Unlocking Subsurface Geology: A Case Study with Measure-While-Drilling Data and Machine Learning

Daniel Goldstein * , Chris Aldrich , Louisa O'Connor , Quanxi Shao

Posted Date: 30 January 2025

doi: 10.20944/preprints202501.2284.v1

Keywords: Measure-While-Drilling (MWD); Artificial Intelligence (AI); Machine Learning (ML); Geophysical; Wireline; Real-time Orebody Analysis; Feature Importance

 doi:10.20944/preprints202501.2284.v1

Article

Unlocking Subsurface Geology: A Case Study with Measure-While-Drilling Data and Machine Learning

Daniel Goldstein [1,*], Chris Aldrich [1], Quanxi Shao [2] and Louisa O'Connor [1]

[1] Western Australia School of Mines, Curtin University, Kalgoorlie, WA 6430, Australia

[2] CSIRO Data61, PO Box 1130, Bentley, WA, 6102, Australia

* Correspondence: daniel.goldstein@postgraduate.curtin.edu.au

Abstract: Bench-scale geological modeling is often uncertain due to limited exploration drilling and geophysical wireline measurements, reducing production efficiency. Measure-While-Drilling (MWD) systems collect drilling data to analyze mining blast hole drill rig performance. Early MWD studies focused on penetration rates to identify rock types. This paper investigates Artificial Intelligence (AI)-based regression models to predict geophysical signatures like density, gamma, magnetic susceptibility, resistivity, and hole diameter using MWD data. Machine Learning (ML) models evaluated include Linear Regression (LR), Decision Trees (DTs), Support Vector Machines (SVMs), Random Forests (RFs), Gaussian Processes (GP), and Neural Networks (NNs). An analytical method was validated for accuracy, and a three-tier experimental method assessed the importance of MWD features, revealing no performance loss when excluding features with less than 2% importance. RF, DTs, and GPs outperformed others, achieving R^2 values up to 0.98 with low RMSE, while LR and SVMs showed lower accuracy. NN performance improved with larger datasets. The study concludes DT, RF, and GP models excel in predicting geophysical signatures. Model selection depends on computational resources and application needs, offering valuable insights for real-time orebody analysis using AI. These findings could be invaluable to geologists who wish to utilize AI techniques for real-time orebody analysis and prediction.

Keywords: Measure-While-Drilling (MWD); Artificial Intelligence (AI); Machine Learning (ML); geophysical; wireline; real-time orebody analysis; feature importance

1. Introduction

The geological profiling of orebodies must be accurate and precise in order to define and achieve feasible grade and tonnage requirements of mining production. Traditional methods of accomplishing this are frequently expensive due to their reliance on resource-definition drill holes. For example, the traditional method for profiling an iron-ore deposit requires the use of an instrument on a wireline called a sonde to obtain geophysical response values in Reverse Circulation (RC) drill holes [1]. This method not only introduces inefficiencies due to a physical limitation of the sonde, but also raises concerns for field personnel due to potential exposure to radioactive sources from several sondes. Furthermore, the high costs of resource definition drilling leave gaps of approximately 50-100 meters between drill holes, resulting in an inaccurate depiction of the subsurface due to interpolation [2]. As a result, a more cost-effective approach that allows for comprehensive data collection is required to enable high-resolution delineation of subsurface geological conditions.

Measurement While Drilling (MWD) technology provides an effective solution to this geological modelling uncertainty. It was originally developed for the petroleum sector before being integrated into open pit mining blast hole drilling systems in the 1970s [3]. Continuous data gathering is enabled by installing a blast hole drill rig with MWD sensors, which provides insights into subsurface

 doi:10.20944/preprints202501.2284.v1

penetration performance [4]. In the context of operations involving repetitious drilling and blasting, such as open-pit mining, construction, and tunnelling, a wealth of MWD data points can be generated [5–7]. For example, a high-output blast rig in an open-pit iron ore mine can generate approximately 10,000 MWD data points per day, and high-volume mines generate even more [8].

Historically, to interpret the complex, nonlinear correlations between drilling responses and subsurface composition from such abundant MWD data, manual methods were used [5,8–13]. Previous MWD research focused predominantly on rock-type detection to improve blast fragmentation [5,8,10,12–14]. However, these findings do not adequately characterize smaller-scale geological conditions to optimize open pit orebody characterization. In contrast to previous manual interpretation methods, recently there have been attempts to apply Artificial Intelligence (AI) and Machine Learning (ML) approaches due to the improvements in computing power and availability [15–19]. Despite these advancements, only few studies have applied analytical methods to MWD data for geological boundary identification [20,21], but none have used a suitable method to evaluate the importance of each MWD metric for predicting geological features.

Principal Component Analysis (PCA) has been used to evaluate the feature importance of MWD metrics for orebody attributes [14,22]. However, its application is problematic due to its inability to determine feature importance. PCA is a method that allows one to reduce the dimensionality of data by identifying the principal components responsible for most of the data variance [23]. Unfortunately, the most variable characteristics are not always the most important, resulting in an incorrect application of PCA to determine feature importance from MWD data [24]. Therefore, this study employs appropriate feature-importance-based algorithms, Multivariate Adaptive Regression Splines (MARS) and Projection Pursuit Regression (PPR) on MWD data as well as ML techniques to determine the most significant features and methods for predictive modelling.

The present investigation focused on the geological characteristics of mineralized ore deposits at an open cut mine located in the Pilbara, Australia using MWD data. A method is presented to assess the feature importance of input drilling variables that will support feature selection for predictive geological modelling using MWD data. In addition, a comparative analysis of the predictive performance of various regression-based ML algorithms is included in this study. Through sophisticated analytics, such as assessment of feature importance and machine learning predictive modelling, it was possible to derive a more accurate representation of an orebody from MWD data. In comparison to resource development RC drill hole-based geological models, the findings represent an order of magnitude increase in spatial resolution previously unavailable without significant additional RC drilling.

2. Methods

2.1. Mine Site

The Western Australian region of Pilbara is renowned for being the main exporter of iron ore in Australia. In 2021, the state exported a remarkable 874 million tons [25]. The iron-ore deposits investigated in this study are in the Hammersley Group's Marra Mamba and Brockman Formations, which have been identified as important contributors to Pilbara's economically viable iron ore [26]. Approximately 2.5 billion years ago, extensive sequences of mineral-rich Banded Iron Formation (BIF) were interlayered with shale layers resulting in these formations [27]. For example, the Marra Mamba Formation comprises the Mount Newman Member, which is overlain by the West Angelas Member, which is dominated by shales. In contrast, the Brockman Formation is made up of the mineralized Dales Gorge BIF and shale bands.

The current investigation focuses on two pits that reflect the geological features of the Marra Mamba and Brockman Formations. 50-meter spaced resource development drillholes were used to delineate the geological characteristics of each pit's orebody. The Brockman Pit (*BR*) consisted of 211 RC drill holes totaling 16,880 meters and an average depth of 80 meters per hole. On the other hand, the Marra Mamba Pit (*MM*) included 167 RC drill holes totaling 13,957 meters and an average length

 doi:10.20944/preprints202501.2284.v1

of 83 meters per hole. To describe each pit's geology, wireline-based geophysical measurements of density (t/m^3), gamma (API), magnetic susceptibility (m^{3}kg^{-1}), resistivity (Ωm), and hole diameter (cm) were recorded at 0.01-meter intervals in the *BR* and *MM* resource-definition RC holes. No additional data engineering was undertaken on the resource-definition data, as the mining company's Quality Assurance and Quality Control (QA/QC) procedure had scrutinized these datasets.

2.2. Geological Qualities from Geophysical Measurements

This study considers various geophysical measurements, namely radioactive (gamma and density), electrical (resistivity and magnetic susceptibility) and physical (hole diameter), which are measured from their respective downhole sondes. The number of observations used after data processing is listed in Table 1.

Table 1. Number of observations used in each dataset after data processing.

Geophysical Measurement	Observations		
	BR	***MM***	***COM***
Density	45,813	5,789	51,602
Gamma	71,126	7,791	78,917
Magnetic Susceptibility	71,012	8,261	79,273
Resistivity	3,202	3,798	7,000
Caliper	61,666	7,505	69,171

Gamma and density wireline logging uses an active radioactive source to assess the bulk densities of subsurface materials as well as their react to the *gamma* radiation emanating from a regulated source housed within the logging instrument [28]. These responses are used for several purposes. For example, density (*dens*) is predominantly used as a proxy for ore grade. It can also be employed to estimate the tonnage of overburden stripping or as a measure of porosity. In contrast, the prevalent association of *gamma* radiation with clay minerals has led to using *gamma* as an indicator of shale or clay.

Resistivity and magnetic susceptibility are types of electrical logging that measure the electrical attributes of a rock formation. Resistivity (*res*) defines its capacity to resist the flow of electric current. Alterations in the rock's electrical properties can be attributed to factors, such as the content of clay minerals, water content and porosity, temperature variations, and conductivity of water [29]. Consequently, resistivity logs assist in interpreting conductive material properties and are predominantly employed to estimate salinity and demarcate lithology for hydrogeological studies. Magnetic susceptibility (*magsus*) quantifies the magnetization level of the stratigraphy in a drill hole when exposed to a magnetic field using electromagnetic induction [30]. Magsus data is useful for characterizing the degree of magnetization of subsurface material encountered in a drill hole exposure to differentiate and infer the mineralogy or lithology of a formation.

The caliper (*cal*) log, also referred to as the hole diameter log, is a physical measurement tool, in which one or more tensioned mechanical arms to measure the dimensions of the drill cavity [28]. Certain physical characteristics of the drill hole, for example, hole diameter, hole wall roughness and drilling mud thickness, influence other geophysical measurements. By interrogating the drill hole wall, *cal* can be used in conjunction with other geophysical measures to gain an improved understanding of subsurface geology.

2.3. MWD Drilling Systems

The research employed MWD for data collection, using a total of 22 rotary blast hole drill rigs that were outfitted with Tungsten Carbide Insert bits. The drilling fleet was comprised of ten Atlas Copco (Epiroc) Pit Viper 271 rigs, two Terex SKS 12 rigs, a single Bucyrus SKS 13 rig, and two Sandvik 460 rigs. These were deployed to drill production blast holes with a diameter of 0.229-meter (Figure

 doi:10.20944/preprints202501.2284.v1

1a). In addition, one Cubex QXR 920 rig, one Sandvik 560 rig, and five Atlas Copco (Epiroc) D65 drill rigs were deployed for drilling 0.165-meter wall control blast holes (Figure 1b). The bench heights in the studied iron-ore pits ranged from 8 to 12 meters, with sub-drilling extending roughly 2 meters below the bench floor. The spacing and burden between production blast holes averaged at 8 meters and 7 meters, respectively.

