Journal of Engineering Innovations & Technology

journal homepage: www.journal-eit.online

Research article

The prediction of deep coal mining based on grey prediction

Zhaowei Shen[1*]

Tibet University

17604891434@163.com

*Corresponding author

ABSTRACT

With increasing coal mining depth, the likelihood of rock bursts has significantly risen, posing a major threat to coal mine safety in China. This paper aims to develop classification and prediction models to identify and predict rock burst precursor signals, thereby mitigating this hazard in deep mining. Using acoustic emission (AE) and electromagnetic radiation (EMR) data, we extracted time-frequency domain features and constructed decision tree and grey prediction models with a sliding window approach. For interference signal identification, we analyzed class C signals, determining their mean, kurtosis, and spectral peak values. For predicting class B signal trends, similar features were used to construct a decision tree model, which successfully identified precursor signals. The model demonstrated excellent classification performance in ROC curve tests. This research provides a scientific basis for preventing and controlling rock bursts, reducing engineering efforts, and enhancing coal mine safety.

Key words: decision tree model, discrete type Fourier transform, sliding window, spectral peak, kurtosis, ROC curve, gray prediction model, wavelet transform

I. Background

In view of the characteristics of China's energy resources are "rich in coal, lack of oil and gas", coal resources have always been the main source of energy supply in China. It plays a key role in many fields, such as coal-fired power generation, industrial production and chemical raw materials. In 2020 or so, coal-fired power generation will account for 52 percent of the country's total power generation. It is expected that coal will remain the main force of China's energy supply in the foreseeable future. However, the safety problems in the process of coal mining cannot be ignored, especially the sudden formation pressure phenomenon of rock burst. The occurrence of rock burst is often accompanied by the sudden destruction of rock mass, the instantaneous release of a large amount of elastic potential energy, the huge sound and vibration, and sometimes will trigger a large injection of rock or ore. All of these pose a serious threat to the safety of mine production. At present, the rock burst has become one of the most serious disasters facing China's coal mine production, which may lead to heavy casualties and property losses. Therefore, it is of great significance to predict the occurrence of rock burst in advance or to adopt technical means to protect the safety of people's lives and property and promote the stable development of China's economy. Although it is difficult to directly prevent the occurrence of rock burst by technical

means, it has become the most feasible solution by prediction.

In recent years, researchers have conducted extensive in-depth research in this field. The study revealed that the electromagnetic radiation and acoustic emission data showed some trend precursors for about seven days before the rock burst occurred. Based on the analysis of these precursor features, we can construct mathematical models to predict the danger of the earth burst, and thus mitigate its potential impact on production and life. This research direction not only has the theoretical value, but also has an important practical application prospect.

II. Characteristic analysis and identification of interference signals

In this section, we will explore the characteristics of interference signal data in electromagnetic radiation and acoustic emission.

2.1 Screscreen and visualize the data

For feature extraction of interference signals, we first need to screen out the interference signal data of two signals, the data of type C, and then visualize them and roughly observe their changing trend. As shown in Figure 1:

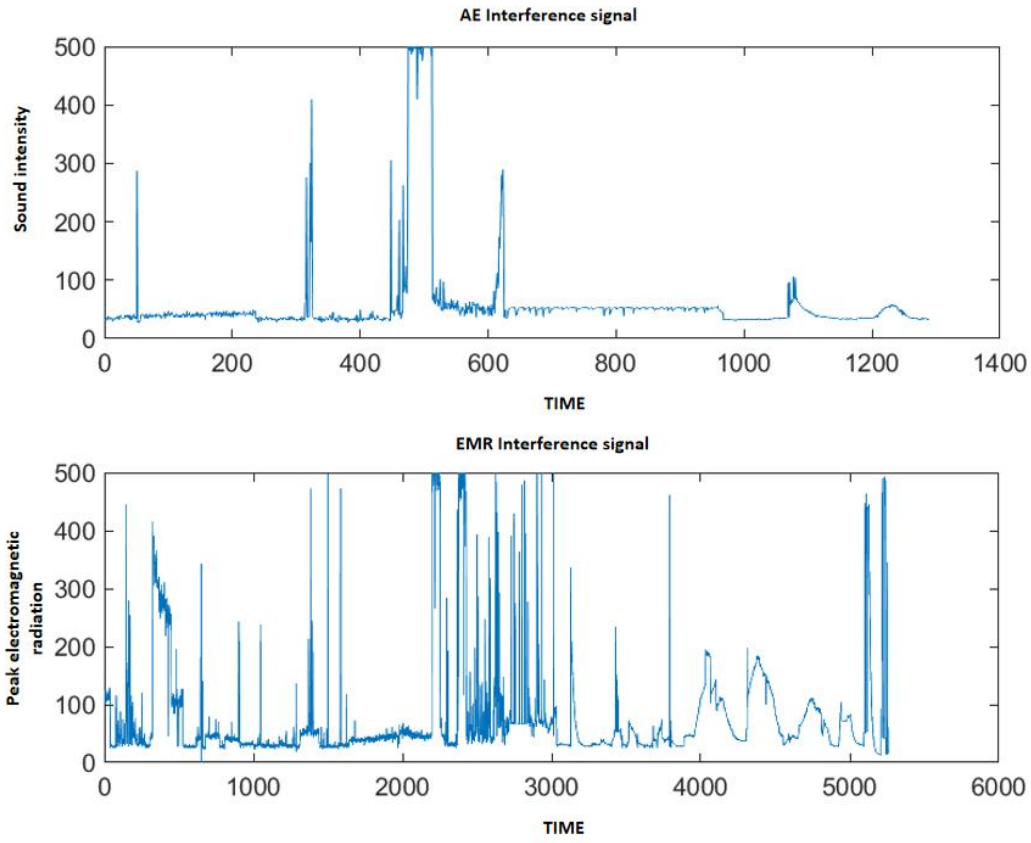

Figure. 1 Visualization of AE interference signals and EMR interference signals

2.2 Analyze the visual data

By analyzing the visualization results of interference signals, it is easy to get the distribution of electromagnetic radiation amplitude and acoustic intensity at different time points. First, the frequency of interference signals shows randomness, and lacks consistent periodicity or interval, increasing the complexity of subsequent relevant algorithms. Secondly, the amplitude change of the interference signal is variable and has no definite period, and its intensity value fluctuates in the range of 0 to 500. Sometimes the amplitude is large, and sometimes it is small. Such significant stochastic fluctuations may be the result of a combination of

multiple interfering factors, with each interfering source having an effect on the signal with its characteristic pattern and strength. In the visual analysis of the data, we noted prominent peaks that may correspond to significant disturbances occurring at a particular time point.

2.3 Features using statistical analysis and time-frequency analysis to extract the signal data

Use statistical analysis to extract time domain features. Related features mainly include mean and peak

(1) Mean value, also known as mean value (Mean), is a statistic that describes the trend in a data set. It adds up all the values in the data set and divides them by the total number of values. For the selected set of interference signal data, the average value is solved as follows: $X = \{x_1, x_2, ..., x_n\}$.

$$A = \frac{1}{n}\sum_{i=1}^{n} x_1 \tag{1}$$

(2) Kurtosis: In statistics, kurtosis (Kurtosis) is a statistic that measures the shape of the data distribution and describes the "sharpness" or "weight of the tail" of the data distribution. The kurtosis is relative to the peak state of the normal distribution (Gaussian distribution), and the kurtosis of the normal distribution is defined as 0. For the selected interference signal data set, the kurtosis K is calculated as:$X = \{x_{g1}, x_{g2}, ..., x_{gn}\}$[1].

$$K_{urosis} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_{gi}-\bar{x}_g\right)^4}{\sigma^4} \quad (2)$$

(3) Bring the values into the calculation. Results are shown in Table 1 and Table 2.

Table 1. Time domain characteristics of the EMR interference signal

feature	A	K
bear fruit	77.9585	12.9982

Table 2. AE time-domain interference signal characteristics

feature	A	K
bear fruit	59.9691	26.0055

Time-frequency analysis method is used to extract the frequency domain features of interference signals.

It is the signal characteristics obtained after converting the time domain (or spatial domain) of the signal into the frequency domain through mathematical transformation. They describe the performance and properties of the signal at different frequencies. These features include the frequency component, amplitude spectrum, phase spectrum, power spectrum, center frequency, the spectral energy of the signal, frequency distribution, and so on. In the frequency domain, the frequency domain characteristics of signals can be analyzed and represented in a variety of ways, among which the most commonly used is the Fourier transform (Fourier Transform, FT). Fourier transform (Fourier Transform) is an

important signal processing tool for converting time domain signal into frequency domain signal. Since the data in Annex 1 and Annex 2 are both discrete, we adopt the discrete type Fourier transform with the following formula:

$$X[m] = \sum_{m=0}^{N-1} x[k] \cdot e^{-j\frac{2\pi}{N}mk} \tag{3}$$

$X[m]$ is a discrete representation on the frequency domain, $x[k]$ is a discrete representation on the time domain, N is the signal length, and k is the frequency sequence number.

First, the original time-domain signal is appropriately preprocessed to reduce noise and improve signal quality. Next, the signal is converted from the time domain to the frequency domain by the Fourier transform to generate a frequency spectrum map. The figure shows the amplitude of the different frequency components. In the spectral map, we can calculate the total spectral energy, as well as analyze the distribution properties of the frequency components. The feature changes before and after the spectrum changes.

