and testing phases of the MLP regressor exhibit similarly low predictive performance ($R^2 \approx 0.36$), with comparable RMSE values.

This indicates that the MLP failed to adequately capture the relationship between the input features and penetration rate. Such behavior is characteristic of underfitting, suggesting that the selected network architecture and configuration were not well suited to the complexity and noise characteristics of the MWD dataset.

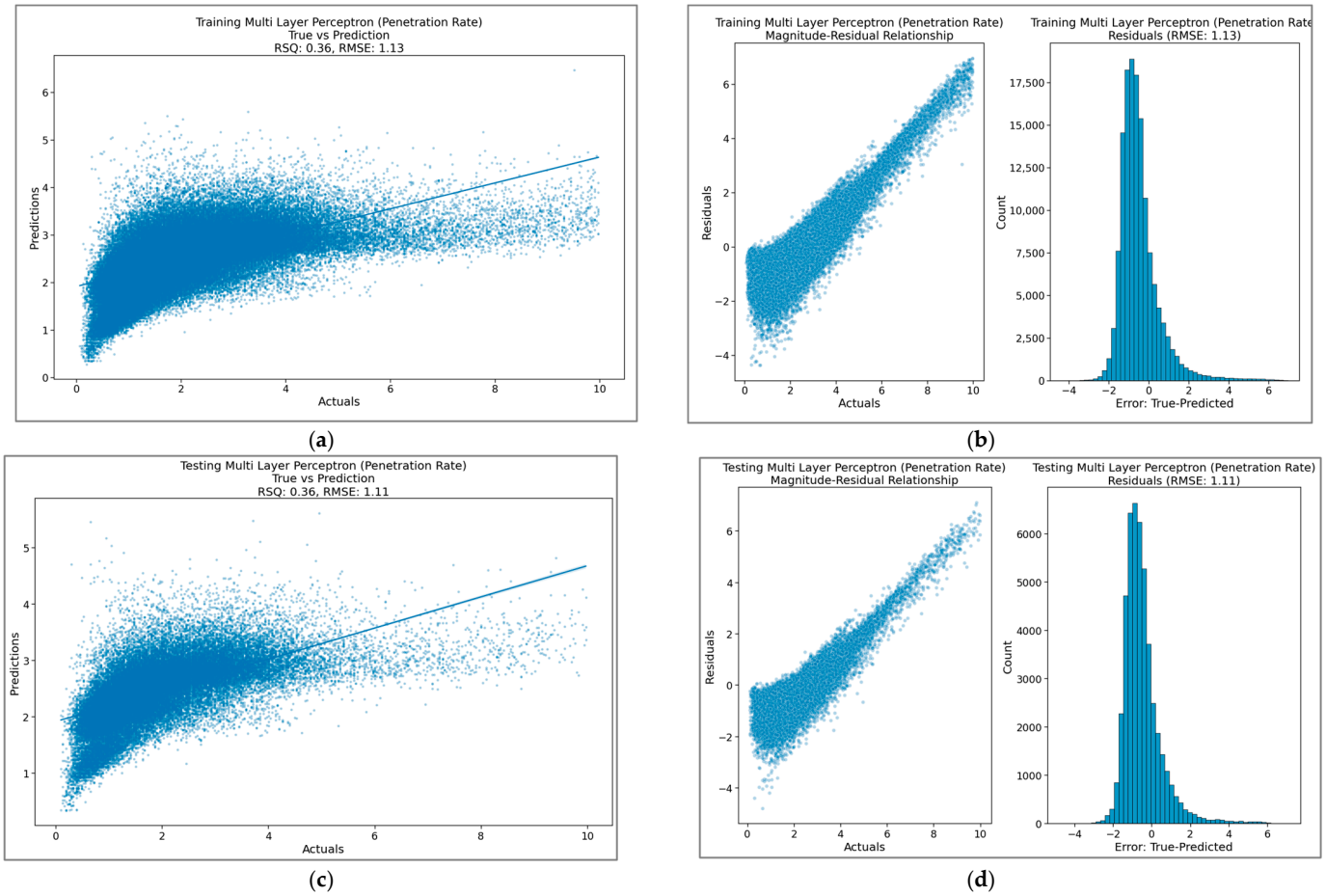


Figure 11. Multi-Layer Perceptron Regressor performance for penetration rate prediction. (**a**) True vs. Predicted PR for the training dataset, showing strong correlation and close fit to the ideal line (R^2 = 0.36, RMSE = 1.13). (**b**) Residual distribution for the training set, indicating low bias and stable performance across penetration magnitudes. (**c**) True vs. Predicted PR for the testing dataset, showing weak generalization and underestimation at higher penetration rates (R^2 = 0.36, RMSE = 1.11). (**d**) Residual distribution for the testing dataset, displaying large variance and diffuse residual patterns indicative of model overfitting.

