



Original Paper

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: Comparison of Eight Single and Ensemble Models

Xin Yin^{1,2}, Quansheng Liu^{1,2,4}, Yucong Pan^{1,2,4}, Xing Huang³, Jian Wu^{1,2} and Xinyu Wang^{1,2}

Received 28 August 2020; accepted 27 November 2020
Published online: 3 January 2021

Rockburst is a common dynamic geological hazard, severely restricting the development and utilization of underground space and resources. As the depth of excavation and mining increases, rockburst tends to occur frequently. Hence, it is necessary to carry out a study on rockburst prediction. Due to the nonlinear relationship between rockburst and its influencing factors, artificial intelligence was introduced. However, the collected data were typically imbalanced. Single algorithms trained by such data have low recognition for minority classes. In order to handle the problem, this paper employed stacking technique of ensemble learning to establish rockburst prediction models. In total, 246 sets of data were collected. In the preprocessing stage, three data mining techniques including principal component analysis, local outlier factor and expectation maximization algorithm were used for dimension reduction, outlier detection and outlier substitution, respectively. Then, the pre-processed data were split into a training set (75%) and a test set (25%) with stratified sampling. Based on the four classical single intelligent algorithms, namely k -nearest neighbors (KNN), support vector machine (SVM), deep neural network (DNN) and recurrent neural network (RNN), four ensemble models (KNN-RNN, SVM-RNN, DNN-RNN and KNN-SVM-DNN-RNN) were built by stacking technique of ensemble learning. The prediction performance of eight models was evaluated, and the differences between single models and ensemble models were analyzed. Additionally, a sensitivity analysis was conducted, revealing the importance of input variables on the models. Finally, the impact of class imbalance on the prediction accuracy and fitting effect of models was quantitatively discussed. The results showed that stacking technique of ensemble learning provides a new and promising way for rockburst prediction, which exhibits unique advantages especially when using imbalanced data.

KEY WORDS: Rockburst prediction, Ensemble learning, Stacking technique, Class imbalance, Outlier detection and substitution.

¹The Key Laboratory of Geotechnical and Structural Engineering Safety of Hubei Province, School of Civil Engineering, Wuhan University, Wuhan 430072, China.

²State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China.

³State Key Laboratory of Geomechanics and Geotechnical Engineering, Institute of Rock and Soil Mechanics, Chinese Academy of Sciences, Wuhan 430071, China.

⁴To whom correspondence should be addressed; e-mail: li-uqs@whu.edu.cn, pyc1991@whu.edu.cn

INTRODUCTION

Rockburst is a common geological disaster and a dynamic instability phenomenon in deep engineering excavation and deep resource extraction; it is accompanied with rock loosening, peeling, ejection

tion and even throwing (Adoko et al. 2013; Zhou et al. 2018; Zhang et al. 2020a). Its occurrence is controlled by multiple factors, both internal and external causes (Zhang et al. 2020b). External causes include mainly ground stress and excavation disturbance. Rockburst usually occurs in high ground stress zone. Due to disturbance, the stress state existing in rock mass is changed, leading to stress redistribution and stress concentration. Internal causes such as rock strength, brittleness and integrity determine energy storage capacity of rock mass. When the elastic strain energy stored in rock mass exceeds that consumed by deformation and fracture, the rest will be converted into kinetic energy. In this situation, rock pieces are ejected at a certain velocity, namely rockburst. From the aspect of internal causes, rockburst tends to occur in the intact and hard rock mass. Additionally, rockburst is extremely destructive, which directly threatens the safety of workers and equipment, affects construction progress and even destroys an entire project and induces earthquakes (Xu et al. 2018).

Since the first recorded rockburst at the Leipzig coal mine in the United Kingdom in 1738 (He et al. 2017), rockbursts have been reported in many countries around the world. In South Africa (Zhou et al. 2012), almost all gold mines have suffered rockburst. It occurred only seven times in 1908 while the number of rockbursts rose to 680 in 1975. During this period, gold mines gradually shifted from shallow mining to deep mining. In Germany (Baltz and Hücke 2008), the number of hazardous rockbursts reached 283 only at the Luer mining area between 1910 and 1978. In Canada (Pu et al. 2018), rockbursts have taken place in many copper–nickel mines from the mid-twentieth century and consequently several mines were closed. In China (Cai et al. 2018), rockbursts happened at the Shengli coal mine in Fushun in 1933 for the first time. According to statistics, there were more than 2000 rockbursts in mines from 1949 to 1997 (Shi et al. 2005). Besides mining industry, rockbursts also frequently occurred in hydropower, transportation and other fields. For example, on November 28, 2009, a highly strong rockburst occurred in the drainage tunnel of Jinping II Hydropower Station (Zhou et al. 2016b); support systems were destroyed, tunnel boring machines worth 120 million were permanently buried and seven people were killed. With the development of underground space, rockburst has become a major challenge for the safety of deep underground engi-

neering (Pu et al. 2019a). It is urgent to carry out studies on rockburst prediction.

Early research mainly focused on rockburst prediction from a single influencing factor (Ouyang et al. 2015). Several empirical criteria from the perspective of strength theory, stiffness theory, energy theory, stability theory, fractal theory and catastrophe theory have been proposed, such as Russenes criterion (Russenes 1974), Barton criterion (Barton 2002) and Hoek criterion (Hoek and Brown 1980). As research continued, it was recognized that the occurrence of rockburst is affected by multiple factors, not just by a single factor (Ma et al. 2018). There is a significant nonlinear relationship between rockburst and these factors. In addition, these factors are interactive such that it is difficult to achieve high prediction accuracy using traditional empirical criteria. Artificial intelligence provides a powerful tool for solving such problems and it has been used widely in geotechnical engineering (Sun et al. 2019a, b, c, 2020a, b). Feng and Wang (1994) first used artificial neural network to predict rockburst, and later, some scholars (e.g., Jia et al. 2013; Sun et al. 2009) carried out related studies. Zhou et al. (2016a) utilized supervised learning to build ten rockburst prediction models and compared their performance. Pu et al. (2019b) employed support vector machine (SVM) to predict rockburst in kimberlite pipes at a diamond mine and achieved good results. Besides, many other intelligent models (Zhou et al. 2012; Adoko et al. 2013; Dong et al. 2013; Li et al. 2017a, b; Sousa et al. 2017; Li and Jimenez 2017; Pu et al. 2018; Roohollah and Abbas 2019; Wu et al. 2019; Xue et al. 2020; Zhou et al. 2020) have been applied for rockburst prediction (Table 1).

With the improvement of data acquisition approaches, massive data have emerged. However, data is generally imbalanced in practice. If trained by such data, single algorithms have low recognition for minority classes (Ganganwar 2012). Some researches (Díez-Pastor et al. 2015; Salunkhe and Mali 2016) indicated that ensemble learning has better performance when using imbalanced data. As seen in Table 1, there is little research about ensemble learning for rockburst prediction. In this paper, stacking technique (Wolpert 1992) of ensemble learning was adopted to build rockburst prediction models. This fills the gap and demonstrates the superiority of ensemble learning when using imbalanced data to predict rockburst.

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1797 Comparison of Eight Single and Ensemble Models

Table 1. Intelligent models used for rockburst prediction

Algorithm	Input parameters	References
Linear discriminant analysis	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Quadratic discriminant analysis	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Partial least-squares discriminant analysis	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Naïve Bayes	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
k -nearest neighbor	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Multilayer perceptron neural network	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Classification tree	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Support vector machine	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Random forest	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Gradient-boosting machine	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Zhou et al. (2016a)
Bayesian network	$H, \sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Li et al. (2017a)
Random forest	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Dong et al. (2013)
Support vector machine	$\sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, B_2, W_{et}$	Pu et al. (2019b)
Decision tree	SCF, B_1, W_{et}	Pu et al. (2018)
Support vector machine	$H, \sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, W_{et}$	Zhou et al. (2012)
Naïve Bayesian classifier	TSUP, K, D, σ_c , ORIENT	Sousa et al. (2017)
Tree-augmented naïve Bayesian classifier	TSUP, K, D, σ_c , ORIENT	Sousa et al. (2017)
Augmented naïve Bayesian classifier	TSUP, K, D, σ_c , ORIENT	Sousa et al. (2017)
Fuzzy inference system	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Adoko et al. (2013)
Adaptive neuro-fuzzy inference system	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Adoko et al. (2013)
Artificial neural network	$\sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, W_{et}$	Zhou et al. (2020)
Least squares support vector machine	$\sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, W_{et}$	Wu et al. (2019)
Extreme learning machine	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Li et al. (2017b)
Logistic regression classifier	$H, \sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Li and Jimenez (2017)
Extreme learning machine	$\sigma_\theta, \sigma_c, \sigma_t, SCF, B_1, W_{et}$	Xue et al. (2020)
Emotional neural network	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Roohollah and Abbas (2019)
Gene expression programming	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Roohollah and Abbas (2019)
Decision tree-based C4.5 algorithm	$\sigma_\theta, \sigma_c, \sigma_t, W_{et}$	Roohollah and Abbas (2019)

σ_θ is maximum tangential stress of surrounding rock; σ_c is uniaxial compressive strength of rock; σ_t is uniaxial tensile strength of rock; SCF is stress concentration factor; B_1 and B_2 are rock brittleness index; W_{et} is elastic strain energy index; H is depth; TSUP is type of support; K is ratio of horizontal stress to vertical stress; D is equivalent diameter of tunnel; ORIENT is orientation of the burst in the periphery of excavation

Stacking technique of ensemble learning can combine advantages of several single models to form a stronger one (Dietterich 2000). In this study, k -nearest neighbors (KNN), support vector machine (SVM), deep neural network (DNN) and recurrent neural network (RNN) were chosen as single models, which are also called base models. Based on them, four ensemble models (KNN–RNN, SVM–RNN, DNN–RNN and KNN–SVM–DNN–RNN) were set up by stacking technique. In total, 246 sets of data were collected. After dimension reduction with principal components analysis (PCA), outlier detection with local outlier factor (LOF) and outlier substitution with expectation maximization algorithm (EM), the data were divided into a training set (75%) and a test set (25%) by stratified sampling. The prediction performances of the eight models were evaluated based on the test set and the differences in performances between ensemble models and single models were analyzed. Moreover, a sen-

sitivity analysis was conducted through permutation importance to reveal the contribution of input variables on the models. Finally, the impact of class imbalance on prediction accuracy and fitting effect was discussed quantitatively. Figure 1 shows the flowchart of this study.