The MWD system on the drill rigs tracked metrics including rate of penetration (*rop*; m/s), rotary pressure or torque (*tor*; Nm), force on bit (*fob*; kgf)–also called weight on bit, thrust or pulldown pressure, bit air pressure or flushing air medium (*bap*; kgf/cm^2), and rotary speed (rpm). However, due to irregularities in the onboard sensor, the rpm data was only available for approximately a quarter of the sample points, leading to the exclusion of rpm from the drilling variables. The collection of MWD metrics was facilitated by a mix of manually operated rigs and semi-autonomous machines, with the latter being remotely overseen from an off-site Operations Centre. The drilling system logged the MWD time-series data at about 0.1-meter intervals along the blast hole depth.

Figure 1. Representative drilling rigs employed in the collection of MWD data [31]: a) Terex SKS 12, utilized for the drilling of 0.229-meter production blast holes and b) Epiroc D65, used in the creation of 0.165-meter wall control blast holes.

MWD data were collected from two distinct pits, *BR* and *MM*, each characterized by unique geological conditions. The *BR* pit provided a dataset encompassing 75,470 blast holes totaling 844,855 meters, while the *MM* pit contributed a dataset comprising 18,887 holes totaling 208,705 meters. A combined dataset (*COM*) was generated using *BR* and *MM* data. This study concentrated on MWD data ranging from 2 meters below the hole collars to the bottom of the blast holes, as the initial two meters of the borehole may not accurately represent the in situ geochemical properties of the rock due to possible toe charge effects during the previous bench's blasting.

2.3.1. MWD Data Pre-Processing

The efficacy of MWD data is affected by a variety of factors, such as subsurface composition, drill rig management system, and external circumstances, which can result in abnormal response values [32]. Consequently, these discrepancies can potentially lead to inaccurate MWD response values and erroneous interpretations of the data [33]. Accordingly, the noise-to-signal ratio in the analyzed mining MWD dataset is substantial, as the data had not been subjected to a thorough QA/QC process.

As a result, the MWD data in this study required feature engineering. Because collaring effects at the start of the shaft and potential blast damage from previous holes could skew the in situ rock representation, the initial MWD dataset omitted the first 2 meters of each drilling hole. Then, any data points with negative *rop*, *tor*, *fob*, or *bap* values were removed. Using linear interpolation, quartile detection methods, and a 1.5-factor threshold, the voids in the MWD data were subsequently filled.

 doi:10.20944/preprints202501.2284.v1

The data from the blast holes were smoothed with a Gaussian filter with a smoothing factor of 0.3 to reduce the local impacts of noise.

The MWD features obtained after performing feature engineering on the first four MWD responses are shown in Table 2. These variables contain the original MWD features, derived ratios of the original features (e.g., *rop* divided by *tor*, indicated as *roptor*) and a moving standard deviation across 0.5 meters for the original features (e.g., *ropS*).

Table 2. MWD features investigated in this study.

Type	**MWD Features**			
Recorded	*rop*	*tor*	*fob*	*bap*
Ratio	*roptor*	*torrop*	*fobrop*	*baprop*
	ropfob	*torfob*	*fobtor*	*baptor*
	ropbap	*torbap*	*fobbap*	*bapfob*
Standard Deviation	*ropS*	*torS*	*fobS*	*bapS*

The drilling datasets for blast-hole MWD and exploration hole were transformed from drill hole interval formats to point data, including geospatial coordinates and associated dataset values for each data point. The point data for exploration holes were generated utilizing downhole wireline logged desurvey data, which recorded the azimuth and dip of each hole every 10 meters until the final depth. On the other hand, the blast hole MWD data were not desurveyed, due to the production nature of the holes and the location of each point was determined by presuming a straight line from the hole's collar to its end. To merge these two datasets, a K-Nearest Neighbor distance-based search technique was used to calculate the distance between each point in the MWD and exploration data. Each exploration drilling data point was associated with the nearest MWD data point to conduct supervised machine learning. Horizontal and vertical distance thresholds were utilized to further refine the outcomes.

2.4. Feature-Importance-Based Methods

PCA has frequently been used to determine the most important MWD features [14,17,19,22,34]. In contrast, this study employs feature importance algorithms to establish the relative importance of each MWD variable identified for geophysical measurements such as *dens*, *gamma*, *magsus*, *res*, and *cal*. Non-parametric approaches, such as MARS and PPR, were applied to the pre-processed and merged *BR*, *MM*, and *COM* datasets. Both techniques do not make any assumptions on the relationships between the input and output variables. However, they evaluate feature significance differently.

MARS, a non-parametric approach to regression, disentangles complex variable interactions through a succession of piecewise linear regressions [35]. It identifies crucial features by fitting the model iteratively with each feature both included and omitted and measuring the performance variation. The MARS algorithm selects the MWD input that leads to the greatest improvement in the model as the most important as follows:

$$\hat{f}(x) = \sum_{j=1}^{J} a_j B_j(x) \quad (1)$$

where $\hat{f}(x)$ is a spline approximation of the function of interest $f(x)$ given by respective constant coefficients, a_j, and a linear combination of basis functions, B_j (x) for (j = 1, 2, …, J), which consist of a constant and a hinge function [36]. The *earth* package in *R*, which uses the MARS technique, was used with default hyperparameters to generate models that match the data distribution and to assess feature relevance of correlations between MWD variables [37].

On the other hand, PPR, a non-linear regression technique, reveals the most informative data projections into a lower-dimensional subspace [38]. In contrast to MARS, it identifies the most influential characteristics by analyzing the impact of each variable on these projections and

 doi:10.20944/preprints202501.2284.v1

determining which variables contribute the most to informative estimates. The PPR formula consists of:

$$\hat{f}(x) = \sum_{m=1}^{M} S_{\alpha m}\left(\sum_{i=1}^{n} \alpha_{im} x_i\right) \quad (2)$$

where $\alpha_{im} x_i$ denotes the inner product iteratively created in three steps: 1) initializing the residual to the response variable and the term counter M to zero, 2) using numerical optimization, S that maximize the figure of merit are determined, and 3) if the merit score falls below a particular threshold, the last term is eliminated. The *R* package *stats* [39], which incorporates PPR, was used with default hyperparameters to determine the goodness of fit for each variable.

MARS and PPR were utilized to quantify the feature importance of drilling metrics with the goal of understanding the complex, multi-variate relationships between MWD features and in situ geochemical signatures. These feature-importance-based methods were applied to MWD data for both short and big n-terms (basic functions of 101 and 201 for MARS, and terms of 5 and up to 50 for PPR, respectively). Furthermore, the purpose was to determine if complex models with larger n-terms would model links between geochemical assays and MWD data better than simpler models with smaller n-terms.

2.5. Regression-Based ML Methods

Neural Networks (NN) are the only regression-based ML algorithm that has been used to address subsurface geophysical intensity, with moderate success [19]. However, this study employed a variety of regression-based ML techniques, including Support Vector Machines (SVMs), Random Forests (RFs), Gaussian Process Regression (GPR), Linear Regression (LR), and Decision Trees (DTs), to investigate the effectiveness of these models to correlate geophysical properties with MWD data, as shown in Table 3.

Su et al. defined LR as a simple linear model that attempts to fit a line to a given dataset [40]. LR works well when the input and output variables are linearly related. However, LR algorithms may miss complex multivariate relationships. In contrast, DTs use recursive partitioning and key attributes to divide the data into smaller subgroups [41]. These trees can efficiently capture nonlinear interactions, but improper pruning may lead to overfitting. SVMs seek to determine the optimal hyperplane for classifying data [42]. They can effectively manage nonlinear relationships and high-dimensional data using kernel methods. Breiman defines RFs as a combination of several DTs to improve efficiency, prevent overfitting and manage nonlinear relationships [43]. GPR examines the output variable as a Gaussian distribution to identify the function that most closely approximates the data [44]. The GPR method considers nonlinear interactions and provides a probabilistic prediction of the outcome. Bishop describes NNs as flexible nonlinear models because they are modelled after the human brain, consisting of interconnected layers of neurons [45]. NNs reflect complicated relationships and work well with high-dimensional data but are susceptible to overfitting if not adequately regulated.

Table 3. The regression-based ML classes and subclasses utilized include Linear Regression (LR), Decision Trees (DTs), Support Vector Machines (SVMs), Random Forests (RFs), Gaussian Process Regression (GPR), and Neural Networks (NNs).

Class	Linear Regression (LR)	Decision Trees (DTs)	Support Vector Machines (SVMs)	Random Forests (RFs)	Gaussian Process Regression (GP)	Neural Networks (NNs)
Subclass	-Linear -Interactions -Robust -Stepwise	-Fine -Medium -Coarse	-Linear -Quadratic -Cubic -Fine Gaussian -Medium Gaussian -Coarse Gaussian	-Boosted -Bagged	- Squared Exponential - Matern 5/2 - Exponential - Rational Quadratic	- Narrow - Medium - Wide - Bilayered - Trilayered

 doi:10.20944/preprints202501.2284.v1

This study evaluated the predictive ability of the regression-based ML algorithms and performed the calculations on a Pawsey Supercomputer Nimbus cloud Ubuntu instance with 8 vCPUs and 32GB of Memory. The *RegressionLearner Toolbox* in *MATLAB*, was used with default hyperparameters and no optimization for each respective regression-based ML method to generate models and assess prediction performance [46]. The coefficient of determination (R^2) and root mean square error (RMSE) metrics were used to compare the performance of various models, defined by the following criteria:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \left(\frac{\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \right) \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (y_i - x_i)^2} \quad (4)$$

where *RSS* is the sum of squares of residuals, *TSS* is the total sum of squares, *N* is the sample size, y_i is the measured value, x_i is the predicted value, and $\bar{y}$ is the mean.

3. Results

A preliminary investigation of the MWD features (*rop*, *tor*, *fob*, and *bap*) collected from sensors on the drills was conducted to comprehend the data's range and frequency. Significant variations in *rop* and *fob* can be attributed many factors, including inconsistencies in mining equipment, operator competence, bit degradation, and rock mass properties [47]. Therefore, a single variable analysis of these features may not adequately capture the nonlinear correlations between MWD and geophysical measurements.

