mathematics

Article

Microseismic Data-Driven Short-Term Rockburst Evaluation in Underground Engineering with Strategic Data Augmentation and Extremely Randomized Forest

Shouye Cheng [1,2], Xin Yin [3,*], Feng Gao [1,2] and Yucong Pan [3]

[1] Research Institute of Mine Construction, Tiandi Science and Technology Company Limited, Beijing 100013, China
[2] State Key Laboratory of Intelligent Coal Mining and Strata Control, Beijing 100013, China
[3] School of Civil Engineering, Wuhan University, Wuhan 430072, China
[*] Correspondence: xinyin@whu.edu.cn

Abstract: Rockburst is a common dynamic geological disaster in underground mining and tunneling engineering, characterized by randomness, abruptness, and impact. Short-term evaluation of rockburst potential plays an outsize role in ensuring the safety of workers, equipment, and projects. As is well known, microseismic monitoring serves as a reliable short-term early-warning technique for rockburst. However, the large amount of microseismic data brings many challenges to traditional manual analysis, such as the timeliness of data processing and the accuracy of rockburst prediction. To this end, this study integrates artificial intelligence with microseismic monitoring. On the basis of a comprehensive consideration of class imbalance and multicollinearity, an innovative modeling framework that combines local outlier factor-guided synthetic minority oversampling and an extremely randomized forest with C5.0 decision trees is proposed for the short-term evaluation of rockburst potential. To determine the optimal hyperparameters, the whale optimization algorithm is embedded. To prove the efficacy of the model, a total of 93 rockburst cases are collected from various engineering projects. The results show that the proposed approach achieves an accuracy of 90.91% and a macro F_1-score of 0.9141. Additionally, the local F_1-scores on low-intensity and high-intensity rockburst are 0.9600 and 0.9474, respectively. Finally, the advantages of the proposed approach are further validated through an extended comparative analysis. The insights derived from this research provide a reference for microseismic data-based short-term rockburst prediction when faced with class imbalance and multicollinearity.

Keywords: underground engineering; rockburst prediction; microseismic monitoring; strategic data augmentation; extremely randomized forest

MSC: 74L10

Citation: Cheng, S.; Yin, X.; Gao, F.; Pan, Y. Microseismic Data-Driven Short-Term Rockburst Evaluation in Underground Engineering with Strategic Data Augmentation and Extremely Randomized Forest. *Mathematics* **2024**, *12*, 3502. https://doi.org/10.3390/math12223502

Academic Editor: Vladimir Balan

Received: 2 October 2024
Revised: 31 October 2024
Accepted: 8 November 2024
Published: 9 November 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

Rockburst is a typical dynamic geological disaster, frequently occurring in underground mining and tunneling engineering [1–4]. Rockburst is caused by the sharp release of elastic energy accumulated in rock mass due to the unloading effect and external disturbances, often accompanied by the violent ejection of rock fragments [5]. Rockburst brings serious threats to the safety of operators and equipment and also results in the deformation and even destruction of supporting structures and excavation systems. The world's first recorded rockburst occurred in 1738 at a tin mine in England [6], and, after that, rockburst has been reported in many countries. Between 1936 and 1993, the United States experienced 172 rockburst incidents, followed by 78 fatalities [7]. From the early 1980s to the mid2000s, Germany went through 42 recorded rockburst incidents with death and injuries [8]. In addition, there were also many rockburst incidents in Australia, Canada, South Africa, and other mining countries [9–11]. The occurrence of rockburst is closely related to the characteristics of minerals and rocks. In general, the higher the content of siliceous or other hard minerals, the more prone to rockburst. In terms of rock types, highly brittle rocks such as quartzite, granite, marble, and sandstone are prone to rockburst. Due to the randomness in space and the suddenness in time, rockburst prediction is faced with great challenges.

The microseismic monitoring technique is developed based on acoustic emissions and seismology [12–16]. When underground rocks fracture due to human or natural factors, seismic waves will be generated and spread around. By arranging multiple detectors in the three-dimensional space around the fracturing zone, the microseismic monitoring technique can realize the real-time in situ observation of the fracturing process. Starting from the initial deformation stage, the microseismic monitoring technique can capture the full failure process of rock mass, including fracture initiation, fracture propagation, and structure instability. This makes it so that the microseismic monitoring technique can detect the precursor to rockburst and provide an important basis for its early warning [17–20].

However, as a real-time monitoring technique, massive microseismic data bring great challenges to the timeliness and accuracy of rockburst early warning. To this end, many researchers have combined artificial intelligence with microseismic monitoring, that is, using machine learning to mine microseismic data and establish the mapping relationship with rockburst. Jin et al. [21] extracted six parameters from microseismic data, including cumulative number, cumulative energy, cumulative apparent volume, changing rate of cumulative number, changing rate of cumulative energy, and changing rate of cumulative apparent volume, and used a support vector machine to build the short-term rockburst prediction model. Similarly, based on these six microseismic parameters, Liang et al. [22] employed a gradient boosting decision tree and a random forest to build the short-term rockburst prediction model, and Feng et al. [23] adopted a probabilistic neural network to construct the short-term rockburst prediction model. Subsequently, Zhou et al. [24] and Liang et al. [25] added a new factor, namely incubation time, to the above parameter system and developed the ensemble learning-based short-term rockburst prediction model. In practical applications, these models have made indelible contributions. Class unbalance and multicollinearity are two common data defects, and their effects on model performance cannot be ignored [26–31]. Consequently, this study, on the basis of the comprehensive consideration of class imbalance and multicollinearity, proposes a novel hybrid intelligent modeling approach for short-term assessment of rockburst potential.

In this approach, the local outlier factor-guided synthetic minority oversampling technique and extremely randomized forest with C5.0 decision trees are proposed from the data and algorithm aspects, respectively. Additionally, the whale optimization algorithm is embedded to synergistically optimize the critical hyperparameters of the data strategy and algorithm strategy. The rest of the paper is organized as follows: Section 2 describes the proposed modeling framework in detail; Section 3 depicts the collected database; Section 4 gives the modeling results and conducts some profound discussion; Section 5 summarizes the main conclusions.

2. Proposed Modeling Framework

Class imbalance and multicollinearity are two common types of data defects in geotechnical and geological engineering. In order to remove their adverse effects on the prediction performance of machine learning models, improvements are made at the data level and algorithm level. At the data level, we propose local outlier factor-guided synthetic minority oversampling (LOF-SMO). At the algorithm level, we propose an extremely randomized forest with C5.0 decision trees (C5.0DT-ERF). To determine the optimal hyperparameters in the LOF-SMO and C5.0DT-ERF, the whale optimization algorithm (WOA) is merged. As a result, the WOA-LOF-SMO-C5.0DT-ERF modeling framework is devised (Figure 1).

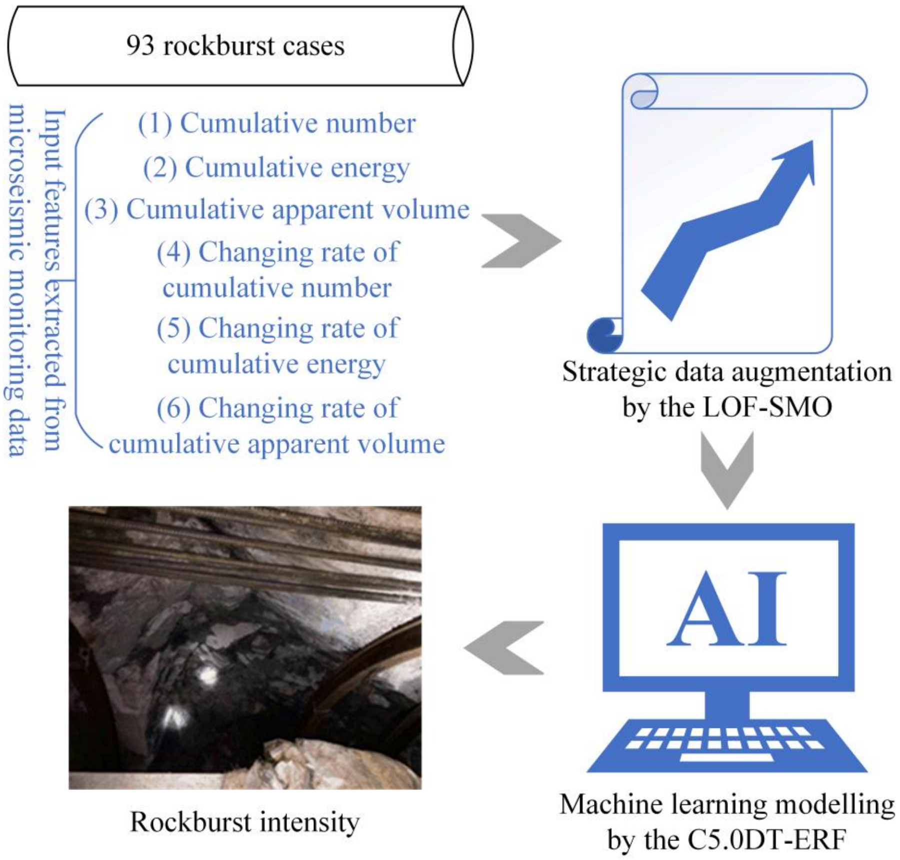

93 rockburst cases
Input features extracted from microseismic monitoring data
(1) Cumulative number
(2) Cumulative energy
(3) Cumulative apparent volume
(4) Changing rate of cumulative number
(5) Changing rate of cumulative energy
(6) Changing rate of cumulative apparent volume
Strategic data augmentation by the LOF-SMO
AI
Machine learning modelling by the C5.0DT-ERF
Rockburst intensity

Figure 1. Proposed modeling framework.

