



## Article

# A Novel Identification Approach Using RFECV–Optuna–XGBoost for Assessing Surrounding Rock Grade of Tunnel Boring Machine Based on Tunneling Parameters

Kebin Shi, Renyi Shi , Tao Fu, Zhipeng Lu and Jianming Zhang 

College of Water Resources and Civil Engineering, Xinjiang Agricultural University, Urumqi 830000, China; xndsg@sina.com (K.S.); futaoscedu@outlook.com (T.F.); xndlzp@sina.com (Z.L.); zjmchnxj@163.com (J.Z.)

\* Correspondence: xjshiry@163.com; Tel.: +86-17-799-638-660

**Abstract:** In order to solve the problem of the poor adaptability of the TBM digging process to changes in geological conditions, a new TBM digging model is proposed. An ensemble learning prediction model based on XGBoost, combined with Optuna for hyperparameter optimization, enables the real-time identification of surrounding rock grades. Firstly, an original dataset was established based on the TBM tunneling parameters under different surrounding rock grades based on the KS tunnel. Subsequently, the RF–RFECV was employed for feature selection and six features were selected as the optimal feature subset according to the importance measure of random forest features and used to construct the XGBoost identification model. Furthermore, the Optuna framework was utilized to optimize the hyperparameters of XGBoost and validated by applying the established TBM dataset of the KS Tunnel. In order to verify the applicability and efficiency of the proposed model in surrounding rock grade identification, the prediction results of five commonly used machine learning models, Optuna–XGBoost, Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Decision Tree (DT), XGBoost, and PSO–XGBoost, were compared and analyzed. The main conclusions are as follows: the feature selection method based on RF–RFECV improved the accuracy by 8.26%. Among the optimal feature subset, T was the most essential feature for the model’s input, while PR was the least important. The Optuna–XGBoost model proposed in this paper had higher accuracy (0.9833), precision (0.9803), recall (0.9813), and F1 score (0.9807) than other models and could be used as an effective means for the lithological identification of surrounding rock grade.

**Keywords:** TBM; tunnel; surrounding rock; tunneling parameters; RFECV; XGBoost; Optuna; machine learning; ensemble learning



**Citation:** Shi, K.; Shi, R.; Fu, T.; Lu, Z.; Zhang, J. A Novel Identification Approach Using RFECV–Optuna–XGBoost for Assessing Surrounding Rock Grade of Tunnel Boring Machine Based on Tunneling Parameters. *Appl. Sci.* **2024**, *14*, 2347. <https://doi.org/10.3390/app14062347>

Academic Editor: Nikolaos Koukouras

Received: 3 February 2024

Revised: 26 February 2024

Accepted: 5 March 2024

Published: 11 March 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, the development and utilization of underground space have attracted increasing attention, involving many ultra-long tunnel projects in various major water diversion and transfer engineering schemes [1]. This has significantly promoted tunnel boring machine (TBM) application and development. Consequently, TBM construction has become the preferred method for super-long tunnel construction [2]. Super-long tunnel construction often faces challenges such as excessive length, deep buried depths, and complex geological conditions. Therefore, TBM’s efficient excavation and construction safety have always been of great concern to related personnel. Compared with the traditional drilling and blasting method, TBM’s continuous mechanized operation offers various advantages, including safety, speed, and high-quality outcomes. It is also environmentally friendly, reduces labor intensity, and enhances construction safety. The TBM construction speed can reach three to ten times that of the drill and blast method [3,4]. However, there is often a significant difference between the actual and designed surrounding rock-grade conditions in practical construction. TBM tunneling shows poor adaptability when geological conditions change.

Traditional methods of advanced geological forecasting often require additional tunneling time and equipment costs. In most cases, the relationship between the rock–machine interaction relies on the operator’s experience and judgment, often failing to adjust the excavation parameters promptly according to changes in the surrounding rock grades of TBM tunneling. This leads to an inability of the TBM to utilize its efficiency advantages during tunneling. It would significantly reduce construction risks if it were possible to continuously monitor the changes in the surrounding rock conditions at the cutter head in real-time during TBM tunneling, thus selecting better excavation parameters.

Tunnel surrounding rock classification is fundamental to understanding geological problems in tunnel engineering and remains a significant research area in underground construction. The common tunnel surrounding rock classification techniques are based on the stability of the surrounding rock. Traditional classification methods used domestically and internationally, such as the RQD [5], RMR system [6], Q system [7,8], and BQ grading method [9,10], are often suitable for traditional drilling and blasting construction. To meet engineering needs, numerous scholars have proposed various rock mass classification methods suitable for TBM construction. For instance, N. Barton [11], based on the Q classification system and data from 145 tunnel constructions and geological information, proposed the QTBM model from the perspective of the rock–machine interaction. This model is used to predict the net tunneling rate and construction speed. Refining and modifying existing rock mass grades for TBM excitability classification has been widely applied in the TBM rock mass classification field [12–15]. In addition, Z.T. Bieniawski [16] proposed the RME scoring system based on the difficulty of rock mass excavation. Based on the proposed system, the comprehensive evaluation of whether the rock mass is easy to excavate, TBM selection cost estimation, and tunneling performance, predictions are completed during the actual construction process. Xue Y [17] integrated TBM excavation and surrounding rock adaptability classifications. They took the TBM construction speed as the classification index to obtain a comprehensive classification method for surrounding rock in TBM construction.

However, existing rock mass classification methods have shortcomings, such as having numerous influencing factors, complex grading systems, and a singular focus on excavation performance evaluation indicators. On the other hand, owing to the complex geological environment in TBM construction and the high cost of sampling and testing, it requires effort to obtain rock mass parameters accurately. Consequently, this hinders the ability to perform real-time and accurate judgments of the surrounding rock grades based on rock mass parameters.

A substantial amount of research has confirmed that there is a significant correlation between rock mass parameters and TBM excavation performance. J. Rostami [18] proposed the CSM model, which uses UCS and BTS to predict the net excavation rate of TBMs. A. Bruland [19] proposed the NTNU model, which achieves TBM performance prediction through the regression analysis of various rock mass parameters and TBM excavation parameters. F. Xiong [20] established a relationship model between surrounding rock parameters (UCS, DPW, and  $\alpha$ ) and FPI based on PSO-SVR and, based on the correspondence between the FPI value and the excavatability of the surrounding rock, established a grading method for the surrounding rock applicable to TBM construction.

Many empirical models and artificial intelligence methods demonstrate a strong correlation between surrounding rock and TBM tunneling parameters. These methods suggest that TBM tunneling parameters can be used in inverse analysis to achieve real-time perception and the identification of surrounding rock information during tunneling.

For example, Q. Zhang [21] established a new surrounding rock classification system by clustering TBM tunneling parameters and then effectively identifying the TBM tunnel surrounding rock grade based on SVM. Using the AdaBoost algorithm, Q. Liu [22] proposed a method for predicting the HC classification of surrounding rock based on the TBM excavation parameters. H. Li [23] used FPI and TPI as input parameters and improved the surrounding rock grade identifying model through the SOM-SVM algorithm. M. Xi [24],

analyzed TBM tunneling parameters and loaded distribution patterns under different surrounding rock conditions, using the RF algorithm with inputs such as N, T, PR, and F and with the surrounding rock grade as the output to establish a surrounding rock grade recognition model. Z. Wu [25] first analyzed the correlation between rock mass parameters and various TBM excitability assessment indices. Then a capability classification of surrounding rock based on TBM tunneling performance using the TOPSIS method effectively perceived and identified the proposed rock mass excitability grades used for four TBM tunneling parameters: F, PR, N, and p.

The above studies have indicated that machine learning techniques exhibit promising applications in the recognition of TBM surrounding rock grade recognition. However, there are still some areas for improvement in the current research. Most studies primarily focus on the comparison of different machine learning algorithms, while overlooking the influence of variations in hyperparameters on the prediction and recognition outcomes of these algorithms. Although optimization algorithms have been employed in a few studies to identify the optimal parameter combinations for specific machine learning algorithms, thereby enhancing model optimization, it is important to note that the applicability of a fixed model in rock mass information perception cannot be guaranteed. Therefore, the optimal algorithm and its corresponding parameter combination is crucial for the accurate identification of the TBM surrounding rock grades.

In response to these challenges, the study is based on a real-time dataset of TBM tunneling parameters obtained from a water supply tunnel in Xinjiang (referred to as the KS Tunnel). For this study, we proposed a novel model for identifying surrounding rock grades based on RFECV–Optuna–XGBoost. The method combined recursive feature elimination (RFE) with cross-validation (CV) to select a superior feature subset. It uses a random forest (RF) as the base model to evaluate feature importance and further reduce dimensionality to obtain the optimal feature subset, thereby enhancing the accuracy of the TBM surrounding rock grade recognition model.