Figure 2a-d depict the first four MWD datapoints from the *COM* dataset, respectively. The *COM rop* displayed a balanced distribution, averaging 0.0248 m/s with a standard deviation of 0.010 m/s. Similarly, as shown in Figure 2b, the *COM tor* responses also have a typical distribution, with a mean of 3.41 Nm and a standard deviation of 1.06 Nm. On the other hand, Figure 2c depicts the skewed distribution of the *COM fob* with a mean of 97,945 kgf and a standard deviation of 78,524 kgf. Figure 2d displays a normal distribution for the *COM bap*, with values ranging from 230,300 kgf/cm² to 439,400 kgf/cm², a mean value of 335,550 kgf/cm², and a standard deviation of 49,384 kgf/cm².

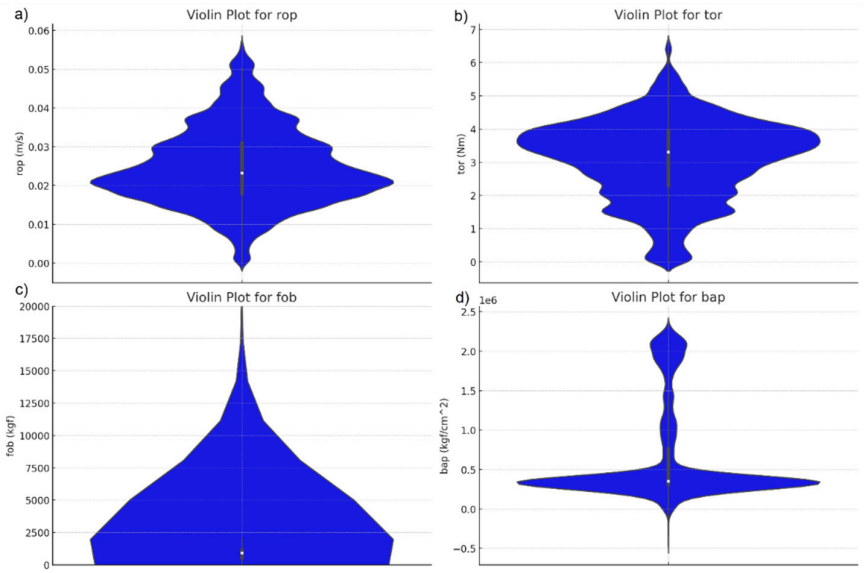

Figure 2. Violin plots showing frequencies and ranges of MWD data for a) *COM rop*, b) *COM tor*, c) *COM fob* d) *COM bap.*

 doi:10.20944/preprints202501.2284.v1

Upon examination of the *COM* geophysical data presented through violin plots in Figure 3, distinct patterns were observed. The *COM dens* measurements exhibit a uniform distribution, suggesting consistent rock densities across the studied region. In contrast, the *COM res* data is notably skewed towards lower values, indicating predominant low resistivity, with occasional higher resistivity regions. In addition, the *COM gamma* and *COM magsus* measurements present more variable distributions, signifying a diverse range of rock properties. Lastly, the *COM cal* data demonstrates symmetry, indicating consistent borehole sizes. While certain measurements like *dens* and *cal* indicate uniformity, *gamma* and *magsus* highlight variability in subsurface geophysical conditions.

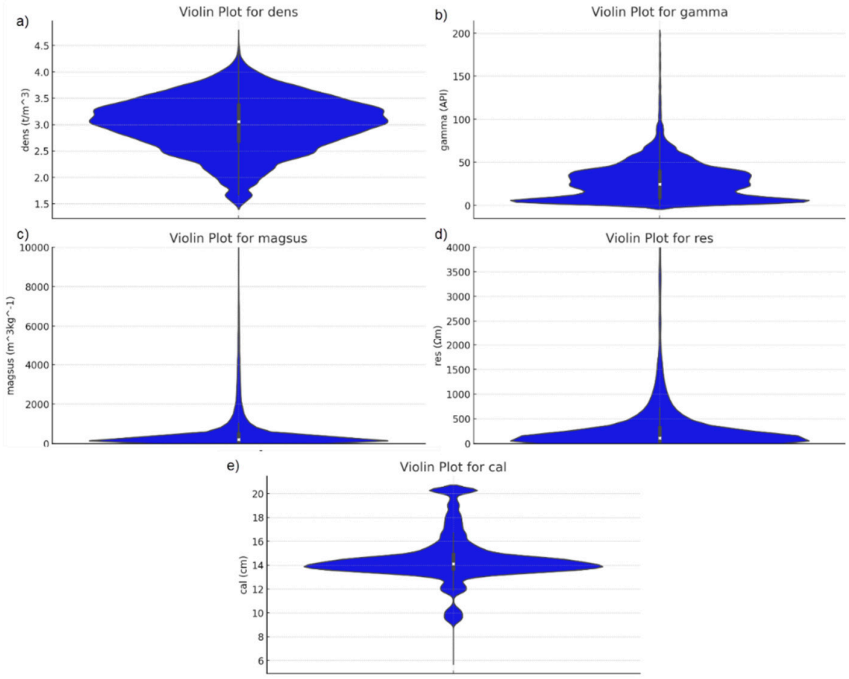

Figure 3. Violin plots showing frequencies and ranges of *COM* geophysical data for a) *dens*, b) *gamma*, c) *magsus* d) *res* and e) *cal.*

3.1. Feature-Importance-Based Results

The importance of the investigated MWD response features in inferring geophysical measure was investigated. Small and large n-terms MARS and PPR models were developed to determine if more complex ML methods would be advantageous during subsequent predictive modelling.

The percentages presented in Tables 4 and 5 were calculated by adding the relative weights of each attribute and dividing by the total for MARS and PPR, respectively. These percentages can be categorized into three different groups: a) 0% relative feature importance, where MWD features were deemed irrelevant by MARS and PPR in predicting orebody quality measures; b) greater than 0% but less than 5% relative feature importance (minor importance), indicating a slight influence on the prediction of orebody quality measures; and c) exceeding 5% relative feature importance.

 doi:10.20944/preprints202501.2284.v1

Table 4. MARS-derived feature importance of MWD measures in predicting *COM* geophysical values. Importance is expressed as a relative percentage of the cumulative value for each specific scenario. The scenarios considered include both small (101) and large (201) n-terms (basis functions).

MWD Feature	Density		Gamma		Magnetic Susceptibility		Resistivity		Caliper	
MARS n-terms	101 (%)	201 (%)	101 (%)	201 (%)	101 (%)	201 (%)	101 (%)	201 (%)	101 (%)	201 (%)
rop	7	7	10	10	6	6	3	4	11	11
tor	7	7	9	9	8	8	8	7	0	0
fob	1	1	0	0	4	4	1	1	0	0
bap	5	4	2	2	0	0	10	10	0	0
roptor	7	7	7	7	6	6	9	9	18	18
ropbap	6	6	7	7	0	0	8	10	0	0
ropfob	6	6	5	5	1	0	6	5	0	0
torrop	7	7	8	8	10	10	4	8	2	2
torbap	8	8	5	5	5	5	12	12	0	0
torfob	6	6	4	4	13	13	5	5	7	7
baprop	10	9	8	8	9	9	3	3	15	15
baptor	4	4	9	9	10	10	6	8	0	0
bapfob	2	1	0	0	9	9	0	0	0	0
fobrop	1	1	0	0	0	0	1	1	14	14
fobtor	6	5	5	5	4	4	6	5	7	7
fobbap	9	9	8	8	2	2	10	2	10	10
ropS	0	0	7	7	0	0	0	0	0	0
torS	3	6	6	6	6	6	6	9	5	5
fobS	3	3	0	0	2	2	0	0	12	12
bapS	3	2	0	0	5	5	0	0	0	0

Table 5. PPR-derived feature importance of MWD measures in predicting *COM* geophysical values. Importance is expressed as a relative percentage of the cumulative value for each specific scenario. The scenarios considered include both small (5) and large (<50) n-terms (basis functions).

MWD Feature	Density		Gamma		Magnetic Susceptibility		Resistivity		Caliper	
PPR n-terms	5 (%)	<50 (%)	5 (%)	<50 (%)	5 (%)	<50 (%)	5 (%)	<50 (%)	5 (%)	<50 (%)
rop	0	0	0	0	0	0	0	0	0	0
tor	0	0	0	0	0	0	0	0	0	0
fob	16	18	11	17	21	21	4	8	3	9
bap	7	4	7	8	5	8	1	18	3	9
roptor	0	0	0	0	0	0	0	0	0	0
ropbap	0	0	0	0	0	0	0	0	0	0
ropfob	0	0	0	0	0	0	0	0	0	0
torrop	19	2	25	9	18	5	9	3	21	12
torbap	0	0	0	0	0	0	0	0	0	0
torfob	0	0	0	0	0	0	0	0	0	0
baprop	28	5	29	6	13	12	11	11	49	6
baptor	7	4	8	4	8	4	7	4	7	19
bapfob	4	8	3	13	7	10	19	16	3	24
fobrop	11	12	4	33	14	28	11	20	6	14
fobtor	5	37	8	6	9	9	34	1	3	5
fobbap	0	0	0	0	0	0	0	0	0	0
ropS	0	0	0	0	0	0	0	0	0	0
torS	0	0	0	0	0	0	0	0	0	0
fobS	3	9	2	3	5	1	2	17	1	1
bapS	0	1	4	2	1	1	2	2	3	1

The findings suggest that the relative importance of features remains stable across models with both small and large term counts. Despite the apparent consistency in both n-term models, MARS

 doi:10.20944/preprints202501.2284.v1

and PPR analyses assigned varying degrees of importance to different features. For instance, when using the *COM* dataset, the MARS method identified 15 out of 20 of the MWD measures as crucial in inferring *dens* (Figure 4a), with the exceptions being *fob*, *bapfob*, *fobrop*, *fobS* and *bapS*. The identification of most features as important due to the MARS method searching for relationships between variables. Conversely, the PPR approach determined that only half of the MWD features were important with only five variables greater than or around ten percent: *fobrop*, *fob*, *bapfob*, *torrop* and *bap*. (Figure 4b). PPR implies that the remaining MWD features exert minimal to no influence on the prediction of orebody quality, possibly due to the lack of consideration for nonlinear interaction among features in the PPR model. Thus, to encompass all potential significant MWD features, MARS and PPR methodologies were employed.