2.1. Local Outlier Factor-Guided Synthetic Minority Oversampling

2.1.1. Local Outlier Factor

Local outlier factor (LOF) is a critical indicator to distinguish outliers and non-outliers in the database [32]. An object is determined to be an outlier if its LOF exceeds the threshold of 1.5. In order to capture the dominant topology of the data in subsequent oversampling, synthetic minority oversampling is implemented in the database without outliers. The calculation principle of the LOF is described as follows:

(1) Determine the k-distance neighborhood $N_k(O)$ of the object O, consisting of the k nearest neighbors of O in the database D.

(2) Calculate the k-distance $dist_k(O)$ of O by:

$$dist_k(O) = \max\{dist(O,P)|P \in N_k(O)\} \tag{1}$$

where P is a neighbor from $N_k(O)$; $dist(O,P)$ is the Euclidean distance between O and P.

(3) Calculate the reachable distance $dist_{reach}(O,P)$ between O and P by:

$$dist_{reach}(O,P) = \max\{dist_k(O), dist(O,P)\} \tag{2}$$

(4) Calculate the local reachable density $lrd(O)$ of O by:

$$lrd(O) = \frac{k}{\sum_{P \in N_k(O)} dist_{reach}(O,P)} \tag{3}$$

(5) Calculate the LOF $lof(O)$ of O by:

$$lof(O) = \frac{\sum_{P \in N_k(O)} \frac{lrd(P)}{lrd(O)}}{k} \tag{4}$$

where $lrd(P)$ is the local reachable density of P.

2.1.2. Synthetic Minority Oversampling

Synthetic minority oversampling (SMO) is an oversampling method used to deal with class imbalance and increased sample size, especially when the number of minority samples is small [33]. Its basic principle is to generate new minority samples by linear interpolation, as shown in Figure 2. The detailed flow of the SMO is described as follows:

(1) Select minority samples: each minority sample in turn is selected as the root sample for synthesizing new samples.

(2) Find nearest neighbors: for each root sample, using Euclidean distance as the standard, its distance to all other minority samples is calculated to obtain its K nearest neighbors.

(3) Select auxiliary samples: one sample is randomly selected from the K nearest neighbors of each root sample as the auxiliary sample for synthesizing new samples.

(4) Synthesize new samples: between the root sample and the auxiliary sample, a new sample is generated by linear interpolation. The interpolation formula is expressed as:

$$x_i^{ne} = x_i^{ro} + \lambda_i \cdot (x_i^{au} - x_i^{ro}) \tag{5}$$

where x_i^{ne}, x_i^{ro} and x_i^{au} are the i-th feature of the new sample, the root sample, and the auxiliary sample, respectively; λ_i is a random number in [0, 1].

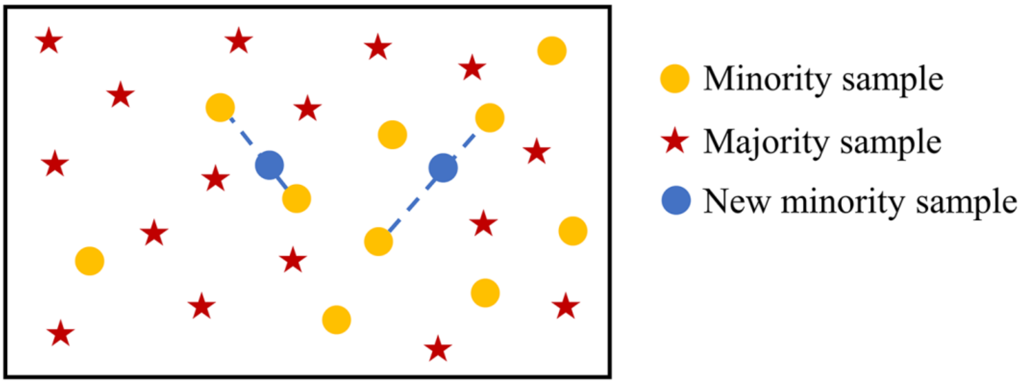


Figure 2. Basic principle of synthetic minority oversampling.

(5) Repeat generation: for each root sample, the above steps are repeated until the number of new samples meets the requirements.

(6) Add new samples: all generated new samples are added to the original dataset, thus increasing the number of minority samples and making the dataset more balanced.

2.2. Extremely Randomized Forest with C5.0 Decision Trees

2.2.1. C5.0 Decision Tree

A decision tree is a group of procedures designed to classify the input data into more homogenous subsets using generated rules [34]. In the training process, the decision tree aims to maximize the information gain and minimize the entropy in the generated subsets. Figure 3 demonstrates the basic components of a decision tree. Initially, all data are gathered in a root node and then split into relatively more homogenous subsets by internal nodes based on feature thresholds. The splitting process continues until the decision tree reaches the leaf nodes. At this final stage, labels are assigned to the leaf nodes.

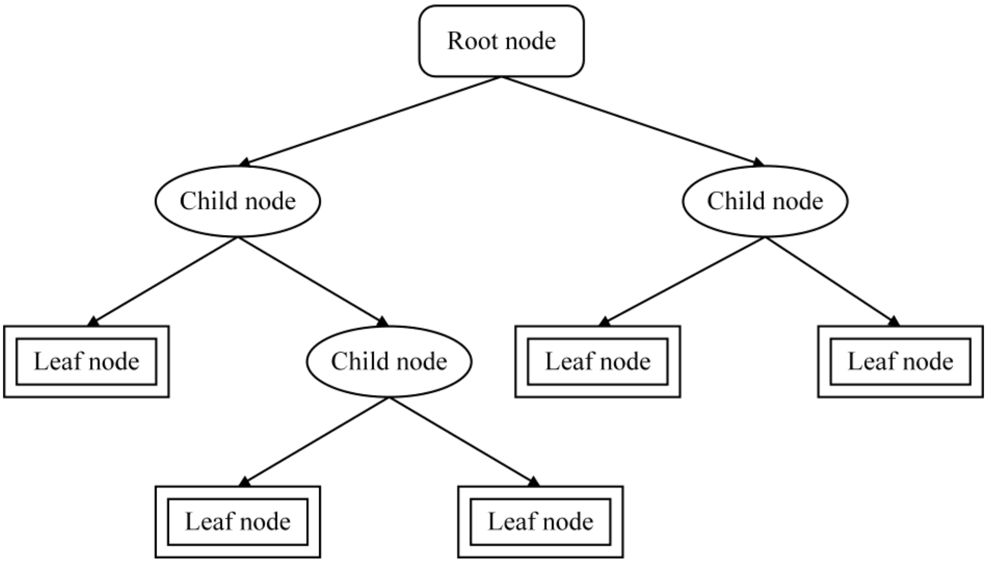


Figure 3. Topology of a decision tree.

A C5.0 decision tree (C5.0DT) [35] is a representative decision tree algorithm that uses the information gain ratio to select split attributes and thresholds in nodes and finish tree growth [36]. In this process, the decrease in the entropy is maximized, and consequently the purity of leaf nodes reaches the highest. Referring to Equation (6), the calculation principle of the information gain ratio is presented. In order to reduce the error when processing new data after tree growth, a decision tree must be pruned. In the C5.0DT, pruning automatically starts in reverse from the leaf node and extends upwards to the entire tree based on the information gain ratio [37].

$$Gain\_ratio(S, A) = \frac{Gain(S, A)}{-\sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}} \tag{6}$$

where $Gain\_ratio(S, A)$ is the information gain ratio obtained by splitting the node S using the attribute A; $Gain(S, A)$ is the information gain obtained by splitting the node S using the attribute A, calculated as [38]; v is the number of child nodes generated by splitting the node S; S_i is the i-th child node generated by splitting the node S; $|\ |$ is the number of samples contained in the node.

2.2.2. Extremely Randomized Forest

An extremely randomized forest (ERF) is made up of a series of standalone trees, as shown in Figure 4 [39]. Each tree generates the prediction results independently according to the input data, and then on this basis, the prediction results of the ERF are determined by voting on all trees, that is, majority rules. In this study, a C5.0 decision tree is used as the basic tree unit of the ERF for the subsequent analysis. The modeling process of the ERF is described as follows:

(1) Conduct bootstrap sampling: bootstrap sampling is performed on the balanced dataset of Section 2.1 so as to generate L subdatasets, where the number of subdatasets is the same as the number of tree models in the ERF.

(2) Build tree models: L tree models are built individually based on L subdatasets. To determine the appropriate split attribute and split value during tree growth, the information gain ratio is used as the metric.

(3) Increase the diversity: To increase the diversity of the tree models, the extremely randomized strategy is implemented during modeling. Specifically, for each tree model, the optimal split attribute is produced by winning from the attribute subset

randomly selected in all candidate attributes, and the split value is from the value subset randomly generated in the candidate range.

(4) Integrate tree models: to determine the output of the ERF, voting is carried out on the basis of the prediction results of all tree models.

