RGNZ/MGPB

Benchmarking Machine Learning Models for Penetration Rate Prediction in Rotary Drilling: A Case Study from the Sungun Copper Mine

Rudarsko-geološko-naftni zbornik
(The Mining-Geology-Petroleum Engineering Bulletin)
DOI: 10.17794/rgn.2026.5.12

Original scientific paper

Ali Nemati Vardin[1], Masoud Monjezi[1], Mojtaba Rezakhah[1], Manoj Khandelwal[2,3,4*], Hasel Amini Khoshalan[5]

[1] *Department of Mining Engineering, Faculty of Engineering, Tarbiat Modares University, Tehran, Iran.*
[2] *Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, VIC, Australia.*
[3] *Laboratory of Sustainable Development in Natural Resources and Environment, Institute for Advanced Study in Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam.*
[4] *Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam.*
[5] *Department of Mining Engineering, Faculty of Engineering, University of Kurdistan, Sanandaj, Iran.*

Abstract

Mining operations require an accurate prediction of rotary drilling penetration rate (ROP) to optimize efficiency and reduce costs. ROP estimation is challenging due to complex, nonlinear relationships among rock properties, equipment specifications, and operational parameters. This study applies six machine learning models, Random Forest, CatBoost, LightGBM, XGBoost, Support Vector Machine, and Extra Trees to forecast ROP using data from 85 drillholes at the Sungun copper mine, Iran. Input features include mining rock mass rating, Schmidt hammer rebound number, rotation pressure, and weight on bit. The Extra Trees model achieved the best performance, with R^2 values of 1 (training) and 0.93 (testing) and RMSE of 0 and 0.937, respectively, demonstrating high accuracy and generalization. The results confirm that machine learning, particularly Extra Trees, can effectively model complex ROP dependencies, enabling improved equipment selection, operational planning, and cost estimation in rotary drilling.

Keywords:
Rate of penetration; Drilling; Mining; Machine learning models; Sungun copper mine

1. Introduction

Accurate estimation of drilling equipment performance is essential for mining operations, as the penetration rate (ROP) directly influences mine design, development, economics, and drilling expenditure avoidance (**Inanloo Arabi Shad et al., 2018**). ROP is a useful indicator of how various criteria affect drilling pace in mining and construction, yet basic estimation remains problematic for equipment selection, operational planning, and product pricing (**Ataei et al., 2015**). Drilling rate formulae can estimate overall drilling expenses and assist machine selection. ROP is influenced by a broad range of factors, including rock mass properties, material characteristics, and drilling equipment specifications. While drilling parameters can be controlled, geological conditions and rock features are beyond human influence **(Ataei et al., 2015; Darbor et al., 2019a; Inanloo et al., 2018; Kazemi et al., 2024; Lawal et al., 2021**).

Although understanding ROP is critical for improving drilling efficiency, building an accurate prediction model is difficult (**Elmgerbi et al., 2021**). Empirical approaches using field data have been developed to handle terrain-related factors, and numerous studies on rotary and percussive drilling have confirmed that ROP depends on rock characteristics. The relationship between rock properties and ROP is complex and nonlinear. Researchers have constructed statistical models to predict ROP from experimental data (**Ersoy and Walker, 1995; Bilim, 2011; Howarth and Rowlands, 1987; Mensa-Wilmot et al., 1999; Singh, 1976; Yarali and Kahraman, 2011**). However, these statistical frameworks often fail to generalise across diverse drilling scenarios or different geological environments and equipment types (**Pathankar and Misra, 1980; Bond, 1990; Rabia, 1982**). They only partially capture the impact of rock characteristics on ROP, and few can predict ROP using only a limited set of rock properties.

The Norwegian Institute of Technology developed the Drilling Rate Index (DRI) for percussive drillability assessment (**Norwegian Institute of Technology, 1981**).

* Corresponding author: Manoj Khandelwal
e-mail address: m.khandelwal@federation.edu.au; khandelwalmanoj@tdtu.ac.vn
Received: 3 April 2026. Accepted: 28 May 2026.
Available online: 26 October 2026

Linear relationships have been observed between ROP and rock properties such as Young's modulus, compressive strength, tensile strength (**Cheniany et al., 2012**), bulk density, apparent porosity, P-wave velocity, and Schmidt hammer rebound (**Hoseinie, Aghababaei, and Pourrahimian, 2008**). Several rock hardness tests (compressive and tensile strengths) also show substantial correlation with ROP (**Thuro, 1997**). Multiple regression analyses have been performed for rotary, down-the-hole (DTH), and hydraulic top hammer drills (**Kahraman, 1999**). The Rock Penetrability Index (RPi) incorporates UCS, Schimazek's F-abrasivity, Mohs hardness, rock texture, particle size, and Young's modulus. Density (**Bilim, 2011**), Rock Quality Designation (RQD), and other discontinuities also significantly affect drilling performance (**Krúpa et al., 2018**). Researchers have proposed drillability models considering UCS, quartz content, bit rotational speed, joint dip, alternation, and fragment size. Theoretical and experimental studies have shown that ROP can be predicted from specific energy, design, and operational factors of the drilling machine (**Hustrulid and Fairhurst, 1971**), with correlations between ROP and drilling thrust, rock strength, bit shape, rig power, and geological discontinuities (**Thuro, 1997**). These studies yield empirical models based on uniaxial compressive strength, analyses of geological influence on ROP, and applications of AI for engineering prediction (**Bauer and Crosby, 1990; Fish, 1968; Kahraman et al., 2003**).

Statistical techniques may not adequately handle non-linear and complex problems in ROP prediction (**Zhao et al., 2022**). Multicollinearity among independent variables undermines the predictive stability of multiple regression models and complicates their ability to address non-linearity. To overcome these limitations, artificial intelligence models have been developed and have improved prediction accuracy over time (**Koopialipoor et al., 2019**). The growth of information technology has enabled advanced drilling and completion technologies, and the use of machine learning (ML) for ROP estimation is increasing. ML techniques use data to find correlations between drilling factors (**Pacis et al., 2023**), offering a significant advantage over analytical ROP models by allowing flexible model forms controlled by hyperparameters. Unlike analytical models, ML does not require predetermined equations. However, increased model complexity can reduce interpretability and raise overfitting risk. Neural networks may face challenges such as unstable learning and slow convergence when input variables exhibit multicollinearity.

An artificial neural network (ANN) model outperformed regression models in predicting diamond drilling penetration rates (**Kahraman, 2016**). Darbor et al. found that MLP-ANN was superior to nonlinear multiple regression (NLMR) for ROP prediction. **Fattahi and Bazdar (2017)** used five hybrid ANN models to estimate DRI, with ANN-SSA achieving the best performance ($R^2 = 0.97$). Shad et al. (**Inanloo et al., 2018**) investigated rock mass parameters and machine specifications, introducing the percentage of iron oxide as a new metric for rotary drill ROP measurement. **Kamran (2021)** developed a probabilistic approach using AdaBoost, Random Forest (RF), and Decision Tree (DT) for DRI estimation, with Monte Carlo simulations improving likelihood distribution estimation. **Sakız et al. (2021)** focused on rock strength and abrasivity to estimate DRI using a fuzzy inference system (FIS), significantly improving prediction accuracy and efficiency.

Recent literature (2023–2025) has further demonstrated the strong potential of ensemble tree-based and gradient boosting algorithms for ROP prediction. **Feng et al. (2023)** developed an explainable ensemble model for geothermal drilling, with CatBoost achieving the best performance. **Yehia et al. (2023)** compared six ML algorithms, including Random Forest, Extra Trees, XGBoost, and SVM on a geothermal dataset and found that Extra Trees and Random Forest delivered the highest accuracy, directly motivating the inclusion of Extra Trees in our benchmark. **Ghorbani and Yagiz (2024)** estimated tunnel boring machine penetration rates using five gradient boosting algorithms, reporting that XGBoost and CatBoost achieved R^2 values above 0.96. **Al-Sahlanee et al. (2023)** adopted four ensemble algorithms for oil- and gas-well ROP prediction and confirmed the consistent superiority of ensemble methods over single learners. **Zhang et al. (2025)** constructed a stacked ensemble combining SVR, Random Forest, Extra Trees, GBM, and LightGBM, achieving an R^2 of 0.9335, with Extra Trees and Random Forest ranking among the top base learners.

A critical examination of previous ROP prediction studies reveals several recurring limitations. Many earlier works used small datasets without reporting test set performance or cross-validation, raising concerns about overfitting. Most comparisons involved only one or two ML models, leaving practitioners uncertain about the best method for a given dataset. Additionally, a significant number of studies relied on laboratory-scale drilling tests rather than field data, limiting real-time applicability. These gaps motivate the present study, which provides a systematic comparison of six ML models on a field dataset, supported by cross-validation and transparent performance reporting.

ML methods can achieve high accuracy and handle extremely non-linear correlations between predictors and response variables, conditions typical in drilling operations. ML models outperform statistical and empirical methods when trained on large, complex datasets, especially when assumptions of normality and linearity are violated. Moreover, ML models can easily integrate data from real-time sensors to enhance prediction accuracy, a capability that statistical and empirical methods lack.

This study uses data from the Sungun copper mine in Iran to construct ROP prediction models considering

Table 1: Operational information of equipment's

N	Equipment	Max rotation speed	Max rotation pressure (psi)	Max Thrust (psi)	Bit type	Bit diameter
1	ATLAS ROC L8	130	780	750	Flat face	6.5
2	Husherr	120	2200	1700	Flat face	6.5
3	SWDE200	60	2200	800	Flat face	6.5
4	Tamrock 1500	140	1000	1800	Flat face	6.5

Table 2: Descriptive statistics of input and output parameters

	WOB (psi)	rotation pressure (psi)	Schmidt hammer rebound	MRMR	ROP
N	85	85	85	85	85
Avg	1118	1427	37.22	40.89	28.81
Max	1800	2200	70	84	25.101
Min	600	600	20	15	13.10

various rock features and equipment. Six ML methods, Random Forest, CatBoost, LightGBM, XGBoost, Support Vector Machine, and Extra Trees, are applied and compared to forecast rotary drilling penetration rates, contributing a practical benchmark for mining engineering. Existing ROP prediction studies have typically employed one or two ML models, often on laboratory-scale data, making it difficult for practitioners to select the most suitable method. Moreover, no previous study has benchmarked Extra Trees, CatBoost, LightGBM, XGBoost, Random Forest, and SVM on a consistent field dataset. This study fills that gap by providing a head-to-head comparison on real operational data from Sungun copper mine, with a clear recommendation (Extra Trees) for similar settings.