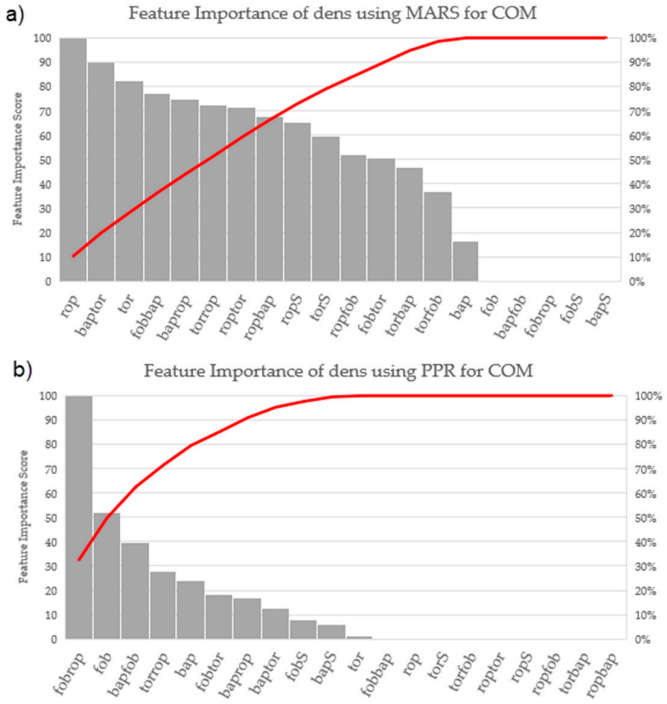

Figure 4. Feature importance scores of MWD features for *COM dens* values employing a) MARS and b) PPR. The total percentage of feature importance is graphically represented as the red lines.

Figure 5 provides a comparative analysis of the significance of the MWD features for *gamma*, *dens*, *magsus*, *res* and *cal COM* dataset, as determined by the MARS and PPR methodologies. These results correspond to the top ten most important MWD features discovered in the *dens* analysis (Figure 4). The MWD characteristic *fobrop* routinely emerges as highly important in this study's datasets along with *bapfob*, *baprop*, *fob*, *bap* and *fobtor*. It is important to note, however, that this does not diminish the importance of other MWD features; rather, it highlights those that are frequently identified as important.

 doi:10.20944/preprints202501.2284.v1

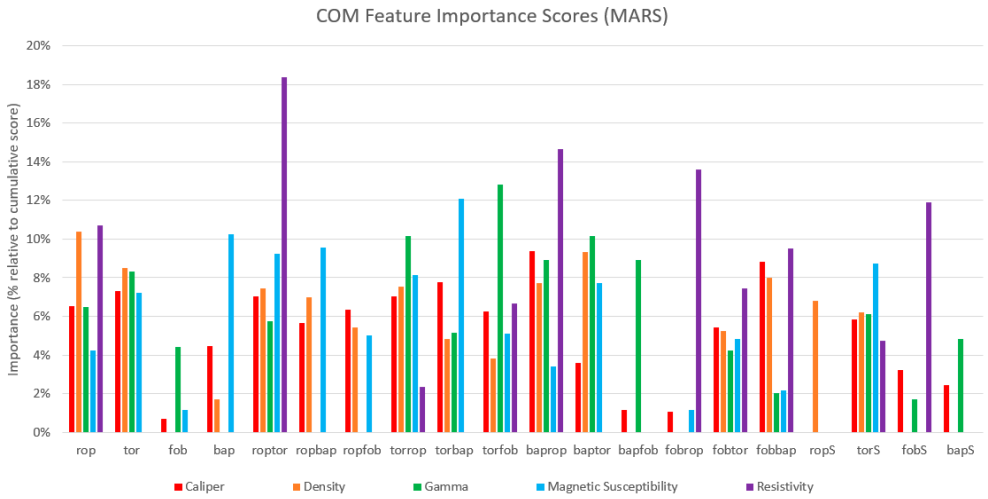

Figure 5. Feature importance-based ML results of MWD variables for the *COM* dataset in predicting geophysical measurements: *cal*, *dens*, *gamma*, *magsus*, and *res*, as determined by the MARS approach. The *y*-axis represents the importance score as a percentage of the overall value for MARS. The MWD features are graphed along the *x*-axis.

As shown in Tables 4 and 5, the importance of features for evaluating *dens* varies across different datasets when the MARS and PPR methodologies are applied. The six MWD variables deemed most important for predicting den from the analyzed BIF deposits are *fobrop*, *fobtor*, *bapfob*, *torrop*, *baprop* and *baptor*, in accordance with the peak importance findings (Figure 5). This result generally corresponds to the importance of MWD features identified for *gamma*, *magsus*, *res*, and *cal* (Figure 4). The ratios *fobrop*, *fobtor*, *bapfob*, *torrop*, *baprop* and *baptor* carried more importance than the primary *rop* and *tor* variables, the *rop*-influenced ratios *roptor*, *ropbap*, and *ropfob*, as well as the variability-related metrics *ropS*, *torS*, and *bapS*. On the other hand, *fob* was identified as highly important from PPR but was very low or missing in the MARS rankings.

3.2. Regression-Based ML Analytical Prediction Results

The following sections evaluate the prediction strength of several kinds of regression-based ML regression models for predicting geophysical measurements of an orebody, such as *dens*, *gamma*, *magsus*, *res*, and *cal*, based on the MWD features described in Table 2. This analysis required validating the proposed ML analytical procedure with a variety of geophysical signatures to establish theoretical precision.

A 10-fold cross-validation technique revealed the training dataset's prediction strength. Datasets were split into 80% Training and 20% Testing. Testing results reported in RMSE and R^2 being the average of the 10 folds during cross validation. In addition, a threshold of twenty-four hours of computation time was established due to practical limitations regarding calculation speed for real-time analysis. Consequently, several GP-based analyses were prematurely terminated and categorized as nonapplicable (N/A).

A preliminary investigation was conducted using the Coarse DT method to determine if additional MWD features beyond *rop, tor, fob* and *bap* would strengthen the prediction performance of regression-based ML models. The results are listed in Table 6. Based on these results of 14 out of 15 models performing better with the additional features, the decision was made to incorporate the ratio and moving standard deviation (Table 2) MWD features in all the investigated predictive ML models. The *res* models did not improve as much as others possibly due to smaller numbers of observations due to this geophysical sonde not used in every drillhole.

 doi:10.20944/preprints202501.2284.v1

Table 6. Prediction performance of DT models using only the 4 measured MWD features and all 20 investigated MWD features. Higher performing models are in **bold.**

Geophysical Measurement	*BR*				*MM*				*COM*			
	Measured		Additional		Measured		Additional		Measured		Additional	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
dens	0.33	0.57	**0.29**	**0.68**	0.37	0.51	**0.34**	**0.59**	0.34	0.56	**0.30**	**0.66**
gamma	15.47	0.50	**14.01**	**0.59**	13.62	0.65	**10.67**	**0.78**	15.53	0.51	**13.67**	**0.62**
magsus	848	0.70	**701**	**0.80**	389	0.24	**376**	**0.29**	845	0.68	**680**	**0.79**
res	582	0.27	**554**	**0.34**	**697**	**0.31**	705	0.29	650	0.29	**631**	**0.33**
cal	1.12	0.68	**0.93**	**0.78**	1.19	0.61	**1.03**	**0.70**	1.16	0.66	**0.94**	**0.78**

3.2.1. Density and Gamma Prediction

The regression results for estimating *dens* and *gamma* values using ML models such as LR, DTs, SVMs, RFs, GP, and NNs are detailed in Tables 7 and 8, respectively. The R^2 values of models utilising *BR* data to predict *dens* and *gamma* were marginally superior to those utilising *MM* and *COM* datasets. This discrepancy may be attributable to the *BR* and *MM* datasets containing different quantities of data.

Table 7. R^2 and RMSE results of regression-based ML models to predict *dens* values from MWD data. The highest performing model results are **bold**. All 20 MWD features were incorporated into the models. Standard deviations (std) from 10-fold cross validation are reported for RMSE and R^2.

Regression-based ML Class	Regression-based ML Suclass	*BR*		*MM*		*COM*	
		RMSE (t/m^3)	R^2	RMSE (t/m^3)	R^2	RMSE (t/m^3)	R^2
LR	Linear	0.49	0.06	0.49	0.13	0.50	0.05
	Interactions	0.55	0.00	0.45	0.28	0.79	0.00
	Robust	0.49	0.06	0.50	0.13	0.50	0.04
	Stepwise	**0.48**	**0.12**	**0.44**	**0.32**	**0.49**	**0.09**
DTs	**Fine**	**0.22**	**0.81**	**0.27**	**0.74**	**0.23**	**0.80**
	Medium	0.25	0.76	0.29	0.69	0.25	0.75
	Coarse	0.29	0.68	0.34	0.59	0.30	0.66
SVMs	Linear	0.49	0.05	0.50	0.12	0.50	0.04
	Quadratic	0.46	0.17	0.41	0.40	0.48	0.13
	Cubic	0.37	0.46	0.31	0.66	0.56	0.00
	Fine Gaussian	**0.31**	**0.63**	**0.25**	**0.77**	**0.32**	**0.61**
	Medium Gaussian	0.40	0.39	0.37	0.50	0.43	0.28
	Coarse Gaussian	0.48	0.09	0.48	0.18	0.49	0.07
RFs	Boosted	0.46	0.19	0.41	0.41	0.47	0.16
	Bagged	**0.21**	**0.83**	**0.24**	**0.80**	**0.21**	**0.82**
GPs	Squared Exponential	0.28	0.70	0.23	0.81	0.27	0.72
	Matern 5/2	0.27	0.72	0.22	0.82	0.26	0.73
	Exponential	0.22	0.82	**0.19**	**0.87**	0.22	0.81
	Rational Quadratic	**0.20**	**0.84**	0.20	0.86	**0.22**	**0.82**
NNs	Narrow	0.44	0.24	0.35	0.56	0.45	0.21
	Medium	0.38	0.42	0.29	0.69	0.41	0.36
	Wide	**0.32**	**0.61**	**0.24**	**0.79**	**0.34**	**0.55**
	Bilayered	0.41	0.36	0.33	0.62	0.43	0.29
	Trilayered	0.40	0.38	0.31	0.65	0.42	0.33

 doi:10.20944/preprints202501.2284.v1

Table 8. R^2 and RMSE results of regression-based ML models used to predict *gamma* values from MWD data. The highest performing model results are **bold**. 20 All MWD features were used in these models. Standard deviations (std) from 10-fold cross validation are reported for RMSE and R^2.