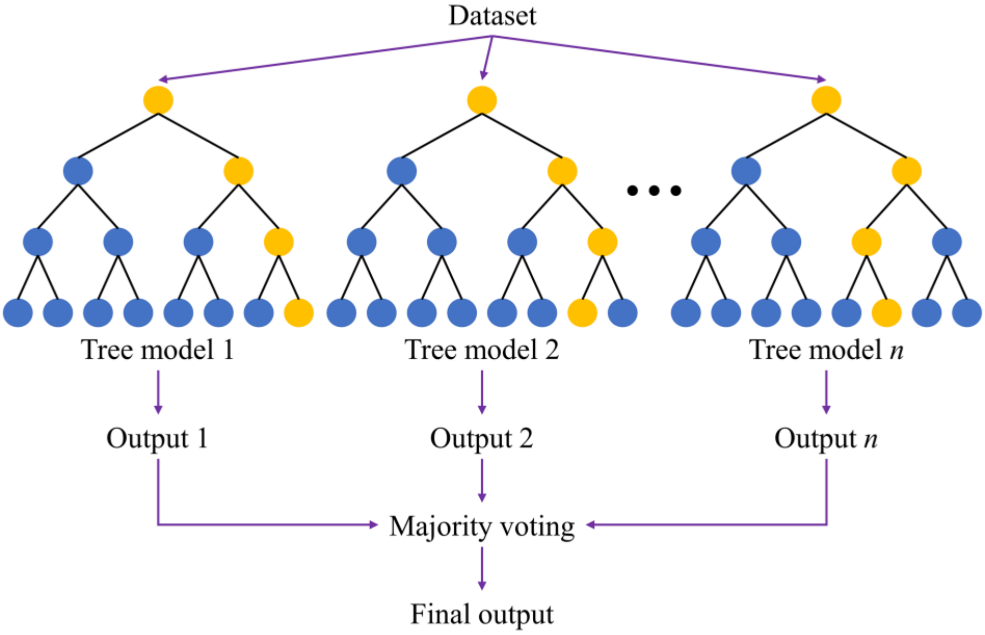


Figure 4. Topology of an extremely randomized forest.

2.3. Hyperparameter Optimization

The whale optimization algorithm (WOA) is a metaheuristic optimization algorithm that simulates the hunting behavior of humpback whales [40]. In the exploration phase, the mathematical model is described as:

$$X_i(t+1) = X_{\text{rand}}(t) - A|C \cdot X_{\text{rand}}(t) - X_i(t)| \tag{7}$$

where $X_i(t)$ and $X_i(t+1)$ are the position of the i-th whale at the t-th and $(t+1)$-th iterations, respectively; $X_{\text{rand}}(t)$ is the position of a whale randomly selected from the population; and A and C are changeable coefficients, defined as:

$$A = 2a \cdot r - a \tag{8}$$

$$C = 2r \tag{9}$$

where a decreases linearly from 2 to 0 with iteration; r is a random number in [0, 1].

In the exploitation phase, the mathematical model is described as:

$$X_i(t+1) = X_{\text{best}}(t) - A|C \cdot X_{\text{best}}(t) - X(t)| \quad \text{when } p < 0.5 \tag{10}$$

$$X_i(t+1) = e^{bl}|X_{\text{best}}(t) - X(t)|\cos(2\pi l) + X_{\text{best}}(t) \quad \text{when } p \geq 0.5 \tag{11}$$

where $X_{\text{best}}(t)$ is the optimal position searched by the population; b is the constant that defines the logarithmic spiral shape; l is the random number in $[-1, 1]$; p is a random number in [0, 1].

In order to balance the global and local search, the exploration and exploitation phases alternate, as demonstrated in Figure 5. In the proposed short-term evaluation framework for rockburst potential, there are five critical hyperparameters that need to be optimized, including the size of the k-distance neighborhood in the LOF (referred to as x_1), the number of used nearest neighbors in the SMO (referred to as x_2), the minimum number of split samples in the C5.0DT (referred to as x_3), the maximum number of leaf nodes in the

C5.0DT (referred to as x_4), and the number of C5.0 decision trees in the ERF (referred to as x_5). Therefore, the position of the whale is a five-dimensional vector, denoted as $\boldsymbol{X} = [x_1,\ x_2, x_3, x_4, x_5]$.

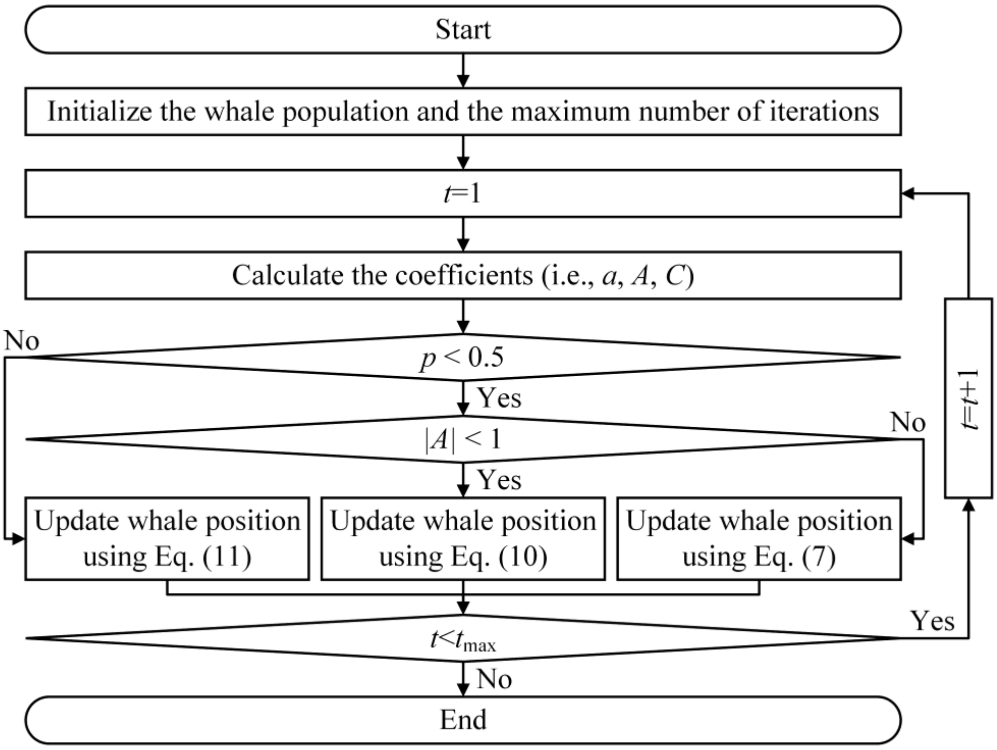

Start
Initialize the whale population and the maximum number of iterations
t=1
Calculate the coefficients (i.e., a, A, C)
No
$p < 0.5$
Yes
$|A| < 1$
No
Yes
Update whale position using Eq. (11)
Update whale position using Eq. (10)
Update whale position using Eq. (7)
t=t+1
t<$t_{\max}$
Yes
No
End

Figure 5. Hyperparameter optimization procedure.

3. Database Description

3.1. Case Collection

To demonstrate the validity of the proposed approach, 93 rockburst cases in total are collected from Feng et al. [41] (see Supplementary Material). According to Table 1, rockburst intensity is split into four grades: none, slight, moderate, and strong. Among these 93 rockburst cases, 34 cases (accounting for 36.56%) belong to none rockburst, 21 cases (accounting for 22.58%) belong to slight rockburst, 25 cases (accounting for 26.88%) belong to moderate rockburst, and 13 cases (accounting for 13.98%) belong to strong rockburst. As observed in Figure 6, the class imbalance is prominent.

Table 1. Classification criteria of rockburst intensity [22,41].

Rockburst Intensity	Failure Characteristics
None	No rockburst occurs. No abnormality in surrounding rock. Normal construction.
Slight	The depth of rockburst crater is <0.5 m. Slight spalling or slabbing. The size of ejected rock fragment is 10~30 cm. Slight cracking sound.
Moderate	The depth of rockburst crater is 0.5~1.0 m. Severe spalling and slabbing. The size of ejected rock fragment is 30~80 cm. Detonator blasting-like sound.
Strong	The depth of rockburst crater is >1.0m. Extensive spalling and slabbing. The size of ejected rock fragment is >80 cm. Explosion-like sound with an impact wave.

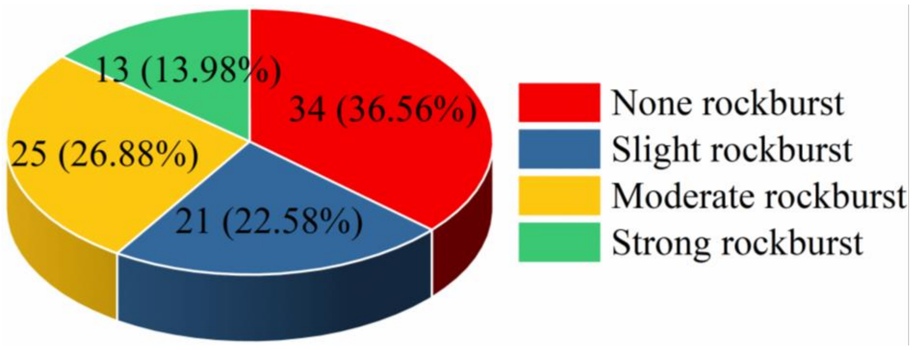


Figure 6. Proportion of different intensities of rockburst.

