Integrating Machine Learning Techniques with Measurement-While-Drilling Data for Subsurface Characterization in Open-Pit Mines.

Jesse Addy

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

in

Mining Engineering

Erik Westman, PhD, Chair

Rohit Pandey, PhD

Nino Ripepi, PhD

December 3, 2025

Blacksburg, Virginia

Keywords: Measurement-While-Drilling (MWD); exploratory data analysis (EDA); supervised learning; unsupervised learning; void detection; lithology; penetration rate; open-pit mining.

Integrating Machine Learning Techniques with Measurement-While-Drilling Data for Subsurface Characterization in Open-Pit Mines.

Jesse Addy

ABSTRACT

Measurement-While-Drilling (MWD) systems generate continuous drilling data that reflect subsurface conditions in real time. With the increasing availability of this data, there is a growing opportunity to use data-driven methods to support geological interpretation and geotechnical risk assessment in mining. However, the complexity and variability of drilling signals require analytical workflows that go beyond traditional threshold-based interpretation.

This thesis integrates machine learning techniques with MWD data to improve subsurface characterization in two open-pit mines, each influenced by different operational and geological conditions. The first research component focuses on identifying zones of disturbed or weakened ground by detecting drilling behavior indicative of voids and compromised rock mass conditions in a mine affected by historic underground workings. The second component applies a structured data preparation and analysis workflow to develop predictive models for lithology and penetration rate in a separate open-pit operation, demonstrating how MWD data can support geological classification and drilling performance evaluation.

Across both studies, the research highlights the importance of exploratory data analysis (EDA), feature engineering, and appropriate model selection. The results show that machine learning offers a scalable and effective way to extract meaningful information from MWD data, enhancing both geotechnical hazard detection and geological modeling. These findings demonstrate the value of integrating modern data science methods into mining workflows, contributing to safer and more informed operational decision-making.

Integrating Machine Learning Techniques with Measurement-While-Drilling Data for Subsurface Characterization

Jesse Addy

GENERAL AUDIENCE ABSTRACT

Modern mines rely on drilling to understand what lies beneath the surface, but the data collected during drilling is often underused. This thesis explores how information recorded automatically during drilling known as Measurement-While-Drilling (MWD) data can be used to better understand ground conditions and make mining safer and more efficient.

Using data from two open-pit mines, this research applies computer-based learning methods to interpret drilling behavior. In the first case, the goal was to find hidden voids and weak areas underground, which can lead to unsafe working conditions if left undetected. In the second case, the same type of drilling data was used to identify different rock types and predict how fast the drill would advance, helping with planning and decision-making.

By combining careful data analysis with modern machine learning techniques, this work shows that drilling data can reveal far more than just depth or hardness. It can help predict geological conditions, highlight safety risks, and reduce uncertainty without slowing down mining operations.

This research demonstrates that mines can gain valuable insight simply by making better use of the data they already collect.

Dedication

Dedicated to:

My late dad

My mum

My grandma

My siblings

My aunties and uncles

I appreciate their love, prayers, patience, support, and sacrifices throughout my academic journey.

Acknowledgments

First, I give thanks to God for granting me good health, wisdom, and strength throughout this journey.

I am deeply grateful to my advisor, Dr. Erik Westman, for his mentorship, guidance, and unwavering support throughout my graduate studies. His encouragement, thoughtful feedback, and commitment to my development were instrumental in the successful completion of this work. I sincerely acknowledge CASERM Research for providing the funding, resources, and research environment that supported and advanced this study.

I would like to thank my committee members, Dr. Rohit Pandey and Dr. Nino Ripepi, for their time, valuable feedback, and constructive suggestions that helped strengthen this research. I am also thankful to the faculty, staff, and colleagues in the Virginia Tech Mining and Minerals Engineering Department, including Cemile Dilara Bag, with whom I was in the same research group, for fostering a supportive and collaborative academic environment. for fostering supportive and collaborative academic environment.

I am especially grateful to Dr. Festus Animah for his guidance and counsel during key moments of my graduate school journey. I also sincerely thank Ishmael Anafo, with whom I collaborated on one of my thesis projects, for his dedication and instrumental contributions that greatly supported the successful completion of this research.

I extend my sincere thanks to Stephen Kwabena Oppong and Joshua Ninepence for their mentorship and steadfast support behind the scenes throughout this journey. Their guidance, encouragement, and willingness to share their experiences contributed significantly to my personal and professional growth.

I would also like to express my heartfelt appreciation to my uncle, Francis Odoom, and his wife, Loreen Wutoh, whose support was instrumental in helping me navigate challenges and settle seamlessly upon my arrival in the United States. Their encouragement to pursue opportunities here, along with their unwavering support, played a vital role in my academic and personal journey.

Finally, I thank my family and friends for their unwavering encouragement, patience, and belief in me throughout this process.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

MWD Measurement-While-Drilling

EDA Exploratory Data Analysis

ROP Rate of Penetration

WOB Weight on Bit

RPM Revolutions Per Minute

MSE Mechanical Specific Energy

ML Machine Learning

MLP Multi-Layer Perceptron

RF Random Forest

XGBoost Extreme Gradient Boosting

IF Isolation Forest

UCS Uniaxial Compressive Strength

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

Chapter 1: General Introduction

1.1 Background

Measurement-While-Drilling (MWD) systems collect real-time drilling data such as penetration rate, torque, rotation speed, air pressure, and weight-on-bit. These parameters are increasingly used in mining for operational monitoring, material characterization, and geotechnical assessment [1]. However, the usefulness of MWD data depends on the ability to correctly interpret drilling responses under varying subsurface conditions [1], [2].

In open-pit mining environments with historical underground workings, voids and weakened ground pose safety risks including ground collapse, equipment instability, and unplanned downtime. Conventional detection methods provide limited spatial coverage and are not suitable for real-time hazard identification [3]. The first manuscript in this thesis presents a machine learning framework for detecting void-prone zones from production drilling data. The approach combines feature engineering, unsupervised anomaly detection, and supervised classification to identify intervals associated with disturbed ground conditions. This workflow demonstrates how MWD-based drilling signatures may be used to support geotechnical hazard identification in areas affected by legacy underground workings.

A related challenge is the development of machine learning models that can classify lithology and predict penetration rate using MWD data. Model performance depends on the consistency, structure, and quality of the input data [4]. Exploratory Data Analysis (EDA) plays a critical role in understanding data behavior, identifying outliers, and selecting reliable features [5]. The second manuscript demonstrates how EDA improves model accuracy and interpretability. Using structured data exploration and filtering, the study develops machine learning models that achieve high performance in lithology classification and penetration rate prediction.

Together, the two manuscripts show that MWD data can be used both to detect hazardous ground conditions and to support geological modeling. Both applications require a data-driven workflow that integrates domain knowledge, feature analysis, and machine learning.

1.2 Objectives

This thesis aims to:

i. Develop a machine-learning-based method for detecting void-prone zones using MWD data to support geotechnical hazard identification.
ii. Demonstrate the role of exploratory data analysis in preparing drilling data for predictive modeling.
iii. Evaluate machine learning models for lithology classification and penetration rate prediction using processed MWD data.

1.3 Thesis Structure

The thesis is organized as follows:

i. Chapter 1 presents the background, motivation, and research objectives.
ii. Chapter 2 provides a review of literature covering MWD systems, drilling analytics, void detection, exploratory data analysis, and machine learning.
iii. Chapter 3 contains the first manuscript, focused on detection of void-prone zones using MWD data.
iv. Chapter 4 contains the second manuscript, focused on exploratory data analysis and predictive modeling.
v. Chapter 5 concludes the thesis with findings and future research directions.

1.4 References

[1] H. Schunnesson, "Rock characterisation using percussive drilling," *International Journal of Rock Mechanics and Mining Sciences*, vol. 35, no. 6, pp. 711–725, Sept. 1998, doi: 10.1016/S0148-9062(97)00332-X.

[2] V. Isheyskiy and J. A. Sanchidrián, "Prospects of Applying MWD Technology for Quality Management of Drilling and Blasting Operations at Mining Enterprises," *Minerals*, vol. 10, no. 10, p. 925, Oct. 2020, doi: 10.3390/min10100925.

[3] Mines Occupational Safety and Health Advisory Board (Western Australia), "Open Pit Mining Through UG Workings." State of Western Australia, 2000.

[4] A. Abbaszadeh Shahri, C. Shan, S. Larsson, and F. Johansson, “Normalizing Large Scale Sensor-Based MWD Data: An Automated Method toward A Unified Database,” *Sensors*, vol. 24, no. 4, p. 1209, Feb. 2024, doi: 10.3390/s24041209.

[5] I. O. Muraina *et al.*, “The Necessity of Exploratory Data Analysis How are preprocessing activities beneficial to Data Analysts and Professional Researchers in Academia,” *IJSRCSE*, vol. 11, no. 3, pp. 22–28, June 2023, doi: 10.26438/ijsrcse/v11i3.2228.

Chapter 2 : Review of Relevant Literature

2.1 Measurement-While-Drilling Systems

Measurement-While-Drilling systems record drilling parameters in real time, providing continuous data on subsurface conditions during blasthole drilling. These systems typically measure parameters such as penetration rate, rotation pressure, weight on bit, torque, feed force, flushing pressure, and depth, enabling quantification of drilling behavior as the rock mass is penetrated [1]. Modern MWD systems integrate sensors, onboard logging, and digital communication, allowing fully automated data capture without interrupting production .

MWD has become widely adopted in mining because it delivers spatially continuous information at a higher resolution than traditional geological sampling. It is used to improve blast design, monitor drilling performance, classify rock types, and support ore control models [2]. In open-pit operations, MWD has proven useful for identifying material variability, estimating blastability, and even predicting chemical properties when paired with machine learning workflows [3].

The fundamental value of MWD lies in the relationship between drilling responses and rock mass properties. Numerous studies show that penetration rate, torque, and drilling pressures correlate strongly with rock hardness and strength . Beyond hardness, drilling parameters have been used to infer rock mass quality and even support selection of tunnel support systems, demonstrating the flexibility of MWD in both mining and civil engineering contexts [4]. MWD has also been applied for geotechnical risk detection - for example, identifying overbreak-prone zones by linking anomalous drilling behavior to weak or fractured rock [5].

Although MWD provides valuable data, the raw signals often include noise, pauses caused by drill rod changes, collaring, and other operational artifacts. As a result, data preprocessing has become an important component of MWD analysis. Recent work emphasizes filtering and normalization to ensure that recorded signals reflect actual drilling and not machine-related effects [6]. After preprocessing, MWD data can be reliably used for large-scale rock mass characterization, structural interpretation, and predictive modeling.

The technology now enables field-scale, drill-pattern-wide characterization workflows. Studies have demonstrated that MWD datasets can be analyzed to automatically infer rock conditions, detect structural changes, and assess ground behavior in real time, making MWD a cornerstone of data-driven mining.

2.2 Previous Studies Using MWD Data in Mining

Several studies have demonstrated the value of MWD data for subsurface interpretation in mining environments. Khorzoughi showed that variations in drilling parameters could be linked to changes in rock mass quality, providing early evidence that MWD could be used for geotechnical characterization beyond operational monitoring [7].

Lithological classification using MWD data has been a major theme in later studies. Li et al. demonstrated that drilling responses captured during blasthole drilling can be used to identify lithological boundaries in real time, improving geological mapping and enabling more responsive mine planning [8]. Other researchers have confirmed that MWD parameters reflect differences in rock conditions and can be used to infer geological contacts at high spatial resolution. Mechanical property estimation is another important application. Liaghat et al. applied drilling data to predict rock hardness, confirming that drilling forces and penetration rate are effective indicators of strength-related behavior [9].

MWD has also been used to detect zones of structural weakness or overbreak risk. Lundberg and Saiang demonstrated that abnormal drilling behavior can indicate broken or fractured ground, enabling early identification of unstable zones in tunneling environments [5]. Similar approaches have been applied to open-pit blasting. Isheyskiy and Sanchidrián showed that MWD-derived information can improve blast quality control and reduce adverse outcomes such as poor fragmentation [2]. Navoyski et al. extended this work by using MWD-based models to estimate blast damage profiles [10].

Because raw MWD signals may contain noise and operational artifacts, preprocessing techniques have also been developed. Navarro et al. proposed filtering and normalization to ensure that variations in MWD responses reflect geological rather than mechanical effects [6].

These studies collectively demonstrate that MWD is a mature and versatile tool for lithological interpretation, rock strength estimation, and geotechnical hazard recognition.

2.3 Machine Learning Methods

Machine learning (ML) refers to a class of computational methods that learn patterns from data and make predictions or decisions with minimal human intervention. ML approaches are commonly grouped into supervised, unsupervised, semi-supervised, and reinforcement learning, depending on the availability of labeled data and learning goals [11]. The different learning categories commonly used in machine learning are illustrated in Figure 2.1.

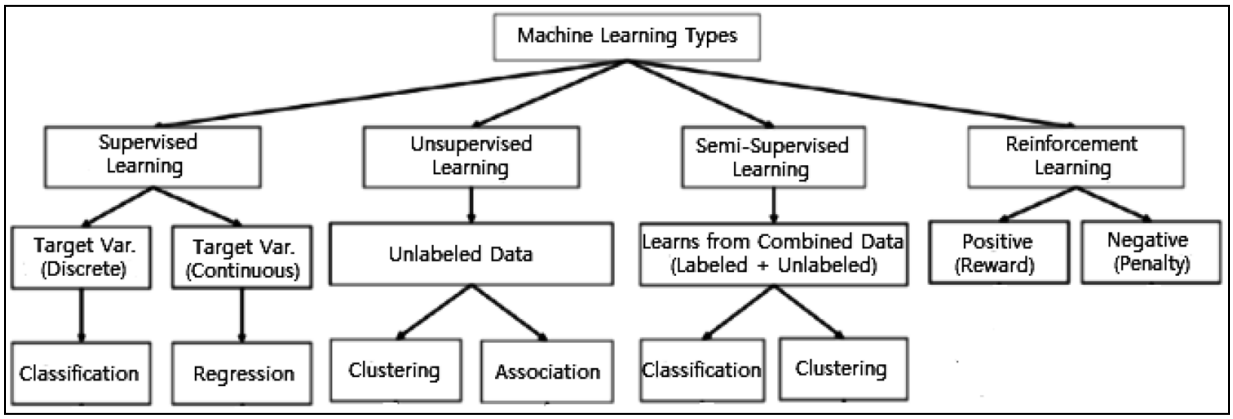

Figure 0.1 Types of Machine Learning Algorithms
Adapted from: "Types of Machine Learning Algorithms," 2020

Supervised learning is used when labeled examples are available and the goal is to predict a target variable. It includes both classification (discrete output) and regression (continuous output) tasks. Algorithms such as logistic regression, random forests, multilayer perceptron (MLP), and gradient-boosted decision trees are widely used examples [11].

Unsupervised learning operates on unlabeled data, discovering structure through methods such as clustering or density estimation. Algorithms like k-means, isolation forest, and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) identify groups or anomalies without known target values [12]. These approaches are valuable when ground-truth labels are scarce or unavailable.

Semi-supervised learning bridges supervised and unsupervised paradigms by exploiting both labeled and unlabeled data to improve generalization. This approach is particularly useful in real-world conditions where labeling is expensive, time-consuming, or incomplete [13].

A general workflow for machine learning includes problem formulation, data preprocessing, training–testing split, algorithm selection, model training, hyperparameter optimization, and performance evaluation. This process is iterative, often requiring multiple tuning cycles before a model is finalized [11]. This workflow is shown in Figure 2.2.

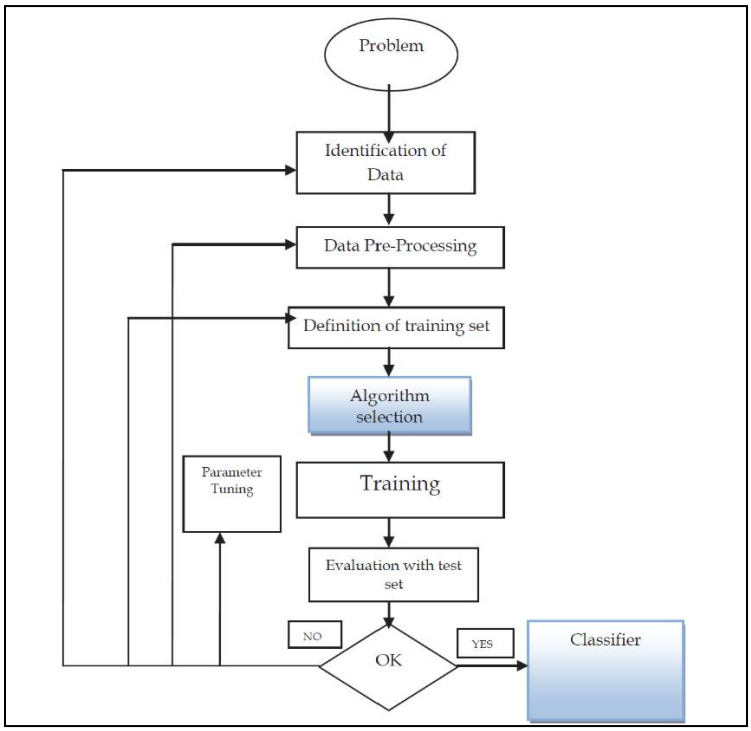

Figure 0.2 General Machine Learning Workflow
Adapted from: "Machine Learning Algorithms: Real World Applications and Research Directions," 2018

The machine learning algorithms employed in this thesis are described in the following subsections. Each subsection introduces the core concept of the algorithm, presents its basic formulation, and explains its relevance to MWD data and the specific application within this research.

2.3.1 Logistic Regression

Logistic Regression is a supervised classification algorithm that models the probability of class membership as a function of a linear combination of input features transformed through a sigmoid function:

$$p(y = 1|x) = \frac{1}{1 + \exp[-(\beta_0 + \beta^T x)]}$$

where x is the feature vector and β represents the model coefficients [14].

Logistic Regression is computationally efficient and interpretable, making it suitable as a baseline classifier. In this thesis, it is used as a reference model for void-prone interval classification and for examining the direction and relative magnitude of feature influence on drilling behavior.

2.3.2 Multilayer Perceptron (MLP)

A Multilayer Perceptron is a feed-forward artificial neural network composed of interconnected layers of neurons capable of learning nonlinear decision boundaries. For a given layer l, the forward propagation is:

$$a^{(l)} = f\left(W^{(l)} a^{(l-1)} + b^{(l)}\right)$$

where $f(\cdot)$ is a nonlinear activation function, and parameters are optimized using backpropagation [15].

MLPs can model complex nonlinear relationships but require careful tuning to avoid overfitting. In this work, MLP models are evaluated for lithology classification and penetration-rate prediction to assess whether increased model complexity improves performance relative to tree-based methods.

2.3.3 Random Forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees trained on bootstrapped subsets of the data with randomized feature selection. For classification, predictions are obtained by majority voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_M(x)\}$$

and for regression by averaging individual tree outputs:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m\,(x)$$

[16].

Random Forest models are robust to noise, capture nonlinear feature interactions, and provide feature importance metrics. These properties make them well suited to heterogeneous and operationally noisy MWD datasets. Random Forest serves as a primary supervised model in this thesis for lithology classification, penetration-rate prediction, and void-prone interval detection.

2.3.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient-boosted decision tree algorithm that builds trees sequentially, with each new tree correcting errors from previous iterations through gradient-based optimization. The model minimizes a regularized objective function:

$$\mathcal{L} = \sum_i l\,(y_i, \hat{y}_i) + \sum_k \Omega\,(f_k), \qquad \Omega(f) = \gamma T + \frac{\lambda}{2} \parallel w \parallel^2$$

where $l(\cdot)$ is the loss function and $\Omega(\cdot)$ penalizes model complexity [17].

XGBoost is particularly effective for structured tabular data and imbalanced classification problems. In this thesis, it is applied to void-prone interval classification where high recall is critical for minimizing missed detections.

2.3.5 Isolation Forest (IF)

Isolation Forest is an unsupervised anomaly detection algorithm based on the principle that anomalous observations are easier to isolate than normal data points. Anomaly scores are derived from the average path length $E[h(x)]$ within randomly generated isolation trees:

$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}}$$

where $c(n)$ is the expected path length for a dataset of size n [18].

Isolation Forest does not assume any underlying data distribution and performs well in high-dimensional settings. In this research, it is the primary method used to identify anomalous drilling behavior indicative of disturbed or void-prone ground in the absence of comprehensive ground-truth labels.

2.3.6 HDBSCAN

HDBSCAN is a density-based clustering algorithm that extends DBSCAN by allowing clusters of variable density and explicitly identifying noise points. A key concept is the mutual reachability distance:

$$d_{\text{mreach}}(a, b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a, b)\}$$

Clusters are extracted from a hierarchical structure based on stability [19].

HDBSCAN is evaluated in this thesis as an alternative unsupervised approach for detecting localized drilling anomalies within MWD datasets.

2.3.7 K-Means Clustering

K-means is an unsupervised partitioning algorithm that groups data into a predefined number of clusters by minimizing within-cluster variance:

$$\arg\min_{\{C_i\}_{i=1}^{k}} \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2$$

where μ_i is the centroid of cluster C_i [11].

Although sensitive to the choice of k, K-means provides a useful exploratory baseline. In this thesis, it is used to assess whether natural groupings exist in the MWD feature space prior to supervised modeling.

2.3.8 Summary

The combination of supervised and unsupervised algorithms employed in this thesis addresses the key challenges of MWD data, including nonlinearity, operational noise, and limited labeled observations. Unsupervised methods are used to identify anomalous drilling behavior, while supervised models generalize these patterns across large production datasets to support geotechnical hazard detection and predictive modeling.

2.4 Machine Learning Applications to MWD Data in Mining

Machine learning has become a central tool for extracting geological and geotechnical information from MWD data. As MWD produces large, continuous datasets that are impractical to manually

interpret, ML methods enable automated classification, prediction, and detection of rock mass behavior. Early applications focused on lithological prediction using drilling responses, with models demonstrating that supervised learning can outperform threshold-based or rule-based systems [20].

Subsequent research expanded ML applications to include rock mass characterization, ore–waste discrimination, and rock property estimation. For instance, Klyuchnikov et al. applied logistic regression, gradient boosting, and neural networks to distinguish shale-rich rock from other units, with boosting methods achieving the highest accuracy [21]. Similarly, Silversides and Melkumyan compared neural networks, Gaussian processes, and boosting models for stratified deposits, showing that multiple ML methods can successfully classify geological units when trained on MWD parameters.

Beyond rock type prediction, ML has been used to infer mechanical properties such as uniaxial compressive strength (UCS), hardness, and rock mass quality. A comparative study demonstrated that random forest models outperform linear methods such as regression when predicting UCS from MWD data [22]. In marble mining, logistic regression and random forest were evaluated for predicting material quality classes; the latter achieved significantly better performance, highlighting the value of non-linear ensemble methods [23].

More recent studies combine unsupervised and supervised approaches. Komadja et al. integrated clustering with extreme XGBoost to group and classify rock strength classes with accuracy exceeding 98% [24]. This synergistic workflow is particularly useful in cases where labeled data are limited, and it closely aligns with the methods used in Manuscript 1 of this thesis.

Machine learning has also been applied to data cleaning and preprocessing. Huang et al. showed that algorithms such as Isolation Forest, One-Class SVM, and DBSCAN outperform traditional statistical filters when removing anomalies from drilling data, improving downstream model performance, and reducing manual processing effort [25].

Collectively, these studies demonstrate that machine learning enhances the interpretive power of MWD systems supporting lithology classification, rock mass property estimation, and automated

data cleaning. These advancements form an important foundation for the work presented in this thesis, which extends ML applications to void detection and data-driven lithological modeling.

2.5 Exploratory Data Analysis in MWD-Based Subsurface Characterization

Exploratory Data Analysis (EDA) was originally introduced by Tukey as a systematic approach for examining data prior to formal modeling, with the aim of revealing structure, identifying anomalies, and developing intuition about variable behavior through visualization and summary statistics [26]. Rather than testing predefined hypotheses, EDA emphasizes flexible and iterative exploration, allowing analysts to understand what the data contains and how they behave before imposing statistical or predictive models. This philosophy established EDA as a foundational step in data analysis, especially for complex datasets where underlying assumptions are not known in advance.

Subsequent literature has expanded EDA into a more structured analytical phase that bridges data collection and modeling. Komorowski et al. describe EDA as an essential step following preprocessing, during which distributions, outliers, and relationships among variables are examined to assess data quality and guide hypothesis generation [27]. This framing positions EDA not as an informal activity, but as a deliberate methodological stage that directly informs feature selection, transformation, and model design. The authors emphasize that many modeling failures can be traced to inadequate exploration of data characteristics early in the analytical workflow.

Empirical studies of analyst practice further clarify the goals and challenges of EDA. Wongsuphasawat et al. identify two dominant objectives of exploratory analysis: profiling, which focuses on understanding data quality and structure, and discovery, which aims to uncover new patterns and insights [28]. Their findings show that profiling activities such as checking distributions, missing values, and variable ranges occur in nearly all analyses, regardless of domain. This highlights the importance of EDA for validating data integrity before higher-level interpretation or modeling.

As datasets grow in size and dimensionality, tools have been developed to support and streamline exploratory workflows. Software frameworks such as EDAssistant and SmartEDA provide structured ways to summarize data, generate visualizations, and examine variable relationships within interactive computing environments [29], [30]. While these tools improve efficiency and

reproducibility, the literature consistently emphasizes that EDA remains an interpretive process that relies on analyst judgment. Automated summaries do not replace the need for domain knowledge when interpreting patterns or distinguishing meaningful signals from artifacts.

Recent work has also examined how emerging computational methods can complement traditional exploratory practices. Reviews of exploratory workflows note that advanced statistical and algorithmic techniques can assist in highlighting correlations, grouping behavior, and variable importance during EDA, but must be applied cautiously and interpreted within context [31]. The emphasis remains on EDA as a preparatory stage that informs, rather than replaces, subsequent analysis.