DATA ACQUISITION AND PREPROCESSING

Data Acquisition

The database was compiled from Pu et al. (2019b) and Zhou et al. (2016a), which indicates it is reliable and valid. It consisted of 246 sets of rockburst data from more than 20 engineering projects, including Jinping II hydropower station, Chengchao iron mine, Dongguashan copper mine, Daxiangling tunnel, Tongyu tunnel, Qinling tunnel and others

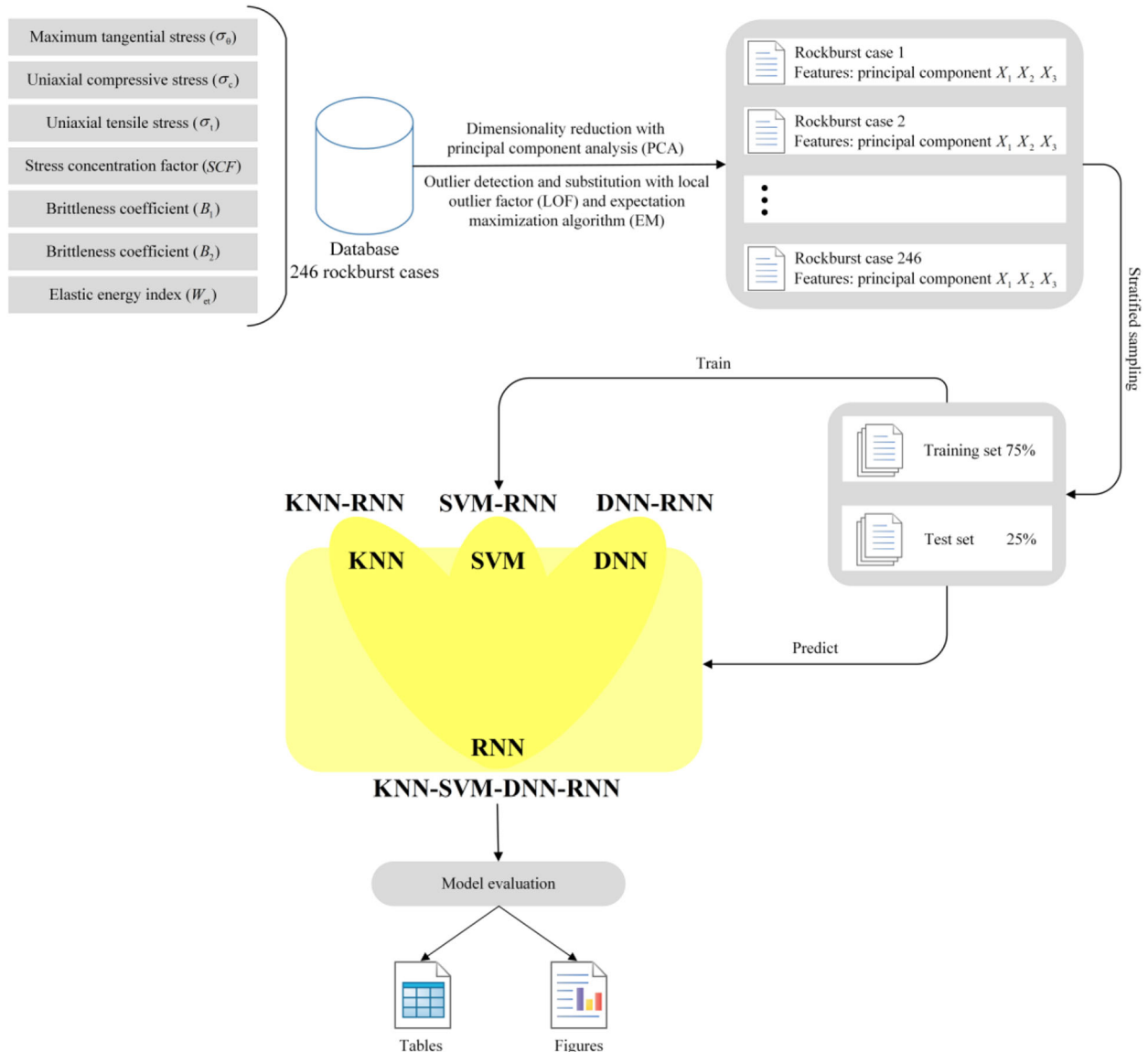


Figure 1. Flowchart of this paper.

(Zhou et al. 2012). Each set of data was composed of seven variables and a corresponding level of rockburst intensity. In this study, rockburst intensity was classified into four levels: none rockburst, low rockburst, moderate rockburst and high rockburst. The employed classification criterion was in accordance with Zhou et al. (2016a), and it is reported in Table 2. In the database, there were 43 sets with none rockburst, 78 sets with low rockburst, 81 sets with moderate rockburst, and 44 sets with high rockburst. Figure 2 shows the proportion of rock-

burst at four levels. Because the number of rockburst cases at each level was different, the database was imbalanced.

The seven variables were maximum tangential stress of surrounding rock (σ_{θ}), uniaxial compressive strength of rock (σ_c), uniaxial tensile strength of rock (σ_t), stress concentration factor (SCF), rock brittleness indices (B_1, B_2) and elastic strain energy index (W_{et}). Among these variables, σ_c, σ_t, B_1 and B_2 represent rock mechanical properties, σ_{θ} reflects ground stress, W_{et} is rock-energy storage capacity,

Table 2. Classification criterion of rockburst intensity (Zhou et al. 2016a)

Rockburst intensity	Failure characteristics
None	There are no sounds and rockburst activity
Low	The surrounding rock is deformed and cracked. There is no ejection or it is not obvious. A weak sound can be heard
Moderate	Large-scale surrounding rock bursts and several rock chips are ejected. There is a crisp crackling sound
High	The surrounding rock bursts severely. Rock fragments are thrown out suddenly or ejected strongly. There is a roaring sound

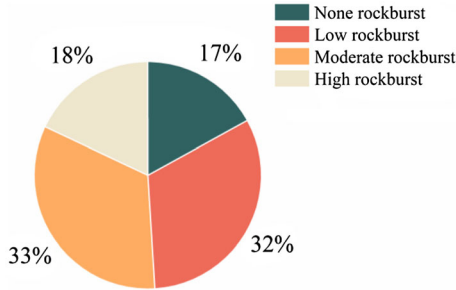


Figure 2. Proportion of rockburst at four levels.

and *SCF* stands for rock mechanical properties as well as ground stress. According to the mechanism of rockburst generation, the occurrence of rockburst depends not only on rock mechanical properties, but also on W_{et} and σ_θ (Afraei et al. 2019). Values of *SCF*, B_1 , B_2 and W_{et} can be calculated, respectively, as:

$$SCF = \frac{\sigma_\theta}{\sigma_c} \quad (1)$$

$$B_1 = \frac{\sigma_c}{\sigma_t} \quad (2)$$

$$B_2 = \frac{\sigma_c - \sigma_t}{\sigma_c + \sigma_t} \quad (3)$$

$$W_{et} = \frac{\phi_{sp}}{\phi_{st}} \quad (4)$$

where ϕ_{sp} and ϕ_{st} represent the stored elastic strain energy and the dissipated elastic strain energy, respectively, in a hysteresis looping test (Pu et al. 2019b).

Numerous empirical criteria for rockburst prediction have been raised based on one or several of the above seven variables. In this paper, σ_θ , σ_c , σ_t , *SCF*, B_1 , B_2 and W_{et} were all considered as the input variables to quantitatively predict rockburst. The

basic statistics of all the variables in the database are shown in Table 3.

Correlation Analysis and Dimension Reduction

Correlation Analysis with Pearson Correlation Coefficient

When there is a correlation between variables, the redundancy of information will increase, which will increase the time for model training and prediction. Hence, it is necessary to perform correlation analysis on the selected variables. Pearson correlation coefficient (Mu et al. 2018) is one of the most widely used methods to measure the correlation. Normally, it is defined as:

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (5)$$

where r_{xy} is the Pearson correlation coefficient for two variables X and Y , and \bar{x} and \bar{y} are the means of X and Y , respectively. The range of r_{xy} is $[-1, 1]$. When X and Y are positively correlated, the sign of r_{xy} is positive; otherwise, it is negative. The relationship between Pearson correlation coefficient and correlation strength is shown in Table 4 (Mohamed Salleh et al. 2015).

By analyzing 246 sets of data using Eq. 5, values of r_{xy} for pairs of variables are obtained (Table 5). The absolute value of r_{xy} is > 0.4 for variable pairs σ_θ and *SCF*, σ_θ and W_{et} , σ_c and σ_t , σ_t and B_1 , σ_t and B_2 , B_1 and B_2 , suggesting relatively strong correlations between these variables. Especially for σ_θ and *SCF*, the absolute value of r_{xy} reaches 0.92, indicating that σ_θ is very strongly correlated with *SCF*.

Table 3. Basic statistics of variables in the database

	σ_θ	σ_c	σ_t	SCF	B_1	B_2	W_{et}
Unit	MPa	MPa	MPa	–	–	–	–
Average	58.004	111.535	7.169	0.588	20.543	0.866	5.161
Minimum	2.6	20	1.3	0.1	0.15	0.19	0.81
Maximum	297.8	304.2	22.6	4.87	80	0.98	30
Standard deviation	54.075	42.661	4.182	0.672	14.241	0.098	4.156
Skewness	2.924	0.777	0.888	3.819	2.037	-3.539	3.247
Kurtosis	9.432	1.815	0.579	16.543	4.947	19.349	13.817

Table 4. Relationship between Pearson correlation coefficient and correlation strength

Absolute value of Pearson correlation coefficient	Strength of correlation
0.00–0.19	Very weak
0.20–0.39	Weak
0.40–0.59	Moderate
0.60–0.79	Strong
0.80–1.00	Very strong

Table 5. Pearson correlation matrix for the seven variables used in this study

	σ_θ	σ_c	σ_t	SCF	B_1	B_2	W_{et}
σ_θ	1	0.034	0.33	0.92	– 0.25	– 0.22	0.47
σ_c	0.034	1	0.42	– 0.27	0.076	0.23	0.19
σ_t	0.33	0.42	1	0.16	– 0.62	– 0.47	0.33
SCF	0.92	– 0.27	0.16	1	– 0.26	– 0.25	0.33
B_1	– 0.25	0.076	– 0.62	– 0.26	1	0.53	– 0.11
B_2	– 0.22	0.23	– 0.47	– 0.25	0.53	1	– 0.082
W_{et}	0.47	0.19	0.33	0.33	– 0.11	– 0.082	1

Dimension Reduction with Principal Component Analysis

Principal Component Analysis (PCA) is a kind of feature extraction technique for multivariate data (Cai et al. 2016). Its basic idea is to map high-dimensional space into low-dimensional space via mathematical transformation. In the field of data mining, it is used to reduce dimension.

From the correlation analysis, relatively strong correlations exist between several variables. Dimension reduction is an effective way to eliminate correlations. The original data used are 7-dimensional. In such high-dimensional situation, dimension disasters are prone to occur, such as sparse data and complex distance calculation. Thus, PCA was implemented to avoid the above problems in this study. The detailed steps were as follows:

1. Build the original data matrix $X = (x_{ij})_{m \times n}$, where m is number of samples, n is number of variables, and x_{ij} denotes the value for variable j of sample i .
2. For eliminating dimension effects and making variables comparable, standardize the original data as:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{6}$$

where \bar{x}_j and s_j are the mean and standard deviation of the j th variable, respectively.