In order to predict rockburst intensity, six microseismic source parameters are used as the input parameters of the model, including cumulative number (unit), cumulative energy (J), cumulative apparent volume (m^3), changing rate of cumulative number (unit/day), changing rate of cumulative energy (J/day), and changing rate of cumulative apparent volume (m^3/day). During the rock fracturing process, the microseismic monitoring technique captures the stress wave emitted by the source in real time and records it as the microseismic event [42,43]. The cumulative number of microseismic events indicates the microseismic activity, the cumulative energy represents the microseismic strength, and the cumulative apparent volume reflects the damage degree of rock mass [24]. On the whole, these three microseismic source parameters jointly characterize the frequency, strength, and scale of rock fracture and are able to comprehensively describe the fracture situation inside the rock mass. Considering that rockburst incubation is a dynamic process, three microseismic source parameters with the time effect are also taken into account in the input system, namely the changing rate of cumulative number, changing rate of cumulative energy, and changing rate of cumulative apparent volume [21].

Specifically, the microseismic energy and apparent volume are calculated by Equations (12) and (13), respectively [21,44,45]. The time interval for the calculation of the cumulative value is the incubation time of rockburst [24]. Assuming that the incubation time is N days, the cumulative value C_v is the sum of the value c_n $(n = 1, 2, \cdots, N)$ of each day in these N days, calculated as Equation (14). Correspondingly, the changing rate C_r is the ratio of the cumulative value C_v to the incubation time N, calculated as Equation (15) [21]. Taking the strong rockburst that occurred at milestone SK8+709 of the Jinping II hydropower station on 11 January 2011 as an example, Figure 7 shows the evolution of the cumulative number of microseismic events from 6–10 January, where $C_v = 49$ and $C_r = 12.25$ according to Equations (14) and (15), respectively [41].

$$E = 8\pi\rho v \int_0^\infty s^2(f)df \tag{12}$$

$$V_A = \mu P^2/E \tag{13}$$

where E is the microseismic energy; V_A is the apparent volume; ρ is the density of rock mass; v is the wave velocity; $s^2(f)$ is the velocity power spectrum; f is the frequency; μ is the shear modulus of rock mass; P is the seismic potency.

$$C_v = c_1 + c_2 + \cdots + c_N \tag{14}$$

$$C_r = C_v/N \tag{15}$$

where C_v is the cumulative value; C_r is the changing rate; c_n $(n = 1, 2, \cdots, N)$ is the individual value of each day during rockburst incubation; N is the incubation time of rockburst.

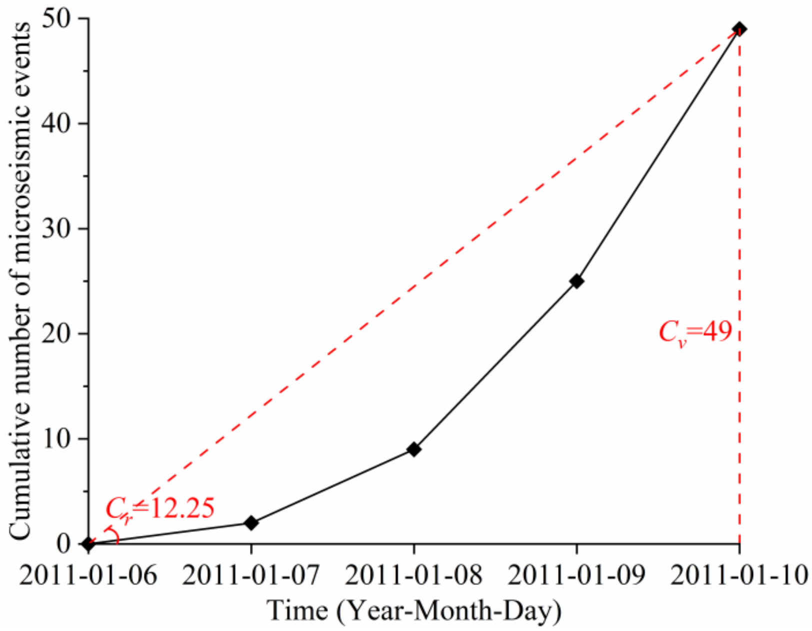


Figure 7. Evolution of cumulative number of microseismic events (taking strong rockburst occurring at milestone SK8+709 of Jinping II hydropower station on 11 January 2011 as example) [41].

Table 2 shows the statistical description of input parameters for various rockburst intensities, and the visual distribution of input parameters is presented in Figure 8.

Table 2. Statistical description of input parameters.

Rockburst Intensity	Input Parameter	Statistical Index						
		Number	**Minimum**	**Maximum**	**Mean**	**Median**	**Skewness**	**Kurtosis**
None	*CN*	34	1	17	3.94	3	2.04	5.22
	CE	34	0.78	5.82	3.16	3.55	−0.14	−1.30
	CAV	34	2.51	4.86	3.62	3.58	0.27	−0.87
	CNR	34	0.11	2.50	0.85	0.76	1.21	0.94
	CER	34	0.18	4.78	2.51	2.81	−0.14	−1.20
	CAVR	34	1.67	4.31	2.97	2.96	−0.19	−0.40
Slight	*CN*	21	3	29	10.14	8	1.48	2.02
	CE	21	3.54	5.56	4.54	4.53	0.05	−0.13
	CAV	21	3.50	4.94	4.18	4.13	0.12	−0.57
	CNR	21	0.54	4.00	1.48	1.11	1.29	1.23
	CER	21	2.84	4.80	3.71	3.67	0.46	−0.36
	CAVR	21	2.39	3.99	3.35	3.50	−0.64	−0.54
Moderate	*CN*	25	3	36	15.12	14	0.80	1.61
	CE	25	3.54	5.98	5.13	5.10	−0.93	0.96
	CAV	25	3.52	4.87	4.48	4.57	−1.41	3.32
	CNR	25	0.43	4.00	1.70	1.71	0.88	1.92
	CER	25	2.29	5.08	4.12	4.25	−1.01	0.60
	CAVR	25	2.67	4.02	3.51	3.55	−0.58	0.61

Table 2. *Cont.*

Rockburst Intensity	Input Parameter	Statistical Index						
		Number	Minimum	Maximum	Mean	Median	Skewness	Kurtosis
Strong	*CN*	13	10	70	37.31	42	−0.09	−0.80
	CE	13	4.11	7.09	5.94	6.15	−1.02	1.02
	CAV	13	3.62	5.17	4.87	4.98	−2.76	8.72
	CNR	13	1.25	12.25	4.53	3.73	1.48	2.88
	CER	13	3.41	5.89	5.01	5.15	−0.92	−0.03
	CAVR	13	2.93	4.39	3.94	4.08	−1.52	3.07

Note: The *CN*, *CE*, *CAV*, *CNR*, *CER*, and *CAVR* denote cumulative number, cumulative energy, cumulative apparent volume, changing rate of cumulative number, changing rate of cumulative energy, and changing rate of cumulative apparent volume, respectively. Particularly, the values of the *CE*, *CAV*, *CER,* and *CAVR* are expressed in logarithmic form with base 10.

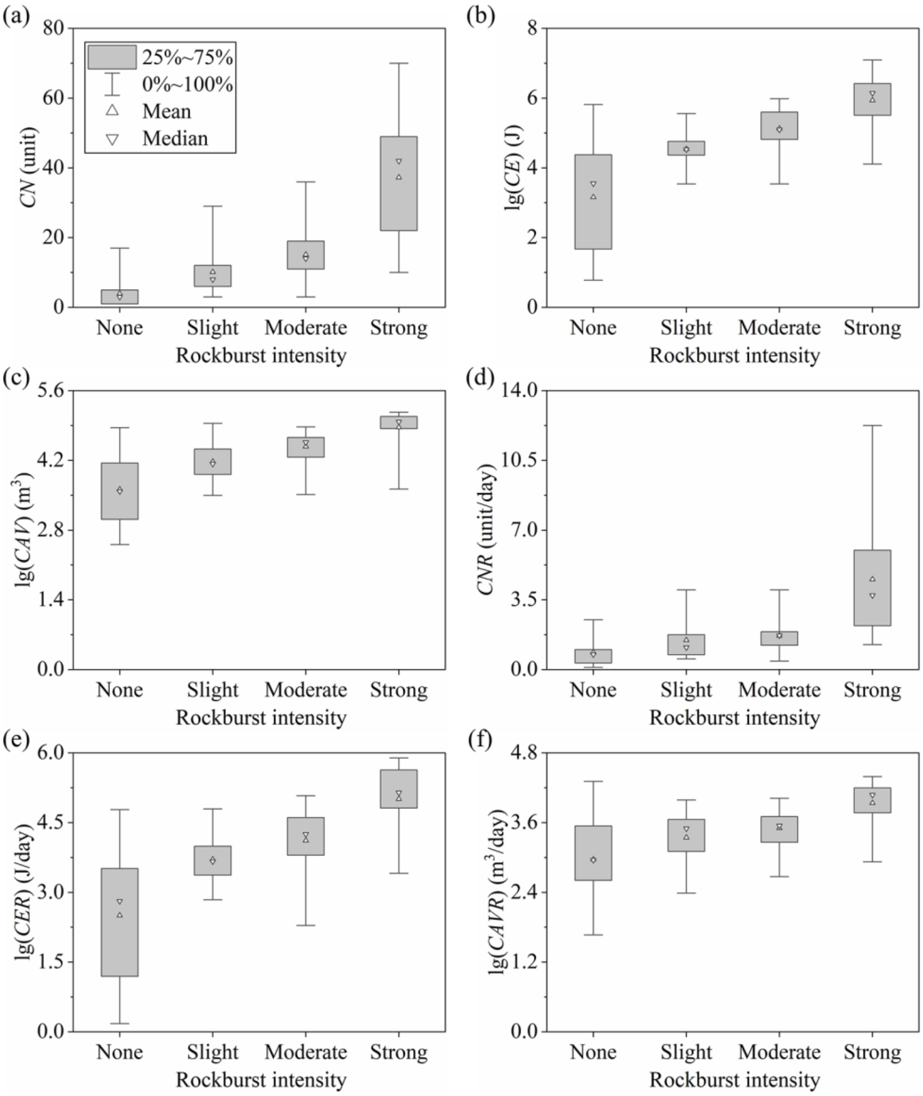


Figure 8. Visual distribution of input parameters: (**a**) cumulative number; (**b**) cumulative energy; (**c**) cumulative apparent volume; (**d**) changing rate of cumulative number; (**e**) changing rate of cumulative energy; (**f**) changing rate of cumulative apparent volume. Particularly, the values of cumulative energy, cumulative apparent volume, changing rate of cumulative energy, and changing rate of cumulative apparent volume are expressed in logarithmic form with base 10.

