Interference and precursor signal analysis based on random forest and improved logistic regression model

Dongping Sheng, Jie Yang, Hongliang Wang, Chun Su*

Changzhou Institute of Technology, Changzhou, China

ABSTRACT

Coal is the main energy and industrial raw material in China. In order to prevent the risks of coal mining and ensure the safety and efficiency of coal mining, this paper established random forest model, ADF and MK test model and logistic regression model, and used auto correlation function algorithm, Fourier transform and sliding window transformation of time series data to analyze and optimize the characteristics of interference signals and precursor characteristics. For the first work, the huge data is preprocessed, tested, and outliers are removed, and the time characteristics are taken to aggregate and classify the data. For work (1.1), data parameters were sorted out, Fourier transform prediction signal model was initially established, and characteristic values were extracted by wavelet transform. Finally, autocorrelation function algorithm was used to optimize and analyze the results. The final results were shown in Table 4 for the data characteristics of electromagnetic radiation (EMR) interference signals and Table 5 for the interference signals of acoustic propagation signals. For work (1.2), a random forest model is established, with sliding window transformation of time series data, window size is specified, boundary processing mode is specified, and label value is defined as interference and normal. The data set is divided into training set and test set in the way of eighty-two allocation. After AE training, there are three stages of training. Finally, random forest algorithm is used to calculate the electromagnetic radiation interference signal interval. For work (2.1), the model and algorithm of work 1 are used to identify the trend characteristics of the data before the occurrence of electromagnetic radiation and acoustic emission signals, and the signal data are judged to have a "slightly rising" trend. The statistical values are compared by KS trend test, and the trend characteristics of normal and precursor signals are calculated by using the autocorrelation function algorithm. A trend feature table is obtained for the precursory feature data of electromagnetic radiation and acoustic propagation signal. For work (2.2), Augmented Dickey-Fuller and MK test models of the system were established. For work 3, a logistic regression model is established to predict the probability of precursor feature data appearing at the last moment of multiple time periods. By using maximum likelihood estimation to train model parameters, the characteristics of the last data collection moment of each time period are predicted, and the probability of precursor feature appearing at each moment is output. As shown in Table 14, the probability of precursor features appearing at the time when the data of multi-classification logistic regression is located is obtained.

Keywords: Random forest, ADF and MK test, Autocorrelation function, Logistic regression, Sigmoid function

1 INTRODUCTION

In order to ensure the life safety of coal miners and the stable development of enterprises, it is necessary to deepen the research on rock burst and improve the accuracy and effectiveness

of monitoring and warning [1]. In addition, it is also necessary to strengthen international cooperation and exchanges, learn from international advanced experience and technical means, and jointly cope with the challenges brought by rock burst disasters [2]. In the electromagnetic radiation and acoustic emission signal data of the field working face, there are interference signals caused by other operations or equipment of the working face [3]. These interferences have an adverse effect on the subsequent signal processing analysis. Using the given data attachment, a mathematical model is first established to analyze the characteristics of interference signals in electromagnetic radiation and acoustic emission, and at least three characteristics are identified [4]. Based on these features, a model is built to identify the time interval of electromagnetic radiation signals from May 1 to May 30, 2022, and the interference signals in acoustic emission signals from April 1 to May 30, 2022, and from October 10 to November 10, 2022 [5]. Finally, the time interval of the first five interference signals in electromagnetic radiation and acoustic emission is given. About 7 days before the danger of rock burst, the electromagnetic radiation and acoustic emission signal will increase with time cycle, which is the precursor characteristic signal. Rock burst may occur within 7 days of the signal, so preventive measures should be taken [6].

2 RELATED WORK AND ASSUMPTION

2.1 Work one

In work 1 (1.1), electromagnetic radiation and acoustic emission signals with interference are analyzed, and the characteristics of interference signal data in electromagnetic radiation and acoustic emission are given respectively. the data is preprocessed, including data cleaning, aggregation, classification, etc., and the statistical parameters of the mean, median and other signals are calculated by the algorithm [7]. In addition to these general statistical parameters, Fourier transform and wavelet transform are also used to solve the practical works of the signal. When Fourier transform is used, the base signal of various frequencies is compared with the unknown signal to find the base signal with the highest similarity to the position model. By using wavelet transform, the infinite trigonometric function base signal in Fourier transform is replaced by a finite wavelet base that will decay. Then the results obtained by these two methods are compared and the eigenvalue of the interference signal is obtained. In work 1 (1.2), the feature obtained in the previous work is required to be used [8]. Random forest model is used to construct a decision tree for each sub-data set, and the features sought above are taken as a feature subset to determine the best splitting rule until the base signal that best fits the target interference signal is split.

2.2 Work two

In the second work, the work requires the analysis of the precursor characteristic signals in the electromagnetic radiation and acoustic emission signals, focusing on the change trend of the signals, and respectively giving the trend characteristics of the electromagnetic radiation and acoustic emission signals before the danger occurs [9]. To solve this work, we still follow the model and algorithm in work 1.1, and find out the characteristics of interference signals through comparison [10]. The second work is to identify the precursory characteristics of the electromagnetic radiation and acoustic emission signals in a certain period of time, and give

the time interval where the first 5 precursory characteristic signals of electromagnetic radiation and acoustic emission signals are located.

2.3 Work three

Work 3 requires the identification of precursor characteristic signals and the prediction of which are interference characteristic signals, which is actually the solution of the probability of danger occurrence [11]. A logistic regression model is established. The features of the interference signal in input work 1 are weighted and the Sigmoid function is used to transform this linear combination, so that its output value is between (0,1) [12]. The closer it is to 1, the greater the probability of danger, and the smaller it is otherwise. Then the model of work 1 and work 2 was used to verify the results, and a perceptron neural network model with Sigmoid activation function was adopted.

