



ARTICLE

Explainable Machine Learning Framework for Tunnel Boring Machine Penetration Rate Prediction

Halil Karahan* and Devrim Alkaya

Department of Civil Engineering, Pamukkale University, Denizli, Turkey

*Corresponding Author: Halil Karahan. Email: hkarahan@pau.edu.tr

Received: 09 May 2026; Accepted: 04 September 2026; Published: 21 October 2026

ABSTRACT: Accurate prediction of the tunnel boring machine (TBM) rate of penetration (ROP) is essential for improving excavation efficiency, project planning, and cost management. This study proposes an integrated machine learning and explainable artificial intelligence (XAI) framework for ROP prediction using Random Forest (RF), Least Squares Boosting (LSBoost), Bagged Trees (BT), and the Generalized Additive Model (GAM). To evaluate the robustness and applicability of the proposed framework, analyses were conducted using two independent geotechnical datasets from the New York Queens and Tehran–Karaj tunnel projects. Multicollinearity was addressed through Variance Inflation Factor (VIF)-based feature reduction, and model performance was evaluated using five-fold cross-validation under both full- and reduced-feature scenarios. The results demonstrate that ensemble learning models consistently provide high predictive accuracy and stable performance across both datasets. RF and LSBoost achieved the most reliable predictions, while BT also exhibited strong predictive capability with limited sensitivity to feature reduction. Explainability analyses based on SHapley Additive exPlanations (SHAP), Permutation Feature Importance (PFI), Partial Dependence Plots (PDP), and Local Interpretable Model-Agnostic Explanations (LIME) revealed mutually consistent interpretations of model behaviour. The dominant explanatory variables differed between datasets, with the rock brittleness index (BI) being the primary predictor in the New York Queens dataset, whereas Distance Between the Planes of Weakness (DPW), together with rock strength parameters, played a more influential role in the Tehran–Karaj dataset, highlighting the influence of geological conditions on ROP prediction. A key contribution of this study is the systematic investigation of the basis dimension (K) parameter in GAM. Although VIF-based feature reduction initially resulted in substantial performance degradation, sensitivity analyses demonstrated that increasing spline flexibility markedly improved predictive performance in both datasets. These findings indicate that appropriate specification of the basis dimension is critical for achieving reliable GAM performance and that the observed deterioration under reduced-feature scenarios primarily reflects insufficient model capacity rather than an inherent limitation of the GAM methodology. Overall, the proposed framework provides a reliable and interpretable approach for TBM ROP prediction. The results demonstrate the value of integrating machine learning with explainable artificial intelligence while highlighting the importance of hyperparameter optimization and independent validation for developing robust predictive models across different geotechnical conditions.

KEYWORDS: Tunnel boring machine (TBM); rate of penetration (ROP); ensemble learning; generalized additive model (GAM); explainable artificial intelligence (XAI); SHapley additive exPlanations (SHAP); feature selection; geotechnical engineering

1 Introduction

In contemporary transportation, energy, and infrastructure projects, tunnels are extensively utilized to ensure continuity of intercity connectivity, support water supply and wastewater management systems, and enable the safe transmission of energy. Constructed under highly complex geotechnical conditions, these structures require advanced technical expertise, meticulous planning, and rigorous safety measures throughout the design, construction, and operational phases. In this respect, tunneling projects provide a powerful application domain that highlights the importance of interdisciplinary collaboration and multidimensional problem-solving skills in engineering education.

Tunnel Boring Machines (TBMs) constitute one of the most significant technological tools within this multidisciplinary framework. The prediction of TBM performance relies on the integrated knowledge of several disciplines, including civil engineering, geology, mechanical engineering, data science, and project management. Key operational parameters such as Rate of Penetration (ROP) and Advance Rate (AR) significantly influence the efficiency and safety of TBM operations. Therefore, accurate estimation of these parameters is critical not only for operational safety but also for cost and time management. Successful TBM operations require effective communication and coordinated collaboration among geologists, geotechnical engineers, mechanical engineers, data analysts, and field operators.

Early studies on TBM penetration rate prediction began with simple empirical models based on rock strength properties [1]. Subsequent research focusing on penetration behaviour in sedimentary rock formations [2] and statistical regression approaches [3] contributed to a better understanding of the relationship between rock properties and ROP. Advanced regression analyses reported relatively high correlation values ($r = 0.82$), leading to the development of empirical equations applicable in practical engineering contexts [3]. However, the predominantly linear structure of these early models limits their ability to capture the complex interactions between geotechnical conditions and machine-ground behaviour.

To overcome these limitations, heuristic optimization algorithms have been increasingly adopted for more realistic modeling of TBM performance. Initially, Yagiz and Karahan proposed a Particle Swarm Optimization (PSO)-based model aimed at minimizing the discrepancy between measured and predicted ROP values [4]. Subsequent studies expanded methodological diversity by employing algorithms such as Harmony Search (HS), Differential Evolution (DE), and Grey Wolf Optimizer (GWO) [5]. Khoshzaher et al. [6] demonstrated that TBM penetration rate is influenced by both machine-related and geological factors and showed, through a comparative study between Firefly and PSO algorithms, that the Firefly algorithm achieved superior performance. These advancements highlight the importance of integrating data analytics, optimization techniques, and computational intelligence within a multidisciplinary research framework for TBM performance prediction.

Accurate prediction of TBM advance rate is also critical for planning and optimization of tunneling projects. Machine learning (ML) methods have been increasingly employed to model complex and nonlinear relationships between TBM operational parameters and advance rate, offering flexible predictive capabilities for complex engineering problems [7–10]. The performance of ML models is strongly influenced by the selection of appropriate input variables, model architecture, and hyperparameters, which should be carefully optimized to achieve reliable predictive performance [11–13].

Fuzzy Logic approaches [14,15], hybrid soft computing techniques [16,17], and various machine learning methods [18–22] have been extensively applied in TBM performance prediction. Bo et al. [20] proposed a PSO-optimized XGBoost model to predict tunnel deformation using rock-machine interaction data obtained during TBM excavation. Wang et al. [21] employed an improved Binary PSO (BPSO) algorithm

with enhanced stochastic search capability and demonstrated that the BPSO-XGBoost model outperformed both PSO-XGBoost and BPSO-Random Forest approaches in terms of prediction accuracy.

Yang et al. [22] developed two models, Evolutionary Polynomial Regression (EPR) and Random Forest (RF), to predict TBM performance (advance rate and penetration index) using four geomechanical parameters. Comparative analysis showed that RF achieved higher predictive accuracy, whereas EPR provided a simpler and more interpretable formulation. Both models exhibited strong generalization capabilities.

Overall, the literature indicates that machine learning methods can effectively capture nonlinear relationships and handle uncertainty, leading to high predictive accuracy in TBM performance estimation. However, effective application of these methods and reliable interpretation of results require a solid understanding of traditional empirical models and regression-based approaches, as well as the influence of input variables on model outputs.

In this context, Karahan and Alkaya [13] developed an SVR-based model for TBM penetration rate prediction and optimized its hyperparameters using Bayesian Optimization (BO), Grid Search (GS), and Random Search (RS). The results showed that BO achieved higher accuracy with fewer evaluations and lower computational cost. In addition, feature importance analyses using Random Forest (RF), Bagged Trees (BT), Support Vector Machines (SVM), and Generalized Additive Models (GAM) revealed that BI, UCS, ALPHA, and DPW significantly influence ROP, thereby enhancing model interpretability.

In subsequent studies, the authors comparatively analyzed parametric models (M1–M6) and machine learning (ML) approaches for TBM penetration rate prediction [18]. The results indicated that the M6 model, which includes interaction terms, achieved the highest accuracy among parametric models, while the GAM provided the best predictive performance among ML methods. This study provides a significant contribution by jointly evaluating parametric and data-driven approaches together with the influence of explanatory variables on ROP.

Finally, in a study conducted on the Tehran–Karaj tunnel, Karahan and Alkaya [19] predicted TBM penetration rate using both parametric and machine learning models based on a small and heterogeneous dataset. SHAP and PDP analyses revealed that UCS, BI, and DPW were the most influential variables, demonstrating that reliable and interpretable predictions are achievable even with limited data.

In this study, the effect of multicollinearity and feature redundancy among input variables on model performance and feature importance distribution is systematically investigated within the framework of TBM (Tunnel Boring Machine) excavation processes and ROP prediction. The analysis is structured into two scenarios based on the Variance Inflation Factor (VIF). In the first scenario, all independent variables are included to establish a baseline modeling framework. In the second scenario, variables exhibiting high multicollinearity are removed to obtain a more parsimonious and less correlated dataset. This design enables a comparative assessment of the impact of variable dependency on model performance.

In the modeling stage, Random Forest (RF), Least Squares Boosting (LSBoost), Bagged Trees (BT), and Generalized Additive Models (GAM) are employed for both scenarios. Furthermore, SHAP, Permutation Feature Importance (PFI), Partial Dependence Plots (PDP), and LIME-based explainability techniques are integrated to enhance interpretability. This comprehensive framework enables a comparative evaluation of both predictive performance and variable contributions under different modeling strategies and data structures.

2 Materials and Methods

2.1 Dataset Description

Two datasets were used in this study. The first dataset was obtained from field and laboratory measurements conducted for a tunneling project located in the southwestern part of Queens, New York City, USA [3]. This dataset has been widely used in previous studies of TBM performance and rate of penetration (ROP) prediction [3–5]. It contains the principal geotechnical and discontinuity-related parameters required for modeling TBM performance, including Uniaxial Compressive Strength (UCS), Brazilian Tensile Strength (BTS), Brittleness Index (BI), Distance Between the Planes of Weakness (DPW), discontinuity orientation (Alpha), and ROP.

The second dataset was obtained from the Karaj–Tehran water conveyance tunnel project in Iran and was reported by Mansouri and Moomivand [23]. This independent dataset was used to assess whether the developed models could be effectively applied to an independent dataset from a different tunneling project and geological setting. Thus, the second dataset provides an additional evaluation of the generalizability and applicability of the developed models beyond the primary dataset.

For the first dataset, the variation of the model input variables and ROP along the tunnel alignment is presented in Fig. 1, while the statistical distributions of the same variables are illustrated using histogram plots in Fig. 2.

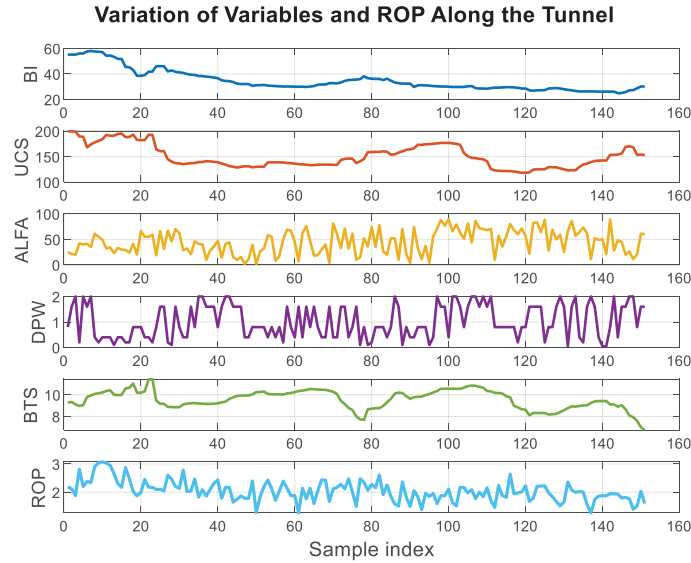


Figure 1: Variation of model variables and ROP values along the tunnel length.

The considered parameters include Uniaxial Compressive Strength (UCS), Brazilian Tensile Strength (BTS), Brittleness Index (BI), Distance Between the Planes of Weakness (DPW), and discontinuity orientation (Alpha). These variables constitute the primary input features for modeling TBM performance and estimating ROP.

The variation of model inputs and ROP along the tunnel alignment is presented in Fig. 1, while the statistical distributions of the same variables are illustrated through histogram plots in Fig. 2.

As illustrated in Fig. 1, the relationships between the input variables and ROP exhibit a highly complex and nonlinear structure. Similarly, the histograms presented in Fig. 2 indicate that the penetration rate cannot be adequately represented by a simple functional form or a univariate relationship. These observations

highlight the necessity of employing more advanced modeling approaches that can effectively capture nonlinear dependencies and interactions among variables.

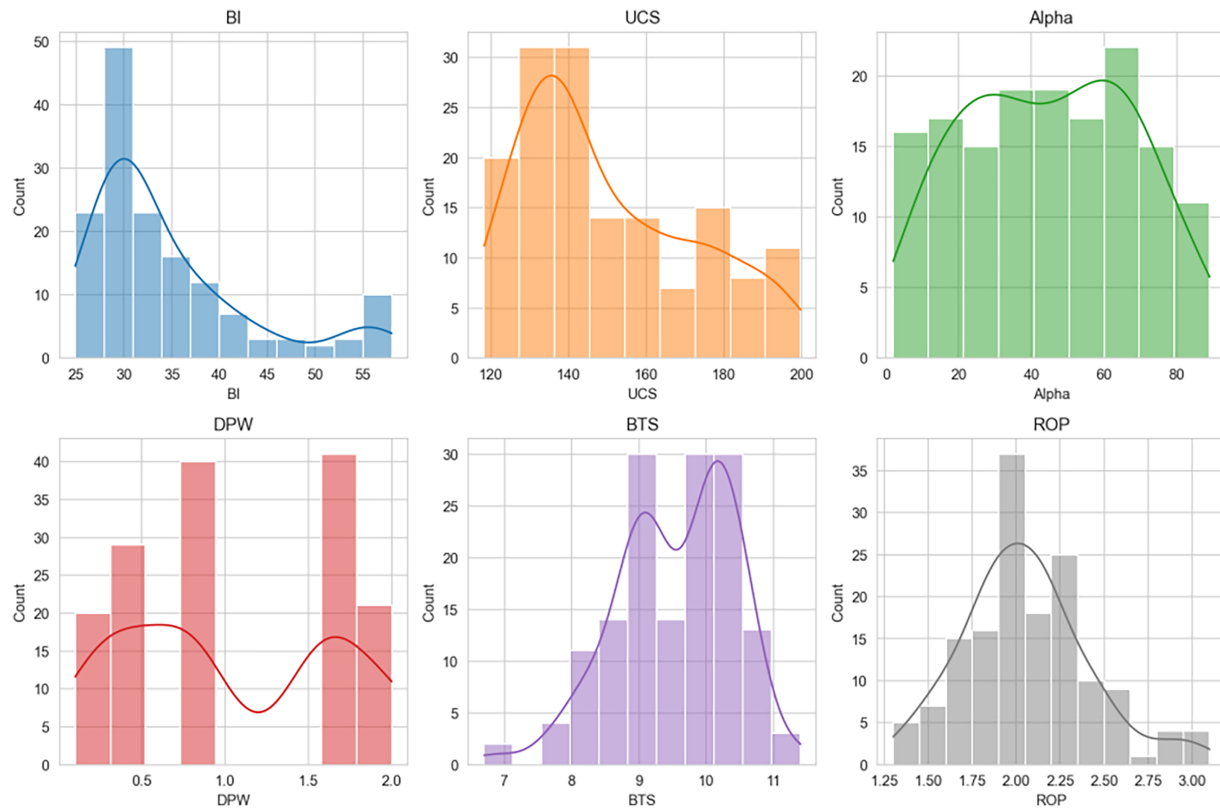


Figure 2: Histograms of model variables and ROP values.

Boxplot Analysis and Outliers

A boxplot analysis was conducted to comprehensively evaluate the distributional characteristics, variability, and potential outliers within the dataset. The obtained results are presented graphically in Fig. 3.

The boxplots presented in Fig. 3 illustrate the median, interquartile range, and minimum–maximum values of each variable, thereby providing a comprehensive overview of the overall data distribution. This analysis plays an important role in identifying potential outliers and supporting the data cleaning process. Since outliers may directly affect the predictive accuracy and generalization capability of machine learning models, they should be carefully considered during the modeling stage. In this sense, boxplot analysis provides an initial assessment of dataset suitability and, when combined with multicollinearity diagnostics such as the Variance Inflation Factor (VIF), contributes to a more robust and integrated analytical framework that enhances model reliability [24–27].

