The 2023 World Congress on
Advances in Structural Engineering and Mechanics (ASEM23)
GECE, Seoul, Korea, August 16-18, 2023

Estimation of NTNU/SINTEF drillability test indices using soft computing techniques based on rock properties

* Tae Young Ko [1)]

[1)] *Department of Energy and Resources Engineering, Kangwon National University, Chuncheon 24341, Korea*
[1)] tyko@kangwon.ac.kr

ABSTRACT

This study presents an innovative approach to estimating the drillability rate index (DRI) for rocks. Utilizing a dataset of 88 instances, the research employs a diverse range of models, including traditional linear regression, machine learning-based regression methods such as Ridge, Lasso, ElasticNet, and a unique application of symbolic regression. The symbolic regression's strength lies in its ability to produce human-readable mathematical expressions, enabling a transparent understanding of the underlying relationships between rock properties such as uniaxial compressive strength (UCS), Brazilian tensile strength (BTS), equivalent quartz content (EQC), and Cerchar abrasivity index (CAI). Performance evaluation using mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2) reveals symbolic regression's superior predictive performance. The results offer valuable insights into rock drillability, contributing to more accurate and efficient tunneling operations, and highlight the potential of symbolic regression as a robust, interpretable modeling technique.

1. INTRODUCTION

Tunneling operations represent an essential area in the field of civil and mining engineering. An important aspect of these operations is the ability to accurately evaluate the properties of rock and soil materials to ensure optimal productivity and cost-effectiveness. Over time, the drillability test indices developed by the Norwegian University of Science and Technology (NTNU) and the Foundation for Scientific and Industrial Research (SINTEF) have gained international recognition as an invaluable tool in this aspect. These indices offer standardized measurements to assess vital factors such as cutter wear, penetration rates, and time and cost estimates.
However, as beneficial as the NTNU/SINTEF drillability test indices have proven to be, there are certain inherent limitations that should be considered. These limitations primarily involve the considerable complexity that is found inherent in the testing procedures and the restrictions in locations where these experimental tests can be

[1)] Professor

feasibly performed, which narrows down their applicability to certain specific conditions. It is within this context that the present study seeks to advance the field by proposing an innovative model to estimate the drilling rate index (DRI) of the NTNU/SINTEF drillability test indices using various rock properties.

The model leverages parameters such as the uniaxial compressive strength (UCS), Brazilian tensile strength (BTS), equivalent quartz content (EQC), and Cerchar abrasivity index (CAI) to estimate DRI. In an attempt to enhance the accuracy of predictions, the proposed model employs symbolic regression, a sophisticated machine learning method. Unlike traditional machine learning regression models, symbolic regression aims to formulate DRI as mathematical equations based on rock properties, thereby potentially offering clear insights into the underlying relationships among variables.

It is hoped that this study will not only address the aforementioned limitations but also introduce a robust and precise predictive model. Through this, we aim to make significant strides towards more accurate and efficient tunneling operations, thereby contributing to the broader goal of improved operational efficiency in civil and mining engineering sectors.

2. DATA COLLECTION AND EXPLORATORY DATA ANALYSIS

The key to developing an effective model for estimating DRI using rock properties is the compilation of a comprehensive dataset. This study employed a thorough literature review as the primary means of data acquisition. Due to the scarcity of test results, the dataset comprised a limited number of data points; specifically, we gathered data of 88 instances for DRI. The rock properties considered in this study included UCS, BTS, EQC, and CAI (Aligholi et al. 2017; Eide 2014; Macias et al. 2015; Majeed et al. 2020).

After the data collection, we constructed a database incorporating not only the rock properties and DRI but also the rock types. The database formation laid the foundation for exploratory data analysis (EDA), which facilitated a deeper understanding of the data characteristics and their relationships.

The results of the EDA, detailed in the following section, offer insightful observations about the distribution, correlation, and potential anomalies in the dataset. This comprehensive understanding of the data aids in developing a robust model and provides a basis for further analysis. The dataset and its properties form the backbone of the proposed symbolic regression model for estimating DRI, highlighting the importance of thorough data collection and preliminary analysis.