In applied engineering and geoscience contexts, EDA plays a critical role in understanding sensor-based and time-series datasets. In coal seam gas wells, exploratory data analytics have been used to examine multivariate time-series behavior, identify anomalies, and guide feature extraction prior to further analysis [32]. These studies demonstrate that EDA is particularly valuable for operational datasets affected by noise, missing values, and heterogeneous sampling, where raw measurements cannot be directly interpreted without careful exploration.

Exploratory techniques have also been applied to high-dimensional scientific datasets outside traditional engineering domains. In ocean-world analog mass spectrometry studies, EDA has been used to characterize correlation structure, identify dominant sources of variance, and detect irregularities before dimensionality reduction or clustering [33]. Such applications reinforce the role of EDA as a general-purpose methodology for understanding complex systems when ground-truth labels are limited or unavailable.

Within mining and drilling applications, EDA has been widely employed to interpret Measurement-While-Drilling (MWD) data. Studies using MWD parameters demonstrate that exploratory analysis of penetration rate, torque, weight-on-bit, and pressure measurements can reveal lithological variability and drilling response trends prior to predictive modeling [34]. These approaches rely on visual inspection, statistical summaries, and correlation analysis to establish baseline behavior and identify departures associated with changes in ground conditions.

Across these domains, the literature consistently shows that EDA is not merely descriptive, but foundational to robust analysis. Decisions made during exploratory analysis shape how data are

cleaned, transformed, and modeled, directly affecting the validity and interpretability of results [35]. Despite its importance, EDA is often treated implicitly in applied studies, with limited documentation of the exploratory reasoning that precedes modeling. This thesis addresses that gap by explicitly foregrounding EDA as a core methodological component, ensuring that subsequent unsupervised and supervised analyses are grounded in a transparent and physically informed understanding of the data.

2.6 Summary and Research Gap

MWD technology provides continuous, high-resolution drilling data that reflects subsurface variability at a scale unmatched by traditional sampling methods. Previous studies have demonstrated the value of MWD parameters such as penetration rate, torque, weight-on-bit, and drilling pressures for characterizing rock hardness, estimating rock mass quality, improving blasting efficiency, and supporting operational decision-making in mining environments. More recent work has expanded these applications to include lithology classification, strength prediction, and structural interpretation using data-driven techniques.

Machine learning has become a central approach for extracting value from MWD data. Supervised models such as Random Forests, gradient boosting methods, and neural networks have been applied to predict lithology, uniaxial compressive strength, and rock mass rating. Unsupervised techniques, including clustering and anomaly detection, have been used to identify discontinuities, weak zones, and geotechnically significant features. In some cases, hybrid workflows incorporating feature engineering, semi-supervised learning, and domain-specific preprocessing have been explored to improve model performance.

Despite these advances, several gaps remain in existing literature. Many studies emphasize predictive performance while giving limited attention to the exploratory analysis required to understand data quality, parameter behavior, and operational variability prior to modeling. Assumptions of clean, uniformly sampled data are common, with limited consideration of operational noise, drilling pauses, collaring intervals, or inconsistent data structure inherent in production drilling environments. Furthermore, most work focuses on either geological classification or mechanical property estimation, with comparatively few studies addressing geotechnical hazard detection, particularly in the context of legacy underground workings and

sparse ground-truth information. Direct comparisons between unsupervised anomaly-based approaches and supervised models trained on limited verified labels are also rare.

This thesis addresses these gaps by developing two complementary MWD-based machine learning frameworks grounded in systematic exploratory data analysis. The first framework focuses on identifying void-prone drilling conditions using both unsupervised and supervised learning approaches. The second demonstrates how rigorous exploratory data analysis improves the performance, robustness, and interpretability of lithology classification and penetration rate prediction models. Together, these contributions advance the use of MWD data as a reliable subsurface characterization tool for both geotechnical hazard assessment and geological modeling in modern open-pit mining.

2.7 References

[1] P. Rai, H. Schunesson, P.-A. Lindqvist, and U. Kumar, "An Overview on Measurement-While-Drilling Technique and its Scope in Excavation Industry," *J. Inst. Eng. India Ser. D*, vol. 96, no. 1, pp. 57–66, Apr. 2015, doi: 10.1007/s40033-014-0054-4.

[2] V. Isheyskiy and J. A. Sanchidrián, "Prospects of Applying MWD Technology for Quality Management of Drilling and Blasting Operations at Mining Enterprises," *Minerals*, vol. 10, no. 10, p. 925, Oct. 2020, doi: 10.3390/min10100925.

[3] R. N. Khushaba, A. Melkumyan, and A. J. Hill, "A Machine Learning Approach for Material Type Logging and Chemical Assaying from Autonomous Measure-While-Drilling (MWD) Data," *Math Geosci*, vol. 54, no. 2, pp. 285–315, Feb. 2022, doi: 10.1007/s11004-021-09970-w.

[4] J. Van Eldert, H. Schunnesson, D. Johansson, and D. Saiang, "Application of Measurement While Drilling Technology to Predict Rock Mass Quality and Rock Support for Tunnelling," *Rock Mech Rock Eng*, vol. 53, no. 3, pp. 1349–1358, Mar. 2020, doi: 10.1007/s00603-019-01979-2.

[5] J. Navarro, J. A. Sanchidrián, P. Segarra, R. Castedo, E. Costamagna, and L. M. López, "Detection of potential overbreak zones in tunnel blasting from MWD data," *Tunnelling and Underground Space Technology*, vol. 82, pp. 504–516, Dec. 2018, doi: 10.1016/j.tust.2018.08.060.

[6] J. Van Eldert, H. Schunnesson, D. Saiang, and J. Funehag, “Improved filtering and normalizing of Measurement-While-Drilling (MWD) data in tunnel excavation,” *Tunnelling and Underground Space Technology*, vol. 103, p. 103467, Sept. 2020, doi: 10.1016/j.tust.2020.103467.

[7] M. B. Khorzoughi, “USE OF MEASUREMENT WHILE DRILLING TECHNIQUES FOR IMPROVED ROCK MASS CHARACTERIZATION IN OPEN-PIT MINES”.

[8] K. Li *et al.*, “Real-time lithology identification while drilling based on drilling parameters analysis with machine learning,” *Geomech. Geophys. Geo-energ. Geo-resour.*, vol. 11, no. 1, p. 44, Dec. 2025, doi: 10.1007/s40948-025-00951-5.

[9] S. Liaghat, S. H. Hoseinie, and N. Al Ansari, “Predicting Rock Hardness Using Measurement While Drilling (MWD) Data in a Case Study of Rotary Drilling in an Open-Pit Mine,” *Indian Geotech J*, Aug. 2025, doi: 10.1007/s40098-025-01337-w.

[10] “Measurement While Drilling technology for blasting damage calculation.”

[11] B. Mahesh, “Machine Learning Algorithms - A Review,” *IJSR*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/ART20203995.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sept. 1999, doi: 10.1145/331499.331504.

[13] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Mach Learn*, vol. 109, no. 2, pp. 373–440, Feb. 2020, doi: 10.1007/s10994-019-05855-6.

[14] H.-A. Park, “An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain,” *J Korean Acad Nurs*, vol. 43, no. 2, p. 154, 2013, doi: 10.4040/jkan.2013.43.2.154.

[15] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, “Multilayer Perceptron and Neural Networks”.

[16] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, June 2024, doi: 10.58496/BJML/2024/007.

[17] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A Comparative Analysis of XGBoost,” 2019, doi: 10.48550/ARXIV.1911.01914.

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” in *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy: IEEE, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[19] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, “A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases,” *IJCA*, vol. 3, no. 6, pp. 1–4, June 2010, doi: 10.5120/739-1038.

[20] O. Akyildiz, H. Basarir, and S. L. Ellefmo, “The development of a lithology prediction model using measurement while drilling data in a quartzite quarry,” *International Journal of Mining, Reclamation and Environment*, vol. 39, no. 2, pp. 93–109, Feb. 2025, doi: 10.1080/17480930.2024.2362577.

[21] T. F. Hansen, G. H. Erharter, Z. Liu, and J. Torresen, “A comparative study on machine learning approaches for rock mass classification using drilling data,” *Applied Computing and Geosciences*, vol. 24, p. 100199, Dec. 2024, doi: 10.1016/j.acags.2024.100199.

[22] Y. Xie, X. Li, and Z. Min, “Comparison of machine learning models for rock UCS prediction using measurement while drilling data,” *Sci Rep*, vol. 15, no. 1, p. 8434, Mar. 2025, doi: 10.1038/s41598-025-93111-4.

[23] O. Akyildiz, H. Basarir, V. S. Vezhapparambu, and S. Ellefmo, “MWD Data-Based Marble Quality Class Prediction Models Using ML Algorithms,” *Math Geosci*, vol. 55, no. 8, pp. 1059–1074, Nov. 2023, doi: 10.1007/s11004-023-10061-1.

[24] G. C. Komadja, E. Westman, A. Rana, and A. Vitalis, “Predicting rock mass strength from drilling data using synergistic unsupervised and supervised machine learning approaches,” *Earth Sci Inform*, vol. 18, no. 3, p. 325, Sept. 2025, doi: 10.1007/s12145-025-01837-6.

[25] F. Huang, H. Qin, M. Manafi, B. Juett, and B. Evans, “Machine learning approaches for automatic cleaning of investigative drilling data.

[26] “Exploratory-Data-Analysis-1977-John-Tukey.”

[27] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, “Exploratory Data Analysis,” in Secondary Analysis of Electronic Health Records, Cham: Springer International Publishing, 2016, pp. 185–203. doi: 10.1007/978-3-319-43742-2_15.

[28] K. Wongsuphasawat, Y. Liu, and J. Heer, “Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study,” Nov. 01, 2019, arXiv: arXiv:1911.00568. doi: 10.48550/arXiv.1911.00568.

[29] X. Li, Y. Zhang, J. Leung, C. Sun, and J. Zhao, “EDAssistant: Supporting Exploratory Data Analysis in Computational Notebooks with In-Situ Code Search and Recommendation,” Dec. 15, 2021, arXiv: arXiv:2112.07858. doi: 10.48550/arXiv.2112.07858.

[30] S. Putatunda, K. Rama, D. Ubrangala, and R. Kondapalli, “SmartEDA: An R Package for Automated Exploratory Data Analysis,” JOSS, vol. 4, no. 41, p. 1509, Sept. 2019, doi: 10.21105/joss.01509.

[31] F. C. Oettl et al., “The artificial intelligence advantage: Supercharging exploratory data analysis,” Knee surg. sports traumatol. arthrosc., vol. 32, no. 11, pp. 3039–3042, Nov. 2024, doi: 10.1002/ksa.12389.

[32] F. Saghir, M. E. Gonzalez Perdomo, and P. Behrenbruch, “Application of Exploratory Data Analytics EDA in Coal Seam Gas Wells with Progressive Cavity Pumps PCPs,” in SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Bali, Indonesia: SPE, Oct. 2020, p. D031S032R002. doi: 10.2118/196528-MS.

[33] V. Da Poian et al., “Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry,” Front. Astron. Space Sci., vol. 10, p. 1134141, May 2023, doi: 10.3389/fspas.2023.1134141.

[34] O. Akyildiz, H. Basarir, V. S. Vezhapparambu, and S. Ellefmo, “MWD Data-Based Marble Quality Class Prediction Models Using ML Algorithms,” Math Geosci, vol. 55, no. 8, pp. 1059–1074, Nov. 2023, doi: 10.1007/s11004-023-10061-1.

[35] I. O. Muraina et al., “The Necessity of Exploratory Data Analysis How are preprocessing activities beneficial to Data Analysts and Professional Researchers in Academia,” IJSRCSE, vol. 11, no. 3, pp. 22–28, June 2023, doi: 10.26438/ijsrcse/v11i3.2228.

Chapter 3: Detecting Void-Prone Zones Near Historic Underground Workings in an Open-Pit Mine Using MWD Data and Machine Learning

3.1 Abstract

Historical underground workings pose significant geotechnical risks to open-pit mining, including ground collapse, equipment instability, and unsafe working conditions. Traditional detection methods provide limited coverage and lack real-time capability. This study presents a data-driven framework for identifying void-prone zones using Measurement-While-Drilling (MWD) data collected from production drill patterns. MWD parameters including rate of penetration, torque, rotary speed, weight-on-bit, and air pressure were preprocessed and transformed into engineered features that capture drilling responses sensitive to weak or disturbed ground. Unsupervised learning methods were applied to characterize anomalous behavior, with Isolation Forest producing the most coherent spatial anomaly patterns and strongest cluster separation. These pseudo-labels were then used to train supervised models for large-scale classification of void-like drilling intervals. For independent validation, a limited set of confirmed void intervals derived from downhole video logs was spatially matched to MWD coordinates and used to train separate supervised models. While performance on this small, highly imbalanced dataset was lower, this outcome is expected given the limited number of confirmed void intervals, the strong class imbalance between void-like and non-void-like samples, and the uncertainty associated with mapping sparse downhole video observations to depth-continuous MWD records. These factors reduce the ability of supervised models to learn stable decision boundaries and increase sensitivity to local drilling variability. Despite these limitations, the models reproduced the general drilling-response trends observed in the pseudo-labeled workflow. Overall, the results demonstrate that machine learning applied to MWD data can effectively delineate void-prone areas across the bench and provide a scalable tool for improving geotechnical hazard identification in regions influenced by historical underground workings.

Keywords: Measurement-While-Drilling; machine learning; void-prone zones; historical underground workings; geotechnical risk: supervised learning; unsupervised learning.

3.2 Introduction

Open-pit mines advancing toward legacy underground workings face significant uncertainty in ground conditions, drilling performance, and blasting outcomes. Historical stopes, development drifts, and partially backfilled voids can create abrupt changes in confinement, rock mass strength, and drilling resistance that pose elevated geotechnical and operational risks. Numerous case studies have shown that unverified or inaccurately mapped underground voids have contributed to equipment instability, unexpected ground collapse, and poor blast fragmentation in active surface mines [1], [2]. As a result, the ability to reliably detect disturbed or void-prone zones ahead of excavation remains a critical requirement for safe and cost-effective mine planning.

Geophysical methods such as electrical resistivity tomography, GPR, microgravity surveying, and borehole radar have been used to identify abandoned underground voids in a range of mining environments [3], [4], [5], [6], [7], [8]. While these techniques can provide valuable information, their effectiveness is often limited by survey coverage, terrain constraints, signal interference, and the cost of repeated deployment. Operational mines therefore increasingly seek continuous, drilling-based indicators, particularly those collected during routine blast-hole drilling to complement or replace geophysical investigations. Variations in feed pressure, torque, flushing behavior, rotary speed, and rate of penetration frequently accompany drilling through weak, broken, or voided ground [9], [10], providing practical real-time markers of geotechnical anomalies.

MWD systems offer an efficient means to capture these drilling responses at high resolution. By recording depth-indexed parameters such as torque, ROP, WOB, and air pressure, MWD data provide direct insight into bit–rock interaction and rock mass conditions. Numerous studies have demonstrated the potential of MWD data to infer rock strength, characterize geotechnical domains, and identify disturbed ground [10], [11], [12], [13], [14]. Ongoing research continues to refine MWD data preprocessing, parameter normalization, and feature engineering to improve interpretability and reduce the impact of noise, machine variability, and operational changes [11], [14].

Beyond operational geology, machine learning has become increasingly important in extracting meaningful patterns from MWD datasets. Applications range from supervised classification of drilling conditions and rock types [15] to anomaly detection in multivariate drilling signals [16],

to broader orebody-knowledge enhancement workflows [17]. Ensemble-based ML developments in mining such as the BoxRF algorithm, which improves grade-estimation performance through robust box-based partitioning and hybrid decision structures demonstrate the potential of advanced classifiers to model complex, high-variability datasets typical of production drilling [18]. These approaches highlight the growing value of combining engineered drilling features with modern ML architectures for anomaly detection and geotechnical interpretation. Field-scale studies further stress the importance of variable standardization and cross-pattern consistency to ensure generalization across mining environments [12].

Despite these advancements, reliably detecting void-prone or disturbed zones using MWD data remains challenging. Drilling responses are influenced by natural geological heterogeneity, operational changes, and equipment dynamics, making it difficult to separate meaningful anomalies from normal variability. Traditional threshold-based interpretation often fails to generalize across benches or drill pattern, [9], [10]. Acoustic-based monitoring approaches have shown promise in identifying sudden changes in confinement [8], but such systems require specialized sensors not deployed in most production rigs. Therefore, there is a growing need for integrated, data-driven workflows capable of capturing the complexity of multivariate drilling behavior, detecting anomalies at fine spatial scales, and validating predictions with independent sources of information.

This study addresses these gaps by developing a machine-learning–enabled workflow for detecting drilling intervals potentially influenced by historical underground workings using routine production MWD data. The framework integrates mechanical-response features, unsupervised anomaly detection, supervised classification, and validation from downhole video logs. The goal is to evaluate how MWD parameters behave in void-affected or disturbed ground and to assess the capability of machine-learning methods to map such intervals across multiple benches and drilling patterns. The findings provide practical insight into detecting weak or void-prone ground in active open-pit operations and demonstrate how MWD-driven analytics can enhance orebody knowledge, improve blast design, and support geotechnical risk management.

3.3 Materials and Methods

3.3.1 Study Area and Dataset Overview

The data used in this study were collected from a large open-pit mining operation located in North America, where current phases of pit expansion are approaching areas underlain by historical underground workings. These legacy excavations consisting of old stopes, development drifts, and partially backfilled voids introduce uncertainty in local ground conditions and may influence drilling performance, blast execution, and overall geotechnical risk. As mining progresses toward these zones, operators increasingly rely on MWD information as a practical means of characterizing the rock mass and identifying intervals where drilling behavior deviates from expected patterns.

The dataset analyzed in this work consists of 44 production drill patterns containing 4,117 blast holes drilled across multiple benches. For each hole, MWD parameters were recorded continuously with depth during routine production drilling. These measurements include rate of penetration, torque, rotary speed, weight-on-bit, and air pressure, providing high-resolution insight into the drill's mechanical and hydraulic response as it advances through different ground conditions. After combining all pattern files into a single dataset, the resulting compilation contained 610,906 depth-referenced drilling records, representing a comprehensive view of drilling performance across the operation.

In addition to the production dataset, four video logs were available from holes where drillers observed backfills, voids, or other ground irregularities during drilling or post-drilling inspection. These video logs provided depth-specific confirmation of anomalous subsurface conditions and were used to help interpret the drilling signatures associated with void-prone or disturbed ground.

The combined dataset offers both the spatial coverage and depth resolution needed to evaluate how drilling responses vary across the operation and to support the development of data-driven methods for identifying zones potentially influenced by historical underground workings.

Figure 3.1 presents a plan-view layout of blast-hole collar locations for the 44 production drill patterns analyzed in this study. The figure is plotted in mine coordinate space to preserve the true spatial relationships between drill patterns, but it does not include an underlying mine map or pit geometry. The layout provides relative positional context by illustrating the lateral extent of

drilling, pattern spacing, and overlap between adjacent production areas. This representation establishes the spatial coverage of the MWD dataset used in subsequent analyses and avoids inclusion of mine infrastructure or geological boundaries that were not available for release.

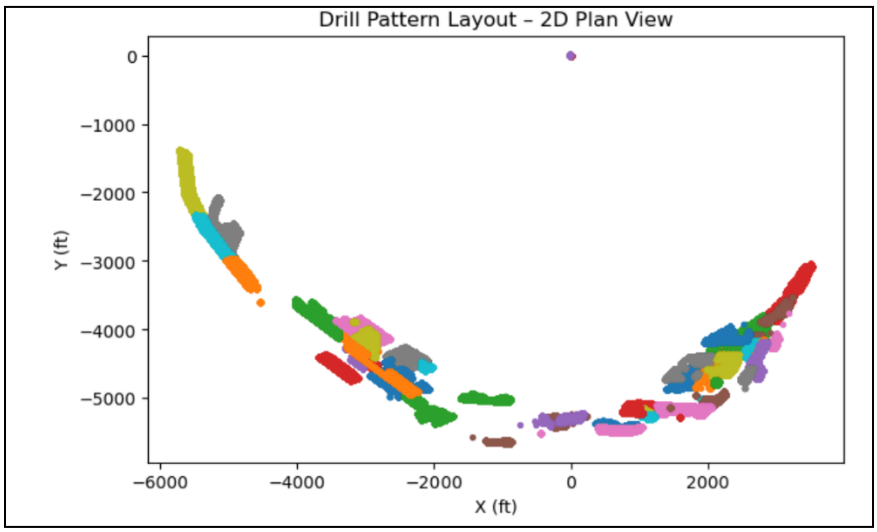

Figure 0.1 Plan-view layout of blast-hole collar locations for the 44 production drill patterns used in this study. Colors indicate individual drill patterns, illustrating the spatial distribution and geometric variability of drilling across the active mining area.

3.3.2 Data Cleaning and Preprocessing

Prior to analysis, the consolidated MWD dataset was organized and conditioned to ensure that drilling responses were internally consistent and suitable for interpretation. Because depth-based drilling measurements are inherently sequential, the full dataset was first sorted by hole identifier and increasing depth. This step preserved the natural order of drilling progression in each hole and ensured that downstream calculations and interpolations reflected the true drilling sequence.

The primary MWD parameters namely, rate of penetration, rotary speed, torque, weight-on-bit, and air pressure, occasionally contained gaps arising from brief sensor interruptions or intermittent recording. These gaps were associated with short-duration operational artifacts such as sensor dropouts, logging delays, or brief non-drilling events, and were not found to correspond to geological features or anomalous ground conditions. These missing values were filled using depth-ordered interpolation performed within each hole only, preventing the introduction of information from other holes or patterns. This method maintained the internal continuity of drilling response while preserving realistic transitions in the measured parameters.

Time-related fields associated with hole initiation and section intervals were standardized by converting all entries to a consistent datetime format. This ensured that any variations in timestamp formatting across the original files did not affect later calculations or the interpretation of drilling progression. Records containing no valid information in any field were removed, as they provided no usable drilling or geological context.

To remove non-representative drilling data, holes in which all major MWD sensors reported near-zero values for most of their recorded length were identified and excluded, as such sequences reflect non-drilling activity or invalid acquisition. Depth intervals were then calculated for each record, and only positive intervals below 2 ft were retained to represent realistic drilling advance. After applying these final filters and re-sorting by hole and depth, the cleaned dataset consisted of 608,553 depth-based drilling intervals, representing approximately 99.6% of the original 610,906 records. The remaining 0.4% of intervals were removed due to non-representative drilling activity or invalid measurements, resulting in a coherent and reliable foundation for exploratory data analysis and subsequent modeling.

3.3.3 Exploratory Data Analysis

Exploratory data analysis was conducted to understand how MWD parameters behaved under typical hard-rock drilling conditions in the open-pit mine and how these responses changed in zones thought to be influenced by old underground workings, backfill, or voids. MWD parameters are extremely sensitive to changes in ground condition, and identifying consistent departures from normal drilling response is the first step in detecting the presence of voids or weak zones. The EDA presented here characterizes (i) the operational ranges of each variable, (ii) natural correlation structure among the parameters, and (iii) time and depth-based drilling signatures indicating potential anomalies.

3.3.4 Descriptive Statistics of MWD Parameters

The descriptive statistics in Table 3.1 summarize the expected operational ranges for rotary speed, weight-on-bit, torque, rate of penetration, and air pressure. These values represent “typical” drilling conditions in competent rock within the mine.

Table 0.1 Descriptive statistics for key MWD parameters in the cleaned dataset.

MWD Parameter	Count	Mean	Std. Dev.	Min	25%	Median	75%	Max
Start Depth (ft)	610,767	28.51	17.64	0.00	13.78	27.89	42.32	158.79
End Depth (ft)	610,767	28.88	17.64	0.00	14.11	28.54	42.65	158.89
RPM	610,767	49.33	27.35	0.00	41.10	53.60	64.30	141.10
Weight-On-Bit (kN)	610,767	319.33	137.55	0.00	198.12	336.87	447.43	616.07
Torque (Nm)	610,767	5.39	1.88	0.00	4.20	5.40	6.50	27.70
ROP (m/s)	610,767	0.0160	0.0332	0.00	0.0080	0.0115	0.0161	1.6667
Air Pressure (kPa)	610,767	347.76	185.31	0.00	275.80	310.90	340.20	1597.20

Several parameters exhibit wide ranges and skewed distributions, reflecting natural geological variability across benches and blast patterns. These ranges establish the baseline against which abnormal drilling episodes can be compared. The drilling systems used in this study operate under automated control loops for key parameters such as feed force, rotary speed, and weight-on-bit; however, operator interaction remains present, particularly during collaring, parameter set-point adjustments, and responses to changing ground conditions. As a result, the recorded MWD signals reflect a combined response to subsurface geology, machine control logic, operator decision-making, and tool condition. In particular:

i. Weight-on-bit and torque exhibit high variability as the bit transitions through stronger and weaker rock units, with additional short-term fluctuations introduced by automated control adjustments and operator interventions during changing drilling conditions.

ii. Rate of penetration is typically low under hard-rock conditions, with sudden increases often marking softer ground, fractured zones, or potentially voided material. While bit wear can reduce penetration rate over time by decreasing cutting efficiency and increasing load requirements, such effects develop gradually and produce smooth, monotonic trends rather than abrupt, depth-localized changes. The anomalous drilling responses of interest in this study characterized by sharp penetration-rate increases coincident with reductions in load-related parameters are therefore inconsistent with progressive bit wear.

iii. Air pressure acts as an indicator of flushing efficiency, where unusual drops may occur when the bit enters air gaps, loose backfills, or permeable zones that allow air to escape under reduced confinement.

Operator-specific identifiers, shift information, bit-life indices, or control-mode flags were not available in the dataset; therefore, operator and tool-condition effects could not be explicitly modeled as independent variables. Nevertheless, their influence is implicitly captured within the measured MWD parameters and reflected in the observed distributions. If such metadata were available, they could be incorporated as categorical or hierarchical features in future machine learning models to further separate operator- or tool-induced variability from geological response.

3.3.5 Distribution of MWD Parameters

The distribution of the MWD parameters across all drill holes (Figure 3.2) was examined to understand the expected operational ranges and to identify potential indicators of abnormal ground.