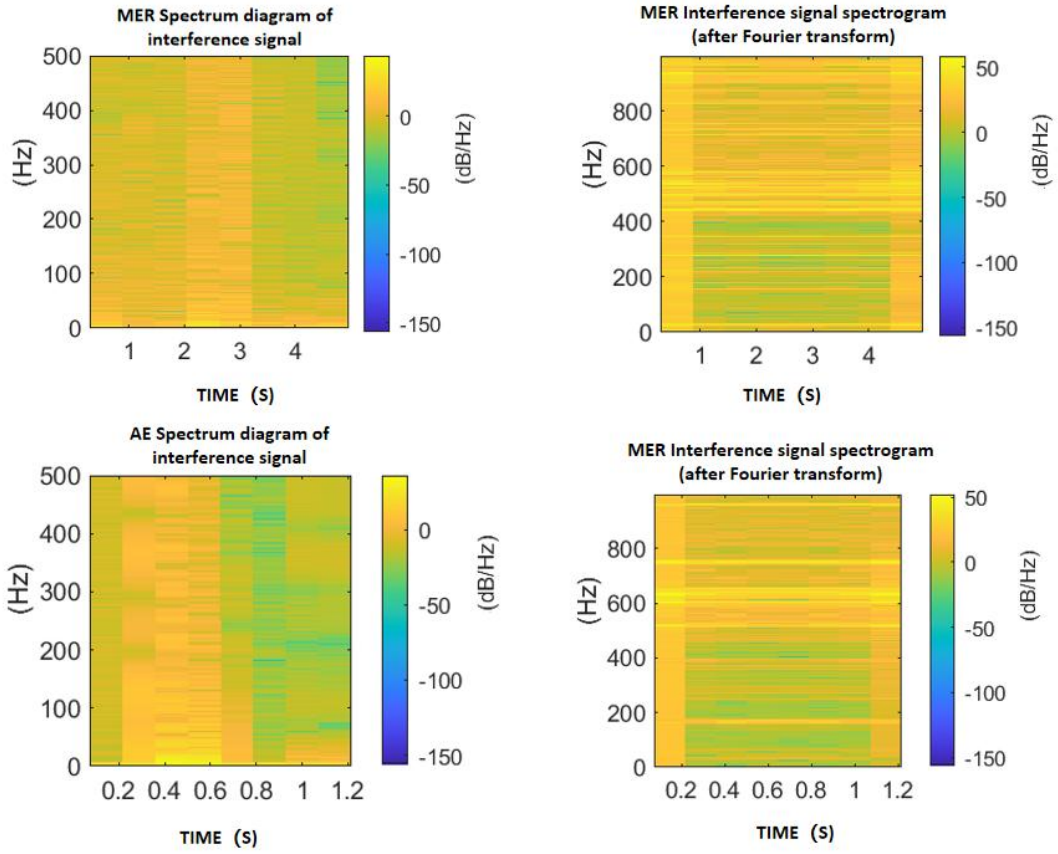

Figure. 2 Spectrum diagram before and after Fourier transform of EMR interference signal and AE interference signal

Spectral peaks refer to the points or multiple points with the largest amplitude on the spectrum map in the frequency domain analysis, which represent the most dominant frequency component in the signal. The specific calculation process is described as follows:

1. Performing a Discrete Fourier Transform (DFT): Formula (6) is applied to the discrete-time signal to obtain a representation of its frequency domain$x[k]X[m]$.
2. Calculated spectral amplitude: For each k, the calculated amplitude, which is obtained by the module:$X[m]X[m]X[m]$.

$$|X[m]| = \sqrt{Re\,(X[m])^2 + Im\,(X[m])^2} \tag{4}$$

3. Calculate the corresponding frequency:

$$f_m = \frac{m}{N} f_s \tag{5}$$

Bring the data into the matlab for the solution, as shown in Table 3.

Table 3 The solution

Time domain characteristics of AE interference signal	Spectrum peak
bear fruit	77300.177
Time domain characteristics of EMR interference signal	Spectrum peak
bear fruit	409983.773

2.4 Summarize these time domain features and frequency domain features

Table 4 Time domain features and frequency domain features

The EMR interference signal signature	A	K	Spectrum peak
bear fruit	77.9585	12.9982	409983.773
The AE interference signal signature	A	K	Spectrum peak
bear fruit	59.9691	26.0055	77300.177

III. Give the interval of the electromagnetic radiation and acoustic emission in the time

3.1 Model building

Sliding window (Sliding Window) is a technology commonly used technique in signal processing, image processing and natural language processing, which is used to extract feature values from data. This approach captures local features by sliding a window over the data and analyzing the data within the window. This approach is particularly suitable for capturing the characteristics of time-varying signals because

it can adapt to local changes in the signal while reducing the computational amount and providing real-time or near-live analysis results.