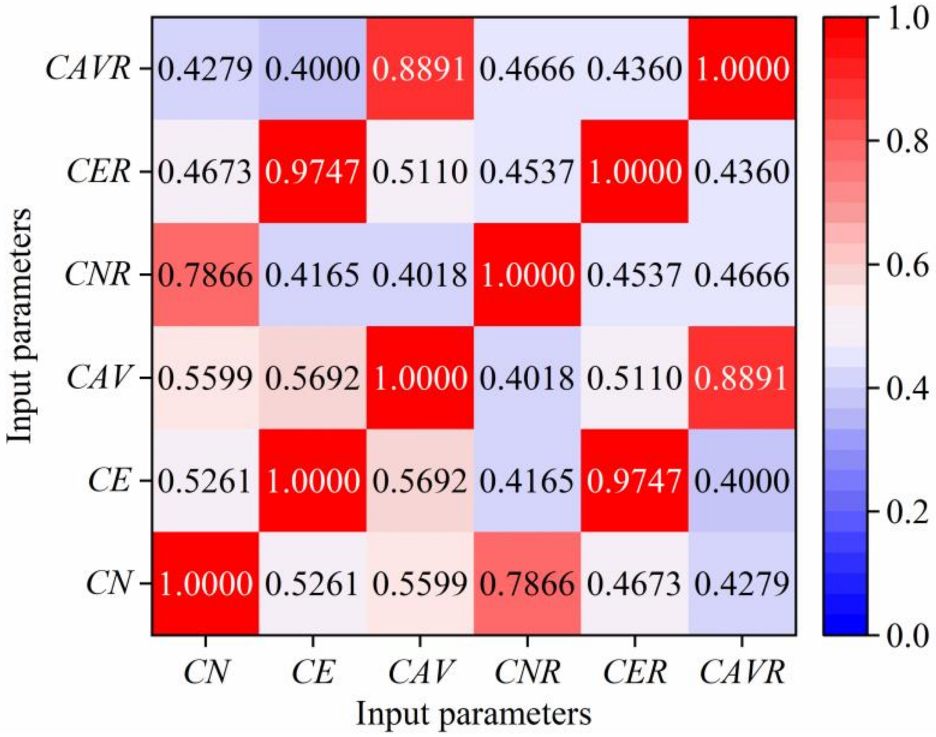
3.2. Correlation Analysis

Generally, input parameters should be independent of each other, and the stronger the parameter correlation, the higher the information redundancy. The Pearson correlation coefficient is adopted to analyze the correlation of the six input parameters selected in Section 3.1, calculated by Equation (16) [46]. Based on the magnitude of the Pearson correlation coefficient, the correlation strength is divided into five levels, among which 0.0~0.2, 0.2~0.4, 0.4~0.6, 0.6~0.8, and 0.8~1.0 indicate extremely weak, weak, moderate, strong, and extremely strong correlation, respectively. In particular, a positive Pearson correlation coefficient represents positive correlation, and a negative Pearson correlation coefficient stands for negative correlation.

$$\rho_{X,Y} = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2 \sum\left(Y - \overline{Y}\right)^2}} \tag{16}$$

where $\rho_{X,Y}$ is the Pearson correlation coefficient between input parameters X and Y; $\overline{X}$ and $\overline{Y}$ are the mean of input parameters X and Y, respectively.

Referring to Figure 9, there exists a strong correlation between cumulative number and its changing rate, with the Pearson correlation coefficient reaching 0.7866. Similarly, a strong correlation can also be found between cumulative energy and its changing rate, as well as cumulative apparent volume and its changing rate, with the Pearson correlation coefficient reaching 0.9747 and 0.8891, respectively. Regarding other parameter pairs, no obvious correlation appears. In essence, cumulative number, cumulative energy, and cumulative apparent volume are static indicators, while their changing rate, as a dynamic indicator, reflects the evolution trend of static indicators during rockburst development. Therefore, static and dynamic indicators have an inevitable physical correlation. Since both static and dynamic metrics have clear physical meaning, no additional measures, such as dimensionality reduction, are taken to eliminate correlations herein. Moreover, the dimension of the current input combination is 6, which does not cause a dimensional disaster in modeling.



	CN	*CE*	*CAV*	*CNR*	*CER*	*CAVR*
CAVR	0.4279	0.4000	0.8891	0.4666	0.4360	1.0000
CER	0.4673	0.9747	0.5110	0.4537	1.0000	0.4360
CNR	0.7866	0.4165	0.4018	1.0000	0.4537	0.4666
CAV	0.5599	0.5692	1.0000	0.4018	0.5110	0.8891
CE	0.5261	1.0000	0.5692	0.4165	0.9747	0.4000
CN	1.0000	0.5261	0.5599	0.7866	0.4673	0.4279

Figure 9. Calculation results of Pearson correlation coefficient.

3.3. Multicollinearity Analysis

When there is multicollinearity between input parameters, small changes in the data can lead to significant changes in the parameter estimates, resulting in a decrease in the stability of the model. Variance inflation factor (VIF) is a common measure of multicollinearity severity, and VIF greater than 10 is generally considered to indicate strong multicollinearity [47]. VIF is calculated by:

$$VIF_i = \frac{1}{1 - {R_i}^2} \tag{17}$$

where R_i is the determination coefficient of the linear fit of the input parameter X_i with respect to other input parameters.

Referring to Figure 10, cumulative energy, cumulative apparent volume, changing rate of cumulative energy, and changing rate of cumulative apparent volume behave with strong multicollinearity, with VIF reaching 153, 35, 130, and 29, respectively. And, cumulative number and changing rate of cumulative number behave with weak multicollinearity, with VIF reaching 6 and 5, respectively.

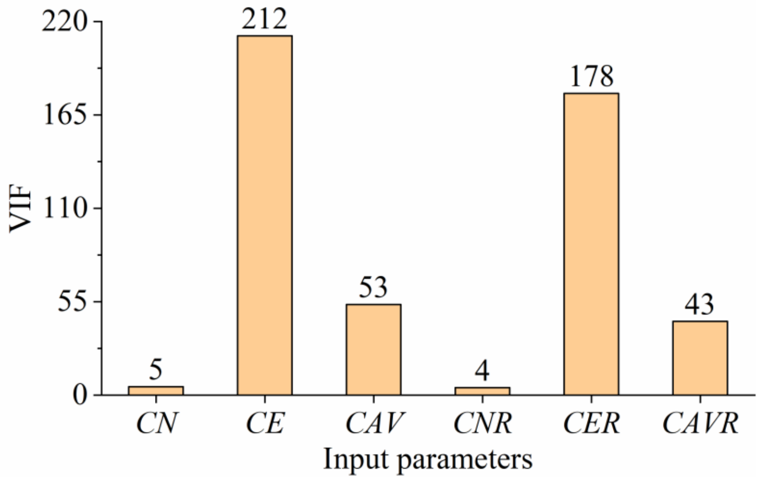


Figure 10. Calculation results of variance inflation factors.

4. Results and Discussion

4.1. Evaluation Metrics

In order to quantitatively evaluate the generalization performance of the proposed model, two evaluation metrics are selected, including the accuracy and macro F_1-score. The accuracy is calculated by:

$$\text{accuracy} = \frac{\sum_{k=1}^{K} u_{k,k}}{\sum_{i=1}^{K}\sum_{j=1}^{K} u_{i,j}} \tag{18}$$

where K is the number of sample classes; $u_{i,j}$ is the number of samples belonging to the i-th class but predicted as the j-th class.