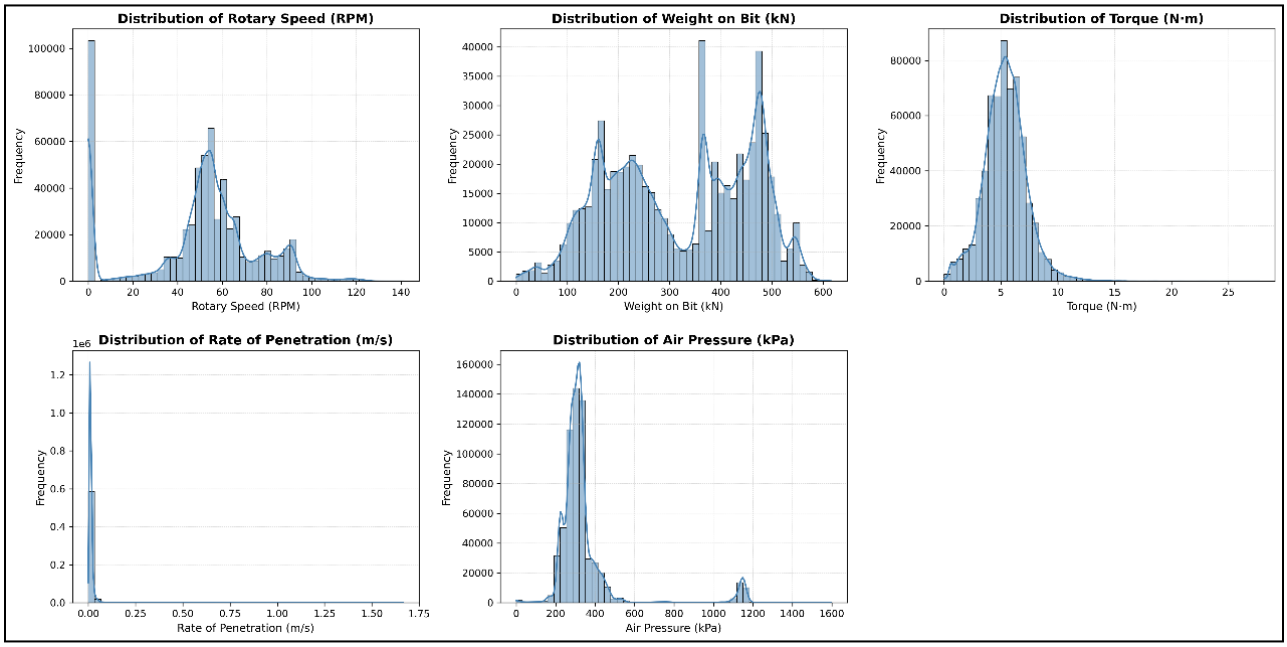

Figure 0.2 Distributions of MWD parameters across all holes, illustrating typical operating ranges and variability in rotary speed, weight-on-bit, torque, rate of penetration, and air pressure.

Key practical observations include:

i. Rotary Speed (RPM): Stable, with most drilling occurring between 40–80 RPM. Sharp increases or collapses in RPM during drilling can reflect abrupt losses of bit resistance, potentially linked to voids or soft fill.

ii. Weight-on-Bit (WOB): Exhibits multiple operational peaks, consistent with automatic control adjustments during bit loading. Extremely low WOB values at depth may indicate sudden loss of resistance as the bit enters loose or hollow material.
iii. Torque: Shows a well-defined central distribution; torque drops are often reliable indicators of reduced bit-to-rock interaction, including zones of backfill, caved material, or voids.
iv. Rate of Penetration (ROP): Strongly right skewed; increases in ROP tend to occur when the bit encounters fractured zones or unconsolidated material. Sharp, isolated ROP spikes are characteristic of drilling into void-prone intervals.
v. Air Pressure**:** Typically exhibits tight clustering, reflecting stable compressor operations. Sudden pressure reductions or instability can occur when flushing air escapes into larger openings or void spaces.

Together, these patterns illustrate how different MWD parameters respond to changes in ground competency and show the potential for detecting void-related anomalies.

3.3.6 Correlation Analysis

Correlation analysis was conducted to examine how key MWD parameters vary relative to one another and how these relationships change under different drilling conditions. The analysis uses the Pearson correlation coefficient, which quantifies the degree to which two variables increase or decrease together.

In simple terms, the Pearson correlation coefficient compares how deviations of two variables from their respective averages relate to one another and normalizes this relationship by their variability. It can be expressed as:

$$r = \frac{\text{covariance}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of variables X and Y. The resulting value ranges from -1 to $+1$. A value near $+1$ indicates that both parameters tend to increase or decrease together, a value near -1 indicates that one parameter increases as the other decreases, and values near zero indicate little or no linear relationship.

The resulting correlation matrix, shown in Figure 3.3, summarizes the pairwise relationships among rotary speed, weight-on-bit, torque, rate of penetration, and air pressure.

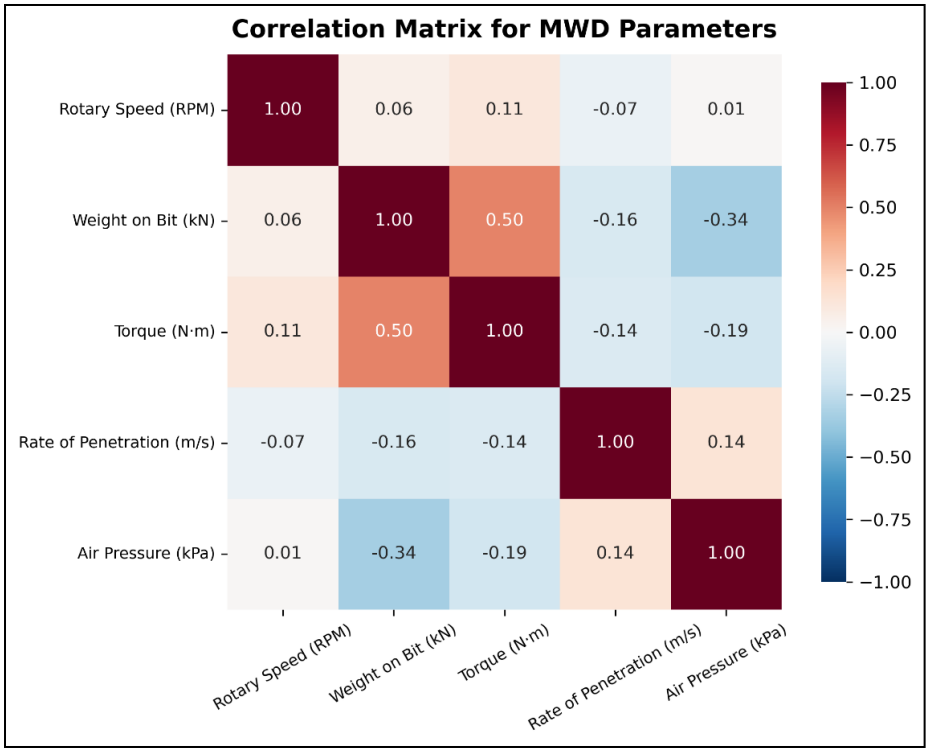

Figure 0.3 Correlation matrix summarizing the pairwise relationships among the MWD parameters.

The correlations are low to moderate, indicating that each parameter reflects a different aspect of bit performance and ground resistance. The moderate positive correlation between weight-on-bit and torque represents normal drilling, where increased bit loading produces higher cutting resistance. Local weakening of this relationship at depth can indicate reduced bit engagement, which is consistent with transitions into softer, disturbed, or void-prone material.

The weak negative correlations between rate of penetration and the resistance-related parameters (WOB and torque) reflect conditions where the bit advances rapidly through low-resistance zones. Air pressure also shows weak negative relationships with these parameters, and reductions in pressure can occur when air escapes into permeable or open ground.

Overall, the correlation trends illustrate how mechanical and pneumatic drilling responses begin to decouple when the bit encounters ground that lacks competent confinement, supporting their use in identifying anomalous or void-prone intervals.

3.3.7 Time-Series Profile of MWD Parameters

Figure 3.4 presents a representative drilling time-series profile selected because it shows a clear transition from stable drilling to an interval of reduced ground resistance. In the upper portion of

the hole, RPM, torque, and weight-on-bit remain stable and well correlated, consistent with normal bit–rock interaction. In the lower interval, weight-on-bit, and torque decrease while penetration rate increases and air pressure becomes unstable. This interpretation is supported by the data: under normal drilling mechanics, reduced applied load would be expected to decrease penetration rate, not increase it. The observed inverse response therefore indicates a loss of mechanical resistance at the bit–rock interface, consistent with weak, fractured, or poorly confined material.

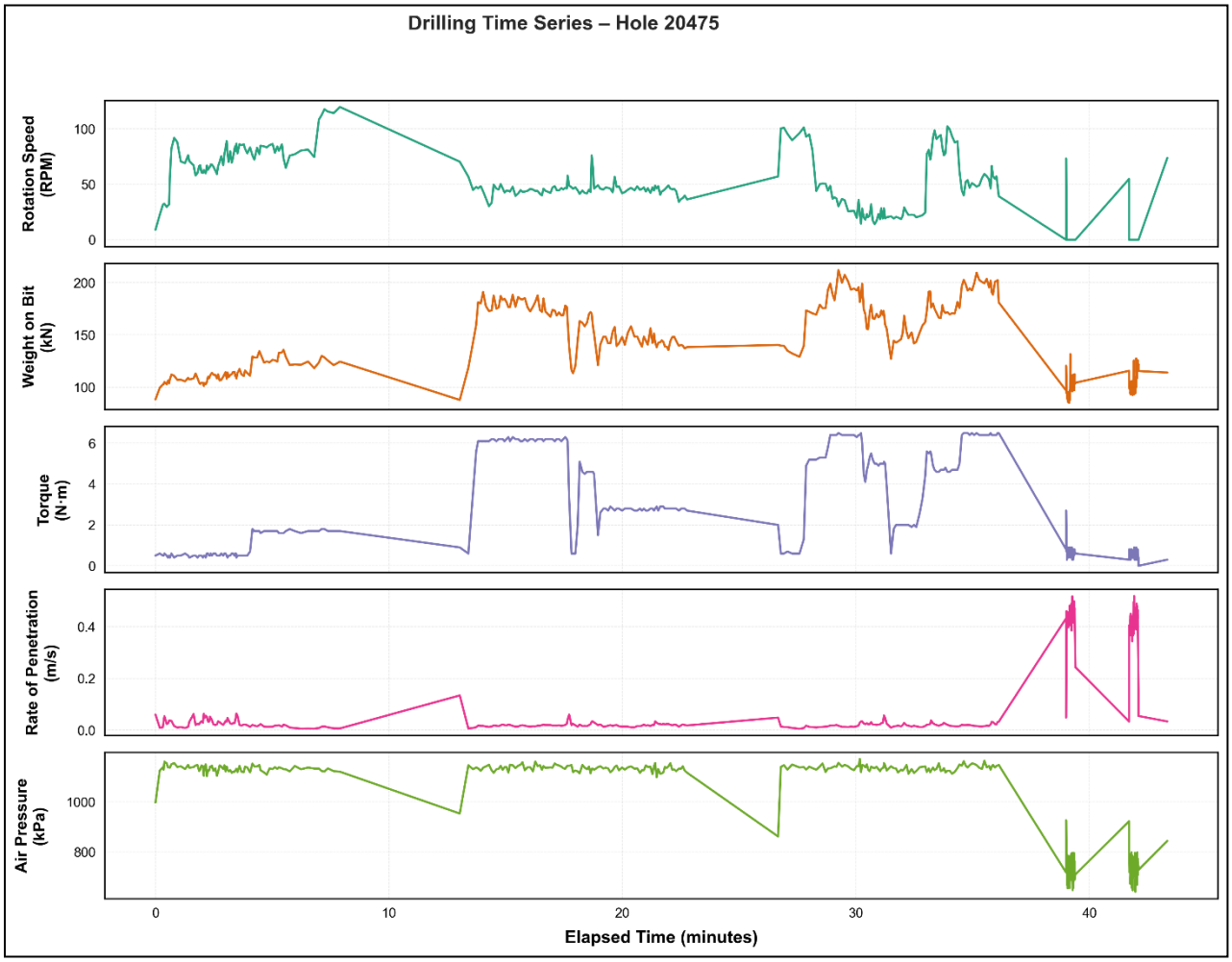

Figure 0.4 Representative drilling time-series profile showing a late-hole response consistent with reduced ground resistance typical of void-prone intervals.

3.3.8 Depth-Based Behavior of MWD Parameters

Figure 3.5 presents depth-referenced MWD responses for a representative hole. The upper portion of the profile shows consistent drilling: torque and WOB increase gradually with depth, while

RPM and air pressure remain stable which is typical of uniform, competent rock. A marked behavioral shift occurs at greater depth, where torque and WOB drop sharply, ROP increases, and air pressure becomes less stable. This coordinated response reflects a reduction in ground resistance and is characteristic of drilling through weak, broken, or void-prone intervals. Depth-based profiles therefore complement the time-series view by highlighting the exact depths where drilling behavior departs from normal.

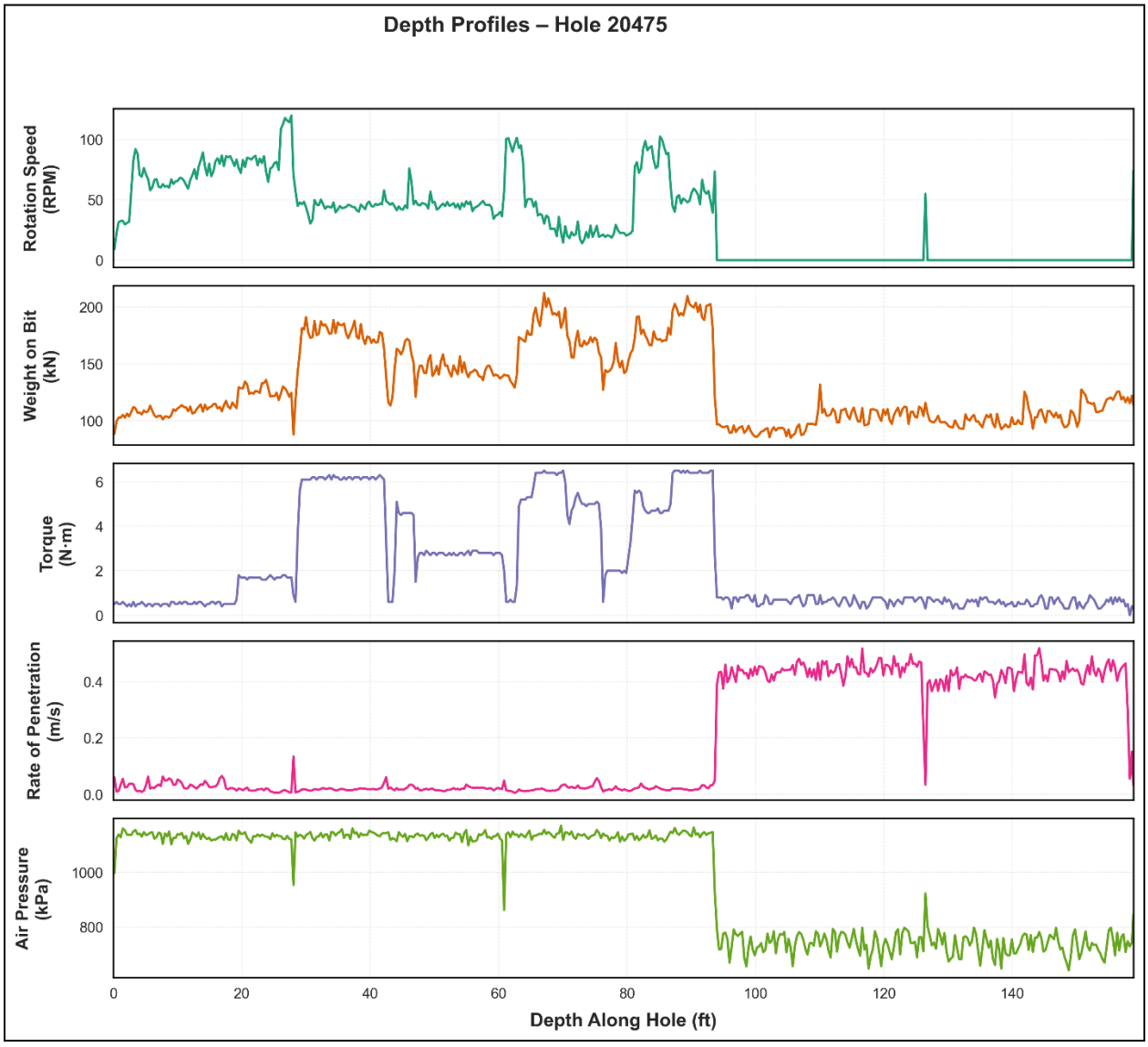

Figure 0.5 Depth-based MWD profiles showing stable drilling in the upper section and a lower interval with reduced resistance consistent with void-prone ground.

3.4 Modeling Framework

The modeling workflow was structured to detect anomalous drilling behavior indicative of weak, disturbed, or void-prone ground. The approach combined engineered features, unsupervised anomaly detection, supervised classification, and validation using independently derived ground-truth intervals.

3.4.1 Feature Engineering

Feature engineering was applied to strengthen the predictive value of the MWD dataset. Alongside the primary drilling measurements (ROP, RPM, torque, WOB, and air pressure), derived indicators were calculated to describe drilling efficiency, mechanical response ratios, and simplified mechanical specific energy (MSE). These features captured variations in bit loading, rotational resistance, and energy demand that commonly shift when drilling encounters soft, fractured, or void-affected material. All features were standardized to ensure comparability across holes and drilling patterns. A summary of engineered features is provided in Table 3.2.

Table 0.2 Summary of features used in the modeling framework.

Category	Feature Name	Description	Relevance to Void Detection
Primary MWD Inputs	ROP, RPM, Torque, WOB, Air Pressure	Raw depth-based drilling measurements recorded by the drill system.	Baseline mechanical responses describing bit–rock interaction and drilling conditions.
Mechanical Ratios	• Drilling Efficiency • WOB per RPM • Torque per WOB	Ratios expressing penetration performance relative to applied load or rotational resistance.	Elevated or unstable ratios may indicate reduced confinement, soft fill, or void-prone ground.
Energy-Based Indicator	Simplified Mechanical Specific Energy (MSE)	(Torque × RPM) ÷ ROP.	Lower energy demand often corresponds to weak, fractured, or voided ground conditions.
Standardization	Feature Scaling	Z-score transformation applied to all features.	Ensures equal weighting and improves comparability across drilling intervals.

3.4.2 Unsupervised Learning

Multiple unsupervised algorithms namely K-means, Isolation Forest (IF), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) were applied to the engineered feature set to identify intervals exhibiting abnormal drilling signatures. Each method was evaluated based on spatial coherence, depth consistency, and sensitivity to known problem areas. The best-performing algorithm produced a set of pseudo-labels representing "anomalous" and "non-anomalous" drilling behavior, which served as the training targets for the supervised learning stage.

3.4.3 Supervised Learning

Supervised classification models were trained to learn the MWD patterns associated with abnormal drilling response. Three algorithms were evaluated - Random Forest (RF), Logistic Regression (LR), and XGBoost using the pseudo-labels derived from the unsupervised stage. This allowed the models to generalize drilling behavior across holes and patterns and to predict intervals potentially associated with weak or void-prone ground.

3.4.4 Ground-Truth Label Creation and Evaluation

A limited set of verified labels was extracted directly from downhole video logs. Void intervals were manually identified pattern-by-pattern and mapped to their corresponding MWD coordinates using nearest-neighbor matching, yielding 21,680 confirmed ground-truth samples from four logged patterns. These labels were used solely to train and independently evaluate supervised models against known void-prone conditions.

3.5 Results

This section presents the outcomes of the unsupervised anomaly detection, and supervised classification of void-prone drilling intervals. The results are structured to first examine how the different clustering algorithms behaved on the full MWD dataset, followed by an evaluation of supervised models trained using the most reliable pseudo-label set. Finally, these confirmed void-prone intervals were used to train and evaluate supervised models independently of the pseudo-labels generated from the unsupervised stage.

3.5.1 Unsupervised Models for Anomaly Detection

The three unsupervised methods produced noticeably different anomaly patterns when visualized in plan view. Although the MWD data are inherently three-dimensional, plan-view visualization was used as a simplified representation to evaluate lateral clustering of anomaly classifications across the drill pattern at the bench scale; this approach does not capture vertical continuity but allows spatial coherence between neighboring holes to be assessed. Each depth-indexed interval was first classified by the unsupervised models and then projected to plan view using blast-hole collar coordinates.

K-Means generated the largest anomaly footprint, classifying 104,355 intervals (≈17%) as void-like (Figure 3.6). These appear as broad, continuous clusters across much of the drilling horizon, with red-labelled intervals forming large patches that cover extensive sections of the drilling traces. This behavior arises from K-Means' global, variance-minimizing partition of the feature space, which groups large numbers of moderately similar drilling responses into a single cluster; when projected to plan view, this produces laterally extensive anomaly patches rather than localized detections.

HDBSCAN produced the sparsest anomaly distribution, identifying only 5,145 intervals (<1%) as anomalies (Figure 3.7). These appear as isolated red markers scattered intermittently along the drill paths. Most intervals are assigned to a single dominant cluster, with only highly localized, density-isolated responses labeled as anomalous.

Isolation Forest resulted in an intermediate pattern, labeling 67,036 intervals (≈11%) as anomalous (Figure 3.8). The red points form coherent streaks and localized clusters along the drilling traces, producing spatially continuous but not overly widespread anomaly zones. This reflects the method's focus on isolating deviations from typical drilling behavior rather than globally partitioning the dataset, resulting in anomaly footprints that are more extensive than HDBSCAN but more focused than K-Means.

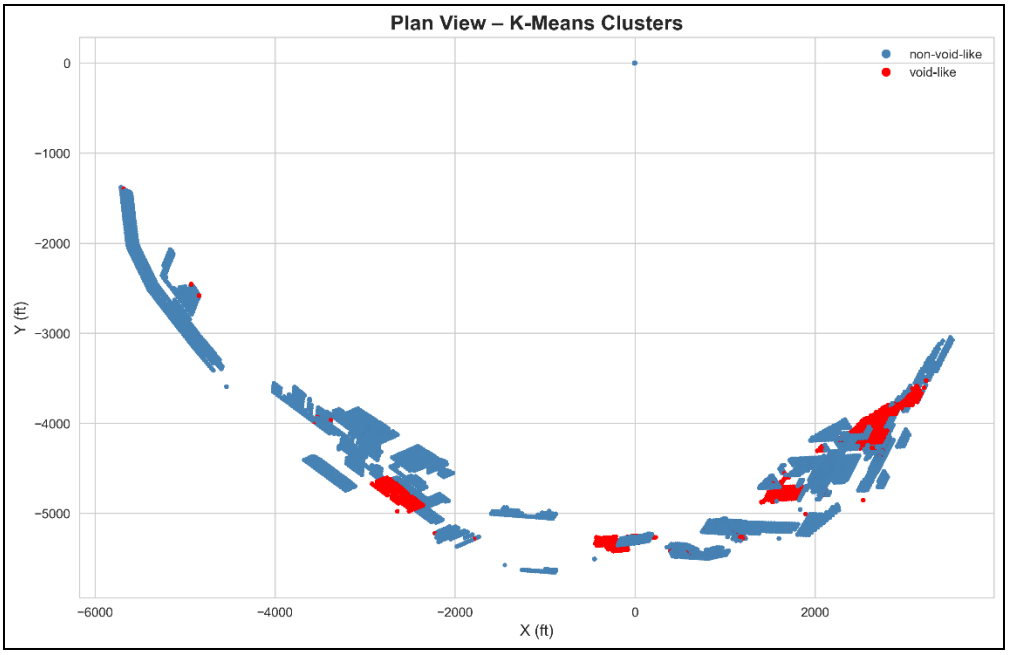

Figure 0.6 Plan-view spatial distribution of void-like intervals identified by the K-Means model.

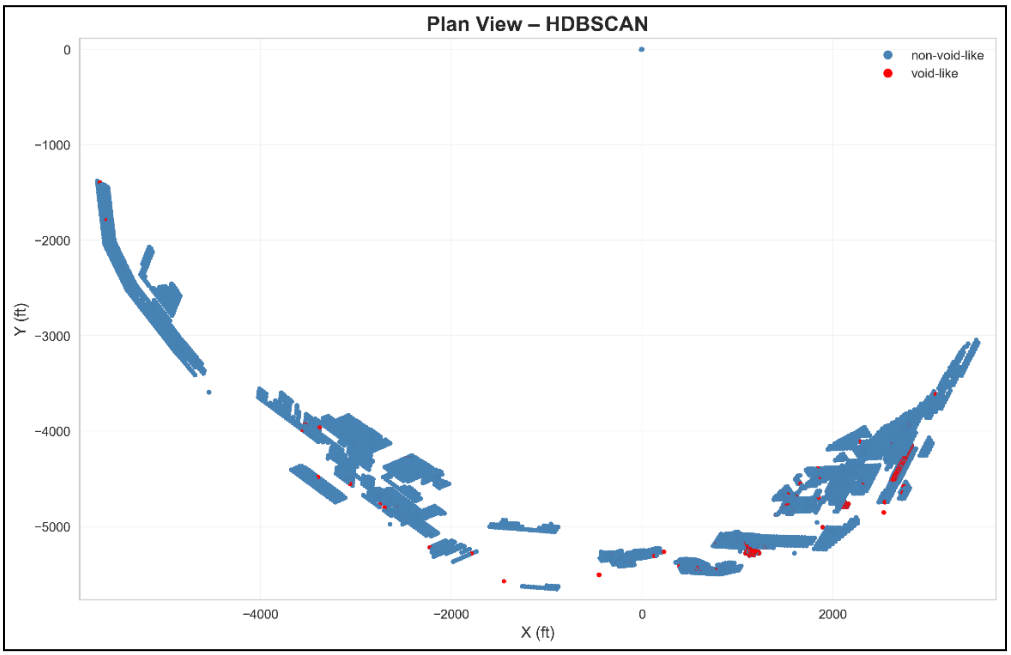

Figure 0.7 Plan-view spatial distribution of void-like intervals produced by HDBSCAN.