3. Calculate correlation coefficient matrix $R = (r_{ij})_{n \times n}$ for the standardized data, where r_{ij} is the Pearson correlation coefficient for variables i and j .

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1801 Comparison of Eight Single and Ensemble Models

4. Calculate eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ and eigenvectors v_1, v_2, \dots, v_n of the matrix R .
5. Select appropriate principal components as new variables instead of original variables to achieve dimension reduction. In general, the first few principal components, whose eigenvalues are more than 1 and cumulative contribution rate exceeds 80%, are elected. The cumulative contribution rate of the first k principal components can be computed as:

$$\eta_k = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_n}. \quad (7)$$

In the results of PCA (Table 6), the eigenvalues of the first three principal components are greater than 1. Moreover, their cumulative contribution rate reaches 82.07% ($> 80\%$), which indicates that they contain more than 80% of the original variable information. Therefore, the first three principal components can be used as new variables to replace the original variables, which are separately denoted as X_1, X_2 and X_3 . Table 7 shows the data after dimension reduction.

Outlier Detection and Substitution

Outlier Detection with Local Outlier Factor

Outliers are defined as data that seriously deviate from others in a group, which may result from random error, artificial error, variation and so on (Knorr and Ng 1998). They can interfere with the training of models. Thus, it is essential to conduct outlier detection in a database.

In recent years, outlier detection algorithms have received more and more attention in the field of data mining, and they can be divided into two categories: distance-based and density-based. Distance-based algorithms can only detect global outliers but not local outliers in comparison with density-based algorithms. Local outlier factor (LOF) (Breunig et al. 2000) is a classic density-based outlier detection algorithm. When using LOF to detect whether an object is an outlier, it is critical to calculate the LOF of the object and compare it with a threshold. If it exceeds the threshold, the object is determined as an outlier. Given a database D and an

object O , the detailed steps for calculating the local outlier factor of O were as follows:

1. Find a set $N_k(O)$ including k nearest neighbors of O in the database D .
2. Calculate the k -distance of O as:

$$\text{dist}_k(O) = \max\{\text{dist}(O, P) | P \in N_k(O)\} \quad (8)$$

where $\text{dist}(O, P)$ denotes Euclidean distance between O and P .

3. Calculate the reachability distance between O and P as:

$$\text{dist}_{\text{reach}}(O, P) = \max\{\text{dist}_k(O), \text{dist}(O, P)\} \quad (9)$$

4. Calculate the local reachability density of O as:

$$\text{lrd}(O) = \frac{k}{\sum_{P \in N_k(O)} \text{dist}_{\text{reach}}(O, P)} \quad (10)$$

5. Calculate the LOF of O as:

$$\text{lof}(O) = \frac{\sum_{P \in N_k(O)} \frac{\text{lrd}(P)}{\text{lrd}(O)}}{k} \quad (11)$$

In this paper, k was set to the default value of 6. It was found after analysis that the LOF was distributed in the range $[0, 6.5]$ for about 99% of the data in the database. Only three (i.e., 1%) among the 246 sets of data had LOF of > 7 , which evidently deviated from the others. Referring to the definition of outliers, these are a minority in the database compared to normal points. For this reason, the threshold was fixed at 6.5. Finally, three outliers were detected, namely $(0.2114, -1.1297, -3.0158)$, $(-0.1164, 2.3938, -4.3962)$ and $(-0.0638, -1.8155, -4.7533)$. The results are shown in Figure 3.

Outlier Substitution with Expectation Maximization Algorithm

To explore fully the value of data, outlier substitution was conducted instead of directly removing outliers from the database. However, for minimizing disturbance to the data, substitution variable should be first determined. With regard to the variables X_1, X_2 and X_3 , they were separately analyzed by LOF. The calculation results of LOF indicated that the

Table 6. Principal component analysis results

Principal component	Eigenvalue	Contribution rate (%)	Cumulative contribution rate (%)
X_1	2.8139	40.20	40.20
X_2	1.5222	21.75	61.95
X_3	1.4085	20.12	82.07
X_4	0.6151	8.79	90.86
X_5	0.4433	6.33	97.19
X_6	0.1651	2.36	99.55
X_7	0.0318	0.45	100

Table 7. Data after dimension reduction with PCA

Case number	X_1	X_2	X_3
1	-0.1417	-0.1280	0.1023
2	-0.0671	-0.7754	-0.5554
3	0.0241	-0.2805	0.1599
4	0.1017	-0.4990	0.0446
...
246	-0.0638	-1.8155	-4.7533

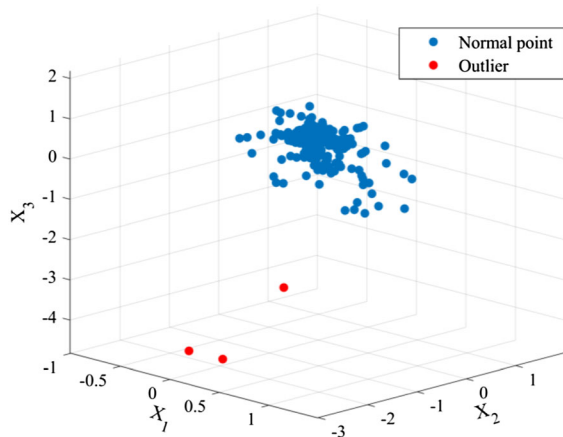


Figure 3. Data visualization after outlier detection with LOF.

abnormality of three outliers was induced mainly by X_3 (Fig. 4). Then, the expectation maximization (EM) algorithm (Dempster et al. 1977) was used to substitute X_3 so that the outliers turn into normal points. Its basic idea is as follows.

1. Calculate the expectation $\eta^{(0)}$ of the variable X_3 for normal points in each class.
2. According to the class to which an outlier belongs, replace the outlier in variable X_3

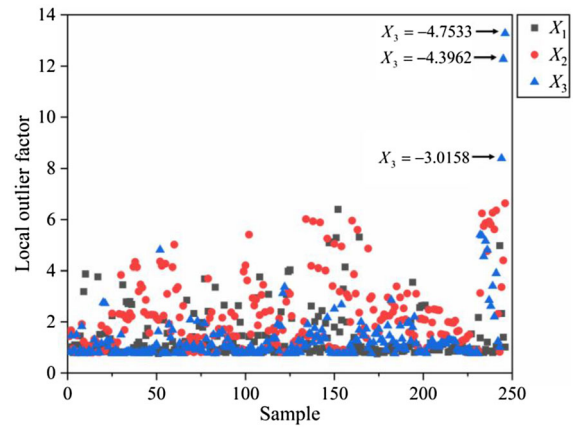


Figure 4. Calculation results of LOF for X_1 , X_2 and X_3 .

with the expectation $\eta^{(0)}$ of the corresponding class.

3. Denote $\eta^{(0)}$ as the initial iteration value and $\eta^{(i)}$ as the i th iteration value.
4. Calculate $Q(\eta, \eta^{(i-1)})$ in each iteration:

$$Q(\eta, \eta^{(i-1)}) = \log P(Y, Z|\eta)P(Z|Y, \eta^{(i-1)}) \quad (12)$$

where Y and Z represent observable and latent variables, separately, in the database; $P(Y, Z|\eta)$ indicates the joint probability distribution of Y and Z ; $P(Z|Y, \eta^{(i-1)})$ signifies the conditional probability distribution of Z given Y and $\eta^{(i-1)}$.

5. Maximize $Q(\eta, \eta^{(i-1)})$ and then determine $\eta^{(i)}$:

$$\eta^{(i)} = \arg \max_{\eta} Q(\eta, \eta^{(i-1)}) \quad (13)$$

6. Repeat steps (4) and (5) until the iteration converges.

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1803 Comparison of Eight Single and Ensemble Models

EM was implemented on SPSS23.0 software. The outliers (0.2114, - 1.1297, - 3.0158), (- 0.1164, 2.3938, - 4.3962) and (- 0.0638, - 1.8155, - 4.7533) were substituted by (0.2114, - 1.1297, 0.9177), (- 0.1164, 2.3938, 1.8912) and (- 0.0638, - 1.8155, 1.4495), respectively. The distribution of LOF after outlier substitution is illustrated in Figure 5. The outlier substitution was successful, as no LOF exceeded the afore-mentioned threshold.

To ensure that data in the training set and test set were sufficiently representative, the database was divided into two parts by stratified sampling after outlier substitution: 75% (185 sets of data) for training and 25% (61 sets of data) for testing. Because stratified sampling does not change data structure, the ratio between the numbers of rockburst at different levels in the training set is consistent with that in the database. Namely, the training set was imbalanced, too. Analyzing the prediction performance of models trained by imbalanced data was the focus of this paper.

CONSTRUCTION OF CLASSIFICATION MODELS

Eight classification models were built to carry out the prediction study of rockburst. The first four models were based on KNN, SVM, DNN and RNN; the rest—KNN-RNN, SVM-RNN, DNN-RNN and ALL (KNN-SVM-DNN-RNN)—were based on the stacking technique of ensemble learning. From analysis, RNN had the best performance among four single models. Thus, RNN was used as the base

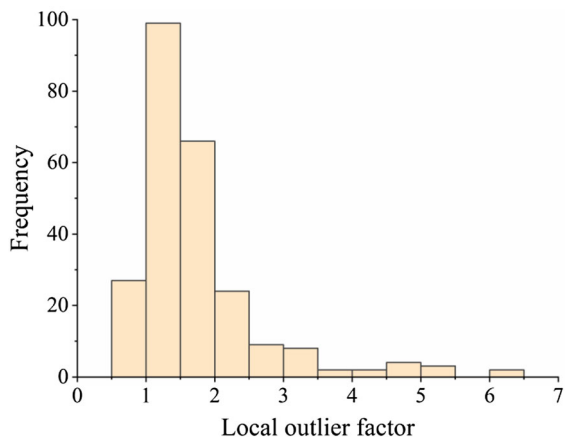


Figure 5. Distribution of LOF after outlier substitution.

learner of four ensemble models simultaneously. A detailed introduction to the eight classification models is given below.

In view of the nature of hyper-parameters, it can be divided into continuous and discrete ones. As for continuous hyper-parameters, grid search method (GSM), particle swarm optimization (PSO), cat swarm optimization (CSO), genetic algorithm (GA) and so on are commonly used as optimization algorithms (Zhou et al. 2012; Xue et al. 2020). With respect to discrete hyper-parameters, hold-out method, ten-fold cross-validation, five-fold cross-validation and leave-one-out method are frequently adopted for tuning (Zhou et al. 2016a, 2020). In this study, considering that c and g in SVM belong to continuous hyper-parameters, GSM was used as optimization algorithm. Regarding the number of nearest neighbors in KNN, the number of neurons in the hidden layer in DNN and RNN, they belong to discrete hyper-parameters, and so leave-one-out method was chosen as optimization algorithm. Another reason for employing leave-one-out method was that it was more reasonable and effective for relatively small database while hold-out method, ten-fold cross-validation and five-fold cross-validation were more used commonly for big database. The hyper-parameters of base learners of ensemble models inherit the ones of corresponding single models.