2. Material and Data Description

2.1. Case Studies

Varzghan city, which is located in East Azerbaijan Province, is the location of the Sungun copper complex. It is situated at a longitude of 46 degrees 43 minutes and a latitude of 38 degrees 43 minutes north. The distance between this complex and either Varzghan or Tabriz is estimated to be 25 kilometers by car and 100 kilometers by car, respectively. This mountainous region has a maximum height of 2700 meters and an average height of 2000 meters above sea level.

2.2. Data Description

Finding the best predictor for ROP at Sungun mine is the goal of this research, which aims to build intelligent models utilizing Random Forest, CatBoost, LightGbm, XGBoost, SVM and Extra Tree. An effort was made to achieve this purpose by compiling a dataset that included 85 drillholes that were carried out in the Sungun mine. The weight on Bit (WOB), the Rotation pressure (RP), the mining rock mass rating (MRMR) and the Schmidt hammer rebound number are the input parameters that are used in the process of ROP prediction, respectively. Operational information of equipment and descriptive statistics of input and output parameters are presented in **Tables 1** and **2**. The objective of this research was to identify a system that is not only easy to use but also suited for fieldwork and can be installed in a short amount of time. In light of the findings of the investigations, the features that were deemed to be of utmost significance were selected after discussions with mine authorities and operators of drilling rigs. For this research, data obtained from drilling at the Sungun copper mine were used. **Table 1** presents data from the Husherr and Tamrock 1500 systems, respectively. The total number of holes drilled by each of the four pieces of equipment was 85, with each equipment producing 19, 22, 23, and 21 drill holes, respectively. The data collection process involves the collection of a variety of ground conditions. The fact that the ML approach needs a substantial quantity of training data in order to create reliable predictions was taken into consideration when the data was partitioned for training and testing (**Khandelwal and Armaghani, 2016**). On the other hand, if there is an excessive amount of training data, the model could overfit, which would result in poor performance when dealing with previously collected data (**Shan et al., 2022**). Whether it be cross-validation or the percentage split strategy, the method that is used to divide the dataset into training and validating sets might have an effect on the performance of the model (**Armaghani et al., 2015**). This is an essential factor to keep in mind. In the event that the training set is insufficiently large, the model could not have sufficient data from which to learn, which might result in unacceptable performance (**Anvari et al., 2022**). On the other hand, if the testing set is too small, it may not be a realistic representation of the complete dataset, and the performance of the model might not be evaluated accurately. The percentage split strategy is a method that is often used and is beneficial for ML modelling. This is even though there is no established agreement on the appropriate ratio for splitting the data. For the present data partitioning, this method

was selected as the best option. **Figures 1** and **2** illustrate the heat map correlation matrix and the all data samples scatter matrix. The spatial organization of these parameters is shown in the matrix of correlations graph shown in **Figure 2**. This graph also displays the link between ROP and the other variables that were input. It is clear from **Figure 2** that there is a pairwise relationship between the variables, with adequate values for correlation, as well as the median distribution of frequencies for each component. It has the ability to calculate the correlation coefficient, which is shown by the triangle in the top right corner of the graph, while the diagonal portion of the graph depicts the statistical range for each and every characteristic. As shown in **Figure 2**, ROP is affected in a variety of ways, dependent on a number of different conditions.

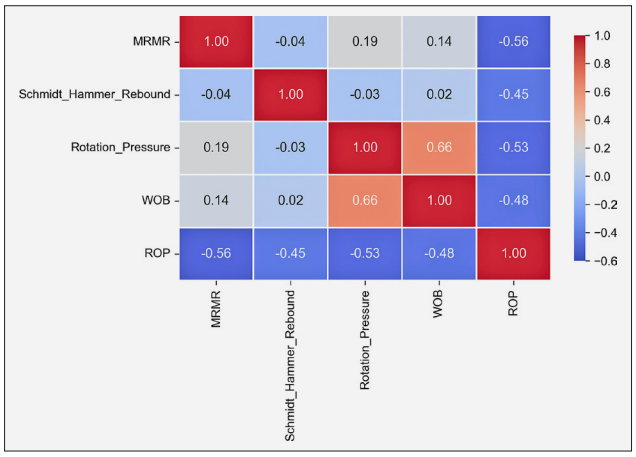


Figure 1: Mapping the correlations between all data samples using a heat map

Although the dataset comprises 85 drill holes, which may appear limited for machine learning applications, it is important to note that the sample-to-feature ratio is approximately 21:1 (85 samples with four input features). This ratio is generally considered sufficient for tree-based ensemble methods (Random Forest, CatBoost, XGBoost, Extra Trees, LightGBM) as these algo-

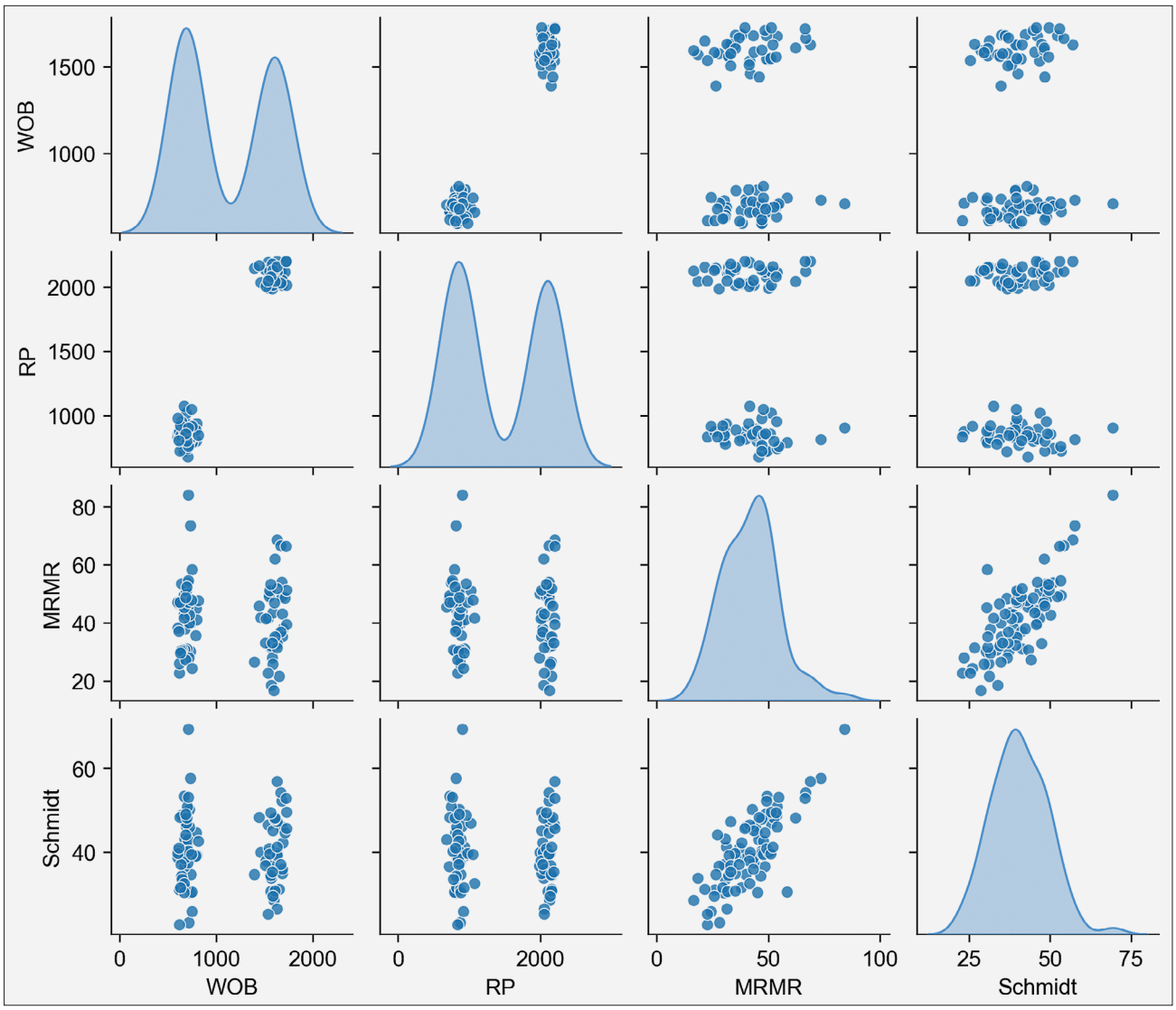


Figure 2: A summary of the scatter plot for every single data point

rithms are inherently robust to overfitting when combined with proper regularisation and cross-validation. Moreover, previous studies on ROP prediction have successfully employed similar or even smaller datasets (**Darbor et al., 2019b; Kazemi et al., 2024**). Nevertheless, it is acknowledged that a larger and more diverse dataset would further improve model generalisability, particularly for capturing extreme operating conditions.

3. Model Development

The machine learning techniques employed to develop models are discussed in this section. For model training and evaluation, the dataset was randomly split into 80% training (68 samples) and 20% testing (17 samples) using a fixed random seed (random_state=42). This ratio was chosen to provide sufficient data for training while retaining a representative hold-out set for unbiased performance assessment.

To mitigate overfitting, especially given the limited dataset (85 samples), the following measures were implemented for all tree-based models: (i) maximum tree depth was limited to 10; (ii) minimum samples per leaf was set to 3; (iii) 10-fold cross-validation was performed to monitor performance stability; (iv) training and test metrics are reported side-by-side to quantify the generalisation gap. For Extra Trees, the randomised split selection further reduces the risk of learning spurious patterns. No extensive hyperparameter tuning was performed to avoid over-optimisation on the small dataset.

