



OPEN ACCESS

EDITED BY

Zhang Cong,
Central South University Forestry and
Technology, China

REVIEWED BY

Jiawei Xie,
The University of Newcastle, Australia
Yipeng Xie,
Ocean University of China, China
Jie Li,
East China Jiaotong University, China

*CORRESPONDENCE

Jingjing Yu,
✉ 2150471020@email.szu.edu.cn

RECEIVED 25 August 2025

ACCEPTED 15 September 2025

PUBLISHED 13 October 2025

CITATION

Dang P, Chang L, Tang P, Yu J and Li Z (2025)
Intelligent recognition of surrounding rock
grades based on TBM tunneling parameters.
Front. Earth Sci. 13:1692577.
doi: 10.3389/feart.2025.1692577

COPYRIGHT

© 2025 Dang, Chang, Tang, Yu and Li. This is
an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with
these terms.

Intelligent recognition of surrounding rock grades based on TBM tunneling parameters

Pengliang Dang¹, Le Chang¹, Peishuo Tang¹, Jingjing Yu^{2*} and Zeliang Li²

¹Sinohydro Bureau 14 Co., Ltd., Kunming, Yunnan, China, ²College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, Guangdong, China

Rapid, accurate, and efficient prediction of surrounding rock grades is crucial for ensuring the safety and enhancing the efficiency of tunnel boring machine (TBM) construction. To achieve intelligent perception of surrounding rock grades based on TBM tunneling parameters, this study leverages data from the TBM1 construction phase of the Luotian Reservoir-Tiegang Reservoir Water Diversion Tunnel Project, integrating geological records and tunneling parameters to establish models for different rock grades. First, raw data were cleaned and denoised using box plots, followed by the selection of eight critical parameters—including thrust, torque, penetration rate (PR), rotation speed (RS), et al—through a hybrid approach combining “knowledge-driven” and “data-driven” criteria. The dataset was partitioned into training, testing, and validation sets at a 7:2:1 ratio. Three data processing methods were applied, and machine learning algorithms (XGBoost, Random Forest, CatBoost, and LightGBM) were employed to construct surrounding rock classification models, with Optuna hyperparameter optimization implemented to enhance model performance. The result reveals that the CatBoost model, optimized via SMOTE (Synthetic Minority Oversampling Technique) and hyperparameter tuning, delivered superior performance, achieving 99% validation accuracy with no misclassification across adjacent surrounding rock grades. This research provides actionable insights for advancing intelligent TBM construction practices.

KEYWORDS

TBM tunneling parameters, surrounding rock classification, ensemble learning algorithms, hyperparameter optimization, model performance

1 Introduction

Under the continuous impetus and leadership of China's informatization and digitization, intelligent construction and smart tunneling are inevitably becoming the mainstream development direction in the tunnel engineering field (Fang et al., 2025; Aston et al., 1988; Xie et al., 2025), particularly for shield tunnels (Barton, 2012; Fang et al., 2023; Zheng et al., 2016). Shield tunnels, constructed using tunnel boring machines (TBMs), are equipped with numerous sensors that monitor and collect various mechanical, electrical, and environmental parameters in real time at specific frequencies during construction. This establishes a foundation for data-driven intelligent TBM construction. Furthermore, with the rapid advancement of computer technology, artificial intelligence (AI) has been widely applied across numerous fields such as finance, healthcare,

TABLE 1 The Main characteristic parameters of TBM.

Parameter	Value	Parameter	Value
TBM type	Earth Pressure/TBM Dual Mode Shield	Maximum cutterhead rotation speed $n_{\max}$ (r/min)	6.1
Diameter of cutterhead D (mm)	6,730	Data acquisition frequency (Hz)	1
Maximum thrust $F_{\max}$ (kN)	40860	Number of disc cutter N (pcs)	43
Maximum torque $T_{\max}$ (kN·m)	9,818	Breakout torque T (kN·m)	11781
Maximum advance speed $v_{\max}$ (mm/min)	100(Earth Pressure)/120(TBM)	Drive Power (kW)	2000

transportation, and manufacturing. The emergence of open-source machine learning libraries (e.g., Scikit-learn, PyTorch, TensorFlow) has also provided more accessible channels for researchers in tunneling to learn and apply AI algorithms.

Currently, AI technologies based on big data are increasingly being introduced into tunnel engineering to assist TBM construction. Key research focuses include TBM efficiency optimization, intelligent perception of surrounding rock grades, tunneling performance prediction, and adverse geological condition forecasting (Tang et al., 2024; Chen et al., 2021; Li et al., 2023a; Li et al., 2023b). In the domain of surrounding rock grade classification, existing studies have employed various machine learning algorithms (Afradi and Ebrahimabadi, 2020; Guo et al., 2022a; Guo et al., 2022b; Ghorbani and Yagiz, 2024; Feng et al., 2021; Hou et al., 2022; Kohestani et al., 2017; Liu et al., 2020; Xiong, 2014; Zhu et al., 2021)—such as ensemble methods (Stacking, Random Forest, Adaboost/Adacost, Decision Trees, GBDT), clustering techniques (KNN, SVM), and deep learning models (MLP, DNN)—to develop intelligent surrounding rock grade perception models based on TBM tunneling parameters. These models incorporate parameters such as gripper pressure, gear seal pressure, advance displacement, cutterhead power, shield pressure and rolling force (Zhu et al., 2020; Mao et al., 2021; Zhang et al., 2019; Wu et al., 2021; Liu et al., 2021; Yin et al., 2022; Prechelt, 2002; Chen et al., 2015). However, due to data imbalance, these models exhibit poor prediction accuracy for Grade II and V surrounding rocks. Although the SMOTE oversampling method has been applied to balance datasets, its effectiveness remains limited (Wu et al., 2021; Liu et al., 2021; Yin et al., 2022; Prechelt, 2002; Chen et al., 2015). Current feature selection methods generally fall into two categories: “data-driven” and “knowledge-driven.” Data-driven approaches analyze correlations between input tunneling parameters and target variables using techniques like Random Forest, Pearson correlation analysis, PCA, and XGBoost, ultimately selecting modeling parameters based on correlation strength. Knowledge-driven methods rely on personal expertise and experience to choose input parameters, which introduces subjective biases. While using averaged stable-phase data for modeling can improve accuracy to some extent, it conflicts with real-time prediction requirements, rendering such models inadequate for guiding actual construction. Although AI technologies have advanced in TBM applications, existing models lack universality due to variations in tunneling parameters across different TBM

types (e.g., cutterhead diameter, tool configuration) and geological conditions.

In summary, this study is based on the Luotian Reservoir-Tiegang Reservoir Water Diversion Tunnel Project (hereafter referred to as the Luotie Project), utilizing comprehensive geological data and tunneling parameters from its TBM1. First, feature parameters were selected by integrating data-driven and knowledge-driven criteria according to data characteristics and prediction targets. The raw data were then cleaned using box plots and partitioned into training (70%), testing (20%), and validation (10%) sets. Subsequently, three data imbalance mitigation strategies were applied to construct surrounding rock classification models using XGBoost, RF, CatBoost, and LightGBM algorithms. This research enables rapid, efficient, and intelligent perception of surrounding rock grades, guiding shield drivers to make parameter adjustments, ensuring safer and more efficient tunneling.