KNN is a commonly used supervised learning algorithm. Its primary aim is to calculate the distance between an object and its neighbors with known labels using Eq. 14:

$$d(x, y) = \left(\sum_{i=1}^q |x_i - y_i|^p \right)^{1/p} \quad (14)$$

where x is the object, y is its neighbor with known label; when p is 1 and 2, the distance is Manhattan distance and Euclidean distance, respectively. Then, KNN proceeds to assign the label with the highest frequency among the selected k nearest neighbors to this object. Compared with other supervised learning algorithms, KNN has no explicit training process so that the execution time of training process is zero. When receiving the test sample, training samples are processed. The disadvantage is that training samples need to be reprocessed each time a new test sample is received. The number of nearest neighbors (k) has a significant impact on the classification results of the model. Figure 6 shows the relationship between

k and the prediction performance of the model. It was found that the prediction performance of the model was optimal when k was set to 7.

SVM is a statistical learning algorithm with superior performance in classification tasks. Its basic idea is to find a hyperplane in the sample space to separate samples with different labels. The hyperplane can be defined as:

$$w^T x + b = 0 \tag{15}$$

where $w = (w_1, w_2, \dots, w_n)$ is the normal vector, which determines the direction of hyperplane; and b is the displacement term, which determines the distance between hyperplane and the origin.

However, when the sample space is not linearly separable, no hyperplane can divide correctly the samples into different classes. Because there is a significant nonlinear relationship between rockburst and its influencing factors, rockburst prediction is a typical linear inseparable problem. To solve this problem, radial basis function (RBF) kernel was adopted to map the original sample space to a higher-dimensional space where samples are linearly separable. The form of RBF kernel is expressed as:

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{16}$$

where σ is the kernel width.

Two parameters (c and g) play an important role in SVM: c is the penalty coefficient and represents the tolerance of errors; and g comes with RBF

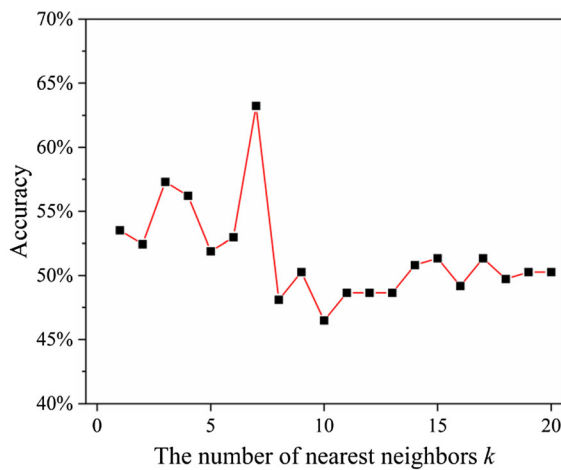


Figure 6. Effect of the number of nearest neighbors k on KNN prediction performance.

kernel and determines the number of support vectors. They are optimized by GSM and the optimization process is shown in Figure 7. The values of c and g were finally taken as 1.320 and 0.002, respectively.

DNN is a representative deep learning algorithm. It has more hidden layers in comparison with traditional artificial neural networks, such as standard BP neural network, RBF network. Theoretically, the higher the complexity of the model, the more complex learning tasks it can complete. In this work, DNN was designed as feedforward neural network with double hidden layers and it was trained with gradient descent using error back-propagation. During feedforward process, the output of neurons was calculated by Eq. 17.

$$y = f\left(\sum_{i=1}^n w_i x_i - \theta\right) \tag{17}$$

where $x = (x_1, x_2, \dots, x_n)$ and y are the input and output of neurons, respectively; $w = (w_1, w_2, \dots, w_n)$ and θ are the weights and bias of network, respectively; and $f(*)$ is the activation function. The weights and bias of network are updated through a back-propagation process, thus:

$$w_i = w_i - \eta \frac{\partial E}{\partial w_i} \tag{18}$$

where $w = (w_1, w_2, \dots, w_n)$ θE is the error between output and the actual value; and $\eta \in (0, 1)$ is the learning rate.

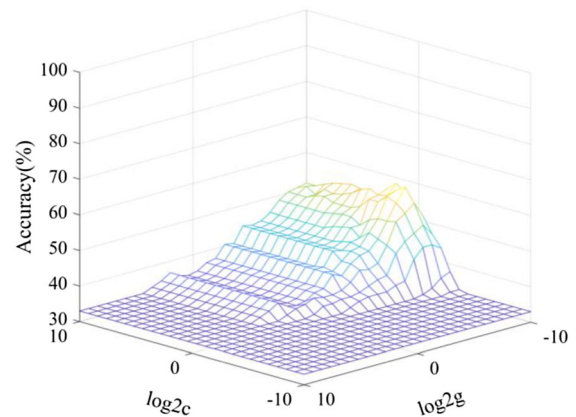


Figure 7. Optimization process to search optimal c and g for SVM.

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1805 Comparison of Eight Single and Ensemble Models

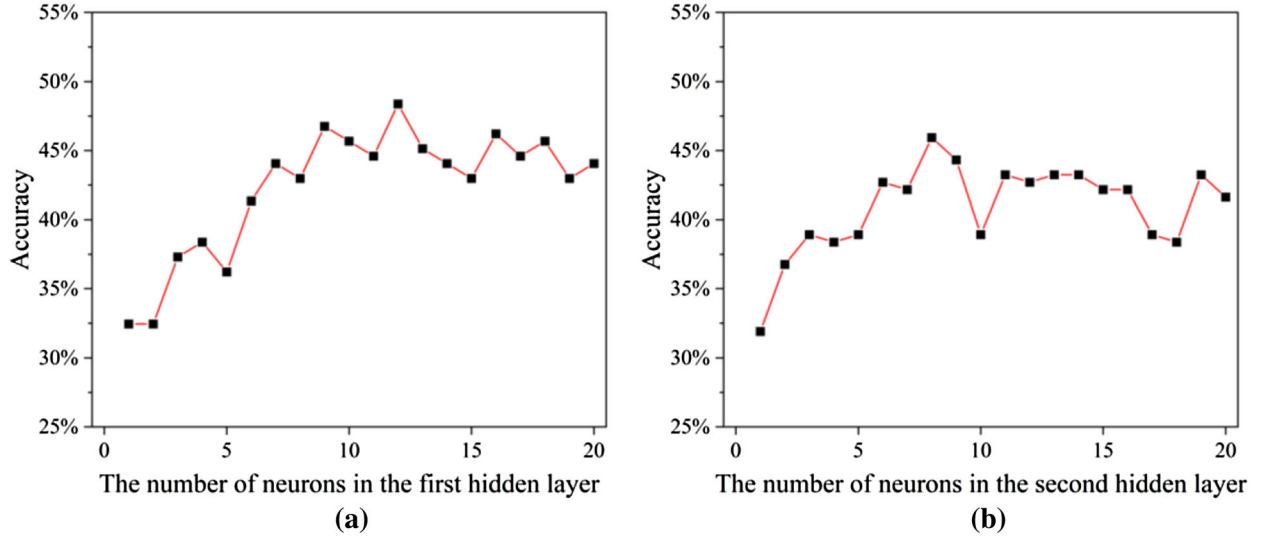


Figure 8. Effect of the number of neurons in two hidden layers on DNN prediction performance.

Referring to Figure 8a and b, the first hidden layer consisted of 12 neurons and the second was made up of eight neurons. At this time, the prediction performance of the model was the best. In addition, sigmoid tangent function was used as the activation function for the neurons.

RNN is another representative deep learning algorithm. It is different from feedforward neural networks, in which loops are allowed to exist so that the output of some neurons can be fed back as input. The Elman network, one of the most widely used RNN frameworks, is composed of four layers: input layer, hidden layer, state layer and output layer. The state layer allows the hidden layer to see its own previous output. Consequently, the subsequent behavior is affected by both the current input and the previous output of the hidden layer. The non-linear state space of the Elman network can be expressed as Eqs. 19, 20 and 21.

$$y(k) = g(w^3x(k)) \quad (19)$$

$$x(k) = f(w^1x_c(k) + w^2u(k - 1)) \quad (20)$$

$$x_c(k) = x(k - 1) \quad (21)$$

where u is input vector, y is output vector, x is vector of hidden layer, x_c is vector of state layer; w^3 represents the weights between output layer and hidden layer; w^2 represents the weights between input layer and hidden layer; w^1 represents the weights between

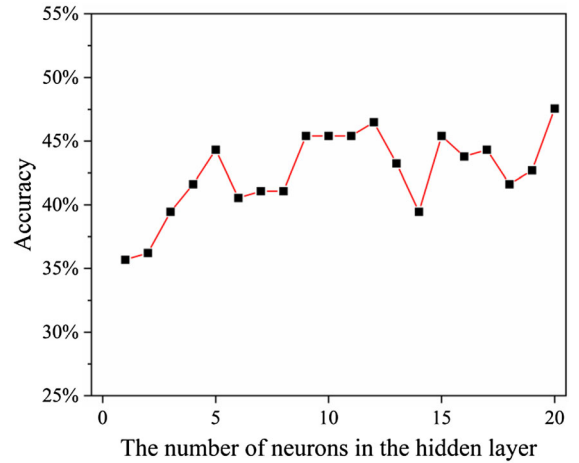


Figure 9. Effect of the number of neurons in the hidden layer on RNN prediction performance.

state layer and hidden layer; and $g(*)$ and $f(*)$ are the activation functions of the output and hidden layers, respectively.

The number of state layer units is equal to the number of output variables. As shown in Figure 9, the model achieves the best prediction performance when the number of hidden layer units was set to 20. Furthermore, the Levenberg–Marquardt algorithm was selected to the train Elman network, which is greatly useful for solving nonlinear problems.

Ensemble learning is the first of the four research directions of machine learning. It can en-

hance notably the generalization performance of models. Its basic idea is to accomplish learning tasks by combining multiple learners. According to combination strategies, ensemble learning can be classified into stacking (Wolpert 1992), bagging (Breiman 1996) and boosting (Friedman 2001). Bagging and boosting directly put several weak learners together to form a strong learner. The output of a strong learner is determined by averaging or voting on the outputs of weak learners. However, there are differences between bagging and boosting. The weak learners of the former are independent and can be generated in parallel while the ones of the latter have strong dependencies and they must be generated serially. Random Forest and AdaBoost are representative algorithms for bagging and boosting, respectively. Unlike bagging and boosting, stacking utilizes a meta-learner to gather a group of base learners. The combination of prediction results of base learners on the training set is regarded as new training set to train the meta-learner and that on the test set is used as new test set. The final output of stacking model is determined by the prediction results of meta-learner on the new test set. Its process is demonstrated in Figure 10.