3.1. Soft Computing Techniques

3.1.1. Support Vector Machine (SVM)

In the realm of classification and regression applications, the SVM is a dependable supervised learning technique. For it to function, it must first identify the hyperplane that is most suitable for maximizing the margins between the various classes in the feature space. An SVM is a popular option for a variety of classification problems because of its capacity to handle high-dimensional data. To address non-linear classification problems, the SVM model was built in this work using the SVR with a radial basis function (RBF) kernel. The training set was used to assess the model after it had been trained on the training set (**Alam et al., 2024; Cortes and Vapnik, 1995**).

3.1.2. CatBoost

Yandex created a gradient boosting algorithm called CatBoost with the intention of handling categorical data in a manner that is both efficient and reliable. The processing of categorical characteristics is handled in a novel manner, which eliminates the need for human encoding and reduces the likelihood of overfitting. As a result of CatBoost's capacity to manage categorical variables, it is especially well-suited for datasets that include mixed data types, which are typical in AS mapping. During the process of model training, the approach makes use of an ordered boosting strategy, which efficiently reduces the amount of target leakage. **Equation 1** illustrates how this is accomplished via the use of a unique mathematical formulation that modifies the values of features depending on the statistics related to the target (**Dorogush, Ershov, and Gulin, 2018**).

$$x_{p,k} = \frac{\sum_{j=1}^{n}\left[x_{j,k} = x_{i,k}\right] Y_j}{\sum_{j=1}^{n}\left[x_{j,k} = x_{i,k}\right]} \tag{1}$$

In order to improve the reliability of the data that is used for training choices, **Equation 1** is used to determine the conditional expectation of the target variable. In addition, according to **Equation 2**, CatBoost uses permutations as part of the training process in order to further enhance feature values. This is accomplished by adjusting to the natural order of data inputs.

$$x_{\sigma(p),k} = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma(j),k} = x_{\sigma(i),k}\right] Y_{\sigma(j)} + \beta P}{\sum_{j=1}^{p-1}\left[x_{\sigma(j),k} = x_{\sigma(i),k}\right] + \beta} \tag{2}$$

In this context, the symbols β and P represent the weight and the previous value, respectively.

This approach makes it possible to include both historical data and real-world observations in an effective updating process. In order to properly handle high-dimensional categorical data, avoid overfitting, and increase generalization across a variety of datasets, CatBoost is able to make use of these mathematical notions (**Cetinkaya and Kocaman, 2024**).

3.1.3. XGBoost

The gradient boosting tree (GBDT) serves as the foundation for the XGBoost algorithm, which is described as a unique and efficient boosting technique. Within the framework of the second-order Taylor function approximation, this approach employs Newton's method to tackle the problem of the loss function's extreme value. Within the context of the loss function, a regularization term is specifically included (**Suaza-Medina et al., 2024**). XGBoost is a boosting algorithm that comprises three key components: the loss function, which quantifies the discrepancy between the predicted outcomes and the actual values; the weak learner, commonly represented by decision trees (such as Classification and Regression Trees and Dropout Additive Regression Trees), which are developed according to various boosting methodologies employing distinct tree construction techniques; and the aggregated output, which refers to the results produced by the specific integration algorithm. **Equation 3** may be used to determine the objective function, which is as follows:

$$Obj = \sum_{i=1}^{n} l(y_i', y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

In the following notations: y_i' denotes the predicted value for sample i, y_i signifies the actual value for sample i, $l(y_i', y_i)$ indicates the loss function, and $\Omega(fk)$ represents the regularization term.

As opposed to the more common practice of using gradient boosting trees, XGBoost aims to minimize the goal function. Two components make up the goal function for every tree. An important part of the model is the differentiable loss function, which measures the discrepancy between the expected and observed values and controls the model's empirical risk. The regularization term is the second component, and it is necessary in order to prevent the model from being overfit (**Yang Liu, Jinming Yang, Xi Chen, Junqiang Yao, 2024**); (**Wenju Zhao, Zhaozhao Li, Haolin Li, 2024**).

3.1.4. Random Forests (RF)

Breiman (**Breiman, 2001**) was the first person to present the concept of random forests, which was based on an earlier study by Ho. Random forests are an ensemble method that may be used to improve decision tree models by combining the results of many decision trees using bagging or bootstrapping. Predictions are averaged out on fresh data after training decision trees on tiny samples taken from the training data (B times with replacement)(x ′) with **Equation 4**:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x'). \tag{4}$$

In many different kinds of industrial applications, random forests are used rather often. It was decided to build quantile regression random forests as an extension of the traditional regression random forest. Beyond this, companies also benefit from being able to quantify uncertainty and incorporate a wider range of acceptable levels of uncertainty in their projection. When comparing conditional distributions, quantile regression forests vary most noticeably from classical random forests in that, instead of only storing information about the mean of all observations at a node, they include the value of all observations in that node (**Balestra et al., 2024**).

3.1.5. Extra tree regression (ETR)

The RF model, which was first presented by Geurts et al. (**Geurts, Ernst, and Wehenkel, 2006**), was the basis for the development of the Extra Tree Regressor (ETR), which represents a considerable improvement in ensemble learning. A collection of unpruned regression trees, each constructed using a traditional top-down process, is used by the ETR technique. The RF model uses a two-stage process for regression analysis called bagging and bootstrapping, which is different from the outlined technique. During the creation of individual trees, the ETR model makes use of an approach that is deterministic in nature. Instead of using a selection method, as RF does to find the best split from a random subset of attributes at each node, ETR chooses a split point for each feature at random and then chooses the best split from those alternatives (**Qayyum et al. 2022; Sharafati, Asadollah, and Hosseinzadeh, 2020**). **Equation 5** is the mathematical formulation of this technique, and it is as follows:

$$Split_{ETR} = \arg\min_{f,s}[\mathrm{Error}(f,s)] \tag{5}$$

Within the framework of this discussion, the variable $Split_{ETR}$ represents the split that was chosen by the ETR algorithm. For a given feature, the symbol f stands for a feature, and the sign s stands for a split point that was chosen at random for that feature. It is the responsibility of the $\mathrm{Error}(f,s)$ function to calculate the reduction in error that is brought about by the split. This error is minimized by the algorithm, which selects the pair of f and s that minimizes it. The RF algorithm's bagging stage typically involves each ensemble tree casting a vote, with the final prediction often being derived by averaging these votes. An approach that is similar to this one is used by the ETR technique; however, it makes use of a collection of unpruned trees that are more diverse. **Equation 6** is a concise expression that may be used to describe the mathematical representation of the output of the final emission trading model.

$$Y_{ETR} = \frac{1}{N}\sum_{i=1}^{N} T_i(X) \tag{6}$$

Three variables, the input feature vector (X), the number of trees (N), and the i-th tree (T_i) in the ensemble, determine the expected output (Y_{ETR}). The ETR algorithm offers a novel regression method that uses mathematical components to achieve a balance between accuracy and unpredictability.

3.1.6. Light Gradient Boosting Machine (LightGBM)

Regression and classification issues may be resolved with the help of the LightGBM method, which is a gradient-based decision tree. One of its benefits is that it can be applied to large amounts of data, and it also has a fast processing speed. The continuous feature values are divided into discrete bins using a histogram-based method, which then builds feature histograms by making use of these bins throughout the training process. This approach does not locate separation points in the continuous feature values. The continuous factors that are included in the training set are transformed into discrete variables via the use of this technique, which results in a reduction in the amount of time and memory that the model requires to operate (**Cakiroglu et al., 2023**).

3.2. Performance Indicators

To precisely evaluate the predictive performance of the models utilized in this study, five principal assess-

ment metrics were employed: Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Mean Absolute Error (MAE), Variance Account For (VAF), and, Mean Absolute Percentage Error (MAPE). **Equations 7-11** represent the mathematical expiration of these metrics. The metrics were computed for both the training and testing stages. This review aimed to assess the models' efficacy and to choose the best appropriate model for actual engineering applications. The five performance metrics, extensively referenced in contemporary research (**Hasanipanah et al., 2016; Li et al., 2022; Ramesh Murlidhar et al., 2021**), are delineated as follows:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{ROP\text{-}i}-y_{ROP\text{-}i})^2} \quad (7)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{ROP\text{-}i}-y_{ROP\text{-}i})^2} \quad (8)$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_{ROP\text{-}i}-y_{ROP\text{-}i}| \quad (9)$$

$$\mathrm{VAF}=\left[1-\frac{\mathrm{var}(y_i-t_i)}{\mathrm{var}(y_i)}\right]\times 100\% \quad (10)$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-t_i}{Y_i}\right| \quad (11)$$

4. Result and Discussion

For this investigation, the performance of a number of ML algorithms was assessed to anticipate the drilling performance measure, more precisely, the rate of penetration (ROP) in a rock formation dataset obtained from the Sungun copper mine. There are several different supervised learning strategies that are included in the models that have been chosen for this assignment. These techniques include linear and non-linear regression methods. The objective was to determine which model provides the most accurate prediction of ROP, which is an essential factor in optimizing drilling operations.

Collectively, **Figures 3-8** reveal a clear hierarchy in model performance. The Extra Trees (ETR) model (see **Figure 7**) exhibits the tightest clustering around the diagonal in the testing phase, confirming its superior generalisation capability. Random Forest (RF) (see **Figure 6**) and CatBoost (see **Figure 4**) also show reasonable clustering, though with more scatter. XGBoost (Figure 5), despite near-perfect training, displays a noticeable drop in testing accuracy, indicative of overfitting. LightGBM (see **Figure 8**) and SVM (see **Figure 3**) perform poorly, with wide dispersion and low R^2 values, suggesting that these algorithms are not well-suited to the small, noisy dataset. The visual patterns are fully consistent with the quantitative metrics presented in **Table 3** and the ranking in **Table 5**.

SVM achieved training $R^2 = 0.262$ and RMSE = 17.3, and test $R^2 = 0.132$ and RMSE = 23.7. The wide scatter around the diagonal, particularly in the testing phase, indicates that the SVM model failed to capture the nonlinear relationships between the input parameters and ROP. This poor performance is attributed to the small dataset size (85 samples) and the lack of extensive hyperparameter tuning.