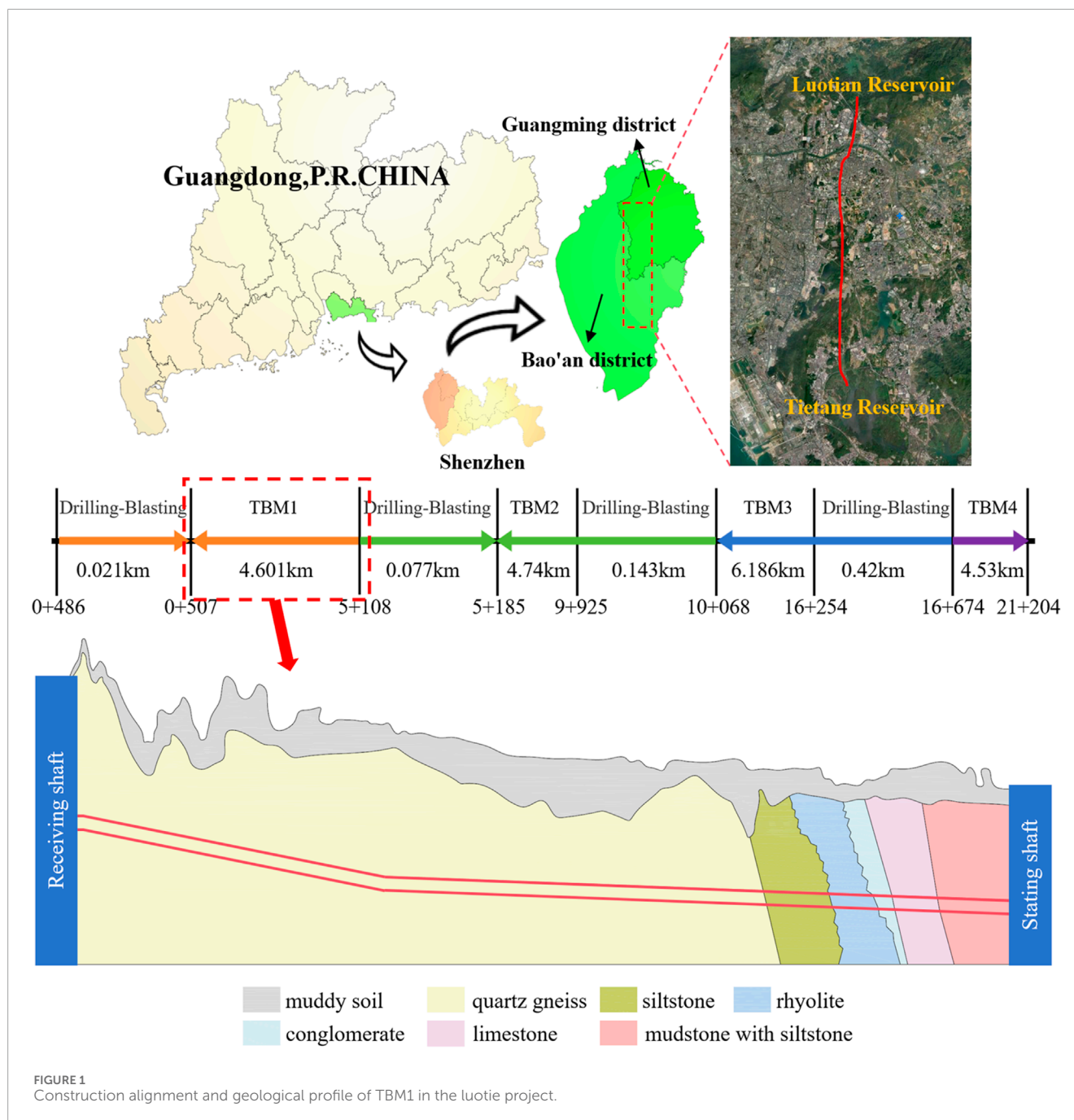
2 Luotie Project

2.1 Description of the project

This study is based on the Luotie Project, a component of the Shenzhen section of the Pearl River Delta Water Resources Allocation Project. The project has a total water delivery capacity of 2.6 million m³/day, with a main tunnel spanning 21.68 km. The target study section is TBM Construction Section 1, starting and ending at chainage K0+486.74 to K5+108.64. This section employs a dual-mode TBM (designated TBM1) with a diameter of 6,730 mm. Key technical specifications are summarized in Table 1.

2.2 Data sources and composition

In the Luotie Project, four TBMs are deployed concurrently, with TBM1 operating in a heterogeneous geological formation predominantly composed of quartz gneiss and sandstone, and characterized by significant groundwater presence. The alignment and longitudinal geological profile of TBM1 are presented in Figure 1. The dataset for this study includes tunneling parameters and geological records derived from the completed excavation segments of TBM1. The tunneling parameters were recorded at

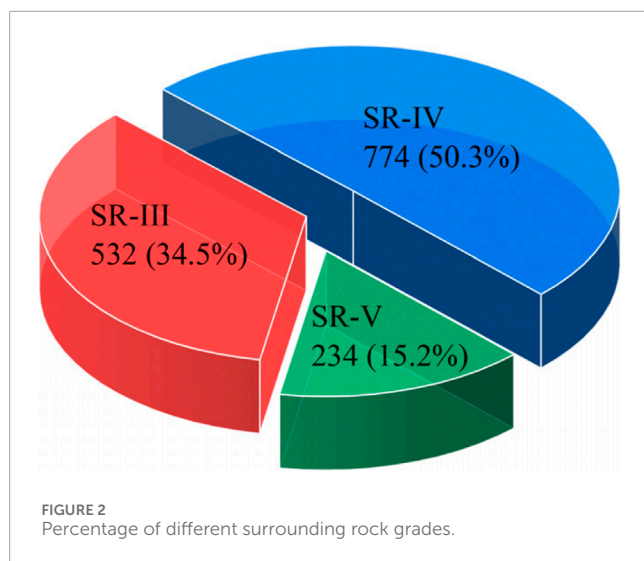


a frequency of 1 Hz, and the distribution of surrounding rock grades along the tunnel is depicted in Figure 2. As shown, the majority of the tunnel encounters SR-IV (Surrounding Rock Grade IV), which constitutes 50.3% (774 segments) of the total excavation. This is followed by SR-III (Surrounding Rock Grade III), representing 34.5% (532 segments), while SR-V (Surrounding Rock Grade V) is the least encountered, comprising only 15.2% (234 segments). This distribution highlights that the tunnel primarily traverses SR-IV, followed by SR-III, with SR-V occurring in a smaller portion of the excavation. Subsequent model development and analysis were based exclusively on the tunneling parameters collected from these segments.

3 Data processing

3.1 Data cleaning

Invalid data were removed from the collected tunneling parameters by retaining only non-zero entries, i.e., any data row containing a zero value in any parameter was deleted according to the criterion defined in Equation 1. After preliminary data cleaning, boxplots were used for secondary outlier removal. A boxplot eliminates abnormal data based on the median, lower quartile (Q1), upper quartile (Q3), maximum value ($Q3+1.5(Q3-Q1)$), and minimum value ($Q1-1.5(Q3-Q1)$). In this study, the stable



excavation phase was specifically extracted to analyze tunneling parameter variations. Notably, even during stable excavation, manual operational decisions introduce fluctuations in two critical parameters—thrust and rotation speed—which subsequently affect other operational indicators. To address this challenge, a refined segmentation method for the stable excavation phase was implemented using a stopping criterion method (Breiman, 2001), as defined in Equation 2. The processed distributions of four key tunneling parameters are shown in Figure 3, where distinct unimodal distributions under different surrounding rock grades demonstrate the effectiveness of the data cleaning methodology.

$$f(F) \cdot f(T) \cdot f(N) \cdot f(v) = 0 \quad (1)$$

$$P_k(t) = 1000 \cdot \left(\frac{\sum_{t_0=t-k+1}^t E(t_0)}{k \cdot \min_{t_0=t-k+1}^t E(t_0)} - 1 \right) \quad (2)$$

3.2 Feature parameter selection and dataset splitting

Pearson correlation analysis was first conducted to assess the linear relationships between all tunneling parameters and surrounding rock classification. The objective of this analysis was to quantify the strength of correlation between each parameter and the surrounding rock grade, facilitating the identification of the most relevant features. Based on the results of this analysis, and employing a hybrid “knowledge-driven” and “data-driven” approach, eight critical feature parameters were selected: Thrust, Rotation Speed (RS), Torque, Penetration Rate (PR), 1# Middle Shield Retracting Gripper Shoe Pressure (1#SSP), 2# Middle Shield Retracting Gripper Shoe Pressure (2#SSP), Foam Pressure (1#FP), and Surrounding Rock Classification (SR). The Pearson correlation coefficients for these parameters are presented in Figure 4. Notably, the absolute correlation coefficients for the other seven features with surrounding rock grade were all greater than 0.15, indicating

a significant linear relationship. In contrast, Thrust exhibited a correlation coefficient of only 0.05, reflecting a weak association with surrounding rock grade. However, given that Thrust is a critical control parameter in TBM tunneling, it was retained in the final set of features. Following data cleaning and feature selection, the dataset was partitioned into training, testing, and validation sets in a 7:2:1 ratio. The partitioning process involved categorizing the data samples by surrounding rock grade (III, IV, V), and then, within each grade group, randomly selecting 70% of the data for the training set, 20% for the testing set, and the remaining 10% for the validation set.