The results indicate that BI, BTS, and ROP contain a limited number of outlying observations. Although these outliers do not significantly distort the overall distribution of the dataset, their potential influence on model behaviour should not be ignored. In contrast, UCS, Alpha, and DPW exhibit more stable and consistent distributions, contributing to the overall homogeneity of the dataset [16–18].

As summarized in Fig. 4, the Variance Inflation Factor (VIF) analysis reveals a notable degree of multicollinearity among certain variables. In particular, UCS (VIF = 80.41) and BTS (VIF = 52.21) show strong correlations with other predictors, indicating the presence of redundant information within the dataset.

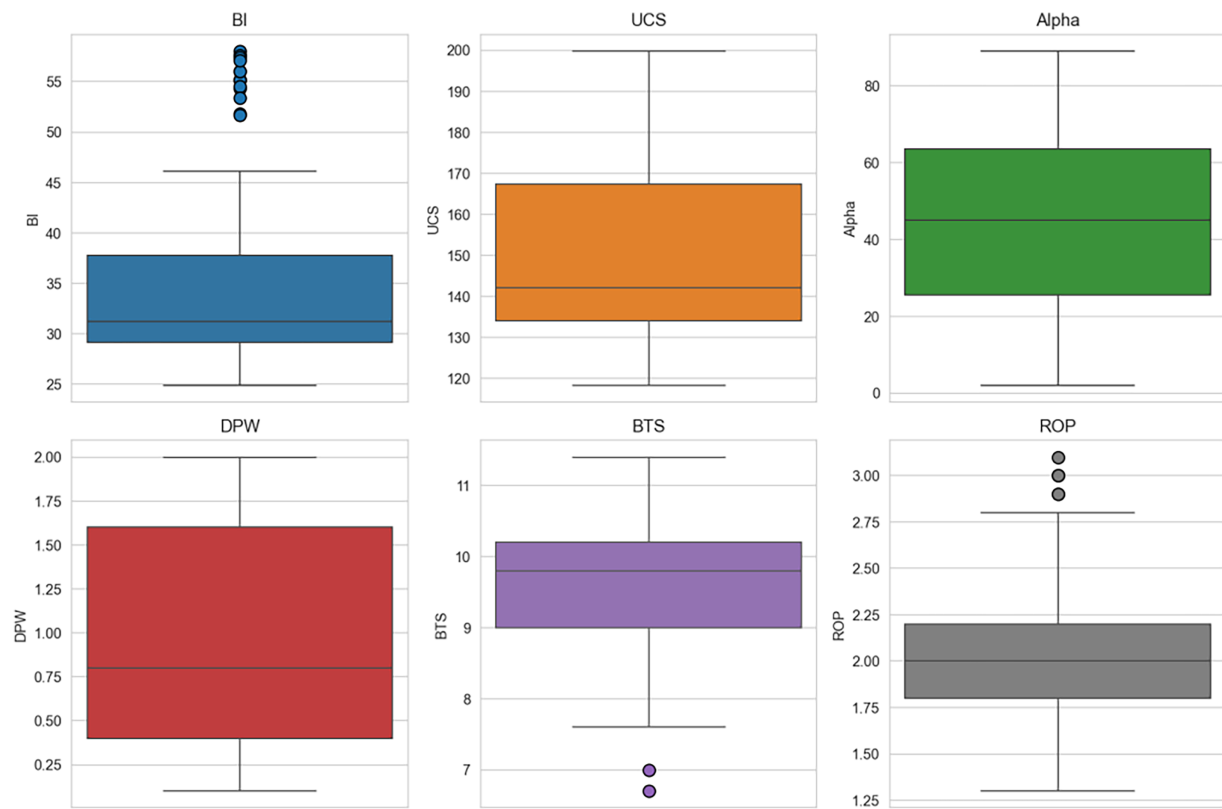


Figure 3: Distribution and outliers of TBM penetration rate and variables.

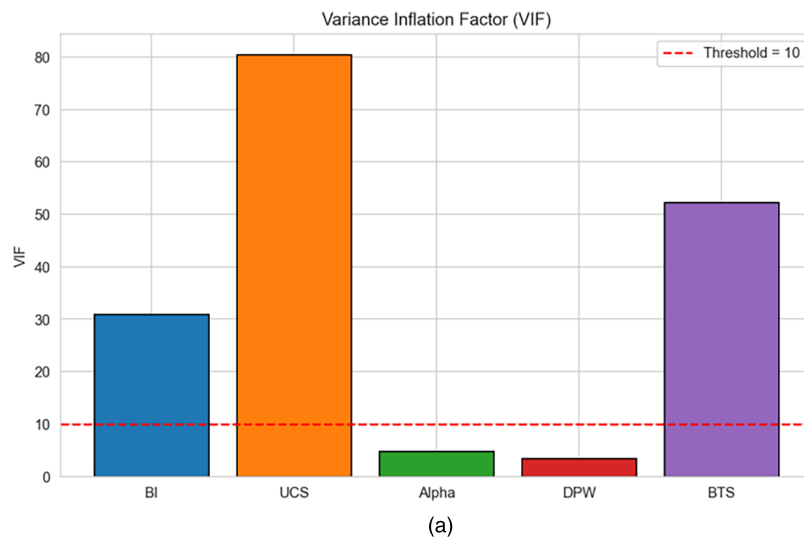


Figure 4: (Continued)

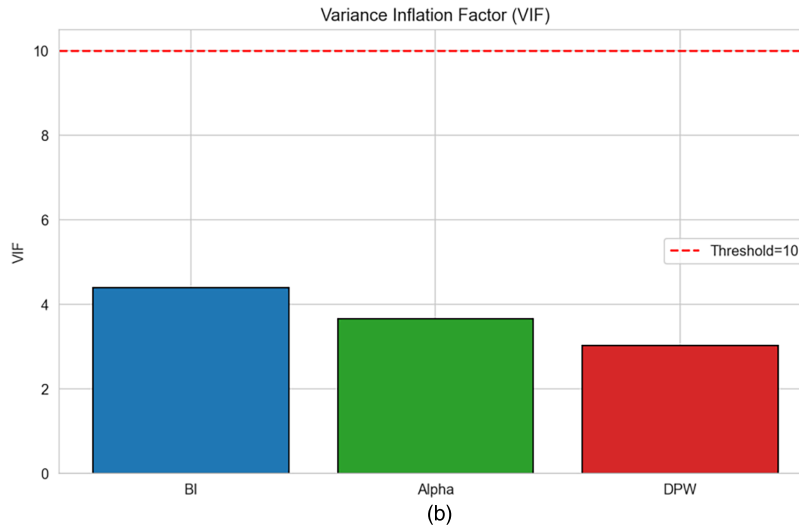


Figure 4: Variance inflation factor (VIF) values of the model variables before and after multicollinearity-based feature reduction: (a) VIF values for all model variables; (b) VIF values after removing variables with high multicollinearity.

These findings suggest that feature importance is not solely determined by predictive power, but is also closely related to inter-variable dependency structures and correlation levels. Accordingly, variables with high VIF values should be carefully evaluated in terms of their impact on model interpretability and generalization performance.

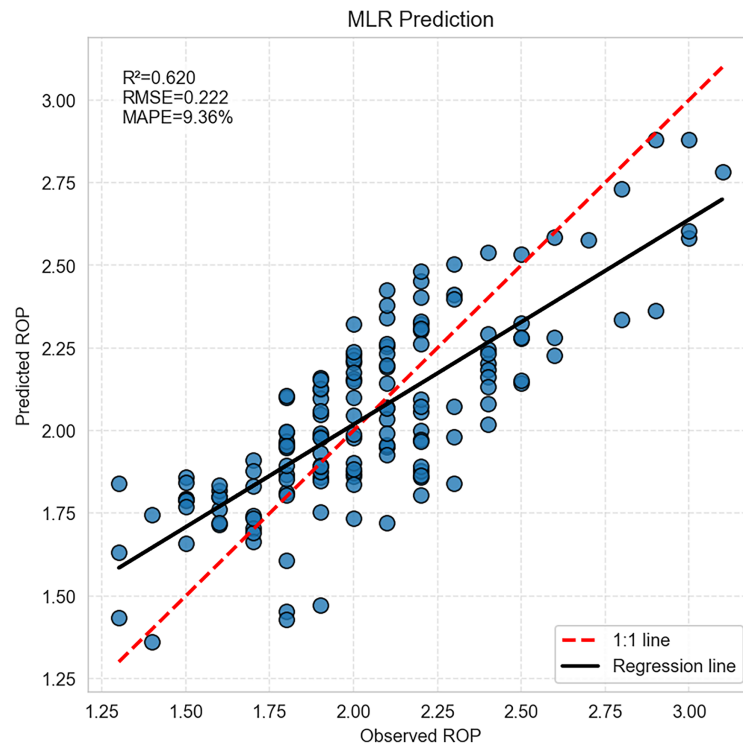
2.2 Selection of Model Inputs and Interactions among Variables

The preceding exploratory and distributional analyses indicate that the selection of input variables and the relationships among them play a critical role not only in predictive performance but also in the interpretability of variable importance. Accordingly, the analyses were conducted under two distinct scenarios.

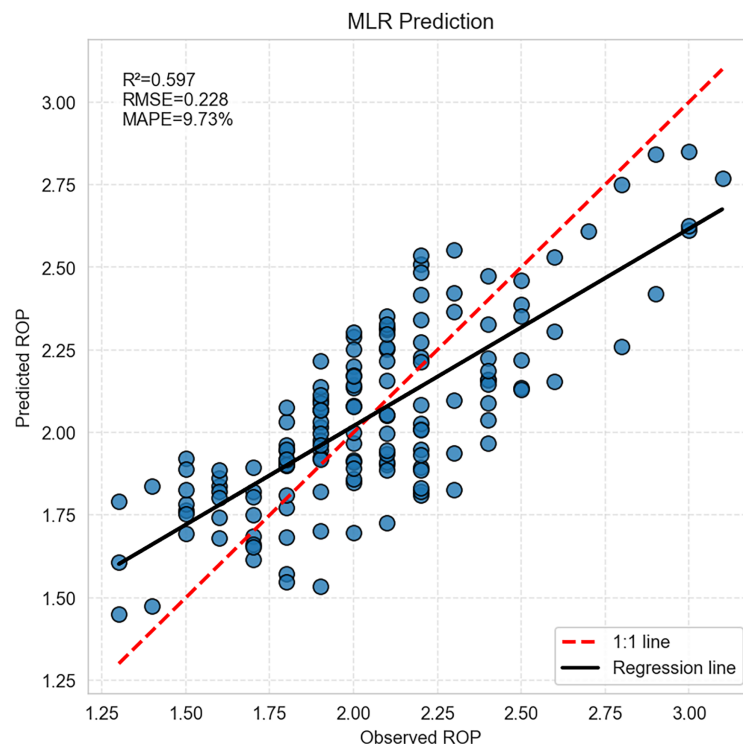
Scenario I represents the baseline case in which all available variables are included in the model. Under this configuration, the Multiple Linear Regression (MLR) model achieved strong predictive performance with $R^2 = 0.620$, $RMSE = 0.222$, and $MAPE = 9.36$. SHAP analysis revealed that BI (38.90%), DPW (26.39%), and Alpha (21.14%) were the dominant contributors to the model, whereas UCS and BTS exhibited comparatively lower influence on the predictions.

In **Scenario II**, the highly collinear UCS variable (characterized by a high VIF value) and the weakly contributing BTS variable were excluded from the model. This modification led to a slight degradation in predictive performance ($R^2 = 0.597$, $RMSE = 0.228$, $MAPE = 9.73$). However, the redistribution of SHAP contributions resulted in a clearer dominance of the remaining key predictors, with BI increasing to 40.55%, followed by DPW at 33.20% and Alpha at 26.25%.

Fig. 5 illustrates the agreement between observed and predicted values of the MLR model for both scenarios. Figs. 6 and 7 further present SHAP-based and absolute contribution analyses, respectively, enabling a comparative assessment of variable importance rankings and their contribution magnitudes. These results provide insight into the effect of variable elimination on model structure while also allowing the observation of shifts in feature importance distributions. It should be noted, however, that these findings are of a preliminary nature, and the primary objective is to further validate these evaluations within machine learning-based modeling frameworks.



(a)



(b)

Figure 5: Observed vs. predicted values for the multiple linear regression (MLR) model across both scenarios. (a) Scenario I; (b) Scenario II.

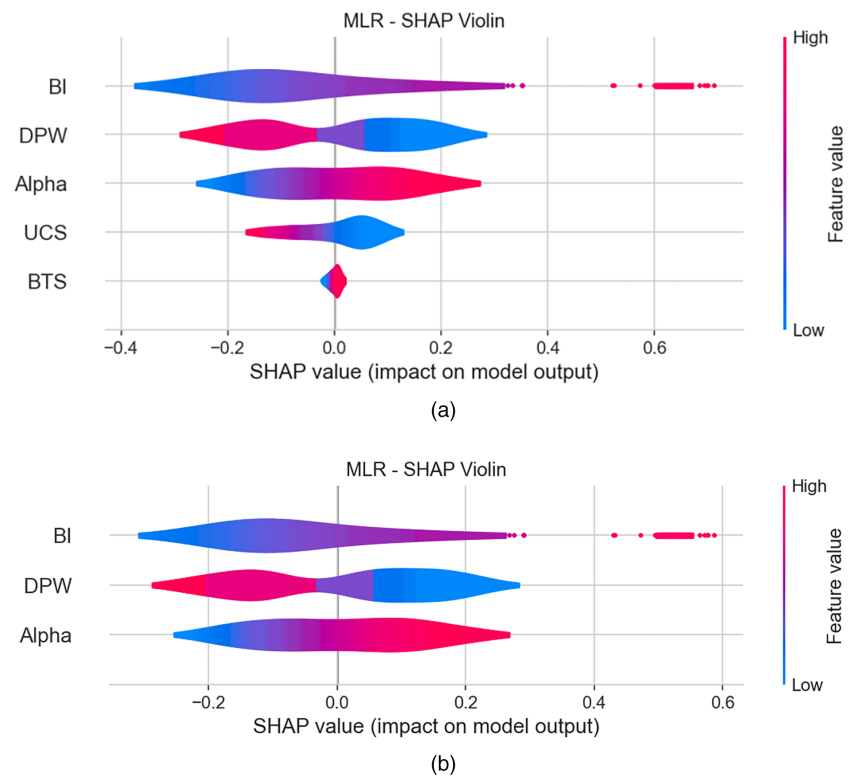


Figure 6: Comparison of feature-wise SHAP value distributions between Scenario I and Scenario II. (a) Scenario I; (b) Scenario II.