Regression-based ML Class	Regression-based ML Suclass	*BR*		*MM*		*COM*	
		RMSE (API)	R^2	RMSE (API)	R^2	RMSE (API)	R^2
LR	Linear	21.30	0.06	20.52	0.20	**21.53**	**0.06**
	Interactions	21.53	0.04	17.81	0.40	22.38	0.00
	Robust	21.47	0.04	22.83	0.01	21.71	0.04
	Stepwise	**20.75**	**0.11**	**16.23**	**0.50**	N/A	N/A
DTs	**Fine**	**10.24**	**0.78**	**0.80**	**0.88**	**10.00**	**0.80**
	Medium	11.90	0.71	8.79	0.85	11.58	0.73
	Coarse	14.01	0.59	10.67	0.78	13.67	0.62
SVMs	Linear	21.49	0.04	22.02	0.08	21.71	0.04
	Quadratic	20.39	0.14	15.45	0.55	20.70	0.13
	Cubic	18.33	0.30	10.05	0.81	27.88	0.00
	Fine Gaussian	**16.13**	**0.46**	**7.56**	**0.89**	**16.02**	**0.48**
	Medium Gaussian	18.63	0.28	12.76	0.69	19.19	0.25
	Coarse Gaussian	20.96	0.09	21.03	0.16	21.24	0.08
RFs	Boosted	19.92	0.18	13.86	0.63	20.03	0.18
	Bagged	**9.65**	**0.81**	**7.04**	**0.91**	**9.43**	**0.82**
GPs	Squared Exponential	14.31	0.58	6.73	0.91	13.30	0.64
	Matern 5/2	13.72	0.61	6.52	0.92	**13.24**	**0.64**
	Exponential	**10.93**	**0.75**	**6.32**	**0.92**	N/A	N/A
	Rational Quadratic	N/A	N/A	6.51	0.92	N/A	N/A
NNs	Narrow	19.70	0.20	11.61	0.74	19.98	0.19
	Medium	18.17	0.32	8.87	0.85	18.61	0.30
	Wide	**16.03**	**0.47**	**6.88**	**0.91**	**16.22**	**0.47**
	Bilayered	18.92	0.26	9.34	0.83	19.19	0.25
	Trilayered	18.84	0.26	9.41	0.83	19.28	0.24

Among all *COM* models, those employing LR and SVMs consistently produced the least accurate predictions for *dens* and *gamma* with R^2 values below 0.50 (Tables 7 and 8). In contrast, models constructed with DTs, RFs, and GP yielded the highest R^2 values for *dens* and *gamma* predictions, with both achieving 0.80. These high-performing DTs, RFs, and GP models yielded an average RMSE of approximately 0.21 t/m^3 for *dens* and 9.42 API for *gamma*. GP displayed the most accurate predictions, with R^2 values of 0.87 and 0.92 and RMSEs of less than 0.19 t/m^3 and 6.32 API for *MM dens* and *gamma*, respectively.

Within each ML algorithm class, significant differences were observed between the subclasses of ML outlined in Table 3. As an example, the Bagged (Bootstrapped Aggregate) Tree method outperformed the Boosted Tree RFs in predicting densities and *gamma*, with R^2 values of approximately 0.82 for both geophysical measurements. Similarly, Wide NNs consistently outperformed other NN types, with peak R^2 values of 0.54 and 0.43 for *dens* and *gamma*, respectively, and the lowest RMSE values of 0.35 t/m^3 for *dens* and 16.68 API for *gamma*. Lastly, Fine Tree DT correlation values of 0.80 for both *dens* and *gamma* were superior to those of Medium and Fine parameters.

Figure 6a-f depicts the ML analytical prediction results compared to actual wireline measured *dens* values for the best-performing LR, DTs, SVMs, RFs, GP, and NNs models. The Bagged RFs models had the strongest correlation with R^2 values of 0.82. The DT models generated slightly weaker R^2 values of 0.80. However, the training speed of the DT models was over 10 times that of the Bagged RF models, around 400 and 35 observations per second, respectively.

 doi:10.20944/preprints202501.2284.v1

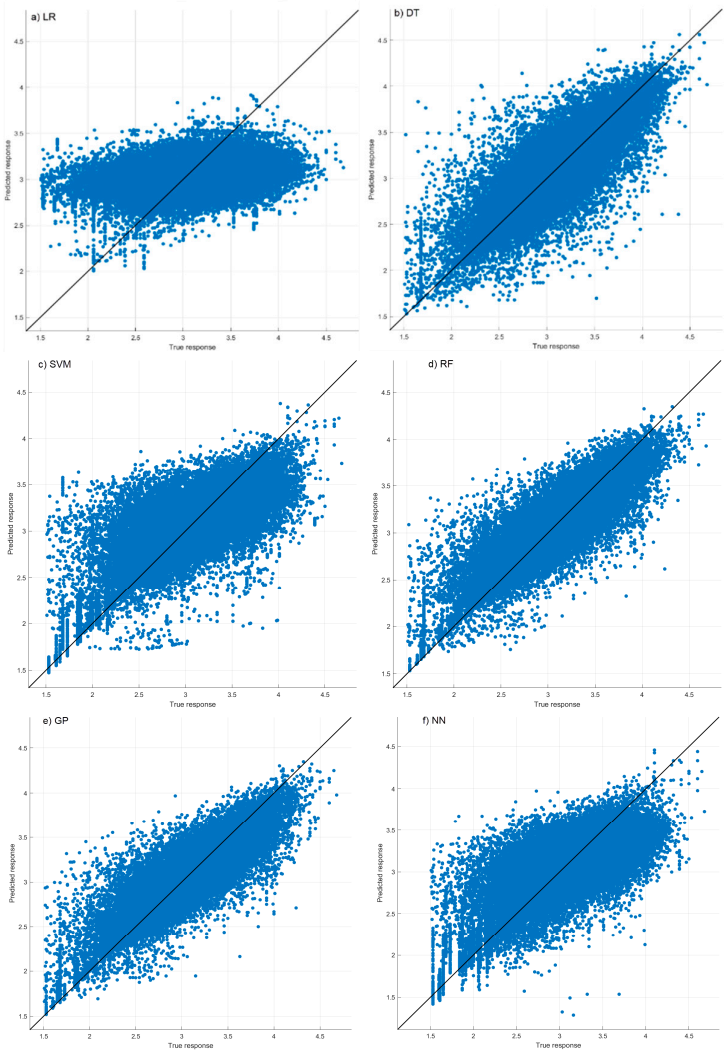

Figure 6. Actual versus predicted R^2 values for *dens* predictions using various machine learning analytical methods, including a) LR, b) DTs, c) SVMs, d) RFs, e) GP and f) NNs.

In addition, a series of three-level experiments with variable input parameters were conducted to correspond to the three primary levels of relative feature significance determined by MARS and PPR in section 3.1, namely 0%, less than 5%, and over 5%. The objective of this method was to determine the effect of omitting MWD features deemed to be of minimal importance. The experimental design included 1) the inclusion of all 20 MWD features, including those with 0% relative importance, 2) the exclusion of MWD features identified as having 0% relative importance, and 3) the removal of MWD features classified as having less than 5% importance, which was designated as minor importance.

Interestingly, the elimination of minor importance features had no effect on prediction performance when compared to the use of all features (Figure 7). With the *gamma COM* dataset, most instances exhibited less than a 0.05 decrease in R^2, whereas the Fine Tree and Medium Tree DT techniques demonstrated a 0.19 and 0.06 enhancement in R^2, respectively. Nonetheless, there was a consistent increase in training speed for *dens* and *gamma* DTs, as well as *dens* NNs, when marginally significant features are used over all features (Figure 7). In contrast, when utilizing all features, *gamma* NNs models demonstrated faster training rates and higher R^2 values. This anomaly may be due to the inherent dynamics of the NN method, which has difficulty establishing relationships between these datasets using marginally significant features, resulting in lower R^2 values.

 doi:10.20944/preprints202501.2284.v1

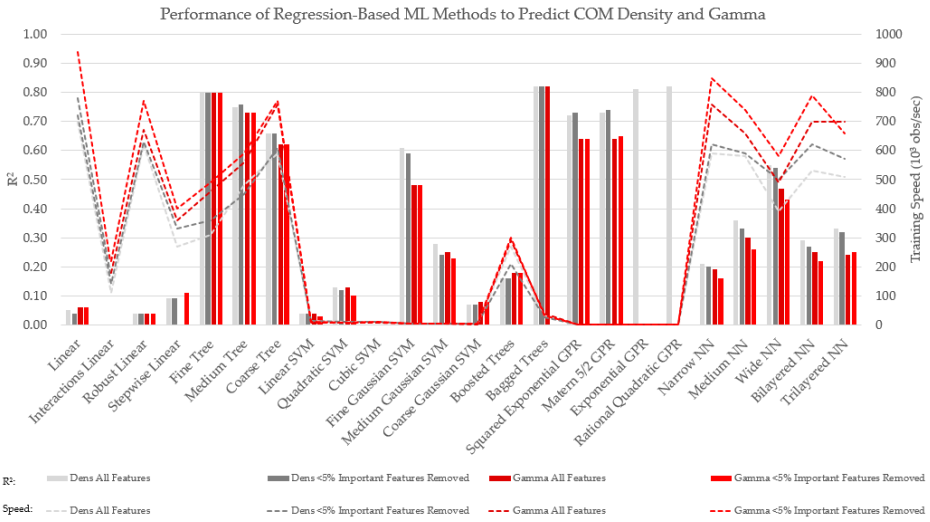

Figure 7. R^2 (columns) and training speeds (lines) of the investigated regression-based ML methods used to predict *dens* and *gamma* orebody geophysical values using MWD features in the *COM*.

3.2.2. Magsus and Res Prediction

The prediction results of *magsus* and *res* geophysical measurements using various ML models, including LR, DTs, SVMs, RFs, GP, and NNs are presented in Tables 9 and 10. Interestingly, DT models showed consistent performance across the *BR*, *MM*, and *COM* datasets, with R2 variances of less than 0.02 between them. In contrast, NN models demonstrated a notable performance improvement on the *MM* dataset. This divergence aligns with observations made for *dens* and *gamma* (see Section 3.1.1) and could be attributed to the disparity in data volume between the *BR* and *MM* datasets, a point that requires more investigation.

Table 9. R^2 and RMSE results of regression-based ML models used to predict *magsus* values from MWD data. The highest performing model results are **bold**. All 20 MWD features were used in these models. Standard deviations (std) from 10-fold cross validation are reported for RMSE and R^2.

