Ensemble Machine Learning for Predicting TBM Penetration Rate with Limited Geotechnical Data

Halil Karahan * and Devrim Alkaya

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.2047.v1

Keywords: Tunnel Boring Machine (TBM); Penetration Rate (ROP) Prediction; machine learning algorithms; Least Squares Boosting (LSBoost); feature importance analysis; SHAP and PDP interpretability; geotechnical data modeling

Article

Ensemble Machine Learning for Predicting TBM Penetration Rate with Limited Geotechnical Data

Halil Karahan * and Devrim Alkaya

Department of Civil Engineering, Pamukkale University, Denizli 20160, Turkey

* Correspondence: hkarahan@pau.edu.tr

Featured Application

The developed LSBoost-based framework enables accurate TBM penetration prediction under limited geotechnical data, providing interpretable insights into key factors and supporting optimized design, parameter selection, and operational decision-making in tunneling projects.

Abstract

This study evaluated the predictive performance of Random Forest, Bagged Trees, Support Vector Machines (SVM), and Least Squares Boosting (LSBoost) for estimating Tunnel Boring Machine (TBM) penetration rate (ROP). While all models achieved acceptable accuracy, LSBoost outperformed the others, showing the highest correlation ($R = 0.965$) and coefficient of determination ($R^2 = 0.909$), along with the lowest RMSE and MAE. Its performance remained robust after Z-score normalization, highlighting its ability to capture nonlinear parameter interactions and generalize well on limited geotechnical datasets. Random Forest and Bagged Trees showed similar performance, with Bagged Trees only slightly improved by normalization. SVM performed less effectively, indicating limited capacity to model complex TBM penetration behavior. Feature importance and SHAP analyses identified discontinuity spacing (DPW) and uniaxial compressive strength (UCS) as the primary controlling factors, while brittleness index (BI) was more influential within the SVM model. Agreement between Jacobian-based derivative analyses and SHAP results confirmed both mathematical sensitivity and engineering interpretability. Overall, TBM penetration prediction is a multivariate and inherently nonlinear problem. LSBoost provides reliable and high-accuracy predictions even under data-constrained conditions. The combination of SHAP- and PDP-based feature importance analyses enhances interpretability, supporting engineering decision-making in TBM design and operation. These findings emphasize the applicability of machine learning approaches for accurate, interpretable, and robust TBM performance prediction.

Keywords: Tunnel Boring Machine (TBM); Penetration Rate (ROP) Prediction; machine learning algorithms; Least Squares Boosting (LSBoost); feature importance analysis; SHAP and PDP interpretability; geotechnical data modeling

1. Introduction

The effective management of time, cost, and operational risks in mechanized tunneling projects largely depends on the accurate prediction of Tunnel Boring Machine (TBM) performance. In this context, reliable estimation of TBM penetration rate (ROP) constitutes a critical engineering challenge, as it directly influences excavation scheduling, operational strategy development, and the mitigation of geotechnical risks. Penetration rate is governed by rock mass characteristics, the spacing and orientation of discontinuities, and machine operational parameters. Owing to the complex interactions among these variables, TBM penetration behavior is inherently nonlinear.

The earliest systematic investigations into TBM penetration prediction employed simple and multiple linear regression techniques to quantify the influence of rock mass parameters. For instance,

 doi:10.20944/preprints202602.2047.v1

the studies by Yağiz [1] and Mansouri & Moomivand [2] demonstrated the applicability of regression-based approaches for estimating penetration rate as a function of rock mass properties. Both studies identified uniaxial compressive strength (UCS), Brazilian tensile strength (BTS), brittleness/hardness, discontinuity spacing, and discontinuity orientation as key determinants of TBM performance.

Yağiz [1] conducted regression analyses based on field and laboratory measurements from the Queens Water Tunnel #3 Stage 2 (7.5 km, hard rock) and reported a strong correlation ($r = 0.82$). Similarly, Mansouri & Moomivand [2] analyzed data from the Karaj–Tehran Water Conveyance Tunnel (4.61 m diameter, 32 km length, schist and siltstone formations), evaluated the combined influence of five parameters, and proposed an empirical formulation for TBM penetration ($r = 0.75$). Despite differences in tunnel location, lithology, and site conditions, these studies confirmed that linear and multiple regression methods can provide reliable performance estimates and highlighted the critical role of rock mass parameters. Variations in data structure and correlation levels were primarily attributed to differences in geological and project-specific conditions.

However, given the inherently nonlinear and complex relationships governing TBM performance, regression-based formulations alone are often insufficient to fully capture system behavior. Consequently, heuristic optimization algorithms have been increasingly adopted to improve predictive accuracy. For example, Yağiz and Karahan [3] developed a Particle Swarm Optimization (PSO)-based model aimed at minimizing discrepancies between measured and predicted ROP values.

Subsequent research expanded methodological diversity through the application of Harmony Search (HS), Differential Evolution (DE), Grey Wolf Optimizer (GWO), and Firefly algorithms, demonstrating that TBM penetration rate is jointly controlled by machine-related and geological factors [4–6]. The difficulty of representing such complex interactions through single-parameter empirical equations, combined with the moderate accuracy of classical models ($R^2 \approx 0.70$), has driven the transition toward advanced soft computing techniques. In recent years, Artificial Neural Networks (ANN) [7–11], Support Vector Machines (SVM) [12–16], fuzzy logic approaches [17–20], and hybrid soft computing frameworks [21–25] have been widely implemented, yielding substantially improved predictive performance.

With the increasing availability of field data and advances in computational capabilities, machine learning (ML) methods have emerged as powerful tools for TBM performance prediction. Owing to their capacity to model multivariate and nonlinear relationships, ML-based approaches offer enhanced flexibility and accuracy in ROP estimation [25].

Numerous studies have successfully applied ML techniques to TBM penetration prediction. Ghorbani and Yagiz [26] implemented Gradient Boosting, XGBoost, LightGBM, AdaBoost, and CatBoost algorithms for ROP estimation, reporting particularly high accuracy for XGBoost and CatBoost models. Similarly, large cross-project datasets have been analyzed using Random Forest, bagging, and stacking ensemble approaches; SHAP analyses revealed that operational parameters such as torque and thrust exert a dominant influence on TBM penetration rate.

Xu et al. [27] compared k-Nearest Neighbor (KNN), decision tree, and ANN models, noting that KNN may exhibit superior performance depending on dataset characteristics. Beyond ROP estimation, ML techniques have also been applied to predict TBM loads and associated risks. Kong et al. [28] demonstrated that a Random Forest model developed for thrust and torque prediction in EPB-type TBMs outperformed theoretical approaches under mixed ground conditions. Kim et al. [29] improved prediction accuracy of surface settlements in urban TBM tunneling by integrating Random Forest with data-driven feature selection. Likewise, Hou et al. [30] employed a Random Forest-based classifier to predict shield jamming risk under limited field data conditions, achieving superior generalization compared with SVM, KNN, and logistic regression models.