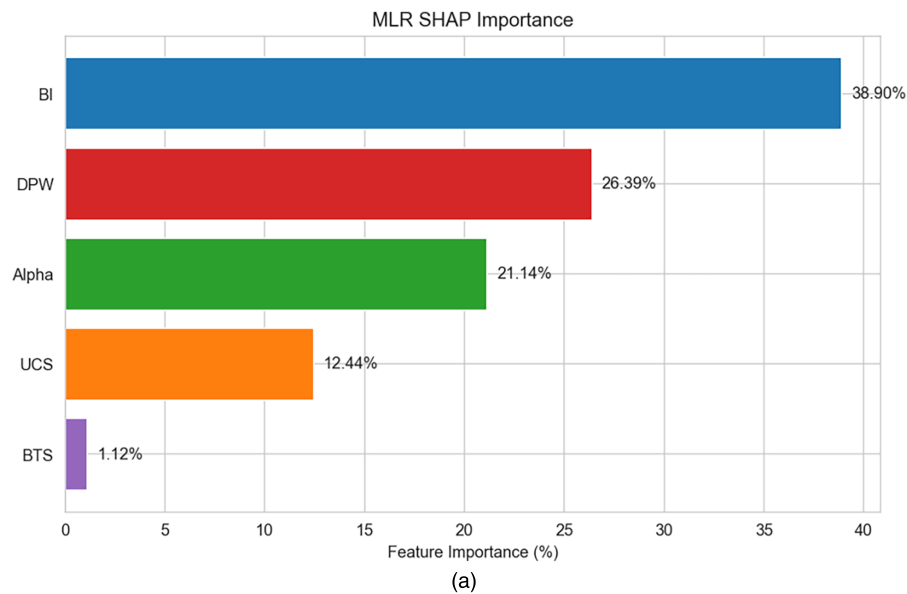


Figure 7: (Continued)

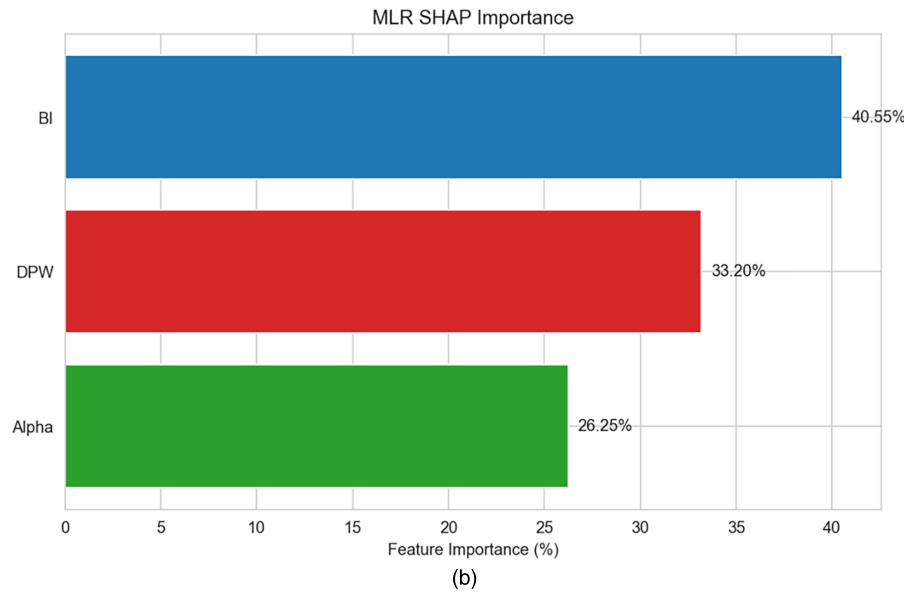


Figure 7: Comparative analysis of feature importance across Scenario I and Scenario II. (a) Scenario I; (b) Scenario II.

Overall, these results indicate that the removal of variables associated with multicollinearity and limited predictive contribution may improve model interpretability and provide a clearer representation of feature importance. However, the effect of feature reduction on predictive accuracy was model-dependent. While the predictive performance of RF, LSBoost, and Bagging Tree models remained relatively stable after variable reduction, the GAM model exhibited a substantial decline in performance in Scenario II, indicating greater sensitivity to the removal of input variables. In particular, the ranking of the primary controlling variables remained largely stable in Scenario II, with a slight increase in the relative contribution of BI, while the overall feature importance structure remained generally consistent. These findings suggest that feature reduction should be carefully evaluated within the specific modeling framework, as an appropriate reduction strategy may improve model transparency and interpretability while maintaining predictive performance for some models, but may lead to substantial performance degradation for models that are more sensitive to the selected predictors.

2.3 Machine Learning–Based Modeling Framework

In recent years, machine learning–based approaches have emerged as powerful alternatives to classical parametric modeling techniques. By learning directly from data without requiring predefined functional relationships, these methods can effectively capture the complex and nonlinear interactions between input variables and the rate of penetration (ROP). Owing to their flexible model structures, high predictive performance, and reduced reliance on restrictive assumptions, machine learning approaches have been widely adopted for ROP prediction and other geotechnical engineering applications [28–32].

In this study, several machine learning algorithms representing different learning paradigms were employed. Specifically, tree-based ensemble methods such as Random Forest (RF) [33,34], LSBoost [35,36], and Bagging Trees (BT) [37], as well as Generalized Additive Models (GAM) [38,39], were evaluated. Tree-based methods are capable of effectively capturing nonlinear relationships and interactions among variables with high representational power, whereas GAM provides a semi-parametric structure that combines flexible

modeling capability with a high degree of interpretability. Accordingly, methods based on different learning philosophies were comparatively analyzed for ROP prediction.

The modeling framework was designed not only with a focus on predictive accuracy but also on generalization capability and interpretability. All models were subjected to identical data preprocessing steps, trained using the same training–testing split, and evaluated using consistent error metrics. This approach ensures a fair and fully comparable evaluation framework across all models.

2.3.1 Model Training Procedure and Hyperparameter Optimization

A consistent training and optimization procedure was applied to all models in this study. The dataset was divided into training and test subsets, and hyperparameter tuning was performed exclusively on the training data. To identify the optimal model configurations, a grid search approach combined with cross-validation was employed.

For the Random Forest model, the number of trees ($n_estimators$), maximum tree depth (max_depth), and minimum number of samples required to split a node ($min_samples_split$) were optimized. In the LSBoost model, the number of trees, learning rate ($learning_rate$), and maximum depth were considered as tuning parameters. For the Bagging Trees model, the number of trees, bootstrap sampling ratio ($max_samples$), and feature sampling ratio ($max_features$) were optimized. For the Generalized Additive Models (GAM), the smoothing parameter (λ) and the number of splines ($n_splines$) were tuned [13,18].

The best hyperparameter combinations for all models were selected based on cross-validation performance, and the final evaluations were conducted on an independent test dataset.

2.3.2 Model Performance Evaluation Metrics

The performance of the models was assessed using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The R^2 metric represents the proportion of variance in the dependent variable explained by the model, thereby indicating the overall goodness of fit. RMSE measures the magnitude of prediction errors by calculating the square root of the mean squared differences between observed and predicted values. Because the errors are squared before averaging, RMSE assigns greater weight to larger prediction errors. MAE provides a more direct and interpretable measure of the average magnitude of prediction errors, with all errors contributing proportionally to the metric. MAPE expresses the prediction error in percentage terms, enabling relative comparison of prediction performance across different datasets [19].

All metrics were used both during the model optimization process and in the final performance comparison stage. Evaluations were conducted on an independent test dataset.

2.3.3 VIF-Based Dual-Scenario Modeling Approach

In this study, the Variance Inflation Factor (VIF) analysis was employed to assess multicollinearity among the independent variables. VIF is a measure that indicates how much the variance of a regression coefficient is inflated due to linear dependence with other predictors, and it is defined as follows [24]:

$$VIF_i = 1/(1 - R_i^2) \quad (1)$$

here, (R_i^2) denotes the coefficient of determination obtained from a regression model in which the i th variable is regressed against all other independent variables. High VIF values indicate strong correlations among variables and suggest redundancy in the model structure.

To evaluate the impact of VIF analysis on both predictive performance and interpretability, two distinct modeling scenarios were constructed. In the first scenario, all independent variables were included in the model, aiming to maximize information content and predictive performance. In this approach, multicollinearity was not explicitly addressed, allowing the model's learning capacity to dominate. Although this often leads to higher predictive accuracy, it may hinder the clear separation of individual variable contributions.

In the second scenario, variables exhibiting high multicollinearity based on VIF analysis were excluded, resulting in a more independent feature set. This approach reduces redundancy within the model, leading to a more stable and interpretable structure. However, the removal of variables may introduce a certain degree of information loss, which can result in slight variations in model performance.

The joint evaluation of these two scenarios provides a balanced analytical framework between predictive accuracy and model interpretability. In this context, the first scenario represents a data-driven, high-capacity learning approach, whereas the second scenario reflects a more statistically consistent and interpretable modeling structure.

2.4 Variable Importance Analysis

Feature importance analysis is a fundamental interpretability technique that provides insights into the extent to which machine learning models rely on different input variables when predicting the target variable. It enables the identification and comparison of the relative contributions of input variables to model predictions [40]. The adopted approaches were Partial Dependence Plot (PDP) analysis, Permutation Feature Importance (PFI), and SHAP (SHapley Additive exPlanations). PDP was used to characterize the average marginal relationship between each input variable and the predicted ROP by systematically varying the value of a given feature while averaging the resulting model predictions over the observed data [40]. PFI was used to assess the importance of each input variable by measuring the change in model predictions resulting from randomly permuting the values of that variable [40]. In contrast, SHAP was used to quantify the contribution of individual input variables to model predictions based on a game-theoretic framework [41]. Thus, PDP provides a global view of the model response to changes in individual input variables, PFI quantifies the predictive importance of variables based on their effect on model performance, and SHAP provides both global information on feature importance and local information on feature contributions for individual observations [41].

The use of multiple interpretability methods provides complementary information about the relationships between input variables and model predictions. Partial Dependence Plots (PDPs) describe the marginal relationship between an individual input variable and the model prediction while accounting for the average effects of the remaining variables. This approach is particularly useful for examining nonlinear response patterns and understanding how model predictions change across the range of an input variable [40]. SHAP provides another perspective by quantifying the contribution of individual input variables to model predictions based on Shapley values. In addition to providing global importance information, SHAP can characterize the direction and magnitude of variable contributions at the individual observation level, thereby providing a more detailed interpretation of model predictions [41].

Recent studies have demonstrated the usefulness of interpretable machine learning approaches for complex engineering and tunneling applications. In particular, the integration of explainability techniques with machine learning models has been shown to improve the interpretation of nonlinear relationships and to identify the variables that most strongly influence model predictions [42]. Similar applications have further demonstrated the value of interpretable machine learning for explaining complex prediction mechanisms in TBM-related engineering problems [43]. Feature importance and feature selection approaches provide

an additional perspective for evaluating the relative contribution and redundancy of input variables. Such approaches can support the identification of informative predictors and improve the interpretability of machine learning models [44]. A comprehensive review of feature selection techniques has likewise emphasized the importance of systematically evaluating the relevance of input variables when developing machine learning models [45]. Further applications of interpretable machine learning in TBM engineering have demonstrated the potential of combining predictive models with explainability techniques to reveal the influence of input variables on complex engineering responses [46]. These findings support the use of complementary interpretability approaches when the objective is not only to obtain accurate predictions but also to understand the relationships underlying model outputs. Taken together, PDP, PFI, and SHAP provide complementary perspectives on model behavior. PDP primarily describes the functional relationship between an input variable and the predicted response, PFI evaluates the change in predictive performance associated with disrupting a variable, and SHAP characterizes the contribution of individual variables to model predictions. Their complementary use can therefore provide a more comprehensive interpretation of machine learning models than reliance on a single interpretability approach.

2.4.1 SHAP Analysis

SHAP (SHapley Additive exPlanations) is a method built upon Shapley values from game theory for explaining model outputs. Its primary objective is to quantify the marginal contribution of each feature to the prediction in a fair and consistent manner. The method is based on computing the average marginal contribution across all possible feature subsets, and the Shapley value is defined as follows [47,48]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (2)$$

here, F represents the set of all features, while S denotes the subsets that do not include feature i . The term $f(S)$ refers to the model output obtained using only the features in S , whereas the value (ϕ_i) indicates the average contribution of feature i to the model. The SHAP approach decomposes the model output within an additive framework:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (3)$$

In this equation, ϕ_0 represents the model's expected (baseline) output, while ϕ_i denotes the individual contribution of each feature. In this way, the model prediction is decomposed into interpretable components, enabling transparent interpretation.

2.4.2 PFI Analysis

Permutation Feature Importance (PFI) is a model-agnostic variable importance method that quantifies the contribution of an input variable by measuring the deterioration in model performance after randomly permuting its values. If a feature is important for the model, randomly shuffling its values is expected to increase the prediction error. In this study, PFI was evaluated using the mean squared error (MSE) as the loss function. The PFI of the j -th feature can be expressed as:

$$PFI_j = (\mathcal{L}(f(X^{\pi_k(j)}), y) - \mathcal{L}(f(X), y)) \quad (4)$$

where L denotes the MSE loss function, $f(X)$ represents the model predictions using the original dataset, and $X^{\pi(j)}$ denotes the dataset in which the values of the j -th feature are randomly permuted. A positive PFI

value indicates that permuting the corresponding feature increases the prediction error, indicating a greater contribution of that feature to model performance [49,50].

To reduce the influence of randomness associated with a single permutation, the procedure was repeated K times and the resulting importance values were averaged:

$$PFI_j = \frac{1}{K} \sum_{k=1}^K (\mathcal{L}(f(X^{\pi_k(j)}), y) - \mathcal{L}(f(X), y)) \quad (5)$$

where K represents the number of independent permutations and $\pi_k(j)$ denotes the k -th permutation of the j -th feature. Thus, larger positive PFI values indicate a greater deterioration in model performance after permutation and, consequently, a higher relative importance of the corresponding feature. As a model-agnostic approach, PFI can be applied to different machine learning algorithms. However, when input variables are strongly correlated, permuting one variable independently may disrupt the original dependence structure among predictors and may therefore lead to biased or potentially underestimated/overestimated importance estimates.

2.4.3 Partial Dependence Plot (PDP) Analysis

Partial Dependence Plot (PDP) is a model-agnostic explainability method used to characterize the average marginal relationship between a specific input variable and the predictions of a machine learning model [40,51]. PDP was originally introduced as a tool for interpreting complex predictive models by examining how model predictions change as the value of a selected input variable is systematically varied while averaging over the observed distribution of the remaining input variables [40]. Thus, PDP provides a global description of the model response to changes in an individual input variable and can reveal linear, nonlinear, monotonic, or threshold-like relationships between the input variable and the predicted outcome [47].

When the model prediction function is denoted by $f(\mathbf{x})$, the partial dependence function for the j -th input variable can be expressed as [40,47]:

$$\hat{f}_{PDP}(x_j) = \frac{1}{n} \sum_{i=1}^n f(x_j, x_{-j}^{(i)}) \quad (6)$$

where x_j represents a specific value of the j -th input variable, $x_{-j}^{(i)}$ denotes the values of all remaining input variables for the i -th observation, and n represents the total number of observations. This formulation represents the average model prediction obtained when the j -th input variable is fixed at a specific value x_j , while the remaining variables retain their observed values.

In this study, PDP analysis was applied to evaluate the average marginal relationships between BI, UCS, DPW, Alpha, and BTS and the predicted ROP. For each input variable, a set of evaluation values spanning its observed range was generated. The selected input variable was then fixed at each evaluation value across all observations, while the remaining input variables were retained at their corresponding observed values. The trained model was subsequently used to generate predictions for each modified dataset, and the predictions were averaged across all observations to obtain the corresponding PDP value.

The resulting PDP values represent the average model prediction associated with a particular value of the corresponding input variable over the observed dataset. Accordingly, increasing, decreasing, nonlinear, and threshold-like patterns in the PDP curves were interpreted as characteristics of the model's learned relationship between the input variable and ROP. Since PDP summarizes the model response over the

dataset rather than explaining individual observations, it was used as a global interpretability method in this study [40,51,52].

PDP analysis was used as a complementary approach to Permutation Feature Importance (PFI) and SHAP. PFI evaluates the predictive importance of input variables according to the change in model performance following random permutation of each variable, whereas SHAP quantifies the contribution of individual input variables to model predictions at both global and observation-specific levels. In contrast, PDP characterizes how the average model prediction changes across different values of an input variable. Therefore, the combined use of PDP, PFI, and SHAP provides complementary information on both the relative importance of input variables and their learned relationships with ROP predictions [51,52].

2.4.4 LIME Analysis

In addition to the methods described above, the LIME (Local Interpretable Model-agnostic Explanations) approach was employed to improve the interpretability of model predictions. LIME is a model-agnostic local explanation technique developed to interpret individual predictions of black-box models [53,54]. The method constructs a simple and interpretable surrogate model that approximates the behaviour of the original model in the local vicinity of each data point.