3.3 Model training methods

Based on the selected feature parameters and processed data, modeling is performed using the following three methods:

Method (1): Using raw data with default hyperparameters and no additional processing; Method (2): Applying hyperparameter optimization to the data; Method (3): Combining hyperparameter optimization with SMOTE for data processing. The SMOTE oversampling method was specifically employed to address severe data imbalance issues across Grades III, IV, and V surrounding rocks.

Surrounding rock grade classification models were developed using XGBoost, CatBoost, RF, and LightGBM algorithms based on these three modeling methods. Model performance was evaluated using key metrics—Precision, Recall, and F1_score—to assess the effectiveness of each strategy. The formulas for these metrics are defined in Equations 3–5.

$$PRE = \frac{TP}{TP + FP} \quad (3)$$

$$REC = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = \frac{2 \times PRE \times REC}{PRE + REC} \quad (5)$$

In the equations, TP denotes the number of true positive samples (correctly predicted positive instances), FP represents false positive samples (negative instances incorrectly predicted as positive), and FN indicates false negative samples (positive instances incorrectly predicted as negative).

4 Machine learning algorithms

4.1 eXtreme gradient boosting (XGBoost)

The Extreme Gradient Boosting (XGBoost) algorithm, proposed by Chen and Guestrin et al. (Zhang et al., 2022), is a composite algorithm formed by combining base functions and weights to achieve superior data fitting. It belongs to the category of Gradient Boosting Decision Tree (GBDT). The core principle of GBDT lies in combining multiple weak learners (decision trees) to form a stronger predictive model. During training, GBDT iteratively adds new decision trees to correct errors from previous models until convergence or reaching predefined iteration limits [Prokhorenkova et al. (2017) and

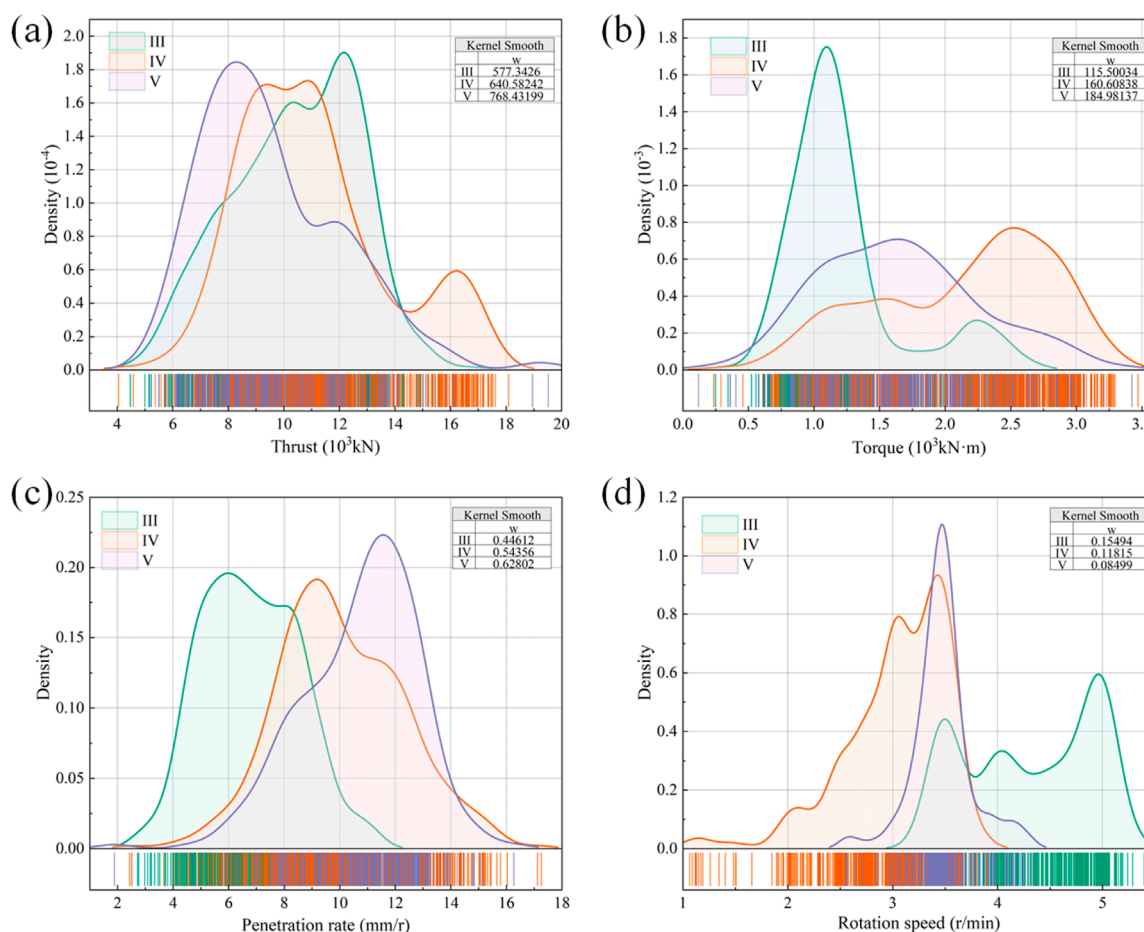


FIGURE 3
Kernel Density Curves of Different Feature Parameters under Corresponding surrounding rock Grades: (a) Thrust; (b) Torque; (c) Penetration rate; (d) Rotation speed.

35]. Unlike traditional GBDT, XGBoost enhances the objective loss function by incorporating regularization terms. To address challenges in calculating derivatives for certain loss functions, XGBoost approximates the loss function using a second-order Taylor expansion, improving computational precision. Additionally, XGBoost employs shrinkage strategies and feature subsampling to prevent overfitting and introduces a sparsity-aware algorithm to handle missing data. It also supports greedy algorithms and approximate learning for node splitting in tree models.

Due to its efficiency in processing large-scale data and complex models, as well as its robustness against overfitting, XGBoost has gained widespread attention and application since its inception.

4.2 Random Forest (RF)

Random Forest (RF), proposed by Breiman in 2001 (Hancock and Khoshgoftaar, 2020), is an ensemble learning method based on decision tree algorithms. The RF model employs a Bagging aggregation strategy, utilizing bootstrapping

(a random sampling method with replacement) for data sampling. These techniques effectively mitigate overfitting risks during model construction. By integrating the results of multiple base estimators, RF achieves superior predictive performance compared to single-estimator models, demonstrating enhanced generalization capability and robustness. Additionally, the model reduces classification errors in imbalanced datasets and exhibits high training efficiency. Consequently, it has been widely applied to both regression and classification problems (Bentéjac et al., 2021).