Regression-based ML Class	**Regression-based ML Suclass**	***BR***		***MM***		***COM***	
		RMSE (m^3kg^{-1})	**R^2**	**RMSE (m^3kg^{-1})**	**R^2**	**RMSE (m^3kg^{-1})**	**R^2**
LR	Linear	1505	0.06	435	0.05	1446	0.05
	Interactions	1490	0.08	**388**	**0.24**	1464	0.03
	Robust	1625	0.00	453	0.00	1552	0.00
	Stepwise	**1450**	**0.13**	N/A	N/A	**1387**	**0.13**
DTs	**Fine**	**518**	**0.89**	**260**	**0.66**	**499**	**0.89**
	Medium	564	0.87	291	0.57	542	0.87
	Coarse	701	0.80	375	0.29	680	0.79
SVMs	Linear	1606	0.00	450	0.00	1537	0.00
	Quadratic	1498	0.07	443	0.01	1468	0.03
	Cubic	1306	0.30	389	0.24	1403	0.11
	Fine Gaussian	**1081**	**0.52**	**375**	**0.29**	**1105**	**0.45**
	Medium Gaussian	1366	0.23	438	0.03	1403	0.11
	Coarse Gaussian	1576	0.00	450	0.00	1517	0.00
RFs	Boosted	1171	0.43	298	0.55	1134	0.42
	Bagged	**486**	**0.90**	**253**	**0.68**	**457**	**0.91**
GPs	Squared Exponential	668	0.82	**252**	**0.68**	646	0.81
	Matern 5/2	667	0.82	254	0.68	590	0.84
	Exponential	504	0.90	N/A	N/A	**471**	**0.90**
	Rational Quadratic	**483**	**0.90**	N/A	N/A	484	0.89
	Narrow	1093	0.51	307	0.53	1114	0.44

 doi:10.20944/preprints202501.2284.v1

	Medium	1019	0.57	287	0.59	1030	0.52
NNs	**Wide**	**865**	**0.69**	**262**	**0.66**	**902**	**0.63**
	Bilayered	997	0.59	295	0.56	1012	0.54
	Trilayered	971	0.61	276	0.62	986	0.56

Table 10. R^2 and RMSE results of regression-based ML models used to predict *res* values from MWD data. The highest performing model results are **bold**. All 20 MWD features were used in these models. Standard deviations (std) from 10-fold cross validation are reported for RMSE and R^2.

Regression-based ML Class	Regression-based ML Suclass	*BR*		*MM*		*COM*	
		RMSE (Ωm)	R^2	RMSE (Ωm)	R^2	RMSE (Ωm)	R^2
LR	Linear	21.30	0.06	20.52	0.20	**21.53**	**0.06**
	Interactions	21.53	0.04	17.81	0.40	22.38	0.00
	Robust	21.47	0.04	22.83	0.01	21.71	0.04
	Stepwise	**20.75**	**0.11**	**16.23**	**0.50**	N/A	N/A
DTs	**Fine**	**10.24**	**0.78**	**0.80**	**0.88**	**10.00**	**0.80**
	Medium	11.90	0.71	8.79	0.85	11.58	0.73
	Coarse	14.01	0.59	10.67	0.78	13.67	0.62
SVMs	Linear	21.49	0.04	22.02	0.08	21.71	0.04
	Quadratic	20.39	0.14	15.45	0.55	20.70	0.13
	Cubic	18.33	0.30	10.05	0.81	27.88	0.00
	Fine Gaussian	**16.13**	**0.46**	**7.56**	**0.89**	**16.02**	**0.48**
	Medium Gaussian	18.63	0.28	12.76	0.69	19.19	0.25
	Coarse Gaussian	20.96	0.09	21.03	0.16	21.24	0.08
RFs	Boosted	19.92	0.18	13.86	0.63	20.03	0.18
	Bagged	**9.65**	**0.81**	**7.04**	**0.91**	**9.43**	**0.82**
GPs	Squared Exponential	14.31	0.58	6.73	0.91	13.30	0.64
	Matern 5/2	13.72	0.61	6.52	0.92	**13.24**	**0.64**
	Exponential	**10.93**	**0.75**	**6.32**	**0.92**	N/A	N/A
	Rational Quadratic	N/A	N/A	6.51	0.92	N/A	N/A
NNs	Narrow	19.70	0.20	11.61	0.74	19.98	0.19
	Medium	18.17	0.32	8.87	0.85	18.61	0.30
	Wide	**16.03**	**0.47**	**6.88**	**0.91**	**16.22**	**0.47**
	Bilayered	18.92	0.26	9.34	0.83	19.19	0.25
	Trilayered	18.84	0.26	9.41	0.83	19.28	0.24

DT approaches produced R^2 values ranging from 0.79 to 0.89 when applied to *COM magsus* data. However, the *res COM* DT models were remarkably lower than the *magsus* data, ranging from 0.32 to 0.58. This discrepancy could be due to the *res* data set having around 10 times less data due to inconsistent wireline logging practices, in which *res* was not measured on the same number of holes as *magsus*.

Certain ML subclasses outperformed others within the same ML class, a pattern consistent with those discovered for *dens* and *gamma*. For example, Fine DTs consistently outperformed Medium and Coarse DTs in both the *COM magsus* and *res* datasets. Moreover, in the predictions for *magsus* and *res*, Bagged RFs performed better than Boosted Tree RFs. Likewise, Wide NNs consistently outperformed other NN subclasses, with R^2 values for *magsus* and *res* reaching a maximum of 0.63 and 0.49, respectively. Fine Gaussian SVMs achieved R^2 values greater than 0.44 for *magsus* and Cubic SVMs of 0.22 for *res*, demonstrating a wide range of outcomes. In contrast, the R^2 values of Linear and Coarse Gaussian SVMs were all 0.00.

With R^2 values exceeding 0.60 for electrical geophysical predictions, GPs consistently generated reliable results across all GP subclasses. Nonetheless, as shown in Figure 8, GP models required a prolonged computation time than methods such as DTs, RFs, and NNs. However, no significant decrease in prediction performance was observed when models excluded features with less than 5% importance in comparison to models that included all features. Certain models, including most

 doi:10.20944/preprints202501.2284.v1

created with Stepwise LR, Exponential GPR, and Rational Quadratic GPR, had to be stopped after 24 hours; therefore, their results are not included here.

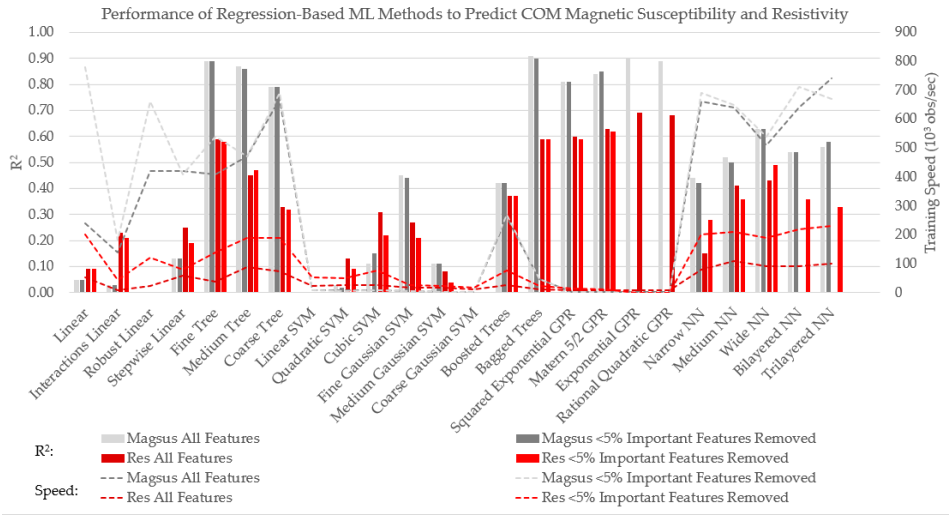

Figure 8. Actual versus predicted R^2 values for *dens* predictions using regression-based ML methods, including a) LR, b) DTs, c) SVMs, d) RFs, e) GP and f) NNs.

3.1.3. Caliper Predictions

Table 11 displays the results of *cal* prediction using LR, DTs, SVMs, RFs, GP, and NNs models. The *BR*, *MM* and *COM* models all had a relatively consistent R^2 and RMSE. These results contrast with the differences between the *BR* and *MM* observed in the *dens*, *gamma*, *magsus* and *res* results, indicating the need for additional investigation to understand the differences between *BR* and *MM* results.

Examining the predictive performance of these models reveals that all variants of LR models, including Linear, Interactions, Robust, and Stepwise, consistently underperformed with R^2 values lower than 0.45. In contrast, models built with SVMs, RFs, GP, and NNs provided more accurate predictions, with maximum R^2 values of 0.86, 0.89, 0.92 and 0.87, respectively, and an average RMSE of approximately 0.63 cm.

Table 11. R^2 and RMSE results of regression-based ML models used to predict *cal* values from MWD data. The highest performing model results are **bold**. All 20 MWD features were used in these models. Standard deviations (std) from 10-fold cross validation are reported for RMSE and R^2.