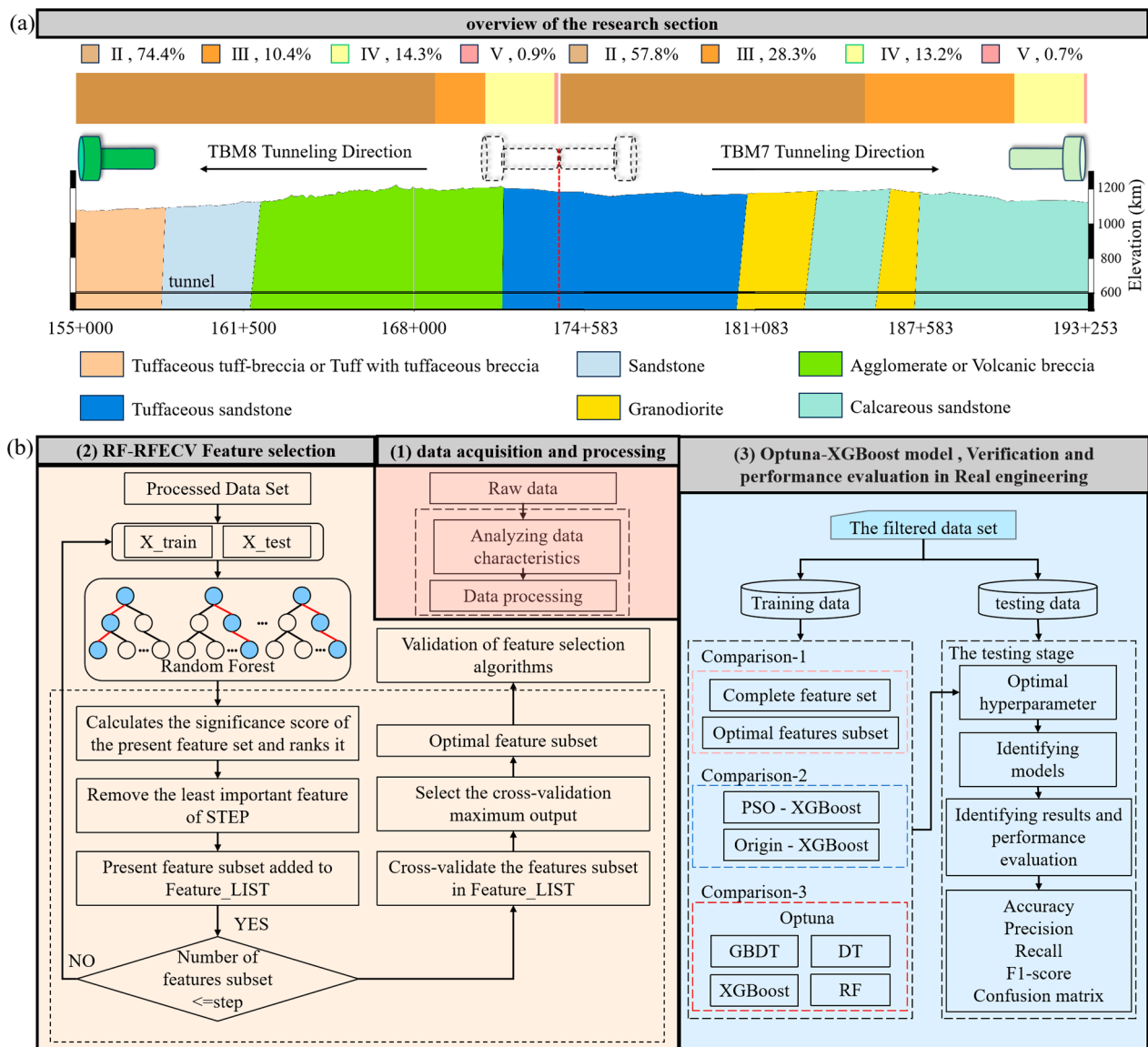
This study employs the Optuna algorithm for hyperparameter optimization in XGBoost to construct an optimal classification model. This model uses the best feature subset as input variables and the surrounding rock grade as the output variable. The tests and the practical application in the KS Tunnel project demonstrate that the proposed model has good predictive accuracy and generalizability and that it offers a more applicable method for the recognition of the surrounding rock grades of TBM excavation in actual engineering projects.

## 2. Research Frameworks and Methodologies

### 2.1. Research Frameworks

The research framework for the surrounding rock grade recognition of the TBM tunneling method proposed in this study is shown in Figure 1. The framework consists of four main parts: (1) data acquisition and processing; (2) feature selection based on RF–RFECV; (3) the establishment of the identification model based on Optuna–XGBoost, with practical engineering validation and performance evaluation; and (4) the measurement of the model features.

(1) By collecting data on the TBM excavation parameters under different surrounding rock grades in the KS Tunnel, an initial dataset for the prediction model was formed after selection and integration. In this dataset, the TBM tunneling parameters were utilized as input variables, and the surrounding rock grade was employed as the output variable. It should be noted that, due to the minimal occurrence of Grade V surrounding rocks (1.6% total) in this project, only Grades II to IV were considered during the identification process. The dataset was divided into a training set and a test set in a 7:3 ratio. The training set was used for model learning, and the test set was used to verify the model's accuracy.



**Figure 1.** Research framework: (a) overview of the research section and (b) four main parts of framework.

(2) To ensure data quality, preprocessing operations such as cleaning, outlier handling, and missing value treatment were conducted on the processed data. Using box plots from mathematical statistics, eight input features were identified as the initial feature set. These features included the thrust (F), rotation speed (N), torque (T), net penetration rate (PR), penetration depth (p), field penetration index (FPI), torque penetration index (TPI), and specific excavation energy (SE). To address the common issues of high complexity and redundancy in the production data, we employed a method combining RFE with CV, using RF as the base model to complete the feature selection; thus, we avoided manual intervention automatically. This method utilized RF and CV to obtain the training score for the current dataset and calculated the importance of each feature. We then eliminated the least essential feature and repeated the training and elimination steps until the dataset was empty. Finally, the dataset with the highest training score was selected as the feature subset after feature selection. This feature selection method ensures data quality while reducing redundant features and shortening the model training time.

(3) The establishment of an Optuna-XGBoost identification model with practical engineering validation and comparison. By automatically optimizing hyperparameters using Optuna, the Optuna-XGBoost model was constructed and applied to the KS Tunnel

to identify the surrounding rock grades of the TBM construction section in the tunnel. The effectiveness of this model was verified based on evaluation metrics, and its superiority was highlighted through performance comparisons with other models.

(4) Model feature measurement. During the training process of the XGBoost model, a global importance measurement of the features was carried out using metrics such as accuracy, F1 score, and recall rate. In addition, comparative experiments were conducted with random forest (RF) and decision tree (DT) models that were optimized through Optuna to highlight the superiority of the model presented in this study.

## 2.2. Research Methodologies

### 2.2.1. Feature Selection: RF–RFECV

#### 1. Random forest (RF)

The random forest (RF) model, introduced by Breiman in 2001, is a supervised learning algorithm based on bagging [26]. Its distinct feature is the introduction of random feature extraction to construct independent training subsets, thereby increasing randomness in the generation of decision trees. It builds multiple decision trees on various resampled data samples and integrates them, with each tree casting a vote to produce the final prediction.

#### 2. Recursive feature elimination and cross-validation (RFECV)

Recursive feature elimination (RFE) [27] is frequently used for feature selection. RFE involves the iterative assessment of the importance of features in a base model, where the least important feature is removed in each iteration until a specified number of features is reached. The sequence of elimination determines the final order of feature importance. Recursive feature elimination with cross-validation (RFECV) extends RFE by incorporating resampling and cross-validation processes. It calculates validation errors for all of the feature subsets and selects the subset with the lowest error rate as the optimal feature subset [28]. The RFECV method includes the following four steps: (a) train a base model using the original feature dataset; (b) calculate the importance scores of each feature; (c) form a new feature subset by removing a feature variable through cross-validation; and (d) repeat steps (a) to (c) until a suitable number of feature variables are determined.

#### 3. RF–RFECV

As mentioned, the random forest algorithm can quickly select features for efficient dimensionality reduction. However, the random forest algorithm only calculates the importance of each feature and cannot determine the optimal number of features; thus, it requires manual setting. This approach somewhat reduces the feasibility of the results, and due to the high randomness of the random forest algorithm, the calculated feature importance also has a degree of uncertainty. Relying solely on a single random forest algorithm to select the optimal feature subset is typically unreliable. Therefore, incorporating the RFECV method into the random forest algorithm when establishing an efficient and accurate prediction model can improve the accuracy of the selection of the optimal feature subset. This method allows for effective feature selection by choosing the best feature subset based on their importance. The feature selection process designed using the RF–RFECV method is illustrated in Figure 1b (2).

### 2.2.2. Optuna

Optuna [29] employs a Bayesian optimization algorithm for hyperparameter space search, making it an efficient method for hyperparameter optimization. The Optuna module was introduced from third-party libraries to achieve efficient and automatic hyperparameter tuning, reduce the burden of manual parameter tuning, and enhance accuracy. Its main features include parallel and distributed optimization, hyperparameter space search in Python syntax, and a lightweight, versatile, and cross-platform architecture. The framework defaults to a tree-structured Parzen estimator for non-standard Bayesian optimization, simulating the generation process through transformation and substituting non-parametric



densities for previously configured distributions. The basic steps for hyperparameter optimization using Optuna are as follows:

(a) Define the search space: Optuna's distribution functions can be used to determine the range of values when defining the hyperparameter search space.

(b) Define the objective function: The objective function is the model to be optimized and can be any callable object, such as a Python function or class method. The objective function takes hyperparameter values as the input and output model performance metrics.

(c) Create an Optuna trial: An Optuna trial object is created by specifying the objective function and search algorithm.

(d) Run the Optuna trial: The Optuna trial is run for the hyperparameter search. After each trial, Optuna updates the hyperparameter values and records the current trial's performance metrics. The number of attempts or the duration can be set to control the search space size and limit the search time.

(e) Analyze trial results: Optuna's visualization tools can analyze the results and select the best combination of hyperparameters after the trial ends.

### 2.2.3. XGBoost

The core idea of the XGBoost [30] algorithm originates from boosting trees. It involves continuously adding boosting trees to form a robust classifier when integrated. The objective function of XGBoost is designed to optimize both the predictive power of the model and its complexity, ensuring a balance between accuracy and overfitting prevention [31–33]. The objective function is as follows:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (1)$$

In the formula,  $l(y_i, \hat{y}_i)$  represents the loss function;  $\hat{y}_i$  represents the predicted output;  $y_i$  represents the actual output;  $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|W\|^2$  represents the regularization term;  $f_k$  represents the  $k$ -th tree model;  $T$  represents the penalty for the number of leaves;  $W$  represents the penalty for the leaf weight values;  $\gamma$  represents the penalized regularization term for the number of leaves; and  $\lambda$  represents the leaf-weighted penalized regularization term.

The objective function of the XGBoost algorithm incorporates regularization terms, including node weights, mainly to reduce the complexity of the model and prevent overfitting. In addition, the loss function uses a second-order Taylor expansion, as shown in Equation (3). This approach effectively enhances the convergence speed and accuracy of the algorithm.

The second-order Taylor expansion of the loss function allows XGBoost to optimize the model more efficiently. By considering not only the first derivative (which represents the direction of the steepest descent) but also the second derivative (which gives information about the curvature of the loss function), XGBoost can make more informed steps during the optimization process. This results in faster convergence towards the minimum of the loss function and generally leads to a more accurate model.

$$L^{(t)} \approx \sum_{i=1}^n \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \Omega(f_t) \quad (2)$$

where  $g_i$  and  $h_i$  are the first derivative and the second derivative of the loss function, respectively.

## 3. Project Overview and Data Preprocessing

### 3.1. Project Overview

The Northern Xinjiang Water Supply Project has a total length of 516.2 km and currently holds the record for the longest water transfer tunnel under construction and in operation worldwide [34], with an average burial depth of 428 m and a maximum of

774 m. The Ka-Shuang Tunnel, which is a part of this project, extends over 283 km and was constructed using 11 open-type TBMs, covering a total excavation length of 226.58 km. This tunnel segment is notable for its TBM cluster excavation, deep burial, and extended single-heading excavation distances. The research section of this study focuses on a critical control segment within the Ka-Shuang Tunnel. Building on the traditional TBM construction experience, a “single tunnel, dual machine” construction mode was employed, where two TBMs were excavated in opposite directions until they connected with other segments. The individual excavation distances for TBMs 7 and 8 were 17.92 km and 19.67 km, respectively, totaling 37.59 km. Geological exploration data showed the proportion of different grades of surrounding rock in the tunnel section [35]; the surrounding rock of the KS Tunnel is primarily composed of rock mass types II and III, which account for 42.65% and 43.91%, respectively. With a circular excavation cross-section, this tunnel segment was constructed using open-type TBMs. The main technical parameters of the TBMs are illustrated in Table 1. The KS Tunnel data comprise the pile numbers KS 155 + 000 to KS 193 + 253; the geological cross-section at these pile numbers is illustrated in Figure 1a.