Optimization-assisted ML frameworks have gained increasing prominence in TBM modeling. Yang, Liu, and Song [31] demonstrated that an Improved Sparrow Search Algorithm-optimized GBRT (ISSA-GBRT) model achieved superior ROP prediction performance. Zhou et al. [32]

 doi:10.20944/preprints202602.2047.v1

emphasized the high accuracy of optimization-enhanced SVM models under hard rock conditions. Karahan and Alkaya [25] showed that Bayesian-optimized SVR and parametric/ML hybrid models improved both predictive accuracy and engineering interpretability in TBM penetration modeling. In a subsequent study [33], Karahan and Alkaya compared parametric and ML approaches using Jacobian-based analyses and Partial Dependence Plots (PDP) to quantify variable effects. While the interaction-based M6 model yielded the highest predictive accuracy, BI and UCS remained dominant factors, and the indirect contributions of DPW and BTS were more clearly elucidated. ML results were consistent with parametric findings, with the Generalized Additive Model (GAM) demonstrating the strongest overall performance. Both studies utilized 151 cross-sectional data points from the Queens Water Tunnel #3 Stage 2 dataset [1,3,4,25,33], collectively establishing a reliable and interpretable hybrid modeling framework for TBM ROP prediction.

While previous studies have primarily been based on relatively large datasets, this study aims to fill a critical gap by focusing on reliable and stable TBM penetration predictions under limited geotechnical data conditions. For this purpose, a smaller dataset of 49 cross-sections—approximately one-third the size of the previously analyzed dataset—from the Karaj–Tehran Water Conveyance Tunnel [2] was employed.

In addition to widely adopted methods such as Random Forest, Bagged Trees, and Support Vector Machines, this study comparatively evaluates the Least Squares Boosting (LSBoost) approach, which has not previously been applied to TBM penetration prediction. By sequentially emphasizing residual errors at each iteration, LSBoost incrementally strengthens weak learners and progressively enhances regression performance. Its ability to effectively model complex and nonlinear relationships renders it particularly suitable for small-sample datasets. Although the method has been reported to provide high accuracy and stability under limited data conditions, applications to TBM ROP prediction remain absent from the literature. Furthermore, systematic comparisons of advanced ML and ensemble techniques on limited field datasets such as Karaj–Tehran are scarce.

Accordingly, this study fills a significant gap by systematically evaluating the performance of ensemble models—including LSBoost—on small-scale TBM datasets. The findings demonstrate that high accuracy, stability, and engineering interpretability can be achieved even under data-constrained conditions. The proposed framework therefore offers a practical, reliable, and performance-oriented predictive approach for researchers and practitioners working with limited field data, contributing both scientifically and operationally to the advancement of TBM performance modeling.

2. Materials and Methods

2.1. Dataset Used in the Study

The dataset used in this study is based on field and laboratory measurements collected by Mansouri and Moomivand from the Karaj–Tehran water conveyance tunnel project [2]. The dataset was obtained through a systematic sampling procedure and includes key parameters such as uniaxial compressive strength (UCS), Brazilian tensile strength (BTS), brittleness index (BI), discontinuity plane spacing (DPW), discontinuity orientation (α), and the Tunnel Boring Machine penetration rate (Rate of Penetration, ROP).

Although this dataset has been relatively limited in its application within TBM performance prediction studies, it encompasses fundamental variables governing penetration behavior. The selected parameters represent both geotechnical characteristics (UCS, BTS, BI) and geometric discontinuity properties (DPW, α), which are widely recognized as critical factors influencing TBM performance. The integration of these parameters enables the development of reliable predictive models for estimating penetration rate.

The variation of model input parameters and TBM penetration rate (ROP) along the tunnel alignment is illustrated in Figure 1, while the primary statistical characteristics of the dataset are summarized in Table 1. The data trends presented in Figure 1 indicate that the relationship between

 doi:10.20944/preprints202602.2047.v1

input variables and ROP is highly complex, suggesting that penetration rate cannot be adequately represented by a simple functional relationship. This complexity highlights the necessity of advanced modelling techniques capable of capturing nonlinear interactions among influencing parameters.

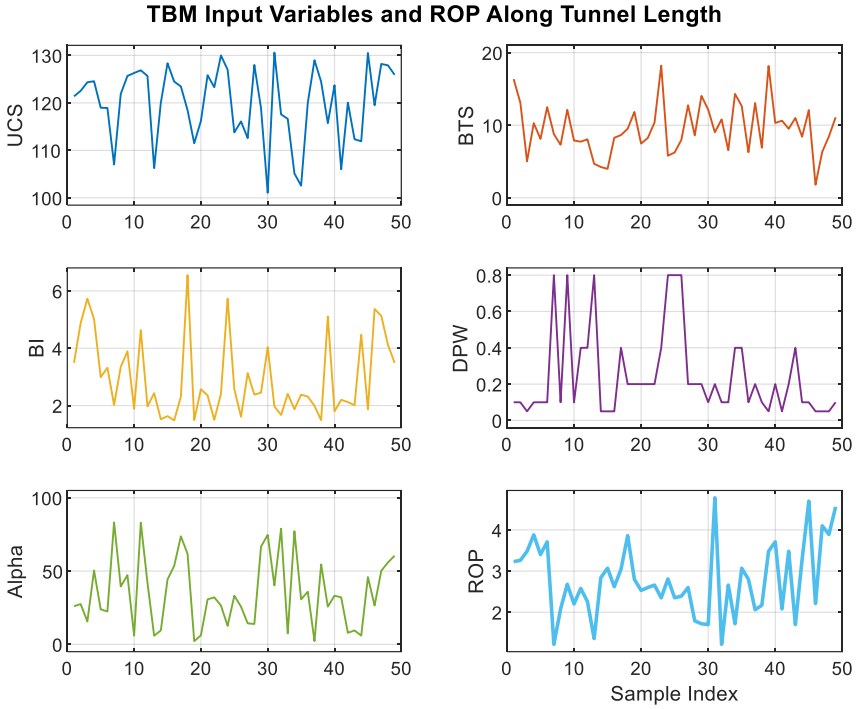


Figure 1. Variation of model variables and ROP values along the tunnel length.

Table 1. Descriptive statistics of the variables used in the study.

Row	Min	Max	Mean	Std
BI	1.48	6.55	2.96	1.38
UCS	101.07	130.59	119.94	7.72
Alpha	2.23	83.19	35.40	23.59
DPW	0.05	0.8	0.24	0.24
BTS	1.81	18.2	9.55	3.48
ROP	1.22	4.78	2.79	0.87

2.2. Selection of Model Input Parameters

In the development of prediction models, five primary input parameters influencing TBM performance were considered: brittleness index (BI), uniaxial compressive strength (UCS), discontinuity plane spacing (DPW), discontinuity orientation (α), and Brazilian tensile strength (BTS). The Tunnel Boring Machine penetration rate (Rate of Penetration, ROP) was selected as the model output variable. The relationships between input variables and ROP, as well as the correlations among model parameters, are illustrated in Figure 2 using a correlation matrix.

