The 2025 World Congress on
Advances in Structural Engineering and Mechanics (ASEM25)
BEXCO, Busan, Korea, August 11-14, 2025

Machine learning-based prediction of TBM utilization factor

Tae Young Ko

Department of Energy and Resources Engineering, Kangwon National University, Chuncheon 24341, Korea
tyko@kangwon.ac.kr

ABSTRACT

A machine learning approach was developed to predict the utilization factor of Tunnel Boring Machines (TBMs) considering two realistic data availability scenarios. The first scenario relies only on TBM type, tunnel geometry, general ground condition, and theoretical advance rate, without RMR. The second scenario includes detailed geological classification, such as rock mass rating (RMR) and rock type. Exploratory analysis revealed that RMR availability leads to more stable and interpretable patterns in TBM performance. Ten regression models were tested, and model performance was compared through five-fold cross-validation. For the first scenario (RMR-excluded), the random forest model achieved the best test R^2 of 0.32. In the second scenario (RMR-included), the gradient boosting model yielded a significantly higher test R^2 of 0.82. Results indicate that geotechnical parameters, particularly RMR, are critical for achieving reliable prediction accuracy. This methodology demonstrates the efficacy of data-driven models for performance forecasting in mechanized tunneling, particularly during preliminary design phases.

1. INTRODUCTION

The utilization factor (UF) of a Tunnel Boring Machine (TBM) refers to the proportion of time during which the machine is actively engaged in boring compared to the total available time. As a multiplier of penetration rate (PR), UF directly determines the advance rate (AR), which is a key indicator of tunneling performance. Accurate utilization factor prediction is essential for project planning, directly impacting schedule reliability, resource allocation, and cost estimation.

Conventional models such as those developed by the Colorado School of Mines (CSM) and the Norwegian University of Science and Technology (NTNU) have offered methods to estimate UF. The CSM model attempts to compute UF by summing detailed downtime components including boring, regripping, cutter change, maintenance, and logistics (Farrokh, 2012). Although theoretically comprehensive, the model requires precise field measurements and detailed breakdowns of activity times, which are typically unavailable during pre-construction phases. The NTNU model,

based on empirical data from completed tunneling projects, simplifies this process by using averaged charts for scheduled downtimes such as mucking, support installation, and maintenance (Bruland, 2000). Nevertheless, this approach inadequately addresses geological variability including squeezing conditions or excessive groundwater inflow, often causing substantial unforeseen delays. Both approaches demonstrate limited adaptability to varying TBM configurations and geological conditions. Performance limitations in challenging geological conditions have been documented in subsequent research (Farrokh, 2012; Rostami, 2016).

Beyond model complexity, geotechnical data availability presents a fundamental constraint in UF prediction. In open or gripper-type TBM projects, the tunnel face is exposed, which allows engineers to observe discontinuities and classify rock mass using systems such as RMR. In contrast, EPB and slurry TBMs operate with pressurized closed faces, making it difficult or impossible to observe in-situ ground conditions during operation. Consequently, numerous urban and soft ground projects lack reliable RMR data, limiting the applicability of conventional rock mass-based prediction models (Frough et al., 2014).

Recent research has explored data-driven and simulation-based approaches as more flexible alternatives. Machine learning methods have been used to model UF based on operational parameters and limited geological information, achieving reasonable predictive performance even without full geotechnical datasets (Yu et al., 2021). Simulation-based approaches have also demonstrated the ability to capture system-level interactions between equipment, tunnel logistics, and downtime events (Khetwal et al., 2021; Moosazadeh et al., 2018).

This research develops machine learning models for TBM utilization factor prediction under two practical scenarios. The first scenario targets cases where RMR is not available and relies solely on operational and project-level parameters such as TBM type, tunnel diameter and length, and general geological condition. The second scenario considers the availability of RMR as part of the input feature set. Unlike real-time monitoring systems, the models in this study are intended to support pre-construction decision-making, particularly in feasibility studies and early project design, where reliable UF estimation is critical yet often difficult to achieve.

The following sections present the datasets, modeling methods, and evaluation results for both scenarios and discuss the implications of the findings for tunnel project planning and performance forecasting in mechanized tunneling.