The method first divides the temporal data into several subsequences and subsequently extracts the data features of the subsequences[3]. The specific steps and the working principle diagram are shown as follows:

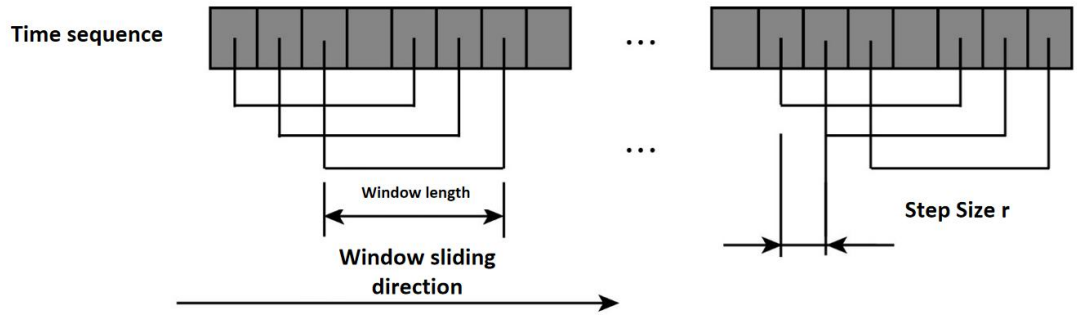

Figure. 3 The working principle diagram

1. The size of the data selection window is 50, and the moving step is 20 (that is, feature extraction in time and frequency domain every 50 data). If the final data does not meet 50, the remaining data will be measured as a group.
2. In order to ensure the correct data, the error rate is set to 0.1.
3. Make the sliding window in, and when the data starting point in Annex 2 starts execution, each source window reads the first value of the corresponding vector.
4. Make all the Windows at the same speed, in sequence, advance a point each time, store the value of the read on the right of the window, remove the value on the original left, and record the source window

and the vector in the corresponding target window after each sliding

5. When the target window reaches the end point of the sequence, the slide ends.

The identification model is established based on the sliding window, which is as follows:

$$judge\ model = \begin{cases} 1(\ Interference\ interval\ ), \\ if(\ any\ (x_i < (1 + \ rate\ ) \cdot \ feature_{\ i}\ and\ x_i > (1 - \ rate\ ) \cdot \ feature_{\ i})) \\ 0\ (Non - interfering\ signal\ interval), \\ if(\ not\ any\ (x_i < (1 + \ rate\ ) \cdot \ feature_{\ i}\ and\ x_i > (1 - \ rate\ ) \cdot \ feature_{\ i})) \end{cases} \tag{6}$$

x_i represents the signal features of each group of interference signals, and $feature_{\ i}$ represents the feature result, and the rate represents the set error rate. After the sliding window feature extraction of the data, the results are visualized as shown in the following figure.

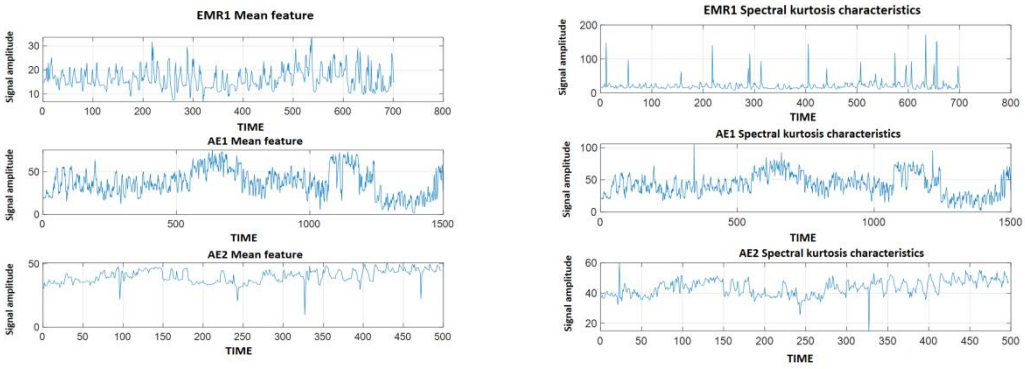

Figure. 4 The visualized results

3.2 Numerical solution

Table 5. Electromagnetic radiation interferes with the signal time interval

order number	Time interval starting point	Time interval end point
1	2022-5-5 1:17:49	2022-5-5 2:25:30
2	2022-5-6 15:22:29	2022-5-6 16:21:05
3	2022-5-12 23:34:25	2022-5-13 0:34:44
4	2022-5-13 3:39:26	2022-5-13 4:37:57
5	2022-5-24 13:04:20	2022-5-24 14:06:32