4.3 Categorical boosting (CatBoost)

The CatBoost algorithm, proposed by Prokhorenkova et al. (2017), Ke et al. (2017), Shen et al. (2025a), is an advanced gradient boosting decision tree (GBDT) algorithm. Building upon GBDT, CatBoost introduces two key enhancements: adaptive learning rates and categorical feature processing, which enable superior performance in both classification and regression tasks. The adaptive learning rate optimizes the contribution of decision trees in each iteration, thereby improving overall model

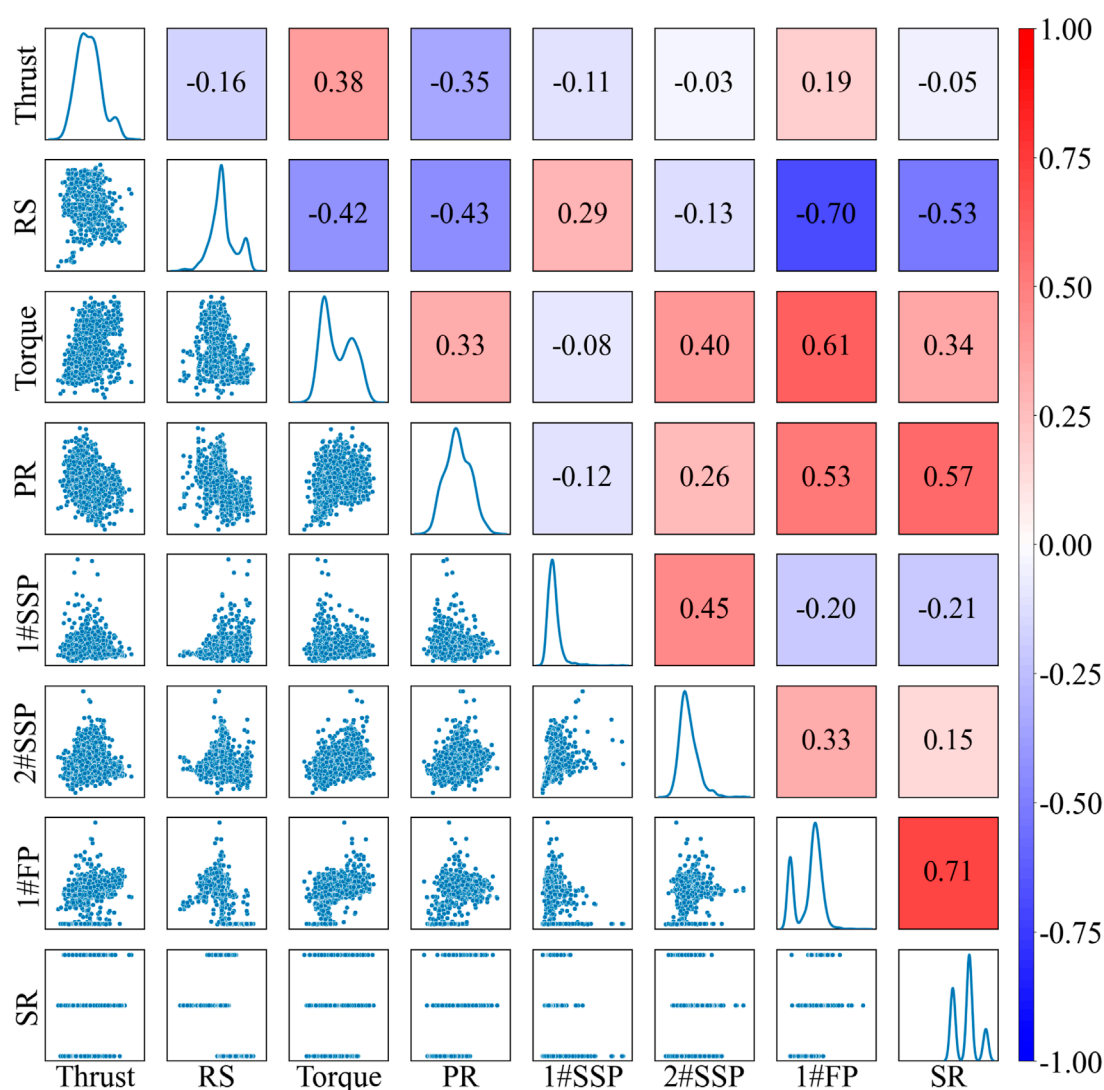


FIGURE 4
Pearson correlation analysis results of tunneling parameters.

accuracy. Its calculation method is detailed in Equations 6, 7. Categorical feature processing employs the Ordered Target Statistics encoding technique to convert categorical features into numerical representations. This method applies hash encoding to categorical values and maps the resulting hash values to numerical equivalents, effectively capturing the influence of categorical features.

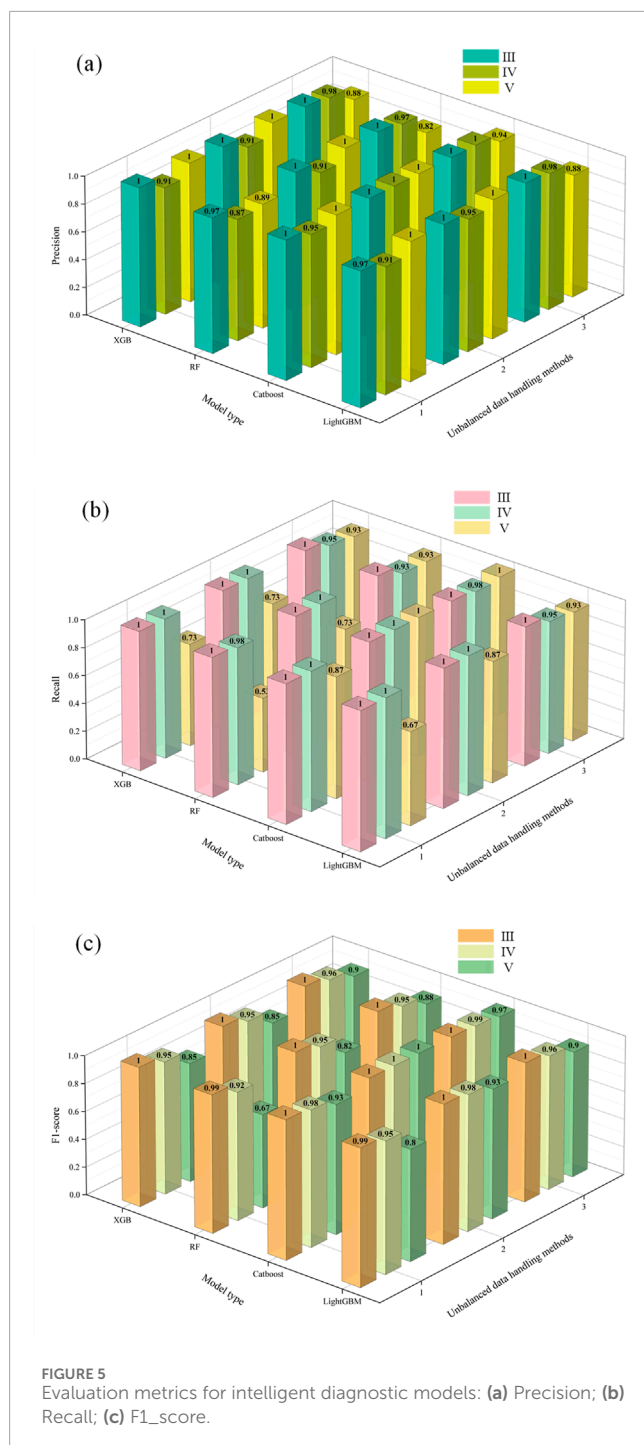
Compared to XGBoost, CatBoost exhibits the following advantages.

1. Higher Model Accuracy: CatBoost often achieves high precision without requiring extensive hyperparameter tuning.
2. Faster Training Speed: Outperforms XGBoost in training efficiency.
3. Superior Prediction Speed: Delivers significantly faster inference times than XGBoost.
4. Lower Memory Consumption: Requires less memory usage on computational hardware.
5. Native Categorical Feature Support: Unlike XGBoost, which relies on OneHot encoding for categorical features, CatBoost directly handles string-type categorical features without preprocessing.

$$\eta_t = \frac{1}{\sqrt{1+t}} \quad (6)$$

$$\alpha_t = \frac{\sum_{i=1}^t \eta_i}{t} \quad (7)$$

In the equations, t denotes the iteration number, η_t represents the learning rate at the t th iteration, and α_t is the average learning rate from the previous iteration.