2. DATA AND METHODOLOGY

Two datasets representing varying levels of geotechnical information availability during tunnel project planning were utilized. The first dataset covers a wider range of ground conditions including both competent rock and soft soils. It includes projects using gripper, double shield, earth pressure balance, slurry, and single shield TBMs. Due to the enclosed nature of shielded machines operating in soft ground, RMR values were not available for this group. The second dataset includes tunneling projects carried out in hard rock conditions using gripper and double shield TBMs. In these

cases, the tunnel face is exposed, allowing reliable classification of rock mass quality. Rock Mass Rating (RMR) values were available for all records in this dataset.

Both datasets contain the utilization factor as the response variable, expressed as a percentage of boring time over total operating time. Independent variables comprised TBM type, tunnel diameter, tunnel length, and ground classification. In the dataset with RMR information, RMR is included as an additional explanatory variable.

2.1 Exploratory Data Analysis

Two datasets were used to reflect different levels of geotechnical information availability. The first dataset does not include rock mass classification and is based only on machine-related and design parameters. The second dataset includes rock mass rating and lithological data, restricted to projects with available geological observations. The dataset without RMR includes the following variables. These are TBM type, tunnel diameter, tunnel length, ground condition categorized as either rock or soil, theoretical advance rate, and utilization factor. Theoretical advance rate was derived from a TBM performance prediction model incorporating machine specifications and ground conditions rather than direct measurement. Tunnel diameter and tunnel length both exhibited right-skewed distributions, with most tunnels having diameters below 7 meters and lengths under 10 kilometers. Utilization factor exhibited wide variability with irregular distribution patterns, displaying a relatively low mean and limited conformity to normal distribution. The distribution characteristics are illustrated in Fig. 1.

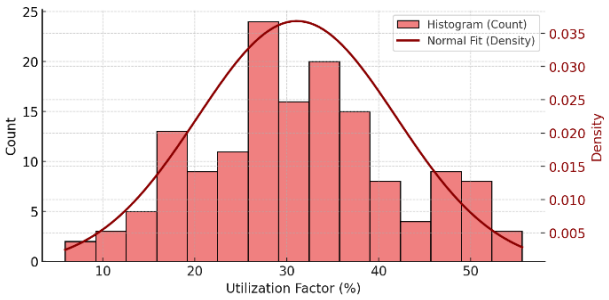

Fig. 1 Distribution of utilization factor in RMR-excluded dataset

The dataset that includes RMR consists of the following variables. These are rock mass rating, rock type, TBM type, tunnel diameter, tunnel length, and utilization factor. RMR values ranged predominantly between 60-80, indicating that most projects occurred in competent rock conditions. Utilization factor demonstrated a more compact and symmetric distribution, with values clustered around a mean of approximately 35%. This pattern indicates enhanced TBM performance predictability when geotechnical data are available. The distribution is illustrated in Fig. 2.

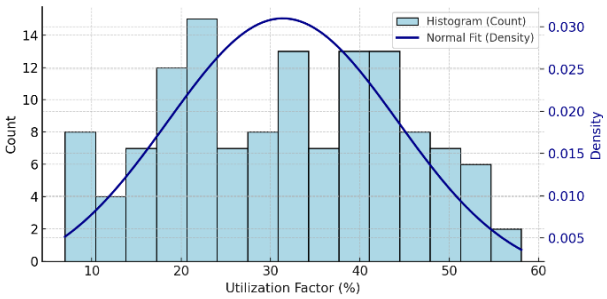

Fig. 2 Distribution of utilization factor in RMR-Included dataset

Correlation analysis was performed separately for each dataset. In the dataset with RMR, a strong positive correlation was observed between RMR and utilization factor. In contrast, in the dataset without RMR, tunnel length showed a weak positive association with utilization, and theoretical advance rate showed a weak negative trend. Tunnel diameter exhibited minimal correlation with utilization in both datasets. Correlation patterns are presented in the heatmaps of Figs. 3 and 4.