As shown in Figure 2, the strongest positive correlation with ROP is observed for UCS ($r = 0.450$), indicating that rock compressive strength plays a dominant role in controlling penetration rate. The brittleness index (BI) demonstrates a moderate positive relationship with ROP ($r = 0.339$), suggesting that rock fracture characteristics contribute to TBM cutting efficiency. In contrast, discontinuity plane spacing (DPW) exhibits a moderate negative correlation with ROP ($r = –0.364$), implying that wider discontinuity spacing may reduce rock mass breakability and consequently decrease penetration efficiency.

 doi:10.20944/preprints202602.2047.v1

Discontinuity orientation (α) and Brazilian tensile strength (BTS) display relatively weak correlations with ROP (r = –0.044 and r = 0.055, respectively). However, despite their limited individual linear correlations, these parameters were retained in the modelling framework because TBM penetration rate is governed by complex and multivariate interactions. Previous studies have demonstrated that parameters with low pairwise correlation may still contribute significantly to predictive performance when nonlinear modelling approaches are employed.

Overall, UCS appears to be the most influential parameter controlling penetration rate, while DPW and BI provide additional explanatory capability and improve model robustness.

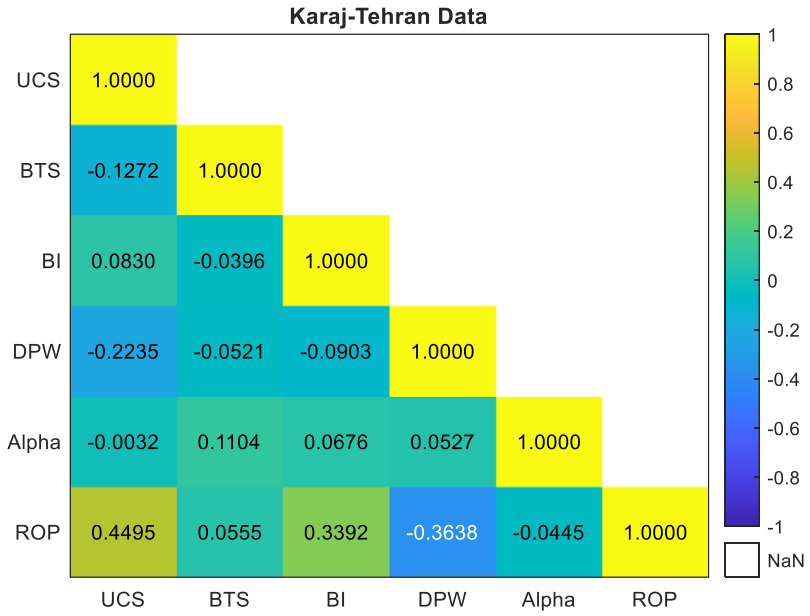


Figure 2. The correlation relationship between ROP and model parameters.

2.3. Machine Learning Methods Used in the Study

After determining the influence of input parameters, both widely used and relatively underexplored machine learning techniques were evaluated for predicting TBM penetration rate (ROP). In this context, three commonly applied machine learning methods—Random Forest (RF), Bagged Trees, and Support Vector Machines (SVM)—were implemented together with the Least Squares Boosting (LSBoost) algorithm, which is applied to TBM penetration rate prediction for the first time in this study.

2.3.1. Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and produces predictions by averaging the outputs of individual trees [34,35]. RF is capable of capturing nonlinear relationships within datasets and improving generalization performance.

The basic prediction formulation is expressed as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \quad (1)$$

where $h_t(x)$represents the prediction of the t-th decision tree and Tdenotes the total number of trees.

2.3.2. Bagged Trees (BT)

The Bagged Trees method involves generating multiple decision trees through bootstrap sampling and averaging their predictions to produce the final output [36]. This approach reduces the

 doi:10.20944/preprints202602.2047.v1

risk of overfitting and enhances model stability. The prediction formulation is similar to that of Random Forest and can be expressed as:

$$\hat{y}_{bag} = \frac{1}{B}\sum_{b=1}^{B} h_b^*(x) \quad (2)$$

where $h_b^*(x)$ represents each bootstrap-trained tree and B denotes the total number of trees.

2.3.3. Support Vector Machines (SVM)

Support Vector Machines are widely recognized for their strong performance, particularly in small- to medium-sized datasets, and are frequently employed in ROP prediction to identify linear or nonlinear decision boundaries [37]. The SVM regression model is defined through the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}\xi_i \quad (3a)$$

subject to:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (3b)$$

where w represents the weight vector, b is the bias term, ξ_i denotes slack variables accounting for prediction errors, and C is the penalty parameter controlling model complexity.

2.3.4. Least Squares Boosting (LSBoost)

LSBoost is an ensemble learning approach that constructs a strong predictive model by sequentially minimizing the residual errors of weak learners [38,39]. In this study, LSBoost is applied to TBM penetration rate prediction for the first time. The fundamental update step is defined as:

$$f_m(x) = f_{m-1}(x) + \nu \cdot h_m(x), m = 1,2,\ldots,M \quad (4)$$

where $h_m(x)$ represents the m-th weak learner, ν is the learning rate, and $f_m(x)$ denotes the updated prediction function. This iterative framework enables LSBoost to effectively capture complex nonlinear relationships within TBM datasets.

2.4. Model Performance Evaluation

To quantitatively assess the agreement between predicted and observed penetration rate values, several statistical performance metrics were employed. These indicators allow comprehensive evaluation of model accuracy, precision, and predictive capability. The performance metrics used in this study are described below.

2.4.1. Coefficient of Determination (R^2)

The coefficient of determination represents the proportion of variance in observed data explained by model predictions and is widely used to evaluate overall model fitness. The R^2 value ranges between 0 and 1, where values closer to 1 indicate better predictive performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

where y_i represents observed values, $\hat{y}_i$ represents predicted values, and $\bar{y}$ denotes the mean of observed values.

2.4.2. Root Mean Square Error (RMSE)

RMSE measures the magnitude of prediction errors by calculating the square root of the average squared differences between predicted and observed values. Lower RMSE values indicate higher prediction accuracy.

 doi:10.20944/preprints202602.2047.v1

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

2.4.3. Mean Absolute Error (MAE)

MAE represents the average absolute difference between predicted and observed values and is less sensitive to extreme outliers compared to RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (7)$$

These statistical indicators were used collectively to conduct a comparative evaluation of the predictive capabilities of the machine learning models. Higher R^2 values reflect a model's stronger ability to account for variability in the observed data, while lower RMSE and MAE values signify improved estimation accuracy. Taken together, this integrated assessment approach supports robust and dependable prediction of TBM penetration rate.

2.5. Model Training and Validation Strategy

To ensure the robustness and generalization capability of the developed machine learning models, a 10-fold cross-validation (CV) procedure was employed during the training phase. Cross-validation is widely accepted as an effective technique for evaluating model performance, particularly when working with limited datasets.

In the 10-fold CV approach, the entire dataset was randomly divided into ten equal subsets. During each iteration, nine subsets were used for model training, while the remaining subset was reserved for validation. This process was repeated ten times so that each subset was used once as validation data. The final model performance was obtained by averaging the results from all folds.

The use of 10-fold cross-validation reduces model bias and variance while providing a reliable estimate of prediction performance. This approach is particularly suitable for small-scale geotechnical datasets, where independent test datasets are often unavailable. Additionally, cross-validation improves model stability and minimizes the risk of overfitting by ensuring that all available data contribute to both training and validation processes.