4.4 Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM), proposed by Ke et al. (2017), Shen et al. (2025b), is an efficient framework for implementing GBDT algorithms. Compared to XGBoost, LightGBM significantly reduces time complexity by converting sample-wise traversal into bin-wise calculations via the histogram algorithm (optimizing from sample-level to bin-level traversal). Additionally, LightGBM employs Gradient-based One-Side

Sampling (GOSS), which retains samples with large gradients while randomly selecting a subset of low-gradient samples, thereby reducing computational load while maintaining gradient distribution stability. Furthermore, its leaf-wise growth strategy prioritizes splitting leaf nodes that yield the greatest loss reduction during the tree-building process, generating deeper asymmetric tree structures to minimize redundant computations. LightGBM also combines optimized feature parallelism and data parallelism methods to accelerate training and introduces the Exclusive Feature Bundling (EFB) algorithm, which merges sparse and mutually exclusive features into single composite features to reduce dimensionality and memory consumption. These innovations enable LightGBM to achieve high efficiency and low resource utilization when processing large-scale, high-dimensional data.

5 Predictive models and results

5.1 Surrounding rock classification model

Surrounding rock classification models were developed using XGBoost, CatBoost, RF, and LightGBM algorithms. First, eight TBM tunneling parameters—Thrust, Rotation Speed (RS), Torque, Penetration Rate (PR), 1#SSP, 2#SSP, 1#FP, and SR—were selected through a hybrid “data-driven” and “knowledge-driven” approach. The data were cleaned according to Equation 1 and denoised using boxplots. The preprocessed dataset was partitioned into training, testing, and validation sets at a 7:2:1 ratio. Three imbalanced data processing methods were applied, with the input parameters being the eight selected TBM tunneling parameters and the output being the surrounding rock grades (III, IV, V), labeled as 3, 4, and 5, respectively.

Before model training, the training set was standardized using Equation 8. The Optuna hyperparameter optimizer was employed for automated hyperparameter tuning, with the optimization cycle set to 100 iterations, yielding final hyperparameters for models based on the three imbalanced data processing methods. Training was then conducted using these optimized hyperparameters. Finally, the trained models were validated using the testing set to obtain the final surrounding rock classification results.

$$x_{new} = \frac{x - \mu}{\sigma} \quad (8)$$

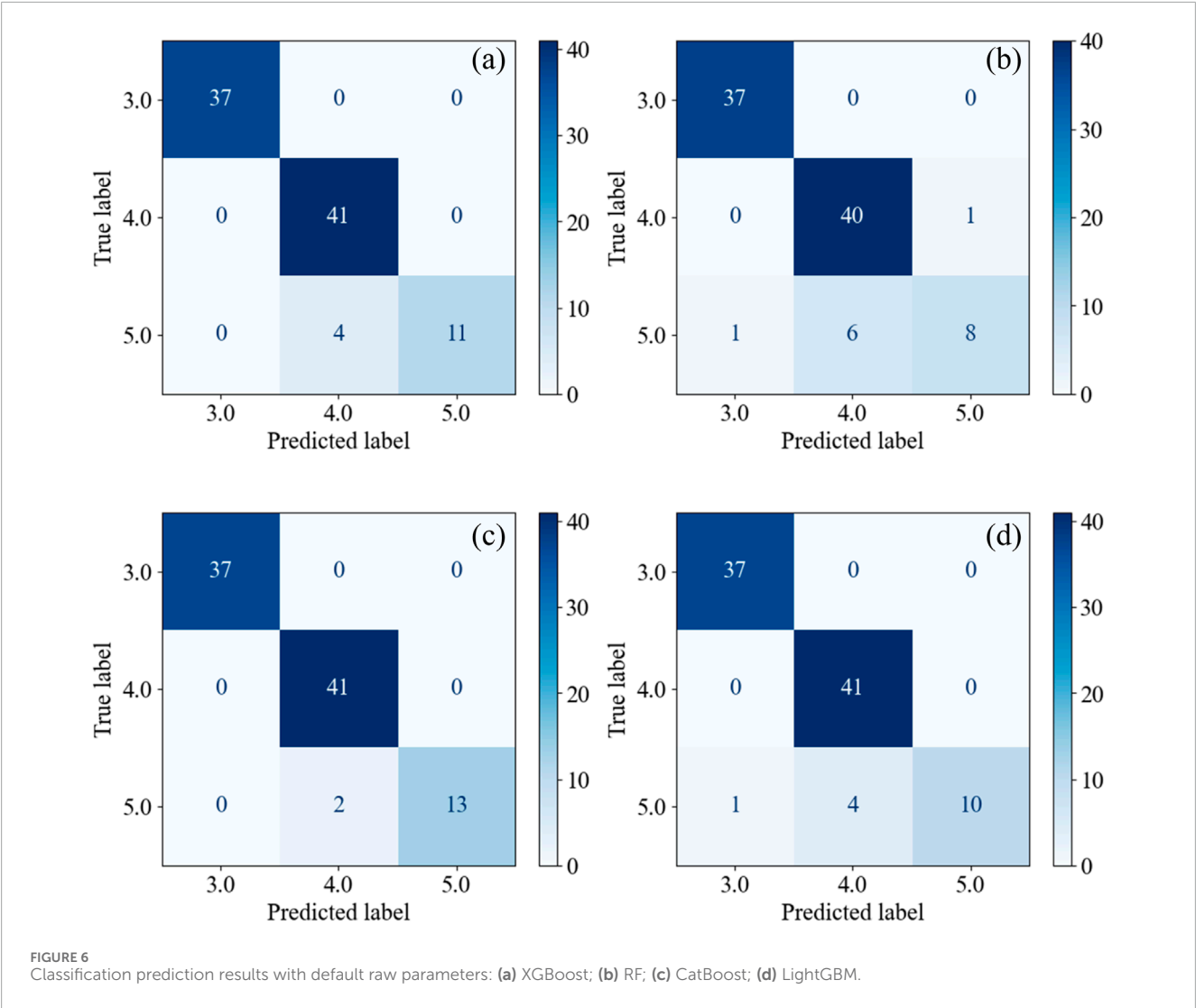
In the equation, x_{new} represents the standardized data, x denotes the raw data, μ is the mean of the sample data, and σ is the standard deviation of the sample data.

5.2 Model performance evaluation

Surrounding rock classification models under three data processing methods were established using XGBoost, CatBoost, RF, and LightGBM ensemble learning algorithms. The prediction results of these models were evaluated using three key metrics: Precision, Recall, and F1_score, as shown in Figure 5. The optimal hyperparameters obtained via Optuna optimization are summarized in Table 2. The results indicate: Method (1)

TABLE 2 The hyperparameter optimization results of the intelligent diagnostic model.

Model type	XGBoost	RF	CatBoost	LightGBM
Unbalanced data handing methods (2)	<i>n_estimators</i> : 1855 <i>max_depth</i> : 9 <i>learning_rate</i> : 0.0468 <i>subsample</i> : 0.8663 <i>colsample_bytree</i> : 0.8882 <i>gamma</i> : 0.5033 <i>min_child_weight</i> : 5 <i>reg_alpha</i> : 0.1713 <i>reg_lambda</i> : 2.6573	<i>n_estimators</i> : 1,559 <i>max_depth</i> : 18 <i>min_samples_split</i> : 3 <i>min_samples_leaf</i> : 3	<i>iterations</i> : 1,142 <i>depth</i> : 4 <i>learning_rate</i> : 5.38e-2 <i>l2_leaf_reg</i> : 0.0188 <i>bagging_temperature</i> : 0.2057 <i>random_strength</i> : 2.3258 <i>border_count</i> : 252	<i>n_estimators</i> : 1,241 <i>max_depth</i> : 0 <i>learning_rate</i> : 0.1681 <i>subsample</i> : 0.7184 <i>colsample_bytree</i> : 0.9625 <i>num_leaves</i> : 55 <i>min_child_samples</i> : 35 <i>reg_alpha</i> : 0.0187 <i>reg_lambda</i> : 1.1492
Unbalanced data handing methods (3)	<i>n_estimators</i> : 455 <i>max_depth</i> : 18 <i>learning_rate</i> : 8.36e-3 <i>subsample</i> : 0.9832 <i>colsample_bytree</i> : 0.9238 <i>gamma</i> : 0.3049 <i>min_child_weight</i> : 9 <i>reg_alpha</i> : 1.003e-4 <i>reg_lambda</i> : 1.181e-4	<i>n_estimators</i> : 1,023 <i>max_depth</i> : 11 <i>min_samples_split</i> : 3 <i>min_samples_leaf</i> : 3	<i>iterations</i> : 542 <i>depth</i> : 6 <i>learning_rate</i> : 2.63e-2 <i>l2_leaf_reg</i> : 0.0610 <i>bagging_temperature</i> : 0.5349 <i>random_strength</i> : 3.6250 <i>border_count</i> : 35	<i>n_estimators</i> : 1,511 <i>max_depth</i> : 14 <i>learning_rate</i> : 9.30e-3 <i>subsample</i> : 0.9086 <i>colsample_bytree</i> : 0.8780 <i>num_leaves</i> : 130 <i>min_child_samples</i> : 4 <i>reg_alpha</i> : 0.0024 <i>reg_lambda</i> : 5.2011