The dataset at hand comprises a range of geomechanical properties linked with varying rock types. The features include the CAI, EQC, UCS, BTS, and DRI. The rock types are encoded into three binary categories, signifying igneous, metamorphic, and sedimentary rocks. Initial exploration of the data revealed no missing values, ensuring the completeness of the dataset. The dataset predominantly consists of igneous rocks, which account for 65 instances, followed by metamorphic rocks (14 instances), and sedimentary rocks (9 instances) (Fig.1). The correlation matrix heatmap shows a moderate correlation between DRI and the rest of the variables (Fig. 2). DRI shows a negative correlation with CAI, UCS, and BTS. This indicates that as these variables increase, DRI tends to decrease. DRI shows a slight positive correlation with EQC. The

histograms of the continuous variables demonstrate diverse distributions. While none of the variables exhibit a perfect normal distribution, they generally tend to concentrate around their central values (Fig.3).

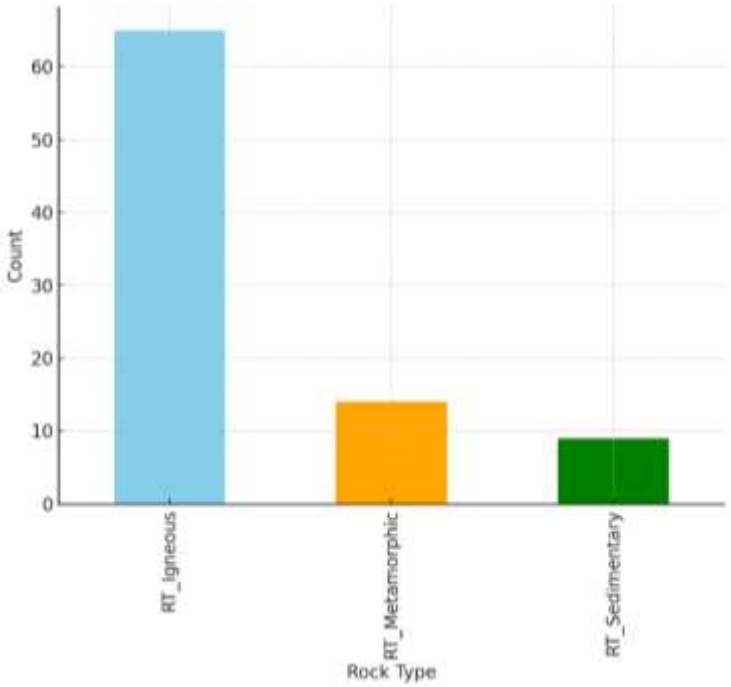

Fig. 1 Count of each rock type for DRI database

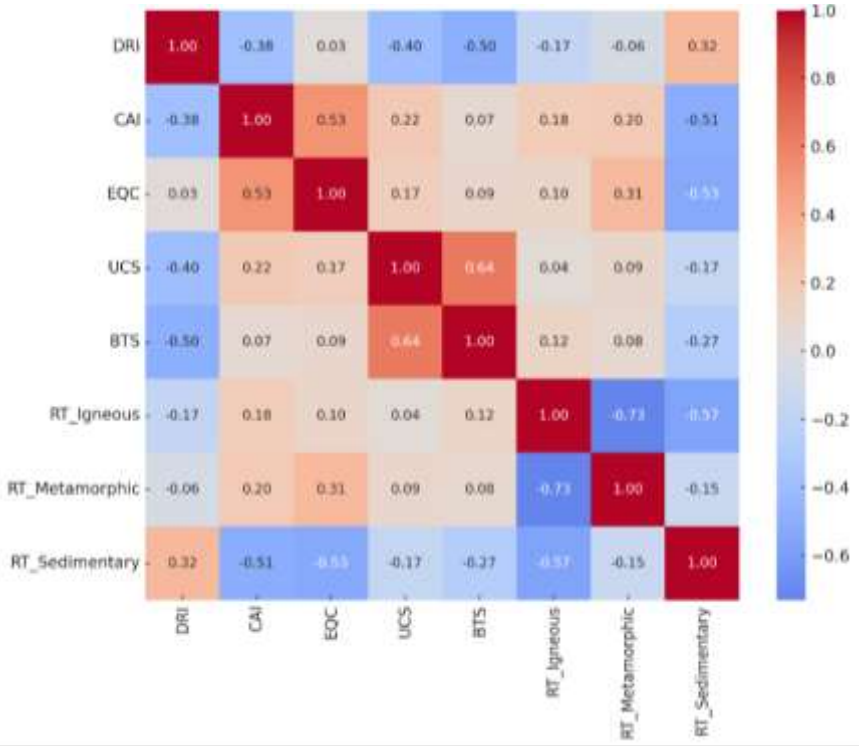

Fig. 2 Correlation matrix for DRI database

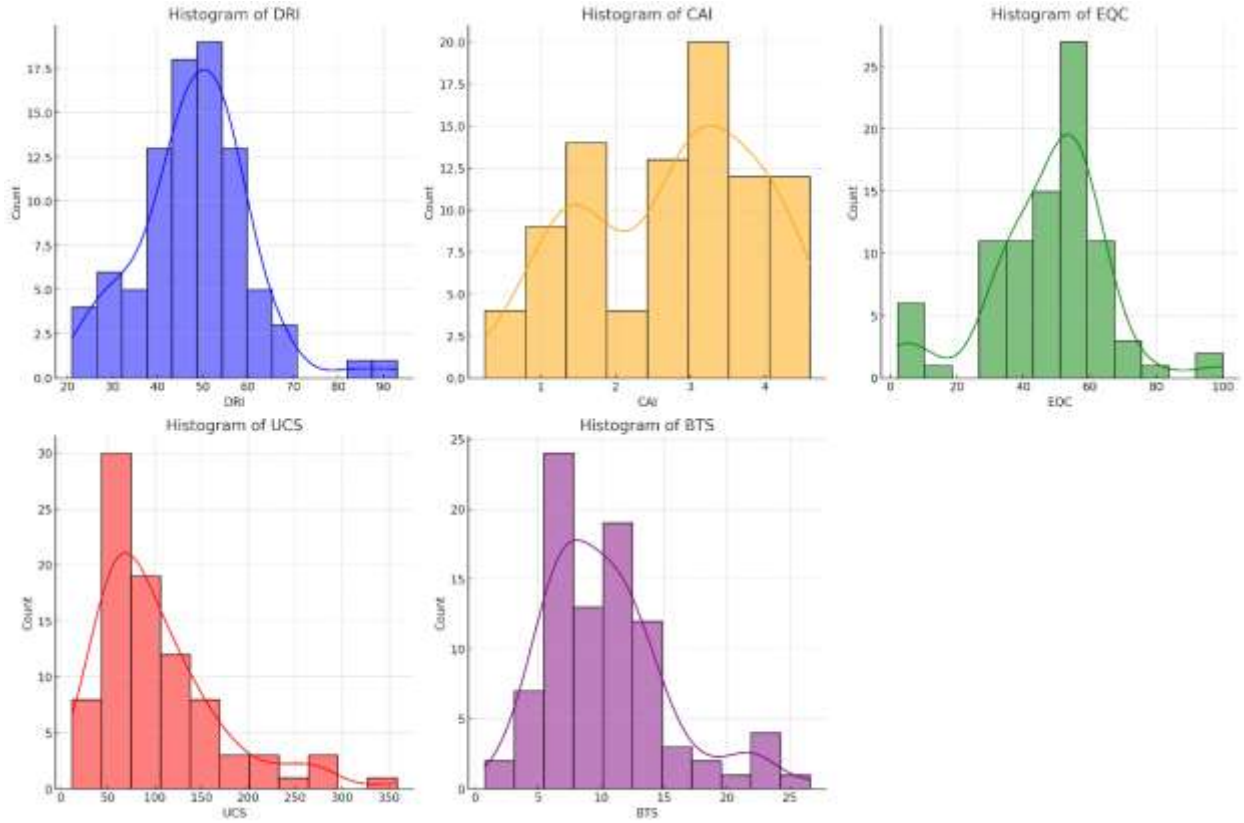

Fig. 3 Histograms of the selected variables for DRI database

The majority of the DRI values are concentrated around 40 to 60, with a slight right skew, indicating a few instances of high DRI values. The CAI values are spread over a broad range, with a notable peak around 2.5 to 3.5. The EQC values appear to be normally distributed around the mean value, with a slight skew towards higher values. The BTS values are skewed towards lower values, with a few instances of higher BTS values.

3. MACHINE LEARNING BASED REGRESSION ANALYSIS

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to find the best-fitting line through the data points that minimizes the sum of the squared errors, and to use this line to predict future values. Linear regression is a powerful and commonly used statistical method that is simple to understand and interpret, making it great for practical applications. It's also fast to compute and can effectively predict outcomes when relationships in the data are linear. However, it does come with some limitations. For instance, it makes several key assumptions, such as linearity and homoscedasticity, and if these assumptions are violated, the results may not be reliable. It can also be sensitive to outliers, which can significantly skew the model's predictions. Overfitting, a situation where the model performs well on training data but poorly on unseen data, is another potential issue.