Residual plots for both the training and testing datasets (Figure 11b,d) exhibited broad dispersion with a mild positive bias, indicating systematic underestimation at higher penetration rates. The histograms were approximately Gaussian but displayed wide tails, confirming increased variance and limited generalization capability of the MLP model.

The weaker generalization of the MLP reflects the sensitivity of neural networks to noise, response imbalance, and hyperparameter configuration in operational MWD datasets, rather than a limitation of neural architectures per se. In contrast, tree-based ensemble models are inherently more robust to such conditions and require less tuning to capture nonlinear parameter interactions. While neural networks can theoretically model complex nonlinear dependencies, the current configuration failed to converge to a meaningful representation under default training conditions.

3.1.5. Summary of Regression Results

Model performance for Penetration Rate (PR) prediction varied with algorithm complexity. The Random Forest Regressor produced the most accurate results (R^2 = 0.83, RMSE = 0.52), followed by the Decision Tree Regressor (R^2 = 0.75, RMSE = 0.63). The Linear Regression model captured general penetration trends but underestimated higher values, while the Multi-Layer Perceptron showed weak generalization (R^2 = 0.08).

Overall, ensemble-based models yielded the most reliable and physically consistent predictions of drilling performance. A summary of regression metrics is presented in Table 4.

Table 4. Summary of regression model performance for Penetration Rate (PR) prediction.

Model	R^2	RMSE	Remarks
Linear Regression	0.42	0.95	Baseline model; underestimates highs
Decision Tree Regressor	0.76	0.62	Captures nonlinear trends
Random Forest Regressor	0.83	0.51	Best performance
Multi-Layer Perceptron Regressor	0.36	1.13	Underfitting evident; weak generalization

3.2. Lithology Classification

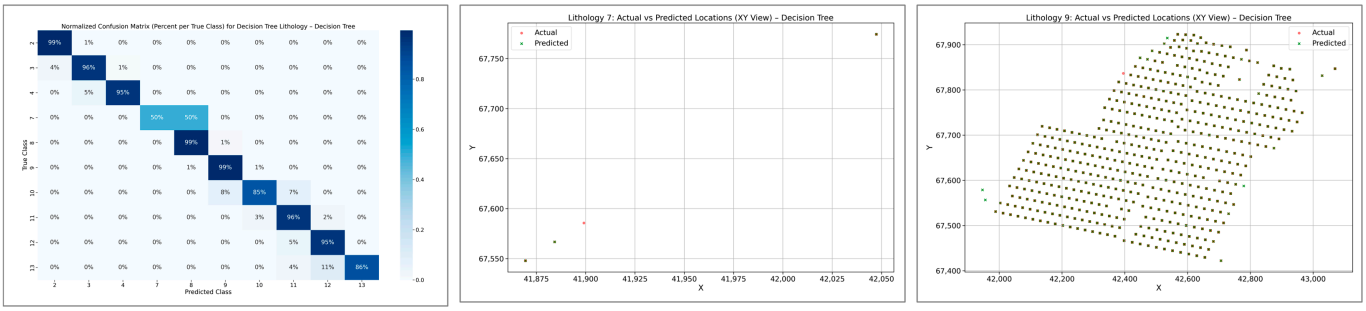
Each model was assessed using accuracy, precision, recall, and F1-score, as well as visual diagnostics from confusion matrices and spatial prediction maps. To visualize both dominant and rare geological units, two representative lithologies were selected:

- Lithology 7, the most dominant;
- Lithology 9, the least frequent unit in the dataset.

These cases provide insight into each model's ability to generalize across class imbalance and capture spatial variability.