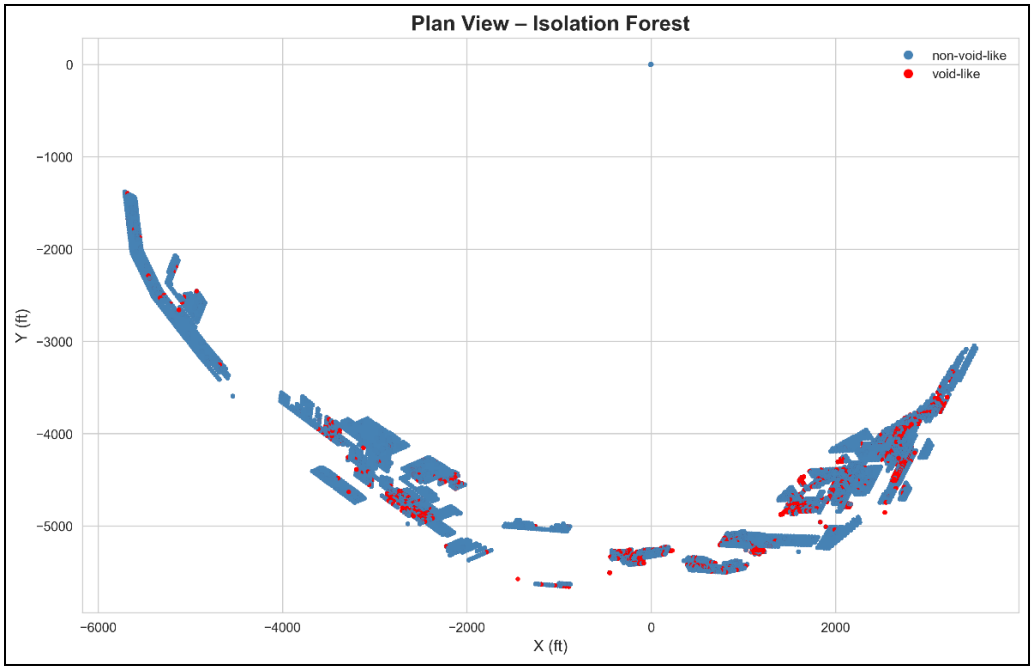

Figure 0.8 Plan-view anomaly distribution generated by the Isolation Forest model.

These visual patterns are consistent with the quantitative metrics. In this context, an "interval" refers to a depth-indexed segment of drilling data classified by its multivariate MWD response and then projected to plan view using the corresponding blast-hole collar coordinates. K-Means achieved modest silhouette scores with no strongly defined cluster separation, reflecting its tendency to group large portions of the feature space into a single cluster. HDBSCAN yielded its highest silhouette values at more conservative parameter settings but assigned almost all points to the main cluster, resulting in sparse anomaly detections. Isolation Forest achieved the highest silhouette scores overall, indicating the clearest separation between normal and anomalous drilling responses at the interval level. The complete silhouette-score results for all tested configurations are summarized in Table 3.3.

Based on the combination of spatial coherence, anomaly coverage, and metric performance, Isolation Forest was selected as the most suitable unsupervised method for generating pseudo-labels. This selection is not intended to imply universal optimality; rather, Isolation Forest demonstrated the strongest alignment with the objectives and constraints of this study. Compared to K-Means and HDBSCAN, it produced spatially coherent yet localized anomaly patterns in plan view while avoiding both excessive anomaly coverage and overly sparse detections. Given the

high dimensionality, operational noise, and lack of ground-truth labels in the MWD dataset, Isolation Forest provided the most balanced and interpretable basis for pseudo-label generation.

Table 0.3 Summary of Silhouette Performance Across Unsupervised Methods

Method	Parameter Settings Evaluated	Best Silhouette Score	Interpretation
K-Means	k = 2–10	0.323 (k = 9)	Moderate separation; limited natural clustering structure in the feature space.
HDSCAN	min cluster size = 20–200; min samples = 5–20	0.478	Improved structure but small anomaly fraction; clusters stable but conservative.
Isolation Forest	contamination = 0.01–0.19	0.680 (0.01)	Strong, well-separated anomaly partition; anomalies highly distinct.

3.5.2 Supervised Model Performance Using Isolation-Forest Pseudo-Labels

Three supervised learners - Logistic Regression, Random Forest, and XGBoost were trained using the pseudo-labels generated by the Isolation Forest method. Model evaluation was performed on a 20% held-out test set.

3.5.2.1 Classification Performance

The confusion matrices illustrate the ability of each model to discriminate between void-like and non-void-like drilling intervals:

i. Logistic Regression (Figure 3.9) shows a relatively high false-positive rate, incorrectly labelling many non-void-like intervals as void-like.
ii. Random Forest (Figure 3.10) achieves very low misclassification counts, showing strong separation between classes.
iii. XGBoost (Figure 3.11) exhibits exceptionally high recall for void-like intervals, with only 34 false negatives in the test set.

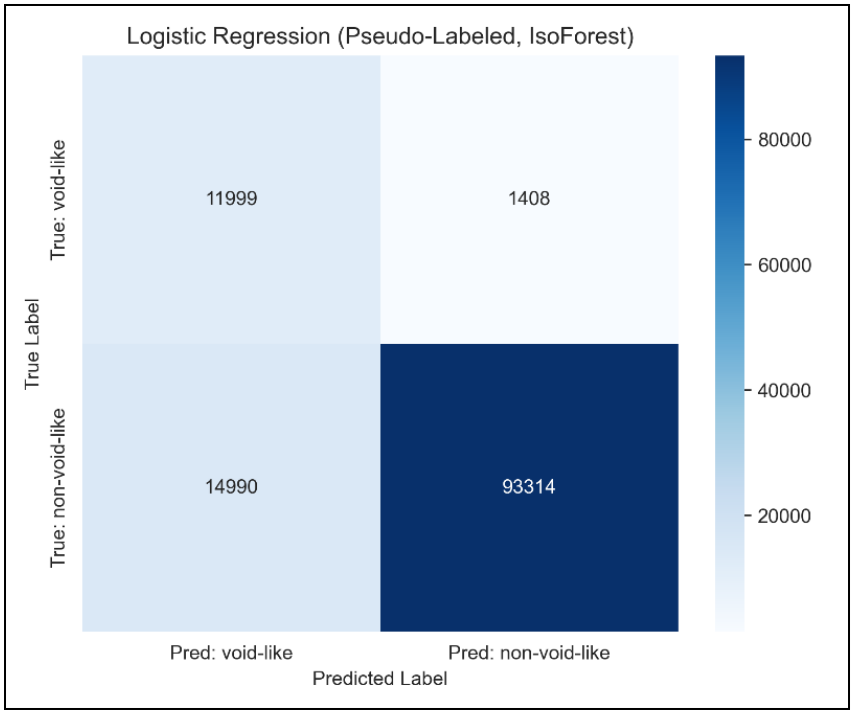

Figure 0.9 Confusion matrix for the Logistic Regression model trained on isolation forest pseudo labels.

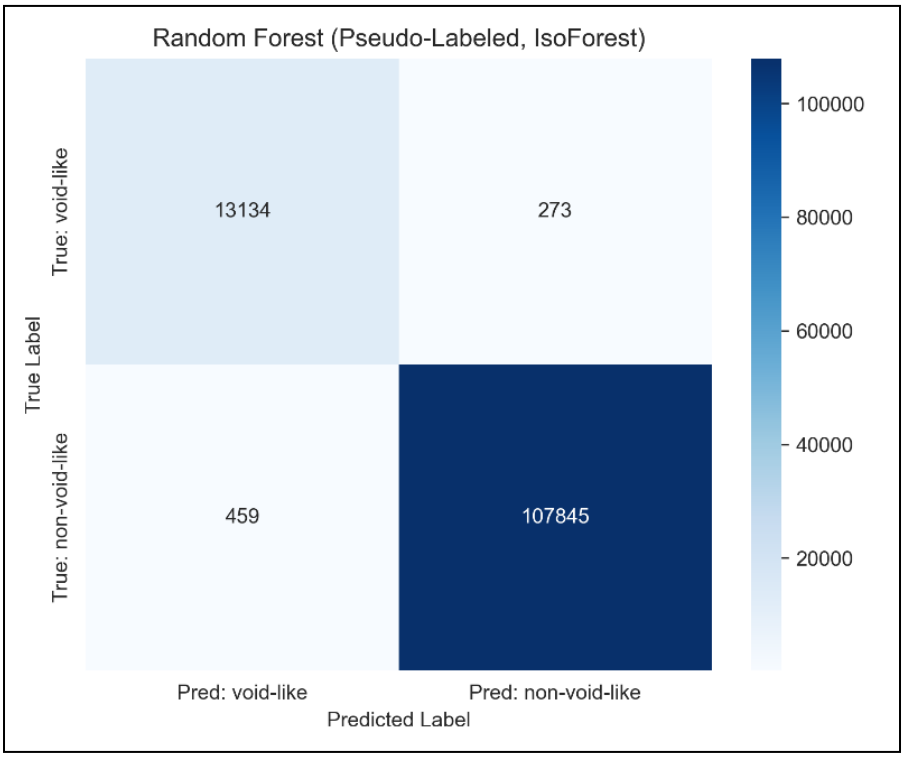

Figure 0.10 Confusion matrix for the Random Forest model trained on isolation forest pseudo labels.

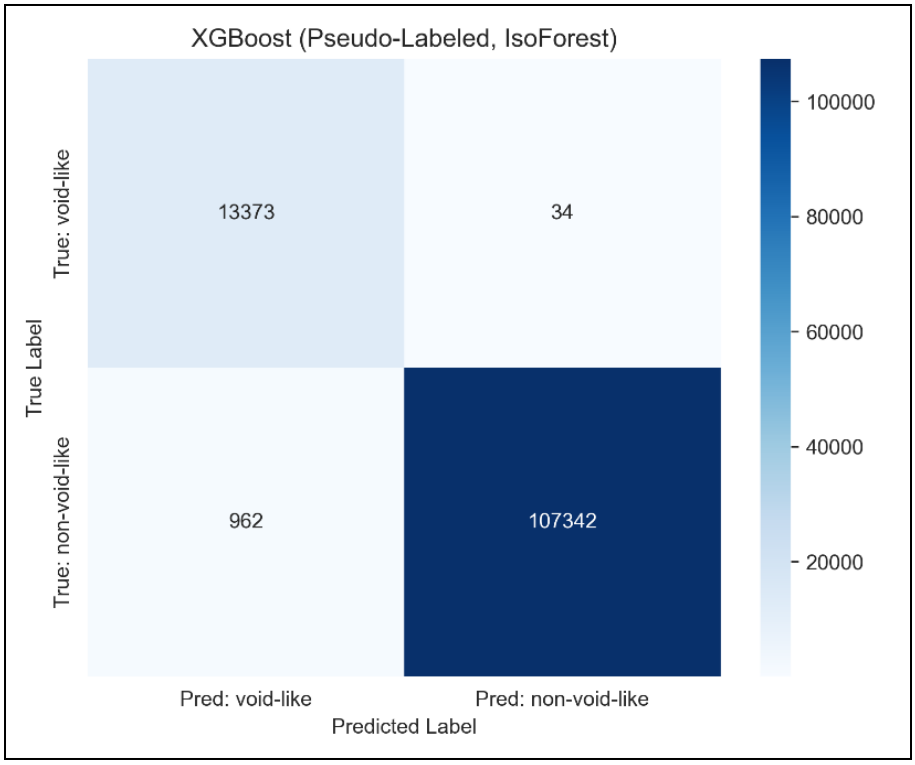

Figure 0.11 Confusion matrix for the XGBoost model trained on isolation forest pseudo labels.

All three supervised models achieved high overall accuracy, with tree-based methods outperforming the linear baseline. As shown in Table 3.3, Random Forest provided the best overall balance, with an F1 score of 0.973, while XGBoost achieved the highest recall (0.997), making it particularly effective for minimizing missed void detections.

Table 3.3 Summary of supervised model classification metrics using pseudo-labels.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.865	0.445	0.895	0.594
Random Forest	0.994	0.966	0.980	0.973
XGBoost	0.992	0.933	0.998	0.964

3.5.2.2 Feature Importance

Figures 3.12–3.14 summarize the feature-importance results for all three supervised models placed side-by-side for comparison. Despite using different learning mechanisms, the models consistently highlighted a similar set of drilling parameters as the most influential for predicting void-like behavior. Logistic Regression (LR) assigned its largest coefficient to MSE, with smaller

contributions from ROP and WOB. Although MSE is a derived parameter computed from multiple drilling variables, its prominence reflects its ability to aggregate several aspects of drilling response into a single, physically meaningful metric.

In LR, each coefficient represents the marginal linear contribution of a feature after accounting for all others; therefore, the individual parameters used to compute MSE may exhibit smaller coefficients due to shared variance and correlation, while the composite MSE feature captures their combined influence more effectively. Random Forest emphasized Weight on Bit and Air Pressure, followed by MSE and Torque, while XGBoost likewise ranked Air Pressure and MSE as its strongest predictors, with additional influence from WOB and WOB-per-RPM. This consistency across models indicates that both raw and derived parameters associated with drilling energy and confinement play a central role in distinguishing void-like behavior.

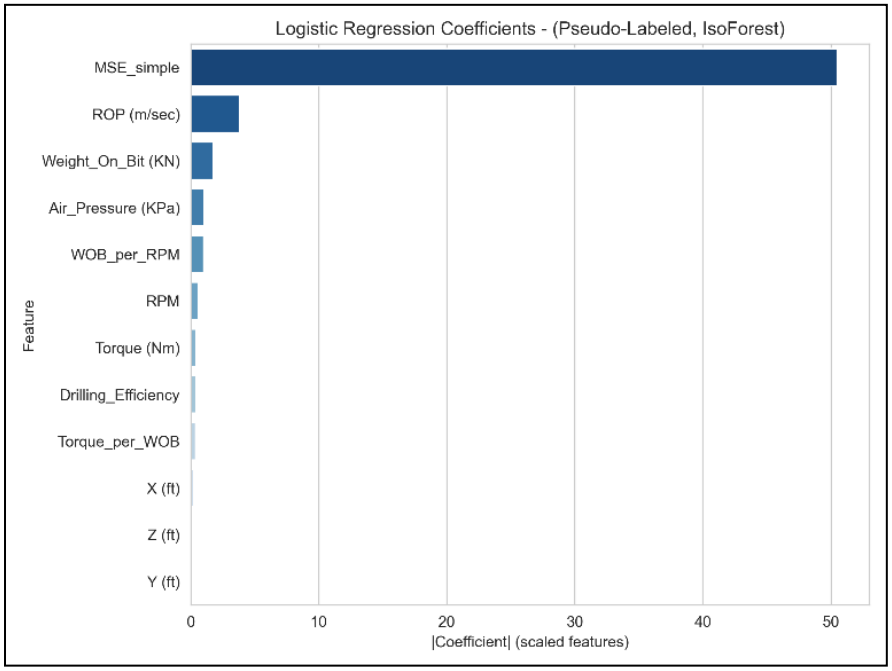

Figure 0.12 Logistic Regression feature coefficients.

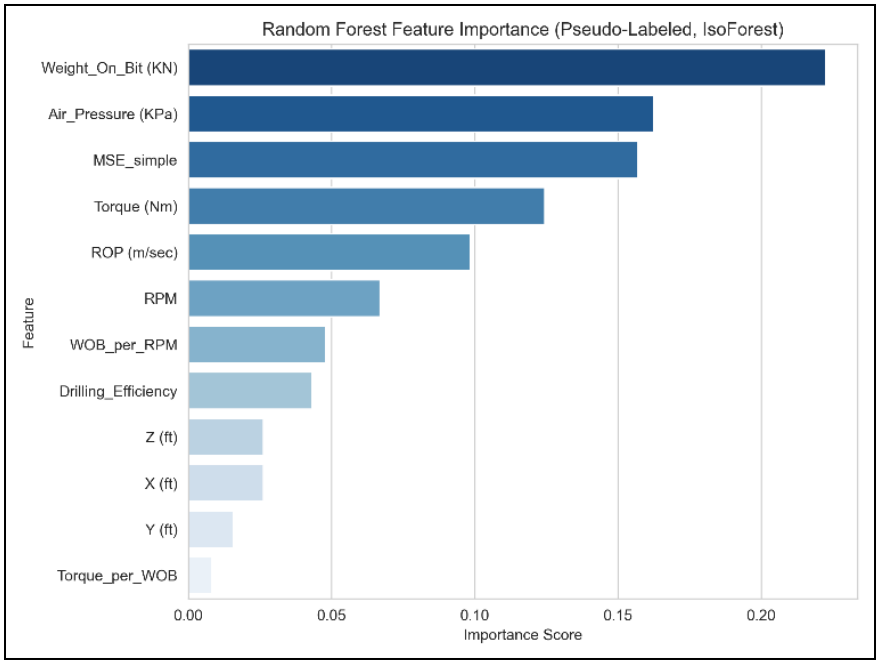

Figure 0.13 Random Forest feature importance scores.

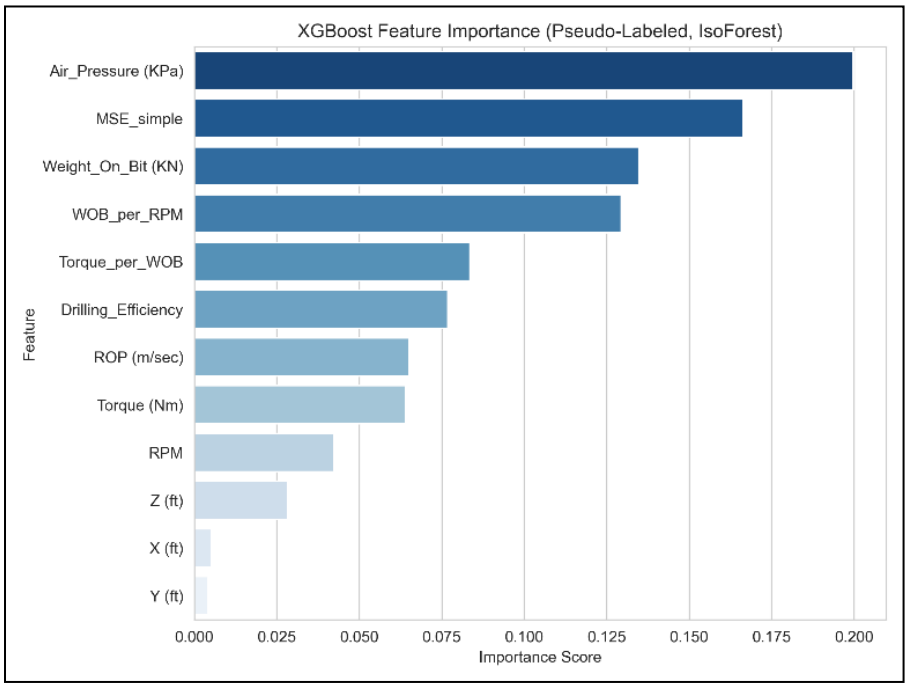

Figure 0.14 XGBoost feature importance scores.

3.5.3 Supervised Model Peformance Using Ground-Truth Labels

To benchmark the quality of the pseudo-labels produced earlier, a separate set of models was trained using only the 21,680 intervals with verified ground-truth labels derived from downhole video logs and mapped voids or structures. The labeled dataset is highly imbalanced, containing 214 void-like intervals (≈1%) and 21,466 non-void-like intervals (≈99%). To address this imbalance, each supervised model incorporated a cost-sensitive learning strategy appropriate to its formulation: Logistic Regression applied class-weight balancing, Random Forest used balanced sampling during tree construction, and XGBoost increased the relative weight of the void class so that rare positive examples exerted greater influence during training. These imbalance-handling approaches are well-established and widely accepted methods for rare-event detection in supervised learning, particularly when false negatives carry higher operational risk than false positives.

The suitability of this methodology is supported by stable model training and improved sensitivity to void-like intervals across all three models, indicating that the proposed approach is appropriate for the objectives and constraints of the available labeled dataset rather than being claimed as a universally optimal solution.

3.5.3.1 Classification Performance

The confusion matrices (Figures 3.15 – 3.17) illustrate how each model distributes errors:

Logistic Regression (Figure 3.15) correctly identifies most void-like intervals (high recall) but generates many false positives. Random Forest (Figure 3.16) and XGBoost (Figure 3.17) classify nearly all void-like samples correctly (recall = 1.00) but still mislabel a small number of non-void intervals as void-like due to the limited positive examples available during training.

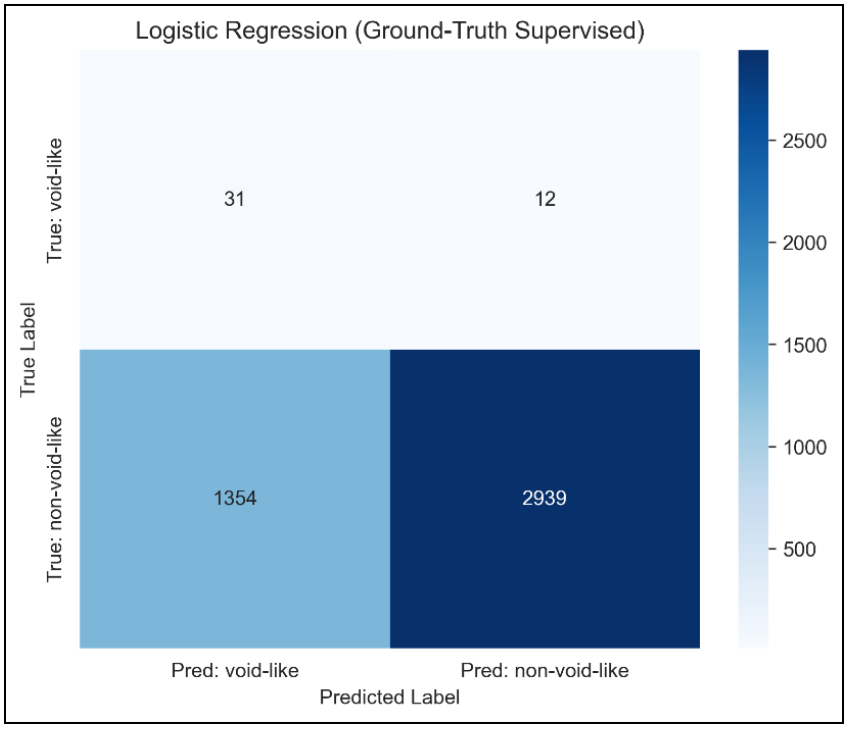

Figure 0.15 Confusion matrix for the Logistic Regression model trained on ground truth labels.

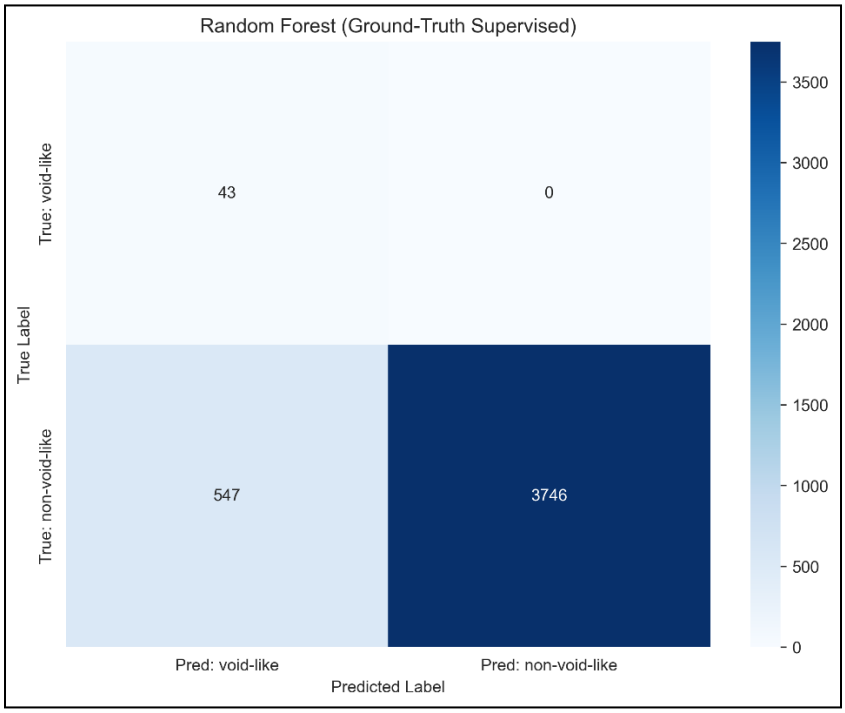

Figure 0.16 Confusion matrix for the Random Forest model trained on ground truth labels.

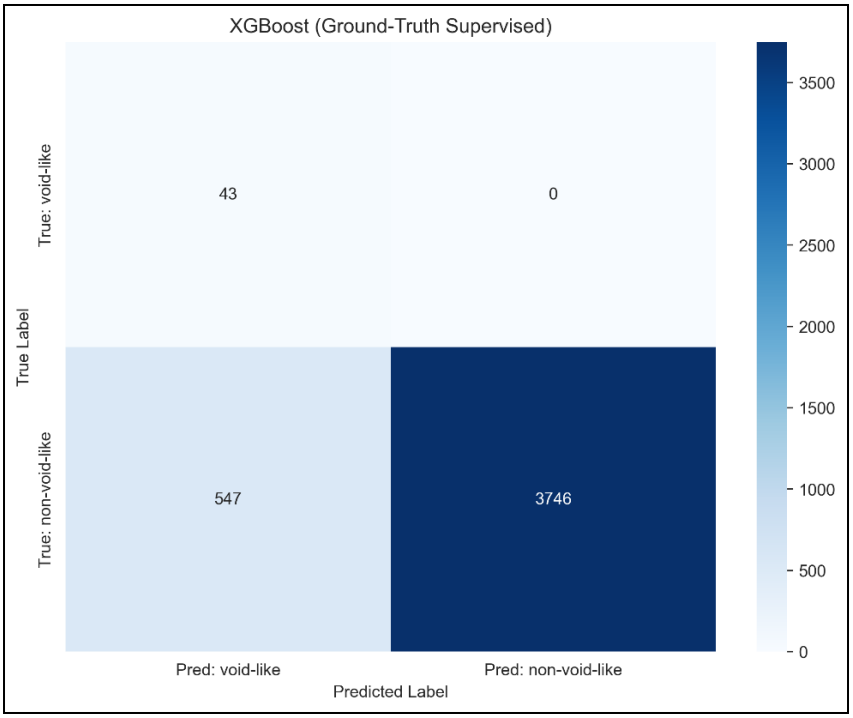

Figure 0.17 Confusion matrix for the XGBoost model trained on ground truth labels.