The macro F_1-score is defined as the mean of the F_1-score in each class, calculated by Equation (19). Unlike the accuracy, the macro F_1-score comprehensively considers the local generalization performance of the model in each class.

$$\text{macro } F_1 - \text{score} = \frac{1}{K}\sum_{k=1}^{K} F_1 - \text{score}_k \tag{19}$$

where $F_1 - \text{score}_k$ is the F_1-score in the k-th class, calculated by:

$$F_1 - \text{score}_k = \frac{2 \cdot Pre_k \cdot Rec_k}{Pre_k + Rec_k} \tag{20}$$

where Pre_k and Rec_k are the precision and recall in the k-th class, respectively, calculated by:

$$Pre_k = \frac{u_{k,k}}{\sum_{i=1}^{K} u_{i,k}} \tag{21}$$

$$Rec_k = \frac{u_{k,k}}{\sum_{j=1}^{K} u_{k,j}} \tag{22}$$

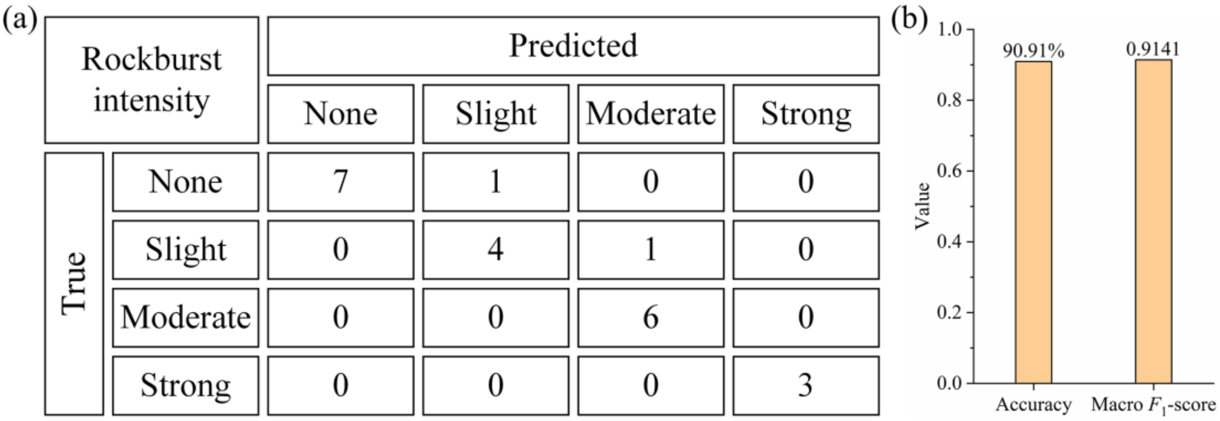
4.2. Evaluation Results

Out of 93 collected rockburst cases, 71 cases (accounting for 75%) were used as the training set to construct the model, and the rest 22 cases (accounting for 25%) were used as the test set to evaluate the model. In order to ensure the representativeness of the training set and the test set, stratified sampling was used to split the database. After that, the LOF-SMO was performed on the training set, as described in Section 2.1. On the one hand, class imbalance was eliminated. On the other hand, the number of training samples was increased from 71 to 104. Based on the balanced and extended training set, the C5.0DT-ERF was established for subsequent rockburst prediction, as described in Section 2.2. Particularly, the data were standardized by Min-Max standardization before modeling, described in Equation (23). In order to determine the optimal hyperparameters in the LOF-SMO and C5.0DT-ERF, including the size of the k-distance neighborhood in the LOF, the number of used nearest neighbors in the SMO, the minimum number of split samples in the C5.0DT, the maximum number of leaf nodes in the C5.0DT, and the number of C5.0 decision trees in the ERF, the WOA was implemented combined with a 10-fold cross-validation on the augmented training set, more technical information of which can be found in Section 2.3.

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{23}$$

where x^* is the standardized value; x is the original value; $x_{\max}$ and $x_{\min}$ is the maximum and minimum, respectively.

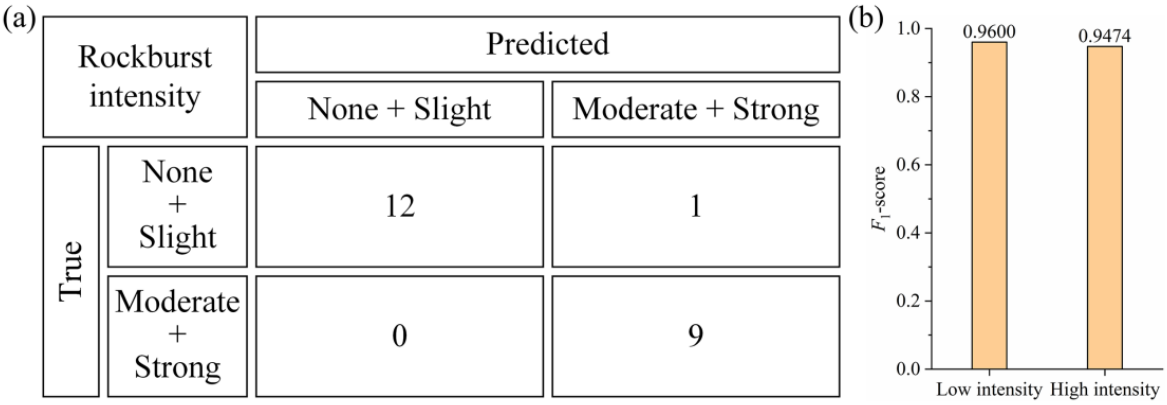
Figure 11a shows the confusion matrix on the test set, which describes the predictive behavior of the model in detail. For six moderate rockburst cases and three strong rockburst cases, there were no wrong predictions. For eight none rockburst cases and five slight rockburst cases, one out of them was wrongly predicted in each. Based on the confusion matrix, the accuracy and macro F_1-score of the model were calculated by Equations (18) and (19), reaching 90.91% and 0.9141, respectively (Figure 11b).



Rockburst intensity		Predicted			
		None	Slight	Moderate	Strong
True	None	7	1	0	0
	Slight	0	4	1	0
	Moderate	0	0	6	0
	Strong	0	0	0	3

Figure 11. Global performance analysis: (**a**) confusion matrix; (**b**) accuracy and macro F_1-score.

Further, the local generalization performance of the model on low-intensity (none + slight) and high-intensity (moderate + strong) rockburst was dissected, as illustrated in Figure 12. For the former, the model achieved the F_1-score of 0.9600. For the latter, the model achieved the F_1-score of 0.9474. The results show that the model has a high prediction reliability for both low-intensity and high-intensity rockburst, with an F_1-score of more than 0.9.



Rockburst intensity		Predicted	
		None + Slight	Moderate + Strong
True	None + Slight	12	1
	Moderate + Strong	0	9

Figure 12. Local performance analysis: (**a**) confusion matrix; (**b**) F_1-score.

In addition, referring to Figures 11a and 12a, misjudged samples are generally from the low intensity category and predicted as high intensity. Compared to low-intensity rockburst, the model has superior sensitivity to high-intensity rockburst. From the perspective of engineering safety and risk management, this misjudgment is conducive to rockburst prevention, but at the same time, it may also cause an unnecessary increase in prevention costs.

4.3. Comparative Analysis

For ease of expression, the model proposed in this study is denoted as the LOF-SMO-C5.0DT-ERF. As the data-level improvement strategy, the LOF-SMO aims to eliminate class imbalance and overcome multicollinearity by increasing the data size. To probe the effectiveness of the LOF-SMO, a comparative analysis was conducted with the single C5.0DT-ERF, in which the LOF-SMO was not integrated. Referring to Figure 13, it can be found that the accuracy of the LOF-SMO-C5.0DT-ERF increased by 18.18% compared with the single C5.0DT-ERF, and, meanwhile, the macro F_1-score improved by 0.2105.

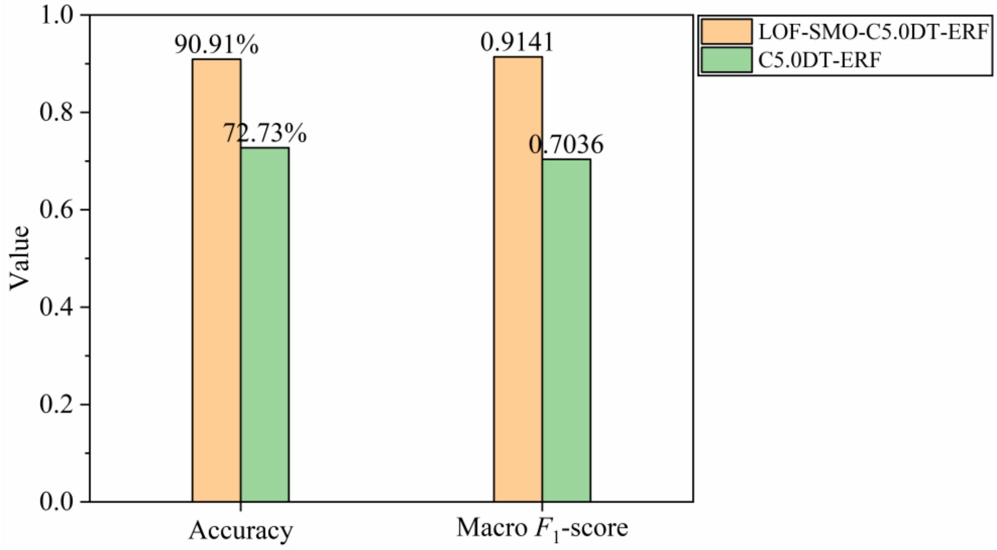


Figure 13. Comparative analysis between the LOF-SMO-C5.0DT-ERF and single C5.0DT-ERF.

As the algorithm-level improvement strategy, the C5.0DT-ERF aggregates multiple standalone C5.0 decision trees through the extremely randomized strategy to accomplish the complicated prediction task. To verify the strength of the C5.0DT-ERF, a comparative analysis between the LOF-SMO-C5.0DT-ERF, LOF-SMO-MLP (multi-layer perceptron), and LOF-SMO--SVM (support vector machine) was conducted, as presented in Figure 14. Particularly for the MLP, a hidden layer was set, and the optimized hyperparameters included the number of neurons in the hidden layer and learning rate. And, for the SVM, the radial basis function was used, and the optimized hyperparameters included the penalty factor and kernel function width. Compared to the LOF-SMO-MLP and LOF-SMO-SVM, the LOF-SMO-C5.0DT-ERF showed an increase of 22.73% and 18.18% in accuracy, respectively. Regarding the macro F_1-score, the LOF-SMO-C5.0DT-ERF achieved an increase of 0.2442 and 0.2082, respectively.