In this research, stacking technique was adopted to establish four ensemble models (KNN-RNN, SVM-RNN, DNN-RNN and KNN-SVM-DNN-RNN). BP neural network using mind evolutionary algorithm served as the meta-learner, whose workflow is illustrated in Figure 11. Before training the BP neural network, its initial weights and bias were optimized via mind evolutionary algorithm, which avoids the randomness. In comparison with other evolutionary computations such as genetic algorithm and evolutionary programming, mind evolutionary algorithm overcomes the shortcomings of slow convergence and premature. To increase the diversity of base learners, output smearing (Breiman 2000) was applied, in which the initial classification output of base learners was transformed into the regression output and then input to the meta-learner.

MODEL EVALUATION METRICS

To measure the prediction performance of models, numerous evaluation metrics have been proposed (Hossin and Sulaiman 2015). Considering class imbalance of the data, accuracy (ACC), sensitivity (SNS), precision (PRC) and F_1 -score (F_1) were picked as evaluation metrics in this study (Luque

et al. 2019). ACC, which is characterized by intuition and simplicity, is the most widely used, indicating the ratio of correctly predicted samples to all predicted samples. SNS represents the proportion of correct predictions among all samples predicted as a certain class while PRC refers to the proportion of correct predictions among all samples belonging to a certain class. F_1 is a comprehensive index combining SNS and PRC, which eliminates their one-sidedness. In particular, F_1 is the harmonic average of SNS and PRC. These metrics can be calculated based on binary confusion matrix. With rockburst prediction being a 4-classification problem, it can be split into four binary classification problems according to one-versus-rest. Correspondingly, four binary confusion matrices were presented. The matrix consisted of four elements, which are named TP (true positive), FP (false positive), TN (true negative) and FN (false negative). Taking moderate rockburst as an example, moderate rockburst is regarded as one class while the rest (none rockburst, low rockburst and high rockburst) is classified as another class. Under the circumstances, the case with actual label being moderate rockburst is called positive; otherwise, it is called negative. The above four elements successively mean both the predicted label and the actual label are moderate rockburst, the prediction label is moderate rockburst but the actual label is not, neither the prediction label nor the actual label is moderate rockburst, the actual label is moderate rockburst but the prediction label is not. The explanation of binary confusion matrix is shown in Figure 12.

Based on the TP, FP, TN and FN, the four evaluation metrics can be as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (22)$$

$$SNS = \frac{TP}{TP + FN} \quad (23)$$

$$PRC = \frac{TP}{TP + FP} \quad (24)$$

$$F_1 = 2 \frac{SNS \times PRC}{SNS + PRC} \quad (25)$$

Because ACC is a global metric, which is applicable to binary classification problems as well as multi-classification problems, its expression can be generalized as:

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1807 Comparison of Eight Single and Ensemble Models

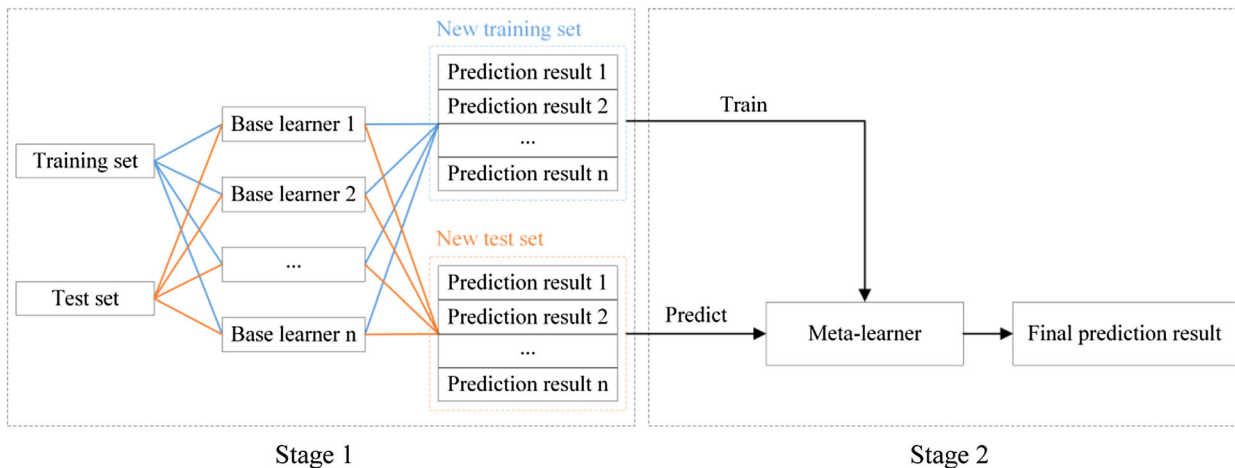


Figure 10. Process of stacking technique.

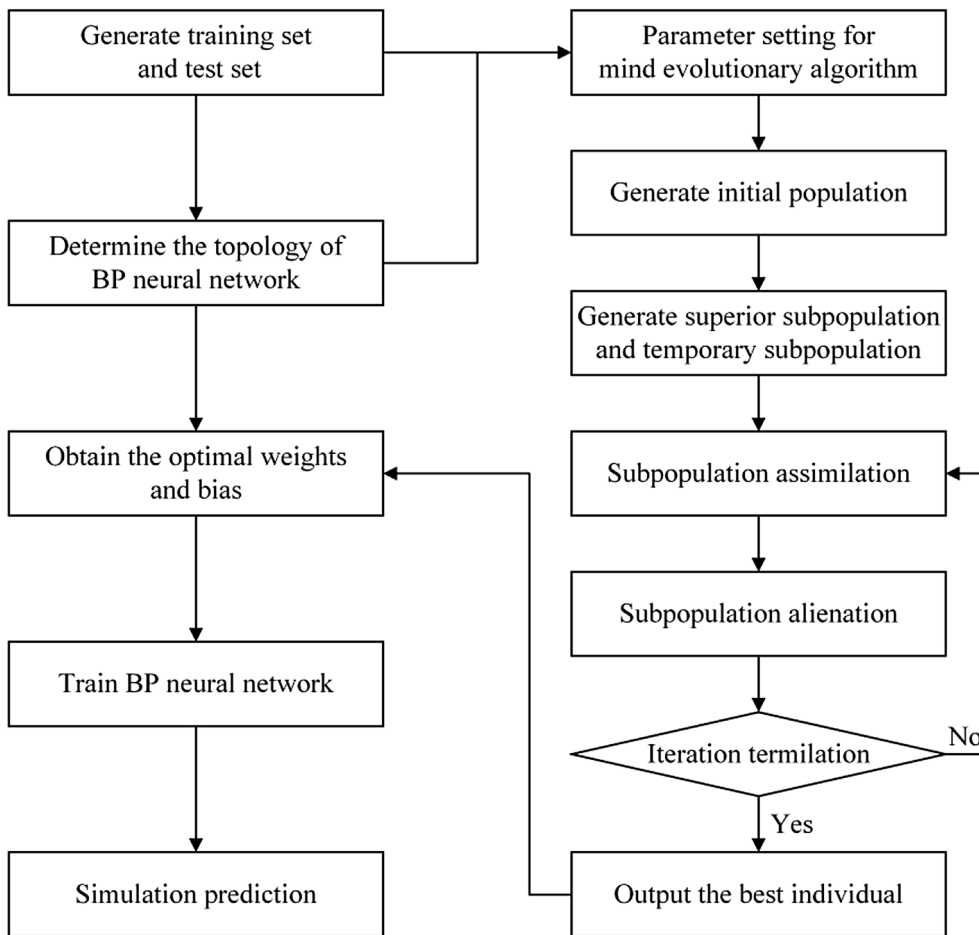


Figure 11. Workflow of meta-learner of stacking technique.

		Predicted label	
		Moderate rockburst	None/Low/High rockburst
Actual label	Moderate rockburst	TP (true positive)	FN (false negative)
	None/Low/High rockburst	FP (false positive)	TN (true negative)

Figure 12. Binary confusion matrix (taking moderate rockburst as an example).

$$ACC = \frac{\text{correct prediction}}{\text{correct prediction} + \text{wrong prediction}} \quad (26)$$

ENGINEERING VALIDATION

In total, 61 rockburst cases were used to test the applicability and practicability of models in engineering. The evaluation metrics introduced in the preceding section were calculated (Table 8). It was found that the maximum of four metrics appeared in ensemble models. Moreover, the metrics of ensemble models were generally larger than that of single models. The results signify that ensemble models achieved better prediction performance than the single models. In particular, ensemble model KNN-RNN had the best prediction performance, whose four metrics simultaneously reached the maximum among the eight classification models.

Based on ACC, SVM had the worst performance with ACC of 47.54%, followed by KNN and DNN with 52.46%; KNN-RNN had the best performance with 91.80%. The ACC of the ensemble models was generally > 80% but those single models were < 55%. Thus, ensemble learning can greatly improve the prediction performance when

using imbalanced data. In addition, RNN achieved the highest ACC among the single learner models. This is why RNN was used as the base learner of the four ensemble models simultaneously.

For detailed analysis of model performance, Figure 13 demonstrates the histograms of errors produced by the eight classification models. Because rockburst prediction is a multi-classification problem, unlike regression problems, its error values were discrete. The range of errors was $\{-3, -2, -1, 0, 1, 2, 3\}$. The magnitude represents the degree of deviation from the actual rockburst level. The positive sign indicates that the predicted level is higher than the actual level. In contrast, the negative sign means that the predicted level is lower than the actual level.

When the predicted level is higher, it will lead to a waste of resources. However, when the predicted level is lower, the existing engineering support measures are not enough to withstand rockburst, which will cause construction safety risks, and even yield huge property damage and casualties. From the perspective of engineering safety, positive errors are better than negative errors. Moreover, the smaller the magnitude of errors, the better the prediction performance.

Regarding the ensemble models, more than 90% of errors were non-negative and only a few test samples had errors of > 1. Especially for KNN-RNN, only one of the 61 test samples had a lower predicted level than the actual level and there were no errors > 1. However, as for single models, the proportion of non-negative errors was about 80% and more test samples had errors > 1. In particular, DNN and RNN had errors reaching -3, which means that the actual level was high rockburst but was predicted as none rockburst. This result is not acceptable in engineering applications.