3.2.1. Decision Tree Classifier

The Decision Tree reached an overall accuracy of 97.5%, establishing a strong baseline for comparison. The confusion matrix (Figure 12a) shows high accuracy along the main diagonal, with only minor misclassifications between neighboring lithologies such as 10–11 and 12–13. Spatial predictions for Lithology 7 (Figure 12b) closely match the true distribution with minimal deviation, while predictions for Lithology 9 (Figure 12c) show slight overextension around the cluster margins.



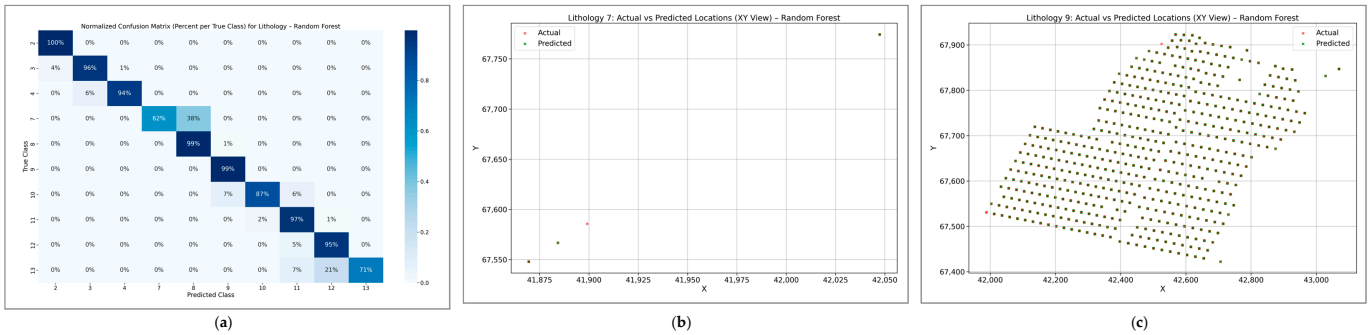
(**a**) (**b**) (**c**)

Figure 12. Decision Tree classification results. (**a**) Normalized confusion matrix. (**b**) Spatial comparison of predicted vs. actual locations—Lithology 7. (**c**) Spatial comparison of predicted vs. actual locations—Lithology 9.

These patterns suggest that the model captured dominant lithologies effectively but tended to overpredict the boundaries of less-represented units.

3.2.2. Random Forest Classifier

Random Forest delivered the highest overall performance, with 98.45% accuracy and an F1 score of 0.95. The confusion matrix (Figure 13a) displays a sharply defined diagonal, confirming highly consistent predictions across all classes. Spatial predictions for Lithology 7 (Figure 13b) show excellent alignment between predicted and true zones, while Lithology 9 (Figure 13c) is also well captured with minimal scatter. These results highlight the ensemble's superior ability to model nonlinear boundaries and maintain stability across both major and rare lithologies.

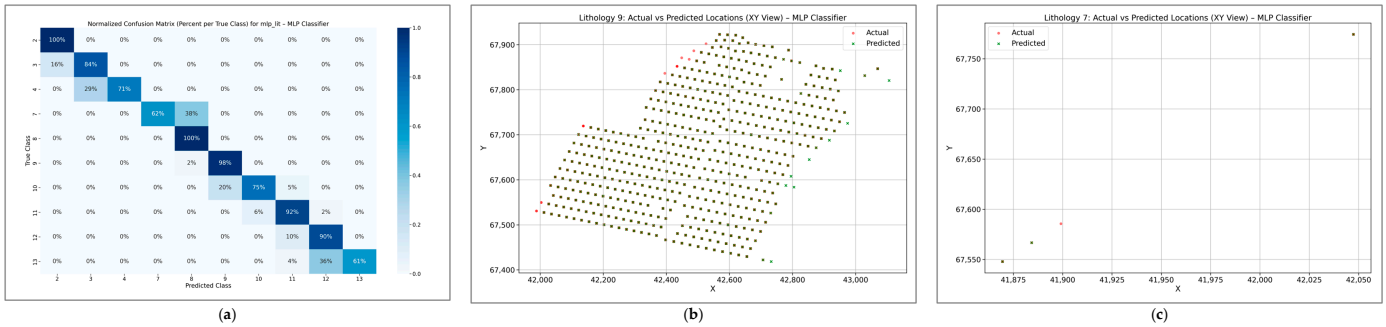


(**a**) (**b**) (**c**)

Figure 13. Random Forest classification results. (**a**) Normalized confusion matrix. (**b**) Spatial comparison of predicted vs. actual locations—Lithology 7. (**c**) Spatial comparison of predicted vs. actual locations—Lithology 9.