3. Model Implementation and Results

3.1. Machine Learning Algorithms

Machine learning is a branch of artificial intelligence that seeks to mathematically characterize the behavior of complex systems and to derive generalizable patterns from historical data. Rather than relying on a predefined functional form, explicit equation, or fixed parametric structure, machine learning algorithms uncover the intrinsic patterns embedded within the data itself. Through this data-driven paradigm, computer systems acquire the capacity to learn from experience and make informed decisions without being explicitly programmed to solve a specific problem [40,41].,

Machine learning methods are commonly categorized into supervised, unsupervised, and semi-supervised learning frameworks, and they can be applied to a wide spectrum of problem types, ranging from regression and classification to clustering and dimensionality reduction. Owing to their flexible model architectures, high predictive accuracy, and relatively modest reliance on strict statistical assumptions compared to conventional approaches, machine learning techniques have found extensive applications across engineering, hydrology, medical sciences, and financial analytics [42–44].

In the present study, a comparative modeling framework was implemented to evaluate the effectiveness of the proposed method in predicting the penetration rate (ROP) of a Tunnel Boring Machine (TBM). Using the same dataset, its performance was benchmarked against widely adopted

 doi:10.20944/preprints202602.2047.v1

machine learning algorithms, including Random Forest (RF) [35,45], Bagged Trees (BT) [36], Support Vector Regression (SVR) [37], and Least Squares Boosting (LSBoost) [38,39].

Table 2 presents a comparative summary of the predictive performance of these machine learning models for TBM penetration rate estimation, based on both the original dataset and the Z-score normalized dataset. Model performance was evaluated using the correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

Table 2. Performance comparison of machine learning methods used for ROP prediction.

Scenario	Model	R	R2	RMSE	MAE
Original	RF	0.873	0.622	0.529	0.427
	BT	0.877	0.632	0.522	0.419
	SVM	0.754	0.554	0.575	0.395
	LSBoost	0.965	**0.909**	**0.260**	0.220
Z-score	RF	0.869	0.609	0.538	0.428
	BT	0.880	0.643	0.514	0.410
	SVM	0.756	0.557	0.573	0.392
	LSBoost	0.965	**0.909**	**0.260**	0.220

As shown in Table 2, the LSBoost model produced the most accurate results across all performance metrics when the original dataset was considered. The obtained values of R = 0.965 and R^2 = 0.909 indicate a very strong agreement between observed and predicted TBM penetration rate (ROP) values. In addition, the low error metrics (RMSE = 0.260 and MAE = 0.220) demonstrate the ability of the LSBoost algorithm to effectively capture the complex and nonlinear relationships governing TBM performance.

The Random Forest (RF) and Bagged Trees (BT) models exhibited comparable predictive performance, with correlation coefficients of approximately R ≈ 0.87 and coefficients of determination in the range of R^2 ≈ 0.62–0.63. These results confirm that tree-based ensemble methods provide reasonable predictive capability for ROP estimation; however, their performance remains noticeably inferior to that of the LSBoost model.

In contrast, the Support Vector Regression (SVR) model yielded lower predictive accuracy on the original dataset (R = 0.754, R^2 = 0.554) compared to the ensemble-based approaches. Nevertheless, an improvement in SVR performance was observed following Z-score normalization (R = 0.775, R^2 = 0.581), indicating that this method is sensitive to data scaling and hyperparameter configuration. Previous studies have reported that the predictive performance of SVM-based models can be significantly enhanced through careful optimization of kernel functions and penalty parameters (C) [25]. In this respect, the limited hyperparameter tuning adopted in the present study may not fully reflect the maximum potential of the SVR model.

While Z-score normalization had a marginal effect on the performance of the RF and BT models, the unchanged performance metrics of the LSBoost model before and after normalization highlight its strong robustness against data scaling. This stability suggests that LSBoost is less sensitive to variations in input magnitude and is capable of maintaining consistent predictive accuracy under different preprocessing conditions.

Overall, the results indicate that the LSBoost algorithm provides the highest accuracy and reliability for TBM penetration rate prediction across both original and Z-score normalized datasets. The predicted ROP values obtained from the LSBoost model were compared with measured values. Figure 3a presents the agreement between predicted and observed ROP values using scatter plots, while Figure 3b illustrates the variation of penetration rate along the tunnel alignment.

 doi:10.20944/preprints202602.2047.v1

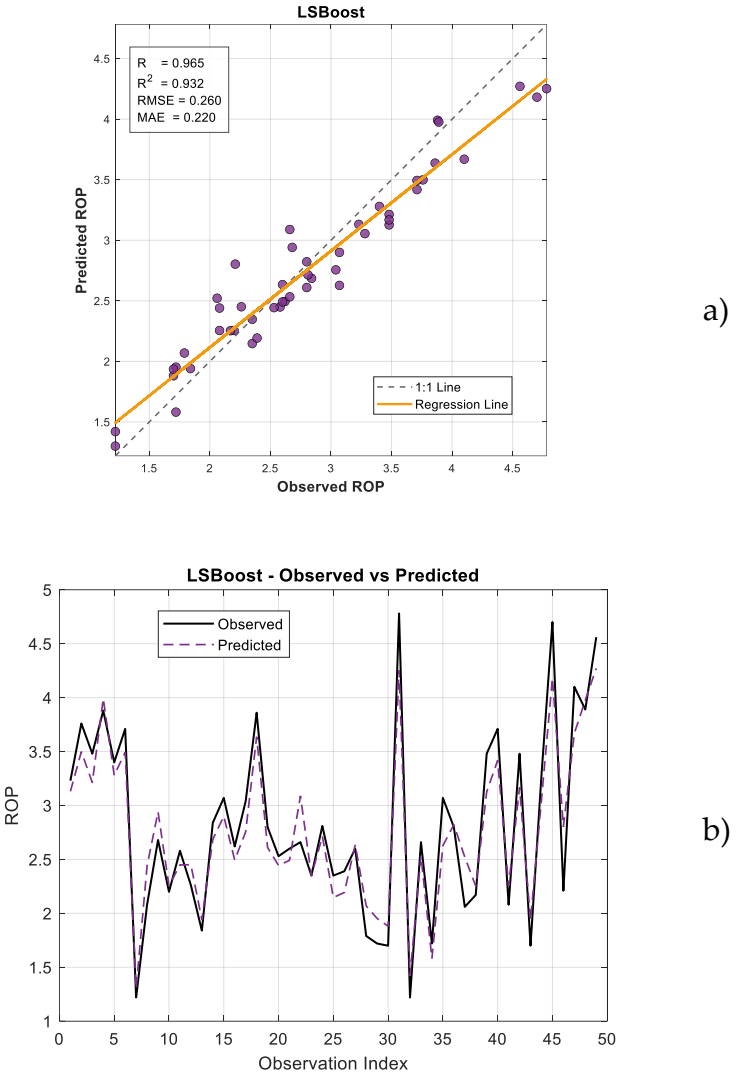


Figure 3. Comparison of ROP prediction results obtained using the LSBoost model: (a) Scatter plot of observed versus predicted values, (b) Variation of observed and predicted ROP along the tunnel length.

