Comparative Analysis of CAI Estimation using Symbolic Regression and Machine Learning Approaches

Tae Young Ko
Department of Energy and Resources Engineering, Kangwon National University, Chuncheon, Korea

Ju-Pyo Hong
Department of Energy and Resources Engineering, Kangwon National University, Chuncheon, Korea

Yun-Seong Kang
Department of Energy and Resources Engineering, Kangwon National University, Chuncheon, Korea

ABSTRACT: The abrasiveness of rocks being excavated is a major challenge in TBM tunneling, as it affects the performance and durability of cutting tools. The Cerchar abrasivity Index (CAI) is a widely used method to assess rock abrasiveness and predict tool wear and cutter life in TBM tunneling. The CAI can be estimated from rock properties, such as compressive strength, tensile strength, and petrographic factors. A novel approach using symbolic regression was proposed to predict CAI. Symbolic regression can generate accurate and interpretable mathematical equations to capture the relationship between inputs and outputs. The proposed approach was compared to traditional machine-learning-based regression models using a dataset obtained from published articles and geotechnical data reports. Various machine-learning-based regression methods were also used to forecast the CAI, and their performances were compared. The proposed symbolic regression-based CAI prediction model has the potential to improve the performance of models for predicting rock abrasivity.

Keywords: Cerchar Abrasivity Index, Wear, Symbolic Regression, Machine Learning.

1 INTRODUCTION

Tunnel boring machines (TBM) have become increasingly common in underground excavation projects. One of the major challenges in TBM tunneling is the abrasiveness of the rocks being excavated, which can have a significant impact on the performance and durability of cutting tools.

The Cerchar abrasivity index (CAI) is a widely used method for assessing the abrasiveness of rocks and has been shown to be an effective tool for predicting tool wear and cutter life in TBM tunneling. One advantage of using CAI to predict rock abrasiveness is that it is a simple and effective method for estimating the abrasiveness of rocks. The CAI can be determined from a relatively small number of simple laboratory tests, which makes it a practical and cost-effective option for many engineering applications. One way to estimate CAI values without direct experimentation is to use rock properties such as compressive strength, tensile strength, and petrographic factors. This

approach can be useful in cases where direct CAI measurements are not available, or where additional information about rock properties is desired.

Several studies have investigated the feasibility of using rock properties to estimate the CAI. The CAI was found to be strongly influenced by the degree of cementation, strength and amount of abrasive minerals, that is, the quartz content or equivalent quartz content in the rocks (Al-Ameen & Waller 1994, Plinninger et al. 2003, Rostami et al. 2014, Moradizadeh et al. 2016, Ko et al. 2016, Yaralı 2017, Ozdogan et al. 2018, Kahraman et al. 2018, Erarslan 2019).

Recently, machine-learning-based regression models have been proposed to estimate CAI using rock properties (Kwak & Ko 2022). These models can handle complex data and nonlinear relationships between variables, and may offer improved accuracy compared to empirical equations. Although machine learning models have shown great potential for predicting the CAI, some limitations still need to be addressed. One issue is the lack of interpretability of the models, which means that it is often difficult to understand how the model arrives at its predictions. Another challenge is the need for large amounts of high-quality training data, which may be difficult and expensive to obtain. Finally, some machine learning models may be overfitted to the training data, which means that they perform well on the training data but poorly on the new data.

To overcome these limitations, we propose a novel approach using symbolic regression for predicting CAI. Symbolic regression is a powerful technique that can generate mathematical equations that accurately capture the relationship between inputs and outputs. It has the advantage of producing models that are not only highly accurate, but also interpretable, making it easier for engineers to understand the underlying relationships between the input and output variables (Zhang et al. 2021).

In this study, we present a symbolic regression-based CAI prediction model and compare its performance with that of traditional machine-learning-based regression models. The proposed symbolic regression-based CAI prediction model has the potential to significantly improve the performance of models for predicting rock abrasivity, which can have important applications in the design and optimization of underground excavation and TBM tunneling operations.

2 DATA PREPARATION

The data used in this study were obtained from published articles and geotechnical data reports on tunneling projects worldwide. The dataset consisted of 417 observations, including CAIs, rock types, and strength parameters such as uniaxial compressive strength (UCS) and Brazilian tensile strength (BTS), as well as petrographic factors such as equivalent quartz content (EQC). The observations were divided into two groups: a training set comprised 70% of the data and a test set comprising the remaining 30%. A histogram of the variables is shown in Figure 1, with CAI as the dependent feature and the others as independent features.