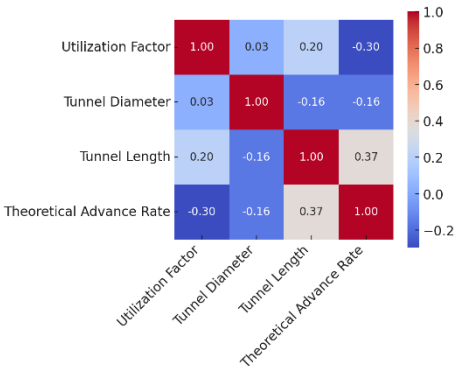

Fig. 3 Correlation heatmap for RMR-excluded dataset

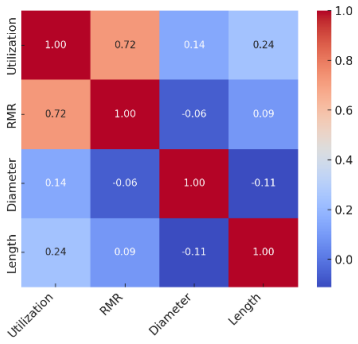

Fig. 4 Correlation heatmap for RMR-included dataset

The results indicate that when only design and machine parameters are available, the ability to explain or predict TBM utilization is limited. However, when rock mass

classification is included, the relationships become more structured and predictive performance improves. These findings support the development of separate modeling strategies depending on the availability of geotechnical data.

2.2 Feature Selection and Preprocessing

Input variables were selected based on the exploratory analysis. In the dataset without RMR, the model used TBM type, tunnel diameter, tunnel length, ground condition, and theoretical advance rate. Theoretical advance rate was calculated from a TBM performance prediction model considering machine specifications and ground conditions. In the dataset with RMR, additional features included rock mass rating and rock type. Utilization factor, expressed as a percentage, served as the target variable for both scenarios. Categorical variables such as TBM type and ground or rock classifications were encoded using one-hot encoding. Records with missing values were excluded. No synthetic data or imputation was applied to preserve the empirical integrity of the dataset.

2.3 Machine Learning Models

Ten regression algorithms were implemented for utilization factor prediction. These included both linear and non-linear algorithms. The linear group consisted of ordinary least squares, ridge, and lasso regression. Tree-based and ensemble methods included decision tree, random forest, gradient boosting, XGBoost, and LightGBM. In addition, support vector regression and k-nearest neighbors were used. Initial model evaluation employed five-fold cross-validation methodology. This approach enabled performance comparison under standardized conditions. The best-performing model for each scenario was then selected based on average validation scores. Following model selection, data were partitioned into training and test sets using an 80:20 split. The selected model was retrained on the training set and optimized through hyperparameter tuning using grid search. The final model was then tested to evaluate its generalization performance. Model performance was assessed using the coefficient of determination and root mean square error. This framework enabled fair algorithm comparison and reliable selection of the optimal predictive model for each data scenario.

3. RESULTS AND MODEL EVALUATION

Results of TBM utilization factor predictive modeling are presented for two scenarios differentiated by rock mass rating availability. In both cases, machine learning models were trained and evaluated using statistical and visual methods.

3.1 Exploratory Analysis of First Scenario (RMR-Excluded)

The first scenario utilized variables typically available during pre-construction phases. These included TBM type, tunnel diameter, tunnel length, general ground condition, and theoretical advance rate. Detailed geotechnical parameters including RMR were unavailable for this scenario. Exploratory analysis identified weak negative correlation between utilization and theoretical advance rate, with nonlinear relationships observed for tunnel length. Tunnel diameter showed no clear correlation with utilization.

Ground condition and TBM type demonstrated notable influence among categorical variables. Projects in rock generally showed higher utilization than those in soil. Similarly, Earth Pressure Balance and Single Shield TBMs exhibited higher average values than Gripper or Double Shield machines. These patterns are summarized in Fig. 5.

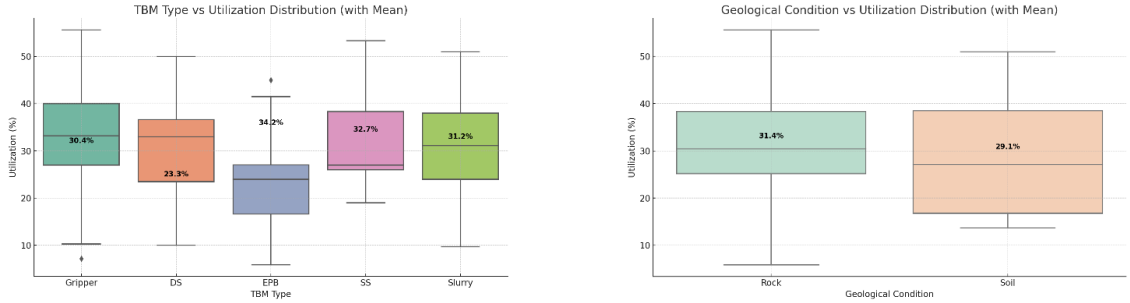

Fig. 5 Boxplot of TBM utilization factor by TBM type and ground condition (RMR-excluded dataset)