The consistently high and stable performance values observed in both scenarios indicate that the LSBoost model is robust against data scaling and is capable of successfully capturing the complex relationships affecting TBM performance. Therefore, LSBoost was selected as the most suitable method for ROP prediction in the subsequent stages of this study. Moreover, even the lowest performance value reported in Table 2 exceeds the prediction accuracy of the best-performing empirical formula proposed in the original study from which the dataset was derived [2]. This finding demonstrates that not only LSBoost but also the machine learning–based approaches employed in this study provide higher accuracy and better generalization capability in TBM penetration rate prediction compared with conventional empirical models.

3.2. Feature Importance Analysis

Feature importance analysis is one of the key explainability approaches used to identify the extent to which each input contributes to the prediction of a target variable by machine learning models. This analysis enables the determination of the relative weight of variables that drive model performance [46–51]. Variable importance metrics are particularly prevalent in tree-based ensemble learning algorithms, where they are commonly applied as part of in-model evaluation [25,49,50].

 doi:10.20944/preprints202602.2047.v1

In the present study, the contributions of geomechanical parameters to ROP prediction were assessed across four different machine learning methods within the TBM modeling framework. Analyses were conducted using both the raw (original) dataset and the Z-score standardized dataset, allowing for an evaluation of the effect of data scaling on variable importance distributions. The intrinsic model importance values are summarized in Table 3.

Table 3. Relative Importance of Input Variables in ROP Prediction Models.

Scenario	Row	BI	UCS	Alpha	DPW	BTS
Original	RF	17.52	21.90	13.01	25.68	21.88
	BT	20.01	22.63	13.11	24.41	19.84
	SVM	28.71	22.79	19.95	14.77	13.79
	LSBoost	16.24	15.43	12.78	**36.43**	19.11
Z-score	RF	21.54	23.06	12.45	23.13	19.82
	BT	17.00	24.21	12.42	28.60	17.77
	SVM	28.71	22.78	19.94	14.77	13.79
	LSBoost	16.24	15.43	12.78	**36.43**	19.11

Examining Table 3, it is evident that in tree-based models (RF, BT, and LSBoost), the DPW and UCS variables exhibit high importance values. In the LSBoost model, DPW demonstrates the highest importance across both data scenarios, whereas in the RF and BT models, UCS emerges as the most influential variable. In the SVM model, the BI variable attains the highest importance in both the original and normalized datasets. The Alpha variable shows relatively lower contribution across all models. Comparing the original and normalized datasets, only minor changes in importance distribution are observed for RF and BT, while importance values remain unchanged for the SVM and LSBoost models.

In addition to intrinsic model importance metrics, variable contributions were further assessed using Jacobian-based numerical sensitivity analysis [33] and the SHAP (Shapley Additive exPlanations) method. SHAP is based on the concept of Shapley values from cooperative game theory [48]. Originally introduced by Lloyd Shapley, the concept was later adapted and extended to machine learning models by Scott Lundberg and Su-In Lee [49]. SHAP provides a unified and consistent framework for local interpretability of models and is currently recognized as a cornerstone methodology in the field of Explainable Artificial Intelligence (XAI). Numerous interpretability techniques in the literature increasingly build upon this approach.

Recently, Karahan and Alkaya [33] proposed a Jacobian-based, derivative-driven model to quantify the effects of individual variables on TBM penetration rate, applying this model to both parametric relationships and four different machine learning methods. In the proposed approach, analytical partial derivatives were employed for parametric models, while numerical partial derivatives were used for the machine learning models. The study reported that the relative importance weights of input variables exhibited similar trends across both parametric and machine learning–based methods.

In the present study, the Jacobian-based method proposed by [33] was applied in a comparative framework alongside the SHAP approach, with the resulting findings presented in Table 4 and Figure 4.,

Table 4. Comparison of Jacbian-Based Sensitivity and SHAP Results.

Future Importance (%)	ML Method	BI	UCS	Alpha	DPW	BTS
Jacobian-Based	RF	16.52	29.07	10.47	27.07	16.87

 doi:10.20944/preprints202602.2047.v1

Sensitivity	BT	15.45	29.27	10.16	27.49	17.63
	SVM	28.57	21.21	19.77	18.52	11.93
	LSBoost	15.66	27.81	12.78	24.72	19.03
SHAP	RF	20.53	25.48	11.36	20.79	21.84
	BT	20.09	25.57	10.50	21.08	22.75
	SVM	32.06	20.34	18.48	18.00	11.12
	LSBoost	21.89	21.13	12.32	22.13	22.54

Examining Table 4, it is evident that both the Jacobian-based sensitivity analysis and the SHAP method reveal largely similar trends in identifying influential variables affecting TBM penetration rate. In both approaches, UCS and DPW exhibit high importance across all models, indicating that rock strength and discontinuity/fracture characteristics play a decisive role in controlling penetration rate.

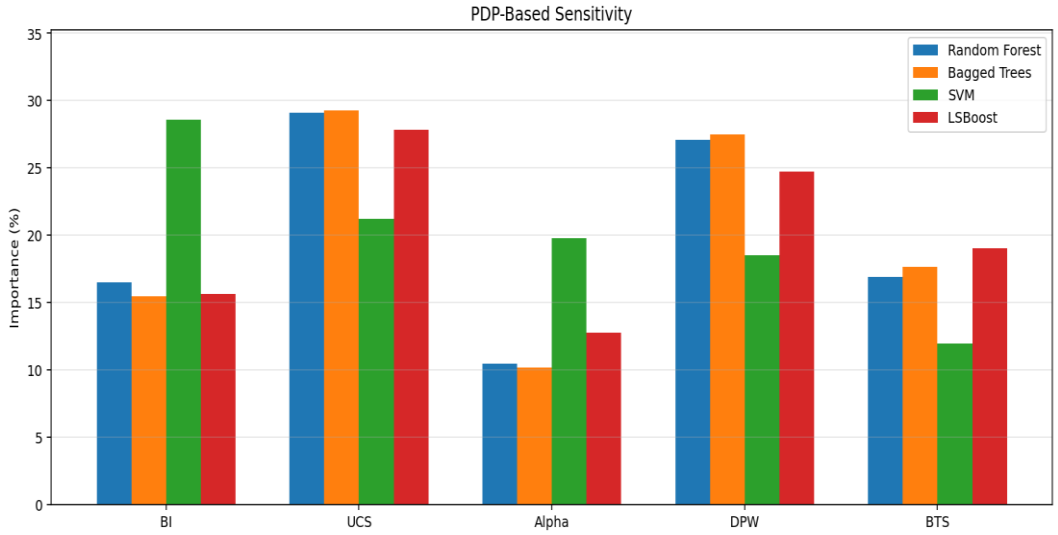
The BI and BTS variables generally display moderate importance, with SHAP analyses highlighting slightly higher contributions for these parameters. This finding suggests that the influence of specific variables on the model can be reflected to varying degrees depending on the methodological approach employed.

For the RF and BT models, both methods produced similar variable rankings, whereas the SVM model exhibited a distinct profile. In the Jacobian analysis, BI contributed most strongly for SVM, while SHAP results emphasized the prominence of Future Importance and UCS. For the other variables (ALFA, DPW, and BTS), SVM importance values were generally lower or ranked differently compared with LSBoost and other ensemble models. This indicates that variable contribution distributions in SVM differ from those in other models, with sensitivity to certain inputs manifesting differently.