CatBoost achieved training $R^2 = 0.999$ and RMSE = 0.456, and test $R^2 = 0.909$ and RMSE = 7.65. The training points align almost perfectly with the diagonal, indicating excellent fitting. However, the testing points show increased scatter, suggesting mild overfitting. Despite this, CatBoost outperforms SVM and LightGBM, ranking third overall among. XGBoost achieved training $R^2 = 1.000$ and RMSE = 0.001 (near-perfect training), and test $R^2 = 0.904$ and RMSE = 7.84. The training points lie exactly on the diagonal, indicating almost perfect memorisation of the training data. The testing performance, while still reasonable, shows more scatter, reflecting a degree of overfitting. XGBoost ranks fourth overall, with strong performance in MAE and MAPE but weaker R^2 compared to Extra Trees and Random Forest. the six models.

RF achieved training $R^2 = 0.981$ and RMSE = 2.78, and test $R^2 = 0.913$ and RMSE = 7.48. The training points are densely clustered around the diagonal, while the testing points show moderate scatter. RF ranks second overall, with strong R^2 and VAF but relatively higher MAE and MAPE, indicating that it captures variance well but may have some systematic bias in absolute error. ETR achieved training $R^2 = 1.000$ and RMSE ≈ 0 (perfect fit), and test $R^2 = 0.937$ and RMSE = 6.36. The training points lie exactly on the diagonal, while the testing points are tightly clustered around the diagonal with minimal scatter. This demonstrates that ETR not only fits the training data perfectly but also generalises exceptionally well to unseen data. The gap between training and testing performance is smaller than for XGBoost and CatBoost, confirming that the randomised split selection in ETR reduces overfitting. ETR ranks first overall across all metrics.

When compared to all of the trained models in Tables 3 and 5, the ETR model was shown to have the highest level of performance. **Figures 3-8** show how well each model's projected values match the training dataset's actual values. The majority of the data points that were generated by the ETR model were distributed closely along the ideal fit line, which is also known as the perfect correlation line, as can be seen in the scatter plot. This demonstrates that the model has a strong capacity for prediction.

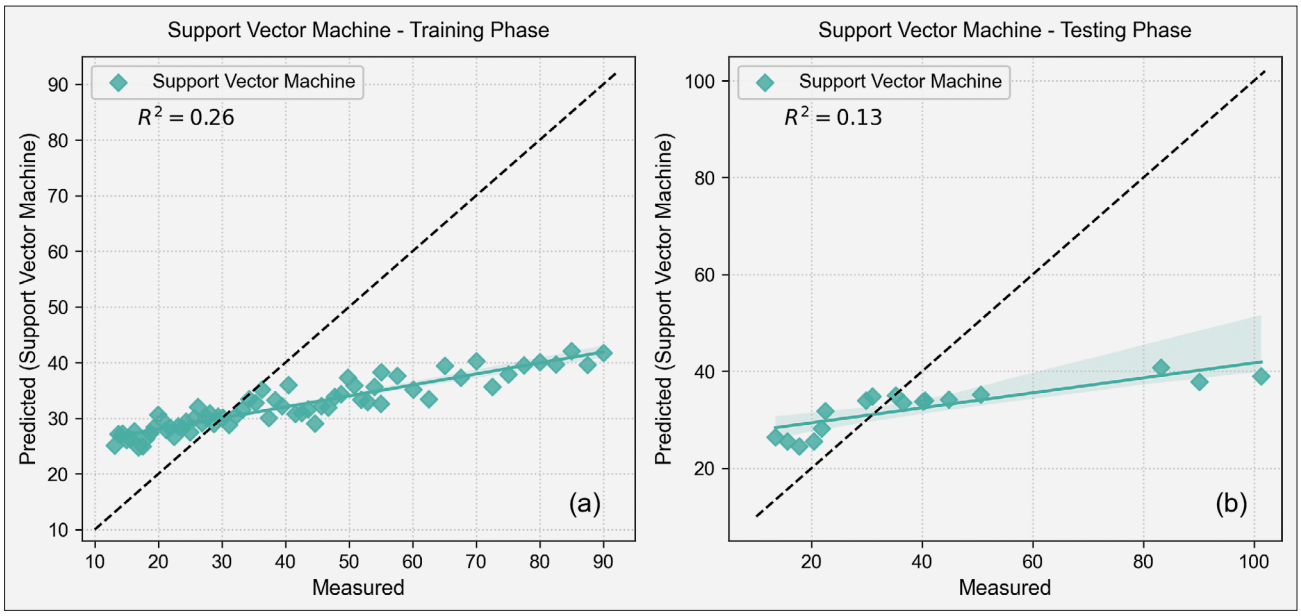


Figure 3: Measured vs. predicted rate of penetration (ROP) in training and testing phases for the SVM Model

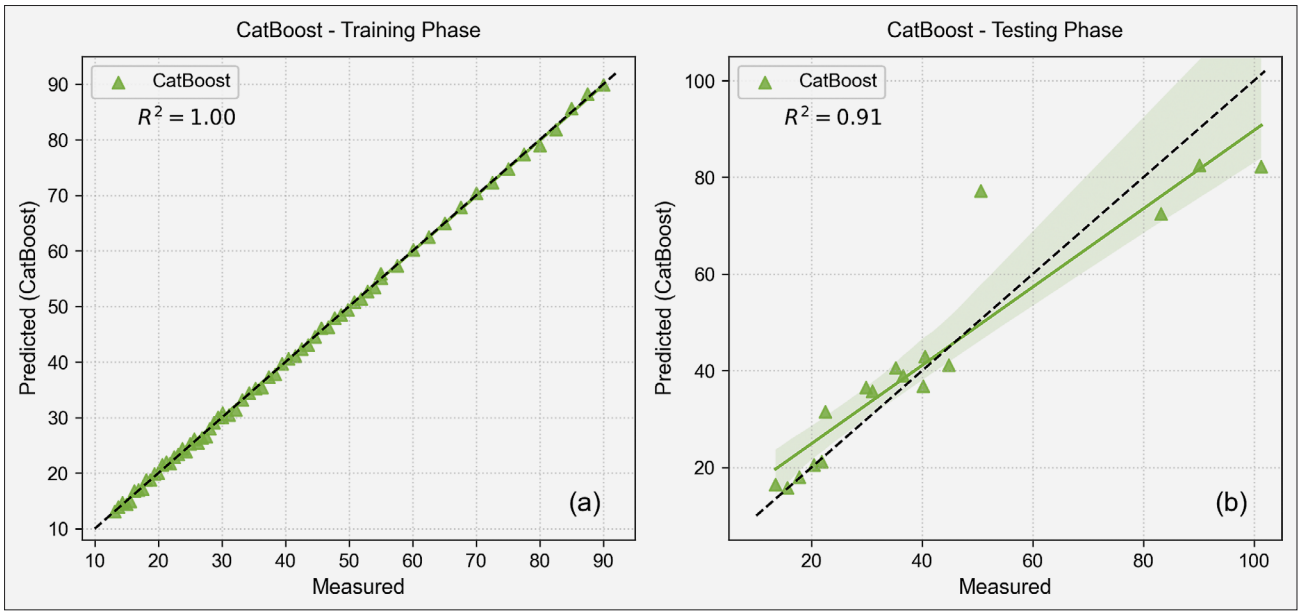


Figure 4: Measured vs. predicted rate of penetration (ROP) in training and testing phases for the CatBoost Model

The evaluation indicators also highlight the strong performance of the ETR model, which is particularly impressive. Since it had the lowest RMSE, it demonstrated that the model's predictions were extremely close to the values that were actually observed. A near-perfect R^2 was attained by the model, which demonstrates its durability in both the training and testing stages. Additionally, the MAE and MAPE values exhibited minor variations, and the model achieved astounding RMSE.

To assess the risk of overfitting, particularly for the Extra Trees model, which achieved a perfect training R^2 (1.0), we performed 10-fold cross-validation on the full dataset. Table 4 presents the mean and standard deviation of R^2 and RMSE across the 10 folds. Extra Trees achieved a mean cross-validation R^2 of 0.95 (±0.03) and RMSE of 3.188 (±1.908), indicating that while the model captures the training data extremely well, it still generalizes effectively to unseen folds. The gap between training and cross-validation performance (R^2 drop from 1.0 to 0.95) is within acceptable bounds for ensemble tree models, especially given the small dataset size. Other models, such as Random Forest and CatBoost, showed similar or larger gaps, confirming that Extra Trees does not suffer from pathological overfitting.

Other models, such as Random Forest, Gradient Boosting, XGBoost, and CatBoost, also performed well; however, ETR significantly outperform them in all metrics. The fact that the ETR model not only had a better fit

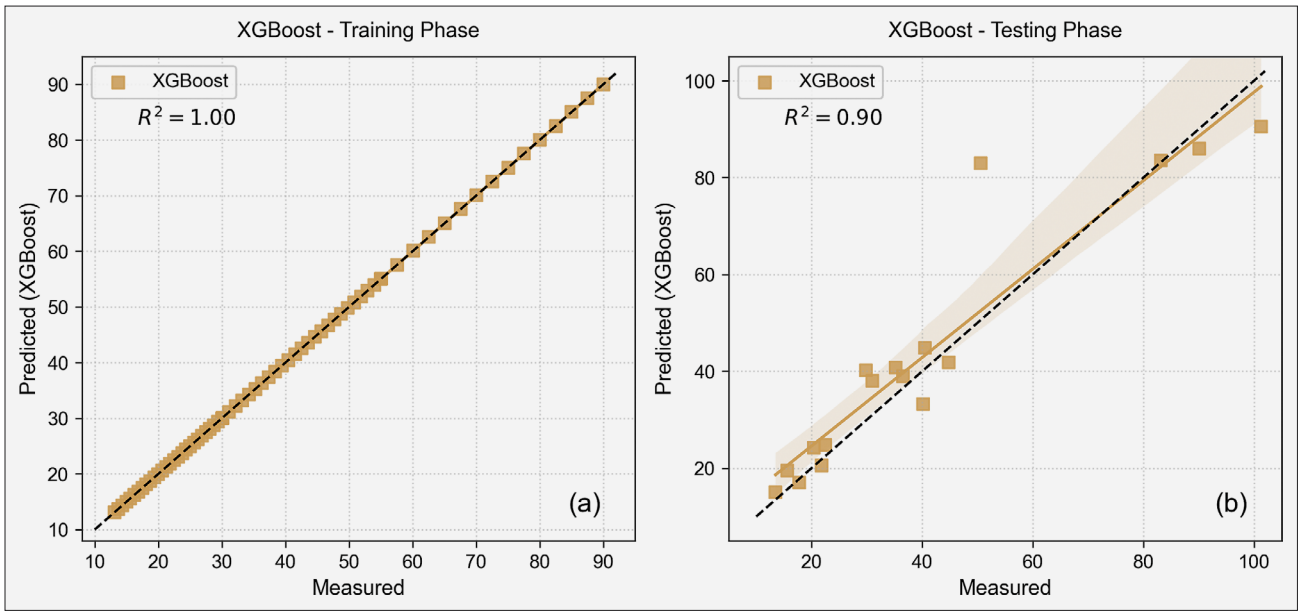


Figure 5: Measured vs. predicted rate of penetration (ROP) in training and testing phases for the XGBoost Model