In this study, LIME was applied independently to selected individual observations representing three ROP conditions: low, mean, and high ROP. Three observations were selected for each condition based on the observed ROP values. Specifically, the three observations with the lowest ROP values were classified as the LOW group, the three observations closest to the mean ROP were classified as the MEAN group, and the three observations with the highest ROP values were classified as the HIGH group. LIME explanations were then generated separately for each selected observation. This sampling strategy enabled the local contributions of the input variables to be examined across different ROP conditions while preserving the instance-level nature of the LIME methodology.

LIME generates perturbed samples around each instance to be explained, and model predictions are obtained for these synthetic samples. The samples are then weighted according to their proximity to the original instance, and the following optimization problem is solved:

$$\arg \min g \in GL(f, g, \pi_x) + \Omega(g) \quad (7)$$

here, f denotes the prediction model, g the local surrogate model, L the local fidelity loss, and $\Omega(g)$ the model complexity penalty. The weighting function π_x is defined using an exponential kernel based on the proximity of samples to the original data point [53,54].

The resulting local surrogate model is typically in the form of linear regression, and the contribution of each feature to the corresponding prediction is interpreted through the model coefficients. In this way, the local behaviour of the model is evaluated separately for each representative point, enabling a comparative analysis of feature contributions. Top of Form

3 Model Implementation

This section presents the detailed implementation procedure of the machine learning models employed for the prediction of the rate of penetration. Within the modeling framework defined in the previous sections, prediction models were developed using Random Forest (RF), LSBoost, Bagging Trees (BT), and Generalized Additive Models (GAM). These models represent different learning paradigms and enable the modeling of complex nonlinear relationships between ROP and the independent variables.

During the modeling phase, the dataset was divided into training and testing subsets. Hyperparameter optimization was performed exclusively on the training data, and the optimal model configurations were determined using a grid search strategy combined with cross-validation. In this context, 5-fold cross-validation was applied to obtain more robust and generalizable estimates of model performance. Each hyperparameter combination was evaluated across different folds to reduce the risk of overfitting.

Based on the results of the VIF-based analysis, the modeling process was conducted under two different scenarios. In the first scenario, a full model structure including all independent variables was established, incorporating BI, UCS, Alpha, DPW, and BTS as predictors. In the second scenario, a reduced variable set was used to mitigate multicollinearity, and the models were trained using only BI, DPW, and Alpha. This enabled a comparative assessment of both the full and simplified model structures in terms of performance and behaviour.

This approach allowed the models to be evaluated not only with respect to predictive accuracy but also in terms of how variations in dataset structure influence model behaviour. The findings regarding model performance and variable importance are presented in the following sections.

3.1 Model Performance

Table 1 presents the model performance results for both scenarios.

Table 1: Comparative model performance across scenarios.

Scenario	Model	R ²	RMSE	MAE	MAPE
I	GAM	0.987	0.040	0.031	1.562
	RandomForest	0.944	0.085	0.066	3.307
	LSBoost	0.942	0.087	0.067	3.364
	Bagging Trees	0.915	0.105	0.076	3.822
II	Random Forest	0.944	0.085	0.067	3.356
	LSBoost	0.931	0.095	0.074	3.696
	Bagging Trees	0.918	0.103	0.081	4.044
	GAM	0.580	0.233	0.186	9.162

When the results presented in Table 1 are evaluated collectively, a scenario-dependent change in model ranking becomes evident. In Scenario I, the GAM model clearly outperforms all other approaches, whereas in Scenario II its performance drops sharply. This contrast highlights the strong dependence of additive models on the availability of the full feature set and indicates that model behaviour is highly sensitive to feature reduction.

In Scenario I, the GAM model achieves the highest predictive accuracy with $R^2 = 0.987$, explaining nearly 99% of the variance in the dependent variable and producing the lowest error values (RMSE = 0.040, MAE = 0.031, MAPE = 1.562). This result indicates that, when all predictors are included, the smooth nonlinear and interaction structures captured by the GAM framework can represent the underlying data structure extremely well. Among the tree-based methods, Random Forest ($R^2 = 0.944$) and LSBoost ($R^2 = 0.942$) show very similar and strong performance, confirming the effectiveness of ensemble learning in modeling nonlinear relationships. Bagging Trees remains the weakest of the ensemble models ($R^2 = 0.915$), suggesting that variance reduction alone is less effective than boosting or random forest strategies.

A markedly different pattern emerges in Scenario II. After the removal of UCS and BTS, the GAM model experiences a dramatic performance decline ($R^2 = 0.580$; $MAPE = 9.162$), indicating that the additive structure heavily relies on the excluded variables and struggles to capture the remaining relationships using only BI, DPW, and Alpha. In contrast, the Random Forest model maintains almost identical performance ($R^2 = 0.944$), demonstrating remarkable robustness and generalization capability under feature reduction. LSBoost shows a moderate decrease in performance ($R^2 = 0.931$), confirming that boosting methods are more sensitive to feature set changes. Bagging Trees exhibits a slight improvement ($R^2 = 0.918$), suggesting that removing redundant or noisy variables provides a modest benefit for this approach.

Considering both scenarios together, two key findings emerge. First, GAM can achieve exceptionally high accuracy when the full predictor set is available, indicating strong compatibility between the additive modeling framework and the complete dataset structure. Second, tree-based ensemble models—particularly Random Forest—demonstrate superior robustness and stability when the feature space is reduced. Overall, Random Forest provides the most consistent performance across scenarios, LSBoost offers competitive but slightly more sensitive performance, Bagging Trees delivers moderate accuracy, and GAM shows strong dependence on the completeness of the predictor set.

3.2 Feature Importance Analysis

In this section, feature importance is analyzed comparatively across the same models (RF, LSBoost, BT, and GAM) and the same explainability techniques (SHAP, PFI, and PDP) under two different scenarios. Scenario I represents the five-variable system (BI, DPW, Alpha, UCS, BTS), while Scenario II corresponds to the reduced three-variable system (BI, DPW, Alpha). This framework enables a direct investigation of how reducing the predictor set influences model interpretability and the distribution of feature importance.

3.2.1 Scenario I

In this scenario, all models were trained using the full set of predictors. The feature importance distributions obtained from the SHAP, PFI, and PDP analyses are summarized in [Table 2](#), while their graphical comparison is presented in [Fig. 8](#).

Table 2: Feature importance analysis for Scenario I using explainable AI methods.

Model	Feature	SHAP (%)	PFI (%)	PDP (%)
RF	BI	40.584	41.860	27.556
	UCS	6.811	6.878	15.868
	Alpha	25.041	22.688	34.352
	DPW	23.363	22.622	13.021
	BTS	4.200	5.952	9.203
LSBoost	BI	36.829	35.273	20.574
	UCS	10.438	13.257	13.150
	Alpha	22.590	23.420	42.043
	DPW	25.166	21.359	13.236
	BTS	4.978	6.691	10.997
	BI	38.828	33.342	23.373
	UCS	7.574	8.222	19.305

(Continued)

Table 2 (continued)

Model	Feature	SHAP (%)	PFI (%)	PDP (%)
BT	Alpha	24.400	28.960	35.415
	DPW	23.931	22.252	6.904
	BTS	5.267	7.225	15.003
GAM	BI	38.528	34.943	40.917
	UCS	12.877	12.735	14.027
	Alpha	19.609	22.657	22.275
	DPW	24.555	25.215	18.704
	BTS	4.431	4.451	4.076

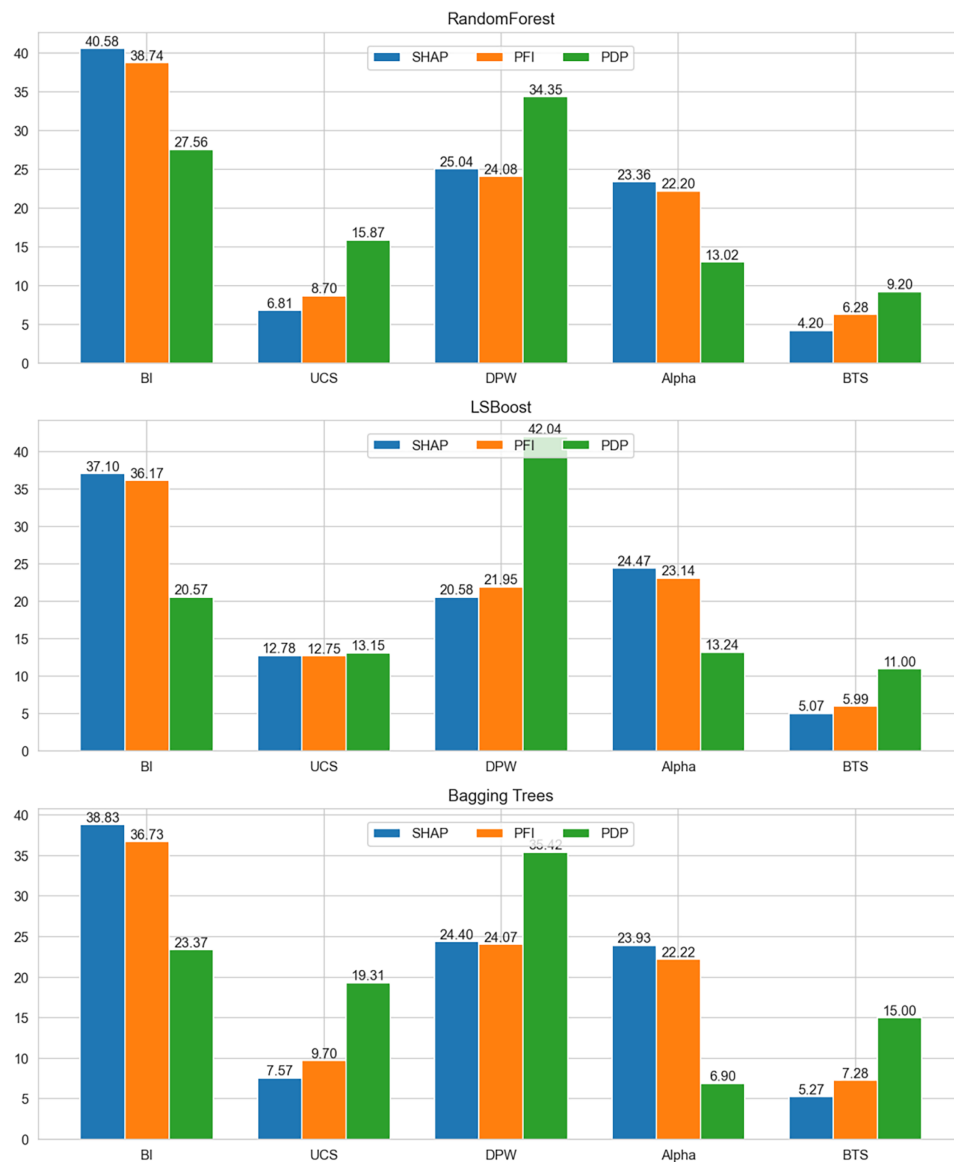


Figure 8: (Continued)

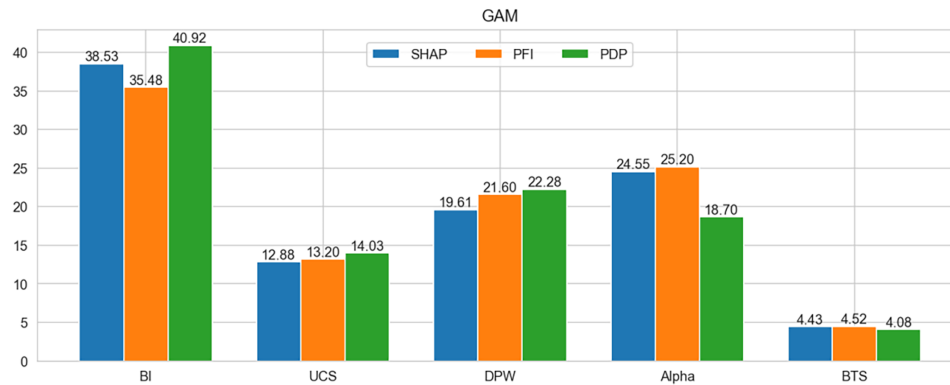


Figure 8: Model-based comparison of feature importance using SHAP, PFI, and PDP under Scenario I.

Together, these explainable AI methods provide complementary perspectives on the contribution of each predictor to the ROP estimation process. Overall, the results presented in Table 2 and Fig. 8 indicate that the relative importance of the predictors varies depending on the explainability method and the model considered. BI emerged as one of the most influential predictors in the SHAP and PFI analyses, whereas Alpha exhibited the highest relative importance in the PDP analysis across the tree-based models (RF, LSBoost, and BT). In particular, the PDP results showed that Alpha had a higher relative importance than BI for RF (34.352% vs. 27.556%), LSBoost (42.043% vs. 20.574%), and BT (35.415% vs. 23.373%). DPW also showed a relatively high contribution, although its importance varied among the models and explainability methods. In contrast, UCS generally exhibited a moderate contribution, while BTS consistently showed the lowest relative importance, indicating a comparatively limited influence on the prediction performance. These findings demonstrate that the interpretation of predictor importance is method-dependent and that combining multiple explainability approaches provides a more comprehensive understanding of the factors governing ROP predictions.

Although the general ranking of variables remains stable, the three interpretability approaches emphasize different aspects of model behaviour. SHAP and PFI produce highly consistent importance distributions, suggesting that the identified feature rankings are robust across different model architectures. PDP, however, places greater emphasis on the marginal influence of Alpha and reveals nonlinear response patterns that are less apparent from SHAP and PFI. These differences demonstrate that the methods should be viewed as complementary rather than interchangeable, as each provides unique insight into the mechanisms underlying model predictions.

Among the evaluated algorithms, the tree-based ensemble models exhibit remarkably similar feature importance patterns, confirming the stability of the learned relationships. The GAM model, however, displays a somewhat different distribution of importance in the PDP analysis, reflecting its additive modelling framework and its distinct representation of nonlinear relationships. Despite these methodological differences, the overall conclusions regarding the dominant and secondary predictors remain consistent across all models.

Taken together, the findings indicate that BI, Alpha, and DPW constitute the primary explanatory variables governing ROP prediction under Scenario I, whereas UCS provides a secondary contribution and BTS has only a limited effect. The close agreement between SHAP and PFI, together with the complementary nonlinear insights provided by PDP, enhances confidence in the robustness and interpretability of the developed models.

In addition to the global feature importance analyses presented in Table 2 and Fig. 8, the local interpretability of the models under minimum, average, and maximum ROP conditions is illustrated in Fig. 9 using the LIME method.