Lastly, linear regression is limited to modeling linear relationships, and if the relationship in the data is not linear, linear regression may not perform well. In such cases, non-linear models may be more suitable.

In this study, we employed a linear regression analysis to predict the DRI based on the following independent variables: CAI, EQC, UCS, BTS, and Rock Types. The model was trained and evaluated on a dataset that was split into a training set (70% of the data) and a test set (30% of the data). The performance of the model was evaluated using the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2) metrics.

The performance metrics for the model on the training set were as follows: MSE = 78.21, RMSE = 8.84, R^2 = 0.45. The performance metrics for the model on the test set were as follows: MSE = 80.82, RMSE = 8.99, R^2 = 0.48.

These results suggest that the model explains approximately 45% and 48% of the variance in the DRI in the training and test sets, respectively. Thus, while the model has a moderate level of predictive power, there is still substantial unexplained variance, indicating that factors not included in the model may also be impacting the DRI.

A scatter plot of the actual versus predicted DRI values for both the training and test sets is showed in Fig. 4. The model's predictions are generally in line with the actual values, although there is some scatter around the line of perfect prediction. The equation of the line of best fit in the scatter plot is given by:

$$\begin{aligned} DRI = {} & 56.23 - 3.49CAI + 0.32EQC - 0.0094UCS - 0.97BTS \\ & -3.39RT\_Igneous - 3.30RT\_Metamorphic + 6.69RT\_Sedimentary \end{aligned} \quad (1)$$

In the regression model, the rock type variable is represented as three separate binary (0 or 1) variables: RT_Igneous, RT_Metamorphic, and RT_Sedimentary. If a rock is of the Igneous type, then RT_Igneous would be 1 and RT_Metamorphic and RT_Sedimentary would be 0.

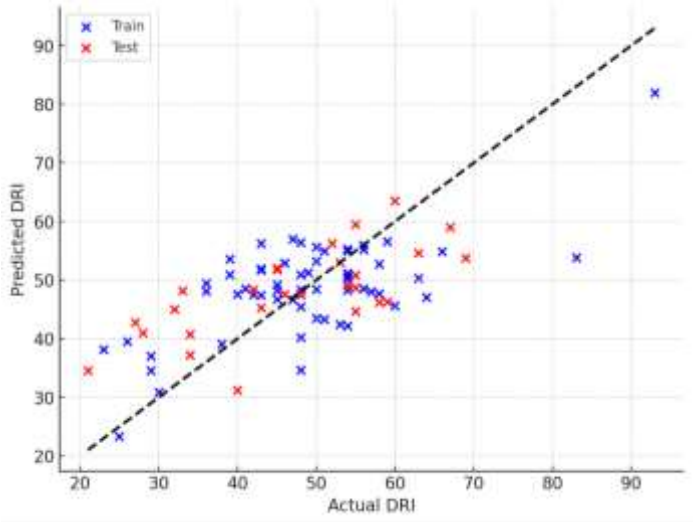

Fig. 4. A scatter plot of the actual versus predicted DRI values

Linear regression is a traditional statistical method that models the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent and dependent variables, and its goal is to find a linear function that predicts the dependent variable values as accurately as possible. While linear regression is simple and interpretable, it may not perform well when there are nonlinear relationships or interactions between variables, or when the variables are highly correlated, a condition known as multicollinearity.

To address these limitations, machine learning-based regression models such as Ridge, Lasso, Elastic Net, Support Vector, Random Forest, Gradient Boosting, AdaBoost, Extra Trees, and XGBoost have been developed. These models can handle more complex relationships and interactions between variables. Ridge, Lasso, and Elastic Net are extensions of linear regression that incorporate regularization terms to prevent overfitting and manage multicollinearity. Support Vector models can handle both linear and non-linear relationships by using kernel functions.

Random Forest, Gradient Boosting, AdaBoost, Extra Trees, and XGBoost are ensemble methods that combine multiple decision trees to improve predictive performance and reduce overfitting. Random Forest and Extra Trees create a multitude of decision trees and aggregate their results, while Gradient Boosting, AdaBoost, and XGBoost construct a sequence of decision trees, where each tree tries to correct the mistakes of the previous one.

These machine learning-based models provide powerful tools for regression analysis, but they also come with their own set of challenges, such as increased computational complexity and decreased interpretability compared to traditional linear regression. The choice of model depends on the specific problem at hand, considering the balance between accuracy and interpretability, the computational resources available, and the nature of the data and the underlying relationships.