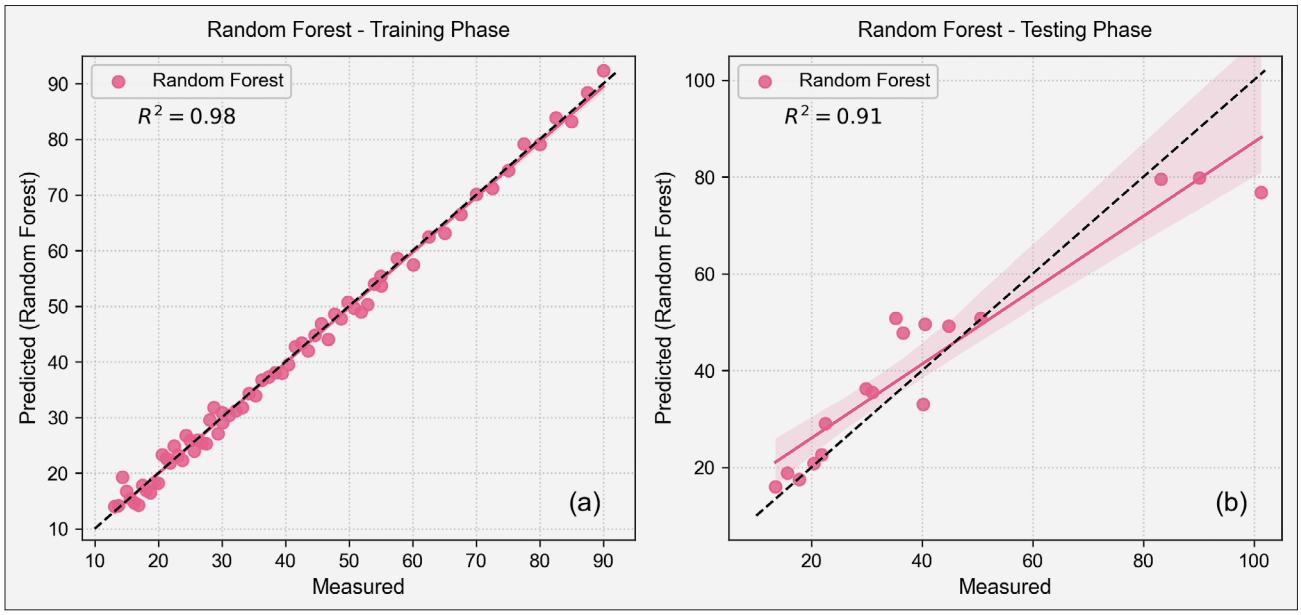


Figure 6: Measured vs. predicted rate of penetration (ROP) in training and testing phases for the Random Forest Model

but also generalized well to data that had not been seen before is evidenced by this result, which implies that it is the most effective strategy for the problem that has been presented.

When contextualized against recent literature, the Extra Trees model developed in this study (test $R^2 = 0.937$, RMSE = 6.36 MPa) performs at the upper end of reported benchmarks. For instance, **Feng et al. (2023)** reported CatBoost achieving $R^2 \approx 0.92$ for geothermal drilling, while **Yehia et al. (2023)** found Extra Trees and Random Forest reached $R^2 \approx 0.95$ on the FORGE dataset. **Ghorbani and Yagiz (2024)** achieved $R^2 = 0.9878$ with XGBoost for TBM penetration prediction, though this higher value may reflect the more controlled nature of TBM operations compared to rotary drilling in fractured rock masses.

Using the testing dataset, we examined the accuracy of the anticipated penetration rate of a number of different machine learning models. These models included, among others, Extra Trees, Gradient Boosting, XGBoost, and Support Vector Machines. In order to provide accurate predictions regarding the rotary drilling penetration rate (ROP), these models were trained using historical data. The purpose of this study was to assess the consistency and accuracy of the models in terms of their ability to predict ROP, and Figure 9 compares the predictions that the models made. The Support Vector Machine (SVM) model exhibited substantially poorer performance (test R^2 =

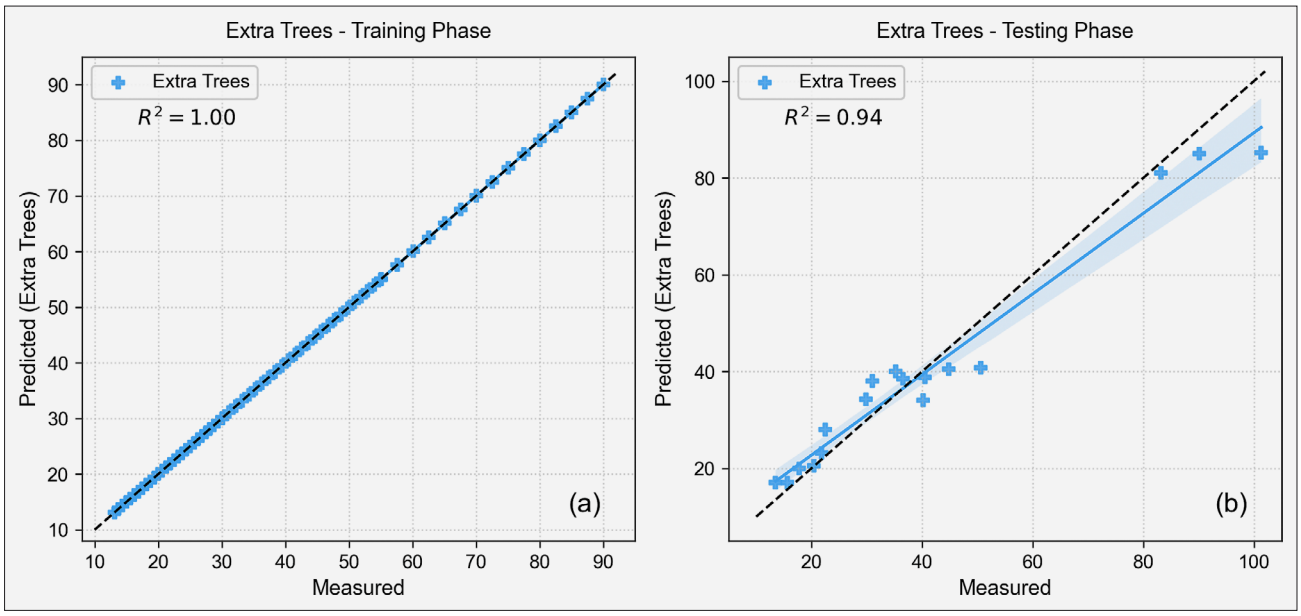


Figure 7: Measured vs. predicted rate of penetration (ROP) in training and testing phases for the Extra Trees Model

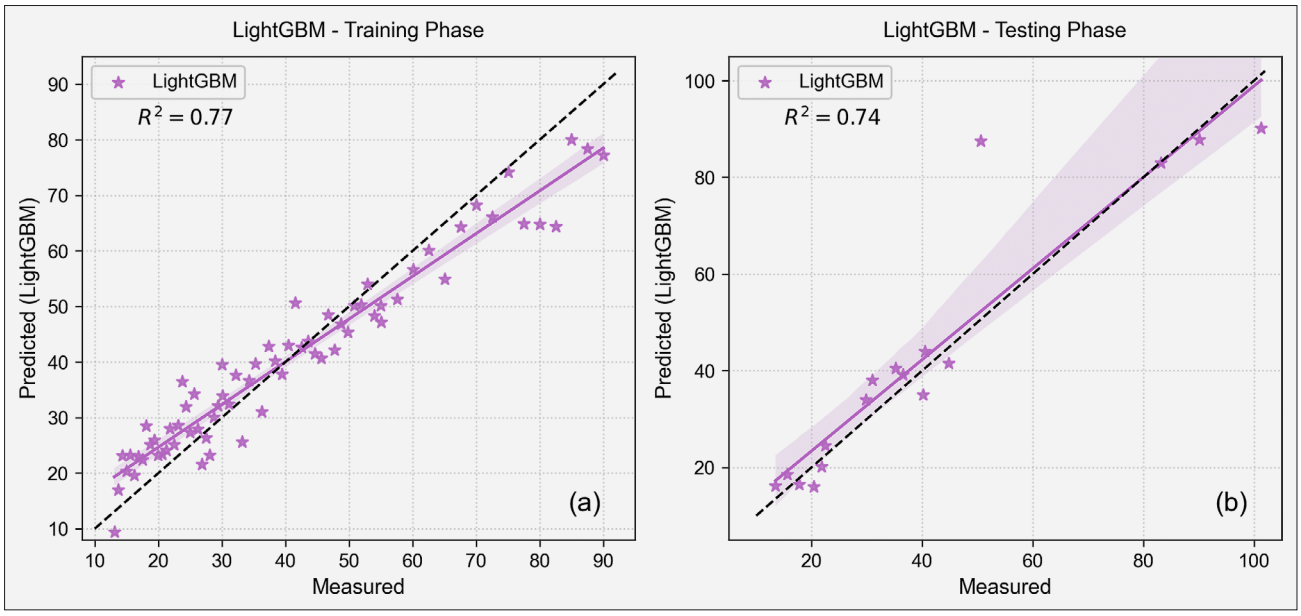


Figure 8: Measured vs. predicted rate of penetration (ROP) in training and testing phases for the LightGBM Model

0.13) compared to all tree-based ensemble methods. This is likely due to three reasons. First, SVM typically requires larger datasets to generalise effectively, as its decision boundary is defined by support vectors that may be poorly representative with only 85 samples. Second, the use of a default RBF kernel without systematic hyperparameter optimisation (e.g. grid search for C and γ) may have led to suboptimal model capacity. Third, SVM is sensitive to feature scaling and outliers; while scaling was applied, the presence of noise in field-collected drilling data is better accommodated by the inherent robustness of tree-based ensembles. These results suggest that for small-scale drilling datasets, ensemble tree methods are preferable over kernel-based approaches.