Table 6. The AE interference signal time period 1 time interval

order number	Time interval starting point	Time interval end point
1	2022-4-2 21:40:01	2022-4-2 22:34:53
2	2022-4-3 15:22:40	2022-4-3 16:10:14
3	2022-4-3 16:10:14	2022-4-3 17:05:05
4	2022-4-3 17:05:05	2022-4-3 17:59:59
5	2022-4-3 17:59:59	2022-4-3 18:51:11

Table 7. The AE interference signal time period 2 time interval

order number	Time interval starting point	Time interval end point
1	2022-10-12 19:12:31	2022-10-12 20:44:00
2	2022-10-12 23:47:00	2022-10-13 1:22:16
3	2022-10-19 0:22:22	2022-10-19 1:53:56
4	2022-10-21 22:29:06	2022-10-22 0:00:32
5	2022-10-23 15:39:01	2022-10-23 17:10:29

IV. Analysis and identification of precursor feature signals

4.1 Model selection

Because the problem makes the trend characteristic, then we can not directly regard the class B signal as a whole. Therefore, it needs to be seen as a process of change. Therefore, we can still build a sliding window model to analyze the trend characteristics of class B signals. The principle is to "slide" through the sliding window, so that the characteristics of the conditional class B signals can also "move" together (namely the trend).

4.2 Model building

The window is still set to be 50 (from that moment, after 49 sets of

data), and then the data characteristics of each group (i. e., the previously specified mean, kurtosis, and spectral peaks). Since I want to analyze the characteristics of class B signal, if I start the characteristics of class B signal at a certain point, and the next 49 sets of data are class B signal, then there is no doubt that the characteristics of this group of data must be the characteristics of class B signal. However, there are also special situations, such as for example, the time point taken at the beginning is not the B class signal, but then the B class signal appears. In order to facilitate calculation, when a set of data is not all class B data, we stipulate that it does not belong to class B signal characteristics. And assuming that a set of data is full of class B signals, then we mark this set as 1. Similarly, assuming that a set of data is all non-type B signals (i. e., signal types A, C, D / E), we mark it as 0. When a set of signals meets our assumptions (i. e., all type B), we record the starting point (i. e., time point) and calculate the characteristics of the data set, and finally "label" the data set. If not satisfied, then directly abandon and measure the next group of data. Treat each required window and thus the feature value as a point (if there are enough data samples), and then wire it to obtain the trend characteristics of its class B signal, as shown in Figure 5.

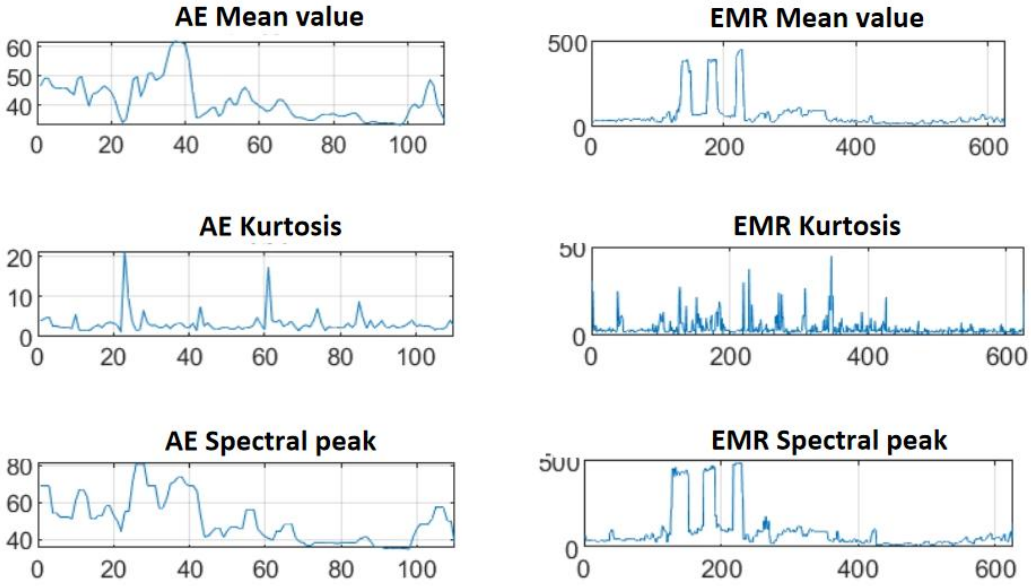

Figure. 5 The trend characteristics of its class B signal

V. Establish a mathematical model, and give the time interval of the first 5 precursors of the electromagnetic radiation and acoustic emission signals respectively

5.1 Model selection

We need to create a classification model to judge. According to the training method, the classification model can be divided into supervised learning and unsupervised learning. Supervised learning involves using datasets with labels to train the model. In the classification task, this means that each training sample has a category label associated with it. The goal of the model is to learn how to predict the correct category labels based on the features of the input data. Unsupervised learning then uses data sets without labels to train the model. In this way of learning, the model needs to discover the structure and patterns in the data. Since

we have previously labeled the data.