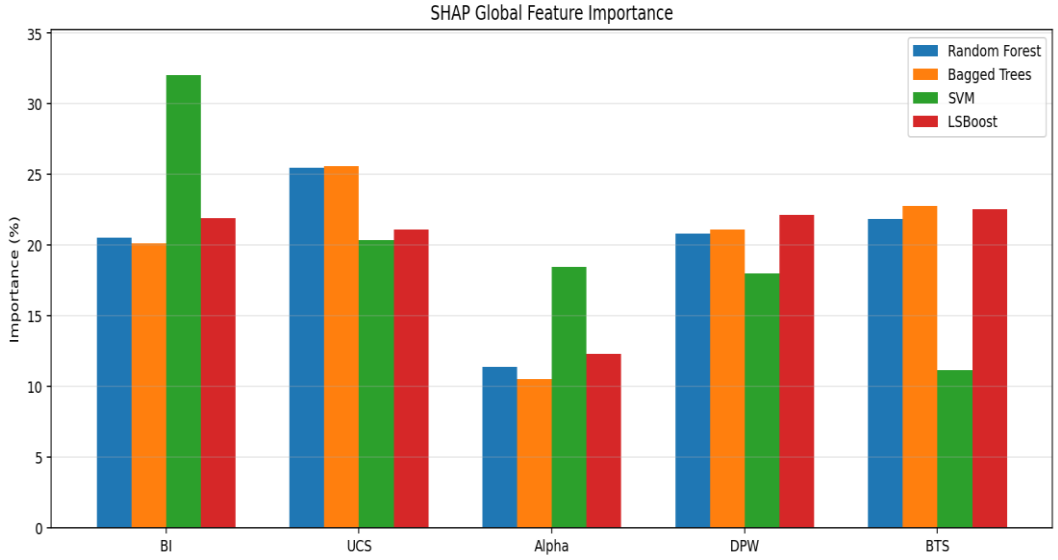
Stability analysis of variables in machine learning models provides insight not only into overall fit across the dataset but also into the model's behavior across different subsets. Such analyses help confirm the consistency of variable contributions, enhancing model reliability and generalizability, while also informing more deliberate learning strategies and hyperparameter selection. This behavior is visualized in Figure 5. Consistent with the findings reported in Table 4 and Figures 4 and 5, it is apparent that the SVM model exhibits a markedly different profile compared with other models, with a more variable learning pattern along the tunnel.

Overall, when Table 4 and Figures 4 and 5 are considered together, it is clear that the Jacobian-based numerical sensitivity analysis shows trends comparable to SHAP and indicates that UCS, DPW, and BI may exert greater influence on TBM penetration rate relative to other parameters. These results support the notion that employing the Jacobian-based approach alongside SHAP in a comparative framework can provide valuable and complementary insights for model interpretation.

 doi:10.20944/preprints202602.2047.v1



a)



b)

Figure 4. Comparison of Jacobian-Based PDP and SHAP Feature Importances for Different ML Models: a) PDP b) SHAP

 doi:10.20944/preprints202602.2047.v1

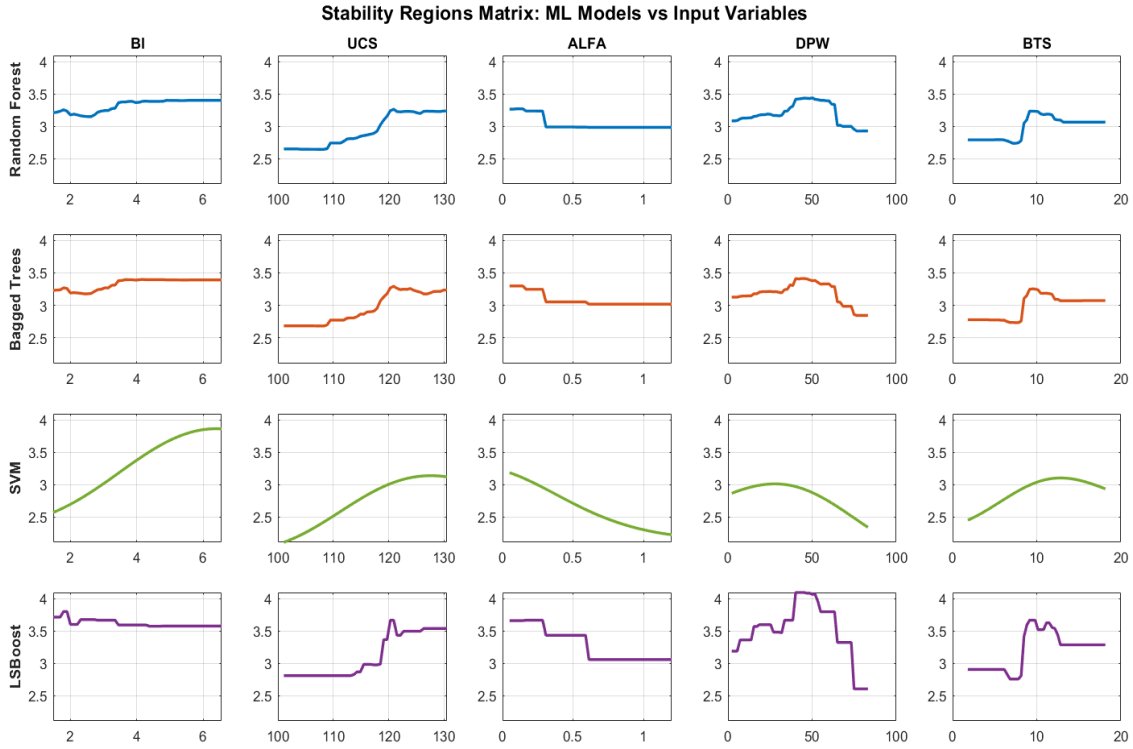


Figure 5. Model-wise sensitivity profiles of input variables for different ML algorithms.

4. Discussion

In this study, the predictive performance of Random Forest, Bagged Trees, Support Vector Machine (SVM), and Least Squares Boosting (LSBoost) algorithms was compared for estimating TBM penetration rate (ROP), and model outcomes were analyzed in the context of rock mechanics parameters. The findings clearly demonstrate that ROP is governed by nonlinear interactions among multiple geomechanical parameters. Both intrinsic model importance metrics and explainability-based analyses such as SHAP indicated that UCS, BI, and DPW play dominant roles in controlling ROP. This consistency supports the interpretation that these parameters are not only statistically significant but also represent the physical determinants of the TBM excavation mechanism.

Analyses were conducted on both the raw (original) dataset and the Z-score standardized dataset. This approach allowed for the assessment of the effect of data scaling on variable importance distributions and model performance. Comparisons between the original and normalized datasets revealed that data scaling had a limited impact on intrinsic variable importance. Minor changes were observed in the Random Forest and Bagged Trees models, yet the overall ranking of variables remained largely unchanged. In contrast, importance values for SVM and LSBoost were unaffected, indicating that these algorithms exhibit relative robustness to data scaling.