In this study, we employed machine learning-based regression models such as Ridge, Lasso, Elastic Net, Support Vector, Random Forest, Gradient Boosting, AdaBoost, Extra Trees, and XGBoost to predict the DRI.

The performance of each model on the training and test sets is as follows:

Table 1. Performance of models

Model	Train MSE	Train RMSE	Train R^2	Test MSE	Test RMSE	Test R^2
Ridge	78.24	8.85	0.45	81.12	9.01	0.48
Lasso	83.52	9.14	0.41	87.22	9.34	0.44
ElasticNet	90.42	9.51	0.36	95.89	9.79	0.39
Support Vector	124.88	11.18	0.12	148.04	12.17	0.06
Random Forest	15.02	3.88	0.89	94.42	9.72	0.40
Gradient Boosting	1.46	1.21	0.99	111.53	10.56	0.29
AdaBoost	22.88	4.78	0.84	116.80	10.81	0.26
Extra Trees	0.00	0.00	1.00	97.52	9.88	0.38
XGBoost	0.00	0.00	1.00	121.78	11.04	0.22
Linear Regression	78.21	8.84	0.45	80.82	8.99	0.48

The performance difference between the training and test sets suggests that all models have a tendency to overfit to the training data to some extent. This is particularly notable in the case of the Extra Trees and XGBoost models, which show nearly perfect performance on the training data but comparatively worse performance on the test data. On the other hand, the Ridge, Lasso, ElasticNet, and Linear Regression models show relatively consistent performance between the training and test data. This suggests that these models are learning a more general pattern in the data that is not specific to the training set. Among these models, the Linear Regression model shows the lowest RMSE and MSE, as well as the highest R^2 on the test data, making it the best performing model on the test data. However, it should be noted that it is also important to consider the complexity and interpretability of the model. While more complex models like Random Forest and Gradient Boosting can capture more complex patterns in the data, they are also more prone to overfitting and are harder to interpret. On the other hand, simpler models like Linear Regression may not capture as much of the complexity of the data, but they are simpler and easier to interpret.

4. SYMBOLIC REGRESSION ANALYSIS

In the pursuit of understanding complex relationships within data, various regression techniques have been employed. Traditional machine learning based regression models, although powerful, often lack transparency in representing the underlying mathematical relationships. In this context, symbolic regression, an evolutionary computation technique that discovers mathematical expressions to fit data, emerges as an advantageous alternative. In this study, we employ symbolic regression to predict DRI based on various geological features.

Unlike many machine learning models that act as black boxes, symbolic regression offers interpretability by producing transparent and human-readable mathematical expressions, providing insights into underlying relationships; flexibility in functional forms as it is not bound by predefined structures and can discover nonlinear and complex relationships; robustness to overfitting through the inclusion of a parsimony coefficient that encourages simpler expressions; and the ability to integrate domain knowledge by allowing researchers to incorporate specific functions, constraints, or relationships, thereby facilitating the discovery of meaningful models (Zhang et al. 2021).

The dataset is divided into a 70% training set and a 30% validation set, ensuring a robust evaluation of the model's predictive performance.

The performance metrics for the model on the training set were as follows: MSE = 59.08, RMSE = 7.69, R^2 = 0.56. The performance metrics for the model on the test set were as follows: MSE = 87.74, RMSE = 9.34, R^2 = 0.49.

Fig. 5 illustrates a scatter plot that compares the actual DRI values with the predicted DRI values, encompassing both the training and test sets, as derived from the symbolic regression model. The prediction equation for DRI according to symbolic regression is as follows:

$$DRI = 62.32 - (RT\_Igneous + BTS + \tan(4.98 \times (CAI - RT\_Metamor) \\ -UCS - 9.01 \times RT\_Sedimen - EQC) + CAI) \quad (2)$$