3 MACHINE LEARNING-BASED REGRESSION ANALYSES

Various machine learning-based regression methods were used in this study to forecast the CAI, such as linear regression, ridge regression, lasso regression, Elastic Net regression, support vector regression (SVR), decision tree regression, k-nearest neighbor (KNN) regression, random forest regression, XGBoost regression, gradient boosting regression, and AdaBoost regression. Each technique has a unique way of modeling the relationship between the dependent variable (CAI) and the independent variables, and this study aimed to compare the performance of these different methods in predicting CAI.

Linear regression attempts to establish a linear relationship between a dependent variable (such as CAI) and one or more independent variables. Ridge regression adds a penalty term to ordinary least squares (OLS) regression to prevent overfitting. Meanwhile, lasso regression adds a penalty term to OLS regression with a different constraint, resulting in feature selection and overfitting prevention. Elastic net regression is a combination of both Ridge and Lasso regression techniques, with the aim of balancing their strengths.

Support vector regression (SVR) uses support vector machines (SVM) to build a nonlinear regression model. Decision tree regression builds a decision tree to model the relationship between independent and dependent variables. KNN regression predicts the value of the dependent variable by examining the values of the KNN. Random forest regression builds an ensemble of decision trees to decrease overfitting and improve prediction accuracy. Gradient boosting regression constructs an ensemble of weak prediction models (usually decision trees) and optimizes them to minimize the loss function. XGBoost regression is a gradient-boosting regression method that utilizes an optimized distributed gradient-boosting library to build a regression model. Finally, AdaBoost regression builds an ensemble of weak prediction models and weights them based on their performance to enhance prediction accuracy.

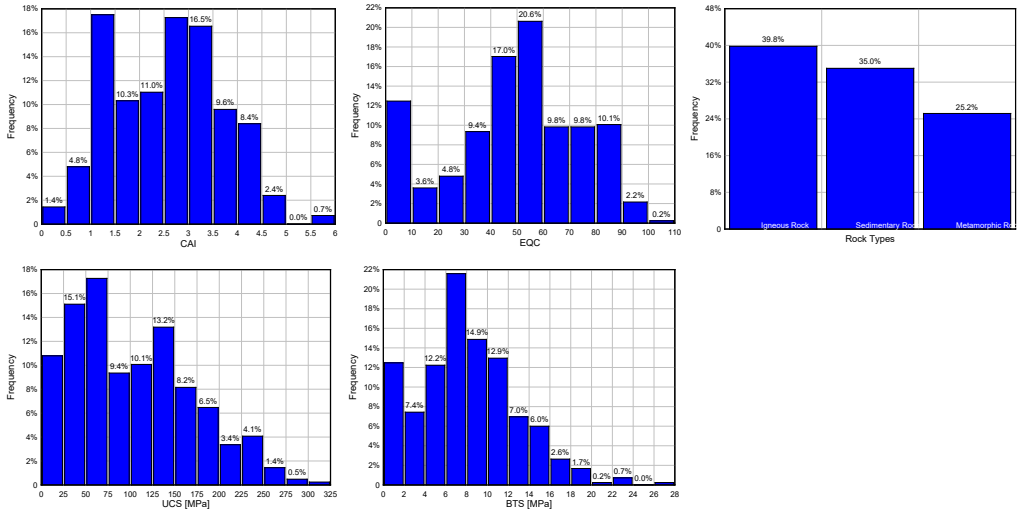

Figure 1. Variable distribution.

Table 1 presents the results of the machine-learning-based regression analyses used for model evaluation. The findings indicated that KNN regression was the most effective among the machine learning-based regression models. The KNN regression achieved an R^2 value of 0.74, which was the highest among all the machine-learning-based regression models evaluated for training data. Additionally, the RMSE value was the lowest among all machine learning-based regression models at 0.55 for the training data. For the test data, the KNN model exhibited the highest R^2 value of 0.63, and the lowest MAPE and RMSE values were 32.44% and 0.71, respectively. The results indicate that KNN regression was the best model for machine learning-based regression.

Table 1. Model evaluation based on machine learning-based regression analysis.

Model	Training data			Test data		
	MAPE(%)	RMSE	R2	MAPE(%)	RMSE	R2
Linear	36.64	0.77	0.48	44.5	0.85	0.46
Ridge	36.64	0.77	0.48	44.5	0.85	0.46
Lasso	36.85	0.77	0.48	44.84	0.86	0.46
Elastic Net	36.79	0.77	0.48	44.73	0.85	0.46
Random Forest	27.74	0.61	0.68	42.46	0.76	0.58
XGB	26.75	0.58	0.71	39.98	0.73	0.61
Decision Tree	22.39	0.57	0.71	34.88	0.72	0.62
KNN	23.92	0.55	0.74	32.44	0.71	0.63
Gradient Boosting	27.17	0.57	0.72	39.67	0.71	0.63
AdaBoost	31.02	0.63	0.66	43.9	0.78	0.56
SVR	37.93	0.78	0.48	45.62	0.85	0.47

Figure 2 shows the KNN regression plots for both training and test sets.