3.2 Model Performance First Scenario (RMR-Excluded)

Performance evaluation of ten regression algorithms employed five-fold cross-validation. Random forest achieved optimal performance with average validation R^2 of 0.38. Consequently, random forest was selected for final model development. Data partitioning employed an 80:20 training-to-test ratio. After hyperparameter tuning through grid search, the model achieved a training R^2 of 0.88 and a test R^2 of 0.32. The root mean squared error was 3.7 on the training set and 9.4 on the test set. Fig. 6 presents the predicted versus actual utilization values. Training set predictions demonstrated close agreement with actual values, whereas test results exhibited increased scatter, indicating potential overfitting.

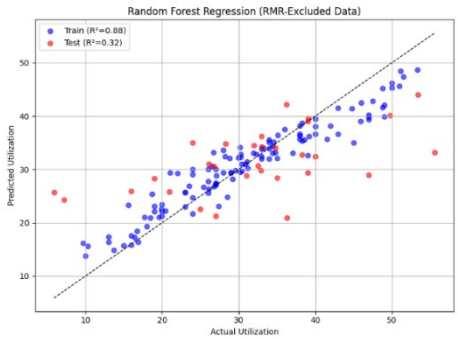

Fig. 6 TBM utilization prediction result without RMR

3.3 Exploratory Analysis of Second Scenario (RMR-Included)

The second scenario incorporated rock mass rating and lithological data alongside tunnel and machine parameters. This enabled comprehensive analysis of geological influence on TBM utilization.

Utilization factor was analyzed across rock types and TBM types. Sedimentary rock showed the highest average utilization, followed by metamorphic and igneous formations. In terms of machine type, Double Shield TBMs recorded higher average utilization than Gripper types. These observations are shown in Fig. 7.

The presence of structured geological variables such as RMR and rock type contributed to more stable and interpretable patterns in the data. This supports their use in predictive modeling of TBM performance.

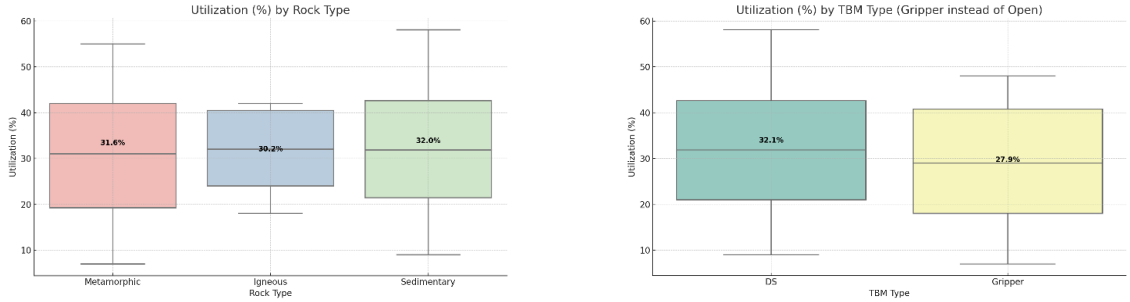

Fig. 7 Boxplot of TBM utilization factor by rock type and TBM type in RMR included dataset

3.4 Model Performance Second Scenario (RMR-Included)

Ten regression algorithms underwent five-fold cross-validation testing. Gradient boosting achieved optimal performance with average validation R^2 of 0.82. This model was selected for final optimization and testing.

The dataset was split into training and test sets in an eight to two ratio. Hyperparameter tuning was performed through grid search with ten-fold cross-validation. The optimized model achieved a training R^2 of 0.92 and a test R^2 of 0.82. The root mean squared error was 3.53 on the training set and 5.40 on the test set.

Fig. 8 shows the comparison between predicted and actual utilization. Conversely to the first scenario, predictions demonstrated close alignment with actual values, indicating enhanced generalization and model reliability.