2.4 Model assumption

(1) When using the sliding window conversion algorithm of time series data, it is assumed that the data given by the work has a relative degree of stationarity;

(2) Suppose that when sliding window conversion of time series data is used, the observed values in the window are assumed to be independent from each other and from the same distribution when the window is small;

(3) To facilitate the extraction of useful features, it is assumed that there is a linear relationship between the data points in the window.

3 MODEL BUILDING AND ANALYZING

3.1 Model establishing and solving for work one

3.1.1 Data preprocessing and visual display

The success rate of the prediction model depends on the quality and quantity of past electromagnetic radiation and sound signal data. Due to the influence of factors such as interference from other operations or equipment on the working face, noise errors may occur, and the fluctuation error of sound data is large. Outliers need to be removed and the data plotted into signal graphs, as shown in Fig.1 and Fig.2. For solving work one, if these data are directly used for prediction, it will lead to a decrease in the success rate of prediction, and even the prediction results will completely deviate from the actual values, and the relevant characteristic values are shown in Tables 1 and 2.

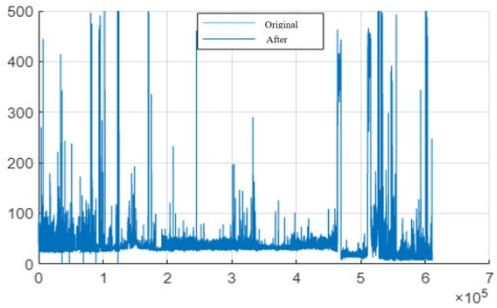

Fig.1: Cleaning data of EMR

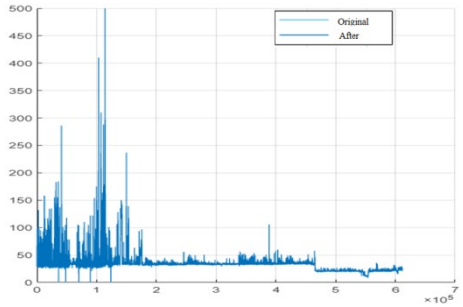

Fig.2: Cleaning data of AE

Table 1: Relevant characteristic values of electromagnetic radiation (EMR)

Type	Minimum value	Maximum value	Mean value	Median	Mode	Standard deviation
Value	0	500	43.9202	31.895	29	60.4712

Table 2: Relevant characteristic values of acoustic emission signals (AE)

Type	Minimum value	Maximum value	Mean value	Median	Mode	Standard deviation
Value	0	500	33.1055	33.49	34	13.601

3.1.2 Data aggregation and classification

Due to the fact that the data is collected by electromagnetic radiation and acoustic emission sensors every 30 seconds, and the data is large and mixed in types, in order to better distinguish the characteristics of different types of electromagnetic radiation and acoustic emission signals, it is necessary to use the aggregation model of machine learning techniques to generate different subsets through data resampling. In this work, taking time characteristics as an example, the data is aggregated every minute, and then the maximum value of sound waves within one minute is used. Fig.3 to Fig.6 show the time series data of D-E category electromagnetic radiation, A category electromagnetic radiation, B category electromagnetic radiation, and C category electromagnetic radiation after classification; Fig.7 to Fig.10 show the time series data of classified D-E class acoustic emission signals, A class acoustic emission signals, B class acoustic emission signals, and C class acoustic emission signals, respectively.

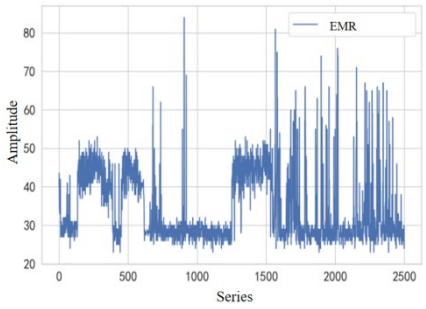

Fig.3: Time series data of D-E category electromagnetic radiation (EMR)

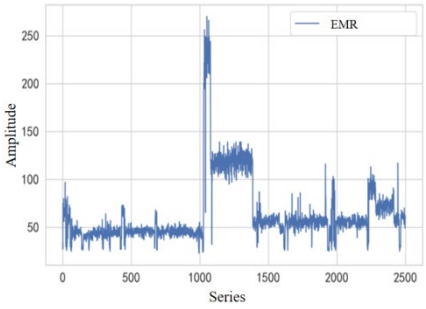

Fig.4: Time series data of electromagnetic radiation (EMR) in category A

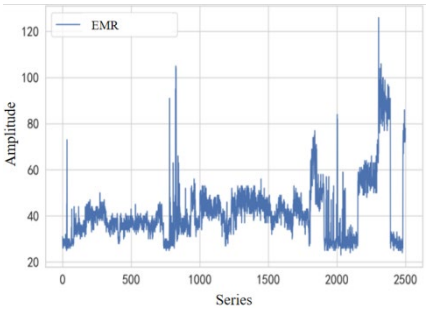

Fig.5: Time series data of Class B electromagnetic radiation (EMR)

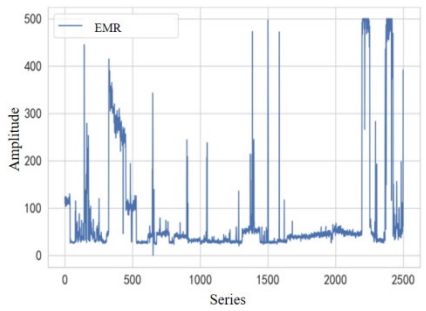

Fig.6: Time series data of Class C electromagnetic radiation (EMR)