Regression-based ML Class	Regression-based ML Suclass	*BR*		*MM*		*COM*	
		RMSE (cm)	R^2	RMSE (cm)	R^2	RMSE (cm)	R^2
LR	Linear	1.92	0.07	1.70	0.20	1.92	0.06
	Interactions	2.56	0.00	1.57	0.31	3.81	0.00
	Robust	2.00	0.00	1.75	0.15	1.99	0.00
	Stepwise	**1.85**	**0.14**	**1.40**	**0.45**	**1.84**	**0.14**
DTs	**Fine**	**0.76**	**0.85**	**0.71**	**0.86**	**0.76**	**0.85**
	Medium	0.79	0.84	0.81	0.82	0.80	0.84
	Coarse	0.93	0.78	1.03	0.70	0.94	0.78
SVMs	Linear	1.98	0.01	1.76	0.14	1.97	0.01
	Quadratic	1.82	0.17	1.34	0.50	1.85	0.13
	Cubic	1.47	0.46	0.97	0.74	1.91	0.07
	Fine Gaussian	**1.08**	**0.71**	**0.70**	**0.86**	**1.12**	**0.68**
	Medium Gaussian	1.58	0.37	1.19	0.61	1.70	0.26
	Coarse Gaussian	1.94	0.05	1.67	0.22	1.94	0.05
RFs	Boosted	1.77	0.21	1.34	0.50	1.79	0.19

 doi:10.20944/preprints202501.2284.v1

	Bagged	**0.71**	**0.87**	**0.62**	**0.89**	**0.70**	**0.87**
GPs	Squared Exponential	0.85	0.82	0.63	0.89	0.83	0.82
	Matern 5/2	0.83	0.83	0.61	0.90	0.81	0.83
	Exponential	**0.71**	**0.87**	**0.53**	**0.92**	**0.70**	**0.88**
	Rational Quadratic	0.75	0.86	0.58	0.91	0.75	0.86
NNs	Narrow	1.67	0.30	1.15	0.63	1.74	0.23
	Medium	1.47	0.45	0.91	0.77	1.57	0.37
	Wide	**1.09**	**0.70**	**0.68**	**0.87**	**1.21**	**0.63**
	Bilayered	1.57	0.37	1.03	0.70	1.64	0.32
	Trilayered	1.50	0.43	1.02	0.71	1.58	0.37

Among these models, Bagged RFs and Wide NNs delivered the best predictive results within their respective ML classes, with RMSEs of 0.62 cm and 0.68 cm, respectively. On the *COM* dataset, DTs produced R^2 values of 0.76, 0.80, and 0.94 for the Coarse, Medium, and Fine parameters, respectively. SVMs displayed the most variable results based on the chosen method, with Fine Gaussian achieving an R^2 of up to 0.86 and Linear, Cubic, and Coarse Gaussian yielding R^2 values below 0.50 and 1.34 cm RMSE.

Experiments that excluded features with less than 5% (minor) importance had no appreciable impact on the *cal* prediction accuracy. Compared to the trials that included all MWD features, as shown in Figure 9, this feature exclusion sped up model training times. DTs and NNs emerged as the quickest training methods with around 600,000 observations per second, whereas GP computations around 300x more time-consuming at around 2000 observations per second. The models utilizing Exponential GPR (for *BR* <5% and *MM* <5%) and Rational Quadratic GPR (for *BR* <5%, *MM* <5%, *COM* all, and *COM* <5%) were discontinued after 24 hours, and their results are therefore not presented.

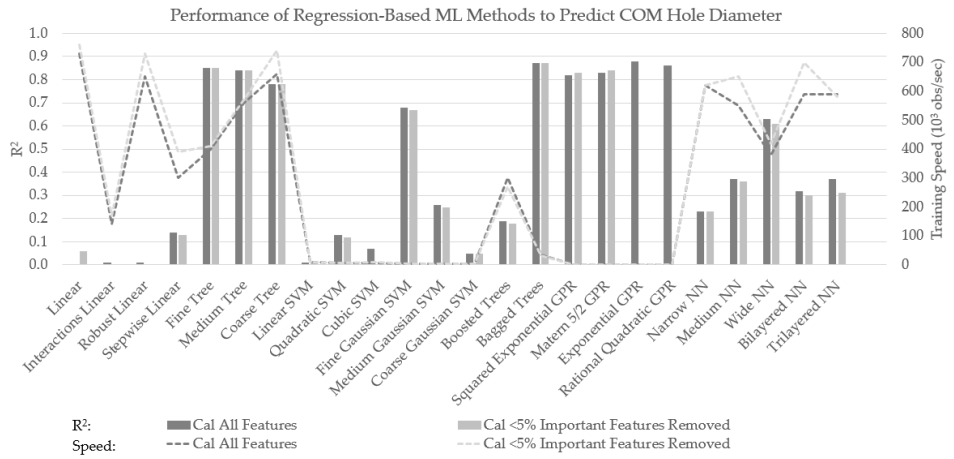

Figure 9. R^2 (columns) and training speeds (represented by lines) demonstrating performance of regression-based ML methods using *COM* MWD data to predict *cal*.

4. Discussion

This study demonstrates the effectiveness of feature-importance-based methods and regression-based ML techniques in estimating subterranean geophysical signatures from MWD data, thereby increasing orebody knowledge. Tables 4 and 5 reveal that the success of predictive modelling of the five investigated geophysical properties depends primarily on two factors: the characteristics of the on-site host rock as represented by MWD data and its distribution across various mining locations, as well as the quantity of data. Though the scope of this study was confined to five geophysical properties, it has the potential to be extended to other measurements such as acoustic, neutron porosity, dip-meter, spontaneous potential, or nuclear magnetic resonance.

 doi:10.20944/preprints202501.2284.v1

This study departs from prior research by showing the importance of MWD ratio features, such as *fobrop*, *fobtor*, *bapfob*, *torrop*, *baprop*, and *baptor*, in addition to *fob*. Earlier research emphasized *rop* and *tor*, utilizing PCA to determine the most important MWD measurements for rock type identification [9,12,14,17,19,34]. In contrast with the PCA-based feature selection in MWD data, feature importance-based ML methodologies such as MARS and PPR, revealed previously unobserved complex relationships between MWD features and rock mass characteristics such as those derived from *bap*.

The differences in feature importance evaluations between MARS and PPR result from the underlying mechanics of these algorithms. The MARS technique evaluates the correlation between MWD features and geophysical signatures, therefore expanding the range of relevant MWD features (Table 4). On the other hand, PPR evaluates the influence of each feature on data projections and determined that half of the drilling features were not important (Table 5).

Moreover, when compared to their smaller equivalents, larger n-term MARS and PPR models did not offer a more robust depiction of correlations between MWD data and geophysical measurements [48]. The consistent performance of DTs and Bagged RFs, with R^2 prediction values exceeding 0.80 across most ML models as shown in Tables 7–11 suggests that complex ML models may not always provide superior predictive capabilities. Furthermore, more complex models like SVMs struggle with non-scaled features and imbalanced datasets where one class of features dominates. In this case, drill rig type and hole diameter may be two factors, as the larger rigs drilled more wider-diameter production holes than smaller rigs drilling narrower holes for wall-control.

This study also examined regression-based ML model prediction performance when minor importance features were eliminated for approximating geophysical signatures. It was found that discarding MWD features of minor significance could increase processing speed without significantly compromising prediction accuracy. The predictive performance of most models remained stable, when less important features were omitted. Despite a slight decline in predictive performance, the accelerated training durations for larger datasets suggest excluding less important features is advantageous.

In addition, the choice of the regression-based ML analytical model, whether GP, NNs, or RFs, had little effect on the ML prediction outcomes, indicating that the predictive accuracy was primarily dependent on the quality of the extracted features. This observation is consistent with findings, which observed comparable prediction abilities among diverse ML models for rock types and geochemical assay results [8,18,31]. In particular, the prediction accuracy of geophysical measurement estimates for *dens*, *gamma*, *magsus*, *res*, and *cal* using DTs, SVMs, RFs, GPs, and NNs models in the *BR* dataset was much higher than in the *MM* or *COM* datasets. A great deal of these variances can be traced to differences in data volume between the two datasets. The differences in geological composition between the Brockman and Marra Mamba Formations may also account for the residual differences in predictive power between *BR* and *MM* studies.

5. Conclusions

This study introduces a method for evaluating subsurface geophysical characteristics by applying feature-importance-based methods and regression-based ML algorithms to MWD data. The ability of feature-importance-based methods to unveil the "black box" nature of ML methods enable greater interpretation and acceptance of these models. A framework was developed to assess the importance of MWD data features in estimating the geophysical properties of an orebody, including density, *gamma*, magnetic susceptibility, resistivity, and hole diameter. Through MARS and PPR feature importance analyses, MWD features were grouped based on their importance as negligible (0%), minor (<5%), or significant (>5%). Notably, several previously unrecognized MWD attributes, such as *fob*, and ratios derived from MWD features—*fobrop*, *fobtor*, *bapfob*, *torrop*, *baprop*, and *baptor*—were found to have significant importance for determining geophysical attributes. Future work will also be extended to the use of other Feature Importance algorithms, such as Shapley value regression, which are increasingly used as tools in variable importance analysis in other fields [49,50].

 doi:10.20944/preprints202501.2284.v1

Considering the varying importance of MWD features, the study compared the prediction performance of various regression-based ML analytical methodologies, omitting specific features at distinct levels, considering the varying importance of the MWD features. The results indicate that omitting MWD attributes classified as having zero to minor importance does not significantly diminish prediction accuracy. Therefore, the elimination of features with low importance can reduce computation time without compromising the accuracy of the ML model's estimates. In addition, empirical data revealed correlations as high as 0.91, between MWD attributes and orebody geophysical predictive values when RF was employed, validating the effectiveness of the proposed method.

These findings have significant implications for the mining industry. By utilizing these models, mining professionals can estimate precise and reliable short-term ore and waste tonnages. This predictive comprehension of orebody geophysical characteristics is crucial for mining operations, as it guides extraction and processing decisions. The high-resolution geological data derived from these models enables the recovery of high-grade ore containing economically valuable minerals. Through the high-resolution orebody representation afforded by the methodologies outlined in this study, mining geologists could better distinguish between high-grade, low-grade, and waste components. As a result, mining engineers can develop optimal excavation strategies that minimize the amount of waste material incorporated into processing facilities.

Acknowledgments: One of the authors (DG) received support through the MRIWA Postgraduate Research Scholarship and the AusIMM Education Endowment Fund Postgraduate Scholarship during his doctoral studies at Curtin University. Furthermore, the research was enabled by the advanced computing resources provided by the Pawsey Supercomputing Research Centre in Perth, Australia.