**Table 1.** Main parameters of TBM 7(TBM 8).

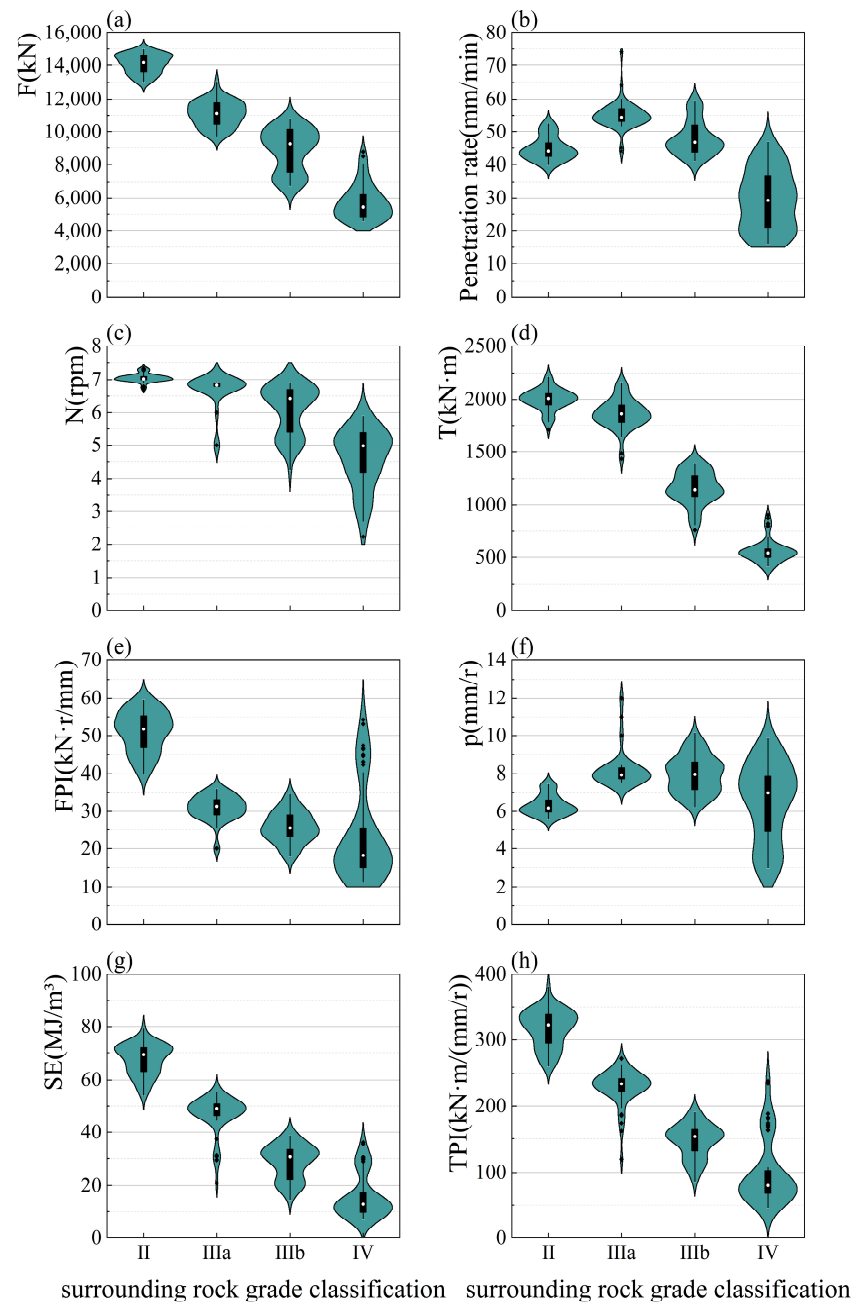
Component Location	Component Names	Component Parameters
Hard rock tunnel boring machine	Machine type	Open-type
	Overall machine length (m)	200
	Tunneling progress (mm)	1800
	Installed net horsepower power (kW)	4400
Cutter tools	Rated total thrust power (kN)	14,373
	Maximum total thrust power (kN)	22,934
	Excavation diameter (mm)	7030
	Cutterhead tool rated load (kN)	315
Cutterhead drive	Drive type	Variable frequency motor drive
	Total power (kN)	$350 \times 8 = 2800$
	Cutterhead rotation speed (rpm)	0–10.9

### 3.2. Data Acquisition and Processing

This study collected 400 sets of TBM tunneling performance data under different surrounding rock grades by combining the on-site excavation data from the KS Tunnel project. These datasets cover four surrounding rock types: II, IIIa, IIIb, and IV. The TBM tunneling parameters primarily include machine parameters, excavation performance parameters, and excavation index parameters. These parameters consist of the thrust (F), rotation speed (N), torque (T), net penetration rate (PR), penetration depth (p), field penetration index (FPI), torque penetration index (TPI), and specific excavation energy (SE). The distribution of each indicator in the engineering cases above is statistically presented in Figure 2. And the statistical characteristics are described in Table 2. The dataset validates the generalization ability of the model established in this study.

**Table 2.** Description of dataset statistics.

Description Indicators	PR	p	F	N	T	FPI	SE	TPI
Max	74.00	12.00	16,618.66	7.47	2762.63	766.04	1613.24	890.00
Min	0.12	0.01	11.63	0.10	27.67	11.21	6.00	45.03
Mean	41.56	6.50	10,923.45	6.38	1518.23	48.96	58.07	297.64
Std	13.46	2.01	3377.00	0.99	619.94	59.73	91.07	531.60
Kurtosis	1.00	1.54	−0.09	8.88	−0.71	78.62	223.92	192.93
Skewness	−1.07	−0.95	−0.64	−2.59	−0.45	7.99	13.92	13.05



**Figure 2.** Distribution of TBM tunneling parameters under different surrounding rock grades: (a) the thrust (F); (b) penetration rate (PR); (c) rotation speed (N); (d) the torque (T); (e) field penetration index (FPI); (f) penetration depth (p); (g) specific excavation energy (SE); and (h) torque penetration index (TPI).

(1) Thrust (F): The distinction between the different surrounding rock grades is outstanding, making it suitable for the classification of surrounding rock grades.

(2) Torque (T) demonstrates good discriminative ability among the IIIa, IIIb, and IV surrounding rock types, but there is an overlap between II and IIIa. Overall, the average torque can be used to distinguish between the different classes of surrounding rock.

(3) Rotation speed (N) shows a high degree of overlap in the distribution among the various classes of IIIa, IIIb, and IV surrounding rock, making it difficult to distinguish between different rock grades effectively. Under II and IIIa surrounding rock conditions, the rotation speed values are relatively concentrated, while in the IIIb and IV conditions, the box in the boxplot is noticeably longer, suggesting a more dispersed range of cutter



head rotation speeds. This dispersion might be due to unstable geological conditions, resulting in the cutter head rotation speed being significantly influenced by the geological environment. Overall, the average rotation speed can effectively reflect the different classes of surrounding rock encountered during the TBM excavation process.

(4) The penetration rate (PR) can effectively differentiate the latter three classes of surrounding rock, with an overlap occurring only between the II and IIIb surrounding rock types. The PR in the IV surrounding rock type is the lowest, with a noticeably elongated box in the boxplot, indicating a more dispersed average net penetration rate.

(5) The penetration depth (p) also shows a high degree of overlap in the distribution among various classes of surrounding rock, making it difficult to effectively distinguish between different rock grades. After analysis, it was observed that the average penetration depth in the II surrounding rock type was significantly lower than that in IIIa, IIIb, and IV. The II rock type has higher strength, necessitating a low penetration depth and high rotation speed during excavation on site. In contrast, the IV rock type, being soft and weak, requires a relatively minor penetration depth to avoid the risk of collapse. As for IIIa and IIIb, these rock classes are more stable and capable of withstanding more prominent engineering disturbances, and they exhibit the highest penetration depth.

(6) The field penetration index (FPI) can effectively identify II surrounding rock but has a lower discriminative ability for IIIa, IIIb, and IV surrounding rock.

(7) The torque penetration index (TPI) and specific excavation energy (SE) can quickly identify the II and IIIa surrounding rock types. However, they face difficulty when distinguishing between the IIIb and IV rock types due to the high overlap.

According to a statistical analysis of the datasets encompassing various features, we derived the subsequent final parameters, which included the thrust (F), torque (T), net penetration rate (PR), field penetration index (FPI), specific excavation energy (SE), and tunnel penetration index (TPI). The parameters can be considered as preliminary indicators that can be used to distinguish between the various grades of the surrounding rock. Using these datasets as an example, the correlation between the features of the datasets was analyzed; the correlation analysis of these features is presented in Figure 3.

As depicted in Figure 3, a strong correlation exists between the features. However, employing all of the features as input features may lead to redundancy and potential contradictions, thereby impacting the accuracy of the output results. Moreover, excessive input variables can impede efficient model training and impose higher data recording requirements, rendering the model less practical for real scenarios.

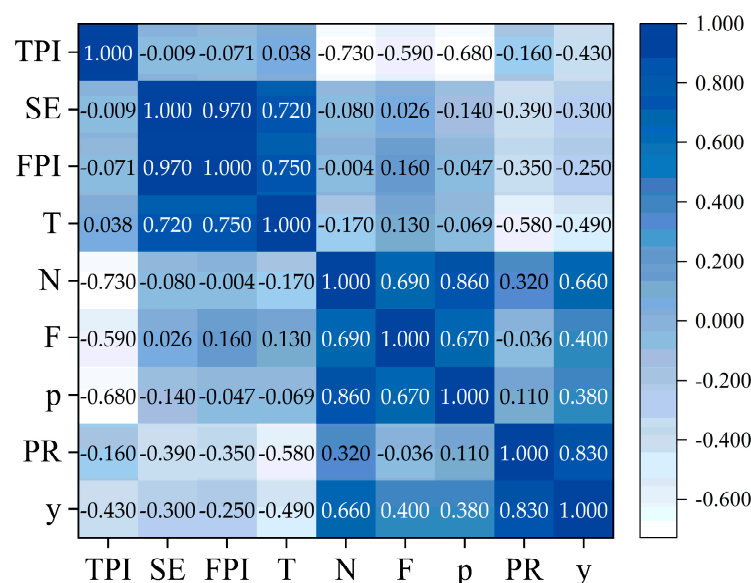


Figure 3. Correlation analysis of the independent features.