3.5.3.2 Feature Importance

Despite being able to handle imbalance, performance remained limited because the number of true void-like examples was extremely small. As summarized in Table 3.4, Logistic Regression showed extremely low precision but high recall for the minority class. Random Forest and XGBoost performed similarly, achieving perfect recall for void-like intervals but still struggling with precision.

Table 0.4 Summary of supervised model classification metrics

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.685	0.022	0.721	0.043
Random Forest	0.874	0.073	1.000	0.136
XGBoost	0.874	0.073	1.000	0.136

The ground-truth dataset contained only a small number of confirmed void intervals and hence feature-importance rankings were not interpreted for these models. Their primary value was in assessing detection performance rather than in deriving geological insights. Also, high recall achieved by tree-based models indicates strong sensitivity to void-like conditions, while the

considerable false-positive rates reflect the limited number of true void examples available for training.

3.6 Discussion

The results from both the unsupervised and supervised workflows highlight the contrasting behaviors of models trained on pseudo-labels versus those trained on verified ground-truth intervals. The Isolation-Forest pseudo-labels provided a large, balanced representation of anomalous drilling signatures, allowing the supervised learners, particularly Random Forest and XGBoost to achieve excellent precision–recall balance and strong overall performance. These pseudo-labels captured continuous anomalous trends across the drilling horizon, giving the models access to thousands of representative samples for both classes.

In contrast, the supervised models trained directly on ground-truth intervals exhibited different behavior due to the extreme rarity of verified void samples, resulting in a highly imbalanced learning problem. Under these conditions, Random Forest and XGBoost achieved perfect recall on the limited true-void examples but produced low precision, reflecting a tendency to flag many intervals as void-like. Such behavior is commonly observed in sparse-positive classification problems, where cost-sensitive training and conservative decision boundaries increase sensitivity to rare events while elevating false-positive rates. This trade-off is inherent to the data distribution rather than indicating model malfunction.

When comparing outputs, the pseudo-label-trained models provide a smoother and more spatially consistent prediction field, useful for operational guidance and early detection of weak zones. Ground-truth-trained models, on the other hand, act as extremely sensitive detectors: they rarely miss a real void but produce false positives, limiting their standalone operational reliability. Together, these findings suggest that pseudo-labels, when carefully validated and spatially coherent, can offer a practical bridge between purely unsupervised structure discovery and highly conservative ground-truth detection performance.

A hybrid workflow may therefore be most effective: pseudo-label models can map broad anomaly regions with high confidence, while ground-truth-trained detectors can be used to verify the most critical segments. The complementary strengths of both approaches point toward future ensemble

strategies where unsupervised anomaly structure and sparse verified events jointly contribute to a more robust void-prediction system.

3.7 Conclusions

This study showed that engineered MWD features, combined with machine-learning methods, can effectively identify drilling intervals associated with void-like conditions. Isolation Forest provided stable pseudo-labels, and supervised models, especially tree-based methods learned predictive patterns that aligned well with independent ground-truth data. While the approach is limited by the small number of validated intervals and reliance on MWD measurements alone, the results demonstrate strong potential for integrating data-driven detection of void-prone zones into drilling workflows.

3.8 References

[1] Mines Occupational Safety and Health Advisory Board (Western Australia), "Open Pit Mining Through UG Workings." State of Western Australia, 2000.

[2] W. Smith and R. Bertuzzi, "Mining through historic underground workings and systematic void management processes at McArthur River Mine," in *SSIM 2021: Second International Slope Stability in Mining*, Australian Centre for Geomechanics, Perth, 2021, pp. 277–286. doi: 10.36487/ACG_repo/2135_17.

[3] W. Johnson, "APPLICATIONS OF THE ELECTRICAL RESISTIVITY METHOD FOR DETECTION OF UNDERGROUND MINE WORKINGS," Lexington, KY.

[4] "Detection of underground voids in Ohio by use of geophysical methods," 1997. doi: 10.3133/wri974221.

[5] F. P. Haeni, L. Halleux, C. D. Johnson, and J. W. Lane, "Detection and Mapping of Fractures and Cavities using Borehole Radar".

[6] R. Benson, R. Kaufmann, Y. Lynn, and R. Hopkins, "Locating and Characterizing Abandon Mines Using MicroGravity." July 28, 2003.

[7] A. K. Bharti, A. Prakash, S. Oraon, P. Jaiswal, and S. K. Mandal, "Electrical Resistivity Tomography study above inaccessible old mine workings for safe erection of high voltage electricity power transmission terrestrial towers: a case study," *Bull Eng Geol Environ*, vol. 83, no. 10, p. 398, Oct. 2024, doi: 10.1007/s10064-024-03893-6.

[8] M. Khoshouei, R. Bagherpour, and M. Yari, “A smart look at monitoring while drilling (MWD) and optimizing using acoustic emission technique (AET),” *Sci Rep*, vol. 14, no. 1, p. 19766, Aug. 2024, doi: 10.1038/s41598-024-70717-8.

[9] W. Liu, J. Rostami, and E. Keller, “Application of new void detection algorithm for analysis of feed pressure and rotation pressure of roof bolters,” *International Journal of Mining Science and Technology*, vol. 27, no. 1, pp. 77–81, Jan. 2017, doi: 10.1016/j.ijmst.2016.11.009.

[10] D. Goldstein, C. Aldrich, and L. O’Connor, “Enhancing Orebody Knowledge using Measure-While-Drilling Data: A Machine Learning Approach,” *IFAC-PapersOnLine*, vol. 58, no. 22, pp. 72–76, 2024, doi: 10.1016/j.ifacol.2024.09.293.

[11] G. C. Komadja, E. Westman, A. Rana, and A. Vitalis, “Predicting rock mass strength from drilling data using synergistic unsupervised and supervised machine learning approaches,” *Earth Sci Inform*, vol. 18, no. 3, p. 325, Sept. 2025, doi: 10.1007/s12145-025-01837-6.

[12] D. Goldstein, C. Aldrich, Q. Shao, and L. O’Connor, “A Field-Scale Framework for Assessing the Influence of Measure-While-Drilling Variables on Geotechnical Characterization Using a Boruta-SHAP Approach,” *Mining*, vol. 5, no. 1, p. 20, Mar. 2025, doi: 10.3390/mining5010020.

[13] T. F. Hansen, Z. Liu, and J. Torresen, “Building and Analysing a Labelled Measure While Drilling Dataset from 15 Hard Rock Tunnels in Norway,” 2024, *SSRN*. doi: 10.2139/ssrn.4729646.

[14] A. Sapronova, T. Marcher, A. Soliman, and F. Klein, “Enhancing Rock Mass Characterization with Advanced Pre-Processing of MWD Data,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 1435, no. 1, p. 012010, Dec. 2024, doi: 10.1088/1755-1315/1435/1/012010.

[15] D. Goldstein, C. Aldrich, Q. Shao, and L. O’Connor, “A Machine Learning Classification Approach to Geotechnical Characterization Using Measure-While-Drilling Data,” *Geosciences*, vol. 15, no. 3, p. 93, Mar. 2025, doi: 10.3390/geosciences15030093.

[16] M. C. Altindal, P. Nivlet, M. Tabib, A. Rasheed, T. G. Kristiansen, and R. Khosravanian, “Anomaly detection in multivariate time series of drilling data,” *Geoenergy Science and Engineering*, vol. 237, p. 212778, June 2024, doi: 10.1016/j.geoen.2024.212778.

[17] D. M. Goldstein, C. Aldrich, and L. O'Connor, "A Review of Orebody Knowledge Enhancement Using Machine Learning on Open-Pit Mine Measure-While-Drilling Data," *MAKE*, vol. 6, no. 2, pp. 1343–1360, June 2024, doi: 10.3390/make6020063.

[18] I. Anafo, R. Ganguli, and N. Sarantsatsral, "BoxRF: A New Machine Learning Algorithm for Grade Estimation," *Applied Sciences*, vol. 15, no. 8, p. 4416, Apr. 2025, doi: 10.3390/app15084416.

Chapter 4: The Role of EDA in Developing Robust Machine Learning Models for Lithology and Penetration Rate Prediction from MWD Data.

4.1 Abstract

Measurement-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.

4.2 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], [5], [6], [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 rate of penetration (ROP), weight 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], [9]. Guidelines such as the Florida Method of Test for MWD [10] 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], [12], [13], [14], [15], [16], [17], [18], [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. [4] 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 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 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. This integration promotes a transparent, reproducible workflow that improves both model generalization and confidence in decision-support systems for drilling and mineral exploration.

4.3 Materials and Methods

4.3.1 Data Source and Description

This study applied an EDA–driven machine-learning workflow to predict lithology and rate of penetration 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 2-inch (5 cm) depth intervals. The raw dataset contained approximately 235,501 valid measurements acquired from 1,436 drill holes. 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 4.1.

Table 0.1 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:

i. 2 – Magnetite-bearing ore
ii. 3 – Claystone or minor alteration zone
iii. 4 – Waste rock or breccia lens
iv. 7 – Thin shale band or unclassified material
v. 8 – Taconite or intermediate banded-iron formation
vi. 9 – Hematite-rich ore
vii. 10 – Banded low-grade taconite
viii. 11 – Chert or siliceous layer
ix. 12 – Hematitic shale
x. 13 – Transitional lithology or boundary rock

4.3.2 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 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.

4.3.3 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.

4.3.3.1 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.

Figures 4.1 and 4.2 illustrate the spatial distribution of boreholes across the study area. The 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 > 39000) highlights denser sampling and lateral lithological continuity, supporting the inclusion of spatial coordinates as predictive features in subsequent analyses.

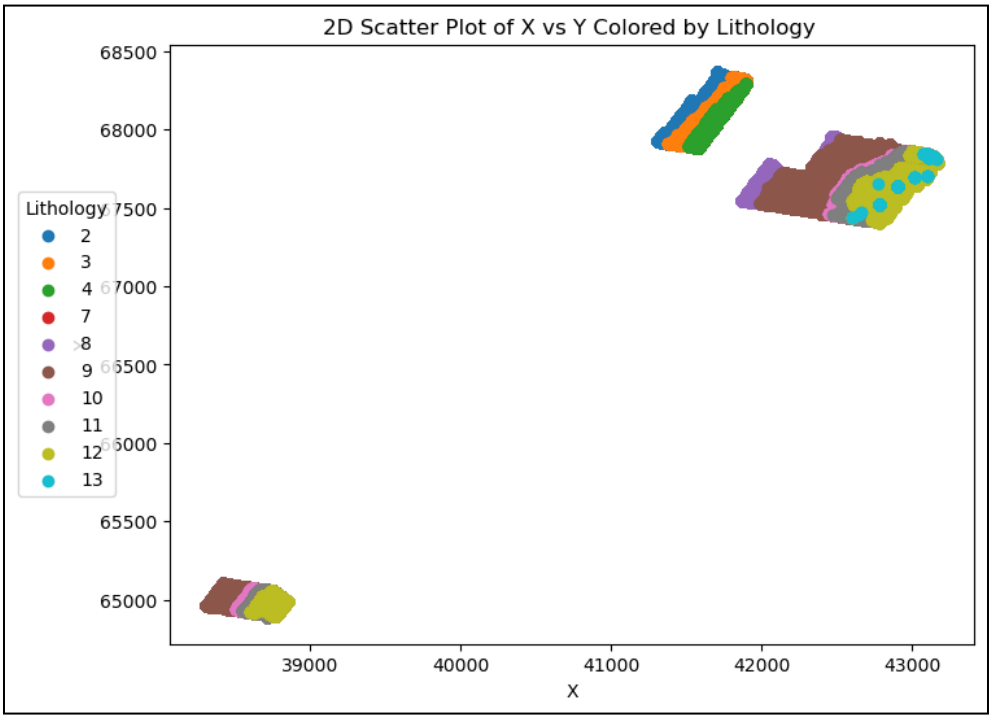

Figure 0.1 2D scatter plot of X–Y coordinates for the full dataset, colored by lithology.

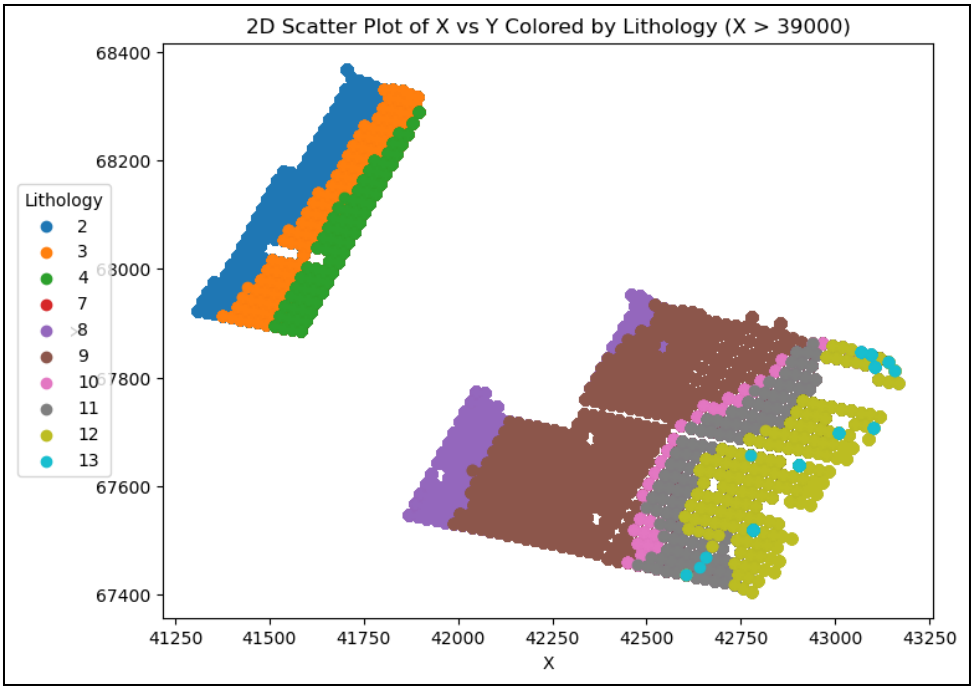

Figure 0.2 2D projection of X–Y coordinates for X > 39 000, colored for lithology.

4.3.3.2 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 4.2 summarizes these statistics, including measures of central tendency, dispersion, and percentile ranges.

Table 0.2 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	235501.00	235501.00	235501.00	235501.00	235501.00	235501.00	235501.00	235501.00
mean	1799.48	2557.54	49.65	2.12	1853.05	66286.50	2767.32	66331.62
std	2.44	512.03	6.87	1.41	298.58	12143.80	565.85	12151.97
min	1652.00	1.45	0.00	0.07	732.45	15987.96	550.00	15998.95
10%	1798.00	2001.54	40.61	0.79	1493.90	52905.24	2062.50	52939.60
50%	1800.00	2686.12	49.31	1.74	1841.99	69976.22	2750.00	70023.77
90%	1801.00	3021.16	58.02	3.84	2210.39	75948.33	3437.50	75999.96
max	1816.00	3125.59	118.93	9.97	4013.23	77719.04	6875.00	77771.92

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 4.3, which visualizes the variability and distribution shape of both spatial and mechanical variables .

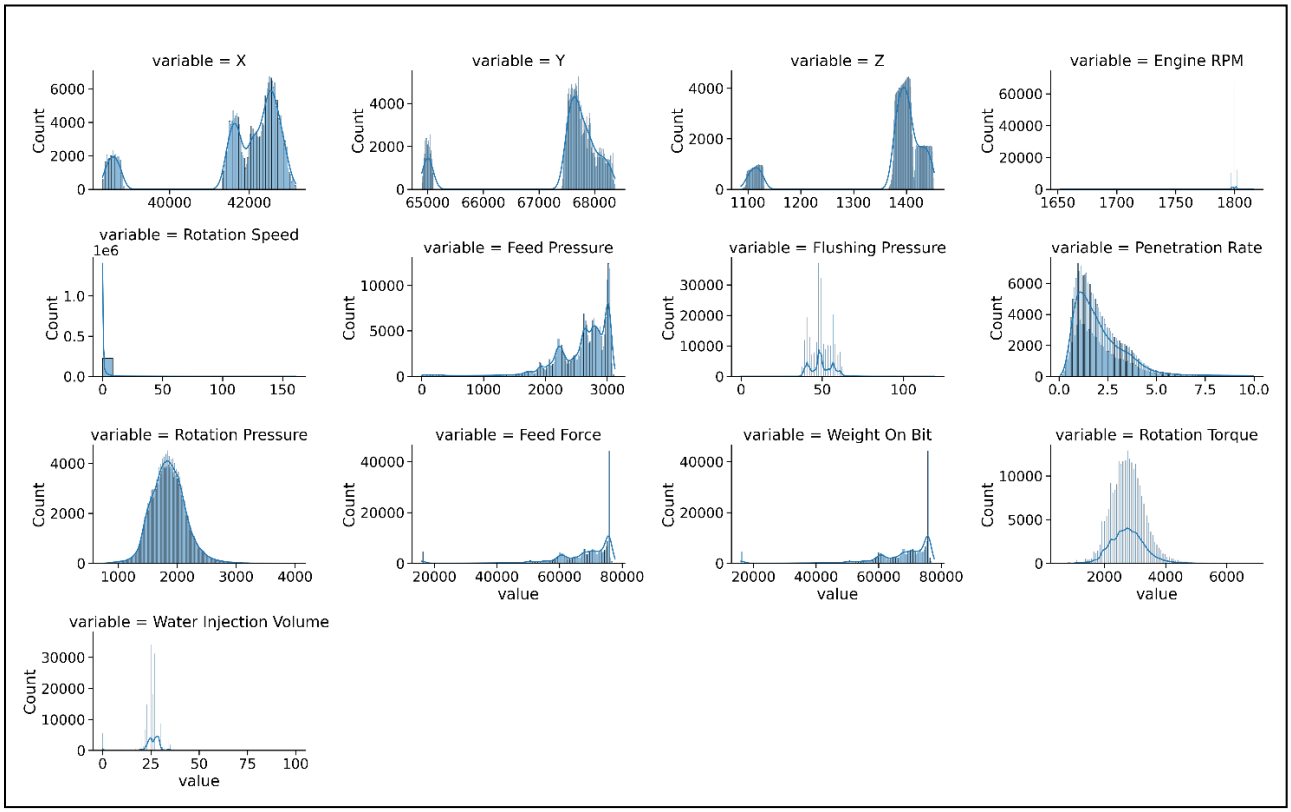

Figure 0.3 Histograms showing univariate distributions of the principal MWD parameters and spatial coordinates in the raw dataset.

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 4.4). 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.

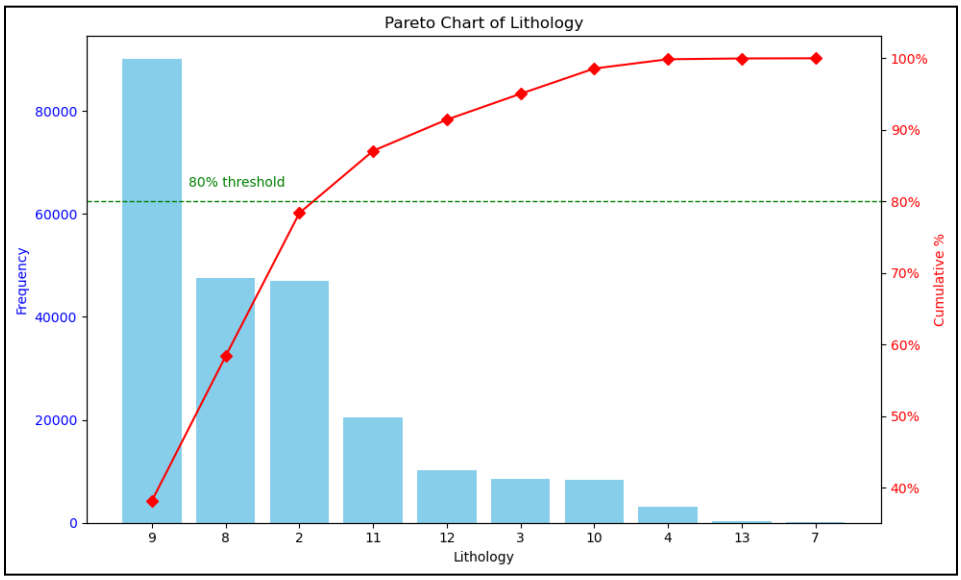

Figure 0.4 Pareto chart of lithology frequency showing that three dominant lithologies (9, 8, and 2) comprise roughly 80% of all samples.

4.3.3.3 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 4.5), 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 4.6) 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.

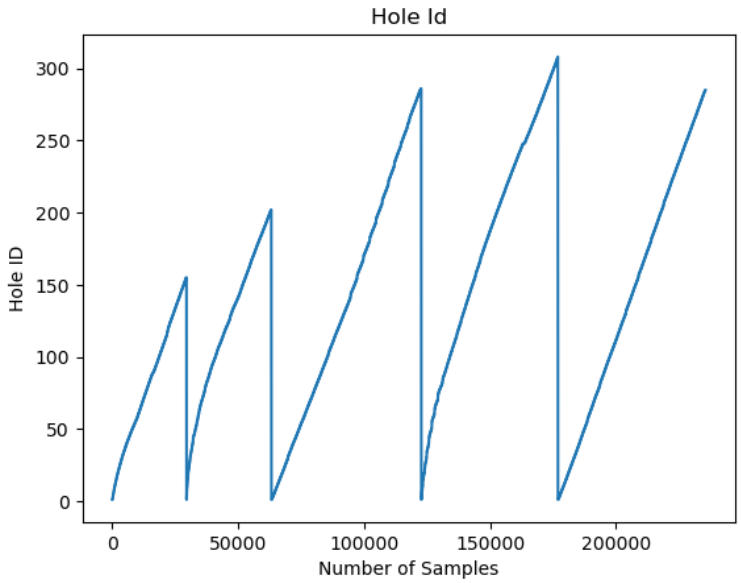

Figure 0.5 Discontinuous Hole ID numbering observed in the raw MWD data across multiple blast patterns.

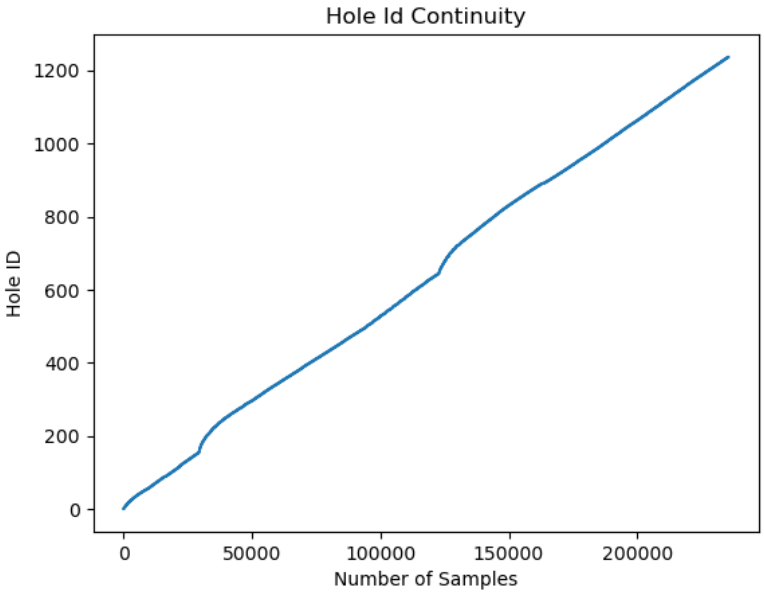

Figure 0.6 Continuous Hole ID sequence after renumbering, ensuring consistent spatial indexing.

4.3.3.4. 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 4.7) 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. 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 4.8) shows the remaining meaningful relationships among 11 retained parameters.

These results confirmed that key operational features such as 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.