Regarding model performance, LSBoost achieved the highest predictive accuracy across both the original and normalized datasets (original data: $R^2 = 0.909$, RMSE = 0.260, MAE = 0.220). Random Forest and Bagged Trees showed similar accuracy levels, with the Bagged Trees model exhibiting only slight improvement after normalization. SVM demonstrated lower predictive accuracy relative to the other algorithms, suggesting that under limited data and parameter configurations, it may not fully capture complex parameter interactions. Conversely, LSBoost consistently produced reliable predictions even with small and heterogeneous datasets, highlighting its applicability in field studies with limited data availability.

Model-based variable importance analyses revealed that the parameters exerting the greatest influence on TBM penetration vary depending on the modeling approach. In the LSBoost model, DPW emerged as the most critical parameter (36.43%), whereas in Random Forest and Bagged Trees, UCS and DPW jointly dominated. In the SVM model, BI accounted for the highest importance

 doi:10.20944/preprints202602.2047.v1

(28.71%). These results indicate that TBM performance cannot be explained by a single rock parameter, but rather by nonlinear interactions among rock strength, brittleness, and discontinuity characteristics.

Comparisons between Jacobian-based sensitivity analyses and SHAP results provide insights from complementary mathematical perspectives. The Jacobian approach highlights variables with high gradient effects by measuring derivative-based changes of model outputs with respect to inputs, whereas SHAP calculates average marginal contributions using Shapley values. SHAP's consideration of variable interactions leads to more balanced contribution distributions, particularly in tree-based and boosting models. For the SVM model, the two methods exhibited higher consistency, while LSBoost showed strong agreement between SHAP and Jacobian results, enhancing sensitivity to local patterns and interpretability.

From an engineering perspective, these findings emphasize the criticality of jointly considering UCS, BI, and DPW in TBM performance prediction. The consistent prominence of these parameters across different models and explainability approaches highlights them as key variables for design and performance optimization. Integrating intrinsic model importance with explainability-based methods enables a more comprehensive and reliable interpretation of TBM penetration performance. While the Jacobian analysis elucidates the model's mathematical sensitivity structure, SHAP presents contributions in an interaction-aware and theoretically grounded framework. This integrated approach supports interpretable and dependable outcomes in machine learning–based geomechanical modeling.

5. Conclusion

This study conducted a comparative evaluation of Random Forest, Bagged Trees, Support Vector Machine (SVM), and Least Squares Boosting (LSBoost) algorithms for predicting TBM penetration rate (ROP). The findings demonstrate that data-driven machine learning approaches can provide effective and reliable solutions for modeling TBM performance.

Performance analyses indicated that LSBoost outperformed the other algorithms, achieving $R = 0.965$ and $R^2 = 0.909$. It also produced the lowest RMSE and MAE values, successfully capturing the nonlinear interactions among parameters influencing ROP, and generating reliable predictions even in small and heterogeneous datasets. Random Forest and Bagged Trees showed comparable performance, with Bagged Trees exhibiting limited improvement after Z-score normalization, indicating partial sensitivity to data scaling. SVM produced lower correlation and explanatory power but offered smoother and continuous prediction curves, suggesting higher generalization capacity while being less capable of capturing local variations.

Feature importance analyses and SHAP results revealed that TBM penetration rate is predominantly governed by discontinuity spacing (DPW) and uniaxial compressive strength (UCS). In the LSBoost model, DPW had the highest importance, highlighting the decisive role of rock mass structural properties on the TBM cutting mechanism. In Random Forest and Bagged Trees, UCS and DPW jointly dominated, supporting the combined influence of rock strength and discontinuity geometry on TBM performance. In the SVM model, the brittleness index (BI) emerged as the most important parameter, emphasizing the impact of rock fracture behavior on penetration performance. SHAP analyses clearly quantified each variable's contribution to individual predictions, confirming DPW and UCS as critical parameters at both global and local scales.,

The effect of data normalization on model performance was limited; however, in variable importance analyses, parameters with narrow value ranges showed noticeable changes. Notably, LSBoost maintained consistent and reliable performance even with small, heterogeneous datasets, reinforcing its applicability in field studies with limited data.

Overall, predicting TBM penetration rate constitutes a multivariate, nonlinear, and complex engineering problem. This study demonstrates that machine learning–based approaches are powerful tools for TBM performance prediction, with LSBoost providing high accuracy and reliability even in data-constrained field applications. Moreover, SHAP and feature importance

 doi:10.20944/preprints202602.2047.v1

analyses render model predictions interpretable from an engineering perspective, providing actionable insights for TBM design and operational decision-making. Another key finding is the strong agreement between derivative-based Jacobian methods and SHAP analyses, confirming the internal consistency and reliability of the model's reasoning.

References

1. Yagiz, S. (2008). Utilizing rock mass properties for predicting TBM performance in hard rock condition. *Tunnelling and underground space technology*, *23*(3), 326-339.
2. Mansouri, M., & Moomivand, H. (2010). Influence of rock mass properties on TBM penetration rate in Karaj-Tehran water conveyance tunnel. *J Geol Min Res*, *2*(5), 114-121.
3. Yagiz, S., & Karahan, H. (2011). Prediction of hard rock TBM penetration rate using particle swarm optimization. *International Journal of Rock Mechanics and Mining Sciences*, *48*(3), 427-433.
4. Yagiz, S., & Karahan, H. (2015). Application of various optimization techniques and comparison of their performances for predicting TBM penetration rate in rock mass. *International Journal of*
5. Khoshzaher, E.; Chakeri, H.; Bazargan, S.; Mousapour, H. The prediction of EPB-TBM performance using firefly algorithms and particle swarm optimization. *Rud.-Geološko-Naft. Zb.* 2023, *38*, 79–86.
6. Benardos, A.G.; Kaliampakos, D.C. Modelling TBM performance with artificial neural networks. *Tunn. Undergr. Space Technol.* **2004**, *19*, 597–605.
7. Benardos, A. Artificial intelligence in underground development: A study of TBM performance. *Undergr. Spaces* **2008**, *102*, 21–32.
8. Javad, G.; Narges, T. Application of artificial neural networks to the prediction of tunnel boring machine penetration rate. *Min. Sci. Technol.* **2010**, *20*, 727–733.
9. Torabi, S.R.; Shirazi, H.; Hajali, H.; Monjezi, M. Study of the influence of geotechnical parameters on the TBM performance in Tehran–Shomal highway project using ANN and SPSS. *Arab. J. Geosci.* **2013**, *6*, 1215–1227.
10. Jung, J. H., Chung, H., Kwon, Y. S., & Lee, I. M. (2019). An ANN to predict ground condition ahead of tunnel face using TBM operational data. *KSCE Journal of Civil Engineering*, *23*(7), 3200-3206.
11. Nikakhtar, L., Zare, S., Nasirabad, H. M., & Ferdosi, B. (2022). Application of ANN-PSO algorithm based on FDM numerical modelling for back analysis of EPB TBM tunneling parameters. *European Journal of Environmental and Civil Engineering*, *26*(8), 3169-3186.
12. Fattahi, H.; Babanouri, N. Applying optimized support vector regression models for prediction of tunnel boring machine performance. *Geotech. Geol. Eng.* **2017**, *35*, 2205–2217.
13. Liu, B.; Wang, R.; Guan, Z.; Li, J.; Xu, Z.; Guo, X.; Wang, Y. Improved support vector regression models for predicting rock mass parameters using tunnel boring machine driving data. *Tunn. Undergr. Space Technol.* **2019**, *91*, 102958.
14. Koopialipoor, M.; Fahimifar, A.; Ghaleini, E.N.; Momenzadeh, M.; Armaghani, D.J. Development of a new hybrid ANN for solving a geotechnical problem related to tunnel boring machine performance. *Eng. Comput.* **2020**, *36*, 345–357.
15. Afradi, A.; Ebrahimabadi, A.; Hallajian, T. Prediction of TBM penetration rate using support vector machine. *GEOSABERES Rev. De Estud. Geoeducacionais* **2020**, *11*, 467–479.
16. [16] Liu, Y., Huang, S., Wang, D., Zhu, G., & Zhang, D. (2022). Prediction model of tunnel boring machine disc cutter replacement using kernel support vector machine. *Applied Sciences*, *12*(5), 2267.
17. Yazdani-Chamzini, A.; Yakhchali, S.H. Tunnel Boring Machine (TBM) selection using fuzzy multicriteria decision making methods. *Tunn. Undergr. Space Technol.* **2012**, *30*, 194–204.
18. Ghasemi, E.; Yagiz, S.; Ataei, M. Predicting penetration rate of hard rock tunnel boring machine using fuzzy logic. *Bull. Eng. Geol. Environ.* **2014**, *73*, 23–35.
19. Minh, V.T.; Katushin, D.; Antonov, M.; Veinthal, R. Regression models and fuzzy logic prediction of TBM penetration rate. *Open Eng.* **2017**, *7*, 60–68.
20. Afradi, A.; Ebrahimabadi, A.; Hallajian, T. Prediction of TBM penetration rate using fuzzy logic, particle swarm optimization and harmony search algorithm. *Geotech. Geol. Eng.* **2022**, *40*, 1513–1536.