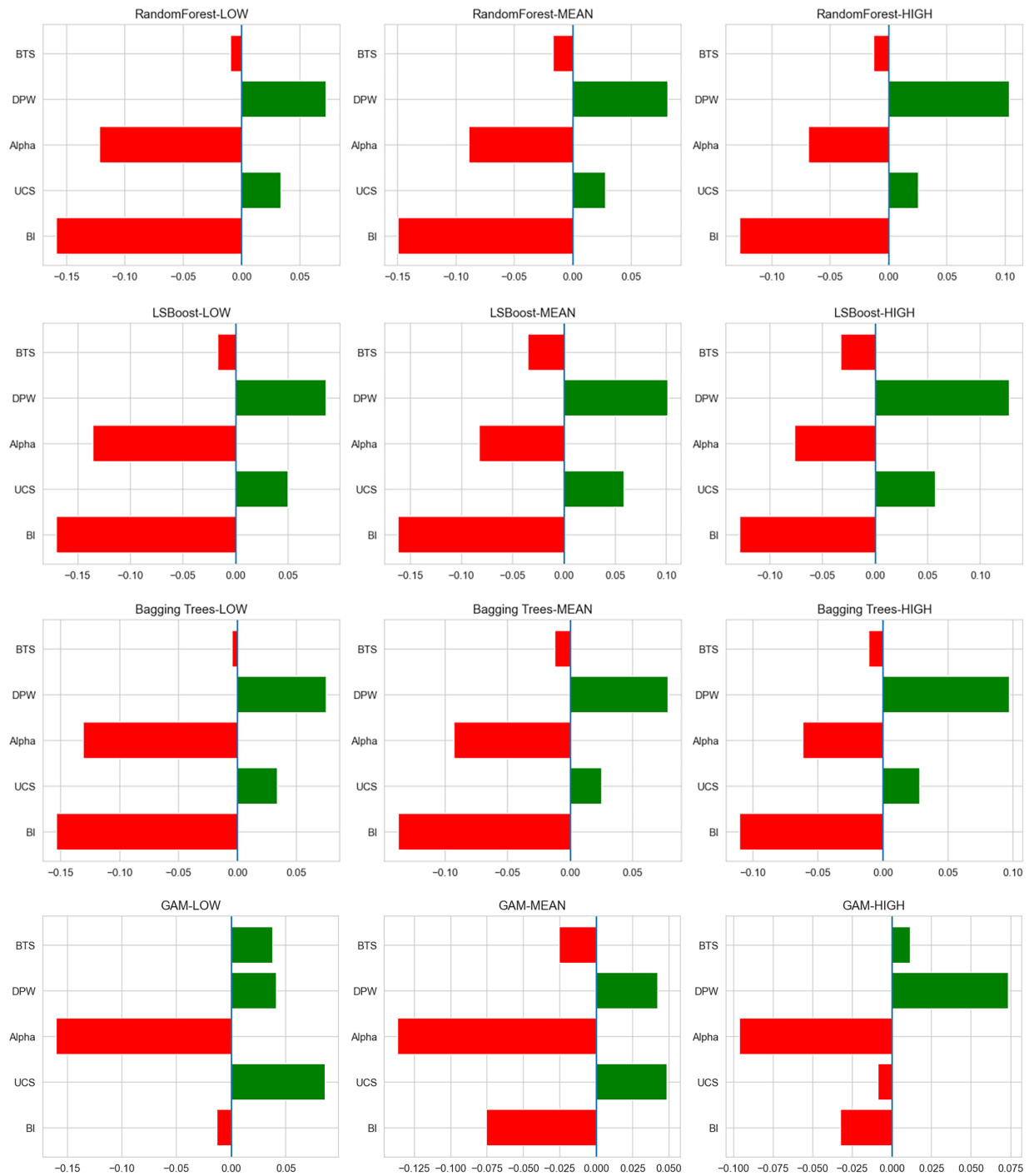


Figure 9: LIME-based local interpretability variations under minimum, average, and maximum ROP conditions in Scenario I.

The analysis is based on the average of the three lowest, mean, and three highest ROP values. As shown in Fig. 9, all machine learning models exhibit broadly consistent local behaviour across the three operating conditions, further supporting the reliability of the identified feature importance patterns.

3.2.2 Scenario II–Three-Variable System

To evaluate the rescaled effects of the variables in the three-variable system, the feature importance distributions obtained from the SHAP, PFI, and PDP analyses are summarized in Table 3, while their graphical comparison is presented in Fig. 10.

Table 3: Feature importance analysis using explainable AI methods for Scenario II.

Model	Feature	SHAP (%)	PFI (%)	PDP (%)
RF	BI	43.215	40.139	31.393
	Alpha	28.029	32.015	49.462
	DPW	28.757	27.846	19.145
LSBoost	BI	40.559	40.290	33.266
	Alpha	28.479	29.611	50.083
	DPW	30.961	30.099	16.650
BT	BI	41.718	42.223	36.026
	Alpha	28.010	27.108	45.649
	DPW	30.273	30.668	18.325
GAM	BI	30.903	29.871	31.175
	Alpha	30.846	34.444	42.445
	DPW	38.252	35.685	26.380

Overall, the results presented in Table 3 indicate that reducing the input space to the three-variable system (BI–Alpha–DPW) does not substantially alter the overall feature importance ranking across the evaluated models. The tree-based models (RF, LSBoost, and BT) maintain highly consistent importance distributions, whereas the GAM model exhibits a comparatively different allocation of feature contributions owing to its additive modelling framework.

Although BI remains the dominant predictor in the tree-based models, the reduced feature space increases the relative contribution of Alpha and DPW. Among the three explainability methods, SHAP and PFI provide highly consistent feature rankings, while PDP places greater emphasis on the nonlinear marginal influence of Alpha, indicating that its contribution becomes more pronounced under the simplified feature configuration.

The GAM model displays the largest deviation from the ensemble tree-based models. Rather than assigning clear dominance to BI, GAM distributes feature importance more evenly between Alpha and DPW, reflecting its ability to represent nonlinear relationships through smooth additive functions. Nevertheless, the overall interpretation remains consistent across models, confirming that BI, Alpha, and DPW constitute the principal explanatory variables governing ROP prediction under Scenario II.

Taken together, the results demonstrate that feature reduction preserves model stability while improving the interpretability of the remaining predictors. The consistent agreement between SHAP and PFI reinforces

the robustness of the identified feature rankings, whereas PDP provides complementary information by highlighting nonlinear effects that are less evident from global importance measures.

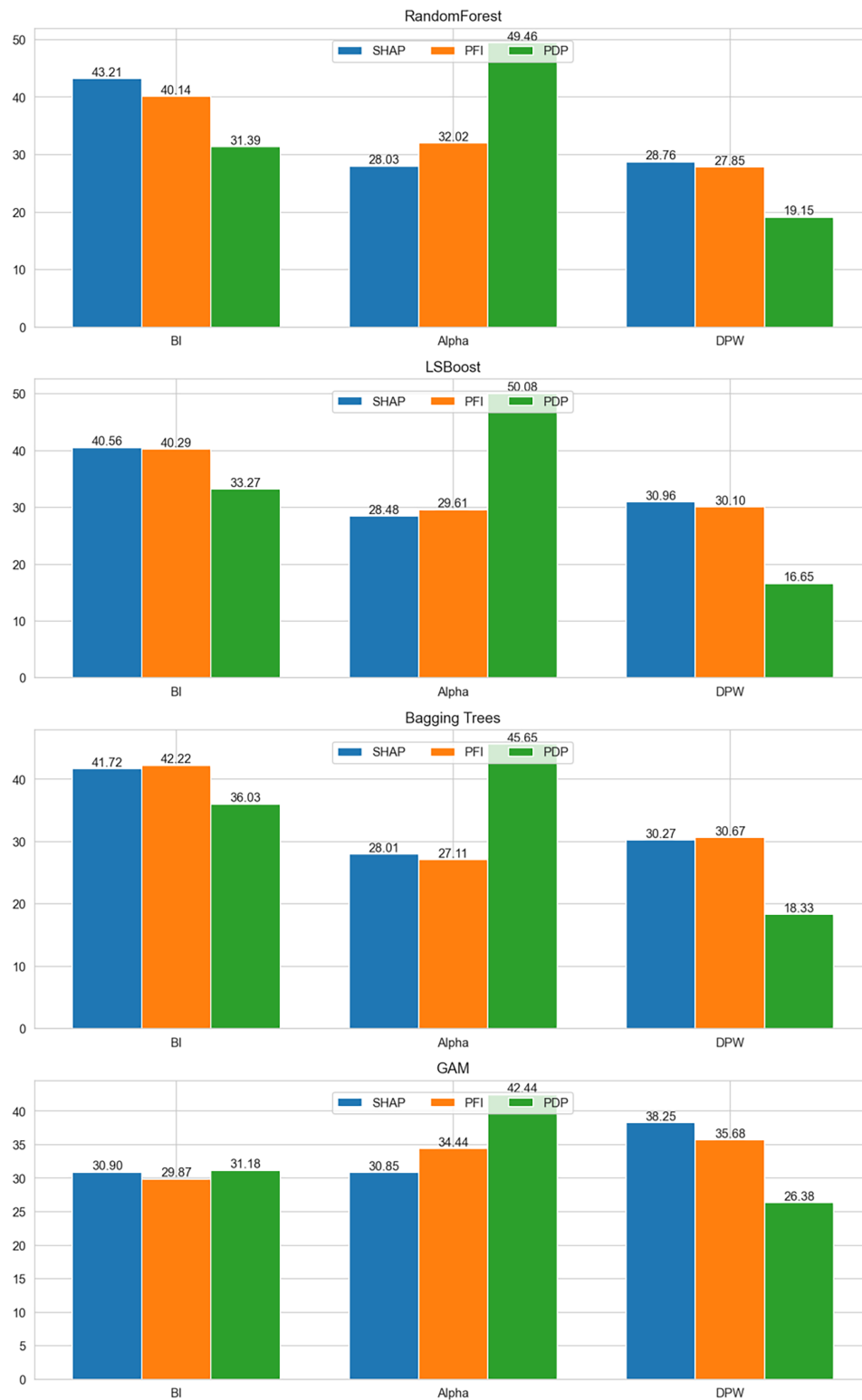


Figure 10: Model-based comparison of feature importance using SHAP, PFI, and PDP for Scenario II.

The graphical comparison presented in Fig. 10 supports the numerical findings reported in Table 3. The similarity among the tree-based models confirms the stability of their learned feature importance patterns, whereas the different distribution observed for the GAM model illustrates the influence of its additive modelling structure on feature attribution.

In addition to the global feature importance analyses, Fig. 11 illustrates the local model responses in Scenario II under minimum, average, and maximum ROP conditions using the LIME method. The local explanations were obtained by considering the average of the three lowest, mean, and three highest ROP observations. Because Scenario II excludes the variables used in Scenario I, the LIME results provide an independent assessment of how the reduced input set contributes to individual model predictions under different ROP conditions. As shown in Fig. 11, the contribution of the retained input variables varies across the three operating conditions, while the models generally maintain similar prediction patterns. These results indicate that the reduced-variable configuration preserves consistent local predictive behaviour across different ROP levels, complementing the feature-specific patterns identified through SHAP, PFI, and PDP analyses.

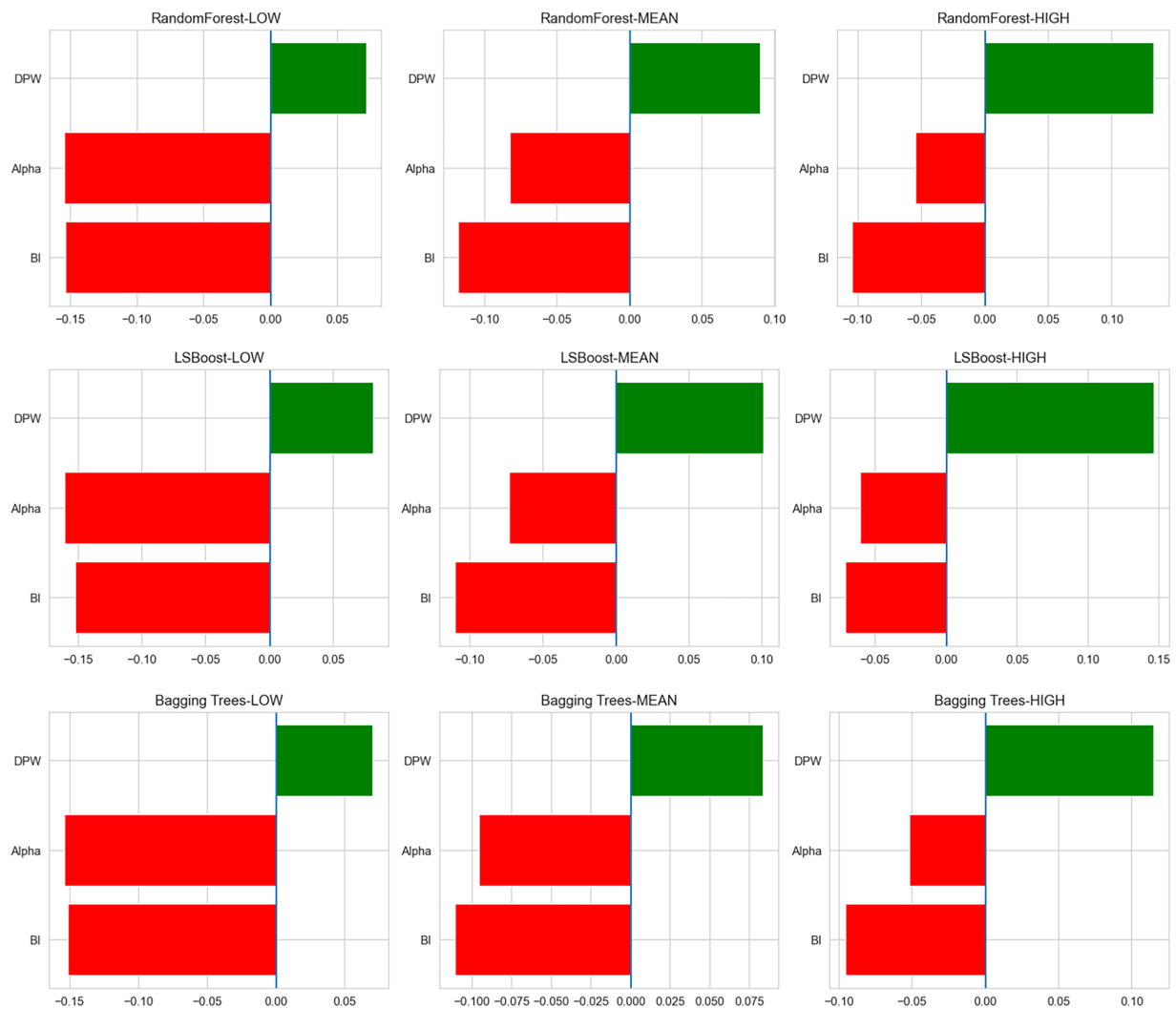


Figure 11: (Continued)

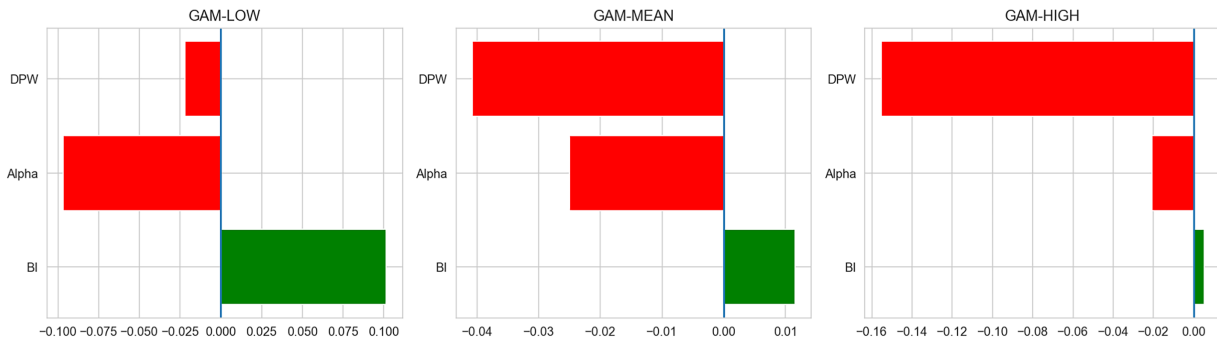


Figure 11: LIME-based local interpretability variations under minimum, average, and maximum ROP conditions in Scenario II.

3.2.3 Cross-Scenario Comparison and Overall Evaluation

When both scenarios are evaluated together, the effect of feature set simplification on model interpretability becomes clearly evident. Scenario I employs the complete feature set (BI, UCS, Alpha, DPW, and BTS), whereas Scenario II excludes UCS and BTS, resulting in a more compact three-variable system. This simplification leads to a redistribution of feature importance while preserving the overall explanatory structure of the models.

A comparison of the results presented in [Tables 2](#) and [3](#) indicates that BI remains the dominant explanatory variable across both scenarios regardless of the explainability method or machine learning algorithm. The tree-based models consistently preserve the same feature importance ranking, whereas the GAM model exhibits a comparatively different distribution of feature contributions owing to its additive modelling framework.