### 3.3. Optimal Feature Subset Selection

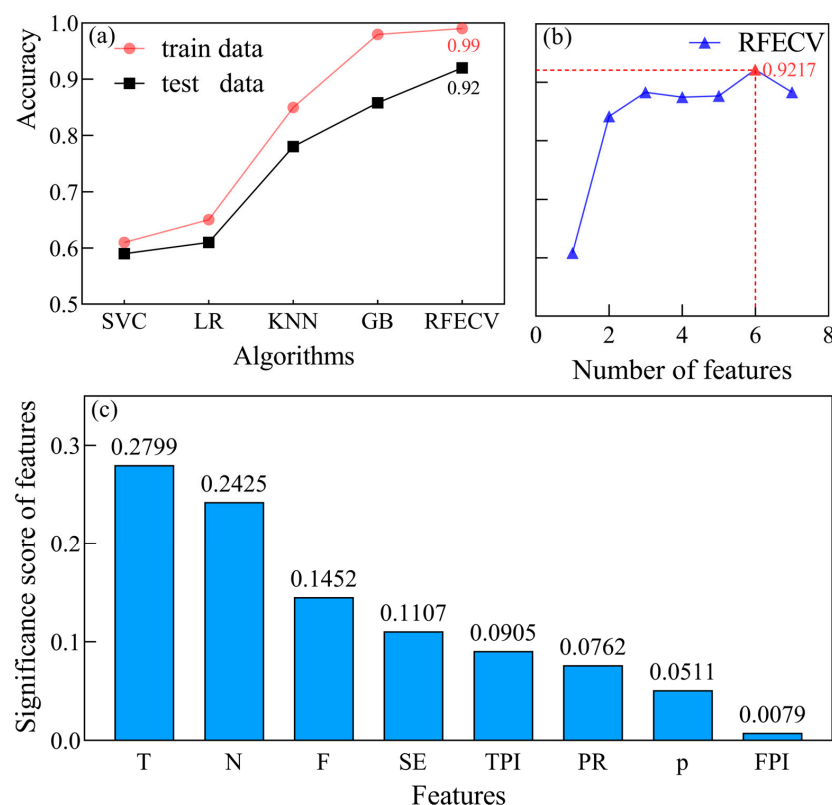
In machine learning tasks, the input features of a model play a crucial role in determining its upper limit. Moreover, the selection of input feature sets also impacts the entire model's training time and generalization ability. Therefore, careful consideration is necessary when selecting input features for a model. This study employs RFECV [36,37] to determine the optimal combination of model features and uses it to evaluate reserved model feature performance. RFECV achieves an adaptive determination of feature combinations through cross-validation, providing more accurate and objective performance evaluations under different feature combinations. Thus, the researcher utilizes RFECV to select the necessary input features that can be used to identify surrounding rock grades.

#### 3.3.1. Feature Selection Based on Recursive Feature Elimination Method

Using RF–RFECV for variable selection, the least important feature can be eliminated in each iteration, and the scores of each feature are adjusted through repeated iterations. Ultimately, the optimal feature subset is selected via cross-validation. When running RFECV, specific hyperparameters are chosen: RF is used as the base model; 10-fold cross-validation is employed; one feature is eliminated at a time (step = 1); and the minimum number of retained features is set to one.

(1) Accuracy rates using these four feature selection methods.

Accuracy measures how often the model is correct when making predictions. In order to evaluate the effectiveness of the RF–RFECV feature selection method, various techniques such as gradient boosting (GB), logistic regression (LR), k-nearest neighbors (KNN), support vector machines (SVCs), and RF–RFECV are used for the selection and extraction of feature variables. The experiments were conducted twice, and the average accuracy was computed to ensure result reliability. Figure 4a (sorted by accuracy) illustrates the achieved accuracy rates of these four feature selection methods.



**Figure 4.** Results of optimal feature subset selection: (a) accuracy rates of these four feature selection methods; (b) classification accuracy corresponding to the number of selected feature variables; and (c) random forest feature importance ranking.

(2) Research on the relationship between the number of feature variables and accuracy.

The RF–RFECV process described earlier was utilized to generate a graphical representation depicting the relationship between classification accuracy and the number of feature variables, as depicted in Figure 4b.

### 3.3.2. Feature Importance Measure for RF

In this study, the random forest algorithm was employed to rank the importance of the features and to further reduce feature dimensionality. The calculation result is illustrated in Figure 4c. The significance of the feature variables was evaluated using the mean decrease in impurity (MDI), which is based on the Gini index. Throughout the feature reduction process, the random forest algorithm utilizes its inherent MDI metric to prioritize feature variables, reflecting their respective roles within the random forest algorithm. The Gini index is calculated and averaged to compare different feature variables, identifying those with a more substantial influence and those with a more efficient ability to recognize surrounding rock grades during the classification process.

$$Gini(p) = \sum_{k=1}^K p_k(1 - p_k) = 1 - \sum_{k=1}^K p_k^2 \quad (3)$$

In the formula,  $K$  represents the overall count of sample categories, and  $p_k$  represents the weights assigned to category  $K$  for sampling.

## 4. Surrounding Rock Grade Identification Based on Optuna–XGBoost

### 4.1. The Process of Constructing a Model

The selection of input indicators plays a crucial role in ensuring the accuracy of predictions. This study carefully chose six key input features through statistical analysis and feature selection: T, N, F, SE, TPI, and PR. These extracted indicators were then utilized as the input vector for the Optuna–XGBoost model to establish a TBM excavation rock grade identification model capable of accurately predicting surrounding rock grades during the TBM tunneling process.

### 4.2. Hyperparameter Optimization Based on Optuna

The performance of the XGBoost classification model can be enhanced through hyperparameter optimization; this is a crucial step in the achievement of the optimal combination of model parameters.

The XGBoost multiclassification algorithm requires the specification of the following metrics in order to proceed: The objective function (objective) should be set as ‘multi:softmax’; the evaluation metrics (eval\_metric) should be set as ‘mlogloss’; and the number of categories should be set as ‘4’ for optimal results.

Regarding the XGBoost algorithm, ‘n\_estimators’, ‘learning\_rate’, ‘max\_depth’, ‘subsample’, ‘colsample\_bytree’, ‘reg\_alpha’, and ‘reg\_lambda’ have a significant impact on the performance of the algorithm.

Optuna optimization assists in finding the best hyperparameters from the search space based on prior evaluations [38]. Unlike hyperparameter optimization methods such as HyperOpt, Scikit-Optimizer, and AutoKeras, Optuna employs adequate sampling and pruning strategies to dynamically construct the hyperparameter search space. These strategies are beneficial for the achievement of high performance with limited resources [39]. Optuna treats the minimization/maximization of the target function as an input for a given hyperparameter search and returns the validation score as an output. All of the other machine learning models proposed in this study were optimized using Optuna. The hyperparameter optimization configurations used for the XGBoost algorithm based on Optuna are presented in Table 3.

**Table 3.** Hyperparameter optimization of XGBoost.

Algorithm	Hyperparameter	Search Space Best	Optimal Hyperparameter
XGBoost	n_estimators	[5, 300]	204
	learning_rate	[0.01, 1]	0.2018
	max_depth	[3, 50]	9
	subsample	[0.7, 1.0]	0.4001
	colsample_bytree	[0.7, 1.0]	0.4221
	reg_alpha	$[10^{-8}, 10]$	0.0024
	reg_lambda	$[10^{-8}, 10]$	$7.01 \times 10^{-7}$

#### 4.3. The Metrics for Evaluating Models

When evaluating machine learning models, using graphical methods for assessment is merely an intuitive approach rather than a rigorous mathematical expression. Therefore, it is essential to select appropriate evaluation metrics to identify and choose the best-performing models. These evaluation metrics, as shown in Table 4, are commonly used to judge the quality of predictive models [40]. It is important to note that different evaluation metrics handle prediction errors differently. Hence, when choosing evaluation metrics, one must consider the characteristics of the data, as well as specific scenarios and business requirements.

**Table 4.** The metrics for evaluating models.

Name	Title 2	Title 3
Accuracy	$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$	<p>True positive (TP) indicates number correctly predicted to be in the positive category</p> <p>False positive (FP) indicates number of incorrectly predicted positive classes</p> <p>True negative (TN) indicates number of correctly predicted negative categories</p> <p>False negative (FN) indicates number of incorrectly predicted negative classes</p> <p>Accuracy represents proportion of correctly categorized samples to all samples</p>
Precision	$Precision = \frac{TP}{TP + FP}$	The parameters are defined as above. Precision represents the proportion of the number of correctly predicted positives to the total number of positive predictions made.
Recall	$Recall = \frac{TP}{TP + FN}$	The parameters are defined as above. Recall represents the proportion of the number of correctly predicted positives to the total number of actual positives.
F1	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	The parameters are defined as above. F1 represents the harmonic mean of precision and recall and is used to measure the accuracy of the model.
Confusion matrix	Comprised of four main components: TP, FP, TN, and FN	The parameters are defined as above. Confusion matrix is used to assess the performance of classification models and represents the relationship between the predicted results for each category and the actual categories.
Kappa	$Kappa = \frac{P_0 - P_e}{1 - P_e}$	<p><math>P_0</math> indicates proportion correctly predicted by the model</p> <p><math>P_e</math> indicates expected accuracy of the model's stochastic predictions</p> <p>Kappa represents a statistical measure used to evaluate the performance of a classifier, especially in cases of imbalanced data. This measure takes into account the effect of random agreement, providing a performance assessment that goes beyond mere accuracy. The value 1 indicates perfect agreement. The value 0 indicates that the observed agreement is no better than the random prediction. Negative values indicate agreement that is worse than the random prediction.</p>

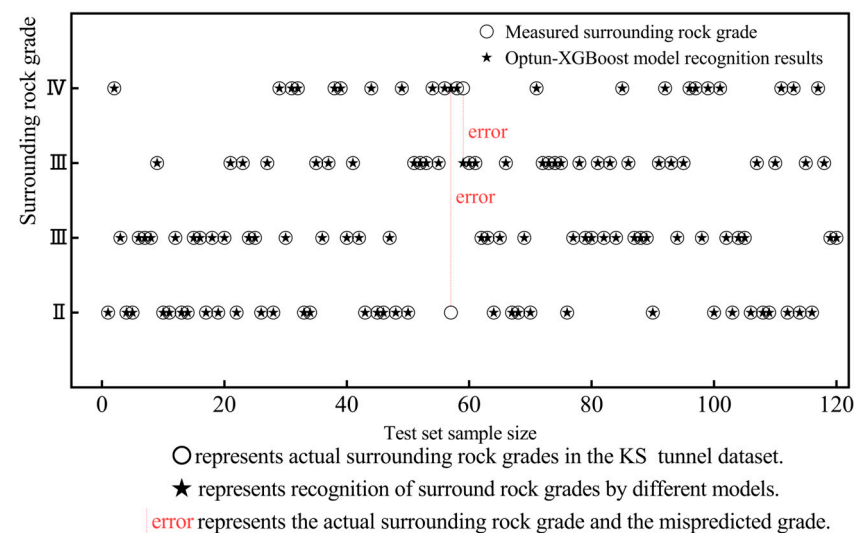
## 5. Results

This section comprehensively evaluates the performances of the different algorithms that were used to identify the surrounding rock grades.