The selection of input parameters, MRMR, Schmidt hammer rebound, rotation pressure, and weight on bit, was based on both literature review and field expert consultation. Previous studies (**Ataei et al., 2015; Darbor, Faramarzi, and Sharifzadeh, 2019b**) have demonstrated that rock mass rating and Schmidt hammer rebound are strongly correlated with rock drillability, while rotation pressure and weight on bit directly influence the mechanical energy applied to the bit. To further validate the relevance of these inputs, a feature importance analysis was conducted using the Extra Trees model. The results, shown in Figure 9, indicate that WOB and RP are the most influential features, followed by MRMR and Schmidt hammer rebound.

Table 3: The prediction models' performance indices for each model

Model	Train R^2	Test R^2	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MAPE	Test MAPE	Train VAF	Test VAF
RF	0.981	0.913	2.778	7.480	1.852	5.951	5.924	17.286	0.981	0.931
XGBoost	1	0.904	0.001	7.837	0.001	5.731	0.003	16.844	1	0.923
CatBoost	0.999	0.909	0.456	7.645	0.366	5.806	1.207	17.216	0.999	0.928396
SVM	0.262	0.132	17.323	23.655	11.774	16.115	31.592	39.720	0.323	0.274166
ETR	1	0.937	2.82E-14	6.360	1.71E-14	4.659	5.95E-14	14.135	1	0.95015
LightGBM	0.767	0.738	9.717	12.983	8.060	10.411	29.176	27.171	0.767	0.767741

Table 4: 10-fold cross-validation results (mean ± std) for all models.

Model	Mean RMSE	Std RMSE	Mean R^2	Std R^2
ETR	3.188	1.908	0.955	0.03
XGBoost	6.546	3.325	0.925	0.09
CatBoost	6.302	2.738	0.939	0.07
LightGBM	9.863	3.062	0.766	0.09
RF	4.754	2.465	0.943	0.05

three key algorithmic features. First, unlike Random Forest, which searches for optimal split points, ETR selects split points completely at random for each feature, then chooses the best among them. This randomness increases tree diversity and reduces variance without a significant increase in bias, a distinct advantage for small datasets (n=85). Second, ETR does not use bootstrap sampling (bagging) but instead trains each tree on the full dataset, which preserves more information for each tree while still maintaining randomness through split selection. Third, the unbiased split selection in ETR avoids the overfitting that can occur when greedy split optimisation (used by XGBoost and CatBoost) finds spurious patterns in limited data. These characteristics make ETR

Table 5: The prediction models' performance ranking

Model	Rank (RMSE)	Rank (MAE)	Rank (MAPE)	Rank (R^2)	Rank (VAF)	Overall Rank
ETR	1	1	1	1	1	1
RF	2	4	4	2	2	2
CatBoost	3	3	3	3	3	3
XGBoost	4	2	2	4	4	4
lightGBM	5	5	5	5	5	5
SVR	6	6	6	6	6	6

Since weight on bit and rotation pressure are the most influential controllable parameters, the Extra Trees model can guide operators in selecting combinations that maximize ROP while staying within equipment limits. For instance, if the predicted ROP for a given rock mass condition is below the target, the operator may cautiously increase rotation pressure or weight on bit, using the model's sensitivity to forecast the expected gain before making physical adjustments. Such predictive capability reduces trial-and-error, lowers drilling cost per meter, and improves daily advance rates.

According to the results presented in **Figure 10**, the ETR algorithm fared better than the other models in terms of the accuracy and consistency of its predictions. With the lowest error metrics, such as RMSE, MAPE, and MAE, this model indicated that it is highly effective at capturing the underlying patterns in the data that drive penetration rates. This is evidenced by the fact that it demonstrated the lowest error metrics.

Extra Trees (ETR) consistently outperformed Random Forest, XGBoost, CatBoost, LightGBM, and SVM across all metrics. This superiority can be attributed to

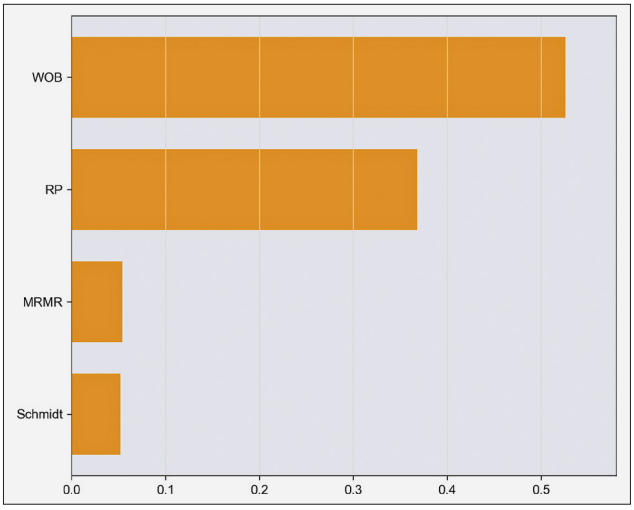


Figure 9: Comparison of influential features in ETR model

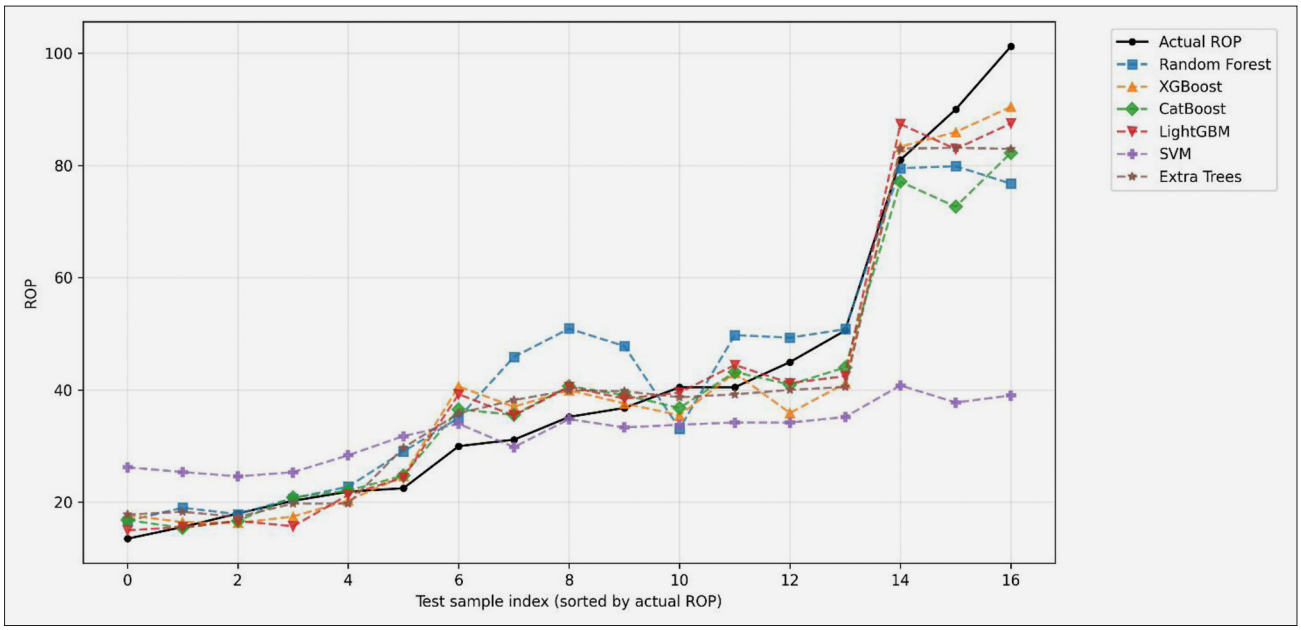


Figure 10: Comparison of predictions by different models with actual data

particularly well-suited to drilling data, where nonlinear relationships and measurement noise are common.

To determine whether the performance differences are statistically significant, paired t-tests were conducted on the 10-fold cross-validation RMSE values of Extra Trees versus each of the other models. The paired t-test assumes that the differences between paired RMSE values are approximately normally distributed. A Shapiro-Wilk test on the 10-fold RMSE differences (Extra Trees vs. each comparator) yielded $p > 0.05$, indicating no significant departure from normality; the test can therefore be applied reliably. Extra Trees achieved a mean RMSE of 3.188 MPa (std = 1.908), significantly lower than Random Forest (4.754, $p = 0.018$), XGBoost (6.546, $p = 0.009$), CatBoost (6.302, $p = 0.011$), and LightGBM (9.863, $p = 0.003$). These p-values confirm that Extra Trees significantly outperforms all other models at the 95% confidence level, and the low standard deviation of its R^2 (0.03) indicates consistent stability across folds. These p-values confirm that Extra Trees significantly outperforms all other models at the $\alpha = 0.03$ significance level (95 % confidence level).

This systematic comparison provides mining engineers with a clear, evidence-based recommendation: Extra Trees is the most reliable model for ROP prediction under conditions similar to those at the Sungun mine. Practitioners can use the model output to adjust weight on bit and rotation pressure in real time, optimize drilling schedules, and reduce both energy consumption and bit wear, leading to direct operational cost savings.