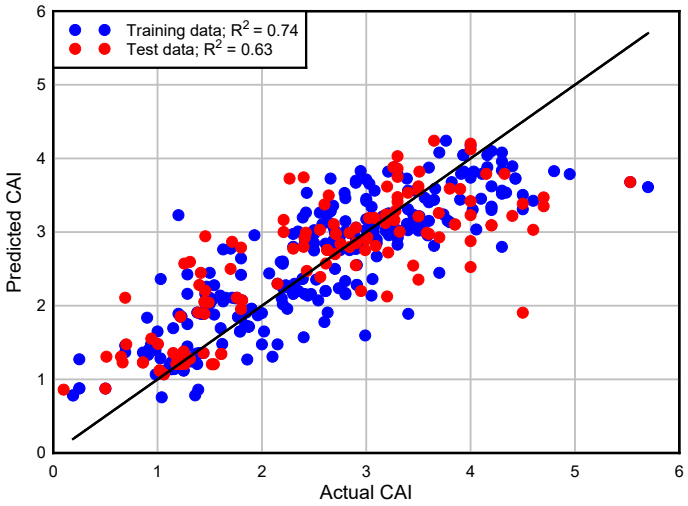

Figure 2. KNN regression plot.

4 SYMBOLIC REGRESSION ANALYSES

Symbolic regression is a machine learning technique that involves automatically searching for mathematical expressions that best fit a given dataset. Its advantages include the ability to discover complex relationships between input variables and output, as well as the potential to generate interpretable models that provide insights into the underlying processes.

Symbolic regression typically uses mathematical expressions that involve basic mathematical operations, such as addition, subtraction, multiplication, division, and exponentiation. Additionally, other functions, such as logarithmic, trigonometric, and exponential functions, are included in the candidate expressions.

To evaluate the fitness of candidate expressions, a fitness function is defined based on the specific problem being addressed. In symbolic regression, the fitness function measures the difference between the predicted output of the candidate expression and the actual output of the training dataset. The goal is to minimize this difference, which is also known as an error or loss function. The selection process involved selecting the best-performing candidate expressions from the current population to be used as parents for the next generation. This is typically performed using a fitness-proportional selection method, such as roulette wheel selection, where the probability of a candidate expression being selected as a parent is proportional to its fitness score. Mutation involves randomly changing the structure of candidate expressions by adding, deleting, or modifying mathematical operations or functions. This introduces diversity in the population and allows the exploration of new areas in the search space. The process of generating new generations of candidate expressions continues until a satisfactory model is found, which is typically defined by a pre-specified stopping criterion, such as the maximum number of generations or a minimum fitness score. Once a satisfactory model is obtained, it can be used to make predictions using new data.

Table 2 presents the results of the model evaluation using the symbolic regression analyses. The training data had an R^2 value of 0.64, and the test data had a value of 0.62. The RMSE and MAPE for the training data were 0.64 and 27.32, respectively, whereas the corresponding values for the test data were 0.72 and 36.27%, respectively.

The CAI prediction model using symbolic regression is as follows:

$$CAI = log10(UCS) + (0.014996 * (EQC + ((RT - 1.97681 * cos(-16.0951 * EQC)) * BTS) + tan(0.795238 * RT)))) \quad (1)$$

where EQC is the equivalent quartz content in %, UCS is the uniaxial compressive strength in MPa, BTS is the Brazilian tensile strength in MPa, and RT = 1 for igneous rocks, 2 for sedimentary rocks, and 3 for metamorphic rocks.

Table 2. Model evaluation based on symbolic regression analysis.

Model	Training data			Test data		
	MAPE(%)	RMSE	R2	MAPE(%)	RMSE	R2
Symbolic	27.32	0.64	0.64	36.72	0.72	0.62

Figure 3 shows the symbolic regression plots for both the training and test sets.

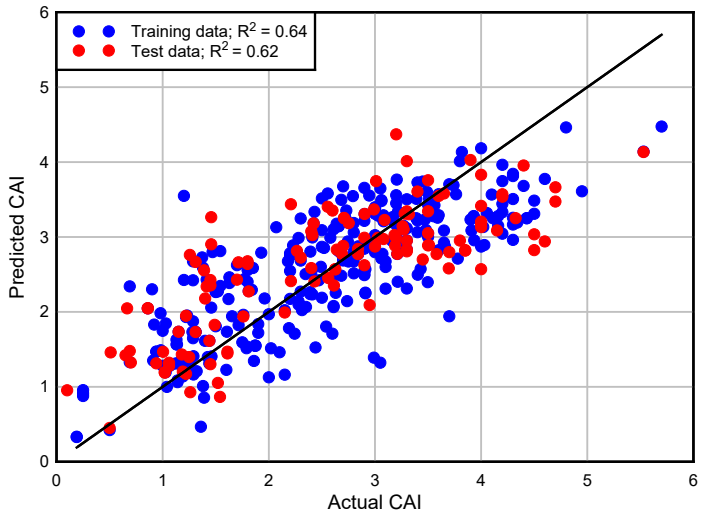

Figure 3. Symbolic regression plot.