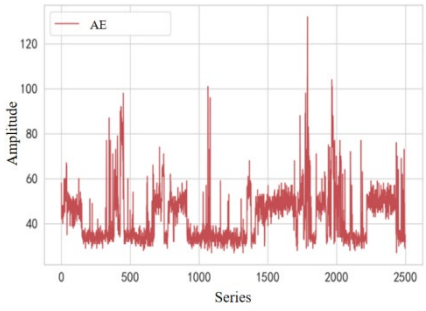

Fig.7: Time series data of D-E category acoustic emission signals (AE)

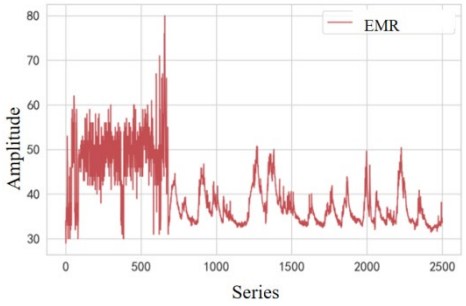

Fig.8: Time series data of Class A acoustic emission signals (AE)

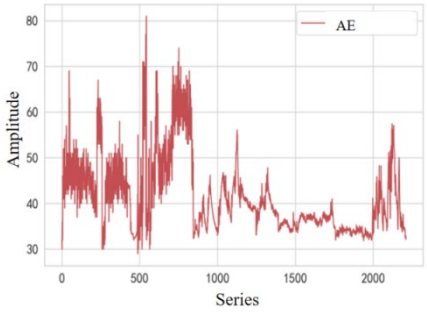

Fig.9: Time series data of Class B acoustic emission signals (AE)

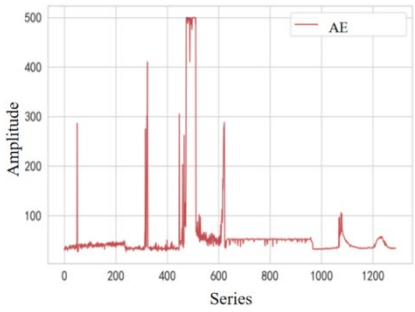

Fig.10: Class C Acoustic Emission Signal

Through data visualization, the goal is to analyze signal datasets of categories A, B, and C, which are sourced from conventional sensor collection. Different intervals have been manually classified as "normal without interference", "precursor stage", and "interference state". The core challenge is to design and extract effective signal features for those time periods marked as non-interfering states.

3.1.3 Data analysis

Comparing the statistical values of normal signal data and interference signal data, a simple feature can be obtained through preliminary analysis, which is that the interference data shows higher variance and mean, as shown in Table 3. Using the autocorrelation function algorithm to analyze and compare the differences between interference and normal data, the specific analysis results are shown in Fig.11 to Fig.12.

Table 3: Comparison and Analysis of Normal Signal and Interference Signal Data

Item	Normal EMR	Disturbed EMR	Normal AE	Disturbed AE
Count	73418.000000	5259.000000	15587.000000	1289.000000
Mean	49.815421	77.958504	37.432989	59.969106
Std	18.279324	90.715876	3.727368	80.421827
Min	9.610000	0.000000	29.000000	27.000000
Max	270.000000	500.000000	80.000000	500.000000

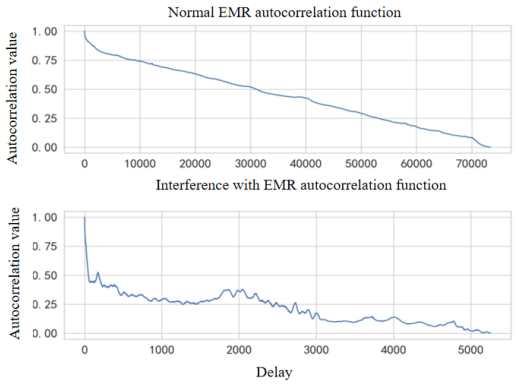

Fig.11: Comparison of Normal EMR and Interference Auto correlation Functions

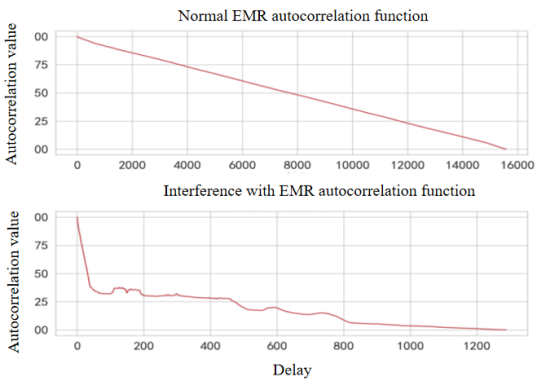

Fig.12: Comparison of Autocorrelation Functions between Normal AE and Interference AE

It could be found that the auto correlation of interference signals exhibits unstable characteristics, specifically reflected in the large fluctuations in the auto correlation function. In terms of its characteristics, the interference signal exhibits high intensity variability and large variance. At the same time, its auto correlation function shows a clear attenuation fluctuation trend, further highlighting the irregularity of these random fluctuations.