Firstly, in Section 3.3, we explored the feature importance and how the number of features affected the classification performances. Then, four types of models, the XGBoost, gradient boosting decision tree (GBDT), decision tree (DT), and random forest (RF) algorithms, were established for the experimental analysis based on the use of the optimal feature subset as input variables. Five standard metrics (Table 4), accuracy, precision, recall, F1 score, and the confusion matrix, were used to evaluate the performance of each model. Additionally, the quality of each model's hyperparameter settings directly affected the final prediction results. We compared each machine learning model's identifying results based on the Optuna optimal hyperparameters that were tuned according to the metrics listed in Sections 4.2 and 5.4, respectively. Lastly, the dataset was divided into a training set and a test set in a 7:3 ratio, and the data were shuffled during training, validation, and testing to ensure that the model training was not affected by the sequence structure of the data. The trained models were then evaluated using the test set. Furthermore, to test the performance of Optuna optimization, the unoptimized XGBoost and PSO-XGBoost models were also included for comparison. The results are presented in Sections 5.1–5.4.

### 5.1. The Test Results of the Optuna–XGBoost Model

Based on the above, the optimal feature subset of T, N, F, SE, TPI, and PR, selected through RFECV, serves as the set of input variables for the Optuna–XGBoost model, with the surrounding rock grade as the output variable. Subsequently, the optimal hyperparameter combination provided in Table 3 configures the classification model, and the optimized model is trained on the training set. Finally, the model's ability to perceive and recognize surrounding rock grades is tested using 120 datasets from the test set, and the results are presented in Figure 5. The metrics used to evaluate Optuna–XGBoost are also presented, as shown in Table 5.



**Figure 5.** Recognition results of the Optuna–XGBoost Model.

**Table 5.** The metrics for evaluating Optuna–XGBoost.

Model	Accuracy	Precision	Recall	F1
Optuna–XGBoost	0.9833	0.9803	0.9813	0.9807

The analysis of the results in Figure 5 and Table 5 reveals that the Optuna–XGBoost model is highly aligned with the actual situation in its predictions of surrounding rock



grades, with only two incorrect predictions. The accuracy of the Optuna–XGBoost model reached 0.983 in the recognition of surrounding rock grades, which demonstrates that the model is entirely precise. Overall, the Optuna–XGBoost model exhibits high metrics in predicting TBM surrounding rock grades. However, for a more scientifically rigorous evaluation of this model, a comparison with other models is necessary.

### 5.2. Validation of Optimal Feature Subset

To validate the efficiency of the RF–RFECV feature selection algorithm, the optimal feature subset derived from the RF–RFECV algorithm was compared with the complete feature dataset. XGBoost served as the classifier in the subsequent analysis, with the test set employed to assess the model’s classification performance based on predefined evaluation metrics. The results of this comparison are presented in Table 6.

**Table 6.** Validation results for the optimal feature subset.

Feature Set	Accuracy	Precision	Recall	F1
Complete feature set	0.9833	0.9803	0.9813	0.9807
Optimal subset of features	0.9083	0.9074	0.9114	0.9090

The utilization of XGBoost as the classification model yields remarkable results, as shown in Table 6, with the accuracy, precision, and F1 score surpassing 90%. This indicates the effectiveness of the identification capability of XGBoost in determining the grade of surrounding rock. The RF–RFECV feature selection algorithm was employed to construct the XGBoost classification model. Compared with the complete feature set, the optimal subset of features as input variables enhanced the four evaluation indexes by 8.26%, 8.03%, 7.67%, and 7.89%, respectively. Hence, this demonstrates that RF–RFECV feature selection is efficacious in reducing feature dimensions and that it is the optimal feature subset to use as input for the subsequent classification models to achieve an accurate recognition of surrounding rock grades.

### 5.3. Performance Comparison of Optuna Hyperparameter Optimization

Adjusting the hyperparameters to suit a specific dataset is critical in enhancing model performance. Hyperparameter optimization methods such as grid search, random search, and Bayesian optimization provide systematic ways to explore and optimize these parameters, thereby identifying the best model configuration to improve performance. To validate the effectiveness of Optuna hyperparameter optimization, unoptimized XGBoost and PSO–XGBoost models were used for comparative analysis with the model presented in this study regarding model performance. The performances of each model on the test set were assessed in terms of accuracy, precision, recall, and F1 scores.

### 5.4. Performance Comparison with other Algorithms

To further validate the superiority of the RFECV–Optuna–XGBoost model in recognizing the surrounding rock grades of TBM excavation, a comparative analysis was conducted using the RF, GBDT, and DT algorithms against the model presented in this study. Moreover, to maximize the performance of each model, Optuna hyperparameter optimization was performed on all three models to obtain their respective optimal hyperparameter combinations. The key hyperparameters of each model, the settings of the optimized hyperparameters, and their ranges are listed in Table 7, and the final optimization results are also presented in Table 7. It should be noted that the same dataset was used for training and testing all three algorithms.

Based on the hyperparameter optimization results shown in the table, after Optuna optimization, the RF algorithm’s optimal hyperparameter combination included ‘n\_estimators’ at 290, ‘learning\_rate’ at 0.4856, ‘max\_depth’ at 13, ‘min\_samples\_split’ at 10, ‘min\_samples\_leaf’ at 6, and ‘max\_features’ at ‘auto’. For the GBDT model, the best

hyperparameter combination obtained through optimization included ‘n\_estimators’ at 51, ‘learning\_rate’ at 0.4674, ‘max\_depth’ at 18, ‘min\_samples\_split’ at 6, and ‘min\_samples\_leaf’ at 9. For the DT model, the best hyperparameter combination obtained through optimization included ‘criterion’ at ‘gini’, ‘max\_depth’ at 21, ‘min\_samples\_split’ at 0.1019, and ‘min\_samples\_leaf’ at 3.

**Table 7.** Hyperparameter optimization of each algorithm.

Algorithm	Hyperparameter	Best Search Space	Optimal Hyperparameter
RF	n_estimators	[50, 300]	290
	learning_rate	[0.01, 1]	0.4856
	max_depth	[5, 30]	13
	min_samples_split	[2, 10]	10
	min_samples_leaf	[0.1, 10]	6
	max_features	['auto', 'sqrt', 'log2']	'auto'
GBDT	n_estimators	[50, 300]	51
	learning_rate	[0.01, 1]	0.4674
	max_depth	[5, 30]	18
	min_samples_split	[0.1, 10]	6
	min_samples_leaf	[0.1, 10]	9
DT	criterion	['gini', 'entropy']	'gini'
	max_depth	[5, 30]	21
	min_samples_split	[0.1, 10]	0.1019
	min_samples_leaf	[1, 10]	3

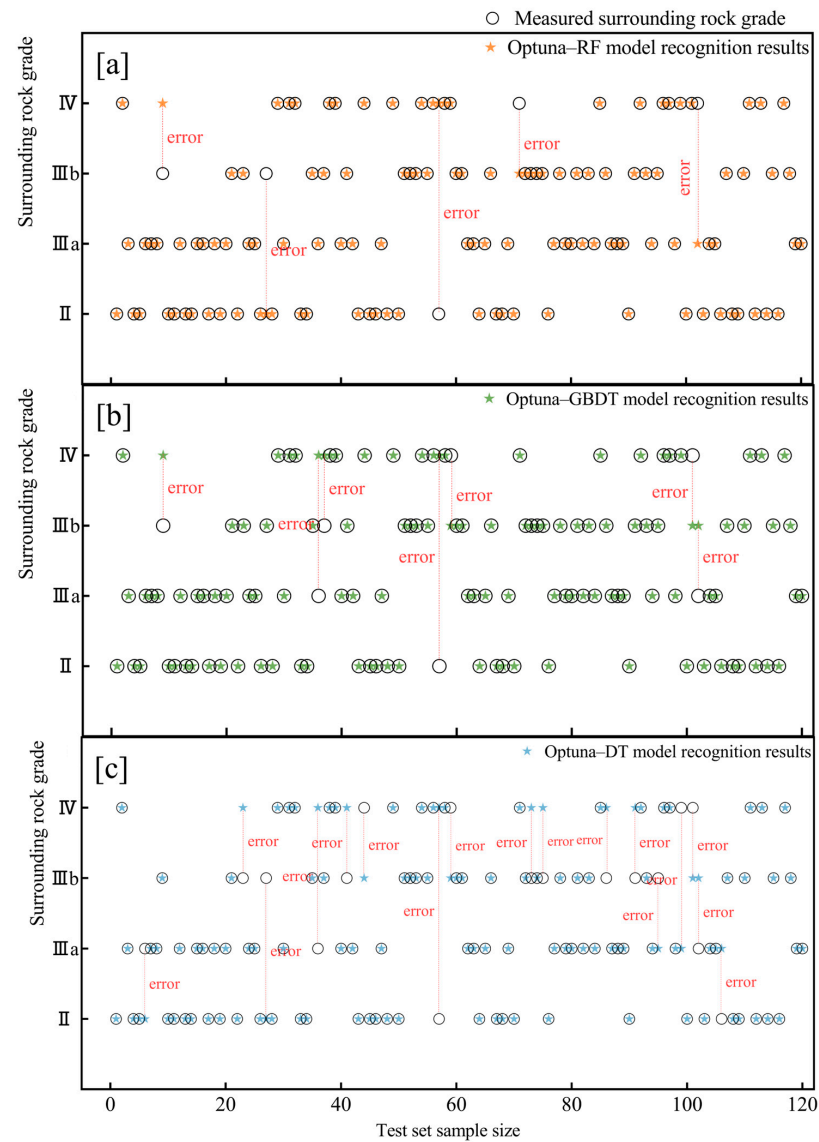
Applying these optimal hyperparameter combinations to the RF, GBDT, and DT algorithms and testing them on the KS Tunnel dataset resulted in the achievement of the accuracy, precision, recall, and F1 scores and the generation of confusion matrices.