In general, classification models perform better on the classes with more training samples (Kautz et al. 2017). Because the data used in this study was imbalanced, it is necessary to analyze the prediction accuracy of the eight classification models on each class. Figure 14 shows the prediction results of the eight classification models on each class. For the same model, the prediction accuracy on none/low/high rockburst was lower than that on moderate rockburst. This is in accordance with the distribution of the data, in which none/low/high rockburst were minority classes and moderate rockburst was majority class. However, ensemble models had higher prediction accuracy on the minority classes in

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1809 Comparison of Eight Single and Ensemble Models

Table 8. Evaluation metrics for the eight classification models

Model	ACC (%)	SNS				PRC				F_1			
		None	Low	Moderate	High	None	Low	Moderate	High	None	Low	Moderate	High
KNN	52.46	0.273	0.421	0.800	0.455	0.375	0.444	0.552	0.833	0.316	0.432	0.653	0.588
SVM	47.54	0.273	0.263	0.800	0.455	0.500	0.417	0.457	0.625	0.353	0.323	0.582	0.526
DNN	52.46	0.364	0.526	0.700	0.364	0.400	0.625	0.500	0.444	0.381	0.571	0.583	0.400
RNN	54.10	0.182	0.579	0.800	0.364	0.500	0.524	0.571	0.500	0.267	0.550	0.667	0.421
KNN-RNN	91.80	0.727	0.947	1.000	0.909	1.000	0.857	0.909	1.000	0.842	0.900	0.952	0.952
SVM-RNN	82	0.636	0.789	1.000	0.727	0.700	0.789	0.833	1.000	0.667	0.789	0.909	0.842
DNN-RNN	88.52	0.636	0.947	1.000	0.818	1.000	0.818	0.870	1.000	0.778	0.878	0.930	0.900
ALL	88.52	0.636	0.947	1.000	0.818	1.000	0.818	0.909	0.900	0.778	0.878	0.952	0.857

The maximum of each metric is highlighted in bold

comparison with single models. Taking none rockburst as example, the average accuracy of single models was 27.27% while the average accuracy of ensemble models was 65.91%. This means that ensemble learning can enhance the recognition ability for minority classes.

RELATIVE IMPORTANCE OF INPUT VARIABLES

In this study, the contributions of input variables to eight classification models were evaluated quantitatively by means of permutation importance (Cortez and Embrechtsb 2013). Its basic idea is to shuffle an input variable in the test set and leave the others in place, then calculate the difference of prediction accuracy on no-shuffled and shuffled test set. Obviously, it was implemented after the model has been trained. The importance score of input variables can be as:

$$S_j = \frac{V_j}{\sum_{i=1}^n V_i} \times 100 \quad (27)$$

where S_j is the importance score of the j -th input variable, V_j is the difference of prediction accuracy on no-shuffled and shuffled test set when shuffling the j -th input variable.

Figure 15 lists the importance scores of input variables on the eight classification models. It was found that the models were sensitive to input variables X_1 , X_2 and X_3 . X_1 was the most important, followed by X_2 and X_3 . Furthermore, X_3 was more important than X_2 . Although the importance score for the same input variable was not consistent on different models, the level of the relative importance was identical.

DISCUSSION

Impact of Class Imbalance on Prediction Accuracy

In previous work (e.g., Daskalaki et al. 2006; Branco et al. 2017), it has been proved that class imbalance has a significant impact on the prediction accuracy of classification models. For quantitatively investigating the impact of class imbalance on the eight classification models established in this paper, it was necessary to conduct the sensitivity analysis of prediction accuracy to class imbalance. Herein, a balanced training set was created based on the original imbalanced training set. The classification models were separately trained by the two training sets and then their prediction accuracy was compared on the same test set.

Sampling technology is an effective means to eliminate class imbalance, which can be roughly divided into two categories: under-sampling and over-sampling. The former is to remove some samples from majority classes to achieve the balance between different classes. The advantage is that the size of training set becomes smaller and the consumed time of training process is reduced. The following disadvantage is that some important information may be lost. The latter achieves the rebalance of classes through adding some samples into minority classes. It should be noted that the same training samples cannot be sampled repeatedly; otherwise, this will result in serious over-fitting. Considering the limited amount of training samples, over-sampling was employed in this study to ensure information integrity. The synthetic minority over-sampling technique (SMOTE) is one of the representative algorithms for over-sampling (Chawla et al. 2011). The basic idea of SMOTE is to increase the minority class samples by interpolation. The

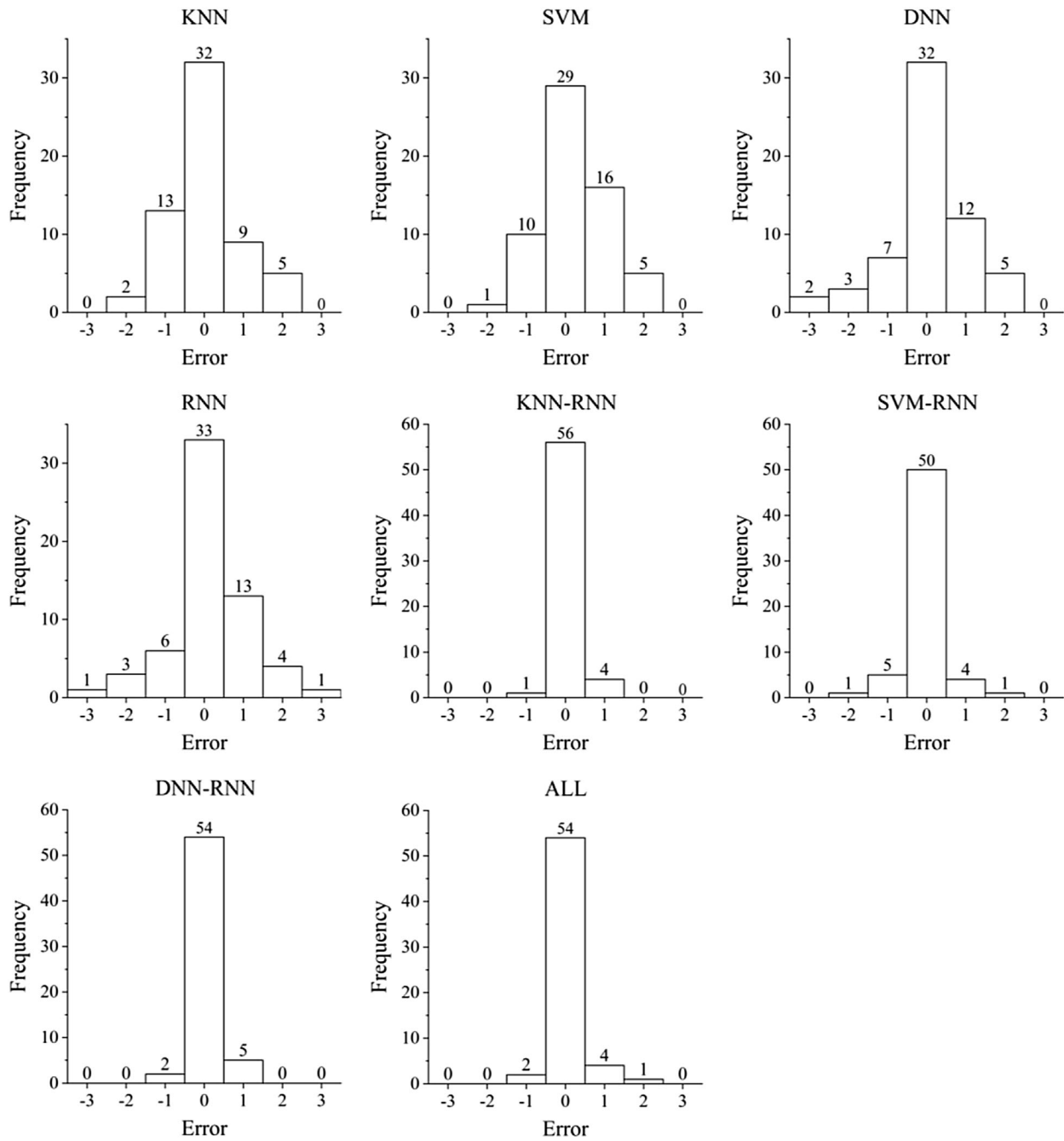


Figure 13. Histograms of errors produced by the eight classification models.

procedure to implement SMOTE is shown below, whereby the number of nearest neighbors was fixed at the default value of 5 (Chawla et al. 2011).

Based on SMOTE, a balanced training set was built. For the original training set, no/low/high rockburst was minority class and moderate rock-

burst was majority class. After SMOTE, the number of no/low/high rockbursts became 61, which is the same as the number of moderate rockbursts.

Table 9 shows the prediction accuracy on the test set for the eight classification models separately trained by different training sets. After eliminating

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1811 Comparison of Eight Single and Ensemble Models

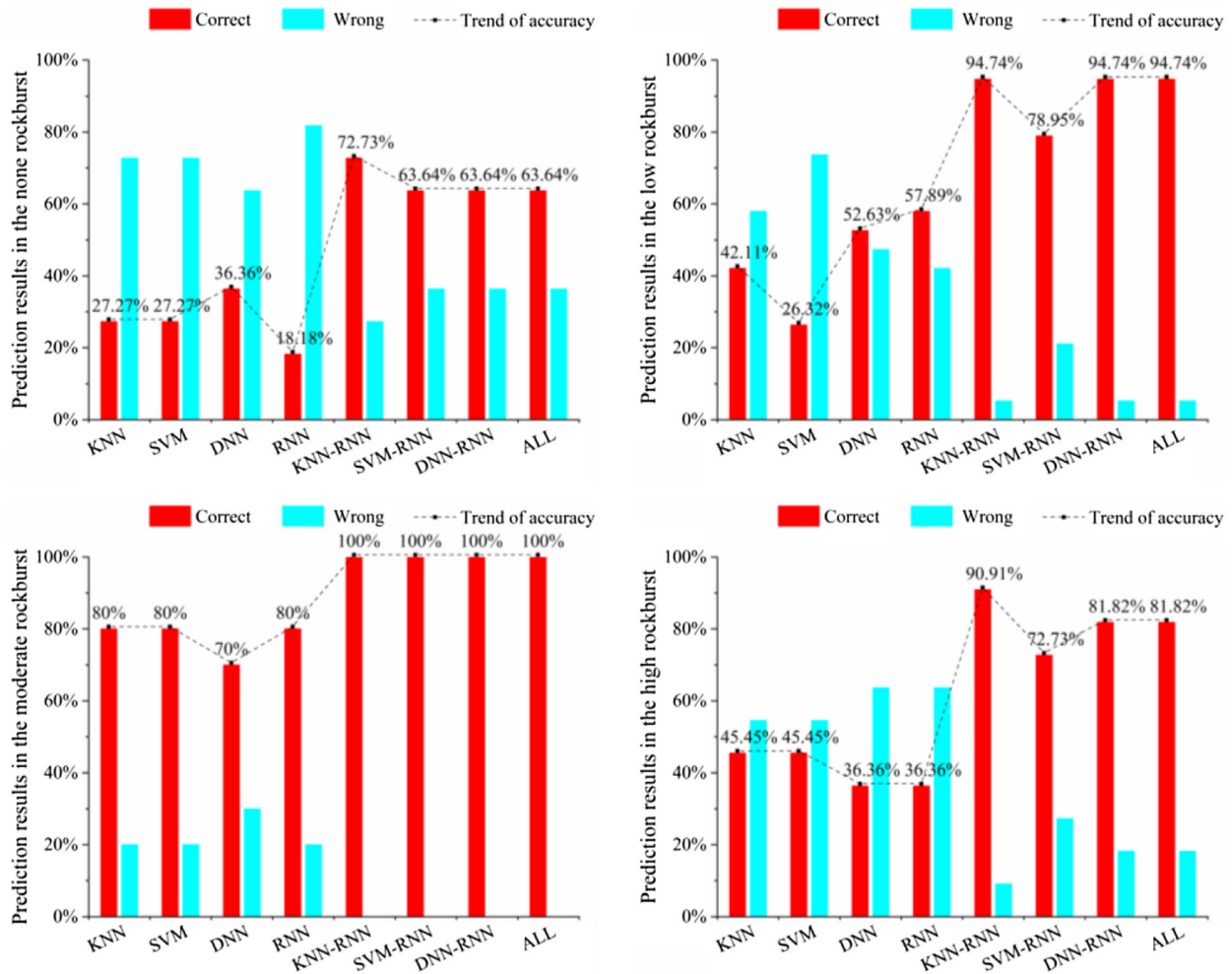


Figure 14. Prediction results of the eight classification models on each class.

class imbalance, the accuracy of single models was improved and the average improvement was 19.93%. However, the accuracy improvement of ensemble models was small; an average of merely 3.27%. Thus, ensemble models are relatively stable against class imbalance while single models are sensitive. Although the accuracy of single models was improved after eliminating class imbalance, it was still less than that of the ensemble models.