3.1.4 Model establishing and solving for work 1.1

3.1.4.1 Statistical parameter characteristics

Establish a mathematical model, divide the preprocessed data into segments, and calculate the statistical parameters of each segment, including mean, variance, skewness, and kurtosis. Compare the original data with the parameters, study the similarity between the original data and the parameters, and treat the group of data with the worst similarity as a signal strength mutation. After comparison, when the signal strength is the sum of the mean and double standard deviation, similarity begins to undergo a sharp turn. Therefore, this data is used as a threshold to determine whether the signal has undergone a mutation. The calculation formulas for variance, skewness, and graduation are as follows:

$$\sigma^2 = \sum(x_i - \overline{x})^2 \tag{1}$$

$$S = \frac{\mu^3}{\sigma^3} \tag{2}$$

$$k = \frac{\mu^4}{\sigma^4} - 3 \tag{3}$$

Where, x_i represents the i^{th} group of sound wave intensity values, $\overline{x}$ is the mean of sound wave intensity data, μ is the center distance, and the calculation formula is

$$\mu = \frac{2}{i}\left[\frac{X_i - E(x)}{1+i}\right] \tag{4}$$

3.1.4.2 Fourier transform

In addition to using statistical parameter comparison methods, Fourier transform can also be used for this work. Fourier transform is a mathematical tool widely used in signal processing, image processing, physics, and other fields. It converts a signal from time or space domain to

frequency domain, thereby revealing the frequency composition of the signal. When using Fourier transform for signal comparison, it usually refers to analyzing the similarity or difference between two signals by comparing their spectra. Here is a spectrum of a 1s long, 5Hz frequency sine wave signal, as shown in Fig.13.

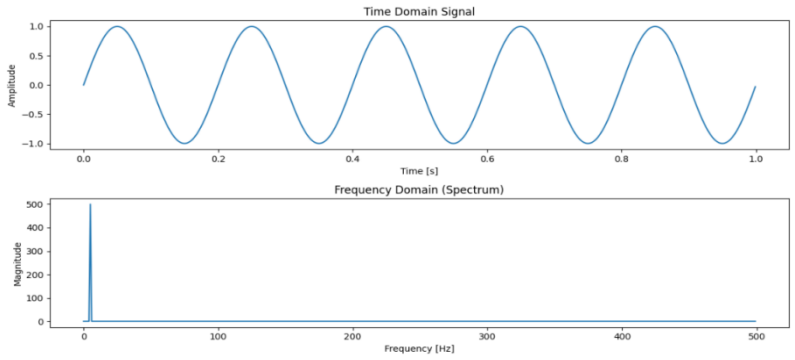

Fig.13: Fourier Transform Spectrum

In this work, compare the base signals of various frequencies with the unknown signal to find the base signal with the highest similarity to the unknown signal. Based on this method, mathematical modeling is carried out to filter out various frequency components in unknown signals. The basic formula for Fourier transforms of continuous time non periodic signals is given below:

$$e^{i\pi} + 1 = 0 \tag{5}$$

$$F(w) = \int_{-\infty}^{\infty} f\,(t) \cdot e^{-iwt} dt \tag{6}$$

$f(t)$ is an unknown spectrum signal; $F(w)$ is the frequency spectrum function after Fourier transform, with the independent variable being the ω frequency. e^{-iwt} is a complex signal.

3.1.4.3 Wavelet transform

When using Fourier transform, it is found that an infinitely long trigonometric function is used as the basis function, which is essentially a decomposition of two orthogonal bases. This basis function is continuously multiplied by the signal to obtain a maximum value, which is the correlation between the signal and the basis function. And here, using wavelet transform, the infinitely long trigonometric basis is transformed into a finite decaying finite basis. Then, by comparing the wavelet coefficients of two signals at the same scale, the similarity between the signals is evaluated by calculating the correlation coefficient. The reference formula and correlation coefficient formula for wavelet transform are as follows:

$$WT(a,t) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f\,(t) \cdot \psi^{(t-\tau)} dt \tag{7}$$

$$r = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{8}$$

Where, y_i is the numerical value of the i^{th} group of base signals, and $\overline{y}$ is the mean of the base signal data.

3.1.4.4 Model optimization to obtain features

When using Fourier transform and wavelet transform, interference signals may cause sudden increase or decrease in signal amplitude, i.e. signal mutation; During signal comparison, there may be distortion of certain wave forms or phases, namely phase and waveform distortion; Interference signals may be concentrated at a certain time frequency, that is, energy concentration. Auto-correlation functions can be used to identify and understand the non-stationarity of sound wave intensity (AE) and time series, thus extracting the above five features as interference signal data, as shown in Tables 4 and Table 5.

Table 4: Characteristics of Electromagnetic Radiation (EMR) Interference Signal Data

Characteristics of Electromagnetic Radiation (EMR) Interference Signal Data
1. Presenting higher variance
2. The value of the interference signal is higher than normal data at 25% after the time-frequency
3. The maximum value of interference signal data is higher than normal data, and the value shows a high and unstable state
4. The autocorrelation function shows a fluctuating downward trend

Table 5: Data characteristics of acoustic emission signal (AE) interference signal

Data characteristics of acoustic emission signal (AE) interference signal
1. Presenting higher variance and mean
2. The interference signal has a value higher than normal data at 50% of the time frequency interval
3. The data values of the interference signal exhibit a high and unstable state
4. The autocorrelation function shows a fluctuating downward trend, and the overall delay is small