3.2.3. Multi-Layer Perceptron (MLP) Classifier

The MLP achieved 97.7% accuracy, comparable to the Decision Tree but with slightly better recall for under-represented lithologies. The confusion matrix (Figure 14a) reveals strong performance on dominant formations but a small reduction in precision for minor classes. Spatial predictions for Lithology 7 (Figure 14b) reproduce the main lithological pattern with high fidelity, while Lithology 9 (Figure 14c) appears somewhat diffuse, indicating the network's sensitivity to class imbalance. Nevertheless, the MLP effectively captured key stratigraphic trends and overall structural continuity.



(**a**) (**b**) (**c**)

Figure 14. MLP classification results. (**a**) Normalized confusion matrix. (**b**) Spatial comparison of predicted vs. actual locations—Lithology 7. (**c**) Spatial comparison of predicted vs. actual locations—Lithology 9.

3.2.4. Summary of Classification Results

As summarized in Table 5, the Random Forest classifier achieved an F1-score of 0.95 and a recall of 0.93, indicating balanced predictive performance across lithological classes, including minority units. A model driven primarily by class dominance would typically exhibit inflated accuracy accompanied by substantially lower F1-scores; the consistently high F1-scores observed across Decision Tree (0.92), Random Forest (0.95), and

MLP (0.93) models confirm robust and well-distributed classification performance. All classifiers achieved high accuracy, exceeding 97%, indicating strong predictive capability for lithological variation from MWD parameters.

Table 5. Summary of Classification Metrics for Lithology Prediction.

Model	Accuracy (%)	Precision	Recall	F1 Score	Remarks
Decision Tree	97.50	0.95	0.90	0.92	Slight overprediction
Random Forest	98.45	0.96	0.93	0.95	Best performance
Multi-Layer Perceptron	97.70	0.92	0.94	0.93	Balanced results

Although the overall lithology classification accuracy exceeds 98%, this performance should be interpreted with caution given the spatial continuity of geological formations and the dominance of a limited number of lithological units within the dataset. To mitigate potential bias from class imbalance and spatial clustering, model performance was evaluated using metrics beyond global accuracy.

The Random Forest Classifier performed best, providing the highest accuracy (98.45%) and most stable classification across lithologies. The Decision Tree produced reliable predictions for dominant lithologies but tended to overextend rare classes, while the MLP Classifier balanced recall and precision with slight dispersion in minority units. A summary of classification metrics is provided in Table 5.

4. Discussion

Exploratory data analysis (EDA) played a central role in shaping the modeling framework and interpreting its outcomes. The initial assessment of the MWD dataset revealed distinct patterns and interdependencies among drilling parameters that directly influenced both lithology classification and penetration rate prediction. Correlation analysis and parameter trend visualization established the mechanical and operational foundations upon which machine learning models were constructed, while systematic data filtering ensured that model inputs reflected only active drilling conditions.

4.1. Influence of Exploratory Analysis on Model Development

The comprehensive correlation study (Figure 5a,b) showed that Feed Pressure, Feed Force, Weight on Bit, and Penetration Rate shared strong positive relationships, forming the core indicators of active drilling behavior. Conversely, Rotation Torque and Rotation Pressure displayed weaker correlations with depth-dependent parameters, highlighting their sensitivity to localized rock–bit interactions rather than overall operational trends. These relationships informed the feature selection process by prioritizing variables with direct mechanical relevance and minimizing redundancy among predictors.

Depth-wise trend analyses (Figure 6a–c) further revealed operational anomalies, most notably the abrupt inverse behavior near sample 37,450, where Penetration Rate spiked as mechanical loads dropped. This pattern, initially attributed to lithological change, was ultimately traced to non-productive drilling activities such as hole cleaning. The subsequent filtration procedure—removing Weight-on-Bit values below the 20th percentile—eliminated such intervals, yielding a dataset that more accurately represented true cutting conditions. The improvement was confirmed visually (Figure 7a–c) and statistically (Table 3), as variability across parameters decreased and relationships between mechanical inputs and penetration response were restored. This filtering step proved critical for reducing noise and enabling more physically meaningful model learning.