Impact of Class Imbalance on Fitting Effect

In the training process, over-fitting frequently occurs. This causes the prediction accuracy on the training set to be much higher than that on the test set, which suggests that the generalization perfor-

mance of models is poor. This situation needs to be avoided. The scatter plot, taking the prediction accuracy on the test set as x axis and the prediction accuracy on the training set as y axis, is a useful tool to examine the fitting effect of models. When a point falls above the line $y = x$, over-fitting tends to occur. Moreover, the larger the deviation from the line $y = x$, the more serious the over-fitting is; it is satisfactory if a point falls on or below the line.

Figure 16 shows the relationship between the prediction accuracy on the training set and test set for the eight classification models. Although all the points were above the line $y = x$, the points after eliminating class imbalance were closer to the line $y = x$. The results from the scatter plot intuitively indicate that eliminating class imbalance can improve fitting effect of models. Next, the impact of

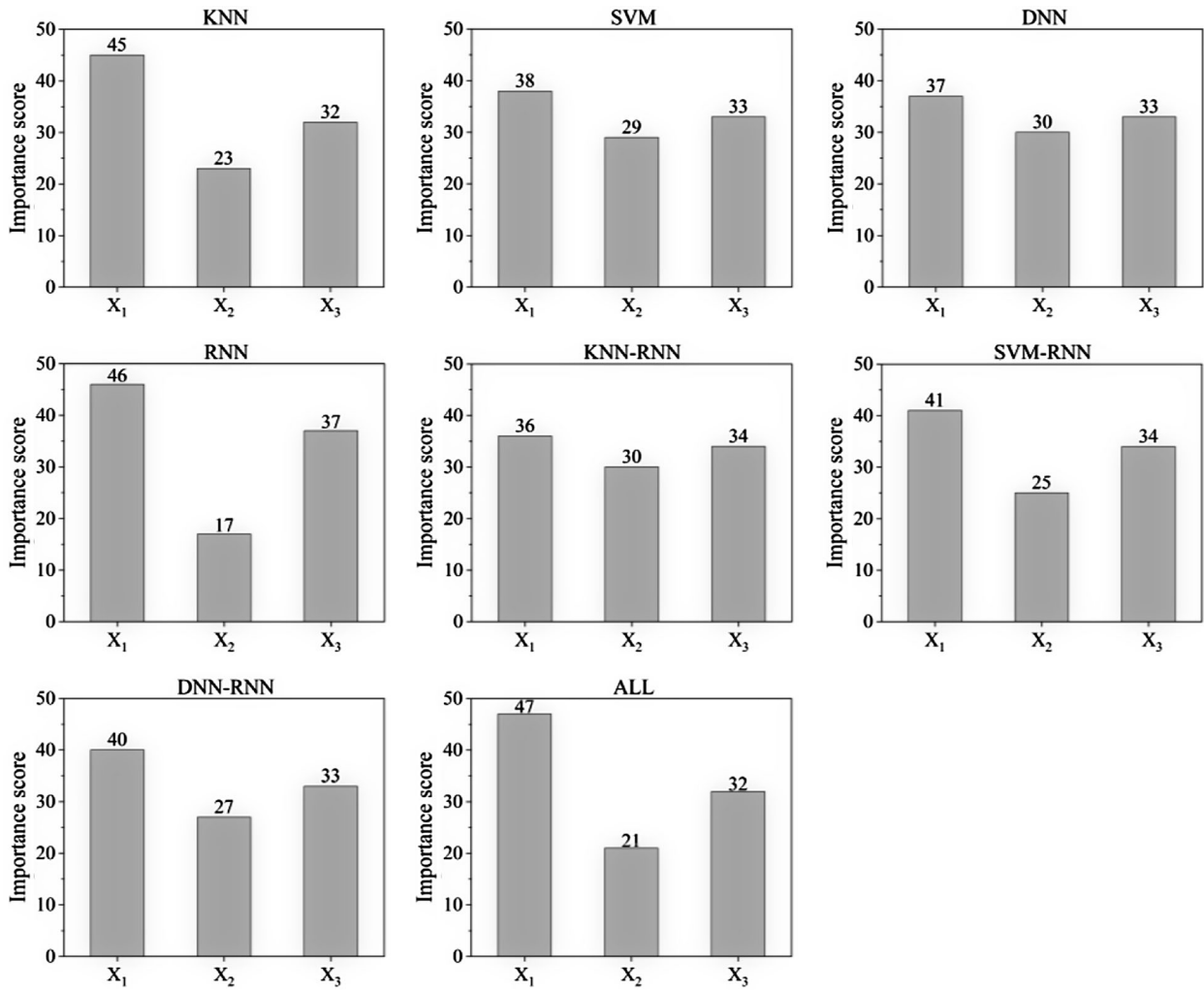


Figure 15. Importance score of input variables on the eight classification models.

Table 9. Comparison of prediction accuracy on the test set before and after eliminating class imbalance

Model	Prediction accuracy		Change (%)
	Before eliminating class imbalance (%)	After eliminating class imbalance (%)	
KNN	52.46	67.21	14.75
SVM	47.54	67.21	19.67
DNN	52.46	63.93	11.47
RNN	54.10	63.93	9.83
KNN-RNN	91.80	95.08	3.28
SVM-RNN	82.00	86.89	4.89
DNN-RNN	88.52	93.44	4.92
ALL	88.52	88.52	0.00

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1813 Comparison of Eight Single and Ensemble Models

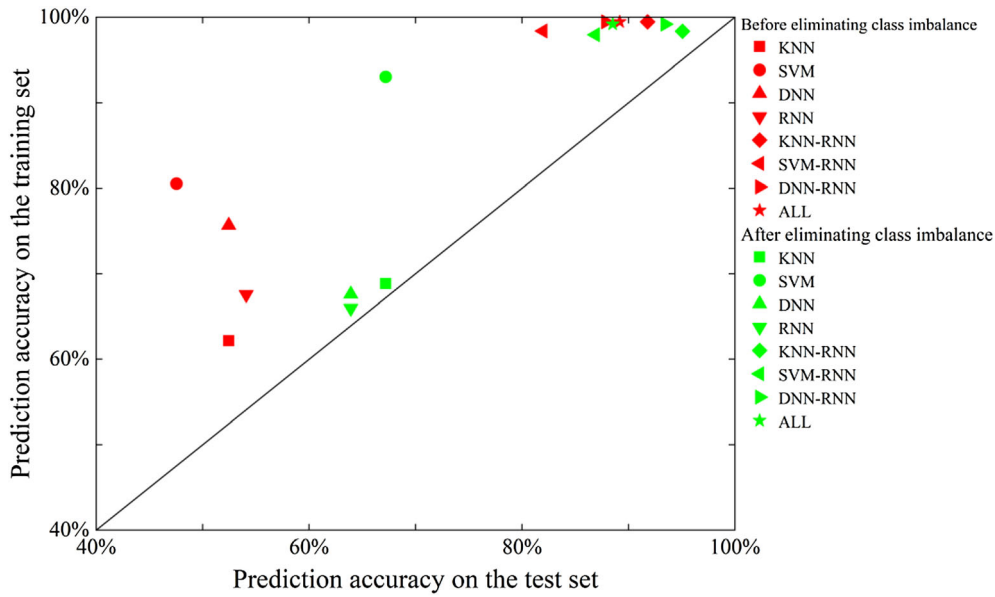


Figure 16. Scatter plot between the prediction accuracy on the training set and test set.

class imbalance on fitting effect was analyzed quantitatively.

The difference between the prediction accuracy on the training set and test set was used as quantitative index to measure fitting effect, which geometrically represents the vertical distance from the points shown in Figure 16 to the line $y = x$. When the difference is larger, the deviation from the points to the line $y = x$ is bigger, which means that overfitting is more serious. Table 10 shows the results of difference before and after eliminating class imbalance. After eliminating class imbalance, the difference for the eight classification models all decreased, which suggests that class imbalance had an impact on fitting effect. Specifically, eliminating class imbalance can improve fitting effect. Among the eight classification models, the change of difference for DNN after eliminating class imbalance was maximal, reaching 21.17%. This means that class imbalance had the greatest impact on DNN.

CONCLUSIONS

The database including 246 sets of data was analyzed in this paper. To eliminate correlation among the variables and avoid dimension disasters, principal components analysis was employed to transform the original data from seven dimensions

to three dimensions. In addition, local outlier factor and expectation maximization algorithm were used to detect and substitute outliers, respectively, in the database. Then, the database was split into the training set (75%) and test set (25%) by stratified sampling.

Eight classification models were established for rockburst prediction, including four single models (KNN, SVM, DNN and RNN) and four ensemble models (KNN-RNN, SVM-RNN, DNN-RNN and KNN-SVM-DNN-RNN). The ensemble models were built based on stacking technique of ensemble learning. Considering class imbalance of the data, accuracy, sensitivity, precision and F_1 score were used as model evaluation metrics. The results indicate that, when using imbalanced data, the ensemble models performed better than the single models. Moreover, the prediction performance on minority classes was improved by stacking technique of ensemble learning. After that, a sensitivity analysis was conducted, which revealed the importance of input variables on the eight classification models.