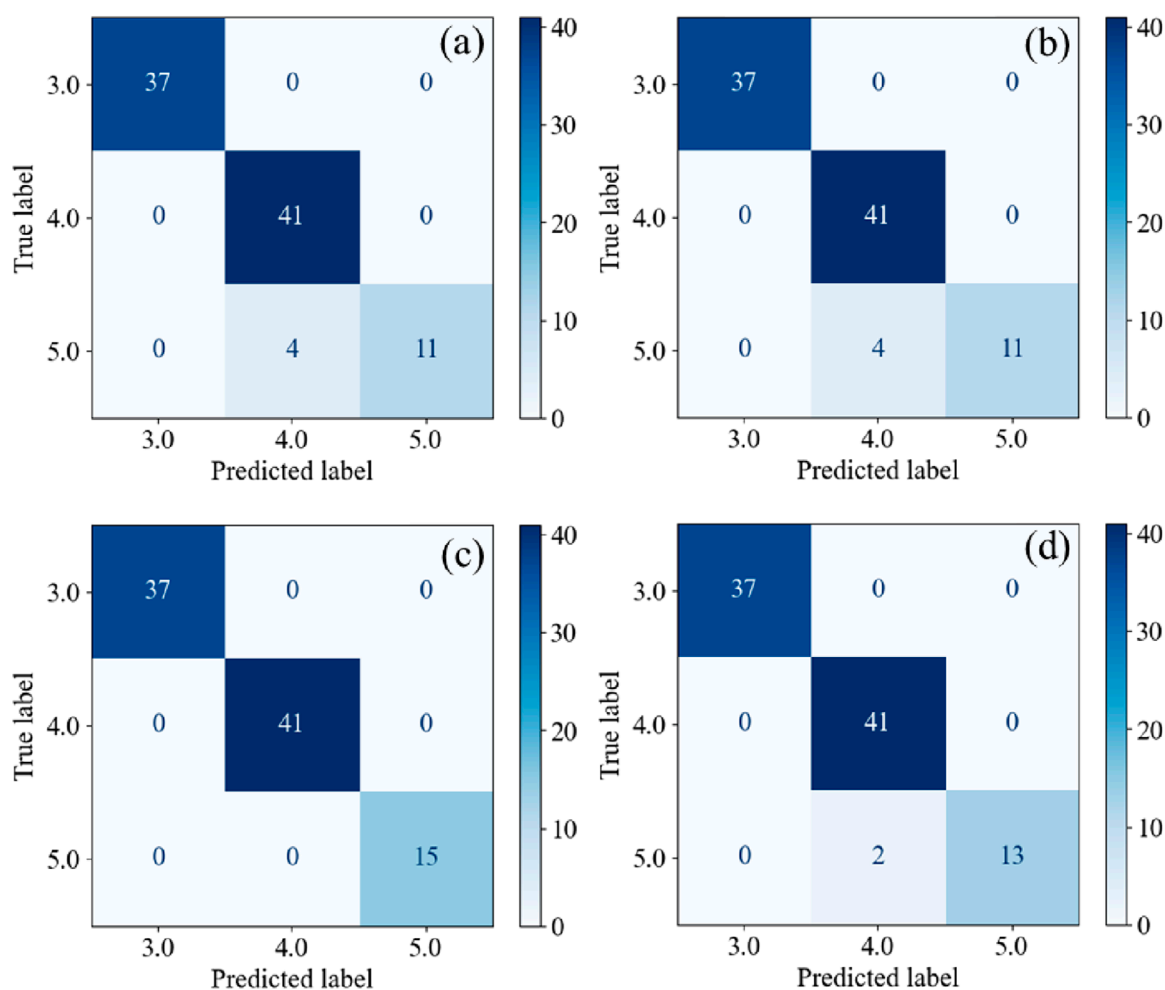


FIGURE 7
Classification prediction results with hyperparameter optimization: (a) XGBoost; (b) RF; (c) CatBoost; (d) LightGBM.

(raw data only): All four models exhibited poor performance. The RF model performed worst for Grade V surrounding rock (Precision:0.67, Recall:0.53, F1:0.89). Method (2) (hyperparameter optimization): Significant accuracy improvements were achieved compared to Method (1). The CatBoost model satisfied all evaluation criteria, indicating reliable predictions, while RF and XGBoost models still underperformed. Method (3) (hyperparameter optimization + SMOTE): All models showed further accuracy gains, with the CatBoost model achieving the best prediction performance.

Performance analysis revealed varying degrees of classification confusion among all models using raw data, as demonstrated by the confusion matrix of the validation set in Figure 6. The RF model exhibited a misclassification rate of 1.92% (1/52) between Grades III and V, with cross-grade misclassifications observed. The misclassification rate between Grades IV and V reached 10.71% (6/56). The XGBoost and LightGBM models showed improved performance, reducing the IV-V misclassification rates to 7.14% (4/56) and 7.27% (4/55), respectively. The CatBoost model outperformed others, achieving

a IV-V misclassification rate of 3.57% (2/56), with prediction accuracies of 100% for Grade III, 100% for Grade IV, and 86% for Grade V. These results confirm that the CatBoost model delivers the best predictive performance under raw data conditions.

After hyperparameter optimization, the confusion matrices of different models on the validation set are shown in Figure 7. The results indicate: RF and XGBoost models exhibited no misclassification between Grades III and V, while the misclassification rate between Grades IV and V decreased to 7.14% (4/56). LightGBM model further reduced the IV-V misclassification rate to 3.57% (2/56). CatBoost model achieved zero misclassification with 100% prediction accuracy for Grades III, IV, and V, demonstrating its superior effectiveness.

When combining SMOTE with hyperparameter optimization, the classification prediction results of the validation set confusion matrix are shown in Figure 8. The results demonstrate: The XGBoost model achieved further improvement, reducing the misclassification rate between Grades IV and V to 5.36% (3/56). The RF and LightGBM models increased the prediction accuracy for Grade V to