5.2 Model building

To meet topic requirements, these selected time periods were targeted using sliding window feature extraction techniques, which allows us to capture dynamically changing trends and patterns from the data. With this approach, we successfully constructed two test set matrices for EMR and a single test set matrix for AE, laying the foundation for further data analysis and pattern recognition.

When the data is filtered, the unbalanced data set is found (i. e., the number of samples in different categories of samples in the data set is very different). Samples of the training set data are imbalanced, resulting in more samples with label 0 and too few samples with label 1. This easily leads to the final identification of almost all zero results. To avoid bias and improve the model performance, we adopt the SMOTE algorithm or random oversampling to balance the sample data. The SMOTE algorithm uses the interpolation strategy to synthesize the minority samples and increase the probability of the minority data. This algorithm can overcome the imbalance of the sample set.

Random oversampling is a data preprocessing technology that increases the number of samples by randomly copying a few categories, making the number of samples more consistent. This problem uses the random oversampling method

To balance the weights of individual features in the model and eliminate the effects of different dimensions, we implemented data normalization processing. The normalisation method chosen is minimum-maximum normalization, a widely adopted technique that maps the raw data within the interval of [0,1] by linear transformation. The formula is as follows[4]:

$$v^{'} = new\_\min_{M} + \frac{v - \min_{M}}{\max_{M} - \min_{M}} \tag{7}$$

In short, the sample feature values are entered, and then the model gives the output (i. e., the label). The decision tree serves as a tree, the root node of the tree is the entire data set space, and each branch node is a test of a single variable, which divides the data set space into two or more blocks. Each leaf node is a record belonging to a single category. Second, the first step is repeated until the records within each leaf node belong to the same class, growing to a complete tree[5]. When constructing the decision tree model, the most commonly used algorithm is the C4.5 algorithm, whose split properties are mainly selected by the information gain rate[6].

The training process is actually a constant iteration of the parameters. Constantly adjust the weight of each input variable. The previously generated electromagnetic radiation (EMR) and acoustic emission (AE) datasets were used as the basis for the training set. Therefore, in order to

train and verify the accuracy of the model, the dataset is divided in a 7:3 ratio, that is, 70% of the total data volume is used to build the decision tree model and learn its classification weight, while the remaining 30% is used in the test stage to evaluate the classification ability of the model. After completing the model training and testing, a series of solution results of the decision tree model are obtained and presented by visual images, as shown in Figure 6 and Figure 7.

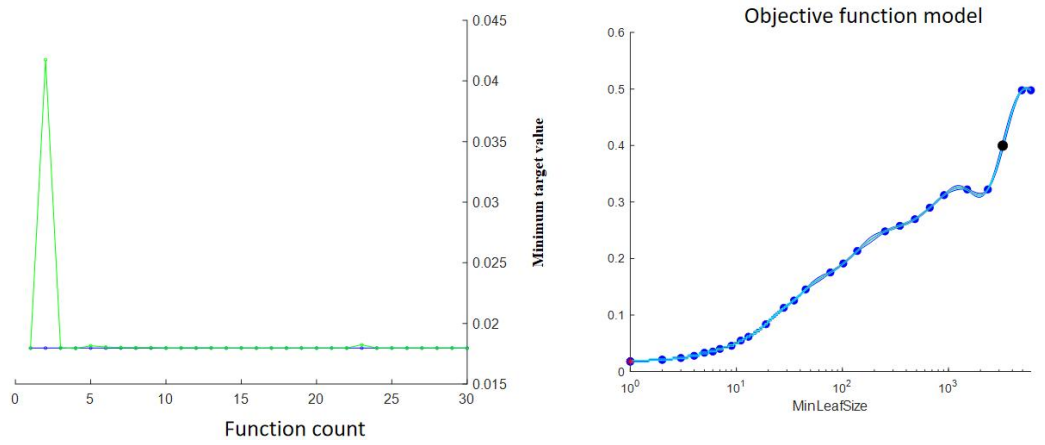

Figure. 6 The EMR decision tree model optimizes the hyperparameter iteration process