Feature reduction increases the relative contributions of Alpha and DPW, particularly in the GAM model, although the dominant predictor varies depending on the interpretability method. In Scenario II, DPW has the highest contribution for GAM according to both SHAP and PFI, whereas Alpha becomes the dominant feature according to PDP. Among the three interpretability approaches, SHAP and PFI generally show consistent feature-ranking patterns across the two scenarios, while differences in the rankings become more evident after feature reduction. In particular, PDP places greater emphasis on the nonlinear influence of Alpha, indicating that its marginal contribution becomes more pronounced in the simplified feature space. These results demonstrate that feature reduction can alter the relative importance of the retained variables and that the identified dominant feature may depend on both the underlying model and the interpretability method used.

The exclusion of UCS and BTS from Scenario II has different effects across the evaluated models, indicating that their explanatory contributions are model-dependent. For the tree-based models, particularly Random Forest and Bagging Trees, the removal of these variables has a relatively limited effect on predictive performance, suggesting that the retained variables provide sufficient information to maintain stable predictions. In contrast, GAM exhibits a substantial performance deterioration after feature reduction, indicating that UCS and BTS contribute more critically to its predictive capability. Therefore, the simplified feature set does not uniformly preserve prediction stability across all models; rather, its effect depends on the model structure and the extent to which the excluded variables contribute to model prediction. Nevertheless, the reduced feature set facilitates a more focused interpretation of the retained variables, particularly for the more robust tree-based models.

The comparative heatmaps presented in Figs. 12 and 13 visually reinforce the numerical findings obtained from Tables 2 and 3. The consistently high intensity associated with BI across both scenarios confirms its dominant contribution to model behaviour. In contrast, the increased intensity observed for Alpha and DPW after feature reduction illustrates the redistribution of explanatory power within the simplified feature space. The similarity of the SHAP- and PFI-based heatmaps further supports the consistency of the global feature importance analyses across different interpretability methods.

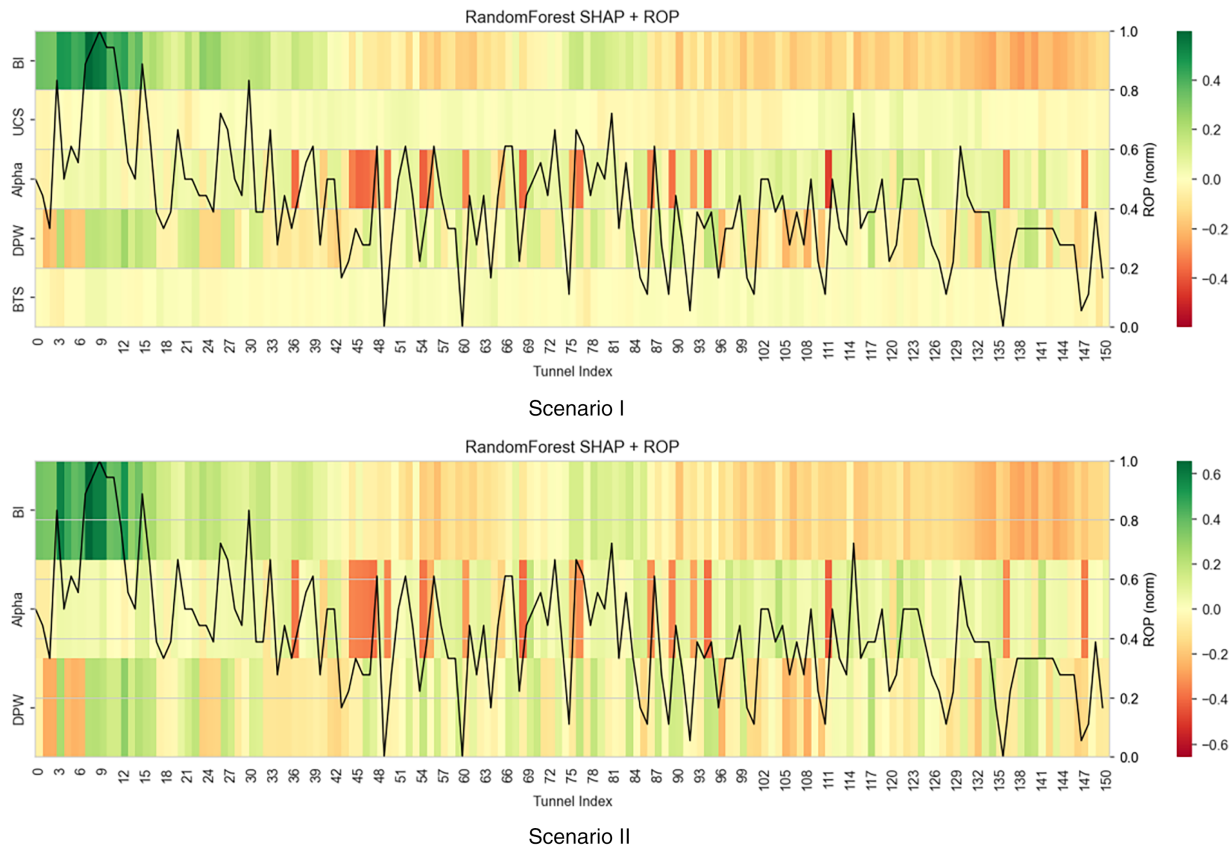


Figure 12: Comparative heatmap showing the distribution of SHAP-based feature importance across Scenario I and Scenario II.

Overall, the cross-scenario comparison demonstrates that simplifying the feature set enhances model interpretability without altering the principal explanatory structure of the prediction models. BI remains the primary controlling variable under both scenarios, while Alpha becomes more influential in describing nonlinear behaviour and DPW continues to provide an important complementary contribution. These findings indicate that the reduced feature set retains the essential predictive information while offering a more compact and interpretable representation of the ROP prediction problem.

3.3 Validation of the Proposed Framework Using Independent Geotechnical Datasets

To evaluate the robustness, reliability, and generalizability of the proposed modeling framework under different data conditions, the methodology was also applied to a second dataset. This dataset was originally compiled by Mansouri and Moomivand from the Karaj–Tehran water conveyance tunnel [23] and was subsequently used by Karahan and Alkaya in their previous study [19]. The same dataset was employed in the

present study to provide an independent evaluation of the proposed modeling framework under different geological and operational conditions.

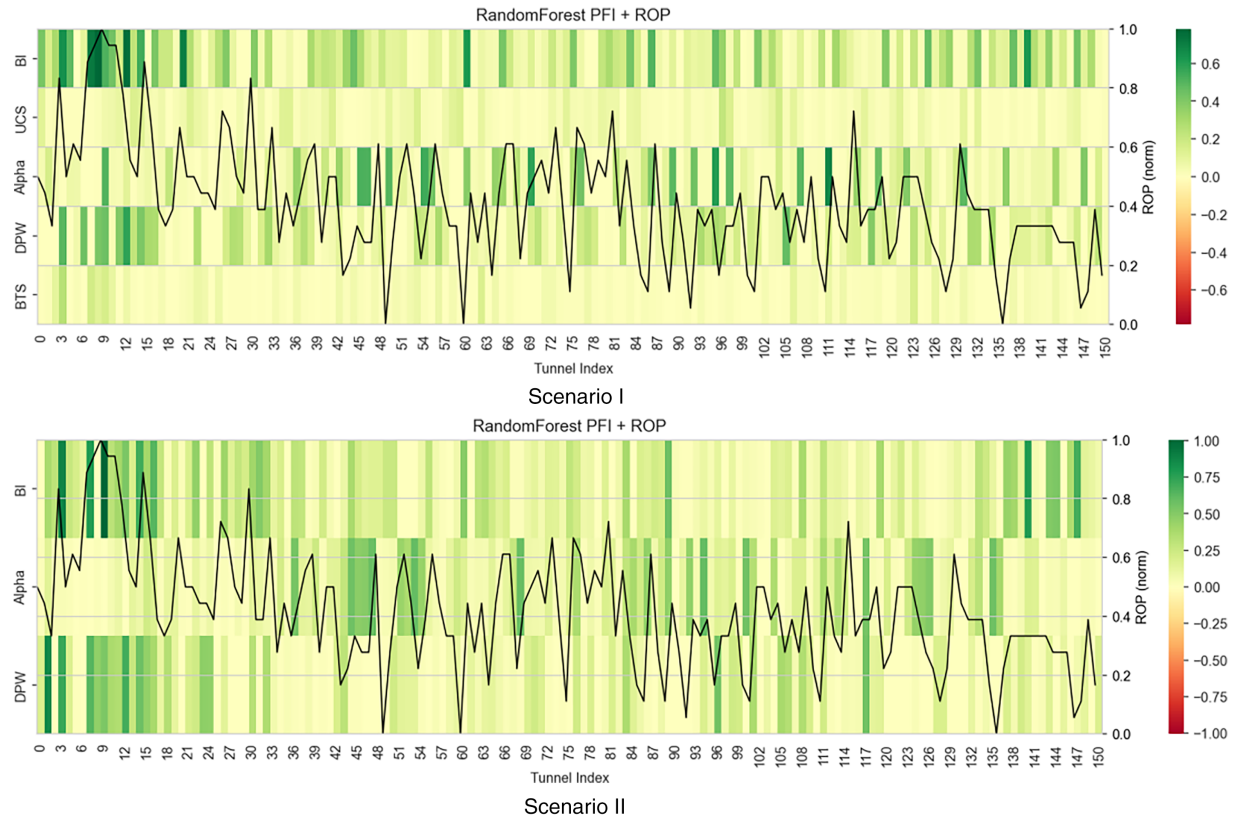


Figure 13: Comparative heatmap illustrating the distribution of PFI-based feature importance across Scenario I and Scenario II.

The primary motivation for selecting this dataset lies in its structural compatibility with the Queens dataset [3,13,18], derived from field and laboratory measurements of a tunnel project located in the southwestern region of New York, as discussed in detail in the previous sections. Specifically, both datasets share similar input and output variables. However, they exhibit notable differences in geotechnical characteristics, data distributions, and sample sizes. This diversity enables a comprehensive assessment of the proposed framework under heterogeneous data conditions and allows for a more rigorous evaluation of its predictive behaviour.

The analysis conducted on these two independent datasets, representing distinct geological and geotechnical environments, is particularly important in demonstrating that the proposed approach is not dataset-specific, but rather capable of producing consistent and reliable results across varying data structures. In this sense, the validation on the second dataset substantiates the generalizability of the proposed framework and contributes to the growing body of literature on explainable machine learning applications in geotechnical engineering.

Both datasets include the uniaxial compressive strength (UCS), Brazilian tensile strength (BTS), rock brittleness index (BI), discontinuity spacing (DPW), discontinuity orientation (Alpha), and tunnel boring machine rate of penetration variables. To ensure comparability of results, all preprocessing and feature reduction steps applied to the Queens dataset were identically replicated for the second dataset. Since the descriptive statistics, outlier analyses, spatial variations of variables along the tunnel alignment, and VIF

results for the second dataset are comprehensively reported in [19], the present study focuses exclusively on modeling and explainability analyses.

The Random Forest (RF), Bagged Trees (BT), and LSBoost models originally employed in [19] were also evaluated in this study. However, the Support Vector Machines (SVM) model used in the previous work was replaced with the Generalized Additive Model (GAM). Accordingly, the predictive performance of the GAM was compared against RF, BT, and LSBoost models in a consistent experimental setting.

Furthermore, a comparative assessment was conducted for both the full-feature (five variables) and VIF-reduced (three variables) configurations using common performance metrics (R^2 , RMSE, MAE, and MAPE), as well as feature importance and interpretability tools, including SHAP, Permutation Feature Importance (PFI), and Partial Dependence Plots (PDP). This design enabled a detailed examination of how feature reduction influences both predictive performance and feature importance distributions, while simultaneously demonstrating that the proposed modeling and explainability framework yields consistent outcomes across datasets with differing geological and geotechnical characteristics.

3.3.1 Scenario-Based Model Performance Comparison for the Karaj–Tehran Dataset

In this section, the performance of the proposed modeling framework is evaluated on the Karaj–Tehran dataset under different variable configurations. To this end, two distinct scenarios are defined: a five-variable configuration obtained prior to VIF analysis (Scenario I) and a reduced three-variable configuration obtained after VIF analysis (Scenario II). Within both scenarios, the performances of the Generalized Additive Model (GAM), Random Forest (RF), Bagged Trees (BT), and LSBoost algorithms are compared in a unified experimental setting.

To ensure a fair and unbiased comparison among all models, the dataset was first divided into training and independent test subsets. Five-fold cross-validation was performed exclusively within the training subset to determine the optimal hyperparameter combinations for each model. Following hyperparameter selection, each model was retrained using the complete training subset with the selected hyperparameters, and its final predictive performance was evaluated on the independent test dataset that was not used during model training or hyperparameter optimization. The reported performance metrics, including R^2 , RMSE, MAE, and MAPE, were therefore calculated from predictions on the independent test dataset. Accordingly, the values presented in Tables 4–6 represent out-of-sample test-set prediction performance rather than cross-validation or training performance. This evaluation procedure was adopted to provide an unbiased assessment of model generalization and to minimize the risk of optimistic performance estimates resulting from model fitting and hyperparameter selection.

Table 4: Scenario-based model performance comparison for the Karaj–Tehran dataset (VIF: 5 vs. 3 Variables).

Scenario	Basis Dimension (K)	Model	R^2	RMSE	MAE	MAPE
I	15	LSBoost	0.9943	0.0647	0.0484	1.7678
		GAM	0.9829	0.1127	0.0802	3.3086
		RandomForest	0.8796	0.2985	0.2362	9.2272
		Bagging Trees	0.8209	0.3641	0.2651	10.3768
II	15	LSBoost	0.9744	0.1375	0.0994	3.6148
		RandomForest	0.8683	0.3122	0.2428	9.6667
		Bagging Trees	0.7738	0.4092	0.2857	10.8095
		GAM	0.0002	0.8602	0.7095	29.0252

Table 5: Effect of basis dimension (K) on GAM performance under VIF scenarios for the Tehran–Karaj dataset.

Scenario	Basis Dimension (K)	R ²	RMSE	MAE	MAPE
I	15	0.9829	0.1127	0.0802	3.3086
	15	0.0002	0.8602	0.7095	29.0252
	20	−0.0036	0.8618	0.7075	28.9817
	25	−0.0205	0.8691	0.7126	29.1064
	26	−0.0548	0.8835	0.7159	28.4218
	27	−0.0556	0.8839	0.7160	28.3579
II	28	0.9271	0.2323	0.1554	5.9371
	29	0.9200	0.2434	0.1545	5.9737
	30	0.9781	0.1273	0.0816	3.0798
	35	0.9867	0.0990	0.0618	2.3400
	40	0.9882	0.0933	0.0581	2.2333
	45	0.9907	0.0830	0.0536	2.0866

Table 6: Effect of basis dimension (K) on GAM performance before and after VIF-based feature reduction for the New York queens dataset.

Scenario	Basis Dimension (K)	R ²	RMSE	MAE	MAPE
I	50	0.9874	0.0404	0.0307	1.5617
	50	0.5798	0.2330	0.1859	9.1620
	60	0.7226	0.1894	0.1559	7.6843
	70	0.7148	0.1920	0.1558	7.6671
	80	0.6987	0.1974	0.1620	7.9736
II	85	0.6841	0.2021	0.1648	8.1271
	86	0.6955	0.2016	0.1650	8.1442
	87	0.9752	0.0567	0.0410	2.1338
	88	0.9784	0.0528	0.0388	2.0193
	90	0.9839	0.0456	0.0324	1.7074

Overall evaluation of [Table 4](#) indicates that all models achieve substantially better performance under the five-variable configuration (Scenario I) prior to VIF-based feature reduction. In particular, the LSBoost model consistently demonstrates the best performance in both scenarios, yielding the lowest error values and the highest R², thereby indicating a strong predictive capability.

In Scenario I, the GAM model outperforms both Random Forest and Bagged Trees; however, it remains inferior to LSBoost. In contrast, after VIF-based reduction to a three-variable structure (Scenario II), the GAM model experiences a severe deterioration in performance, with the R² value dropping to nearly zero and error metrics increasing significantly.