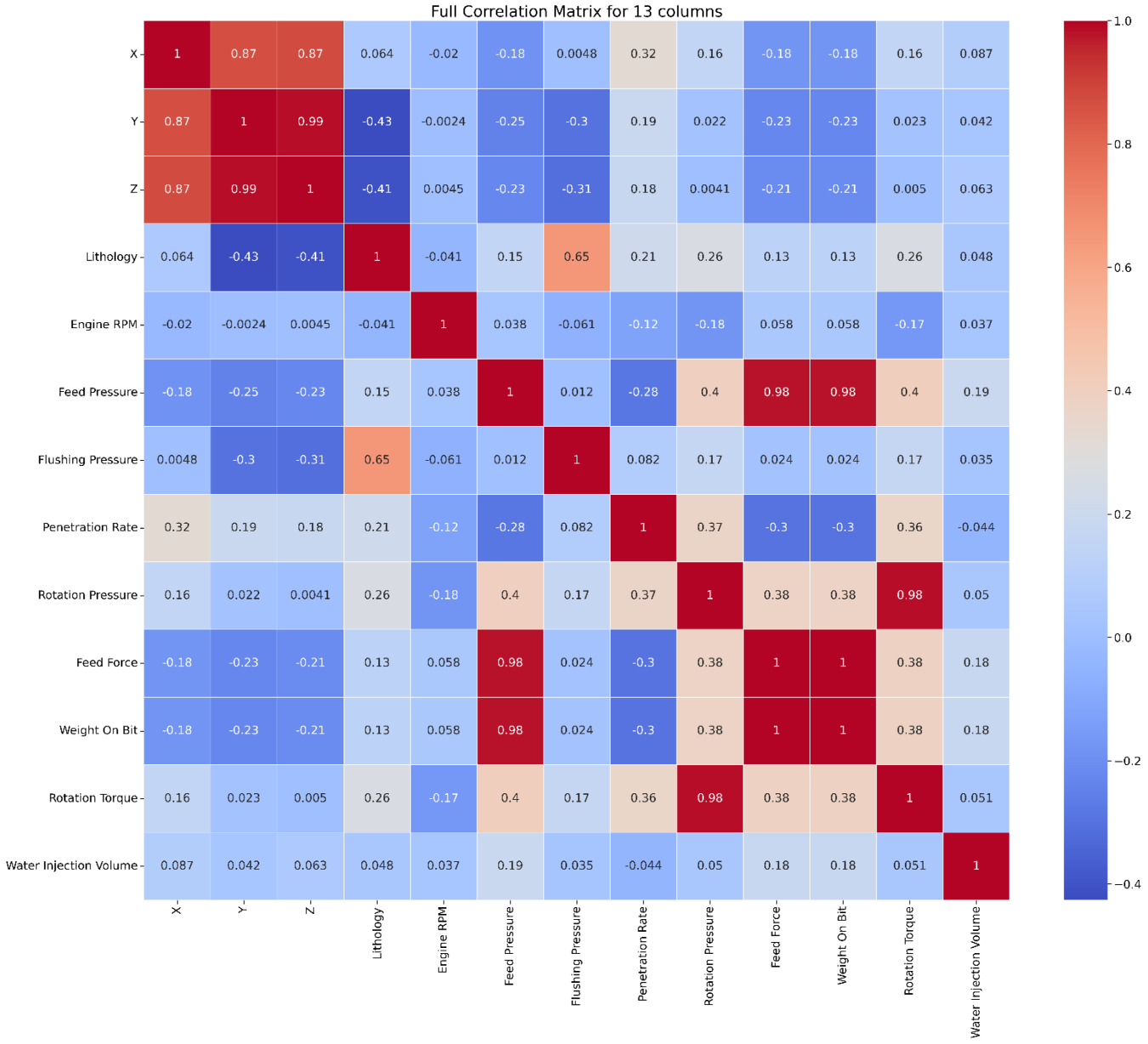

Figure 0.7 Full Pearson correlation matrix for MWD parameters, illustrating all pairwise linear relationships.

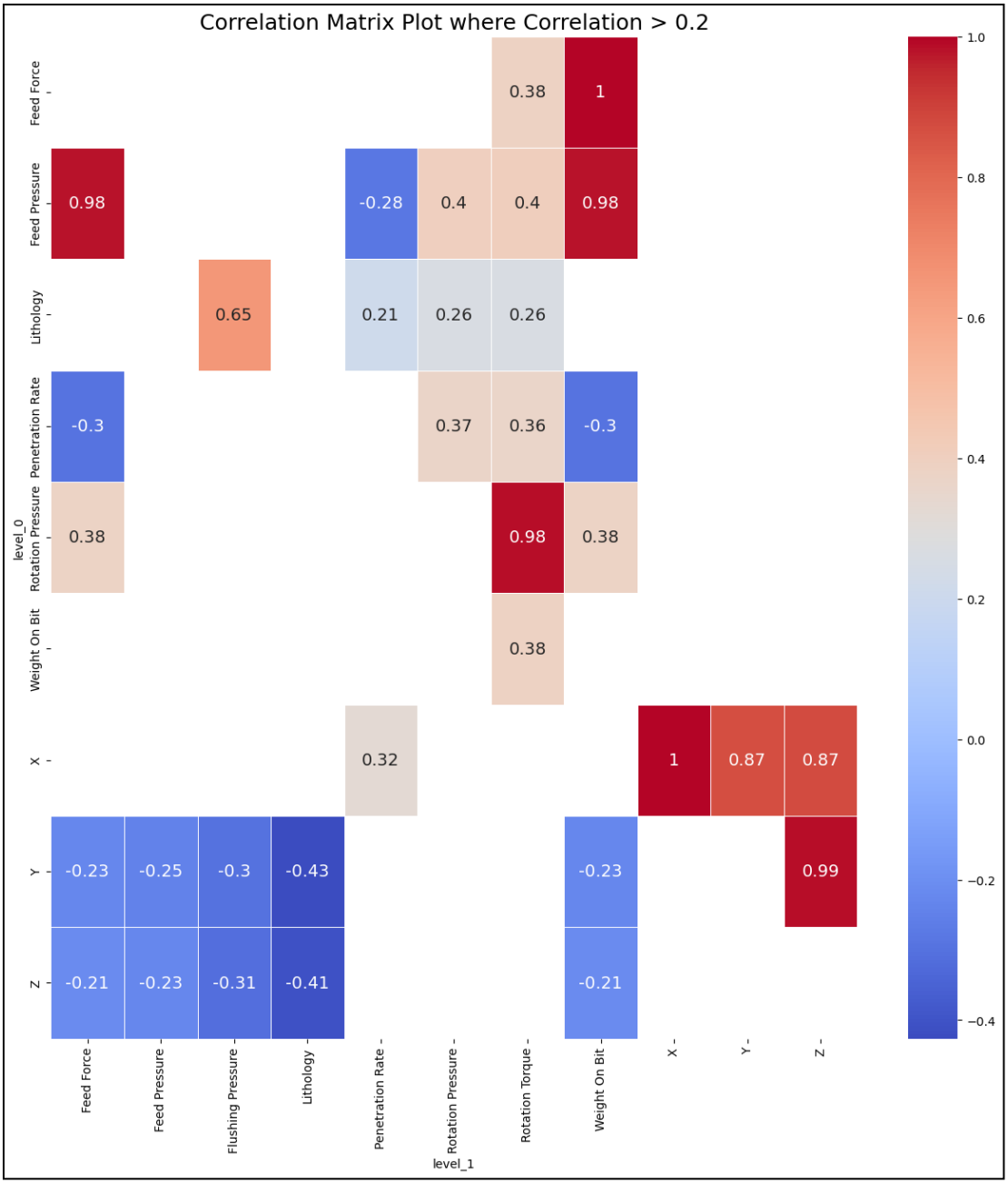

Figure 0.8 Filtered Pearson correlation matrix for MWD parameters, showing only correlations with |r| > 0.2.

4.3.3.4 MWD Parameters Behavior and Operational Anomalies

The earlier correlation analysis revealed that Penetration Rate appeared negatively (though weakly) correlated with Engine RPM, Feed Pressure, Feed Force, and Weight on Bit. This observation was physically unlikely under typical drilling conditions; higher mechanical input should 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 (Figures 4.9 – 4.11).

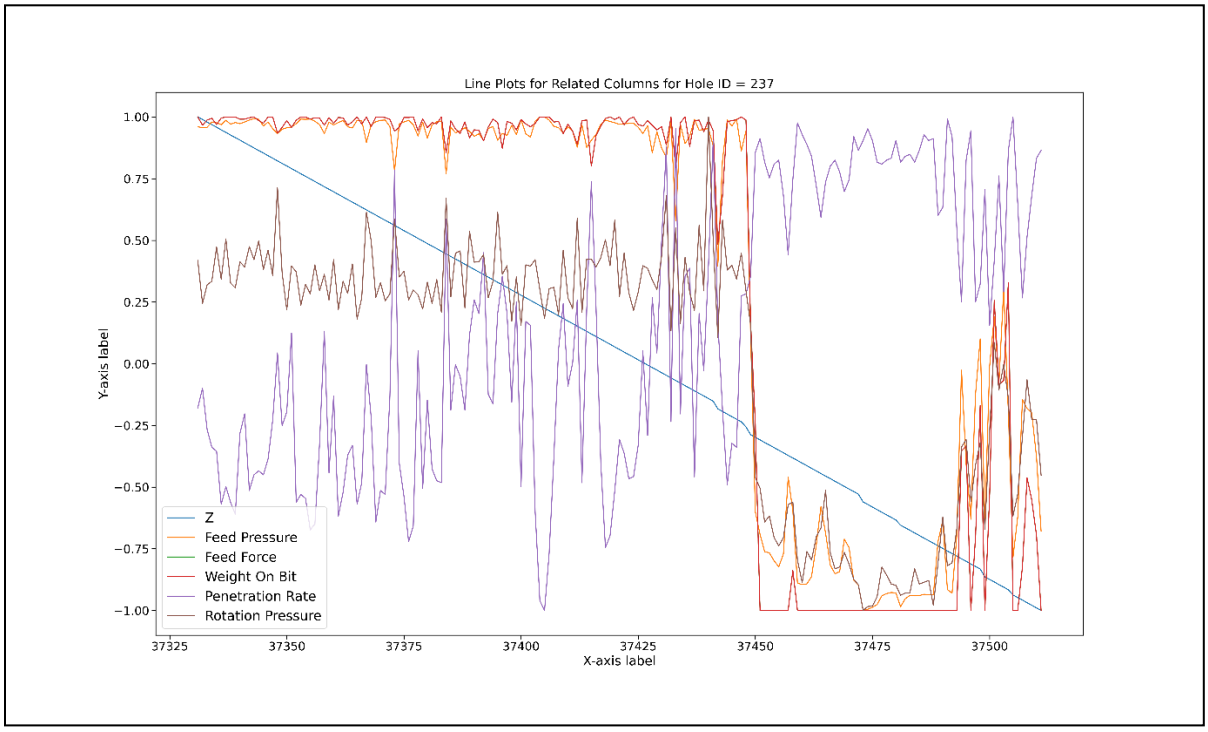

Figure 0.9 Depth-wise variation of feed-, weight-, and pressure-related MWD parameters for Hole ID 237.

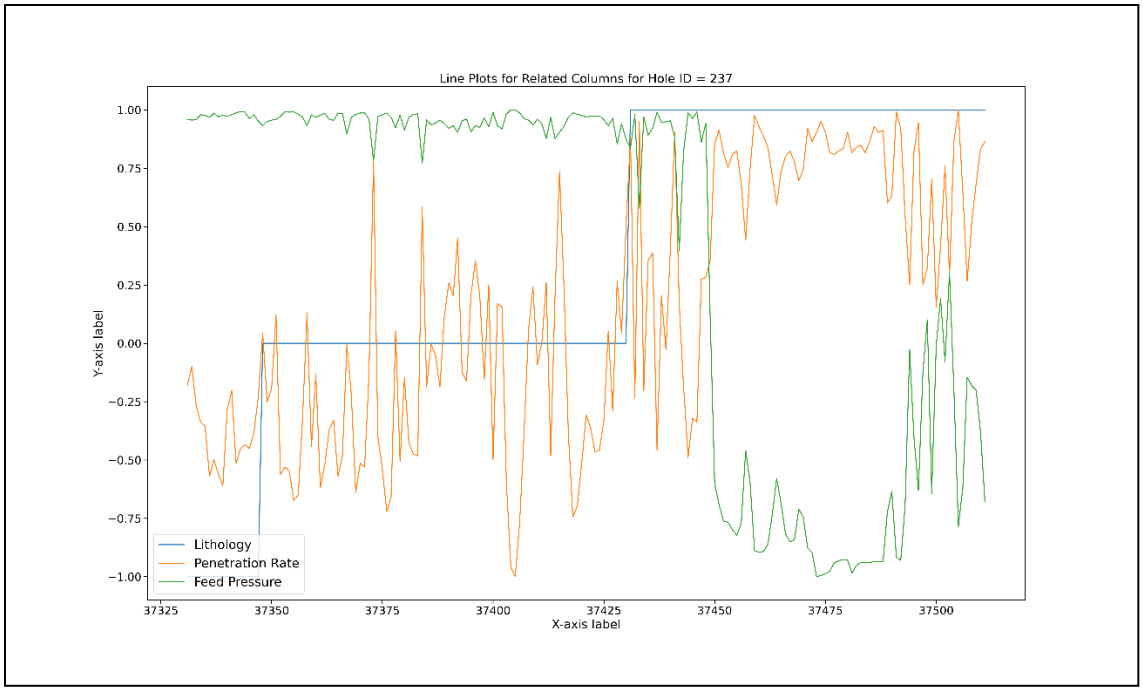

Figure 0.10 Depth-wise variation of lithology, penetration rate, and feed pressure for Hole ID 237.

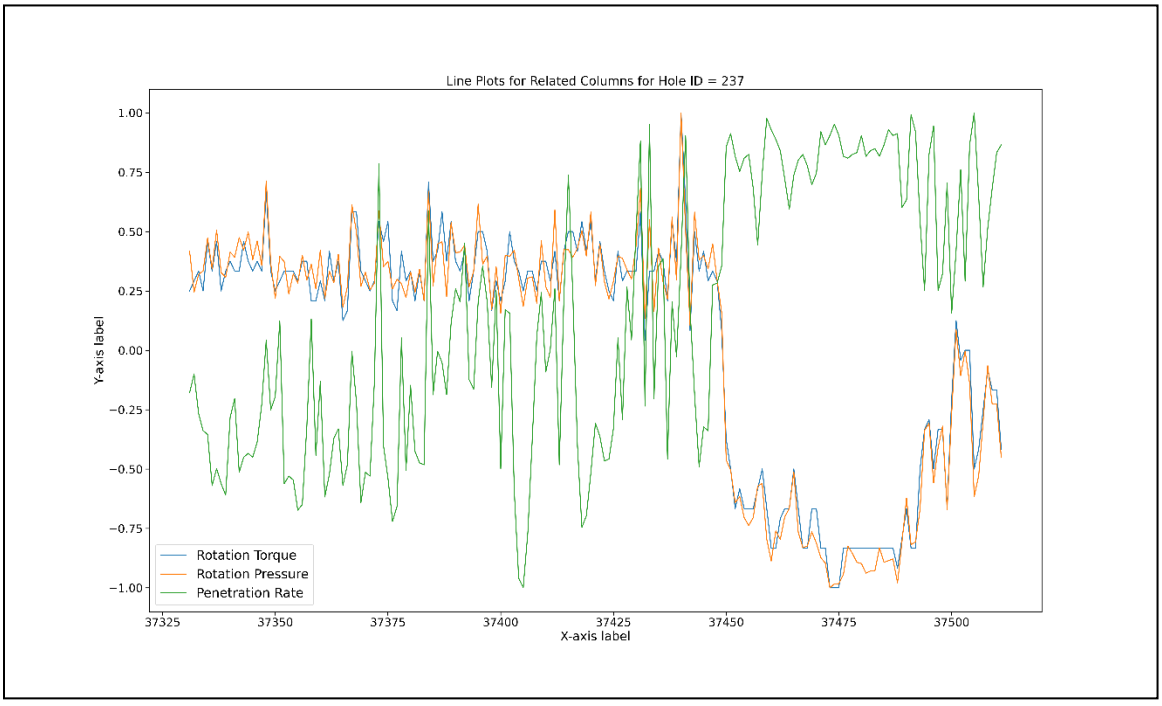

Figure 0.11 Depth-wise variation of torque- and rotation-related MWD parameters for Hole ID 237.

As shown in Figures 4.9 – 4.11, the parameters maintained consistent relationships with depth until approximately sample 37450, 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 4.9, Feed Pressure, Feed Force, Weight on Bit, and Rotation Pressure remain steady and well-correlated through most of the borehole. Near depth 37450, 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 4.10 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 4.11, 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.

4.3.3.5. 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.

Following this refinement, the same parameter relationships were re-examined for Hole ID 237. As shown in Figures 4.12 – 4.14, the anomalous behavior observed earlier near sample 37450 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.

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.

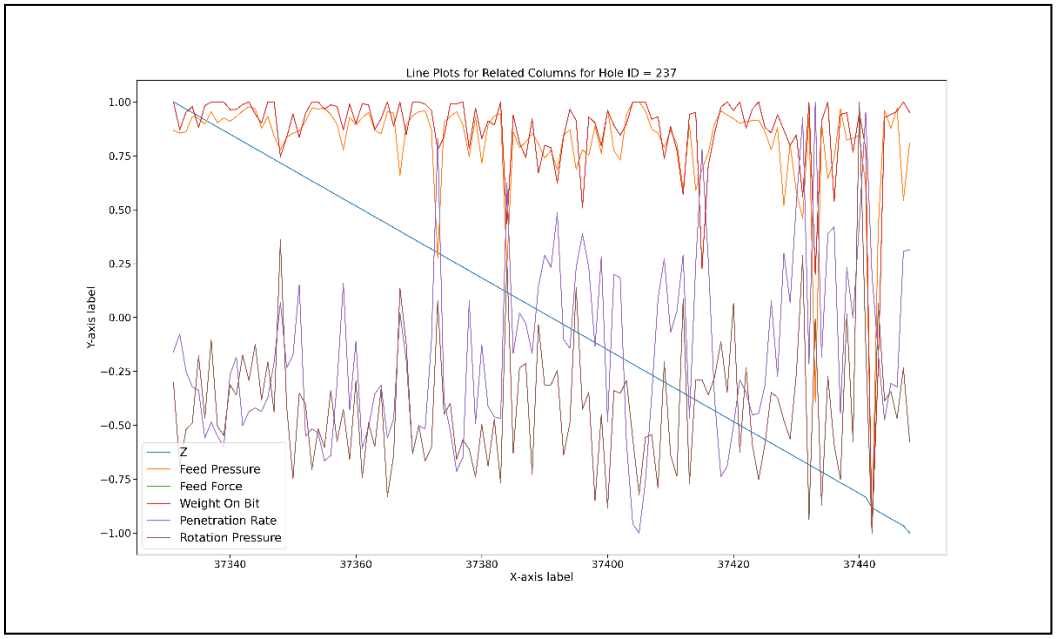

Figure 0.12 Depth-wise variation of feed-, weight-, and pressure-related MWD parameters for Hole ID 237, after filtering out low WOB intervals (< 20th percentile).

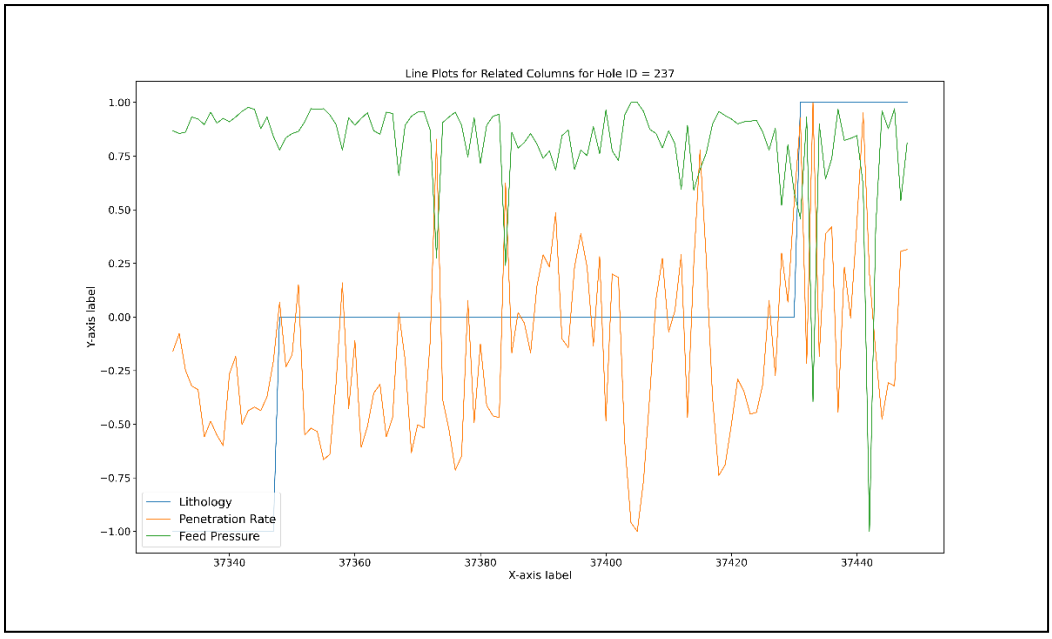

Figure 0.13 Depth-wise variation of lithology, penetration rate, and feed pressure for Hole ID 237, after filtering out low WOB intervals (< 20th percentile).

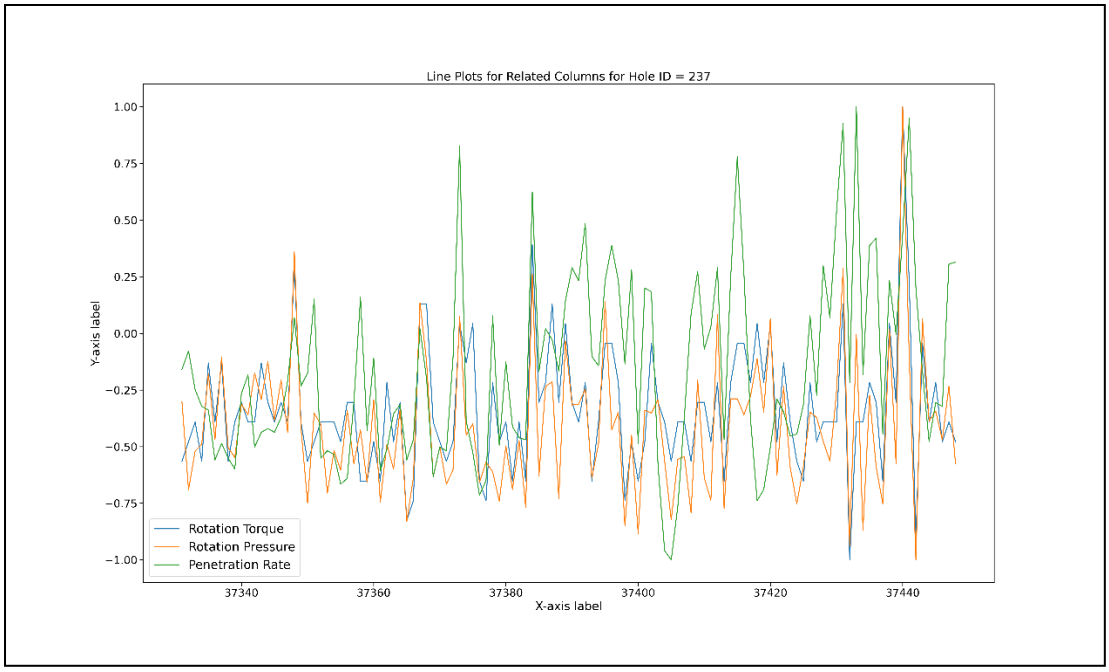

Figure 0.14 Depth-wise variation of torque- and rotation-related MWD parameters for Hole ID 237, after filtering out low WOB intervals (< 20th percentile).

To quantify the effect of filtration, descriptive statistics were calculated for key MWD parameters before and after filtering the data. Table 4.3 summarizes representative metrics (mean, standard deviation, and selected percentiles).

Table 0.3 Comparison of key MWD parameter statistics before and after filtering low WOB 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	66286.5	12143.8	52905.24	69976.22	75948.33	77719.04
	Filtered	70995.02	4739.35	63263.03	71965.67	75948.33	77719.04
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	66331.62	12151.97	52939.6	70023.77	75999.96	77771.92
	Filtered	71043.31	4742.55	63306.29	72014.55	75999.96	77771.92

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.

4.4 Modeling Framework and Approach

The cleaned and filtered dataset 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 using the scikit-learn machine learning library, supplemented with visualization and evaluation tools.

4.4.1 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. 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 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.

4.4.2 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.

4.4.3 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:

i. Penetration Rate Regression: Linear Regression (LR), Decision Tree Regression (DTR), Random Forest Regression (RFR), MLP Regressor
ii. 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.

4.4.4 Model Evaluation and Performance Metrics

Model performance was assessed using both quantitative metrics and diagnostic visualization tools. 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.

4.5 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.

4.5.1 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. Figures 8a–8h illustrate the true-predicted relationships and residual behaviors for each model.

4.5.1.1 Linear Regression (LR)

The Linear Regression model served as the baseline. As shown in Figures 4.15 – 4.16, it captured the general increasing trend between actual and predicted penetration rates but systematically underestimated higher values.

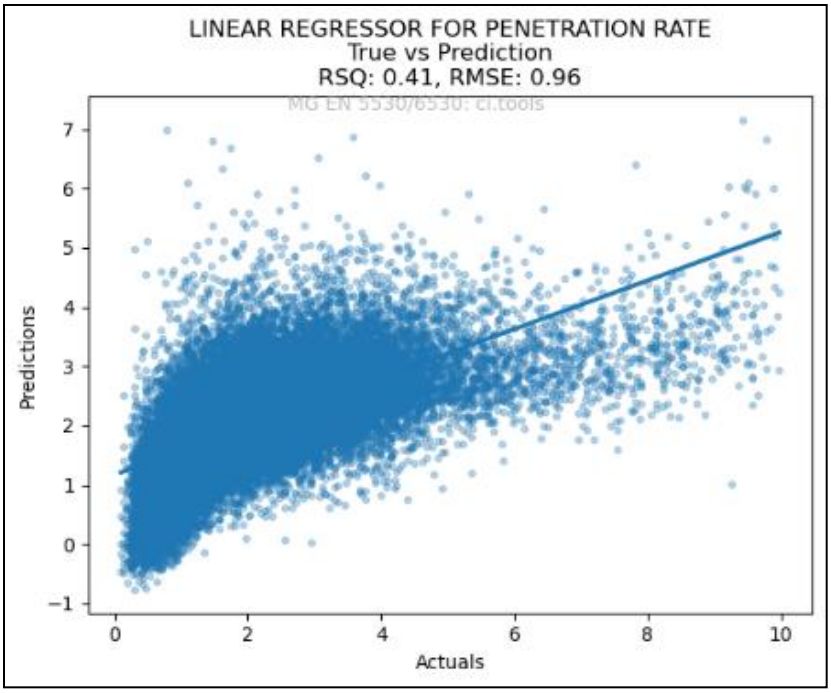

Figure 0.15 True vs. predicted penetration rate using a linear regressor, showing overall trend and underestimation at higher values ($R^2 = 0.41$, RMSE = 0.96).