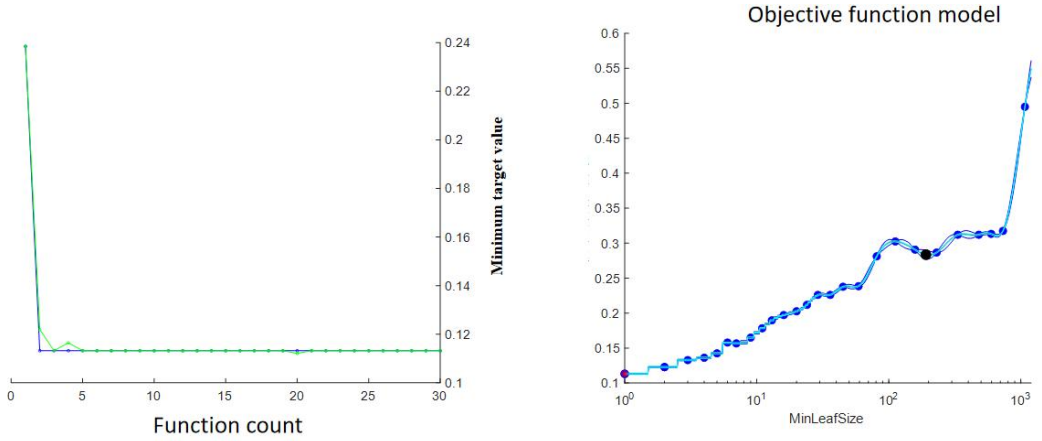

Figure. 7 AE decision tree model optimizes the hyperparameter iteration process

The ROC curve is a graphical tool used to evaluate the performance of a dichotomized model. Is a graphical tool for evaluating the performance of the classification models. It demonstrates the

performance of the model by plotting the true and false positive case rates at different thresholds. The ROC curves provide a way to compare the performance of the different models. In general, the curve near the upper left corner indicates better classification performance. The area below the ROC curve is called the AUC, the integral under the ROC curve. An AUC value of 1 indicates a perfect classifier. An AUC value greater than 0.5 usually indicates that the model has good classification performance. At an AUC value of 0.5, the model performs equivalent to random guessing. A higher AUC value indicates the better performance of the model, and usually the AUC values range between 0.5 and 1.

The ROC curves were visualized using matlab, as shown in Figure 8.

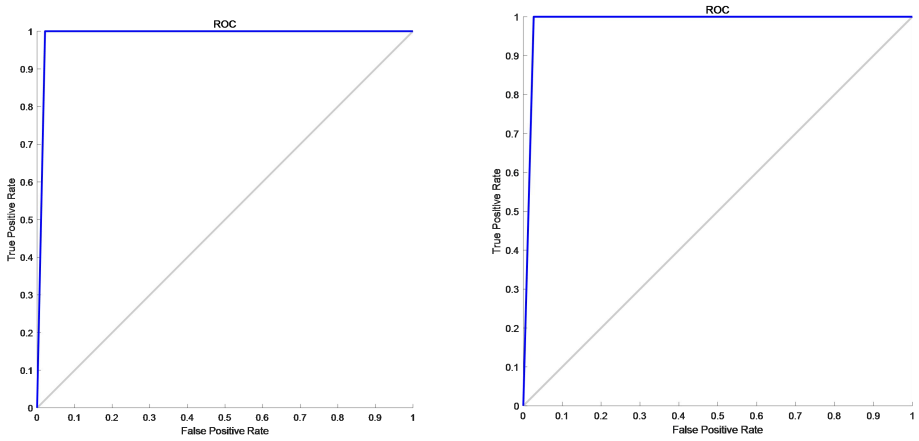

Figure. 8 The ROC curve

According to the visualization results, both the ROC curve of EMR and the ROC curve of AE are close to the upper left corner, thus indicating that the model has good classification performance for the decision tree we constructed

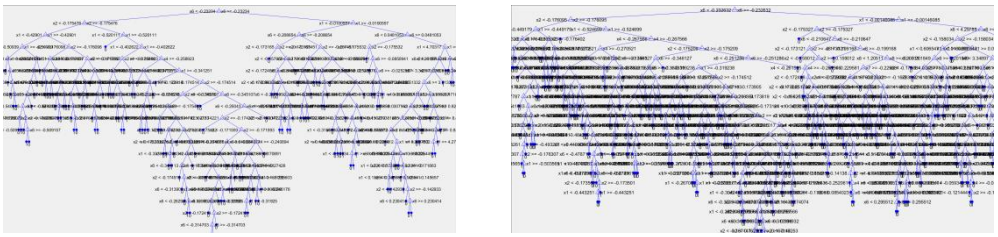

Figure. 9 EMR decision tree model Structure AE decision tree model structure

After the whole process of model training, we used the resulting decision tree model to make classification prediction on the test set of three time periods, and assigned corresponding classification labels to each sample in the test set. In a detailed analysis of the prediction results, we specifically selected all the samples predicted to be positive. Among these positive class samples, we further identified the top five samples. Below are details of the five samples, which were collated into tabular form to facilitate further scrutiny and possible in-depth analysis.