3.1.5 Model establishing and solving for work 1.2

The random forest prediction model is an ensemble learning method that integrates multiple decision trees for prediction to achieve higher accuracy and stability. Its main advantage lies in its ability to process a large number of features and evaluate the importance of each feature, as well as handle non-linear relationships and high-dimensional data. Firstly, collect the electromagnetic radiation from May 1, 2022 to May 30, 2022, as well as the time interval of interference signals in the acoustic emission signals from April 1, 2022 to May 30, 2022 and October 10, 2022 to November 10, 2022, as data. Then, preprocess the missing and abnormal values, perform data type conversion, and other data cleaning tasks. Then, based on the five characteristics of the interference signal data extracted in work 1.1 as variables, divide the dataset into training and testing sets, with data allocation of 80% and 20%. Next, the model is constructed and trained, and a random forest classifier is created using the fit ensemble function. The number of trees, sample sampling ratio, and feature sampling ratio are specified, including signal mutation, phase and waveform distortion, energy concentration, periodicity, baseline drift, and non-stationarity. Afterwards, the random forest model is trained using the training dataset consisting of features and target variables. Next, predict and evaluate the test set, and evaluate the model performance. The performance of the model on the test set is mainly evaluated using indicators such as accuracy, recall, F1 score, mean square error (MSE), and coefficient of determination R. The formula for solving the mean square error (MES) and the coefficient of determination R^2 is given as follows:

$$MES = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (9)$$

$$R^2 = \frac{SSR}{SST} \quad (10)$$

Where, SSR represents the sum of squares of the difference between the predicted value of the interference signal data in the model and its mean predicted value, and SST represents the sum of squares of the difference between the observed value and its mean. The value of R^2 ranges from 0 to 1. The closer it is to 1, the higher the model's fit to the data, indicating that the model can better identify interference signal data. Conversely, the closer it is to 0, the less the model can recognize signal anomalies.

Finally, through data fitting analysis, it was found that signal mutation and waveform distortion contribute the most to the model's prediction. Therefore, on the basis of the original model, the data training set excluding signal mutation and waveform distortion is removed, and the proportion of the training set is increased for cross validation to find the optimal parameter combination.

Based on the characteristics of work (1.1), construct the volatility features of moving average, moving variance, and autocorrelation function on the basis of the signal, then train the learning model, and then use random forest for training. The results obtained from the sliding window conversion algorithm based on temporal data are shown in Fig.14.

	precision	recall	f1-score	support
A	1.00	1.00	1.00	38404
B	1.00	1.00	1.00	5319
C	1.00	1.00	1.00	1773
accuracy			1.00	45496
macro avg	1.00	1.00	1.00	45496
weighted avg	1.00	1.00	1.00	45496

Fig.14: EMR volatility characteristic random forest training results graph

A label value of 1 in the above Figure represents interference, while 0 represents normal. The sensitivity value shown in the above Fig.is 1, indicating that the training test results can fit well. Next, we will proceed with future interference detection, as shown in Fig.15.

```
A    34790
B       12
dtype: int64
```

Fig.15: Future EMR interference signal detection results

The results of the above Fig.indicate that there is a significant difference in the proportion of predicted results. Based on the predicted normal EMR, the interference EMR can be inferred, as shown in Fig.16.

	1电磁辐射 (EMR)	1时间 (time)	moving_avg	moving_var	moving_std	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	predict
15960	0.0	2022-05-14 15:07:39	13.982795	3.179131	1.783012	12.370	13.330	13.541	14.110	14.430	19.540	A
15961	0.0	2022-05-14 15:09:31	13.881805	3.957272	1.989289	12.610	13.110	13.230	13.980	14.630	20.198	A
15962	0.0	2022-05-14 15:11:22	13.784205	4.756318	2.180898	12.730	13.030	13.378	14.260	14.390	19.520	A
15963	0.0	2022-05-14 15:16:57	13.686355	5.533272	2.352291	13.232	13.170	13.310	13.650	14.569	19.570	A
15974	0.0	2022-05-14 15:57:52	13.151330	7.709514	2.776601	11.650	12.910	13.489	13.670	14.050	15.324	A

Fig.16: Display of EMR training results for interference

The time interval of electromagnetic radiation interference signals is shown in Table 6.

Table 6: Time intervals of electromagnetic radiation interference signals

Item	Starting point of time interval	Ending point of time interval
1	2022-05-01 00:00:00	2022-05-14 15:06:36
2	2022-05-14 15:14:37	2022-05-14 15:56:54
3	2022-05-14 16:23:58	2022-05-14 17:42:00
4	2022-05-14 17:42:00	2022-05-14 18:16:40
5	2022-05-14 18:16:40	2022-05-14 19:10:20

Train AE using the same method, as shown in Fig.17 to Fig.19.

	precision	recall	f1-score	support
A	1.00	1.00	1.00	11935
B	0.96	1.00	0.98	964
C	1.00	0.88	0.93	322
accuracy			1.00	13221
macro avg	0.99	0.96	0.97	13221
weighted avg	1.00	1.00	1.00	13221

Fig.17: AE Volatility Characteristics Random Forest Training Map

It can be seen that the fitting effect is quite good, and then the data can be predicted in three stages:

```
A    57895
B    16841
dtype: int64
```

```
C    24471
A      107
B       45
dtype: int64
```

Fig.18: Future AE interference signal detection results

Time period 1 interferes with data, while time period 2 only has interference data. Final prediction result, as shown in Fig.19.