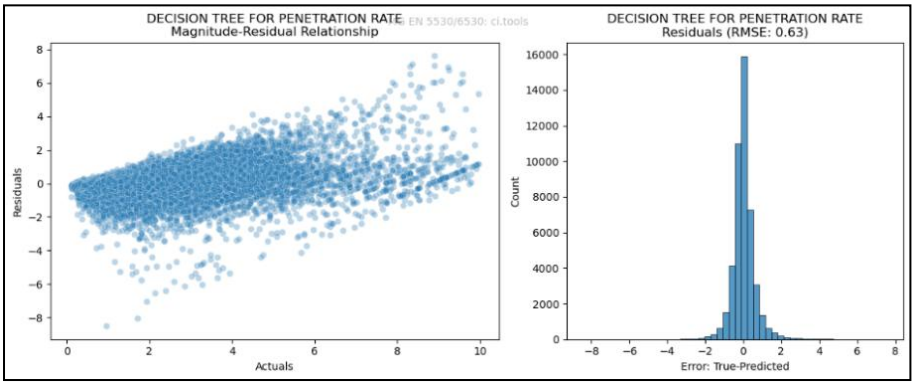

Figure 0.16 Residual analysis for the linear regressor, including magnitude–residual relationship and residual distribution histogram.

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.

4.5.1.2. Decision Tree Regression (DTR)

Introducing nonlinear decision boundaries improved prediction fidelity. As shown in Figures 4.17 – 4.18, residual dispersion narrowed relative to the linear model, and the histogram exhibited a sharp, symmetric distribution.

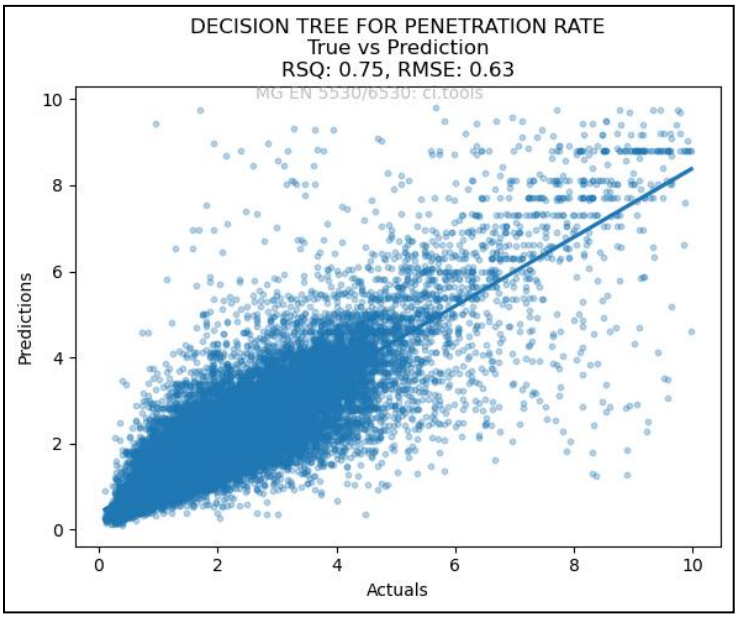

Figure 0.17 True vs. predicted penetration rate using a Decision Tree Regressor, showing improved fit across the full range of observations ($R^2 = 0.75$, RMSE = 0.63).

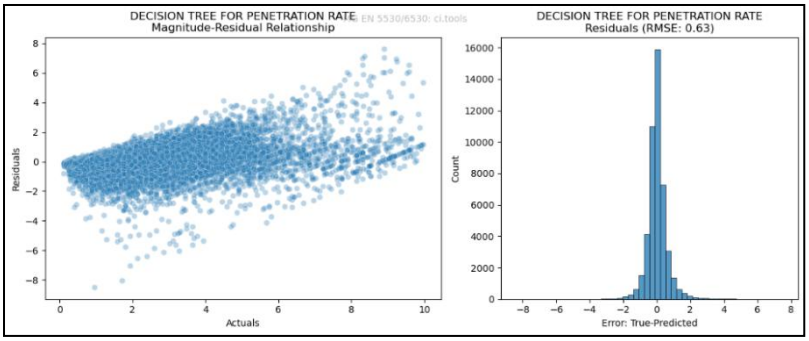

Figure 0.18 Residual analysis for the Decision Tree Regressor, including magnitude–residual relationship and residual distribution histogram.

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.

4.5.1.3. Random Forest Regression (RFR)

The Random Forest ensemble delivered the highest overall predictive accuracy and stability. Figures 4.19 – 4.20 show a near-linear alignment between predicted and actual values with minimal deviation.

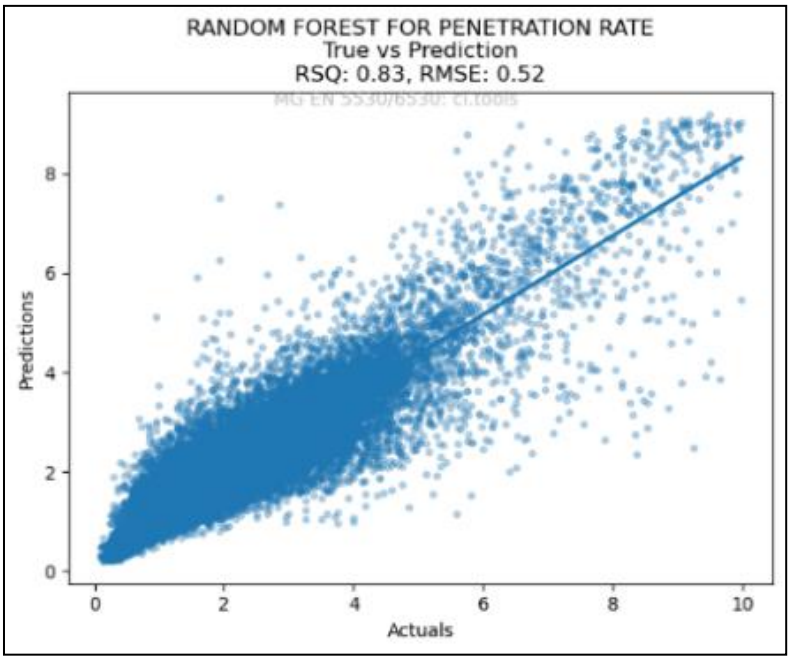

Figure 0.19 True vs. predicted penetration rate using a Random Forest Regressor, showing strong alignment and minimal dispersion ($R^2 = 0.83$, RMSE = 0.52).

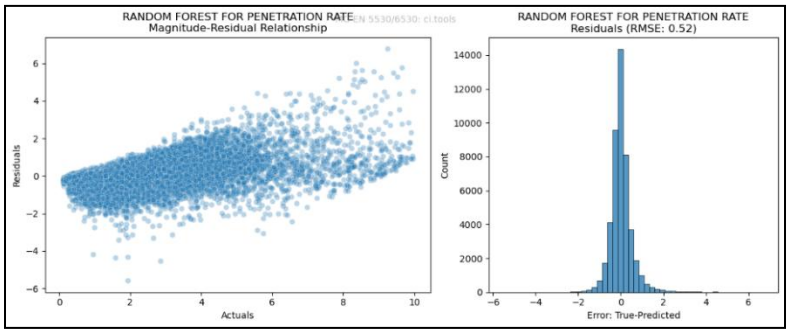

Figure 0.20 Residual analysis for the Random Forest Regressor, including magnitude–residual relationship and residual distribution histogram.

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.

4.5.1.4. Multi-Layer Perceptron (MLP) Regressor

The 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 Figures 4.21 – 4.24, both training and testing phases produced nearly identical outcomes, with low R^2 and high RMSE values. The model failed to capture the underlying nonlinear structure in the data, indicating underfitting rather than overfitting.

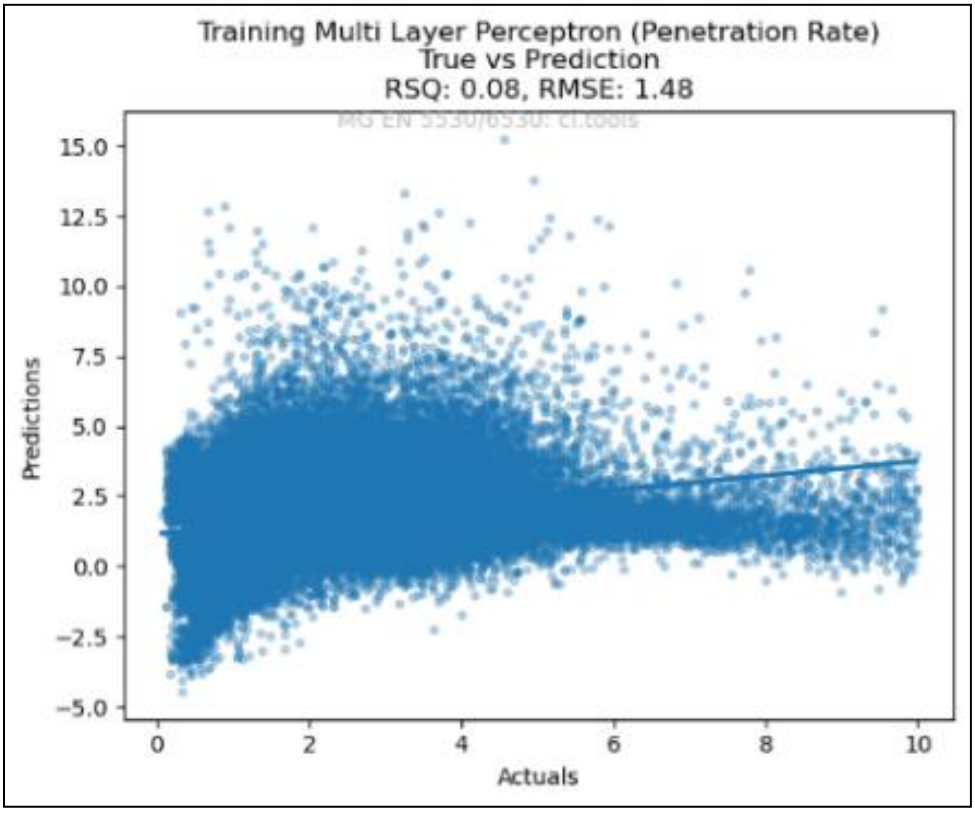

Figure 0.21 True vs. predicted penetration rate for the training dataset using a Multi-Layer Perceptron Regressor, showing strong correlation and close fit to the ideal line (R^2 = 0.08, RMSE = 1.48).

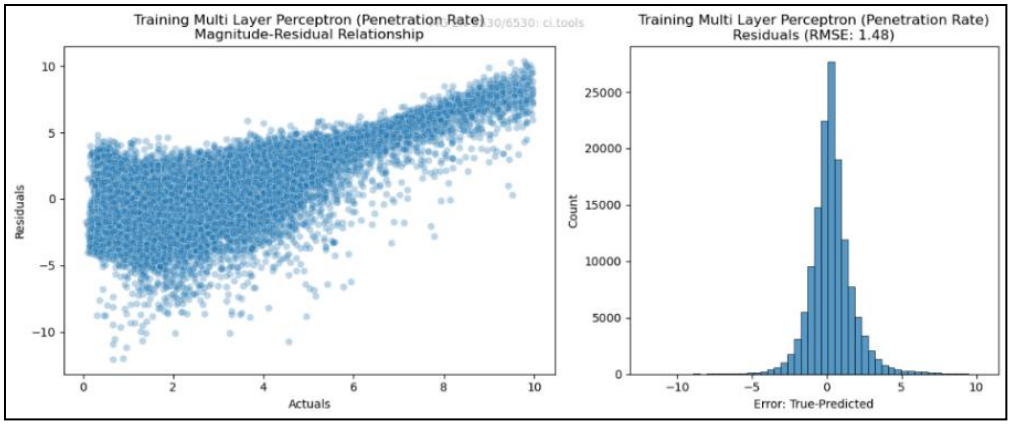

Figure 0.22 Residual distribution for the training dataset, indicating low bias and stable performance across penetration rate magnitudes.

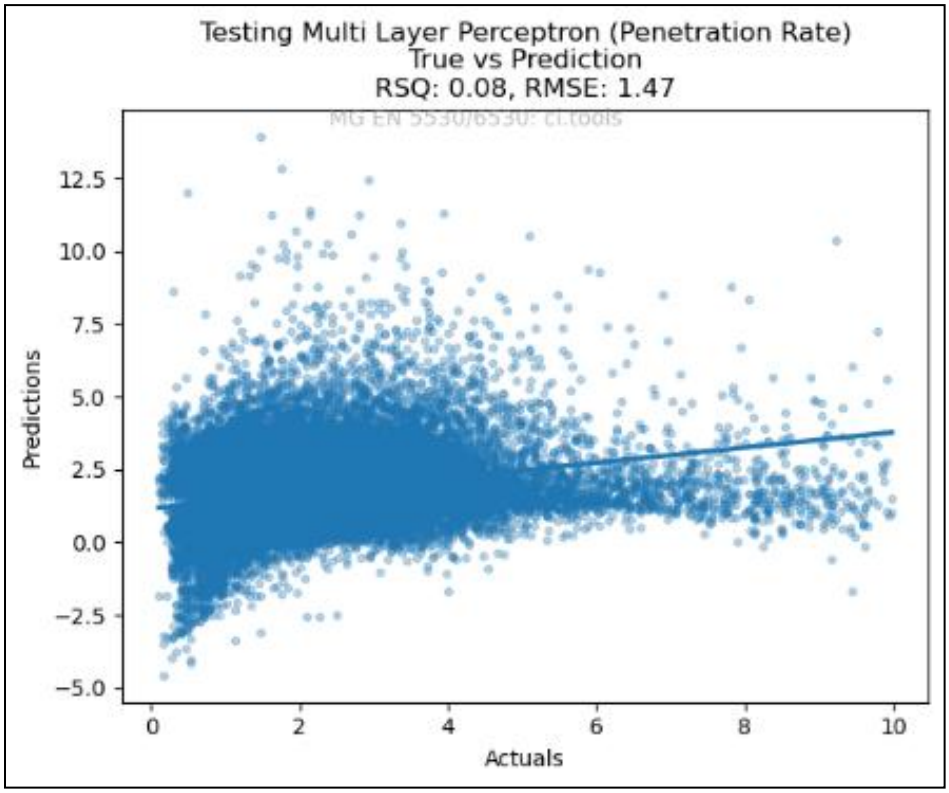

Figure 0.23 True vs. predicted penetration rate for the testing dataset using a Multi-Layer Perceptron Regressor, showing weak generalization and underestimation at higher penetration rates (R^2 = 0.08, RMSE = 1.47).

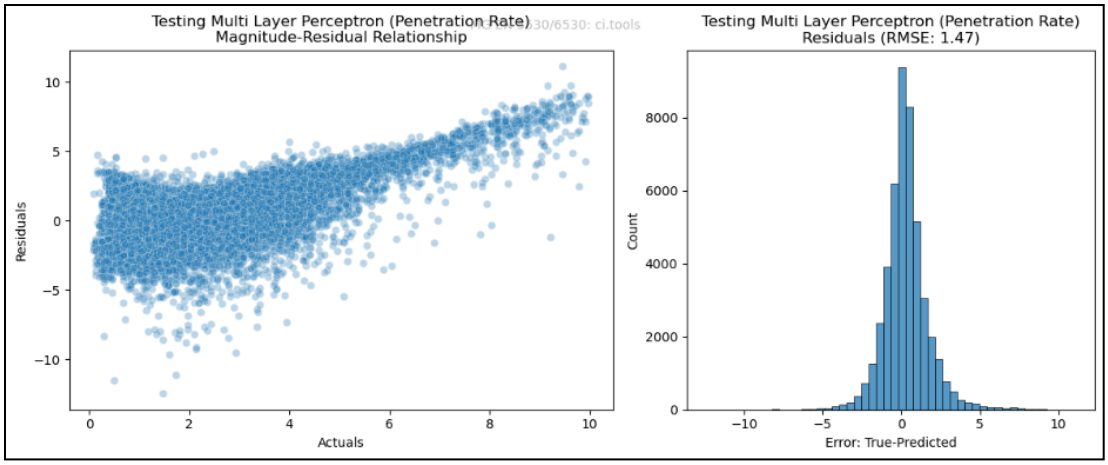

Figure 0.24 Residual distribution for the testing dataset, displaying large variance and diffuse patterns indicative of model overfitting.

Residual plots for both the training and testing datasets (Figures 4.22 and 4.24) 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.

This underperformance likely stems from suboptimal network depth and lack of hyperparameter tuning. While neural networks can theoretically model complex nonlinear dependencies, the current configuration failed to converge to a meaningful representation under default training conditions.

4.5.2 Summary of Regression Results

Model performance for 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 4.4.

Table 0.4 Summary of regression model performance for Penetration Rate (PR)

Model	R^2	RMSE	Remarks
Linear Regression	0.41	0.96	Baseline model; underestimates highs
Decision Tree Regressor	0.75	0.63	Captures nonlinear trends
Random Forest Regressor	0.83	0.52	Best performance
Multi-Layer Perceptron Regressor	0.08	1.48	Overfitting evident; weak generalization

4.5.3 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 frequently intersected formation, and Lithology 9, the rarest unit in the dataset. These cases provide insight into each model's ability to generalize across class imbalance and capture spatial variability.

4.5.3.1 Decision Tree Classifier

The Decision Tree reached an overall accuracy of 97.5 %, establishing a strong baseline for comparison. The confusion matrix (Figure 4.25) 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 4.26) closely match the true distribution with minimal deviation, while predictions for Lithology 9 (Figure 4.27) show slight overextension around the cluster margins. These patterns suggest that the model captured dominant lithologies effectively but tended to overpredict the boundaries of less-represented units.

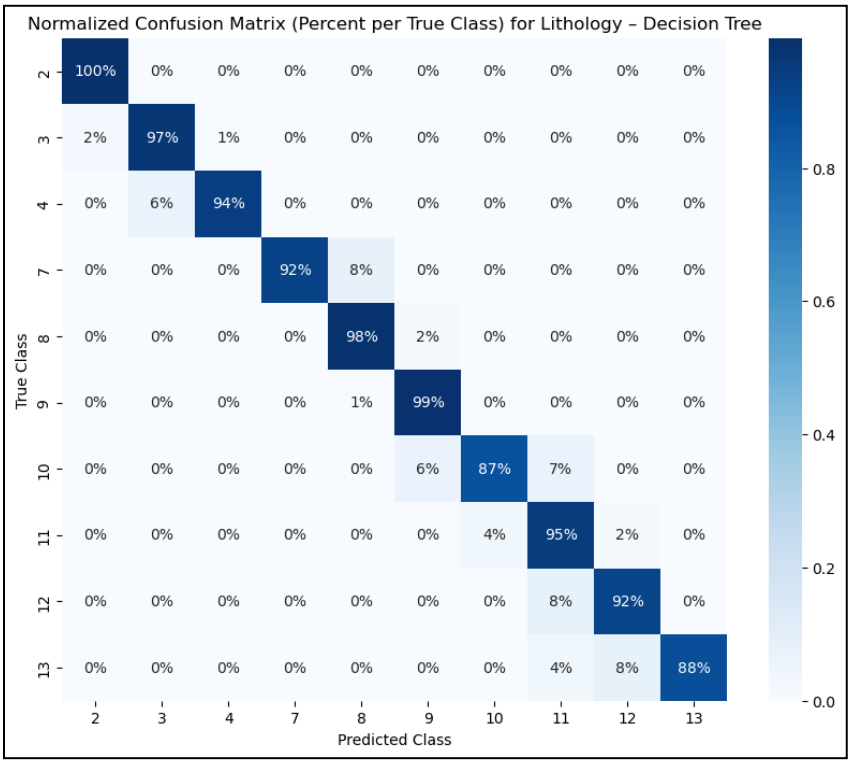

Figure 0.25 Normalized confusion matrix for the Decision Tree classifier.

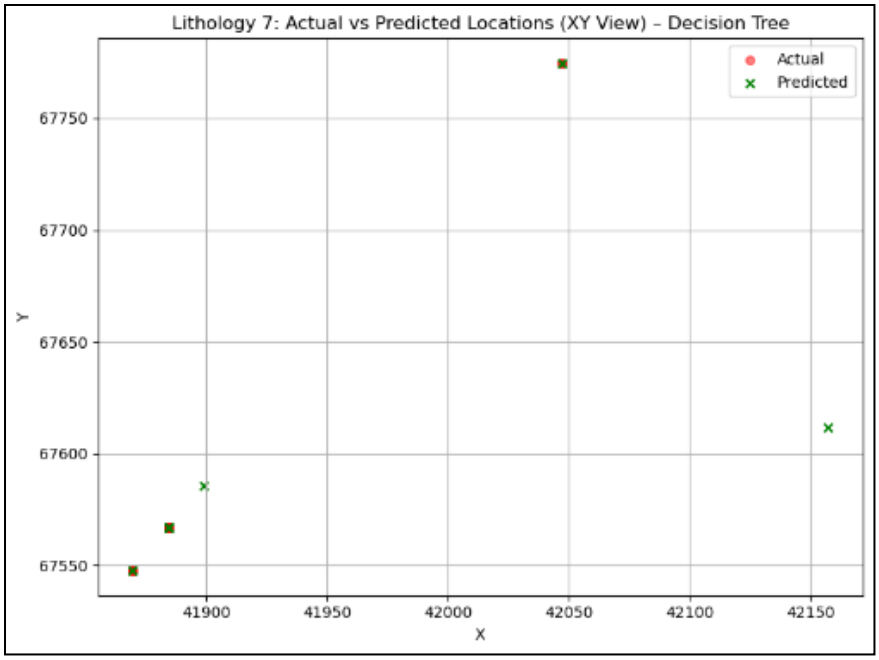

Figure 0.26 Spatial comparison of predicted versus actual locations for Lithology 7 using the Decision Tree classifier.

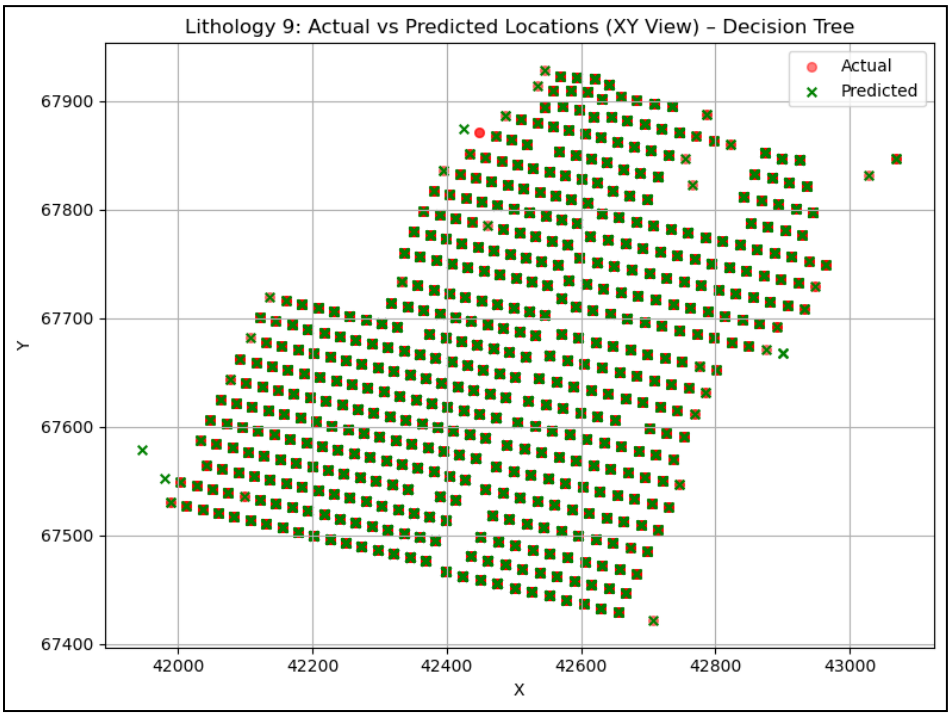

Figure 0.27 Spatial comparison of predicted versus actual locations for Lithology 9.

4.5.3.2 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 4.28) displays a sharply defined diagonal, confirming highly consistent predictions across all classes. Spatial predictions for Lithology 7 (Figure 4.29) show excellent alignment between predicted and true zones, while Lithology 9 (Figure 4.30) 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.

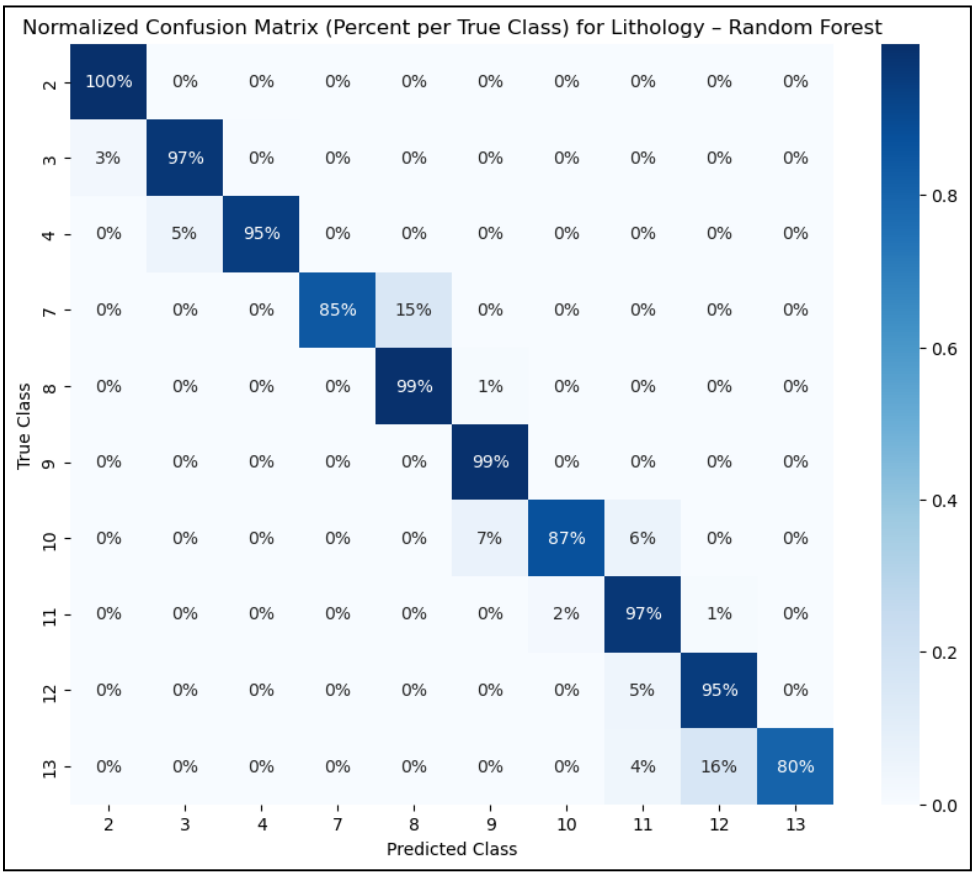

Figure 0.28 Normalized confusion matrix for the Random Forest classifier.