4.2. Interpretation of Regression Model Performance

The performance trends among the regression models reflect the nature of the underlying MWD data and the extent of nonlinearity among variables. The Linear Regression model, though useful as a baseline, underperformed due to its inability to capture the coupled nonlinear effects of feed, torque, and hydraulic parameters on penetration rate. Its residuals exhibited heteroscedasticity, confirming a mismatch between linear assumptions and drilling dynamics.

While neural networks can theoretically model complex nonlinear relationships, their performance in drilling applications is highly dependent on careful hyperparameter optimization and data conditioning, whereas ensemble tree-based methods provide more stable performance under noisy, field-scale conditions.

Tree-based models substantially improved prediction accuracy, with the Decision Tree achieving moderate gains and the Random Forest outperforming all others. The ensemble approach mitigated overfitting by averaging multiple weak learners, leading to smoother residual distributions and superior generalization (R^2 = 0.83). The Random Forest's success aligns with prior studies that demonstrated its robustness in handling high-dimensional, noisy drilling datasets where variable interactions are complex and locally nonlinear.

The Multi-Layer Perceptron (MLP) achieved comparable overall accuracy but exhibited higher variance in predictions at extreme penetration values. This sensitivity likely arises from imbalanced data and limited tuning of hyperparameters. While the MLP captured general nonlinear relationships, its relatively shallow architecture constrained its ability to generalize beyond dominant data regions. Nonetheless, the network's performance validates the potential of deep learning approaches when sufficient optimization and data balancing are implemented.

4.3. Interpretation of Lithology Classification

For lithology prediction, all three classifiers—Decision Tree, Random Forest, and MLP—achieved high accuracies exceeding 97%, underscoring the strong link between MWD parameters and subsurface lithological variation. The Decision Tree provided interpretable decision boundaries but slightly overpredicted rare lithologies, as evident from spatial maps and confusion matrices. The Random Forest, with an accuracy of 98.45%, delivered the most stable and spatially coherent predictions, accurately delineating dominant lithologies (e.g., Lithology 7) while maintaining good representation of rarer units (e.g., Lithology 9).

The MLP classifier achieved accuracy like the Decision Tree but demonstrated improved recall for underrepresented lithologies, suggesting its flexibility in learning subtle transitions within the drilling data. However, mild spatial diffusion in rare lithology predictions indicated that class imbalance and limited training samples still affected the neural network's precision. Overall, the Random Forest provided the most balanced performance, combining predictive accuracy, spatial realism, and robustness against class imbalance.

4.4. Broader Implications

The findings highlight the value of rigorous EDA and preprocessing in developing reliable data-driven models for drilling analytics. The 20th percentile Weight-on-Bit filter proved pivotal in removing operational noise and aligning statistical and physical consistency between parameters. Moreover, the success of ensemble-based methods reinforces the advantage of models that can capture nonlinear, multivariate dependencies without requiring extensive parameter tuning.

From an operational standpoint, these models demonstrate that high-resolution MWD data can serve as a reliable proxy for real-time lithology estimation and penetration rate

prediction. The strong alignment between predicted and actual values suggests potential for adaptive drilling optimization, such as real-time bit selection, feed rate adjustment, or anomaly detection during blasting operations.

4.5. Future Work

This study is based on MWD data from a single iron ore deposit and, as such, does not aim to demonstrate universal model transferability across different geological settings. The primary objective is to illustrate how rigorous exploratory data analysis improves data integrity, physical consistency, and predictive reliability within a real operational dataset. Although the resulting models show strong performance at this site, broader generalization requires validation using multi-deposit and multi-bench datasets, which will be the focus of future work.

While the 20th percentile Weight-on-Bit (WOB) threshold was selected based on a physically interpretable transition between non-productive and active drilling behavior, a comprehensive sensitivity analysis across alternative thresholds (e.g., 10%, 15%, and 25%) was beyond the scope of this study. Future work should explicitly evaluate the robustness of the filtering strategy by systematically varying percentile cutoffs and assessing their impact on model performance and parameter relationships. Such analysis would further strengthen confidence in the general applicability of the proposed EDA-driven filtering approach.

It is acknowledged that the use of a row-based train–test split in spatially correlated MWD data may lead to optimistic performance estimates, as samples from nearby spatial locations can appear in both subsets. Accordingly, the reported classification accuracy should be interpreted as representative of site-scale characterization performance rather than fully independent prediction in unseen geological domains.