16328	0.469503	2022-10-31 04:22:51	1.690620	0.444256	0.666526	0.153767	2.058432	2.042030	1.990774	1.943106	1.972834	A
16329	0.479241	2022-10-31 04:24:41	1.683157	0.451177	0.671697	0.167606	2.059457	2.037417	1.991287	1.941568	1.971809	A
16330	0.491543	2022-10-31 04:26:30	1.675771	0.457848	0.676645	0.181958	2.068683	2.046643	1.990774	1.940543	1.968734	A
16331	0.502819	2022-10-31 04:28:21	1.668467	0.464292	0.681389	0.195797	2.071246	2.038954	1.989236	1.939518	1.963608	A
16332	0.515633	2022-10-31 04:30:10	1.661233	0.470483	0.685918	0.211174	2.073808	2.046643	1.987699	1.939006	1.962583	A
16333	0.526909	2022-10-31 04:32:00	1.654065	0.476447	0.690251	0.250641	2.068170	2.039467	2.047155	1.946694	1.960533	A
16334	0.538186	2022-10-31 04:33:49	1.646963	0.482188	0.694397	0.242440	2.189646	2.039467	1.988211	1.938493	1.958483	A

Fig.19: Display of Training Results for Interference AE Part

The time interval of acoustic emission interference signals is shown in Table 7.

Table 7: Time intervals of acoustic emission interference signals

Item	Starting point of time interval	Ending point of time interval
1	2022-10-10 00:00:30	2022-10-16 01:31:10
2	2022-10-16 02:09:08	2022-10-25 14:31:49
3	2022-10-25 14:31:49	2022-10-31 03:21:29
4	2022-10-31 05:23:43	2022-11-09 09:56:42
5	2022-10-31 04:52:08	2022-10-31 05:03:06

3.2 Model establishing and solving for work two

3.2.1 Model establishing and solving for work 2.1

In work one, by comparing Fourier transform and wavelet transform, it was found that wavelet transform has more advantages in processing interference signals in this work. Therefore, wavelet transform is still used to identify the trend characteristics of electromagnetic radiation and acoustic emission signals before the occurrence of danger by comparing interference signals with base signals. The results obtained from the sliding window conversion algorithm based on temporal data are shown in Fig.20 and Fig.21.

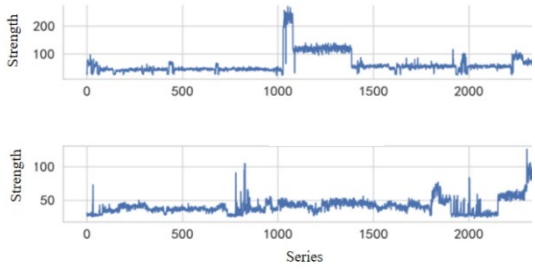

Fig.20: Comparison of Premature EMR and Normal EMR Intensity

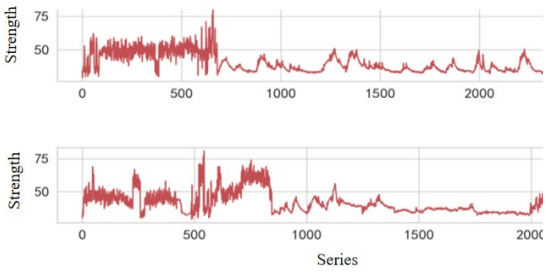

Fig.21: Comparison of intensity between precursor AE and normal AE

As the title suggests, the signal data will have a "slight upward" trend, which can be constructed using the statistical values of KS trend test. Below is a comparison of statistical values, as shown in Table 8.

Table 8: Comparison of precursor signals and normal signal statistics for EMR and AE

Item	count	mean	std	min	25%	50%	75%	max
Normal EMR	73418.000000	49.815421	18.279324	9.610000	42.451000	46.489500	52.000000	270.000000
EMR precursor	12532.000000	71.054808	91.899844	11.670000	30.148750	41.000000	71.674000	491.000000
Normal AE	15587.000000	37.432989	3.727368	29.000000	35.290000	36.661000	38.265500	80.000000
AE precursor	2212.000000	41.668658	8.303728	29.000000	35.410000	39.105000	45.163750	81.000000

The statistical value data is shown in the table above, and there are some discrepancies in the data results. Further reference can be made to the auto-correlation function, as shown in Fig.22 and Fig.23.

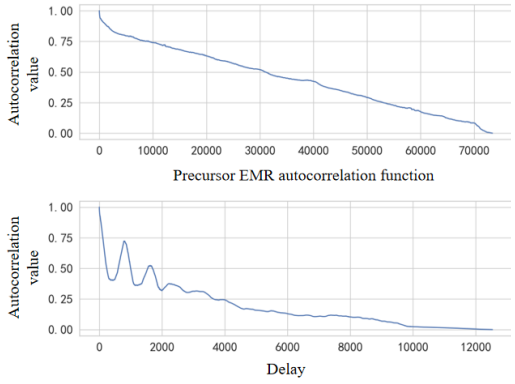

Fig.22: Auto-correlation diagram of EMR

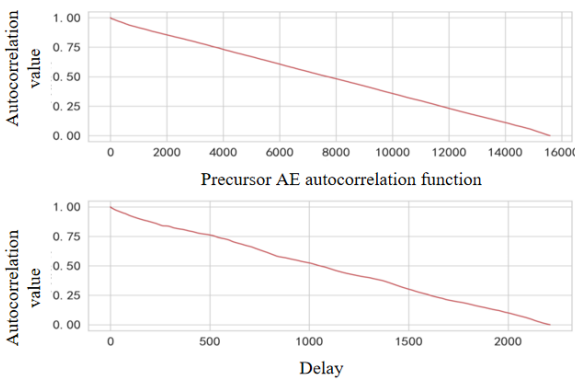

Fig.23: Auto-correlation diagram of AE

As shown in the above figure, the fluctuation of the autocorrelation function of the precursor signal is significantly different from that of the interference signal, that is, the

precursor and normal are relatively similar. According to the autocorrelation function, the trend characteristics of the precursor characteristic data can be obtained, as shown in Tables 9 and Table 10.