 doi:10.20944/preprints202602.2047.v1

21. Armaghani, D.J.; Mohamad, E.T.; Narayanasamy, M.S.; Narita, N.; Yagiz, S. Development of hybrid intelligent models for predicting TBM penetration rate in hard rock condition. *Tunn. Undergr. Space Technol.* **2017**, *63*, 29–43.
22. Wang, K.; Wu, X.; Zhang, L.; Song, X. Data-driven multi-step robust prediction of TBM attitude using a hybrid deep learning approach. *Adv. Eng. Inform.* **2023**, *55*, 101854.
23. Yu, S.; Zhang, Z.; Wang, S.; Huang, X.; Lei, Q. A performance-based hybrid deep learning model for predicting TBM advance rate using attention-ResNet-LSTM. *J. Rock Mech. Geotech. Eng.* **2024**, *16*, 65–80.
24. Yao, M.; Li, X.; Pang, Y.E.; Wang, Y. Prediction model of TBM response parameters based on a hybrid drive of knowledge and data. *Tunn. Undergr. Space Technol.* **2025**, *161*, 106598.
25. Karahan, H., & Alkaya, D. (2026). Integrating SVR Optimization and Machine Learning-Based Feature Importance for TBM Penetration Rate Prediction. *Applied Sciences*, *16*(1), 355. https://doi.org/10.3390/app16010355
26. Ghorbani, E., & Yagiz, S. (2024). Estimating the penetration rate of tunnel boring machines via gradient boosting algorithms. *Engineering Applications of Artificial Intelligence*, *136*, 108985.
27. Xu, H., Zhou, J., Asteris, P. G., Armaghani, D. J., & Tahir, M. M. (2019). *Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate*. *Applied Sciences*, **9**(18), 3715.
28. Kong, X., Ling, X., Tang, L., Tang, W., & Zhang, Y. (2022). Random forest-based predictors for driving forces of earth pressure balance (EPB) shield tunnel boring machine (TBM). *Tunnelling and underground space technology*, *122*, 104373.
29. Kim, D., Pham, K., Oh, J. Y., Lee, S. J., & Choi, H. (2022). Classification of surface settlement levels induced by TBM driving in urban areas using random forest with data-driven feature selection. *Automation in Construction*, *135*, 104109.
30. Hou, S., Liu, Y., Zhuang, W., Zhang, K., Zhang, R., & Yang, Q. (2023). Prediction of shield jamming risk for double-shield TBM tunnels based on numerical samples and random forest classifier. *Acta Geotechnica*, *18*(1), 495-517.
31. Yang, H., Liu, X., & Song, K. (2022). A novel gradient boosting regression tree technique optimized by improved sparrow search algorithm for predicting TBM penetration rate. *Arabian Journal of Geosciences*, *15*(6), 461.
32. Zhou, J., Qiu, Y., Zhu, S., Armaghani, D. J., Li, C., Nguyen, H., & Yagiz, S. (2021). Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Engineering Applications of Artificial Intelligence, 97, 104015.
33. Karahan, H., & Alkaya, D. (2026). Input Variable Effects on TBM Penetration Rate: Parametric and Machine Learning Models. *Applied Sciences*, *16*(3), 1301. https://doi.org/10.3390/app16031301
34. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
35. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
36. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.
37. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.
38. Hastie, T., Tibshirani, R., & Friedman, J. (2008). Boosting and additive trees. In *The elements of statistical learning: data mining, inference, and prediction* (pp. 337-387). New York, NY: Springer New York.
39. Alatefi, S., & Almeshal, A. M. (2021). A new model for estimation of bubble point pressure using a bayesian optimized least square gradient boosting ensemble. *Energies*, *14*(9), 2653.
40. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, UK, 2016.
41. Bahri, M.; Romero-Hernández, R.; Mascort-Albea, E.J.; Soriano-Cuesta, C.; Jaramillo-Morilla, A. Predicting Maximum Surface Displacement from Mechanized Twin Tunnel Excavation in Seville Using Machine Learning and FLAC3D Simulation. Geotech. Geol. Eng. 2025, 43, 70.
42. Karahan, H.; Iplikci, S.; Yasar, M.; Gurarslan, G. River flow estimation from upstream flow records using support vector machines. *J. Appl. Math.* **2014**, *2014*, 714213.,
43. Karahan, H.; Erkan Can, M. A Novel Method to Forecast Nitrate Concentration Levels in Irrigation Areas for Sustainable Agriculture. *Agriculture* **2025**, *15*, 161.

 doi:10.20944/preprints202602.2047.v1

44. Karahan, H.; Cetin, M.; Can, M.E.; Alsenjar, O. Developing a new ANN model to estimate daily actual evapotranspiration using limited climatic data and remote sensing techniques for sustainable water management. *Sustainability* **2024**, *16*, 2481.
45. Breiman, L.; Cutler, A.; Liaw, A.; Wiener, M.; Liaw, M.A. Package 'Randomforest'; University of California, Berkeley: Berkeley, CA, USA, 2018; Volume 81, pp. 1–29.
46. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning.
47. Guyon,I.; Elisseeff, A. An introduction to variable and feature selection. J. Mach. Learn. Res. 2003, 3, 1157–1182
48. Shapley, L. S. (1953). A value for n-person games.
49. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
50. Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, *74*, 851-886.
51. Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304.