Table 8. Electromagnetic radiation precursor characteristic time period 2 time interval

order number	Time interval starting point	Time interval end point
1	2020/4/10 6:07:13	2020/4/10 8:29:34
2	2020/4/11 2:14:44	2020/4/11 3:02:11
3	2020/4/11 6:15:50	2020/4/11 7:03:17
4	2020/4/11 8:41:05	2020/4/11 9:28:31
5	2020/4/11 20:32:52	2020/4/11 21:20:18

Table 9. EMradiation precursor characteristic time period 3 time interval

order number	Time interval starting point	Time interval end point
1	2021/11/25 8:03:42	2021/11/25 8:52:36
2	2021/11/30 21:19:28	2021/11/30 22:06:33
3	2021/12/2 9:26:18	2021/12/2 11:27:40
4	2021/12/2 18:20:49	2021/12/2 19:09:44
5	2021/12/4 15:03:35	2021/12/4 16:21:29

Table 10. The AE precursor characteristic time interval

order number	Time interval starting point	Time interval end point	order number
1	2021/11/2 5:24:25	2021/11/2 6:24:12	1
2	2021/11/2 7:27:36	2021/11/2 8:27:23	2
3	2021/11/2 9:30:51	2021/11/2 10:32:29	3
4	2021/11/2 11:35:54	2021/11/2 12:35:41	4
5	2021/11/2 21:55:45	2021/11/2 22:55:32	5

VI. Conclusion

This study addresses the critical issue of predicting rock bursts in deep coal mining, which pose significant threats to mine safety. By leveraging grey prediction and decision tree models, the research successfully identifies precursor signals using acoustic emission (AE) and electromagnetic radiation (EMR) data. Key contributions include: Signal Characterization: Detailed time-frequency domain analyses were performed to extract mean, kurtosis, and spectral peak features from AE and EMR data. These features were critical in understanding the interference signals and their precursors.Sliding Window Approach: The sliding window method effectively captured dynamic signal changes, enabling real-time analysis and prediction of rock bursts. The selected window size and error rate optimized the detection of interference signals.Decision Tree Model: A robust decision tree model was developed and validated using ROC curves, demonstrating excellent classification performance. This model accurately identified the time intervals of precursor signals, offering a reliable tool for early warning.Practical Implications: The research provides a scientific basis for early prediction

and control of rock bursts, potentially reducing engineering efforts and improving coal mine safety. The predictive models can be integrated into existing monitoring systems to enhance proactive risk management.In conclusion, this study advances the methodology for predicting rock bursts in deep coal mining, offering practical tools for enhancing safety and stability in mining operations. Future work could focus on refining these models and exploring their application in different mining environments.

References

[1] Chen Xingang, Yang Dingkun, Tan Hao, etc. Analysis of Raman spectroscopy of the characteristic gas in transformer oil based on kurtosis [J]. High-voltage technology, 2017,43(07):2256-2262.DOI:10.13336/j.1003-6520.hve.20170628022.

[2] Liu Xiaohong, Gong Ruichun, Tong Xiaomei. Research on the approximate analysis of continuous signal frequency spectrum by using discrete Fourier transform [J]. Computer CD software and application, 2014,17 (21): 137-138.

[3] Tian Teng, Shi Maolin, Song Xuegong, et al. Time-series anomaly detection method based on sliding windows [J]. Instrument Technology and Sensors, 2021, (07): 112-116.

[4] CAI Weiling, Chen Dongxia. The effect of the data normalization method on the K-nearest neighbor classifier [J]. Computer Engineering, 2010,36 (22): 175-177.

[5] Yang Xuebing, Zhang Jun. The decision tree algorithm and its core techniques [J]. Computer Technology and Development, 2007, (01): 43-45.

[6] Luo Jia, Li Mingming. Design and application of student employment prediction model based on decision tree algorithm [J]. Integrated circuit applications, and 2023,40(10):62-64.DOI:10.19339/j.issn.1674-2583.2023.10.024.

[7] Wang Qinghe, Wang Qingshan. Several commonly used digital filtering algorithms in data processing [J]. Metering Technology, 2003, (04): 53-54.

[8] Xie Jiecheng, Zhang Dali, Xu Wenli. Summary of wavelet image denoising [J]. Chinese Journal of Image and graphics, 2002 (03): 3-11.

[9] Zhao Ziyi, Hao Zhongqi, Lu Ying, et al. Improvement of the stability of laser-induced breakdown spectrum detection by two-dimensional wavelet denoising [J / OL]. Journal of Optics, 1-14 [2024-05-03]..

[10] Zhang Jun. Improvement of the grey prediction model and its application [D]. Xi'an University of Technology, 2008.

[11] Yang Hualong, Liu Jinxia, Zheng Bin. Improvement and application of grey prediction GM (1,1) model [J]. The Practice and Awareness of Mathematics, 2011,41 (23): 39-46

[12] Zheng Jian, Chen Bin. GM (1,1) based on time weight series [J]. Control and decision-making, 2018,33(03):529-534.DOI:10.13195/j.kzyjc.2017.0033.