This behaviour suggests that the GAM is more sensitive not only to feature reduction but also to the weakening of the underlying relational structure within the dataset. In particular, spline-based formulations require a sufficiently informative combination of explanatory variables, and the reduction in feature space substantially limits the functional approximation capacity of the model. The marked decline observed

under Scenario II should therefore not be interpreted solely as a consequence of reducing the number of explanatory variables. Rather, it also reflects that the default basis dimension ($K = 15$) is insufficient to adequately represent the nonlinear relationships remaining after VIF-based feature elimination. Accordingly, the substantial reduction in predictive performance motivated a systematic sensitivity analysis of the basis dimension parameter, presented in Table 5, to determine whether the observed behaviour originated from limited spline flexibility or from the intrinsic information loss caused by feature reduction.

By contrast, the LSBoost model maintains high and stable performance across both scenarios, indicating that boosting-based approaches possess stronger generalization capability under varying feature spaces. Random Forest and Bagged Trees models exhibit moderate performance, with a noticeable but less pronounced decline in predictive accuracy following feature reduction. These results further indicate that ensemble tree-based algorithms are comparatively less sensitive to changes in feature dimensionality than spline-based GAM formulations.

To further investigate this sensitivity of the GAM model, a basis dimension (K)-dependent sensitivity analysis was conducted, and the results are summarized in Table 5. Within this framework, model performance was comparatively evaluated across different values of K under both the pre- and post-VIF data structures.

The results presented in Table 5 indicate that the predictive behaviour of the GAM model is highly dependent on the selected basis dimension (K), particularly under the reduced three-variable configuration (Scenario II). Under the pre-VIF scenario (Scenario I), a high level of predictive accuracy is achieved with $K = 15$ ($R^2 = 0.9829$), suggesting that the original five-variable dataset provides sufficient explanatory information for an accurate spline representation without requiring additional model flexibility.

In contrast, after VIF-based feature reduction, the same basis dimension results in a substantial decrease in predictive performance. When K is varied between 20 and 27, the model continues to exhibit poor predictive capability, with negative or near-zero R^2 values. These findings indicate that, within this interval, the spline basis remains insufficient to adequately capture the nonlinear relationships preserved in the reduced feature space.

A pronounced improvement is observed when the basis dimension increases from $K = 27$ to $K = 28$, where the R^2 value rises sharply. Rather than interpreting this abrupt transition as evidence of an immediate improvement in the intrinsic predictive capability of the model, it should be regarded as a threshold behaviour associated with the spline representation. Increasing the basis dimension expands the functional flexibility of the GAM, enabling the model to represent nonlinear relationships that cannot be captured under more restrictive spline configurations. Similar nonlinear transitions in model behaviour have been reported for spline-based learning algorithms when the available basis functions become sufficient to approximate the underlying response surface.

Although predictive performance continues to improve for larger K values, these results should also be interpreted with caution. Increasing the basis dimension enhances model flexibility but simultaneously increases model complexity, which may elevate the risk of overfitting, particularly for datasets with relatively limited sample sizes such as the Karaj–Tehran dataset. For this reason, the performance values reported in Table 5 were obtained using five-fold cross-validation, and all evaluation metrics correspond exclusively to the test subsets rather than the training data. Consequently, the reported improvements represent predictive performance on unseen observations instead of goodness-of-fit to the training samples, thereby reducing the possibility that the observed increase in R^2 is solely attributable to overfitting.

Nevertheless, the consistently increasing performance observed for higher K values suggests that selecting excessively flexible spline structures without independent validation may still compromise model

generalizability. Therefore, the purpose of the present sensitivity analysis is not to recommend the largest possible basis dimension, but rather to demonstrate that the predictive behaviour of GAM is strongly governed by appropriate hyperparameter specification. In practical applications, the basis dimension should be determined through systematic validation procedures that balance predictive accuracy and model complexity.

Overall, the findings demonstrate that the reduced-feature scenario requires greater spline flexibility than the original feature configuration. More importantly, they indicate that the deterioration observed under the default $K = 15$ setting primarily reflects insufficient model capacity rather than an inherent limitation of the GAM methodology itself. Accordingly, the basis dimension should be considered a critical calibration parameter when GAM is applied to geotechnical datasets characterized by nonlinear relationships and varying feature dimensionality.

To determine whether these observations are specific to the Karaj–Tehran dataset or represent a more general characteristic of GAM, the same basis-dimension sensitivity analysis was subsequently performed using the independent Queens dataset. Applying an identical experimental design enables the robustness of the observed K -dependent behaviour to be evaluated across two datasets with different geological characteristics and sample sizes. The corresponding results are presented in Table 6.

The results presented in Table 6 for the New York Queens dataset provide important evidence regarding the combined effects of VIF-based feature reduction and basis dimension on GAM performance. Under Scenario I, in which the complete set of input variables is retained, the GAM achieves a high predictive performance at $K = 50$ ($R^2 = 0.9874$, $RMSE = 0.0404$, $MAE = 0.0307$, and $MAPE = 1.5617$). The same basis dimension provides a direct comparison after VIF-based feature reduction: under Scenario II, where UCS and BTS are excluded, the performance at $K = 50$ decreases substantially to $R^2 = 0.5798$, $RMSE = 0.2330$, $MAE = 0.1859$, and $MAPE = 9.1620$. Thus, the two $K = 50$ entries in Table 6 are intentional and demonstrate the substantial effect of feature reduction on the predictive capability of the GAM under otherwise comparable basis-dimension settings.

Following feature reduction, the GAM exhibits relatively low predictive performance for $K = 50$ –86, with R^2 values ranging from 0.5798 to 0.7226. This indicates that the reduced three-variable configuration substantially limits the ability of the GAM to represent the underlying nonlinear relationships at these basis dimensions. However, a marked improvement is observed when the basis dimension reaches $K = 87$, where R^2 increases to 0.9752, followed by similarly high performance at $K = 88$ and $K = 90$ ($R^2 = 0.9784$ and 0.9839, respectively). These results indicate that increasing the spline basis dimension can compensate, to some extent, for the loss of explanatory information resulting from feature reduction by providing greater flexibility in representing the nonlinear response surface.

All performance metrics reported in Table 6 were obtained using the evaluation procedure described in Section 2.3.1. Specifically, five-fold cross-validation was used within the training subset for hyperparameter optimization, whereas the final performance metrics were calculated from predictions on the independent test dataset. Therefore, the R^2 , $RMSE$, MAE , and $MAPE$ values reported in Table 6 represent out-of-sample predictive performance rather than training accuracy or cross-validation test-fold performance.

The contrasting results between Scenarios I and II demonstrate that GAM performance is influenced not only by the basis dimension but also by the information content of the input variables. In particular, the comparison at $K = 50$ shows that VIF-based feature reduction, although beneficial for reducing multicollinearity and simplifying the model input structure, can substantially reduce predictive performance when applied to GAM. The subsequent recovery of performance at $K \geq 87$ suggests that greater spline flexibility can partially compensate for the information lost through feature reduction. Nevertheless, this improvement

should not be interpreted as evidence that larger K values are universally preferable. The optimal basis dimension remains dependent on the dataset, the available explanatory variables, and the underlying nonlinear relationships. Therefore, K should be treated as a dataset- and model-specific hyperparameter that requires systematic optimization and independent validation to achieve an appropriate balance among predictive accuracy, model flexibility, and generalization capability.

3.3.2 Scenario-Based Feature Importance Analysis for the Karaj–Tehran Dataset and Comparison with the Queens Dataset

Feature importance analyses for the Karaj–Tehran dataset were conducted under both the pre-VIF (five-variable) and post-VIF (three-variable) scenarios using SHAP, Permutation Feature Importance (PFI), and Partial Dependence Plot (PDP) methods. The obtained results are presented comparatively with those of the Queens dataset in [Tables 7](#) and [8](#). This comparative analysis aims to investigate how VIF-based feature reduction influences feature importance patterns, model interpretability, and the relative ranking of explanatory variables across two independent geotechnical datasets.

Table 7: Comparison of SHAP-, PFI-, and PDP-based feature importance scores for five-input-variable models (BI, UCS, Alpha, DPW, and BTS) using the Quens and Tehran–Karaj datasets.

Model	Feature	Quens Data			Tehran-Karaj Data		
		SHAP (%)	PFI (%)	PDP (%)	SHAP (%)	PFI (%)	PDP (%)
RF	BI	40.584	41.860	27.556	21.316	19.740	5.339
	UCS	6.811	6.878	15.868	20.838	18.489	27.763
	Alpha	25.041	22.688	34.352	11.401	8.802	14.419
	DPW	23.363	22.622	13.021	22.615	25.503	31.283
	BTS	4.200	5.952	9.203	23.829	27.467	21.196
LSBoost	BI	36.829	35.273	20.574	21.519	18.447	8.871
	UCS	10.438	13.257	13.150	17.115	21.206	18.292
	Alpha	22.590	23.420	42.043	13.444	12.027	18.381
	DPW	25.166	21.359	13.236	25.414	27.428	32.592
	BTS	4.978	6.691	10.997	22.508	20.892	21.864
BT	BI	38.828	33.342	23.373	15.249	16.434	12.659
	UCS	7.574	8.222	19.305	23.047	27.337	38.048
	Alpha	24.400	28.960	35.415	15.379	9.908	9.533
	DPW	23.931	22.252	6.904	18.434	20.129	26.586
	BTS	5.267	7.225	15.003	27.890	26.193	13.174
GAM	BI	38.528	34.943	40.917	20.049	23.586	21.566
	UCS	12.877	12.735	14.027	21.185	21.004	25.307
	Alpha	19.609	22.657	22.275	13.187	11.448	9.948
	DPW	24.555	25.215	18.704	20.454	21.583	17.803
	BTS	4.431	4.451	4.076	25.125	22.379	25.376

Table 8: Comparison of SHAP-, PFI-, and PDP-based feature importance scores for three-input-variable models (BI, Alpha, and DPW) using the Queens and Tehran–Karaj datasets.

Model	Feature	Queens Data			Tehran-Karaj Data		
		SHAP (%)	PFI (%)	PDP (%)	SHAP (%)	PFI (%)	PDP (%)
RF	BI	43.215	40.139	31.393	31.872	34.031	19.175
	Alpha	28.029	32.015	49.462	19.983	22.333	13.157
	DPW	28.757	27.846	19.145	48.145	43.636	67.669
LSBoost	BI	40.559	40.290	33.266	32.835	33.046	20.514
	Alpha	28.479	29.611	50.083	22.626	15.051	17.863
	DPW	30.961	30.099	16.650	44.539	51.903	61.623
BT	BI	41.718	42.223	36.026	32.098	33.212	11.693
	Alpha	28.010	27.108	45.649	24.838	16.983	20.935
	DPW	30.273	30.668	18.325	43.064	49.805	67.372
GAM	BI	30.903	29.871	31.175	32.719	31.530	34.944
	Alpha	30.846	34.444	42.445	22.993	24.438	12.792
	DPW	38.252	35.685	26.380	44.289	44.031	52.264

Table 7 presents the feature importance results obtained from models developed using five input variables (BI, UCS, Alpha, DPW, and BTS). The results demonstrate that the Queens and Karaj–Tehran datasets exhibit distinctly different feature importance patterns despite sharing the same explanatory variables. In the Queens dataset, the BI variable consistently emerges as the dominant predictor across all machine learning models, yielding the highest importance scores according to both SHAP and PFI analyses. In contrast, the Karaj–Tehran dataset exhibits a more balanced distribution of feature importance, with DPW, BTS, and UCS collectively contributing more substantially to the prediction of ROP.

These differences are consistent with the distinct statistical distributions and geotechnical characteristics of the two datasets. In the Queens dataset, BI consistently appears as the dominant explanatory variable across the applied explainability methods, whereas in the Karaj–Tehran dataset, the greater variability of DPW appears to provide additional discriminatory information for the learning algorithms, thereby increasing its contribution to predictive performance. Furthermore, PDP analysis suggests that DPW and UCS exhibit stronger nonlinear marginal effects on ROP than the remaining explanatory variables in the Karaj–Tehran dataset.

Following VIF-based feature reduction, the explanatory variable set was reduced from five to three variables (BI, Alpha, and DPW), and the corresponding feature importance results are summarized in Table 8.

In the Queens dataset, the BI variable maintained its position as the most influential predictor according to the SHAP and PFI analyses, whereas the PDP results indicate that the Alpha variable exhibits the strongest marginal effect on the predicted ROP. This observation suggests that the influence of Alpha is characterized by a more pronounced nonlinear response despite its comparatively lower global importance.

In the Karaj–Tehran dataset, under the three-variable scenario, the DPW variable consistently achieved the highest importance scores across all machine learning models and all explainability methods. The strong agreement among the SHAP, PFI, and PDP analyses further supports the dominant contribution of DPW to ROP prediction within the reduced feature space. Moreover, the increase in the relative importance of

DPW following the removal of UCS and BTS suggests that DPW becomes the primary explanatory variable after feature reduction, likely absorbing part of the predictive information previously distributed among the removed variables.

Overall, although moderate differences exist in the magnitude of feature importance among the RF, LSBoost, Bagged Trees, and GAM models, the relative ranking of the explanatory variables remains largely consistent. These findings suggest that the overall feature importance pattern is governed primarily by the intrinsic characteristics of the dataset rather than by the specific machine learning algorithm employed. Collectively, the results indicate a BI-dominated importance structure for the Queens dataset, whereas discontinuity-related geotechnical parameters—particularly DPW—play a more influential role in explaining tunnel boring machine penetration performance in the Karaj–Tehran dataset.

4 Discussion

The present study comprehensively evaluated machine learning models for tunnel boring machine (TBM) rate of penetration (ROP) prediction from both predictive and interpretability perspectives. In addition to assessing model performance on the New York Queens dataset, the proposed modeling and explainability framework was independently validated using the Tehran–Karaj dataset. By analysing two geotechnical datasets characterized by different geological conditions, statistical distributions, and sample sizes, the robustness and broader applicability of the proposed framework were systematically investigated.

The results indicate that ensemble learning algorithms provide consistently high predictive performance for ROP estimation across both datasets. Random Forest, LSBoost, and Bagged Trees maintained stable predictive capability under both the original and VIF-reduced feature configurations, suggesting that ensemble tree-based methods effectively capture the complex nonlinear relationships between rock properties and drilling performance. Their relatively limited sensitivity to feature reduction further indicates that the retained variables preserve sufficient explanatory information, allowing these models to maintain strong predictive capability despite reduced feature dimensionality. This behaviour is consistent with the ability of ensemble learning algorithms to model nonlinear interactions while remaining comparatively robust to multicollinearity and variable redundancy.

A markedly different response was observed for the Generalized Additive Model (GAM). Although VIF-based feature reduction initially appeared to substantially reduce predictive performance, the systematic basis-dimension sensitivity analysis demonstrated that insufficient spline flexibility constitutes a major contributing factor to this behaviour. For both datasets, increasing the basis dimension (K) substantially improved model performance, indicating that the deterioration observed under the default parameter settings primarily reflects limited model capacity rather than an inherent limitation of the GAM methodology itself. These findings highlight that spline flexibility and feature dimensionality should be considered jointly when applying GAM to geotechnical datasets characterized by nonlinear relationships.

An important observation is that this behaviour was consistently reproduced in two independent datasets with distinctly different geological characteristics and sample sizes. Although the optimal basis dimension differed between datasets, both exhibited a similar transition from poor to high predictive performance as spline flexibility increased. This consistency suggests that the interaction between feature dimensionality and spline complexity plays a fundamental role in determining GAM performance. Consequently, the basis dimension should be regarded as a dataset-specific hyperparameter requiring systematic optimization to achieve an appropriate balance between predictive accuracy, model complexity, and generalization capability.