From Section 5.3 to Section 5.4, we compare the Optuna–XGBoost algorithm with the RF, GBDT, DT, PSO–XGBoost, and XGBoost algorithms. These results are illustrated in Figures 6 and 7, respectively. The performance of each model on the test set, including the accuracy, precision, recall, and F1 scores and the confusion matrix, is presented in Figure 8.

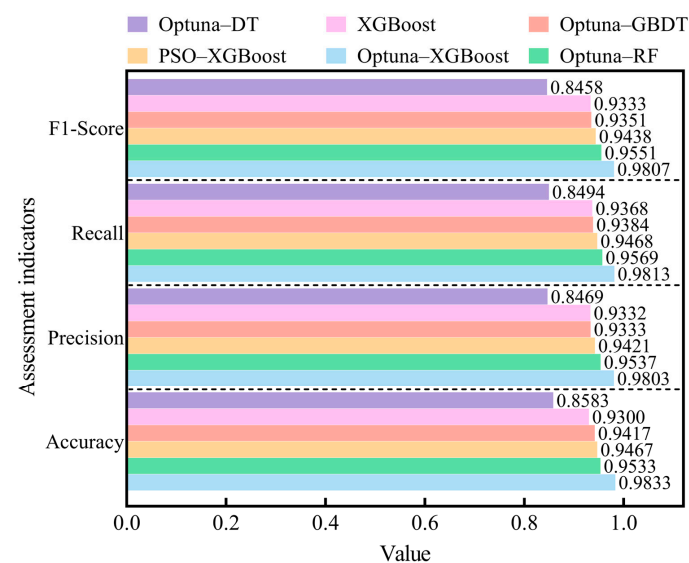
As shown in Figure 7, compared to the unoptimized XGBoost and PSO–XGBoost models, the four evaluation metrics improved by 4.18–5.35%, 3.54–4.64%, 4.16–5.15%, and 3.59–5.45%, respectively. This indicates that Optuna is a more efficient method for hyperparameter optimization than PSO.

Figures 6 and 7 show that among the six algorithms, the Optuna–XGBoost algorithm exhibited a notably superior overall performance. Compared to the underperforming DT model, the RFECV–Optuna–XGBoost model in this study achieved the best scores in all four evaluation metrics on the test set: accuracy, precision, recall, and F1 score. The improvements over the RF, GBDT, and DT in these metrics were 3.15–14.56%, 2.79–15.75%, 2.55–15.53%, and 2.68–15.95%, respectively. Therefore, the RFECV–Optuna–XGBoost model demonstrates excellent overall performance on the test set.

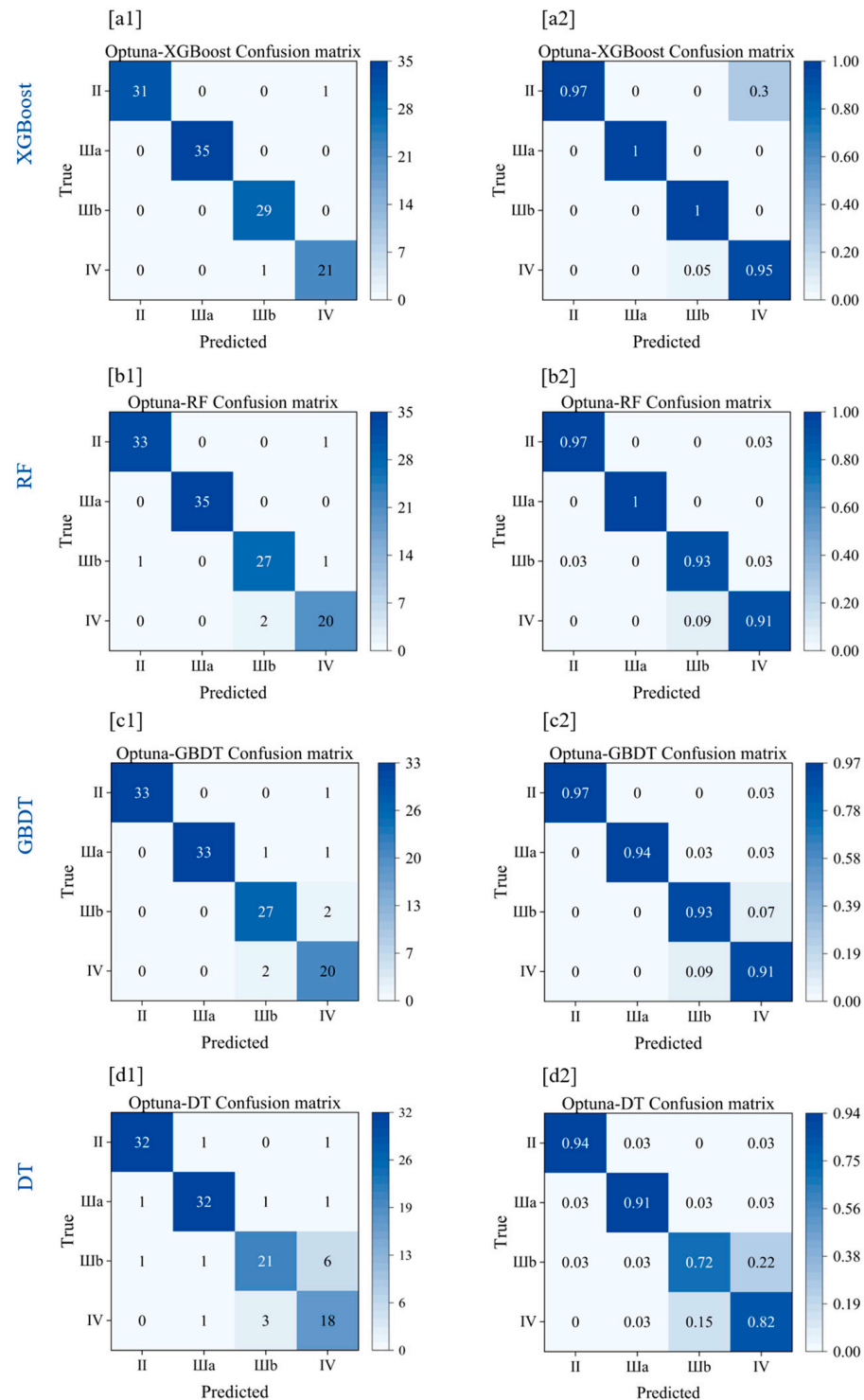
In addition, the confusion matrix results must be considered. Figure 8 shows that only one prediction error was misclassified as IV in cases where the surrounding rock grade was II. When the actual surrounding rock grade was IV, there was only one prediction error, a misclassification as IIIb. These results demonstrate the effectiveness of the Optuna–XGBoost model in accurately predicting surrounding rock grades under various conditions.



**Figure 6.** Recognition results of various models. (a) Optuna-RF model recognition results; (b) Optuna-GBDT model recognition results; and (c) Optuna-DT model recognition results.



**Figure 7.** Performance comparison with other algorithms.



**Figure 8.** Confusion matrices for different models: (a1,a2) XGBoost; (b1,b2) RF; (c1,c2) GBDT; and (d1,d2) DT.

## 6. Discussion

The results of this research can provide necessary guidance for TBM drivers in the actual digging process. In the process of digging, TBM drivers cannot directly observe the surrounding rock palm surface; the actual situation of the surrounding rock cannot be accurately obtained in real time, and it is difficult to quickly and accurately identify the surrounding rock level at the construction site and adjust the TBM digging parameters according to the change in the surrounding rock level to avoid the construction risk.

### 6.1. Effects of Different Features on Identifying Results

Related studies have demonstrated that reducing irrelevant features can reduce model complexity, and reducing data dimensionality through feature selection can improve the model's classification performance and enhance the model's generalization performance [41]. This is consistent with the results of this study. In this study, we investigate the model performance of feature approximation using backward elimination based on importance metrics. The RFECV process allows us to systematically assess the importance of each feature by recursively removing features and evaluating the model's performance using cross-validation. This approach was chosen because it effectively identifies the features that contribute the most to the identification of the enclosure grades while also considering the interactions between the excavation parameters and their joint impact on the identification performance. By selecting these features, the Optuna–XGBoost model can achieve superior performance by focusing on the most informative variables, which reduces the risk of overfitting and improves the generalization of new data.

Through detailed experimental analyses, we demonstrate that the selected features significantly impact model performance. For example, we demonstrate that the model's accuracy significantly improves after applying feature selection. In contrast, the model's complexity is reduced, further validating the effectiveness of our feature selection strategy. As seen from Figure 4a, RFECV–RF extracts features with the highest accuracy among all of the methods on both the training and test sets. From Figure 4b, it can be seen that, by adaptively approximating the input features to an optimal subset of features containing only six metrics, the method achieves an accuracy of 0.9217. The results show that the optimal feature subset obtained by ranking the features in terms of their importance includes T, N, F, SE, TPI, and PR, which is consistent with the preliminary metrics derived from the preliminary statistical analysis. T and N contribute the most to the model among the six input features.

### 6.2. Effects of Different Algorithms on Identification Results

The RFECV–Optuna–XGBoost model, when applied to the data prediction and analysis of the KS Tunnel, demonstrated superior performance on the test set, with an accuracy of 0.9833, a precision of 0.9803, a recall of 0.9813, and an F1 score of 0.9807. The main reason for the different performances of the different recognition algorithms is their different capabilities.

#### 6.2.1. Discussion on the Effectiveness of XGBoost

The first aspect of this section is the algorithmic principle of the model [42]. The DT model fits the training data perfectly, especially when the tree has a significant depth and many nodes. However, this makes the model very sensitive to noise in the training data, leading to poor performances with unseen data. Second, in multiclassification problems, the DT model requires more divisions to correctly classify all of the classes, which can cause the tree to become very complex and extensive. This increases the computational cost and may increase the risk of overfitting. RFs consist of multiple decision trees with voting or averaging to improve predictive accuracy and stability. However, some aspects could be improved when dealing with multicategory problems. For example, in the case of category imbalance [43], random forests tend to give higher importance to the majority category, resulting in poor classification performance for the minority category. This issue is further illustrated in the subsequent presentation of limitations and in the discussion. In addition, the random forest model integrates the results of multiple decision trees, which makes the model interpretation worse. GBDT builds the decision trees step by step, and each tree learns the residuals of the previous tree to reduce the model error. It performs well when dealing with regression and binary classification problems. However, errors generated in the upper layers may propagate to the lower layers in decision trees with multiple classes, leading to less satisfactory classification results [44]. Like the RF model, it is also affected by the imbalance in the number of samples in each category and the difficulty



in interpreting the model. The XGBoost model learns the nonlinear relationships of the data by constructing multiple decision trees and shows remarkable results when dealing with complex data feature structures. In addition, XGBoost has excellent robustness. It can handle missing values and category imbalance datasets with built-in regularization terms, effectively preventing overfitting and improving generalization ability. The results of this study are the same as those reported by Saporetti and Xie. Saporetti combined gradient tree boosting with differential evolution to identify lithology [45]. Xie et al. used five machine learning methods to identify lithology. They concluded that XGBoost performed the best in the ensemble learning method for identification [46]. In addition, in some related studies in the fields of fault diagnosis and classification, XGBoost has shown greater advantages than other models, and the study in this paper is consistent with the results in [47–50].