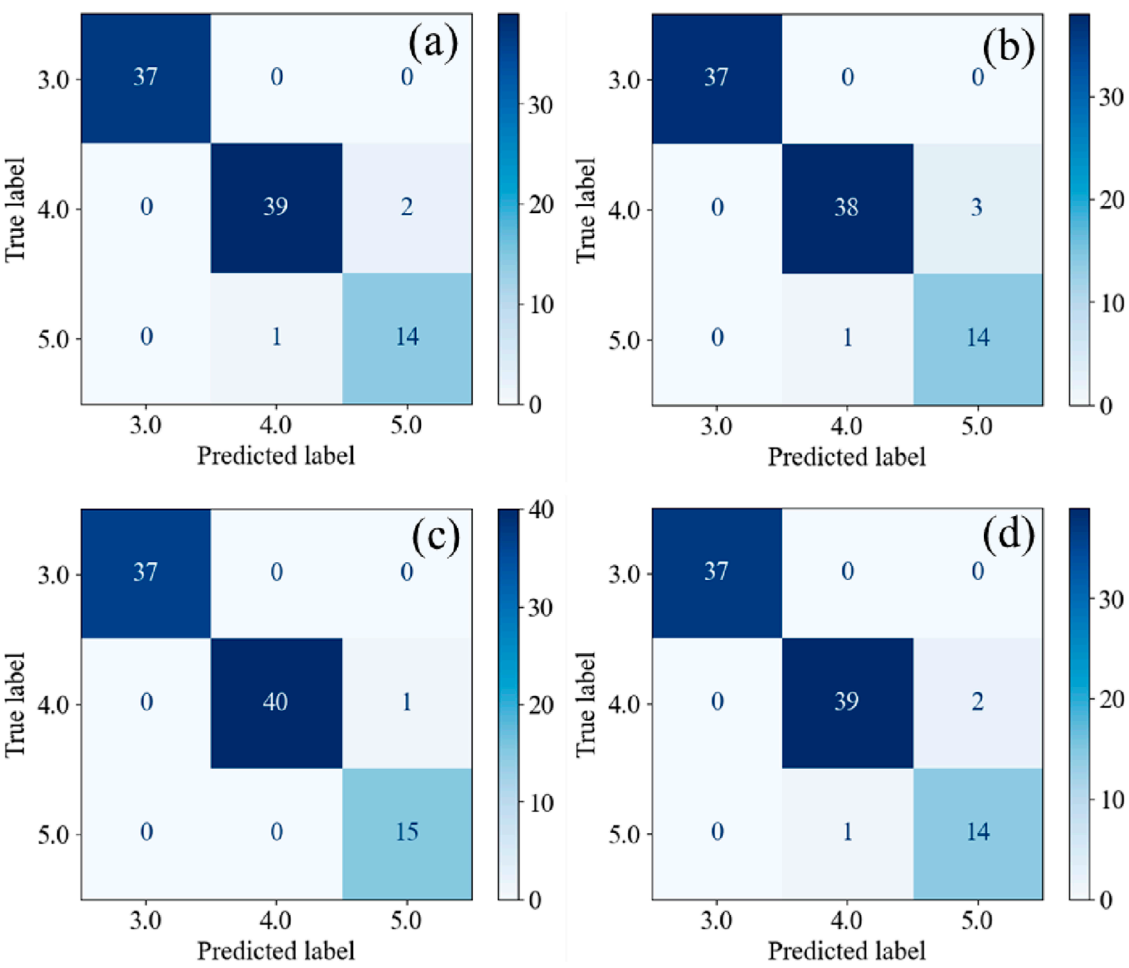


FIGURE 8
Classification prediction results with SMOTE and hyperparameter optimization: (a) XGBoost; (b) RF; (c) CatBoost; (d) LightGBM.

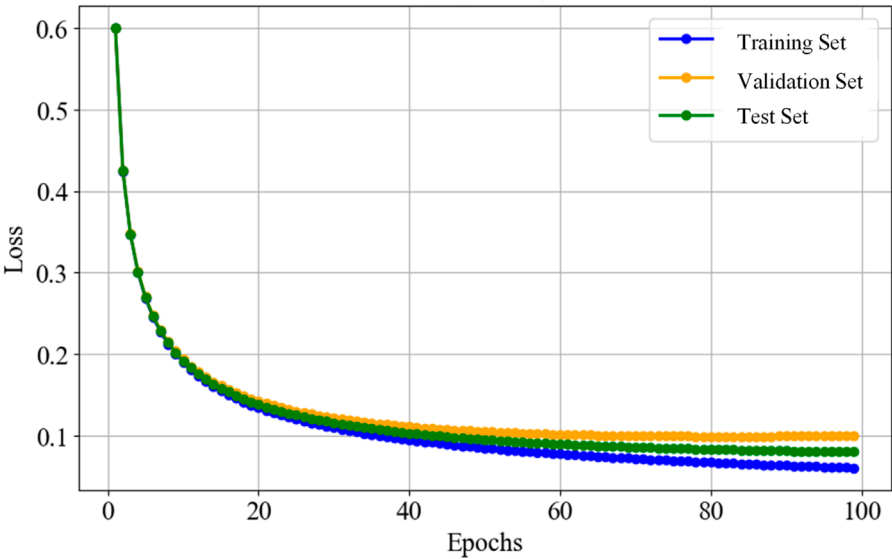


FIGURE 9
Convergence curve of CatBoost model.

93.3% (14/15) with no cross-grade misclassifications. These results indicate that SMOTE-enhanced hyperparameter optimization further improves model accuracy. The CatBoost model, which already achieved excellent performance with hyperparameter optimization alone, showed negligible differences after SMOTE integration.

Therefore, through comprehensive consideration of prediction accuracy, the CatBoost model integrated with SMOTE and hyperparameter optimization was selected as the optimal intelligent diagnostic model in this study. Without further validating the performance curves of the CatBoost model, its convergence curve was plotted as shown in Figure 9. The results indicate that the model exhibits only slight overfitting, and its overall performance remains acceptable.

6 Conclusion

TBM tunneling efficiency and construction safety are highly dependent on real-time perception of surrounding rock conditions and parameter optimization. This study, based on TBM construction data from the Luotie Project, established intelligent diagnostic models for surrounding rock classification using XGBoost, RF, CatBoost, and LightGBM algorithms. These models integrate feature selection methods combining data-driven and knowledge-driven approaches, along with three data processing techniques. Model performance was evaluated using three key metrics: Precision, Recall, and F1_score. The conclusions are as follows.

1. Feature-selected tunneling parameters from TBM1 in the Luotie Project were processed using three methods to develop surrounding rock classification models. The CatBoost model with SMOTE-enhanced hyperparameter optimization achieved the highest prediction accuracy (average 99%).
2. The RF model using raw data exhibited a 1.92% (1/52) misclassification rate between Grades III and V, involving cross-grade errors. Other models (XGBoost, CatBoost, LightGBM) and the hyperparameter-optimized RF model with SMOTE eliminated cross-grade misclassifications, demonstrating perfect differentiation between Grades III and V. This confirms that SMOTE-enhanced hyperparameter optimization significantly improves prediction accuracy.
3. The CatBoost model achieved optimal prediction accuracy with hyperparameter optimization alone, showing negligible performance differences when combined with SMOTE. Both approaches satisfied all key evaluation criteria.

References

Afradi, A., and Ebrahimabadi, A. (2020). Comparison of artificial neural networks (ANN), support vector machine (SVM) and gene expression programming (GEP) approaches for predicting TBM penetration rate. *SN Appl. Sci.* 2 (12), 2004. doi:10.1007/s42452-020-03767-y

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

PD: Writing – review and editing. LC: Writing – review and editing. PT: Writing – review and editing. JY: Writing – original draft, Writing – review and editing. ZL: Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

Authors PD, LC, and PT were employed by Sinohydro Bureau 14 Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Aston, T. R. C., Gilby, J. L., and Yuen, C. M. K. (1988). A comparison of rock mass disturbance in TBM and drill and blast drivages at the donkin Mine, Nova Scotia. *Int. J. Min. Geol. Eng.* 6 (2), 147–162. doi:10.1007/bf00880804