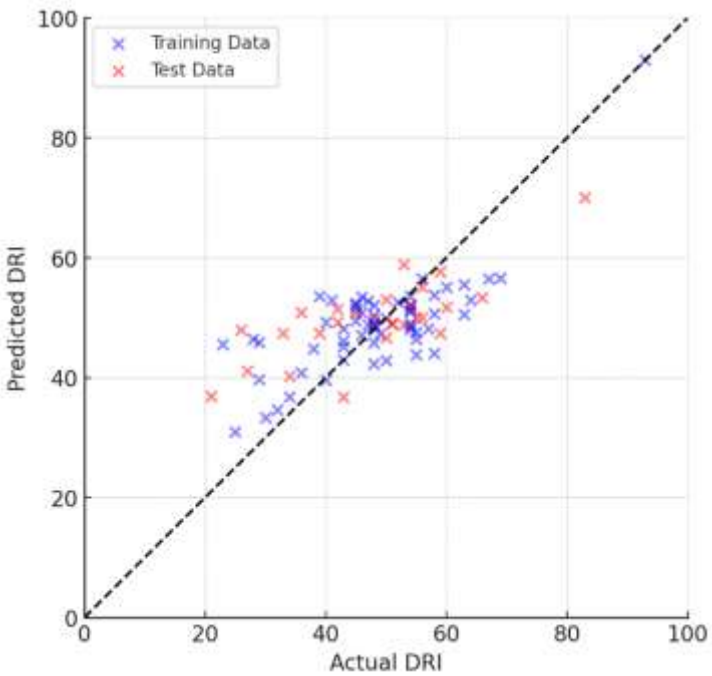

Fig. 5. Comparison of Actual and Predicted DRI Values through symbolic regression analysis

The evaluation of the symbolic regression and linear regression models on the given dataset reveals that the symbolic regression model outperforms the linear regression model across all considered metrics, with lower MSE values of 59.08 and 87.74 for the training and test sets respectively (compared to 78.21 and 80.82 for linear regression), lower RMSE values of 7.69 and 9.34 (compared to 8.84 and 8.99), and higher coefficient of determination (R^2) values of 0.56 and 0.49 (compared to 0.45 and 0.48), indicating a better overall fit to the observed data and potentially better generalization to unseen data, though the specific use case, complexity, interpretability, and computational efficiency of the models should be considered in the final selection.
The symbolic regression model for predicting DRI on the entire dataset exhibited a coefficient of determination (R^2) of 0.5375, reflecting that the model explains approximately 53.75% of the variance, along with a MSE of 67.87 and a RMSE of 8.24. In summary, the symbolic regression model demonstrates a reasonable level of accuracy and fit in predicting DRI for the given dataset. The lower values of MSE and RMSE, along with a relatively higher R^2, indicate a satisfactory predictive performance.

5. CONCLUSIONS

The research undertaken in this paper has provided significant insights into the estimation of the drillability rate index (DRI) for rocks, a vital parameter in the fields of

civil and mining engineering. Through the utilization of a comprehensive dataset and the application of various modeling techniques, including linear regression, machine learning-based regression, and the innovative use of symbolic regression, the study has contributed to the understanding of complex relationships between rock properties. Symbolic regression, in particular, has emerged as a powerful method, offering human-readable mathematical expressions and demonstrating superior performance in predicting DRI. The results not only address limitations in existing methods but also introduce a more robust and precise predictive model. These findings have the potential to lead to more accurate and cost-effective tunneling operations and reflect the promise of symbolic regression as a versatile and interpretable tool in the analysis of complex systems. Future work may focus on further enhancing the model by integrating additional features or exploring other cutting-edge techniques, thereby continuing to advance the field towards improved operational efficiency in civil and mining engineering sectors.

ACKNOWLEDGMENT

This work was supported by a National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT) (No. NRF-2022R1F1A1063228).

REFERENCES

Aligholi, S., Lashkaripour, G. R., Ghafoori, M, and Azali, S. T. (2017), "Evaluating the relationships between NTNU/SINTEF drillability indices with index properties and petrographic data of hard igneous Rocks", *Rock Mech Rock Eng*, **50**, 2929–2953.

Eide, L.N.R. (2014),"TBM tunnelling at the Stillwater mine", Norwegian University of Science and Technology MS Thesis

Macias, F.J., Dahl, F., Bruland, A. (2015), “New rock abrasivity test method for tool life assessments on hard rock tunnel boring: The rolling indentation abrasion test (RIAT)", *Rock Mech Rock Eng*, **49**, pages1679–1693.

Majeed, Y., Abu Bakar, M.Z., Butt, I.A. (2020), “Abrasivity evaluation for wear prediction of button drill bits using geotechnical rock properties”, *Bull. Eng. Geol. Environ*., **79**, 767–787.

Zhang, L., Zhang, Q., Zhou, S., Liu, S. (2021), “Modeling of tunneling total loads based on symbolic regression algorithm”, *Appl. Sci*., **11**, 5671.