#### 6.2.2. Discussion on the Effectiveness of Optuna

Regarding hyperparameter optimization, XGBoost has many tunable hyperparameters, such as the learning rate, maximum depth of the tree, subsample ratio, etc., constituting a high-dimensional parameter space. It is a challenge to perform an effective search in this space. PSO is a population-based optimization algorithm with drawbacks such as high computational cost, slow convergence, and sensitive parameter settings [51]. In contrast, Optuna has an efficient search strategy. Optuna uses advanced search algorithms such as Bayesian optimization and TPE to find the optimal hyperparameters more efficiently and to find parameter combinations with better performance faster than the traditional grid search or random search. Optuna's pruning feature automatically stops the evaluation of unpromising trials. The ability to prune ineffective trials early can lead to faster convergence in a superior model configuration, contributing to the observed performance gains. In addition, Optuna's automated pruning strategy and parallelization support give Optuna the advantage of being able to explore more parameter combinations and run multiple experiments simultaneously with the same computational budget. The Optuna–XGBoost algorithm offers better classification performance, simpler training models, and higher training efficiency and is less susceptible to the effects of training data. This efficiency is likely to be a significant contributor to the improvements in performance metrics. As a result, it exhibits the best classification accuracy [52,53]. By exploiting the above properties, Optuna can achieve significant performance gains when optimizing the hyperparameters of the XGBoost model.

We observed an improvement in all four evaluation metrics for the Optuna–XGBoost recognition model mentioned in this study. In other studies, Sun et al. [54], Chen et al. [55], and Mehdary [56] used machine learning techniques followed by hyperparameter optimization, with the Optuna algorithm obtaining the highest accuracy.

#### 6.3. Discussion of Research Limitations

The classification recognition of surrounding rock based on RFECV–Optuna–XGBoost has a high accuracy. However, like all studies, it has limitations, such as a single data type, insufficient scale, and the fact that geological and rock mass parameters are not considered. If the established model is directly applied to other regions, it may not be able to obtain more optimized results. Therefore, the steps to establish the dataset are proposed, and the following three points are discussed.

(1) Firstly, it is necessary to consider collecting more types of characteristic variable data through advanced geological prediction methods and to establish a dataset that comprehensively considers the rock–machine relationship. All of the data used in this study are numerical data, which will generate other data types, such as image type and text type, in the tunneling process. For example, in previous studies, the characteristics of rock slag were analyzed to verify the feasibility of rock slag in order to judge the surrounding rock conditions, and the identification and classification of the surrounding rock were realized based on the characteristics of the rock slag and the parameters. This means that

when classifying other types of data, the data preprocessing process will be more complex and will require the analysis of multiple and multi-feature data types.

(2) The identification of surrounding rock grades is usually a multiclass classification problem, and the main challenge in solving the identification problem is the phenomenon of class imbalance in the dataset, which will significantly impact the accuracy of the model. Therefore, when constructing the dataset, we thoroughly considered this factor and chose the method of sample size balance. Many approaches to overcoming the class imbalance problem have been proposed [57,58]. The most commonly used methods involve the implementation of various class balancing algorithms, oversampling (such as SMOTE) [59,60], undersampling, cost-sensitive learning [61,62], or ensemble methods [63–65] tailored for imbalanced datasets; these are used to solve the problem of the uneven data scale distribution of different dominant lithologies in the dataset.

(3) The lithologies involved in this study include tuffaceous tuff-breccia or tuff with tuffaceous breccia, sandstone, agglomerate, or volcanic breccia, tuffaceous sandstone, granodiorite, and calcareous sandstone (Section 3.1). This system can provide a reference for similar types of TBM tunnel engineering. Further tests and optimization are needed for the different lithologies of engineering.

In future work, we will focus on developing and optimizing the methods of identifying the surrounding rock grade, exploring the application of model generalization ability in different geological conditions, integrating the rock–machine relationship (including physical and mechanical properties such as rock strength, fracture development degree, and water content), and using field observation information to improve the accuracy and efficiency of the surrounding rock grade identification. Meanwhile, we will also develop a more intelligent real-time identification and early warning system for surrounding rock in TBM tunneling; this system will be able to monitor the real-time changes in the surrounding rock grade in TBM tunneling and adjust the tunneling parameters in time to adapt to different geological conditions. In addition, we also plan to focus on studying the variation law of TBM tunneling parameters before adverse geological disasters such as mud bursts, water inrush, large deformation, and rock bursts in order to provide adequate early warnings.

## 7. Conclusions

This study proposes a model for identifying the grade of surrounding rock based on TBM excavation parameters and Optuna–XGBoost algorithms. First, the RFECV algorithm is used for the global measurement of feature importance in the model’s predictions; ultimately, the optimal feature subset was selected and compared with the complete feature set. Then, the Optuna method was used to optimize the XGBoost model’s hyperparameters to improve the recognition model’s accuracy. Finally, the accuracy, precision, recall, F1 score, and confusion matrix of the Optuna–XGBoost model were compared with those of other algorithms to verify its effectiveness. The main conclusions drawn are as follows:

(1) Feature importance was measured using RF. The results show that T and N contributed the most to the model among the six input features, with contribution scores of 0.2798 and 0.2424, respectively. PR contributed the least to model performance, with a score of only 0.0762.

(2) The use of the RF–RFECV feature selection algorithm to construct the XGBoost classification model was better than the full feature set in each evaluation criterion; the four evaluation indicators were increased by 8.26%, 8.03%, 7.67%, and 7.89%. This indicates that RF–RFECV can enhance the accuracy of identifying surrounding rock grades.

(3) The Optuna–XGBoost algorithm was compared with the RF, GBDT, DT, PSO–XGBoost, and XGBoost algorithms, based on the TBM excavation parameter dataset of the KS Tunnel. The experimental results show that the accuracy of the Optuna–XGBoost model was 0.9833, which was much higher than that of the other algorithms. The superiority of the Optuna–XGBoost algorithm in dealing with transformer fault diagnosis problems is verified. A new method for predicting the grade of surrounding rock based on TBM excavation parameters is proposed.

**Author Contributions:** Conceptualization, R.S. and K.S.; methodology, T.F.; resources, Z.L.; data curation, J.Z.; writing—original draft preparation, R.S.; writing—review and editing, K.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** This study comes from a confidential project and data cannot be disclosed due to privacy concerns.

**Acknowledgments:** Thanks to all the anonymous reviewers for their value comments on this paper, which improved quality of our paper.

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

## References

1. Xinhua News Agency. The 14th Five-Year Plan for National Economic and Social Development of the People's Republic of China and the Outline of the long-range goals for 2035. *China Water Resour.* **2021**, *6*, 1–38. (In Chinese)
2. Hong, K.; Feng, H. Development and thinking of tunnels and underground engineering in China in recent 2 years (from 2019 to 2020). *Tunn. Constr.* **2021**, *41*, 1259–1280. (In Chinese)
3. Du, L. Progresses, challenges and countermeasures for TBM construction technology in China. *Tunn. Constr.* **2017**, *37*, 1063–1075. (In Chinese)
4. Liu, J.; Xiao, X.; Yang, H.; Fu, D. A study on key construction techniques for tunnel boring machines adopted in super-long tunnels. *Mod. Tunn. Technol.* **2005**, *42*, 37–44. (In Chinese)
5. Deere, D.U.; Hendron, A.J.; Patton, F.D.; Cording, E.J. Design on surface and near surface construction in rock. In Proceedings of the 8th U.S. Symposium on Rock Mechanics (USRMS), Minneapolis, MI, USA, 15–17 September 1966.
6. Hamidi, J.K.; Shahriar, K.; Rezai, B. Performance prediction of hard rock TBM using Rock Mass Rating (RMR) system. *Tunn. Undergr. Space Technol. Inc. Trenchless Technol. Res.* **2010**, *25*, 333–345. [\[CrossRef\]](#)
7. Barton, N.; Line, R.; Lunde, J. Engineering classification of rock masses for the design of tunnel support. *Rock Mech. Rock Eng.* **1974**, *6*, 183–236. [\[CrossRef\]](#)
8. Barton, N. *TBM Tunneling in Jointed and Faulted Rock*; Taylor & Francis: Abingdon, OX, UK, 2000; pp. 72–73.
9. Wu, A.; Liu, F. Advancement and application of the standard of engineering classification of rock masses. *Chin. J. Geotech. Eng.* **2012**, *31*, 1513–1523.
10. Cai, F. Discussion about several problems of the use of standard for engineering classification of rock masses. *Rock Soil Mech.* **2003**, *24*, 74–77.
11. Barton, N. Comments on 'A critique of Q TBM'. *T T Int.* **2005**, *7*, 37.
12. Gong, Q.; Lu, J.; Xu, H.; Chen, Z.; Zhou, X.; Han, B. A modified rock mass classification system for TBM tunnels and tunneling based on the HC method of China. *Int. J. Rock Mech. Min. Sci.* **2020**, *137*, 104551. [\[CrossRef\]](#)
13. Ji, F.; Shi, Y.; Li, R.; Zhou, C.; Zhang, N.; Gao, J. Modified O-index for prediction of rock mass quality around a tunnel excavated with a tunnel boring machine (TBM). *Bull. Eng. Geol. Environ.* **2019**, *75*, 3755–3766. [\[CrossRef\]](#)
14. He, F.; Gu, M.; Wang, C. Study on surrounding rock classification of tunnel cut by TBMs. *Chin. J. Rock Mech. Eng.* **2002**, *21*, 1350–1354. (In Chinese)
15. Li, C.; Peng, Y. Discussion about surrounding rock classification of tunnel excavate by TBMs. *J. China Foreign Highw.* **2006**, *26*, 235–237. (In Chinese)
16. Bieniawski, Z.T.; Celada, B.; Galera, J.M. TBM Excavability: Prediction and machine-rock interaction. *Proc. RETC* **2007**, *01*, 1118–1130.
17. Xue, Y.; Li, X.; Diao, Z.; Zhao, F. A novel classification method of rock mass for TBM tunnel based on penetration performance. *Chin. J. Geotech. Eng.* **2018**, *37* (Suppl. S1), 3382–3391. (In Chinese)
18. Postami, I. Development of a Force Estimation Model for Rock Fragmentation with Disc Cutters through Theoretical Modeling and Physical Measurement of Crushed Zone Pressure. Master's Thesis, Colorado School of Mines, Mines, CO, USA, 1997.
19. Bruland, A. Hard Rock Tunnel Boring. Master's Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2000.
20. Xiong, F. Research of the TBM Excavation Efficiency Prediction and Rock Classification Based on the PSO-SVR Algorithm. Master's Thesis, Chang'an University, Xi'an, China, 2016.
21. Zhang, Q.; Liu, Z.; Tan, J. Prediction of geological conditions for a tunnel boring machine using big operational data. *Autom. Constr.* **2019**, *100*, 73–83. [\[CrossRef\]](#)
22. Liu, Q.; Wang, X.; Huang, X. Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data. *Tunn. Undergr. Space Technol.* **2020**, *106*, 103595. [\[CrossRef\]](#)