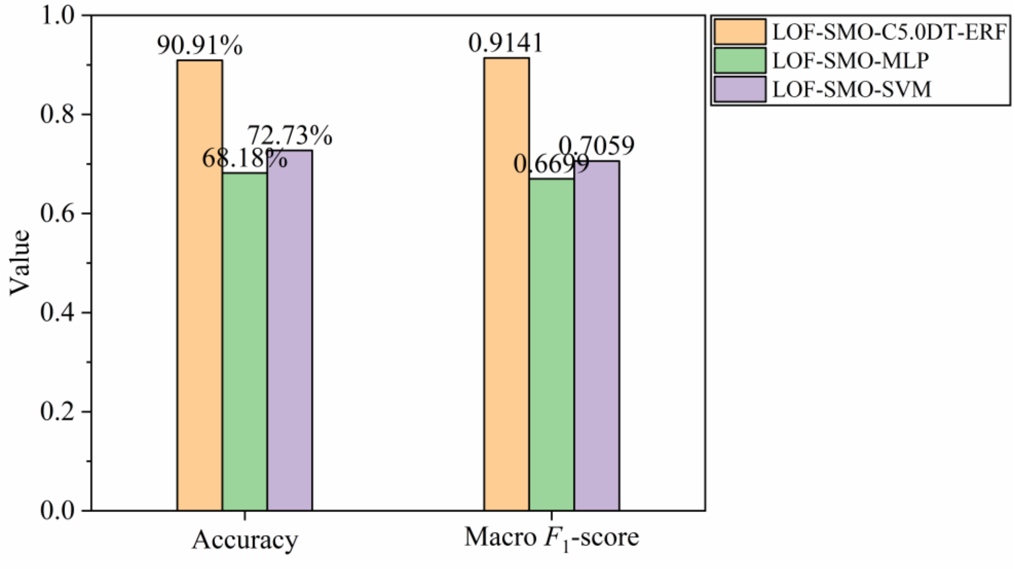


Figure 14. Comparative analysis between the LOF-SMO-C5.0DT-ERF, LOF-SMO-MLP, and LOF-SMO-SVM.

4.4. Sensitivity Analysis

In this section, we analyzed the sensitivity of the model performance for each input parameter. This was achieved by removing each input parameter individually from the input system and then observing the change in model performance, as described in Equation (24). The more significant the performance change, the more sensitive the input parameter. The results showed that the LOF-SMO-C5.0DT-ERF is the most sensitive to cumulative energy, and the accuracy of the model decreased by 13.64% when cumulative energy was removed, followed by cumulative number, cumulative apparent volume, and changing rate of cumulative energy, which caused a 9.09% decrease in accuracy when they were removed. In contrast, the least sensitive input parameters are changing rate of cumulative number and changing rate of cumulative apparent volume, which each reduced the model's accuracy by 4.55% when removed. Figure 15 summarizes the results of the sensitivity analysis.

$$\Delta Acc_u = Acc_S - Acc_{u \notin S} \tag{24}$$

where ΔAcc_u is the change in model accuracy when the input parameter u is removed from the input system; Acc_S is the accuracy of the model using the entire input system; $Acc_{u \notin S}$ is the accuracy of the model using the input system without the input parameter u.

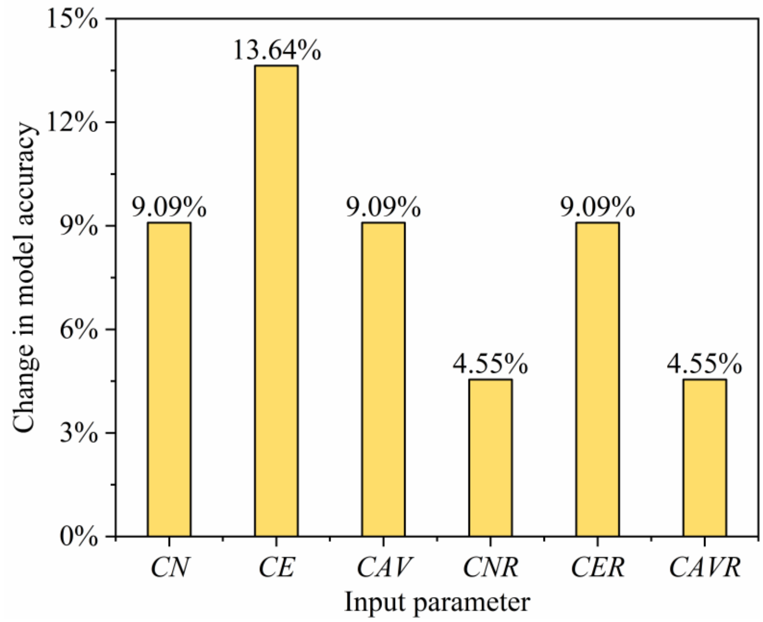


Figure 15. Sensitivity analysis results of input parameters.

5. Conclusions

Rockburst is a common dynamic geological disaster in underground engineering that seriously threatens the safety of workers and equipment and leads to damage to the excavation. As an in situ real-time monitoring technology, microseismic monitoring is widely used to analyze the fracturing behavior of rock mass. This study aims to utilize machine learning to establish the quantitative mapping relationship between microseismic parameters and rockburst intensity so as to achieve the short-term evaluation of rockburst potential.

Class imbalance and multicollinearity are two common types of data defects in geotechnical and geological engineering. To eliminate their adverse effects on rockburst prediction, this study makes some improvements from two levels of data and algorithm. Correspondingly, a WOA-LOF-SMO-C5.0DT-ERF modeling framework is proposed. The results indicate that this model achieves an accuracy of 90.91% and a macro F_1-score of 0.9141. Additionally, the local F_1-scores on low-intensity and high-intensity rockburst are 0.9600 and 0.9474, respectively. Finally, through a comparative analysis with the single C5.0DT-ERF, LOF-SMO-MLP, and LOF-SMO-SVM models, the advantages of the proposed model were validated.

However, this study still has some limitations. First, the size of the used database is limited. It is well known that the generalization performance of machine learning models is closely related to database size. In general, the larger the database size, the better the generalization performance. Second, there is a reverse data structure in the used database. Normally, non-rockburst cases are more than rockburst cases, but here non-rockburst cases are fewer than rockburst cases. This may result in the application of the model producing more false predictions of non-rockburst cases. Third, this study only utilizes microseismic data to predict rockburst, neglecting the geological data (e.g., rock types, stress level, and the orientation of the fault and structural plane) and construction data (e.g., support conditions and the distance to the working face). Fourth, the rockburst type discussed here is dominated by stress rockburst, and there is a lack of in-depth analysis of fault rockburst and stroke rockburst. In future research, we will pay more attention to the establishment of the comprehensive database from the above four aspects, so as to further strengthen the performance of the model by improving the database quality.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math12223502/s1, Table S1: Supplementary Material-Rockburst database.

Author Contributions: Conceptualization, S.C. and X.Y.; Data curation, X.Y.; Funding acquisition, X.Y., F.G. and Y.P.; Investigation, S.C., X.Y., F.G. and Y.P.; Methodology, S.C., X.Y., F.G. and Y.P.; Supervision, X.Y.; Validation, S.C. and X.Y.; Writing—original draft, S.C. and X.Y.; Writing—review & editing, S.C., X.Y., F.G. and Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant Nos. 42472351, 42177140, and 52404127), the Natural Science Foundation of Hubei Province (Grant No. 2024AFD359), and the Young Elite Scientist Sponsorship Program by CAST (Grant No. YESS20230742).

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/s.

Conflicts of Interest: Authors Shouye Cheng and Feng Gao were employed by the company Tiandi Science and Technology Company Limited. 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.