Table 9: Trend characteristics of precursor characteristic data of EMR

Electromagnetic Radiation (EMR)
1. The precursor EMR shows a fluctuating downward trend in the initial stage
2. The overall trend of precursor EMR is decreasing, which is similar to the normal signal
3. The time-frequency of precursor EMR is lower than that of normal EMR
4. The precursor EMR shows a continuous downward trend in the middle and later stages

Table 10: Trend Characteristics of Precursor Characteristic Data of AE

Sound propagation signal (AE)
1. The overall precursor AE is the same as the normal AE
2. The time-frequency of precursor AE is very low
3.The trend of changes in precursor AE is extremely similar to that of normal AE

3.2.2 Model establishing and solving for work 2.2

In the process of solving the model in work 1, it is found that the fluctuation of the auto-correlation function is obviously different from the interference signal, and the similarity between the precursor signal and the normal signal is very high. Therefore, the random forest model is still used in work 2. At the same time, when establishing features, remove auto-correlation features and add ADF stationary sequence test and MK trend test features. The basic model of ADF test is an auto regressive model (AR model), which is mathematically defined as follows:

$$\begin{aligned} \Delta y_i = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \\ \Delta y_{t-2} + \cdots + \delta_p \Delta y_{t-p} + \varepsilon_t \end{aligned} \tag{11}$$

The MK test is a non-parametric statistical test method used to analyze whether there is a trend in time series data. The long-term time series data given in the context of this work, based on symbol comparison of data pairs: set the time series $x_1, x_2, \ldots, x_n$, define:

$$s_{ij} = \begin{cases} 1 & \text{if } x_j > x_i \\ 0 & \text{if } x_j = x_i \\ -1 & \text{if } x_j < x_i \end{cases} \tag{12}$$

The test statistic S is the sum of all s_{ij} pairs:

$$\mathrm{S} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{s}_{ij} \tag{13}$$

In the absence of equal values, the standard normal variable Z is used to test:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{\mathrm{Var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S+1}{\sqrt{\mathrm{Var}(S)}} & \text{if } S < 0 \end{cases} \tag{14}$$

$$\mathrm{Var}(S) = \frac{n(n-1)(2n+5)}{18} \tag{15}$$

The above is a mathematical description of the two testing methods, ADF and MK. Based on this, the electromagnetic radiation from May 1, 2022 to May 30, 2022 and the time interval of the interference signal in the acoustic emission signals from April 1, 2022 to May 30, 2022 and October 10, 2022 to November 10, 2022 can be identified, and the interval of the interference signal can be given.

Kendall's W consistency test was performed and shown in Table 11.

Table 11: Kendall's W Consistency Test Results

Kendall's W analysis results					
Name	Rank mean	median	Kendall's W coefficient	X^2	P
Normal EMR	3.125	48.152	0.283	6.797	0.079*
EMR precursor	3	71.364			
AE precursor	2.188	40.387			
AE precursor	1.688	37.047			

The above table presents the results of model validation. The results of the Kendall coefficient consistency test show that the significance P-value of the overall data is 0.079 *, which does not show significance at the level and cannot reject the null hypothesis. Therefore, the data cannot show consistency. At the same time, the Kendall coordination coefficient value of the model is 0.283, indicating a general degree of consistency in the correlation. To use multi-variable temporal data sliding window conversion for temporal data sliding window conversion, first define the parameters: window size=200, and the results obtained from the sliding window conversion algorithm based on temporal data are shown in Fig.24 to Fig.26.

	precision	recall	f1-score	support
A	1.00	1.00	1.00	38397
B	1.00	1.00	1.00	5319
C	1.00	1.00	1.00	1755
accuracy			1.00	45471
macro avg	1.00	1.00	1.00	45471
weighted avg	1.00	1.00	1.00	45471

Fig.24: EMR timing data window conversion result diagram

As shown in the above figure, the fitting effect is good.

```
A    27933
B      419
Name: count, dtype: int64
```

Fig.25: EMR detection results from April 8, 2020 to June 8, 2020

```
A    27933
B      419
Name: count, dtype: int64
```

Fig.26: EMR results from November 20, 2021 to December 20, 2021

As can be seen from the above figure, there is a significant difference in the predicted results (the window size for this part is 250). Time intervals for the precursor characteristics of electromagnetic radiation could be summarized and shown in Table 12.

Table 12: Time interval of electromagnetic radiation precursor characteristics

Item	Starting point of time interval	Ending point of time interval
1	2020-04-09 14:13:25	2020-04-16 14:13:25
2	2020-05-04 06:12:32	2020-05-11 06:12:32
3	2020-05-25 12:43:08	2020-06-22 12:43:08
4	2021-11-30 08:10:08	2021-12-07 08:10:08
5	2021-12-16 02:08:40	2021-12-23 02:08:40

Train AE using the same method: the window size for this training is window size, which is 250.

	precision	recall	f1-score	support
A	1.00	1.00	1.00	8598
B	1.00	1.00	1.00	821
accuracy			1.00	9419
macro avg	1.00	1.00	1.00	9419
weighted avg	1.00	1.00	1.00	9419

Fig.27: AE Time Series Data Window Conversion Results

As can be seen from the Fig.27, the fitting effect is good. The summarized time intervals for precursor characteristics of acoustic emissions are shown in Table 13.