Also, the current models achieved high predictive accuracy; however, future research should address class imbalance through data augmentation or resampling and explore more advanced architectures such as gradient boosting or deep neural networks. Incorporating temporal features and real-time operational data could further enhance prediction robustness. Additionally, integrating uncertainty quantification would make these models more applicable for field deployment and decision support.

Future studies should also examine model performance across multiple benches or sites to evaluate robustness under varying geological and operational conditions and evaluate borehole-wise or spatially blocked validation strategies to more conservatively assess model generalization and explicitly quantify the impact of spatial autocorrelation on model performance. Cross-site datasets would support development of normalization strategies and provide stronger evidence for broader applicability beyond a single operation. Finally, extending plan-view analysis to three-dimensional representations would improve interpretation of vertical continuity and subsurface geometry, enabling closer integration with geological models and mine planning workflows.

5. Conclusions

This study demonstrated that coupling exploratory data analysis (EDA) with machine learning provides a reliable framework for predicting both penetration rate and lithology from MWD data.

The 20th-percentile Weight-on-Bit filtering effectively isolated intervals of active drilling, enhancing data integrity and model performance. Among regression models, the Random Forest Regressor achieved the best predictive accuracy (R^2 = 0.83), while the Random Forest Classifier provided the most consistent lithology predictions (98.45% accuracy) and spatial coherence.

These results confirm that ensemble-based models outperform linear and single-tree approaches for capturing the nonlinear and multivariate nature of drilling processes. The integration of EDA-driven data refinement, physical interpretability, and robust modeling establishes a practical foundation for real-time lithology identification and drilling optimization in future operations.

Author Contributions: Conceptualization, E.W., J.A. and I.A.; methodology, J.A. and I.A.; software, J.A. and I.A.; validation, J.A. and I.A.; data curation, J.A.; writing—original draft preparation, J.A.; writing—review and editing, I.A., E.W. and J.A.; visualization, I.A.; supervision, E.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Center to Advance the Science of Exploration to Reclamation in Mining (CASERM), a collaborative partnership between Colorado School of Mines and Virginia Tech and part of the National Science Foundation's Industry–University Cooperative Research Centers (IUCRC) program. Funding is provided by the National Science Foundation through the project titled 'IUCRC Phase II+ Virginia Tech: Center to Advance the Science of Exploration to Reclamation in Mining (CASERM)' under Award No. 2310948.