Finally, the impact of class imbalance on prediction accuracy and fitting effect was analyzed quantitatively. The results showed that ensemble models were relatively stable against class imbalance while single models were sensitive to it. Furthermore, eliminating class imbalance can improve fitting effect of models. In conclusion, the use of

Table 10. Comparison of difference between prediction accuracy on the training set and test set before and after eliminating class imbalance

Model	Difference		Change (%)
	Before eliminating class imbalance (%)	After eliminating class imbalance (%)	
KNN	9.7	1.64	8.06
SVM	33	25.82	7.18
DNN	23.22	2.05	21.17
RNN	13.47	3.69	9.78
KNN-RNN	7.66	3.28	4.38
SVM-RNN	16.41	11.06	5.35
DNN-RNN	10.94	5.74	5.2
ALL	10.94	10.66	0.28

stacking technique of ensemble learning serves as a new and promising way for rockburst prediction. Especially in the case of imbalanced data, it shows unique advantages.

ACKNOWLEDGMENTS

This research is supported by National Natural Science Foundation of China under Grant Nos. 41941018 and 41807250, China Postdoctoral Science Foundation Program under Grant Nos. 2019T120686, and National Key Basic Research Program of China under Grant Nos. 2015CB058102. These supports are gratefully acknowledged.

REFERENCES

- Adoko, A. C., Gokceoglu, C., Wu, L., & Zuo, Q. J. (2013). Knowledge-based and data-driven fuzzy modeling for rockburst prediction. *International Journal of Rock Mechanics and Mining Sciences*, *61*, 86–95.
- Afraei, S., Shahriar, K., & Madani, S. H. (2019). Developing intelligent classification models for rock burst prediction after recognizing significant predictor variables, section 1: Literature review and data preprocessing procedure. *Tunnelling and Underground Space Technology*, *83*, 324–353.
- Baltz, R., & Hucke, A. (2008). Rockburst prevention in the German coal industry. In *Proceedings of the 27th international conference on ground control in mining* (pp. 46–50).
- Barton, N. (2002). Some new Q-value correlations to assist in site characterisation and tunnel design. *International Journal of Rock Mechanics and Mining Sciences*, *39*(2), 185–216.
- Branco, P., Torgo, L., & Ribeiro, R. P. (2017). Relevance-based evaluation metrics for multi-class imbalanced domains. In *Advances in knowledge discovery and data mining* (pp. 698–710).
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.
- Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, *40*(3), 229–242.
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104).
- Cai, W., Dou, L. M., Si, G. Y., Cao, A. Y., He, J., & Liu, S. (2016). A principal component analysis/fuzzy comprehensive evaluation model for coal burst liability assessment. *International Journal of Rock Mechanics and Mining Sciences*, *81*, 62–69.
- Cai, W., Dou, L., Zhang, M., Cao, W., Shi, J., & Feng, L. (2018). A fuzzy comprehensive evaluation methodology for rock burst forecasting using microseismic monitoring. *Tunnelling and Underground Space Technology*, *80*, 232–245.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357.
- Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17.
- Daskalaki, S., Kopanas, I., & Avouris, N. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, *20*(5), 381–417.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1–38.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15).
- Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C., & Kuncheva, L. I. (2015). Random balance: Ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems*, *85*, 96–111.
- Dong, L. J., Li, X. B., & Peng, K. (2013). Prediction of rockburst classification using random forest. *Transactions of Nonferrous Metals Society of China*, *23*(2), 472–477.
- Feng, X. T., & Wang, L. (1994). Rockburst prediction based on neural networks. *Transactions of Nonferrous Metals Society of China*, *4*(1), 7–14.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.
- Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.
- He, J., Dou, L. M., Gong, S. Y., Li, J., & Ma, Z. Q. (2017). Rock burst assessment and prediction by dynamic and static stress analysis based on micro-seismic monitoring. *International Journal of Rock Mechanics and Mining Sciences*, *93*, 46–53.

Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: 1815 Comparison of Eight Single and Ensemble Models

- Hoek, E., & Brown, E. T. (1980). *Underground excavations in rock*. London: The Institution of Mining and Metallurgy.
- Hossin, M., & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1.
- Jia, Y., Lu, Q., & Shang, Y. (2013). Rockburst prediction using particle swarm optimization algorithm and general regression neural network. *Chinese Journal of Rock Mechanics and Engineering*, 32(2), 343–348.
- Kautz, T., Eskofier, B. M., & Pasluosta, C. F. (2017). Generic performance measure for multiclass-classifiers. *Pattern Recognition*, 68, 111–125.
- Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th VLDB conference* (pp. 392–403).
- Li, N., & Jimenez, R. (2017). A logistic regression classifier for long-term probabilistic prediction of rock burst hazard. *Natural Hazards*, 90(1), 197–215.
- Li, N., Jimenez, R., & Feng, X. D. (2017a). The influence of bayesian networks structure on rock burst hazard prediction with incomplete data. *Procedia Engineering*, 191, 206–214.
- Li, T. Z., Li, Y. X., & Yang, X. L. (2017b). Rock burst prediction based on genetic algorithms and extreme learning machine. *Journal of Central South University*, 24(9), 2105–2113.
- Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216–231.
- Ma, T. H., Tang, C. A., Tang, S. B., Kuang, L., Yu, Q., Kong, D. Q., et al. (2018). Rockburst mechanism and prediction based on microseismic monitoring. *International Journal of Rock Mechanics and Mining Sciences*, 110, 177–188.
- Mohamed Salleh, F. H., Arif, S. M., Zainudin, S., & Firdaus-Raih, M. (2015). Reconstructing gene regulatory networks from knock-out data using Gaussian Noise Model and Pearson Correlation Coefficient. *Computational Biology and Chemistry*, 59, 3–14.
- Mu, Y., Liu, X., & Wang, L. (2018). A Pearson's correlation coefficient based decision tree and its parallel implementation. *Information Sciences*, 435, 40–58.
- Ouyang, Z. H., Qi, Q. X., Zhao, S. K., Wu, B. Y., & Zhang, N. B. (2015). The mechanism and application of deep-hole pre-cracking blasting on rockburst prevention. *Shock and Vibration*, 2015, 1–7.
- Pu, Y. Y., Apel, D. B., & Lingga, B. (2018). Rockburst prediction in kimberlite using decision tree with incomplete data. *Journal of Sustainable Mining*, 17(3), 158–165.
- Pu, Y. Y., Apel, D. B., Liu, V., & Mitri, H. (2019a). Machine learning methods for rockburst prediction-state-of-the-art review. *International Journal of Mining Science and Technology*, 29(4), 565–570.
- Pu, Y. Y., Apel, D. B., & Xu, H. W. (2019b). Rockburst prediction in kimberlite with unsupervised learning method and support vector classifier. *Tunnelling and Underground Space Technology*, 90, 12–18.
- Roohollah, S. F., & Abbas, T. (2019). Long-term prediction of rockburst hazard in deep underground openings using three robust data mining techniques. *Engineering with Computers*, 35(2), 659–675.
- Russenes, B. (1974). *Analyses of rockburst in tunnels in valley sides*. Trondheim: Norwegian Institute of Technology.
- Salunkhe, U. R., & Mali, S. N. (2016). Classifier ensemble design for imbalanced data classification: A hybrid approach. *Procedia Computer Science*, 85, 725–732.
- Shi, Q., Pan, Y. S., & Li, Y. J. (2005). The typical cases and analysis of rockburst in China. *Coal Mining Technology*, 2, 13–17.
- Sousa, L. R., Miranda, T., Sousa, R. L., & Tinoco, J. (2017). The use of data mining techniques in rockburst risk assessment. *Engineering*, 3(4), 552–558.
- Sun, Y., Li, G., & Zhang, J. (2020a). Developing hybrid machine learning models for estimating the unconfined compressive strength of jet grouting composite: A comparative study. *Applied Sciences*, 10(5), 1612.
- Sun, Y., Li, G., Zhang, N., Chang, Q., Xu, J., & Zhang, J. (2020b). Development of ensemble learning models to evaluate the strength of coal-grout materials. *International Journal of Mining Science and Technology*. <https://doi.org/10.1016/j.ijmst.2020.09.002>.
- Sun, Y., Li, G., Zhang, J., & Qian, D. (2019a). Prediction of the strength of rubberized concrete by an evolved random forest model. *Advances in Civil Engineering*, 2019(3), 1–7.
- Sun, J., Wang, L. G., Zhang, H. L., & Shen, Y. F. (2009). Application of fuzzy neural network in predicting the risk of rock burst. *Procedia Earth and Planetary Science*, 1(1), 536–543.
- Sun, Y., Zhang, J., Li, G., Ma, G., Huang, Y., Sun, J., et al. (2019b). Determination of Young's modulus of jet grouted coalcretes using an intelligent model. *Engineering Geology*, 252, 43–53.
- Sun, Y., Zhang, J., Li, G., Wang, Y., Sun, J., & Jiang, C. (2019c). Optimized neural network using beetle antennae search for predicting the unconfined compressive strength of jet grouting coalcretes. *International Journal for Numerical and Analytical Methods in Geomechanics*. <https://doi.org/10.1002/nag.2891>.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259.
- Wu, S., Wu, Z., & Zhang, C. (2019). Rock burst prediction probability model based on case analysis. *Tunnelling and Underground Space Technology*, 93, 103069.
- Xu, C., Liu, X. L., Wang, E. Z., Zheng, Y. L., & Wang, S. J. (2018). Rockburst prediction and classification based on the ideal-point method of information theory. *Tunnelling and Underground Space Technology*, 81, 382–390.
- Xue, Y., Bai, C., Qiu, D., Kong, F., & Li, Z. (2020). Predicting rockburst with database using particle swarm optimization and extreme learning machine. *Tunnelling and Underground Space Technology*, 98, 103287.
- Zhang, Q., Wang, E., Feng, X., Niu, Y., Ali, M., Lin, S., et al. (2020a). Rockburst risk analysis during high-hard roof breaking in deep mines. *Natural Resources Research*, 29, 4085–4101.
- Zhang, J., Wang, Y., Sun, Y., & Li, G. (2020b). Strength of ensemble learning in multiclass classification of rockburst intensity. *International Journal for Numerical and Analytical Methods in Geomechanics*. <https://doi.org/10.1002/nag.3111>.
- Zhou, J., Koopialipour, M., Li, E., & Armaghani, D. J. (2020). Prediction of rockburst risk in underground projects developing a neuro-bee intelligent system. *Bulletin of Engineering Geology and the Environment*, 79, 4265–4279.
- Zhou, J., Li, X. B., & Mitri, H. S. (2016a). Classification of rockburst in underground projects: Comparison of ten supervised learning methods. *Journal of Computing in Civil Engineering*, 30(5), 04016003.
- Zhou, J., Li, X. B., & Mitri, H. S. (2018). Evaluation method of rockburst: State-of-the-art literature review. *Tunnelling and Underground Space Technology*, 81, 632–659.
- Zhou, J., Li, X. B., & Shi, X. Z. (2012). Long-term prediction model of rockburst in underground openings using heuristic algorithms and support vector machines. *Safety Science*, 50(4), 629–644.
- Zhou, K. P., Lin, Y., Deng, H. W., Li, J. L., & Liu, C. J. (2016b). Prediction of rock burst classification using cloud model with entropy weight. *Transactions of Nonferrous Metals Society of China*, 26(7), 1995–2002.