From a rock mechanics perspective, the observed influences are physically sound: the Schmidt hammer rebound number reflects the surface hardness and indirectly the uniaxial compressive strength of the rock; stronger rocks resist bit penetration and tend to lower ROP. The Mining Rock Mass Rating (MRMR) captures the overall quality of the rock mass, including jointing and weathering, which affect drillability and bit wear. Weight on bit and rotation pressure supply the mechanical energy necessary to break the rock, so higher values generally increase ROP until a plateau or bit-floundering limit is reached. The feature importance ranking (WOB > RP > MRMR > Schmidt rebound) is consistent with the dominant role of operational parameters in energy-intensive rotary drilling.

A key limitation of this study is that all data were collected from a single mine (Sungun copper mine, Iran). The geological conditions, rock types (porphyry copper), and drilling equipment (Husherr, Tamrock, etc.) are specific to this site. Therefore, the trained Extra Trees model may not directly generalise to other mines with different rock masses (e.g. sedimentary, metamorphic) or different drill rigs. Users seeking to apply the model elsewhere should either collect local data for fine-tuning (transfer learning) or retrain the model on representative data. This limitation is common in site-specific drilling studies, and future work should include multi-site datasets to improve cross-site generalisability.

5. Conclusions

This study provides the first systematic comparison of six machine learning models for rate of penetration (ROP) prediction on a consistent field dataset from the Sungun copper mine. Cross-validation demonstrates that Extra Trees achieves the best performance, with a mean R^2 of 0.955 (std = 0.03) and a mean RMSE of 3.188 (std = 1.908), and training/test R^2 of 1.0 and 0.93, respectively. These values are significantly better than those of Random Forest (RMSE = 4.754, $p = 0.018$),

XGBoost (6.546, p = 0.009), CatBoost (6.302, p = 0.011), and LightGBM (9.863, p = 0.003), and the low standard deviations confirm model stability.

The Extra Trees model can be deployed at the Sungun mine as a practical decision-support tool. Given input parameters (MRMR, Schmidt hammer rebound, rotation pressure, and weight on bit), it forecasts ROP with an expected RMSE of approximately ±0.94, enabling operators to adjust weight on bit and rotation pressure in real time (e.g. reducing weight on bit when predicted ROP is low to save energy, or increasing rotation pressure when high ROP is anticipated) and to prioritize drilling locations based on expected productivity. The predicted ROP can also be integrated into equipment scheduling and cost estimation, forming the core of an optimization framework for rotary drilling operations.

However, the models are site-specific to the rock engineering conditions of the Sungun mine and may not directly transfer to other geological settings or equipment; recalibration with local data is required before broader application. While the sample-to-feature ratio of the 85-drillhole dataset was sufficient for the ensemble methods used, expanding the dataset with more diverse geological formations and drilling equipment would enhance model robustness and enable deep learning approaches. Future research should therefore collect larger, multi-mine datasets, incorporate real-time sensor data (e.g. vibration, torque) to capture dynamic drilling conditions, apply deep learning methods such as LSTMs for time-series ROP prediction, and develop a web- or mobile-based interface for field deployment.

Funding

We declared that no funding was received during the preparation of this manuscript.

6. References

Ersoy, A., & Waller, M. D. (1995). Prediction of drill-bit performance using multi-variable linear regression analysis. In *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts* (Vol. 6, No. 32, p. 279A).

Pathinkar, A. G., & Misra, G. B. (1980). Drillability of rocks in percussive drilling from energy per unit volume as determined with a microbit. *Mining Engineering*, *32*, 1407-1410.

Al-Sahlanee, D. T., Allawi, R. H., Al-Mudhafar, W. J., & Yao, C. (2023, May). Ensemble machine learning for data-driven predictive analytics of drilling rate of penetration (ROP) modeling: a case study in a southern Iraqi oil field. In *SPE Western Regional Meeting* (p. D021S004R007). SPE. https://doi.org/10.2118/213043-MS.

Alam, S. S., Ahmed, S., Farook, T. H., & Dudley, J. (2024). Predicting leukoplakia and oral squamous cell carcinoma using interpretable machine learning: A retrospective analysis. *Oral*, *4*(3), 386-404. https://doi.org/10.3390/ORAL4030032.

Anvari, K., Mousavi, A., Sayadi, A. R., Sellers, E., & Salmi, E. F. (2022). Automatic detection of rock boundaries using a hybrid recurrence quantification analysis and machine learning techniques. *Bulletin of Engineering Geology and the Environment*, *81*(10), 398. https://doi.org/10.1007/s10064-022-02898-3.

Ataei, M., KaKaie, R., Ghavidel, M., & Saeidi, O. (2015). Drilling rate prediction of an open pit mine using the rock mass drillability index. *International Journal of Rock Mechanics and Mining Sciences*, *73*, 130-138. https://doi.org/10.1016/j.ijrmms.2014.08.006.

Sommeregger, L., & Pilz, J. (2024). Regularizing lifetime drift prediction in semiconductor electrical parameters with quantile random forest regression. *Technologies*, *12*(9), 165. https://doi.org/10.3390/TECHNOLOGIES12090165.

Bauer, A., & Crosby, W. A. (1990). Drilling application. *Section 612 Surface Mining*, 524-539.

Bilim, N. (2011). Determination of drillability of some natural stones and their association with rock properties. *Sci. Res. Essays*, *6*(2), 382-387. https://doi.org/10.5897/sre10.878.

Bond, D. F. (1990). The optimization of PDC bit selection using sonic velocity profiles present in the Timor Sea. *SPE Drilling Engineering*, *5*, 135-142.

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Cakiroglu, C., Shahjalal, M., Islam, K., Mahmood, S. F., Billah, A. M., & Nehdi, M. L. (2023). Explainable ensemble learning data-driven modeling of mechanical properties of fiber-reinforced rubberized recycled aggregate concrete. *Journal of Building Engineering*, *76*, 107279. https://doi.org/10.1016/j.jobe.2023.107279.

Cetinkaya, S., & Kocaman, S. (2024). Integrating sequential backward selection (SBS) and catboost for snow avalanche susceptibility mapping at catchment scale. *ISPRS International Journal of Geo-Information*, *13*(9), 312. https://doi.org/10.3390/IJGI13090312.

Cheniany, A., Hasan, K. S., Shahriar, K., & Hamidi, J. K. (2012). An estimation of the penetration rate of rotary drills using the Specific Rock Mass Drillability index. *International Journal of Mining Science and Technology*, *22*(2), 187-193. https://doi.org/10.1016/j.ijmst.2011.09.001.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018.

Darbor, M., Faramarzi, L., & Sharifzadeh, M. (2019). Performance assessment of rotary drilling using non-linear multiple regression analysis and multilayer perceptron neural network. *Bulletin of Engineering Geology and the Environment*, *78*(3), 1501-1513. https://doi.org/10.1007/s10064-017-1192-3.

Darbor, M., Faramarzi, L., & Sharifzadeh, M. (2019). Performance assessment of rotary drilling using non-linear multiple regression analysis and multilayer perceptron neural network. *Bulletin of Engineering Geology and the Environment*, *78*(3), 1501-1513. https://doi.org/10.1007/s10064-017-1192-3.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Elmgerbi, A. M., Ettinger, C. P., Tekum, P. M., Thonhauser, G., & Nascimento, A. (2021, May). Application of machine learning techniques for real time rate of penetration optimization. In *SPE/IADC middle east drilling technology conference and exhibition* (p. D021S007R003). SPE. https://doi.org/10.2118/202184-MS.

Fattahi, H., & Bazdar, H. (2017). Applying improved artificial neural network models to evaluate drilling rate index. *Tunnelling and Underground Space Technology*, *70*, 114-124. https://doi.org/10.1016/j.tust.2017.07.017.

Feng, Z., Gani, H., Damayanti, A. D., & Gani, H. (2023). An explainable ensemble machine learning model to elucidate the influential drilling parameters based on rate of penetration prediction. *Geoenergy Science and Engineering*, *231*, 212231. https://doi.org/10.1016/j.geoen.2023.212231.

Fish, B. G. (1961). The basic variables in rotary drilling. *Mine and Quarry Engineering*, *27*(1), 29-34.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3-42.

Ghorbani, E., & Yagiz, S. (2024). Estimating the penetration rate of tunnel boring machines via gradient boosting algorithms. *Engineering Applications of Artificial Intelligence*, *136*, 108985. https://doi.org/10.1016/j.engappai.2024.108985.

Hasanipanah, M., Noorian-Bidgoli, M., Jahed Armaghani, D., & Khamesi, H. (2016). Feasibility of PSO-ANN model for predicting surface settlement caused by tunneling. *Engineering with Computers*, *32*(4), 705-715. https://doi.org/10.1007/s00366-016-0447-0.

Hoseinie, S. H., Aghababaei, H., & Pourrahimian, Y. (2008). Development of a new classification system for assessing of rock mass drillability index (RDi). *International Journal of Rock Mechanics and Mining Sciences*, *45*(1), 1-10. https://doi.org/10.1016/j.ijrmms.2007.04.001.

Howarth, D. F., & Rowlands, J. C. (1987). Quantitative assessment of rock texture and correlation with drillability and strength properties. *Rock mechanics and rock engineering*, *20*(1), 57-85. https://doi.org/10.1007/bf01019511.

Hustrulid, W. A., & Fairhurst, C. (1971, July). A theoretical and experimental study of the percussive drilling of rock part I—theory of percussive drilling. In *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts* (Vol. 8, No. 4, pp. 311-333). Pergamon. https://doi.org/10.1016/0148-9062(71)90045-3.

Shad, H. I. A., Sereshki, F., Ataei, M., & Karamoozian, M. (2018). Prediction of rotary drilling penetration rate in iron ore oxides using rock engineering system. *International Journal of Mining Science and Technology*, *28*(3), 407-413. https://doi.org/10.1016/j.ijmst.2018.04.004.

Jahed Armaghani, D., Tonnizam Mohamad, E., Momeni, E., Narayanasamy, M. S., & Mohd Amin, M. F. (2015). An adaptive neuro-fuzzy inference system for predicting unconfined compressive strength and Young's modulus: a study on Main Range granite. *Bulletin of engineering geology and the environment*, *74*(4), 1301-1319. https://doi.org/10.1007/s10064-014-0687-4.