Data Availability Statement: The data set used in this study is not publicly available due to confidentiality agreements between Virginia Tech and the iron ore mine.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Herrmann, H.; Bucksch, H. Exploration Data Analysis. *Dict. Geotech. Eng. Geotech.* **2014**, 493.
2. Church, R.M. How To Look At Data: A Review of John W. Tukey'S Exploratory Data Analysis 1. *J. Exp. Anal. Behav.* **1979**, *31*, 433–440. [CrossRef]
3. Komorowski, M.; Marshall, D.C.; Salciccioli, J.D.; Crutain, Y. *Secondary Analysis of Electronic Health Records*; Springer Nature: Basingstoke, UK, 2016; pp. 1–427. [CrossRef]
4. Wongsuphasawat, K.; Liu, Y.; Heer, J. Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study. *arXiv* **2019**, arXiv:1911.00568. [CrossRef]
5. Muraina, I.O.; Adesanya, O.M.; Agoi, M.A.; Abam, S.O. The Necessity of Exploratory Data Analysis How Are Preprocessing Activities Beneficial to Data Analysts and Professional Researchers in Academia. *Int. J. Sci. Res. Comput. Sci. Eng.* **2023**, *11*, 22–28. [CrossRef]
6. Oettl, F.C.; Oeding, J.F.; Feldt, R.; Ley, C.; Hirschmann, M.T.; Samuelsson, K. The Artificial Intelligence Advantage: Supercharging Exploratory Data Analysis. *Knee Surg. Sport. Traumatol. Arthrosc.* **2024**, *32*, 3039–3042. [CrossRef] [PubMed]
7. Putatunda, S.; Ubrangala, D.; Rama, K.; Kondapalli, R. SmartEDA: An R Package for Automated Exploratory Data Analysis. *J. Open Source Softw.* **2019**, *4*, 1509. [CrossRef]
8. FDOT Florida Method of Test for Measuring While Drilling (MWD) for Geotechnical Applications. 2022. pp. 1–27. Available online: https://fdotwww.blob.core.windows.net/sitefinity/docs/default-source/materials/administration/resources/library/publications/fstm/methods/fm5-625.pdf(accessed on 2 December 2025).
9. Isheyskiy, V.; Sanchidrián, J.A. Prospects of Applying MWD Technology for Quality Management of Drilling and Blasting Operations at Mining Enterprises. *Minerals* **2020**, *10*, 925. [CrossRef]
10. Isheyskiy, V.; Martinyskin, E.; Smirnov, S.; Vasilyev, A.; Knyazev, K.; Fatyanov, T. Specifics of MWD Data Collection and Verification during Formation of Training Datasets. *Minerals* **2021**, *11*, 798. [CrossRef]
11. Komadja, G.C.; Westman, E.; Rana, A.; Vitalis, A. A Machine Learning Approach to Lithology Classification in Mining Using Measurement While Drilling and Exploration Data. *Min. Metall. Explor.* **2025**, *42*, 1955–1973. [CrossRef]
12. Silversides, K.L.; Melkumyan, A. Machine Learning for Classification of Stratified Geology from MWD Data. *Ore Geol. Rev.* **2022**, *142*, 104737. [CrossRef]
13. Li, K.; Ren, T.; Yao, N.; Roberts, J.; Song, H.; Li, Z.; Liang, C. Real-Time Lithology Identification While Drilling Based on Drilling Parameters Analysis with Machine Learning. *Geomech. Geophys. Geo-Energy Geo-Resour.* **2025**, *11*, 44. [CrossRef]
14. Burak, T.; Sharma, A.; Hoel, E.; Kristiansen, T.G.; Welmer, M.; Nygaard, R. Real-Time Lithology Prediction at the Bit Using Machine Learning. *Geosciences* **2024**, *14*, 250. [CrossRef]
15. Anafo, I.; Ganguli, R.; Sarantsatsral, N. BoxRF: A New Machine Learning Algorithm for Grade Estimation. *Appl. Sci.* **2025**, *15*, 4416. [CrossRef]

16. Akyildiz, O.; Basarir, H.; Ellefmo, S.L. The Development of a Lithology Prediction Model Using Measurement While Drilling Data in a Quartzite Quarry. *Int. J. Min. Reclam. Environ.* **2025**, *39*, 93–109. [CrossRef]
17. Goldstein, D.; Aldrich, C.; O'Connor, L. Enhancing Orebody Knowledge Using Measure-While-Drilling Data: A Machine Learning Approach. *IFAC-PapersOnLine* **2024**, *58*, 72–76. [CrossRef]
18. Hansen, T.F.; Erharter, G.H.; Liu, Z.; Torresen, J. A Comparative Study on Machine Learning Approaches for Rock Mass Classification Using Drilling Data. *Appl. Comput. Geosci.* **2024**, *24*, 100199. [CrossRef]
19. Manzoor, S.; Liaghat, S.; Gustafson, A.; Johansson, D.; Schunnesson, H. Rock Mass Characterization Using Mwd Data and Photogrammetry. In *Mining Goes Digital, Proceedings of the 39th International Symposium 'Application of Computers and Operations Research in the Mineral Industry' (APCOM 2019), Wroclaw, Poland, 4–6 June 2019*; CRC Press: Boca Raton, FL, USA, 2019; pp. 217–225. [CrossRef]
20. Heydari, S.; Hoseinie, S.H.; Bagherpour, R. Prediction of Jumbo Drill Penetration Rate in Underground Mines Using Various Machine Learning Approaches and Traditional Models. *Sci. Rep.* **2024**, *14*, 8928. [CrossRef] [PubMed]
21. Wu, S.; Wang, X.; Yue, Z.Q. Addressing Random Variations in MWD Penetration Rate with the DPM Algorithm. *Sustainability* **2022**, *14*, 13456. [CrossRef]
22. Samanta, B.; Bandopadhyay, S.; Ganguli, R.; Dutta, S. Sparse Data Division Using Data Segmentation and Kohonen Network for Neural Network and Geostatistical Ore Grade Modeling in Nome Offshore Placer Deposit. *Nat. Resour. Res.* **2004**, *13*, 189–200. [CrossRef]