23. Li, H. Prediction and identification method of tunnel boring machine surrounding rock grade based on tunneling parameters inversion. *Tunn. Constr.* **2022**, *42*, 75–82. (In Chinese)
24. Xi, M. Research on Identification of Rock Type and Operating Parameter Decision of TBM Based on Engineering Data Analysis. Master's Thesis, Zhejiang University, Zhejiang, China, 2020.
25. Wu, Z.; Fang, L.; Weng, L. A classification and boreability perception and recognition method for rock mass based on TBM tunneling performance. *Chin. J. Geotech. Eng.* **2022**, *41* (Suppl. S1), 2684–2699. (In Chinese)
26. Breiman, L. Random forest. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
27. Su, X.; Liu, H.; Tao, L. TF entropy and RFE based diagnosis for centrifugal pumps sub-jeet to the limitation of failure samples. *Appl. Sci.* **2020**, *10*, 2932. [\[CrossRef\]](#)
28. Shang, Q.; Feng, L.; Gao, S. A Hybrid Method for Traffic Incident Detection Using Random Forest-Recursive Feature Elimination and Long Short-Term Memory Network With Bayesian Optimization Algorithm. *IEEE Access* **2020**, *9*, 1219–1232. [\[CrossRef\]](#)
29. Shekhar, S.; Bansode, A.; Salim, A. A Comparative study of Hyper-Parameter Optimization Tools. *arXiv* **2022**, arXiv:2201.06433v1.
30. Nguyen, H.; Bui, X.-N.; Bui, H.-B.; Cuong, D.T. Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: A case study. *Acta Geophys.* **2019**, *67*, 477–490. [\[CrossRef\]](#)
31. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [\[CrossRef\]](#)
32. Zhou, J.; Li, E.; Wang, M.; Chen, X.; Shi, X.; Jiang, L. Feasibility of Stochastic Gradient Boosting Approach for Evaluating Seismic Liquefaction Potential Based on SPT and CPT Case Histories. *J. Perform. Constr. Facil.* **2019**, *33*, 04019024. [\[CrossRef\]](#)
33. Chen, T.; He, T. *Xgboost: Extreme Gradient Boosting*, R Package Version 0.4-2. Available online: <https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf> (accessed on 11 March 2019).
34. Deng, M.; Tan, Z. Some issues during TBM trial advance of super-long tunnel group and development direction of construction technology. *Mod. Tunn. Technol.* **2019**, *56*, 1–12. (In Chinese)
35. Deng, M.; Tan, Z. Analysis of adaptability of TBM in trial boring stage of super-long tunnel. *Tunn. Constr.* **2019**, *39*, 1–22. (In Chinese)
36. Ye, X.Q.; Wu, Y.F. Cancer gene selection algorithm based on support vector machine recursive feature elimination and feature clustering. *J. Xiamen Univ. Nat. Sci.* **2018**, *57*, 702–707.
37. Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms and practice. *Neumcomputing* **2020**, *415*, 295–316. [\[CrossRef\]](#)
38. Wu, J.; Chen, X.; Zhang, H.; Xiong, L.; Lei, H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40.
39. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2623–2631.
40. Cui, X.; Shi, D.; Chen, Z.; Xu, F. parallel forestry text classification technology based on XGBoost in spark framework. *Trans. Chin. Soc. Agric. Mach.* **2019**, *50*, 280–287. (In Chinese)
41. Begum, A.M.; Mondal, M.R.H.; Podder, P.; Kamruzzaman, J. Weighted Rank Difference Ensemble: A New Form of Ensemble Feature Selection Method for Medical Datasets. *BioMedInformatics* **2024**, *4*, 477–488. [\[CrossRef\]](#)
42. Chatzilygeroudis, K.; Perikos, I.; Hatzilygeroudis, I. Machine Learning Basics. In *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice*; Eslambolchilar, P., Komninos, A., Dunlop, M., Eds.; ACM: New York, NY, USA, 2021; pp. 143–193.
43. Barulina, M.; Okunkov, S.; Ulitin, I.; Sanbaev, A. Sensitivity of Modern Deep Learning Neural Networks to Unbalanced Datasets in Multiclass Classification Problems. *Appl. Sci.* **2023**, *13*, 8614. [\[CrossRef\]](#)
44. Shaik, K.; Ramesh, J.V.N.; Mahdal, M.; Rahman, M.Z.U.; Khasim, S.; Kalita, K. Big Data Analytics Framework Using Squirrel Search Optimized Gradient Boosted Decision Tree for Heart Disease Diagnosis. *Appl. Sci.* **2023**, *13*, 5236. [\[CrossRef\]](#)
45. Saporetti, C.M.; da Fonseca, L.G.; Pereira, E. A Lithology Identification Approach Based on Machine Learning with Evolutionary Parameter Tuning. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1819–1823. [\[CrossRef\]](#)
46. Xie, Y.; Zhu, C.; Zhou, W.; Li, Z.; Liu, X.; Tu, M. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *J. Pet. Sci. Eng.* **2018**, *160*, 182–193. [\[CrossRef\]](#)
47. Wang, T.; Li, Q.; Yang, J.; Xie, T.; Wu, P.; Liang, J. Transformer Fault Diagnosis Method Based on Incomplete Data and TPE-XGBoost. *Appl. Sci.* **2023**, *13*, 7539. [\[CrossRef\]](#)
48. Lin, H.; Liu, X.; Han, Z.; Cui, H.; Dian, Y. Identification of Tree Species in Forest Communities at Different Altitudes Based on Multi-Source Aerial Remote Sensing Data. *Appl. Sci.* **2023**, *13*, 4911. [\[CrossRef\]](#)
49. Huang, I.-L.; Lee, M.-C.; Nieh, C.-Y.; Huang, J.-C. Ship Classification Based on AIS Data and Machine Learning Methods. *Electronics* **2024**, *13*, 98. [\[CrossRef\]](#)
50. Yang, Y.; Liu, G.; Zhang, H.; Zhang, Y.; Yang, X. Predicting the Compressive Strength of Environmentally Friendly Concrete Using Multiple Machine Learning Algorithms. *Buildings* **2024**, *14*, 190. [\[CrossRef\]](#)
51. Raji, I.D.; Bello-Salau, H.; Umoh, I.J.; Onumanyi, A.J.; Adegboye, M.A.; Salawudeen, A.T. Simple Deterministic Selection-Based Genetic Algorithm for Hyperparameter Tuning of Machine Learning Models. *Appl. Sci.* **2022**, *12*, 1186. [\[CrossRef\]](#)
52. Xu, Y.; Zhen, J.N.; Jiang, X.P.; Wang, J.J. Mangrove species classification with UAV-based remote sensing data and XGBoost. *J. Remote Sens.* **2021**, *25*, 737–752. [\[CrossRef\]](#)

53. Wang, Y.; Wang, J.; Chang, S.; Sun, L.; An, L.; Chen, Y.; Xu, J. Classification of Street Tree Species Using UAV Tilt Photogrammetry. *Remote Sens.* **2021**, *13*, 216. [[CrossRef](#)]
54. Sun, Z.; Jiang, B.; Li, X.; Li, J.; Xiao, K. A Data-Driven Approach for Lithology Identification Based on Parameter-Optimized Ensemble Learning. *Energies* **2020**, *13*, 3903. [[CrossRef](#)]
55. Chen, J.; Deng, X.; Shan, X.; Feng, Z.; Zhao, L.; Zong, X.; Feng, C. Intelligent Classification of Volcanic Rocks Based on Honey Badger Optimization Algorithm Enhanced Extreme Gradient Boosting Tree Model: A Case Study of Hongche Fault Zone in Junggar Basin. *Processes* **2024**, *12*, 285. [[CrossRef](#)]
56. Mehdary, A.; Chehri, A.; Jakimi, A.; Saadane, R. Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection. *Sensors* **2024**, *24*, 1230. [[CrossRef](#)]
57. Siers, M.J.; Md Zahidul, I. Class Imbalance and Cost-Sensitive Decision Trees: A Unified Survey Based on a Core Similarity. *ACM Trans. Knowl. Discov. Data* **2020**, *15*, 4. [[CrossRef](#)]
58. Rekha, G.; Tyagi, A.K.; Reddy, V.K. A Wide Scale Classification of Class Imbalance Problem and its Solutions: A Systematic Literature Review. *J. Comput. Sci.* **2019**, *15*, 886–929. [[CrossRef](#)]
59. Sayegh, H.R.; Dong, W.; Al-madani, A.M. Enhanced Intrusion Detection with LSTM-Based Model, Feature Selection, and SMOTE for Imbalanced Data. *Appl. Sci.* **2024**, *14*, 479. [[CrossRef](#)]
60. Swana, E.F.; Doorsamy, W.; Bokoro, P. Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors* **2022**, *22*, 3246. [[CrossRef](#)] [[PubMed](#)]
61. Ling, C.X.; Sheng, V.S. Cost-Sensitive Learning. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2011.
62. Song, C.; Li, X. Cost-Sensitive KNN Algorithm for Cancer Prediction Based on Entropy Analysis. *Entropy* **2022**, *24*, 253. [[CrossRef](#)] [[PubMed](#)]
63. Li, X.; Zheng, Z.; Dai, H. When services computing meets blockchain: Challenges and opportunities. *J. Parallel Distrib. Comput.* **2021**, *150*, 1–14. [[CrossRef](#)]
64. Xu, Z.; Shen, D.; Nie, T.; Kou, Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *J. Biomed. Inform.* **2020**, *107*, 103465. [[CrossRef](#)]
65. Muntasir Nishat, M.; Faisal, F.; Jahan Ratul, I.; Al-Monsur, A.; Ar-Rafi, A.M.; Nasrullah, S.M.; Khan, M.R.H. A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN over-sampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Sci. Program* **2022**, *2022*, 3649406.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.