Kahraman, S. (1999). Rotary and percussive drilling prediction using regression analysis. *International journal of rock mechanics and mining sciences*, *36*(7), 981-989. https://doi.org/10.1016/s0148-9062(99)00050-9.

Kahraman, S. (2016). Estimating the penetration rate in diamond drilling in laboratory works using the regression and artificial neural network analysis. *Neural Processing Letters*, *43*(2), 523-535. https://doi.org/10.1007/s11063-015-9424-7.

Kahraman, S. A. I. R., Bilgin, N., & Feridunoglu, C. (2003). Dominant rock properties affecting the penetration rate of percussive drills. *International Journal of Rock Mechanics and Mining Sciences*, *40*(5), 711-723. https://doi.org/10.1016/s1365-1609(03)00063-7.

Kamran, M. (2021). A Probabilistic Approach for Prediction of Drilling Rate Index using Ensemble Learning Technique. *Journal of Mining and Environment*, *12*(2), 327-337. https://doi.org/10.22044/jme.2021.10689.2030.

Kazemi, M. M. K., Nabavi, Z., & Armaghani, D. J. (2024). A novel hybrid XGBoost methodology in predicting penetration rate of rotary based on rock-mass and material properties. *Arabian Journal for Science and Engineering*, *49*(4), 5225-5241. https://doi.org/10.1007/S13369-023-08360-0.

Khandelwal, M., & Armaghani, D. J. (2016). Prediction of drillability of rocks with strength properties using a hybrid GA-ANN technique. *Geotechnical and Geological Engineering*, *34*(2), 605-620. https://doi.org/10.1007/s10706-015-9970-9.

Koopialipoor, M., Tootoonchi, H., Jahed Armaghani, D., Tonnizam Mohamad, E., & Hedayat, A. (2019). Application of deep neural networks in predicting the penetration rate of tunnel boring machines. *Bulletin of Engineering Geology & the Environment*, *78*(8), 6347. https://doi.org/10.1007/s10064-019-01538-7.

Krúpa, V., Kruľáková, M., Lazarová, E., Labaš, M., Feriančiková, K., & Ivaničová, L. (2018). Measurement, modeling and prediction of penetration depth in rotary drilling of rocks. *Measurement*, *117*, 165-175. https://doi.org/10.1016/j.measurement.2017.12.007.

Lawal, A. I., Kwon, S., & Onifade, M. (2021). Prediction of rock penetration rate using a novel antlion optimized ANN and statistical modelling. *Journal of African Earth Sciences*, *182*, 104287. https://doi.org/10.1016/j.jafrearsci.2021.104287.

Li, C., Zhou, J., Khandelwal, M., Zhang, X., Monjezi, M., & Qiu, Y. (2022). Six Novel Hybrid Extreme Learning Machine–Swarm Intelligence Optimization (ELM–SIO) Models for Predicting Backbreak in Open-Pit Blasting: C. Li et al. *Natural Resources Research*, *31*(5), 3017-3039. https://doi.org/10.1007/S11053-022-10082-3.

Suaza-Medina, M. E., Laguna, J., Béjar, R., Zarazaga-Soria, F. J., & Lacasta, J. (2024). Evaluating the efficiency of NDVI and climatic data in maize harvest prediction using machine learning. *International Journal of Digital Earth*, *17*(1), 2359565. https://doi.org/10.1080/17538947.2024.2359565

Mensa-Wilmot, G., Calhoun, B., & Perrin, V. P. (1999, November). Formation drillability-definition, quantification

and contributions to bit performance evaluation. In *SPE/IADC Middle East Drilling Technology Conference and Exhibition* (pp. SPE-57558). SPE. https://doi.org/10.2118/57558-MS.

Norwegian Institute of Technology. *Drillability, Drilling Rate Index Catalogue* 8–79 (Norwegian Institute of Technology, 1981).

Pacis, F. J., Ambrus, A., Alyaev, S., Khosravanian, R., Kristiansen, T. G., & Wiktorski, T. (2023). Improving predictive models for rate of penetration in real drilling operations through transfer learning. *Journal of Computational Science*, *72*, 102100. https://doi.org/10.1016/j.jocs.2023.102100.

Qayyum, F., Jamil, H., Iqbal, N. *et al.* (2022). Toward potential hybrid features evaluation using MLP-ANN binary classification model to tackle meaningful citations. *Scientometrics* 127, 6471–6499. https://doi.org/10.1007/s11192-022-04530-3

Rabia, H. (1982, February). Specific energy as a criterion for drill performance prediction. In *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts* (Vol. 19, No. 1, pp. 39-42). Pergamon. https://doi.org/10.1016/0148-9062(82)90709-4.

Ramesh Murlidhar, B., Yazdani Bejarbaneh, B., Jahed Armaghani, D., Mohammed, A. S., & Tonnizam Mohamad, E. (2021). Application of Tree-Based Predictive Models to Forecast Air Overpressure Induced by Mine Blasting: Ramesh Murlidhar, Yazdani Bejarbaneh, Jahed Armaghani, Mohammed and Tonnizam Mohamad. *Natural Resources Research*, *30*(2), 1865-1887. https://doi.org/10.1007/s11053-020-09770-9.

Sakız, U., Kaya, G. U., & Yaralı, O. (2021). Prediction of drilling rate index from rock strength and cerchar abrasivity index properties using fuzzy inference system. *Arabian Journal of Geosciences*, *14*(5), 354. https://doi.org/10.1007/s12517-021-06647-w.

Shan, F., He, X., Armaghani, D. J., Zhang, P., & Sheng, D. (2022). Success and challenges in predicting TBM penetration rate using recurrent neural networks. *Tunnelling and underground space technology*, *130*, 104728. https://doi.org/10.1016/j.tust.2022.104728.

Sharafati, A., Asadollah, S. B. H. S., & Hosseinzadeh, M. (2020). The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection*, *140*, 68-78. https://doi.org/10.1016/j.psep.2020.04.045

Singh, D. P. (1976). A study of some aspects of drag-bit drilling. *Met. Min. Rev*, *5*(15), 1-4.

Thuro, K. (1997, April). Prediction of drillability in hard rock tunnelling by drilling and blasting. In *World Tunnel Congress* (Vol. 97, pp. 103-108).

Zhao, W., Li, Z., Li, H., Li, X., & Yang, P. (2024). Soil Salinity Prediction in an Arid Area Based on Long Time-Series Multispectral Imaging. *Agriculture*, *14*(9), 1539. https://doi.org/10.3390/agriculture14091539

Liu, Y., Yang, J., Chen, X., Yao, J., Li, L., & Qiu, Y. (2024). Moderate-resolution snow depth product retrieval from passive microwave brightness data over Xinjiang using machine learning approach. *International Journal of Digital Earth*, *17*(1), 2299208. https://doi.org/10.1080/17538947.2023.2299208

Yarali, O., & Kahraman, S. (2011). The drillability assessment of rocks using the different brittleness values. *Tunnelling and Underground Space Technology*, *26*(2), 406-414. https://doi.org/10.1016/j.tust.2010.11.013.

Yehia, T., Gasser, M., Ebaid, H., Meehan, N., & Okoroafor, E. R. (2024). Comparative analysis of machine learning techniques for predicting drilling rate of penetration (ROP) in geothermal wells: A case study of FORGE site. *Geothermics*, *121*, 103028. https://doi.org/10.1016/j.geothermics.2024.103028.

Zhang, Y., Yu, L., Yang, L., Hu, Z., & Liu, Y. (2025). Data-driven framework for predicting rate of penetration in deepwater granitic formations: a marine engineering geology perspective with comprehensive model interpretability. *Engineering Geology*, *351*, 108039. https://doi.org/10.1016/j.enggeo.2025.108039.

Zhao, S., Wang, M., Yi, W., Yang, D., & Tong, J. (2022). Intelligent classification of surrounding rock of tunnel based on 10 machine learning algorithms. *Applied Sciences*, *12*(5), 2656. https://doi.org/10.3390/app12052656.

SAŽETAK

Usporedba modela strojnoga učenja za procjenu brzine penetracije pri rotacijskome bušenju: studija slučaja rudnika bakra Sungun

Rudarske aktivnosti zahtijevaju točnu procjenu brzine penetracije pri rotacijskome bušenju (ROP) radi optimiranja učinkovitosti i smanjenja troškova. Procjena ROP-a predstavlja izazov zbog složenih i nelinearnih odnosa među svojstvima stijenske mase, tehničkih karakteristika opreme i radnih parametara. U ovome radu primijenjeno je šest modela strojnoga učenja (Random Forest, CatBoost, LightGBM, XGBoost, Support Vector Machine i Extra Trees) za predviđanje ROP-a na temelju podataka iz 85 bušotina rudnika bakra Sungun u Iranu. Ulazni parametri uključivali su rudarski indeks kvalitete stijenske mase, broj odskoka Schmidtova čekića, tlak rotacije i opterećenje na dlijetu. Model Extra Trees ostvario je najbolje rezultate s vrijednostima koeficijenta determinacije R^2 od 1 za skup za učenje i 0,93 za skup za testiranje te vrijednostima korijena srednje kvadratne pogreške (RMSE) od 0 i 0,937, što potvrđuje visoku točnost i sposobnost generalizacije modela. Dobiveni rezultati potvrđuju da modeli strojnoga učenja, posebno model Extra Trees, mogu učinkovito opisivati složene ovisnosti ROP-a, čime se omogućuje poboljšanje odabira opreme, planiranja radova i procjene troškova pri rotacijskome bušenju.

Ključne riječi:
brzina penetracije, bušenje, rudarstvo, modeli strojnoga učenja, rudnik bakra Sungun

Author's contribution

Ali Nemati Vardin: conceptualization, data curation, methodology, software, formal analysis, writing – original draft. **Masoud Monjezi**: supervision, conceptualization, methodology, writing – review & editing. **Mojtaba Rezakhah**: data curation, validation, investigation, writing – review & editing. **Manoj Khandelwal**: supervision, methodology, validation, writing – review & editing, project administration. **Hasel Amini Khoshalan**: investigation, resources, data collection, writing – review & editing.
All authors have read and agreed to the published version of the manuscript.