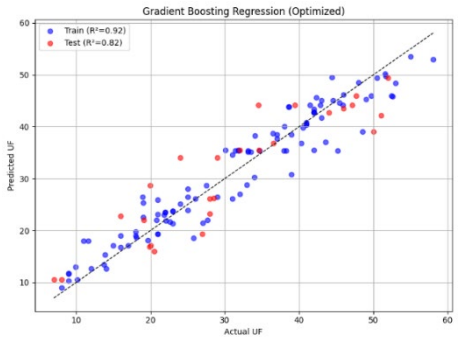

Fig. 8 TBM utilization prediction result with RMR

3.5 Comparison Between Scenarios

Modeling scenarios exhibited distinct predictive performance differences. In the first scenario (RMR-excluded), the best model achieved a test R^2 of 0.32. In the second scenario (RMR-included), the test R^2 increased to 0.82. This improvement of more than 50 percentage points in explained variance highlights the importance of incorporating geological classification into utilization prediction.

The comparison between Figs. 6 and 8 illustrates this contrast. Without RMR, predictions were widely scattered and less reliable. When RMR was used, predicted values closely followed the actual trend across the entire range. This indicates that models with geological input not only improve accuracy but also enhance generalization.

While project and machine parameters provide useful context, they are not sufficient to explain the full variability of TBM performance. Including structured geological information such as RMR significantly improves model stability and allows for more dependable forecasting, particularly in early-stage planning.

4. CONCLUSIONS

Machine learning techniques were applied to develop TBM utilization factor prediction models under two distinct data availability scenarios. The first dataset relied solely on machine type, tunnel dimensions, ground condition, and theoretical advance rate. The second dataset included rock mass rating and lithological classification.

Exploratory analysis revealed more stable and symmetric utilization factor distributions when rock mass rating was available. Conversely, the first scenario (RMR-excluded) exhibited greater variability and reduced predictability. Correlation analysis confirmed strong RMR-utilization relationships, whereas design variables alone offered limited predictive insight.

Ten regression algorithms were implemented for each scenario. In the first scenario (RMR-excluded), random forest demonstrated optimal performance, though predictive capability remained constrained. In the second scenario (RMR-included), gradient boosting achieved high accuracy, yielding test R^2 exceeding 0.82. This confirmed that including geological classification significantly improves model performance.

The results demonstrate that machine learning can provide meaningful predictions of TBM utilization, particularly when supported by reliable geotechnical data. When such information is not available, basic project parameters can still offer approximate estimates, though with reduced reliability. These findings highlight the importance of structured geological input in early-stage planning and support the use of data-driven models in tunnel design and risk assessment.

ACKNOWLEDGMENT

This work was supported by a grant from the Korea Agency for Infrastructure Technology Advancement (KAIA) funded by the Ministry of Land, Infrastructure, and Transport (Grant RS2024-00416524). Additional support was provided by the Energy &

Mineral Resources Development Association of Korea (EMRD) grant funded by the Korean government (MOTIE) (Grant 2021060003, Training Program for Specialists in Smart Mining).

REFERENCES

Bruland, A. (2000), “Hard rock tunnel boring,” Doctoral dissertation, Norwegian University of Science and Technology.

Farrokh, E. (2012), “Study of utilization factor and advance rate of hard rock TBMs,” Ph.D. Dissertation, Colorado School of Mines, Golden, CO.

Frough, O., Torabi, S.R. and Yagiz, S. (2014), “Application of RMR for estimating rock-mass–related TBM utilization and performance parameters: A case study,” *Rock Mech. Rock Eng*., **47**(4), 1267–1283.

Khetwal, A., Mishra, B. and Singh, A. (2021), “Simulation-based estimation of TBM utilization factor using stochastic models,” *Tunn. Undergr. Space Technol*., **108**, 103756.

Ko, T.Y., Kim, T.K. and Lee, D.H. (2019), “Statistical characteristics and rational estimation of rock TBM utilization,” *Tunnel & Underground Space*, **29**(5), 356–366.

Moosazadeh, S., Aghababaie, H., Hoseinie, S.H. and Ghodrati, B. (2018), “Simulation of tunnel boring machine utilization: A case study,” *J. Min. Environ*., **9**(1), 53–60.

Rostami, J. (2016), “Performance prediction of hard rock Tunnel Boring Machines (TBMs) in difficult ground,” *Tunn. Undergr. Space Technol*., **57**, 173–182.

Yu, H., Tao, J., Qin, C., Sun, H. and Liu, C. (2021), “A novel A-CNN method for TBM utilization factor estimation,” *J. Phys.: Conf. Ser*., **2002**, 012049.