Although the performance of the symbolic regression model was slightly lower than that of KNN or other ensemble models based on machine learning, it demonstrated a better prediction ability than linear regression analysis models. Therefore, considering the advantage of representing the prediction model as a mathematical formula, symbolic regression has potential for practical applications.

5 CONCLUSION

In this study, a novel approach was proposed using symbolic regression to predict the CAI of rocks, which is a widely used method for assessing the abrasiveness of rocks in TBM tunneling. Symbolic regression is a powerful technique that can generate mathematical equations that accurately capture the relationship between inputs and outputs.

The proposed symbolic-regression-based CAI prediction model was compared with traditional machine-learning-based regression models. Although the performance of the symbolic regression model was slightly inferior to that of traditional machine learning-based regression models, its ability to provide interpretable mathematical expressions was highlighted. The dataset used in this study consisted of 417 observations, including CAIs, rock types, and strength parameters such as UCS and BTS, as well as petrographic factors such as EQC. Various machine-learning-based regression methods were used in this study to forecast the CAI. The proposed approach has the potential to significantly improve the performance of models for predicting rock abrasivity, which can have important applications in the design and optimization of underground excavation and TBM tunneling operations.

ACKNOWLEDGEMENTS

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

REFERENCES

Al-Ameen, S.I. & Waller, M.D. 1994. The influence of rock strength and abrasive mineral content on the Cerchar Abrasive Index. Eng. Geol. 36 (3-4), pp. 293-301. DOI: 10.1016/0013-7952(94)90010-8

Erarslan, N. 2019. Assessment of Cerchar abrasivity test in anisotropic rocks. *Geomech. Eng*. 17 (6), pp. 527-534. DOI: 10.12989/gae.2019.17.6.527

Kahraman, S., Fener, M., Käsling, H. & Thuro, K. 2018. Investigating the effect of strength on the LCPC abrasivity of igneous rocks. *Geomech. Eng.* 15 (2), pp. 805-810. DOI: 10.12989/gae.2018.15.2.805

Ko, T.Y., Kim, T. K., Son, Y. & Jeon, S. 2016. Effect of geomechanical properties on Cerchar Abrasivity Index (CAI) and its application to TBM tunnelling. *Tunn. Undergr. Space Technol*., 57, pp. 99-111. DOI: 10.1016/j.tust.2016.02.006

Kwal, N.S. & Ko, T.Y. 2022. Machine learning-based regression analysis for estimating Cerchar abrasivity index. *Geomech. Eng*. 29 (3), pp. 219-228. DOI: 10.12989/gae.2022.29.3.219

Moradizadeh, M., Cheshomi, A., Ghafoori, M. & TrighAzali, S. 2016. Correlation of equivalent quartz content, Slake durability index and Is50 with Cerchar abrasiveness index for different types of rock. *Int. J. Rock Mech. Min. Sci*. 86, pp. 42-47. DOI: 10.1016/j.ijrmms.2016.04.003

Ozdogan, M.V., Deliormanli, A.H. & Yenice, H. 2018. The correlations between the Cerchar abrasivity index and the geomechanical properties of building stones. *Arab. J. Geosci.* 11, p. 604. DOI: 10.1007/s12517-018-3958-8

Plinninger, R., Kasling, H., Thuro, K. & Spaun, G. 2003. Testing conditions and geomechanical properties influencing the CERCHAR abrasiveness index (CAI) value. *Int. J. Rock Mech. Min. Sci*., 40, pp. 259–263. DOI: 10.1016/S1365-1609(02)00140-5

Rostami, J., Ghasemi, A., Gharahbagh, E.A., Dogruoz, C. & Dahl, F. 2014. Study of dominant factors affecting Cerchar abrasivity index. *Rock Mech Rock Eng*. 47 (5), pp. 1905-1919. DOI: 10.1007/s00603-013-0487-3

Yaralı, O. 2017. Investigation into relationships between Cerchar hardness index and some mechanical properties of coal measure rocks. *Geotech. Geol. Eng.*, 35, pp. 1605–1614. DOI: 10.1007/s10706-017-0195-y

Zhang, L., Zhang, Q., Zhou, S., & Liu, S. 2021. Modeling of tunneling total loads based on symbolic regression algorithm. *Appl. Sci.*, 11, 5671. DOI: 10.3390/app11125671