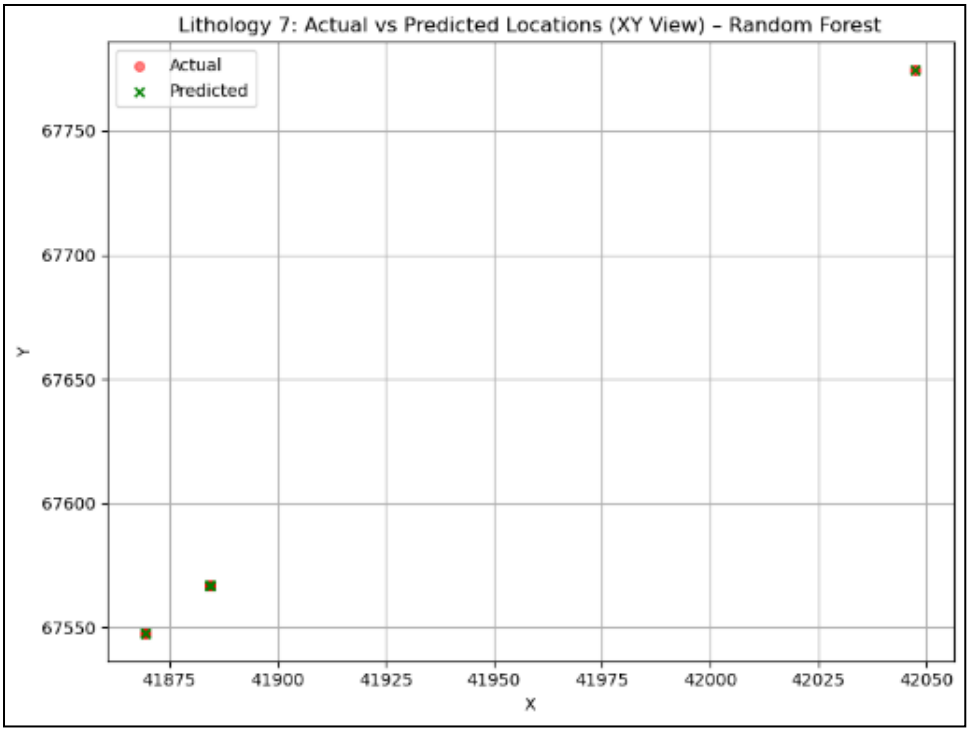

Figure 0.29 Spatial comparison of predicted versus actual locations for Lithology 7 using the Random Forest classifier.

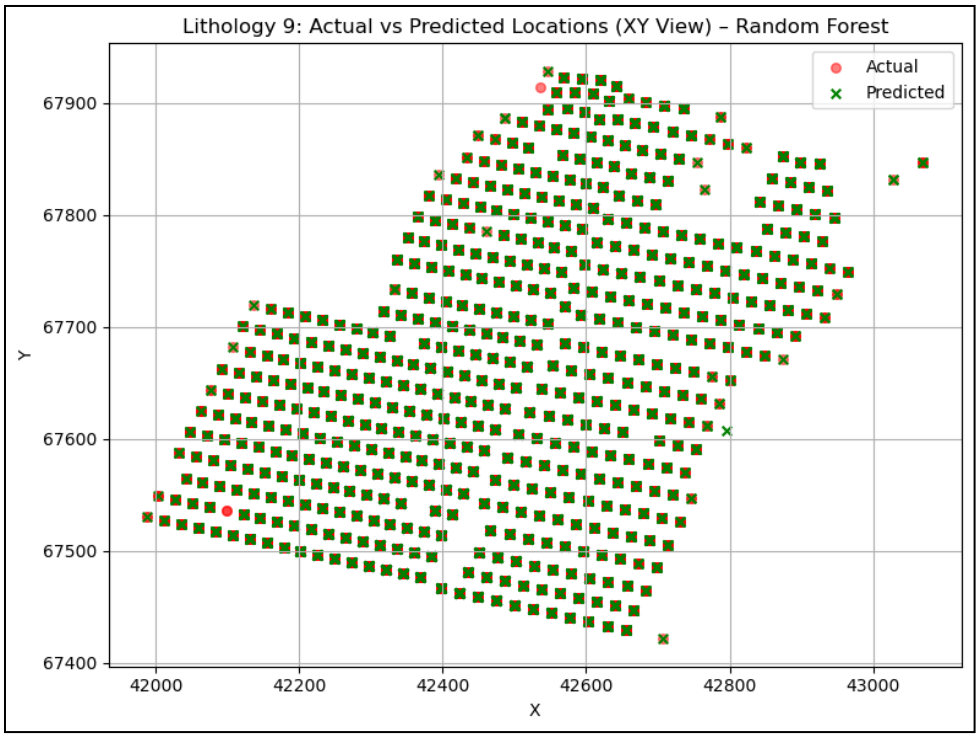

Figure 0.30 Spatial comparison of predicted versus actual locations for Lithology 9 using the Random Forest classifier.

4.5.3.3 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 4.30) reveals strong performance on dominant formations but a small reduction in precision for minor classes. Spatial predictions for Lithology 7 (Figure 4.31) reproduce the main lithological pattern with high fidelity, while Lithology 9 (Figure 4.32) appears somewhat diffuse, indicating the network's sensitivity to class imbalance. Nevertheless, the MLP effectively captured key stratigraphic trends and overall structural continuity.

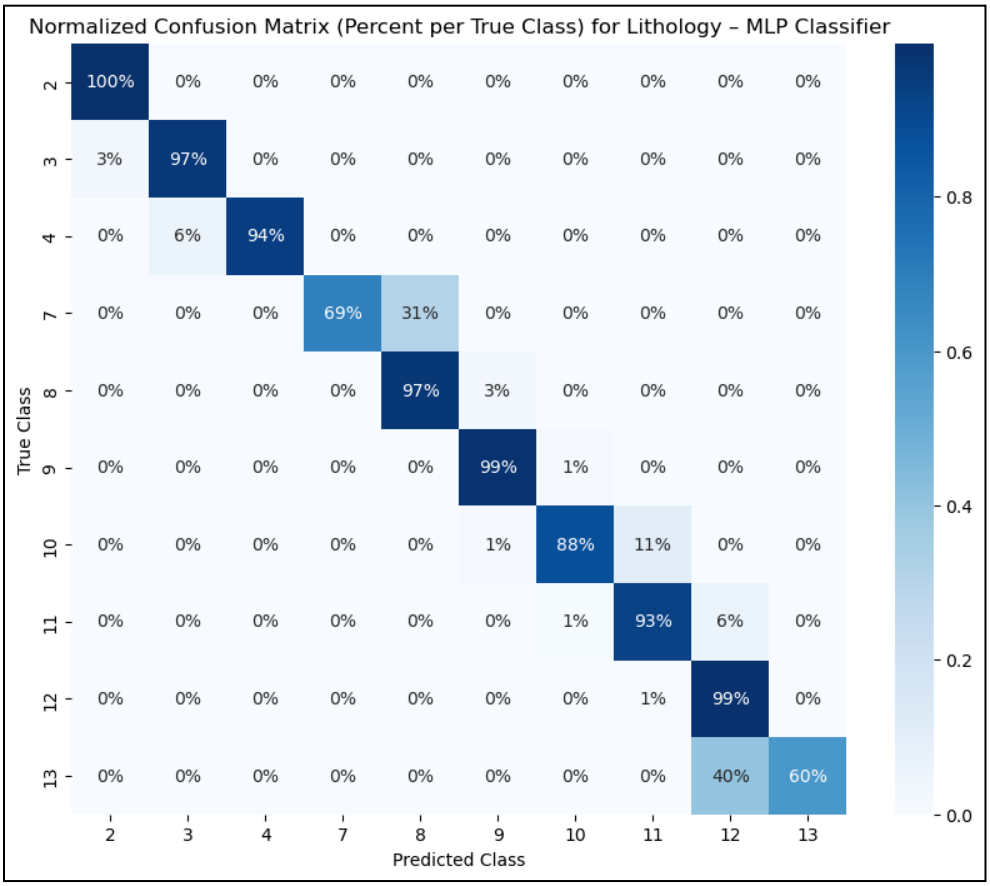

Figure 0.31 Normalized confusion matrix for the MLP classifier.

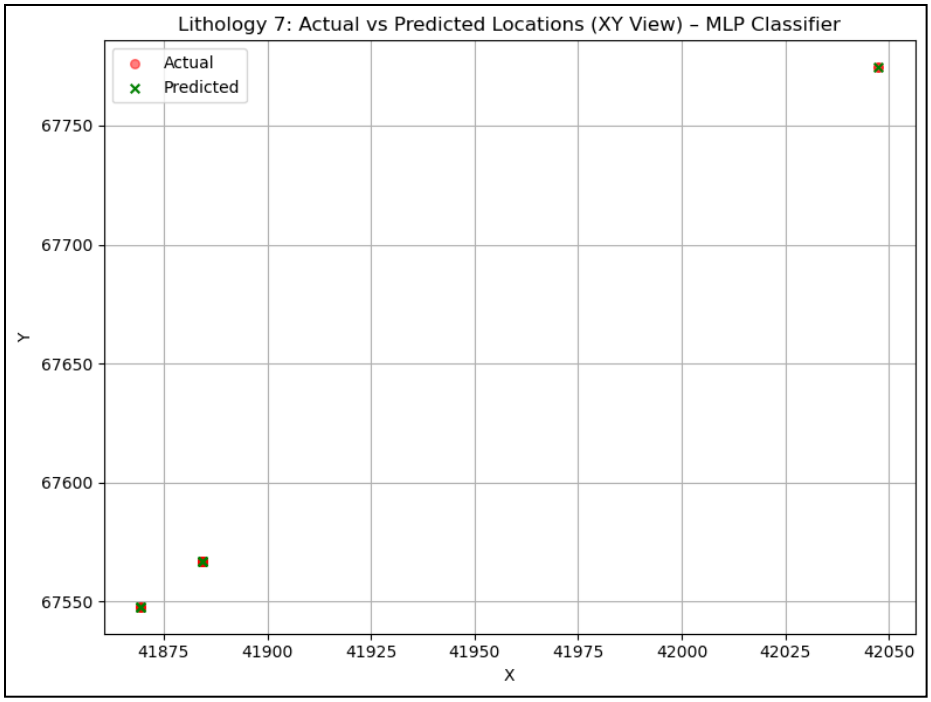

Figure 0.32 Spatial comparison of predicted versus actual locations for Lithology 7 using the MLP classifier.

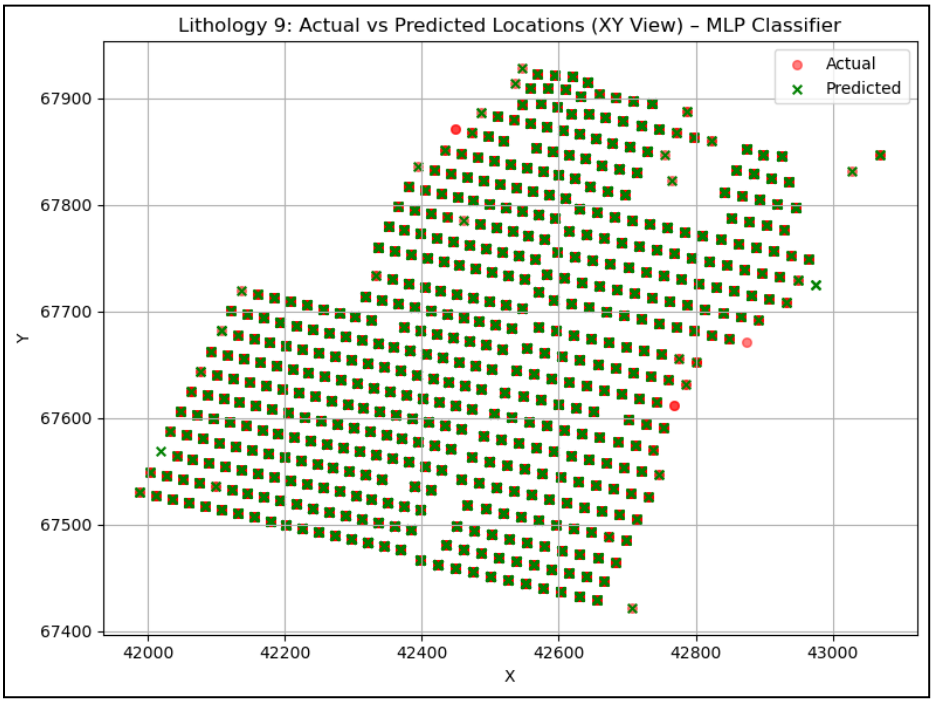

Figure 0.33 Spatial comparison of predicted versus actual locations for Lithology 9 using the MLP classifier.

4.5.4 Summary of Classification Results

All classifiers achieved high accuracy, exceeding 97 %, indicating strong predictive capability for lithological variation from MWD parameters. 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 4.5.

Table 0.5 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

4.6 Discussion

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.

4.6.1 Influence of Exploratory Analysis on Model Development

The comprehensive correlation study (Figures 4.7 – 4.8) 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 (Figures 4.9 – 4.11) further revealed operational anomalies, most notably the abrupt inverse behavior near sample 37450, where Penetration Rate spiked as

mechanical loads dropped. This pattern, initially attributed to lithological change, was traced to non-productive drilling activities such as hole cleaning. The subsequent filtration procedure which involved 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 (Figures 4.12 – 4.14) and statistically (Table 4.3), 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.

4.6.2 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.

Tree-based models 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 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.

4.6.3 Interpretation of Lithology Classification

For lithology prediction, all three classifiers namely Decision Tree, Random Forest, and MLP achieved high accuracies exceeding 97%, indicating 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.

4.6.4 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.

4.7 Conclusions

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

The 20th-percentile WoB 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.

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.

4.8 References

[1] "Exploratory-Data-Analysis-1977-John-Tukey."

[2] R. Church, "HOW TO LOOK AT DATA." Journal of the Experimental Analysis of Behavior, 1979.

[3] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Exploratory Data Analysis," in *Secondary Analysis of Electronic Health Records*, Cham: Springer International Publishing, 2016, pp. 185–203. doi: 10.1007/978-3-319-43742-2_15.

[4] I. O. Muraina *et al.*, "The Necessity of Exploratory Data Analysis How are preprocessing activities beneficial to Data Analysts and Professional Researchers in Academia," *IJSRCSE*, vol. 11, no. 3, pp. 22–28, June 2023, doi: 10.26438/ijsrcse/v11i3.2228.

[5] K. Wongsuphasawat, Y. Liu, and J. Heer, "Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study," Nov. 01, 2019, *arXiv*: arXiv:1911.00568. doi: 10.48550/arXiv.1911.00568.

[6] F. C. Oettl *et al.*, "The artificial intelligence advantage: Supercharging exploratory data analysis," *Knee surg. sports traumatol. arthrosc.*, vol. 32, no. 11, pp. 3039–3042, Nov. 2024, doi: 10.1002/ksa.12389.

[7] S. Putatunda, K. Rama, D. Ubrangala, and R. Kondapalli, "SmartEDA: An R Package for Automated Exploratory Data Analysis," *JOSS*, vol. 4, no. 41, p. 1509, Sept. 2019, doi: 10.21105/joss.01509.

[8] V. Isheyskiy and J. A. Sanchidrián, “Prospects of Applying MWD Technology for Quality Management of Drilling and Blasting Operations at Mining Enterprises,” *Minerals*, vol. 10, no. 10, p. 925, Oct. 2020, doi: 10.3390/min10100925.

[9] V. Isheyskiy, E. Martinyskin, S. Smirnov, A. Vasilyev, K. Knyazev, and T. Fatyanov, “Specifics of MWD Data Collection and Verification during Formation of Training Datasets,” *Minerals*, vol. 11, no. 8, p. 798, July 2021, doi: 10.3390/min11080798.

[10] Florida Department of Transportation, “Florida Method of Test for MWD.”

[11] K. Li *et al.*, “Real-time lithology identification while drilling based on drilling parameters analysis with machine learning,” *Geomech. Geophys. Geo-energ. Geo-resour.*, vol. 11, no. 1, p. 44, Dec. 2025, doi: 10.1007/s40948-025-00951-5.

[12] O. Akyildiz, H. Basarir, and S. L. Ellefmo, “The development of a lithology prediction model using measurement while drilling data in a quartzite quarry,” *International Journal of Mining, Reclamation and Environment*, vol. 39, no. 2, pp. 93–109, Feb. 2025, doi: 10.1080/17480930.2024.2362577.

[13] T. F. Hansen, G. H. Erharter, Z. Liu, and J. Torresen, “A comparative study on machine learning approaches for rock mass classification using drilling data,” *Applied Computing and Geosciences*, vol. 24, p. 100199, Dec. 2024, doi: 10.1016/j.acags.2024.100199.

[14] D. Goldstein, C. Aldrich, and L. O’Connor, “Enhancing Orebody Knowledge using Measure-While-Drilling Data: A Machine Learning Approach,” *IFAC-PapersOnLine*, vol. 58, no. 22, pp. 72–76, 2024, doi: 10.1016/j.ifacol.2024.09.293.

[15] I. Anafo, R. Ganguli, and N. Sarantsatsral, “BoxRF: A New Machine Learning Algorithm for Grade Estimation,” *Applied Sciences*, vol. 15, no. 8, p. 4416, Apr. 2025, doi: 10.3390/app15084416.

[16] G. C. Komadja, E. Westman, A. Rana, and A. Vitalis, “A Machine Learning Approach to Lithology Classification in Mining Using Measurement While Drilling and Exploration Data,” *Mining, Metallurgy & Exploration*, vol. 42, no. 4, pp. 1955–1973, Aug. 2025, doi: 10.1007/s42461-025-01286-1.

[17] K. L. Silversides and A. Melkumyan, “Machine learning for classification of stratified geology from MWD data,” *Ore Geology Reviews*, vol. 142, p. 104737, Mar. 2022, doi: 10.1016/j.oregeorev.2022.104737.

[18] T. Burak, A. Sharma, E. Hoel, T. G. Kristiansen, M. Welmer, and R. Nygaard, “Real-Time Lithology Prediction at the Bit Using Machine Learning,” *Geosciences*, vol. 14, no. 10, p. 250, Sept. 2024, doi: 10.3390/geosciences14100250.

[19] S. Manzoor, S. Liaghat, A. Gustafson, D. Johansson, and H. Schunnesson, “Rock mass characterization using MWD data and photogrammetry,” in *Mining Goes Digital*, 1st ed., London: CRC Press, 2019, pp. 217–225. doi: 10.1201/9780429320774-25.

[20] S. Heydari, S. H. Hoseinie, and R. Bagherpour, “Prediction of jumbo drill penetration rate in underground mines using various machine learning approaches and traditional models,” *Sci Rep*, vol. 14, no. 1, p. 8928, Apr. 2024, doi: 10.1038/s41598-024-59753-6.

[21] S. Wu, X. Wang, and Z. Q. Yue, “Addressing Random Variations in MWD Penetration Rate with the DPM Algorithm,” *Sustainability*, vol. 14, no. 20, p. 13456, Oct. 2022, doi: 10.3390/su142013456.

[22] B. Samanta, S. Bandopadhyay, R. Ganguli, and S. Dutta, “Sparse Data Division Using Data Segmentation and Kohonen Network for Neural Network and Geostatistical Ore Grade Modeling in Nome Offshore Placer Deposit,” *Natural Resources Research*, vol. 13, no. 3, pp. 189–200, Sept. 2004, doi: 10.1023/B:NARR.0000046920.95725.1b.

Chapter 5: Conclusions and Recommendations

5.1 Summary of Research

This chapter integrates the findings from the two research components of this thesis and situates them within the broader context of mining data analytics. The work demonstrates how MWD data, when combined with machine learning techniques, can enhance subsurface characterization in open-pit mines. The chapter provides a consolidated discussion of the results, outlines the contributions of the research, identifies limitations, and proposes directions for future work.

The overall purpose of this thesis was to evaluate how MWD data could be used not only for operational drilling control, but also as a source of geological and geotechnical insight. Two independent studies were carried out on datasets from two different open-pit mines, allowing the methods to be tested under distinct geological and operational conditions.

The first study examined geotechnical hazard detection, focusing on the identification of void-prone zones associated with historic underground workings using MWD data. Through analysis of drilling response anomalies and the application of unsupervised learning methods, the study demonstrated how variations in penetration rate, load-related parameters, and pressure can be used to indicate potentially hazardous subsurface conditions. The second study emphasized exploratory data analysis as the primary approach to geological characterization, using statistical summaries, distributional trends, and inter-parameter relationships to investigate lithological variability and drilling behavior. Based on this EDA framework, supervised learning models were subsequently applied to evaluate the capability of MWD parameters to support lithology classification and penetration rate prediction.

Collectively, the two studies indicate that MWD data contain interpretable subsurface information that can be systematically extracted through appropriate preprocessing, exploratory analysis, and modeling, contributing to both geotechnical risk assessment and geological understanding in mining applications.

5.2 Discussion of Key Findings

The void detection study showed that changes in drilling response can serve as indicators of weakened or disturbed ground. A hybrid modeling framework was developed that combined

unsupervised anomaly detection with supervised learning. This approach proved effective despite limited validated data and demonstrated that reliable hazard indicators can be generated even when direct mapping is incomplete.

The geological modeling component highlighted the importance of data structure and quality. Raw MWD data often contains noise, pauses, and machine artifacts, and cannot be used directly. Through exploratory data analysis and feature engineering, meaningful relationships were uncovered between drilling parameters and subsurface geology. Once processed, the data supported reliable predictions of lithology and penetration rate which are outcomes that are difficult to achieve through manual observation alone.

Taken together, the findings show that:

i. MWD data can be repurposed for both geotechnical and geological interpretation.
ii. Machine learning methods can reveal subsurface patterns that are not evident through visual inspection.
iii. Data preparation is often as important as model selection.

5.3 Contributions

This thesis makes several key contributions to the use of measurement-while-drilling (MWD) data for subsurface characterization in mining applications. It demonstrates a practical and implementable workflow for detecting void-prone ground conditions using MWD data, even in settings where independent ground-truth information is limited. The work further shows that careful data cleaning and exploratory data analysis are essential for revealing meaningful drilling response patterns and for enabling accurate lithology classification and penetration rate prediction using supervised learning. In addition, the thesis establishes MWD data as a multi-purpose source of subsurface information that can support both geotechnical hazard awareness and mine planning decisions. Finally, the proposed methodology is designed to be transferable and can be adapted for use in other open-pit mining operations with similar drilling and data collection practices.

5.4 Limitations

The findings of this thesis must be interpreted considering several limitations related to data availability, operational context, and methodological scope. The primary datasets used in this work

consisted of MWD records collected during production drilling, supplemented by a limited number of downhole video logs, and mapped subsurface features. Although MWD data provided high-resolution, continuous drilling response measurements, independent confirmation of subsurface conditions was available only for a small subset of intervals. Confirmed void occurrences identified from video logs and historical mine records were sparse relative to the total volume of drilling data. This constraint reflects the inherent difficulty of obtaining direct subsurface validation in active mining environments rather than a deficiency in data collection practices.

The mining operation from which the data were obtained is characterized by heterogeneous geology, variable drilling conditions, and operational constraints typical of large-scale production settings. Drilling is influenced by multiple interacting factors, including lithological variability, structural features, equipment configuration, and operator-controlled parameters. Although the MWD system captures the combined response of these factors, metadata describing operator identity, rig-specific characteristics, bit condition, and control-mode settings were not consistently available. The absence of such information limits the ability to explicitly separate geological effects from operational variability and may affect the generalizability of model results across different rigs, operators, or sites.

Data accessibility also posed practical challenges. Historical underground workings and void geometries were incompletely documented, and their spatial uncertainty limited the precision with which anomalous drilling responses could be linked to specific subsurface features. Similarly, lithological labels were available only where core logging or reliable geological interpretation had been performed, constraining the size and representativeness of the labeled dataset used for supervised learning. As a result, lithology prediction performance is dependent on the quality and distribution of existing labels and may degrade in zones where geological contacts are complex, highly fractured, or influenced by groundwater conditions.

These limitations highlight several areas where additional data would enable a more robust and comprehensive analysis. Expanded downhole imaging coverage, improved documentation of historical workings, and systematic recording of operator, rig, and bit-condition metadata would strengthen model interpretability and validation. Furthermore, multi-site datasets spanning different geological and operational contexts would allow for cross-site testing and improved assessment of model transferability. Despite these constraints, the present study demonstrates the

potential of MWD data for subsurface characterization and hazard identification, while also providing a clear framework for future work aimed at enhancing data completeness, validation, and model robustness.

5.5 Future Work

Future work should focus on strengthening validation, improving model robustness, and increasing practical applicability of the proposed workflows. A primary need is expanded subsurface validation, as the limited number of confirmed void intervals constrained direct performance assessment. Increased downhole imaging coverage and improved documentation of historic workings would allow more rigorous evaluation of detection reliability and better calibration of anomaly thresholds across different ground conditions.

Additional operational metadata would further enhance model interpretability and generalization. Drilling response reflects not only geology but also operational factors such as rig configuration, control settings, and tool condition. Systematic recording of rig identifiers, bit condition, and machine control modes would allow these effects to be modeled explicitly, reducing confounding and improving transferability of results across drilling campaigns.

Future studies should also examine model performance across multiple benches or sites to evaluate robustness under varying geological and operational conditions. 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.

5.6 Final Remarks

This thesis demonstrates that MWD data, when paired with machine learning, can significantly improve subsurface understanding in open-pit mines. By enabling both geotechnical hazard detection and geological interpretation, the methods developed here support safer, more efficient, and more informed mining operations. The research contributes to the ongoing shift toward data-driven mining and establishes a foundation for future advances in intelligent drilling systems and automated ground characterization.