- Barton, N. (2012). Reducing risk in long deep tunnels by using TBM and drill-and-blast methods in the same project—the hybrid solution. *J. Rock Mech. Geotechnical Eng.* 4 (2), 115–126. doi:10.3724/sp.j.1235.2012.00115
- Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* 54 (3), 1937–1967. doi:10.1007/s10462-020-09896-5
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324
- Chen, T. Q., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R. package version 0.4-2* 1 (4), 1–4.
- Chen, Z. Y., Zhang, Y. P., Li, J. B., Li, X., and Jing, L. (2021). Diagnosing tunnel collapse sections based on TBM tunneling big data and deep learning: a case study on the Yinsong Project, China. *Tunn. Undergr. Space Technol.* 108, 103700. doi:10.1016/j.tust.2020.103700
- Fang, X. H., Yang, J. S., Zhang, X. M., Zhang, C., Wang, S., and Xie, Y. (2023). Numerical modeling of open TBM tunneling in stratified rock masses using a coupled FDM-DEM method. *Comput. Geotechnics* 156, 105251. doi:10.1016/j.compgeo.2023.105251
- Fang, X. H., Yang, L. Q., Zhang, C., Yang, J., and Zhou, W. (2025). Large diameter shafts in fractured strata: equipment and construction procedures. *J. Constr. Eng. Manag.* 151 (9), 06025001. doi:10.1061/jcemd4.coeng-16745
- Feng, S. X., Chen, Z. Y., Luo, H., Wang, S., Zhao, Y., Liu, L., et al. (2021). Tunnel boring machines (TBM) performance prediction: a case study using big data and deep learning. *Tunn. Undergr. Space Technol.* 110, 103636. doi:10.1016/j.tust.2020.103636
- Ghorbani, E., and Yagiz, S. (2024). Predicting disc cutter wear using two optimized machine learning techniques. *Archives Civ. Mech. Eng.* 24 (2), 106. doi:10.1007/s43452-024-00911-y
- Guo, D., Li, J. H., Jiang, S. H., Li, X., and Chen, Z. (2022a). Intelligent assistant driving method for tunnel boring machine based on big data. *Acta Geotech.* 17 (4), 1019–1030. doi:10.1007/s11440-021-01327-1
- Guo, D., Li, J. H., Li, X., Li, Z., Li, P., and Chen, Z. (2022b). Advance prediction of collapse for TBM tunneling using deep learning method. *Eng. Geol.* 299, 106556. doi:10.1016/j.enggeo.2022.106556
- Hancock, J. T., and Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *J. big data* 7 (1), 94. doi:10.1186/s40537-020-00369-8
- Hou, S. K., Liu, Y. R., and Yang, Q. (2022). Real-time prediction of rock mass classification based on TBM operation big data and stacking technique of ensemble learning. *J. Rock Mech. Geotechnical Eng.* 14 (1), 123–143. doi:10.1016/j.jrmge.2021.05.004
- Ke, G., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., et al. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Adv. neural Inf. Process. Syst.*, 30. doi:10.5555/3294996.3295074
- Kohestani, V. R., Bazarganlari, M. R., and Asgari Marnani, J. (2017). Prediction of maximum surface settlement caused by earth pressure balance shield tunneling using random forest. *J. AI Data Min.* 5 (1), 127–135. doi:10.22044/jadm.2016.748
- Li, J. B., Chen, Z. Y., Li, X., Jing, L. J., Zhang, Y. P., Xiao, H. H., et al. (2023a). Feedback on a shared big dataset for intelligent TBM part I: feature extraction and machine learning methods. *Undergr. Space* 11, 1–25. doi:10.1016/j.undsp.2023.01.001
- Li, J. B., Chen, Z. Y., Li, X., Jing, L. J., Zhang, Y. P., Xiao, H. H., et al. (2023b). Feedback on a shared big dataset for intelligent TBM Part II: application and forward look. *Undergr. Space* 11, 26–45. doi:10.1016/j.undsp.2023.01.002
- Liu, B., Wang, R., Zhao, G., Guo, X., Wang, Y., Li, J., et al. (2020). Prediction of rock mass parameters in the TBM tunnel based on BP neural network integrated simulated annealing algorithm. *Tunn. Undergr. Space Technol.* 95, 103103. doi:10.1016/j.tust.2019.103103
- Liu, Z. B., Li, L., Fang, X. L., Qi, W., Shen, J., Zhou, H., et al. (2021). Hard-rock tunnel lithology prediction with TBM construction big data using a global-attention-mechanism-based LSTM network. *Automation Constr.* 125, 103647. doi:10.1016/j.autcon.2021.103647
- Mao, Y. Z., Gong, G., Zhou, X., Hu, X. Q., and Chen, Z. Y. (2021). Identification of TBM surrounding rock based on Markov process and deep neural network. *J. Zhejiang Univ. Eng. Sci.* 55 (3), 448–454. doi:10.3785/j.issn.1008-973X.2021.03.004
- Prechelt, L. (2002). *Early stopping-but when?* [M]//Neural networks: tricks of the trade. Berlin, Heidelberg: Springer Berlin Heidelberg, 55–69.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Adv. neural Inf. Process. Syst.*, 31. doi:10.5555/3327757.3327770
- Shen, J., Bao, X. H., Chen, X. S., Wu, X. L., Qiu, T., and Cui, H. Z. (2025a). Seismic resilience assessment method for tunnels based on cloud model considering multiple damage evaluation indices. *Tunn. Undergr. Space Technol.* 157, 106360. doi:10.1016/j.tust.2024.106360
- Shen, J., Bao, X. H., Li, J. H., Chen, X. S., and Cui, H. Z. (2025b). Study on the mechanism of EPWP dissipation at the joints of shield tunnel in liquefiable strata during seismic events. *Soil Dyn. Earthq. Eng.* 188, 109089. doi:10.1016/j.soildyn.2024.109089
- Tang, Y., Yang, J. S., You, Y. Y., Fu, J., Zheng, X., and Zhang, C. (2024). Multi-output prediction for TBM operation parameters based on stacking ensemble algorithm. *Tunn. Undergr. Space Technol.* 152, 105960. doi:10.1016/j.tust.2024.105960
- Wu, Z. J., Wei, R. L., Chu, Z. F., and Liu, Q. (2021). Real-time rock mass condition prediction with TBM tunneling big data using a novel rock-machine mutual feedback perception method. *J. Rock Mech. Geotechnical Eng.* 13 (6), 1311–1325. doi:10.1016/j.jrmge.2021.07.012
- Xie, Y. P., Qu, T. M., Yang, J. S., Wang, S., and Zhao, J. (2025). Multiscale insights into tunneling-induced ground responses in coarse-grained soils. *Comput. Geotechnics* 185. doi:10.1016/j.compgeo.2025.107319
- Xiong, X. (2014). Research on prediction of rating of rockburst based on BP neural network. 8, 463, 469. doi:10.2174/1874149501408010463
- Yin, X., Liu, Q. S., Huang, X., and Pan, Y. (2022). Perception model of surrounding rock geological conditions based on TBM operational big data and combined unsupervised-supervised learning. *Tunn. Undergr. Space Technol.* 120, 104285. doi:10.1016/j.tust.2021.104285
- Zhang, N., Li, J. B., Jing, L. J., Liu, C., and Gao, J. (2019). Prediction method of Rockmass parameters based on tunnelling process of tunnel boring machine. *J. Zhejiang Univ. Eng. Sci.* 53 (10), 1977–1985. doi:10.3785/j.issn.1008-973X.2019.10.015
- Zhang, W. G., Zhang, R. H., Wu, C. Z., Goh, A. T., and Wang, L. (2022). Assessment of basal heave stability for braced excavations in anisotropic clay using extreme gradient boosting and random forest regression. *Undergr. Space* 7 (2), 233–241. doi:10.1016/j.undsp.2020.03.001
- Zheng, Y. L., Zhang, Q. B., Zhao, J., Du, J. H., Yin, J. B., and Li, H. Q. (2016). Challenges and opportunities of using tunnel boring machines in mining. *Tunn. Undergr. Space Technol.* 57, 287–299. doi:10.1016/j.tust.2016.01.023
- Zhu, M. Q., Zhu, H. H., Wang, X., and Cheng, P. P. (2020). Study on CART-based ensemble learning algorithms for predicting TBM tunneling parameters and classing surrounding rockmasses. *Chin. J. Rock Mech. Eng.* 39 (9), 1860–1871. doi:10.13722/j.cnki.jrme.2020.0555
- Zhu, M. Q., Gutierrez, M., Zhu, H. H., Ju, J. W., and Sarna, S. (2021). Performance Evaluation Indicator (PEI): a new paradigm to evaluate the competence of machine learning classifiers in predicting rockmass conditions. *Adv. Eng. Inf.* 47, 101232. doi:10.1016/j.aei.2020.101232