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

References

1. Silversides, K.; Melkumyan, A.; Wyman, D.; Hatherly, P. Automated Recognition of Stratigraphic Marker Shales from Geophysical Logs in Iron Ore Deposits. *Comput. Geosci.* **2015**, *77*, 118–125, doi:10.1016/j.cageo.2015.02.002.
2. Wedge, D.; Hartley, O.; McMickan, A.; Green, T.; Holden, E.J. Machine Learning Assisted Geological Interpretation of Drillhole Data: Examples from the Pilbara Region, Western Australia. *Ore Geol. Rev.* **2019**, *114*, 103118–103118, doi:10.1016/j.oregeorev.2019.103118.
3. Barr, M.V. Instrumented Horizontal Drilling for Tunnelling Site Investigation. PhD, University of London: Imperial College of Science and Technology, 1984.
4. Hatherly, P.; Leung, R.; Scheding, S.; Robinson, D. Drill Monitoring Results Reveal Geological Conditions in Blasthole Drilling. *Int. J. Rock Mech. Min. Sci.* **2015**, *78*, 144–154, doi:10.1016/j.ijrmms.2015.05.006.
5. Khorzoughi, M.B. Use of Measurement While Drilling Techniques for Improved Rock Mass Characterization in Open-Pit Mines. MSc, University of British Columbia: Vancouver, 2011.
6. Navarro, J.; Sanchidrian, J.A.; Segarra, P.; Castedo, R.; Paredes, C.; Lopez, L.M. On the Mutual Relations of Drill Monitoring Variables and the Drill Control System in Tunneling Operations. *Tunn. Undergr. Space Technol.* **2018**, *72*, 294–304, doi:10.1016/j.tust.2017.10.011.
7. van Eldert, J.; Schunnesson, H.; Johansson, D.; Saiang, D. Application of Measurement While Drilling Technology to Predict Rock Mass Quality and Rock Support for Tunnelling. *Rock Mech. Rock Eng.* **2020**, *53*, 1349–1358, doi:10.1007/s00603-019-01979-2.
8. Kadkhodaie-Ilkhchi, A.; Monteiro, S.T.; Ramos, F.; Hatherly, P. Rock Recognition from MWD Data: A Comparative Study of Boosting, Neural Networks, and Fuzzy Logic. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 680–684, doi:10.1109/LGRS.2010.2046312.
9. Galende-Hernández, M.; Menéndez, M.; Fuente, M.J.; Sainz-Palmero, G.I. Monitor-While-Drilling-Based Estimation of Rock Mass Rating with Computational Intelligence: The Case of Tunnel Excavation Front. *Autom. Constr.* **2018**, *93*, 325–338, doi:10.1016/j.autcon.2018.05.019.

10. Klyuchnikov, N.; Zaytsev, A.; Gruzdev, A.; Ovchinnikov, G.; Antipova, K.; Ismailova, L.; Muravleva, E.; Burnaev, E.; Semenikhin, A.; Cherepanov, A.; et al. Data-Driven Model for the Identification of the Rock Type at a Drilling Bit. *J. Pet. Sci. Eng.* **2019**, *178*, 506–516, doi:10.1016/j.petrol.2019.03.041.
11. Peck, J.P. Performance Monitoring of Rotary Blasthole Drills. **1989**, 395.
12. Scoble, M.J.; Peck, J.; Hendricks, C. Correlation between Rotary Drill Performance Parameters and Borehole Geophysical Logging. **1989**, *8*, 301–312.
13. Segui, J.B.; Higgins, M. Blast Design Using Measurement While Drilling Parameters.; Hunter Valley, NSW, 2001; pp. 28–31.
14. Navarro, J.; Seidl, T.; Hartlieb, P.; Sanchidrián, J.A.; Segarra, P.; Couceiro, P.; Schimek, P.; Godoy, C. Blastability and Ore Grade Assessment from Drill Monitoring for Open Pit Applications. *Rock Mech. Rock Eng.* **2021**, *54*, 3209–3228, doi:10.1007/s00603-020-02354-2.
15. Akyildiz, O.; Basarir, H.; Vezhapparambu, V.S.; Ellefmo, S. MWD Data-Based Marble Quality Class Prediction Models Using ML Algorithms. *Math. Geosci.* **2023**, doi:10.1007/s11004-023-10061-1.
16. Basarir, H.; Wesseloo, J.; Karrech, A.; Pasternak, E.; Dyskin, A. The Use of Soft Computing Methods for the Prediction of Rock Properties Based on Measurement While Drilling Data. *Proc. Eighth Int. Conf. Deep High Stress Min.* **2017**, 537–551, doi:10.36487/acg_rep/1704_36_basarir.
17. Beattie, N. Monitoring-While-Drilling for Open-Pit Mining in a Hard Rock Environment. Master of Science, Queen's University: Kingston, Ontario, Canada, 2009.
18. Khushaba, R.N.; Melkumyan, A.; Hill, A.J. A Machine Learning Approach for Material Type Logging and Chemical Assaying from Autonomous Measure-While-Drilling (MWD) Data. *Math. Geosci.* **2021**, doi:10.1007/s11004-021-09970-w.
19. Martin, J. Application of Pattern Recognition Techniques to Monitoring-While- Drilling on a Rotary Electric Blasthole Drill at an Open-Pit Coal Mine. MSc, Queen's University: Kingston, Ontario, Canada, 2007.
20. Silversides, K.L.; Melkumyan, A. Multivariate Gaussian Process for Distinguishing Geological Units Using Measure While Drilling Data. In Proceedings of the Minig Goes Digitial; Taylor & Francis Group: London, 2019; pp. 94–100.
21. Silversides, K.L.; Melkumyan, A. Boundary Identification and Surface Updates Using MWD. *Math. Geosci.* **2020**, doi:10.1007/s11004-020-09891-0.
22. Schunnesson, H. Drill Process Monitoring in Percussive Drilling: A Multivariate Approach for Data Analysis. Licentiate, Lulea University of Technology: Lulea, Sweden, 1990.
23. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
24. Goldstein, D.M.; Aldrich, C.; O'Connor, L. A Review of Orebody Knowledge Enhancement Using Machine Learning on Open-Pit Mine Measure-While-Drilling Data. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1343–1360, doi:https://doi.org/10.3390/make6020063.
25. Ker, P. Iron Ore Supply Slump as Rio Runs Late on New Mines. *Aust. Financ. Rev.* 2021.
26. De-Vitry, C.; Vann, J.; Arvidson, H. Multivariate Iron Ore Deposit Resource Estimation–a Practitioner's Guide to Selecting Methods. *Trans Inst Min. Metall. Sect B* **2010**, *119*, 154–165.
27. Jones, H.; Walraven, F.; Knott, G. Natural Gamma Logging as an Aid to Iron Ore Exploration in the Pilbara Region of Western Australia.; Perth, Australia, 1973.
28. Tittman, J.; Wahl, J.S. The Physical Foundations of Formation Density Logging (Gamma-Gamma). *GEOPHYSICS* **1965**, *30*, 284–294, doi:10.1190/1.1439574.
29. Yang, Q.; Tan, M.; Zhang, F.; Bai, Z. Wireline Logs Constraint Borehole-to-Surface Resistivity Inversion Method and Water Injection Monitoring Analysis. *Pure Appl. Geophys.* **2021**, *178*, 939–957, doi:10.1007/s00024-021-02674-6.
30. Elsayed, M.; Isah, A.; Hiba, M.; Hassan, A.; Al-Garadi, K.; Mahmoud, M.; El-Husseiny, A.; Radwan, A.E. A Review on the Applications of Nuclear Magnetic Resonance (NMR) in the Oil and Gas Industry: Laboratory and Field-Scale Measurements. *J. Pet. Explor. Prod. Technol.* **2022**, *12*, 2747–2784, doi:10.1007/s13202-022-01476-3.
31. Goldstein, D.; Aldrich, C.; O'Connor, L. Enhancing Orebody Knowledge Using Measure-While-Drilling Data: A Machine Learning Approach. *IFAC Pap.* **2024**, *58*, 72–76.

32. Khorzoughi, B.M.; Hall, R. Processing of Measurement While Drilling Data for Rock Mass Characterization. *Int. J. Min. Sci. Technol.* **2016**, *26*, 989–994, doi:10.1016/j.ijmst.2016.09.005.
33. van Eldert, J.; Schunnesson, H.; Saiang, D.; Funehag, J. Improved Filtering and Normalizing of Measurement-While-Drilling (MWD) Data in Tunnel Excavation. *Tunn. Undergr. Space Technol.* **2020**, *103*, 103467–103467, doi:10.1016/j.tust.2020.103467.
34. Ghosh, R.; Gustafson, A.; Schunnesson, H. Development of a Geological Model for Chargeability Assessment of Borehole Using Drill Monitoring Technique. *Int. J. Rock Mech. Min. Sci.* **2018**, *109*, 9–18, doi:10.1016/j.ijrmms.2018.06.015.
35. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1–67.
36. Shao, Q.; Traylen, A.; Zhang, L. Nonparametric Method for Estimating the Effects of Climatic and Catchment Characteristics on Mean Annual Evapotranspiration: NONPARAMETRIC METHOD FOR MEAN ANNUAL EVAPOTRANSPIRATION. *Water Resour. Res.* **2012**, *48*, doi:10.1029/2010WR009610.
37. Millborrow, S. *Package 'Earth';* Multivariate Adaptive Regression Splines; 2023;
38. Friedman, J.H.; Stuetzle, W. Projection Pursuit Regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817–823.
39. R Core Team *R Stats Package*; R Foundation for Statistical Computing, 2022;
40. Su, X.; Yan, X.; Tsai, C.-L. Linear Regression. *WIREs Comput. Stat.* **2012**, *4*, 275–294.
41. Kotsiantis, S.B. Decision Trees: A Recent Overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283.
42. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28.
43. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
44. Schulz, E.; Speekenbrink, M.; Krause, A. A Tutorial on Gaussian Process Regression: Modelling, Exploring, and Exploiting Functions. *J. Math. Psychol.* **2018**, *85*, 1–16.
45. Bishop, C.M. Neural Networks and Their Applications. *Rev. Sci. Instrum.* **1993**, *65*.
46. *RegressionLearner Toolbox*; The MathWorks Inc.: Natick, Massachursetts, 2024;
47. Ghosh, R.; Schunnesson, H.; Kumar, U. Evaluation of Rock Mass Characteristics Using Measurement While Drilling in Boliden Minerals Aitik Copper Mine, Sweden. In *Mine Planning and Equipment Selection*; Drebenstedt, C., Singhal, R., Eds.; Springer International Publishing: Cham, 2014; pp. 81–91 ISBN 978-3-319-02677-0.
48. Roe, K.D.; Jawa, V.; Zhang, X.; Chute, C.G.; Epstein, J.A.; Matelsky, J.; Shpitser, I.; Taylor, C.O. Feature Engineering with Clinical Expert Knowledge: A Case Study Assessment of Machine Learning Model Complexity and Performance. *PLOS ONE* **2020**, *15*, e0231300, doi:10.1371/journal.pone.0231300.
49. Aldrich, C. Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework. *Minerals* **2020**, *10*, 420, doi:https://doi.org/10.3390/min10050420.
50. Deng, S.; Aldrich, C.; Liu, X.; Zhang, F. Explainability in Reservoir Well-Logging Evaluation: Comparison of Variable Importance Analysis with Shapley Value Regression, SHAP and LIME. *IFAC Pap.* **2024**, *58*, 66–71, doi:https://doi.org/10.1016/j.ifacol.2024.09.292.