The explainability analyses further strengthen the proposed framework by demonstrating consistent feature importance patterns across multiple interpretation techniques. SHAP, Permutation Feature Importance (PFI), Partial Dependence Plots (PDP), and Local Interpretable Model-Agnostic Explanations (LIME) provided complementary perspectives on model behaviour while yielding mutually consistent interpretations. The agreement between global explanation methods (SHAP, PFI, and PDP) and local explanations (LIME) increases confidence that the relationships learned by the models reflect meaningful patterns within the datasets rather than artifacts associated with a particular interpretation technique.

To further support the LIME interpretation, the average values of the representative LOW, MEAN, and HIGH ROP segments are summarized in [Table 9](#).

Table 9: Average values of variables used in LIME representative segments.

ROP (Lime Analizi)	BI	UCS	Alpha	DPW	BTS
Low	29.133	134.833	6.667	1.467	9.900
Mean	41.500	159.767	43.000	0.933	9.533
High	56.333	185.267	46.000	0.333	10.233

The observed trends are consistent with the global interpretations obtained from SHAP and PDP analyses. Specifically, increasing BI and decreasing DPW are associated with higher predicted ROP values, whereas BTS exhibits comparatively limited variation across the representative instances. The agreement between global and local explanation methods further strengthens the reliability of the proposed explainability framework.

Despite this methodological consistency, the relative importance of the explanatory variables differed considerably between the two datasets. For the New York Queens dataset, the rock brittleness index (BI) consistently emerged as the dominant explanatory variable across nearly all models and explainability methods, indicating that rock brittleness plays a primary role in controlling drilling performance under the geological conditions represented by this dataset. In contrast, the Tehran–Karaj dataset exhibited a more balanced importance structure, in which discontinuity spacing (DPW), uniaxial compressive strength (UCS), and Brazilian tensile strength (BTS) contributed more substantially to ROP prediction. Following VIF-based feature reduction, DPW became the dominant explanatory variable, suggesting that discontinuity-related geotechnical characteristics exert a stronger influence on drilling performance within the reduced feature space. These observations indicate that feature importance is governed primarily by the intrinsic geological characteristics of the dataset rather than by the specific machine learning algorithm employed.

The successful application of the proposed framework to two independent datasets also provides evidence supporting its applicability across different geotechnical environments. Although only two case studies were considered, the consistent modelling behaviour, explainability results, and hyperparameter sensitivity patterns observed in both datasets suggest that the proposed workflow is not restricted to a single geological setting. Nevertheless, additional validation using larger datasets representing different lithological formations, TBM configurations, and excavation conditions would further strengthen the generalizability of the proposed framework.

Overall, the findings indicate that ensemble learning methods provide reliable and robust solutions for ROP prediction owing to their consistently high predictive performance and limited sensitivity to feature reduction. At the same time, the results demonstrate that GAM can achieve comparable predictive accuracy when an appropriate basis dimension is selected while providing an inherently interpretable additive modeling framework. Therefore, model selection for ROP prediction should consider not only

predictive accuracy but also dataset characteristics, interpretability requirements, and the sensitivity of model performance to hyperparameter specification.

5 Conclusions

This study developed and evaluated machine learning-based models, including Random Forest (RF), LSBoost, Bagged Trees (BT), and the Generalized Additive Model (GAM), for tunnel boring machine (TBM) rate of penetration (ROP) prediction. Beyond predictive performance, the proposed framework integrates multiple explainability techniques, including SHAP, Permutation Feature Importance (PFI), Partial Dependence Plots (PDP), and Local Interpretable Model-Agnostic Explanations (LIME), to provide both accurate predictions and interpretable insights into model behaviour. Furthermore, the proposed workflow was independently validated using two geotechnical datasets with different geological characteristics, namely the New York Queens and Tehran–Karaj datasets.

The main findings of this study can be summarized as follows:

1. Ensemble learning models generally demonstrated strong and robust predictive performance across both datasets. RF and LSBoost provided highly accurate and stable predictions, while BT also showed reliable predictive capability. However, GAM achieved the highest predictive performance for the Queens dataset under Scenario I, whereas its performance deteriorated substantially following VIF-based feature reduction. This indicates that the relative performance of ensemble and non-ensemble models depends on the dataset, input configuration, and model structure. In contrast, the relatively limited effect of VIF-based feature reduction on the ensemble tree-based models indicates that RF, LSBoost, and BT are comparatively robust to moderate reductions in feature dimensionality.
2. **The explainability analyses produced mutually consistent interpretations of model behaviour.** SHAP, PFI, PDP, and LIME consistently identified the most influential variables governing ROP prediction, while the agreement between global and local explanation methods increased confidence in the reliability of the learned relationships.
3. **Feature importance patterns were found to be dataset-dependent.** BI was identified as the dominant explanatory variable for the New York Queens dataset, whereas discontinuity-related geotechnical parameters, particularly DPW, played a more influential role in the Tehran–Karaj dataset. These findings indicate that the relative importance of explanatory variables is governed primarily by the geological characteristics of the dataset rather than by the choice of machine learning algorithm.
4. **The basis dimension (K) was identified as a critical hyperparameter for GAM.** Although the reduced-feature scenarios initially resulted in poor predictive performance, systematic sensitivity analyses demonstrated that increasing spline flexibility substantially improved model accuracy in both datasets. These results indicate that the deterioration observed under the default parameter settings primarily reflects insufficient model capacity rather than an inherent limitation of the GAM methodology.

Overall, the successful application of the proposed framework to two independent geotechnical datasets provides evidence supporting its applicability under different geological conditions. The results further demonstrate that combining machine learning with explainable artificial intelligence offers a practical approach for achieving both reliable prediction and transparent model interpretation in TBM performance analysis. While ensemble learning methods provide highly robust predictive solutions, GAM represents a promising interpretable alternative when its basis dimension is appropriately optimized.

Future research should validate the proposed framework using larger datasets representing different lithological formations, TBM configurations, and excavation conditions, while also investigating automated

strategies for optimizing GAM hyperparameters and extending the framework to additional explainable machine learning techniques.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Halil Karahan, Devrim Alkaya; data collection: Halil Karahan, Devrim Alkaya; analysis and interpretation of results: Halil Karahan; draft manuscript preparation: Halil Karahan, Devrim Alkaya. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed in this study were obtained from the publicly available sources cited in the manuscript [3,23].

Ethics Approval: Not applicable in this section.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tarkoy PJ. Rock hardness index properties and geotechnical parameters for predicting tunnel boring machine performance [dissertation]. Champaign, IL, USA: University of Illinois at Urbana-Champaign; 1975.
2. Nelson P. Tunnel boring machine performance in sedimentary rocks [dissertation]. Ithaca, NY, USA: Cornell University; 1983.
3. Yagiz S. Utilizing rock mass properties for predicting TBM performance in hard rock condition. *Tunn Undergr Space Technol.* 2008;23(3):326–39. doi:10.1016/j.tust.2007.04.011.
4. Yagiz S, Karahan H. Prediction of hard rock TBM penetration rate using particle swarm optimization. *Int J Rock Mech Min Sci.* 2011;48(3):427–33. doi:10.1016/j.ijrmms.2011.02.013.
5. Yagiz S, Karahan H. Application of various optimization techniques to predict TBM penetration rate in rock mass and comparison of their performance. *Int J Rock Mech Min Sci.* 2015;80(4):308–15. doi:10.1016/j.ijrmms.2015.09.019.
6. Khoshzaher E, Chakeri H, Bazargan S, Mousapour H. The prediction of EPB-TBM performance using firefly algorithms and particle swarm optimization. *Min Geol Pet Eng Bull.* 2023;38(5):79–86. doi:10.17794/rgn.2023.5.7.
7. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273–97. doi:10.1007/BF00994018.
8. Awad M, Khanna R. Efficient learning machines: theories, concepts, and applications for engineers and system designers. Berkeley, CA, USA: Apress; 2015.
9. Yaghoubi E, Yaghoubi E, Khamees A, Vakili AH. A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Comput Appl.* 2024;36(21):12655–99. doi:10.1007/s00521-024-09893-7.
10. Mackay CT, Nowell D. Informed machine learning methods for application in engineering: a review. *Proc Inst Mech Eng Part C J Mech Eng Sci.* 2023;237(24):5801–18. doi:10.1177/09544062231164575.
11. Mehdiyev N, Majlatow M, Fettke P. Interpretable and explainable machine learning methods for predictive process monitoring: a systematic literature review. *Artif Intell Rev.* 2025;58(12):378. doi:10.1007/s10462-025-11399-0.
12. Tian Y, Shi Y, Liu X. Recent advances on support vector machines research. *Technol Econ Dev Econ.* 2012;18(1):5–33. doi:10.3846/20294913.2012.661205.
13. Karahan H, Alkaya D. Integrating SVR optimization and machine learning-based feature importance for TBM penetration rate prediction. *Appl Sci.* 2025;16(1):355. doi:10.3390/app16010355.
14. Minh VT, Katushin D, Antonov M, Veinthal R. Regression models and fuzzy logic prediction of TBM penetration rate. *Open Eng.* 2017;7(1):60–8. doi:10.1515/eng-2017-0012.

15. Arbabsiar MH, Ali Ebrahimi Farsangi M, Mansouri H. Fuzzy logic modelling to predict the level of geotechnical risks in rock tunnel boring machine (TBM) tunnelling. *Min Geol Pet Eng Bull.* 2020;35(2):1–14. doi:10.17794/rgn.2020.2.1.
16. 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(3):1513–36. doi:10.1007/s10706-021-01982-x.
17. Drucker H. Improving regressors using boosting techniques. In: *Proceedings of the 14th International Conference on Machine Learning*; 1997 Jul 8–12; Nashville, TN, USA.
18. Karahan H, Alkaya D. Input variable effects on TBM penetration rate: parametric and machine learning models. *Appl Sci.* 2026;16(3):1301. doi:10.3390/app16031301.
19. Karahan H, Alkaya D. Ensemble machine learning for predicting TBM penetration rate with limited geotechnical data. *Appl Sci.* 2026;16(7):3451. doi:10.3390/app16073451.
20. Bo Y, Guo X, Liu Q, Pan Y, Zhang L, Lu Y. Prediction of tunnel deformation using PSO variant integrated with XGBoost and its TBM jamming application. *Tunn Undergr Space Technol.* 2024;150(11):105842. doi:10.1016/j.tust.2024.105842.
21. Wang Y, Zhao J, Jiang K, Zhou Q, Kang Z, Chen C, et al. Prediction of TBM operation parameters using machine learning models based on BPSO. *Adv Eng Inform.* 2023;56(7):101955. doi:10.1016/j.aei.2023.101955.
22. Yang J, Yagiz S, Liu YJ, Laouafa F. Comprehensive evaluation of machine learning algorithms applied to TBM performance prediction. *Undergr Space.* 2022;7(1):37–49. doi:10.1016/j.undsp.2021.04.003.
23. Mansouri M, Moomivand H. Influence of rock mass properties on TBM penetration rate in Karaj-Tehran water conveyance tunnel. *J Geol Min Res.* 2010;2(5):114–21.
24. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied linear statistical models.* 5th ed. Boston, MA, USA: McGraw-Hill Irwin; 2005.
25. Belsley DA, Kuh E, Welsch RE. *Regression diagnostics: identifying influential data and sources of collinearity.* Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2005.
26. Montgomery DC, Peck EA, Vining GG. *Introduction to linear regression analysis.* 6th ed. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2021.
27. Hair JF Jr, Black WC, Babin BJ, Anderson RE. *Multivariate data analysis.* 8th ed. Boston, MA, USA: Cengage Learning; 2019.
28. Goodfellow I, Bengio Y, Courville A. *Deep learning.* Cambridge, MA, USA: MIT Press; 2016.
29. Bahri M, Romero-Hernández R, Mascort-Albea EJ, 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(2):70. doi:10.1007/s10706-024-02969-0.
30. Liu Z, Yuan Y, Zhou H, Yang Q. Real-time rock mass classification in TBM tunneling using a dynamically weighted LightGBM model optimized by Bayesian algorithm. *J Rock Mech Geotech Eng.* 2026. doi:10.1016/j.jrmge.2026.01.020.
31. Nie Y, Zhang Q, Hou L, Du L, Zhao X, Xue Y, et al. TBM rock mass classification using XGBoost and interpretable machine learning. *Adv Eng Inform.* 2025;66:103459. doi:10.1016/j.aei.2025.103459.
32. Hu W, Wu K, Liu H, Luo W, Li X, Guan P. Interpretable machine learning approach for TBM tunnel crown convergence prediction with Bayesian optimization. *Front Earth Sci.* 2025;13:1608468. doi:10.3389/feart.2025.1608468.
33. Liaw A, Wiener M. Classification and regression by randomForest. *R News.* 2002;2(3):18–22.
34. RColorBrewer S, Liaw MA. Package ‘randomforest’. Berkeley, CA, USA: University of California, Berkeley; 2018. p. 1–29.
35. Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: *The elements of statistical learning: data mining, inference, and prediction.* New York, NY, USA: Springer New York; 2008.
36. Alatefi S, Almeshal AM. A new model for estimation of bubble point pressure using a Bayesian optimized least square gradient boosting ensemble. *Energies.* 2021;14(9):2653. doi:10.3390/en14092653.
37. Breiman L. Bagging predictors. *Mach Learn.* 1996;24(2):123–40. doi:10.1023/A:1018054314350.

38. Hastie T, Tibshirani R. Generalized additive models. *Statist Sci.* 1986;1(3):297–310. doi:10.1214/ss/1177013604.
39. Wood SN. Generalized additive models. *Annu Rev Stat Appl.* 2025;12(1):497–526. doi:10.1146/annurev-statistics-112723-034249.
40. Molnar C, Freiesleben T, König G, Herbringer J, Reisinger T, Casalicchio G, et al. Relating the partial dependence plot and permutation feature importance to the data generating process. In: *Explainable artificial intelligence*. Cham, Switzerland: Springer; 2023. p. 456–79.
41. Ning Y, Ong MEH, Chakraborty B, Goldstein BA, Ting DSW, Vaughan R, et al. Shapley variable importance cloud for interpretable machine learning. *Patterns.* 2022;3(4):100452. doi:10.1016/j.patter.2022.100452.
42. Sioutas KN, Benardos A. Boosting model interpretability for transparent ML in TBM tunneling. *Appl Sci.* 2024;14(23):11394. doi:10.3390/app142311394.
43. Huang K, Zhao Y, Xiao P, Wang H, Xiong Q, Yu J, et al. Performance comparison and interpretability of machine learning models for TBM penetration rate prediction. *Front Earth Sci.* 2026;14:1803020. doi:10.3389/feart.2026.1803020.
44. Villa-Blanco C, Bielza C, Larrañaga P. Feature subset selection for data and feature streams: a review. *Artif Intell Rev.* 2023;56(1):1011–62. doi:10.1007/s10462-023-10546-9.
45. Theng D, Bhojar KK. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowl Inf Syst.* 2024;66(3):1575–637. doi:10.1007/s10115-023-02010-5.
46. Zhou J, Liu Z, Li C, Du K, Khandelwal M. Interpretable machine learning for predicting cutting force of tunnel boring machine: a study on SVR with meta-heuristic optimization. *Int J Numer Anal Meth Geomech.* 2025;49(18):4412–35. doi:10.1002/nag.70083.
47. Shapley LS. A value for n-person games. Princeton, NJ, USA: Princeton University Press; 1953.
48. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017 Dec 4–9; Long Beach, CA, USA.
49. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics.* 2010;26(10):1340–7. doi:10.1093/bioinformatics/btq134.
50. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res.* 2019;20(301):177. doi:10.1080/01621459.1963.10500830.
51. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist.* 2001;29(5):1189–232. doi:10.1214/aos/1013203451.
52. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. *Proc AAAI Conf Artif Intell.* 2018;32(1):1–9. doi:10.1609/aaai.v32i1.11491.
53. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv Intell Syst.* 2025;7(1):2400304. doi:10.1002/aisy.202400304.
54. Garreau D, Luxburg U. Explaining the explainer: a first theoretical analysis of LIME. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*; 2020 Aug 26–28; Online.