References

1. Zhou, J.; Li, X.B.; Mitri, H.S. Evaluation method of rockburst: State-of-the-art literature review. *Tunn. Undergr. Space Technol.* **2018**, *81*, 632–659. [CrossRef]
2. Yin, X.; Cheng, S.; Yu, H.; Pan, Y.; Liu, Q.; Huang, X.; Gao, F.; Jing, G. Probabilistic assessment of rockburst risk in TBM-excavated tunnels with multi-source data fusion. *Tunn. Undergr. Space Technol.* **2024**, *152*, 105915. [CrossRef]
3. Zhou, S.T.; Zhang, Z.X.; Luo, X.D.; Niu, S.S.; Jiang, N.; Yao, Y.K. Developing a hybrid CEEMDAN-PE-HE-SWT method to remove the noise of measured carbon dioxide blast wave. *Measurement* **2023**, *223*, 113797. [CrossRef]
4. Liu, Q.S.; Lei, Y.M.; Yin, X.; Lei, J.S.; Pan, Y.C.; Sun, L. Development and application of a novel probabilistic back-analysis framework for geotechnical parameters in shield tunneling based on the surrogate model and Bayesian theory. *Acta Geotech.* **2023**, *18*, 4899–4921. [CrossRef]
5. Askaripour, M.; Saeidi, A.; Rouleau, A.; Mercier-Langevin, P. Rockburst in underground excavations: A review of mechanism, classification, and prediction methods. *Undergr. Space* **2022**, *7*, 577–607. [CrossRef]
6. Gong, F.; Dai, J.; Xu, L. A strength-stress coupling criterion for rockburst: Inspirations from 1114 rockburst cases in 197 underground rock projects. *Tunn. Undergr. Space Technol.* **2023**, *142*, 105396. [CrossRef]
7. Mark, C. Coal bursts in the deep longwall mines of the United States. *Int. J. Coal Sci. Technol.* **2016**, *3*, 1–9. [CrossRef]
8. Baltz, R.; Hucke, A. Rockburst prevention in the German coal industry. In Proceedings of the 27th International Conference on Ground Control in Mining, Morgantown, WV, USA, 29–31 July 2008; pp. 46–50.
9. Potvin, Y.; Hudyma, M.; Jewell, R.J. Rockburst and seismic activity in underground Australian mines-an introduction to a new research project. In Proceedings of the ISRM International Symposium, Melbourne, Australia, 19–24 November 2000; p. ISRM-IS-2000-2552.
10. Webber, S.J. Rockburst risk assessment on south african gold mines: An expert system approach. In Proceedings of the ISRM International Symposium, Turin, Italy, 2–5 September 1996.
11. Adushkin, V.V.; Lovchikov, A.V.; Goev, A.G. The Occurrence of a Catastrophic Rockburst at the Umbozero Mine in the Lovozero Massif, Central Part of the Kola Peninsula. *Dokl. Earth Sci.* **2022**, *504*, 305–309. [CrossRef]
12. Zhang, S.C.; Tang, C.A.; Wang, Y.C.; Li, J.M.; Ma, T.H.; Wang, K.K. Review on Early Warning Methods for Rockbursts in Tunnel Engineering Based on Microseismic Monitoring. *Appl. Sci.* **2021**, *11*, 10965. [CrossRef]
13. Zhang, B.H.; Deng, J.H. Microseismic Monitoring Analysis Methods for Disaster Prevention in Underground Engineering. *Disaster Adv.* **2012**, *5*, 1420–1424.
14. Ma, T.H.; Tang, C.A.; Tang, S.B.; Kuang, L.; Yu, Q.; Kong, D.Q.; Zhu, X. Rockburst mechanism and prediction based on microseismic monitoring. *Int. J. Rock Mech. Min. Sci.* **2018**, *110*, 177–188. [CrossRef]
15. Xu, N.W.; Tang, C.A.; Li, L.C.; Zhou, Z.; Sha, C.; Liang, Z.Z.; Yang, J.Y. Microseismic monitoring and stability analysis of the left bank slope in Jinping first stage hydropower station in southwestern China. *Int. J. Rock Mech. Min. Sci.* **2011**, *48*, 950–963. [CrossRef]
16. Yin, X.; Liu, Q.; Lei, J.; Pan, Y.; Huang, X.; Lei, Y. Hybrid deep learning-based identification of microseismic events in TBM tunnelling. *Measurement* **2024**, *238*, 115381. [CrossRef]
17. Tang, C.A.; Wang, J.M.; Zhang, J.J. Preliminary engineering application of microseismic monitoring technique to rockburst prediction in tunneling of Jinping II project. *J. Rock Mech. Geotech. Eng.* **2010**, *2*, 193–208. [CrossRef]

18. Yin, X.; Liu, Q.; Huang, X.; Pan, Y. Real-time prediction of rockburst intensity using an integrated CNN-Adam-BO algorithm based on microseismic data and its engineering application. *Tunn. Undergr. Space Technol.* **2021**, *117*, 104133. [CrossRef]
19. Srinivasan, C.; Arora, S.K.; Benady, S. Precursory monitoring of impending rockbursts in Kolar gold mines from microseismic emissions at deeper levels. *Int. J. Rock Mech. Min. Sci.* **1999**, *36*, 941–948. [CrossRef]
20. Wang, C.; Zhan, K.; Zheng, X.; Liu, C.; Kong, C. A Method for Evaluating the Data Integrity of Microseismic Monitoring Systems in Mines Based on a Gradient Boosting Algorithm. *Mathematics* **2024**, *12*, 1902. [CrossRef]
21. Jin, A.; Basnet, P.M.S.; Mahtab, S. Microseismicity-based short-term rockburst prediction using non-linear support vector machine. *Acta Geophys.* **2022**, *70*, 1717–1736. [CrossRef]
22. Liang, W.; Sari, A.; Zhao, G.; McKinnon, S.D.; Wu, H. Short-term rockburst risk prediction using ensemble learning methods. *Nat. Hazards* **2020**, *104*, 1923–1946. [CrossRef]
23. Feng, G.L.; Xia, G.Q.; Chen, B.R.; Xiao, Y.X.; Zhou, R.C. A method for rockburst prediction in the deep tunnels of hydropower stations based on the monitored microseismicity and an optimized probabilistic neural network model. *Sustainability* **2019**, *11*, 3212. [CrossRef]
24. Qiu, Y.; Zhou, J. Short-term rockburst prediction in underground project: Insights from an explainable and interpretable ensemble learning model. *Acta Geotech.* **2023**, *18*, 6655–6685. [CrossRef]
25. Liang, W.; Sari, Y.A.; Zhao, G.; McKinnon, S.D.; Wu, H. Probability Estimates of Short-Term Rockburst Risk with Ensemble Classifiers. *Rock Mech. Rock Eng.* **2021**, *54*, 1799–1814. [CrossRef]
26. Zhou, S.; Zhang, Z.-X.; Luo, X.; Huang, Y.; Yu, Z.; Yang, X. Predicting dynamic compressive strength of frozen-thawed rocks by characteristic impedance and data-driven methods. *J. Rock Mech. Geotec. Eng.* **2024**, *16*, 2591–2606. [CrossRef]
27. Hosseini, S.; Khatti, J.; Taiwo, B.O.; Fissha, Y.; Grover, K.S.; Ikeda, H.; Pushkarna, M.; Berhanu, M.; Ali, M. Assessment of the ground vibration during blasting in mining projects using different computational approaches. *Sci. Rep.* **2023**, *13*, 18582. [CrossRef] [PubMed]
28. Khatti, J.; Grover, K.S. Assessment of hydraulic conductivity of compacted clayey soil using artificial neural network: An investigation on structural and database multicollinearity. *Earth Sci. Inform.* **2024**, *17*, 3287–3332. [CrossRef]
29. Daniel, C.; Khatti, J.; Grover, K.S. Assessment of compressive strength of high-performance concrete using soft computing approaches. *Comput. Concr.* **2024**, *33*, 55.
30. Yin, X.; Huang, X.; Pan, Y.C.; Liu, Q.S. Point and interval estimation of rock mass boreability for tunnel boring machine using an improved attribute-weighted deep belief network. *Acta Geotech.* **2023**, *18*, 1769–1791. [CrossRef]
31. Zhou, S.T.; Lei, Y.; Zhang, Z.X.; Luo, X.D.; Aladejare, A.; Ozoji, T. Estimating dynamic compressive strength of rock subjected to freeze-thaw weathering by data-driven models and non-destructive rock properties. *Nondestruct. Test. Eval.* **2024**. [CrossRef]
32. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 15–18 May 2000; pp. 93–104.
33. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
34. Talekar, B.; Agrawal, S. A Detailed Review on Decision Tree and Random Forest. *Biosci. Biotechnol. Res. Commun.* **2020**, *13*, 245–248. [CrossRef]
35. Kristóf, T.; Virág, M. EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. *Res. Int. Bus. Financ.* **2022**, *61*, 101644. [CrossRef]
36. Nam, B.H.; Park, K.; Kim, Y.J. Prediction of karst sinkhole collapse using a decision-tree (DT) classifier. *Geomech. Eng.* **2024**, *36*, 441–453. [CrossRef]
37. Akkaş, E.; Akin, L.; Evren Çubukçu, H.; Artuner, H. Application of Decision Tree Algorithm for classification and identification of natural minerals using SEM–EDS. *Comput. Geosci.* **2015**, *80*, 38–48. [CrossRef]
38. Domínguez-Olmedo, J.L.; Toscano, M.; Mata, J. Application of classification trees for improving optical identification of common opaque minerals. *Comput. Geosci.* **2020**, *140*, 104480. [CrossRef]
39. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
40. Mirjalili, S.; Lewis, A. The whale optimization algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]
41. Xiating, F.; Binrui, C.; Chuanqing, Z.; Shaojun, L.; Shiyong, W. *Mechanism, Warning and Dynamic Control of Rockburst Development Process*; Science Press: Beijing, China, 2013.
42. He, Z.; Xu, X.; Rao, D.; Peng, P.; Wang, J.; Tian, S. PSSegNet: Segmenting the P- and S-Phases in Microseismic Signals through Deep Learning. *Mathematics* **2024**, *12*, 130. [CrossRef]
43. Zhang, Q.; Zhang, X.-P.; Liu, Q.; Pan, Y.; Chi, J.; Qiu, J.; Yin, X. Microseismic Monitoring and Rockburst Characteristics in a Deep-Buried Tunnel Excavated by TBM. *Rock Mech. Rock Eng.* **2024**, *57*, 1565–1578. [CrossRef]
44. Basnet, P.M.S.; Jin, A.B.; Mahtab, S. Applying machine learning approach in predicting short-term rockburst risks using microseismic information: A comparison of parametric and non-parametric models. *Nat. Hazards* **2024**. [CrossRef]
45. Jin, A.B.; Basnet, P.; Mahtab, S. Evaluation of Short-Term Rockburst Risk Severity Using Machine Learning Methods. *Big Data Cogn. Comput.* **2023**, *7*, 172. [CrossRef]

46. Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4.
47. Khatti, J.; Grover, K.S. Prediction of compaction parameters for fine-grained soil: Critical comparison of the deep learning and standalone models. *J. Rock Mech. Geotec. Eng.* **2023**, *15*, 3010–3038. [CrossRef]

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.