Table 13: Time interval of acoustic emission precursor characteristics

Item	Starting point of time interval	Ending point of time interval
1	2022-04-06 17:31:45	2022-04-13 17:31：45
2	2022-04-08 14:22:12	2022-04-15 14:22:12
3	2022-04-09 19:10:42	2022-04-16 19:10:42
4	2022-04-10 08:23:19	2022-04-17 08:23:19
5	2022-05-05 19:40:04	2022-05-12 19:40:04

3.3 Model establishing and solving for work three

3.3.1 logistic regression model

In this work, it is necessary to predict the interference characteristic signals, which is actually a solution to the probability of danger occurrence. Establish a logistic regression model, input the feature weighted sum of interference signals in work one, and transform this linear combination through a sigmoid function to output values between (0, 1). The closer it is to 1, the greater the probability of danger occurrence. Conversely, the closer it is to 0, the lower the probability of danger occurrence. The loss function is defined as cross entropy loss, which is used to measure the difference between the predicted probability distribution of the model and the true probability distribution. The mathematical definitions of the sigmoid function and cross entropy loss function are as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{16}$$

$$L = -[x\log(p) + (1 - x)\log(1 - p)] \tag{17}$$

The final mathematical form of the model is:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \tag{18}$$

For this model, maximum likelihood estimation is used to train model parameters, predict the features at the last data collection time of each time period, and output the probability of precursor features appearing at each time period. The likelihood function formula is as follows:

$$L(\theta; D) = \prod_{i=1}^{n} f\left(x_i \mid \theta\right) \tag{19}$$

3.3.2 work 3 training results and analysis

The training results are shown in Tables 14 to Table 16.

Table 14: Probability of Precursor Features

Moment of electromagnetic radiation	Probability of precursor features	Moment of acoustic emission	Probability of precursor features
2023-1-24 23:58:36	8.227057631956858e-06	2023-1-24 23:58:36	0.0007851826910823287
2023-2-11 23:59:20	0.821522141194042	2023-2-11 23:59:20	0.006907113873965548
2023-2-26 23:59:27	0.9966395940550525	2023-2-26 23:59:27	0.38311406634996756
2023-3-10 23:58:14	0.1227736265443528	2023-3-10 23:58:14	0.0199011667138625
2023-3-30 23:58:13	0.998951683646935	2023-3-30 23:58:13	0.8218217011138061

The data in Table 15 is used to evaluate the performance or validate the effectiveness of the model, including likelihood ratio test, p-value, AIC value, and BIC value. If the P-value is less than 0.05, it indicates that the model is effective. The AIC and BIC values are used to compare the advantages and disadvantages of two models, and the smaller the value, the better. The Table 16 shows the classification evaluation indicators, which are further quantified to measure the classification effect of logistic regression.

Table 15: Model Evaluation Indicators

Likelihood ratio value	p	AIC	BIC
183.595	0.000***	199.595	223.68

Table 16: Classification Evaluation Indicators Table

Accuracy	Recall	Accuracy	F1	AUC
0.98	0.98	0.889	0.809	0.89

3.3.3 Optimization and solution of the model

In work 1, the data recording interval is every two minutes, while work 2 involves more dense data, with the recording frequency increased to once per second. For work 2, a sliding window strategy with a time length of 250 seconds was adopted to analyze the data, which is roughly equivalent to a time span of 4 to 5 minutes, in order to capture the dynamic mean, trend, and variance characteristics of the data. Drawing inspiration from this strategy, in the handling of work 3, the same time window size (250 seconds) was used to construct features, aiming to extract feature representations at the end of each window and ultimately obtain the probability output of model prediction. The sigmoid function can effectively squeeze the

output value into the (0,1) interval, thereby providing more accurate and intuitive probability interpretation for the prediction results, improving the reliability and practicality of the model's output probability. The obtained results are shown in Table 17.

Table 17: Probability of Precursor Features

Moment of electromagnetic	Probability of precursor	Moment of acoustic	Probability of precursor
2023-1-24 23:58:36	$8.22700763195485e^{-06}$	2023-1-24 23:58:36	0.0007851827910823287
2023-2-11 23:59:20	0.82112214119704 2	2023-2-11 23:59:20	0.006907013873965548
2023-2-26 23:59:27	0.9966395940550125	2023-2-26 23:59:27	0.38311306634996756
2023-3-10 23:58:14	0.12277382654435288	2023-3-10 23:58:14	0.0199011617138625
2023-3-30 23:58:13	0.998951783646935	2023-3-30 23:58:13	0.8218217000138063

4 CONCLUSIONS

The advantage of logistic regression model is that the model is simple and easy to interpret, which is particularly effective for the processing of a wide range of data. Secondly, the importance of features can be calculated to make the influence of each feature on the prediction result clearer. However, since logistic regression assumes that the data is linearly separable, the nonlinear relationship may not be well modeled, so the feature transformation is needed. At the same time, due to the instability of coefficient estimation, logistic regression may not perform well when dealing with highly correlated features. In addition, when dealing with a large number of features or complex relationships, its accuracy will be reduced compared with other predictive models.

Using logistic regression model, for work three, which is actually a discontinuous version of work two, you can use ADF test and MK test, and use random forest model to separate training set and test set for training. In work one, the data is recorded every two minutes, while work two involves more intensive data and the recording frequency is increased to once per second. For work two, a 250-second sliding window strategy was used to analyze the data, which is roughly equivalent to a 4-5 minutes time span, in order to capture the dynamic mean, trend, and variance characteristics of the data. Drawing on this strategy, in the processing of work three, the same time window size (250 seconds) is used to construct features, aiming to extract the feature representation at the end of each window, and finally get the probability output predicted by the model.

5 ACKNOWLEDGEMENTS

This work is supported by ministry of education industry-university cooperative education project (Grant No.: 231106441092432), the research and practice of integrating "curriculum thought and politics" into the whole process of graduation design of Mechanical engineering major: (Grant. No.: 30120300100-23-yb-jgkt03), research on the integration mechanism of "course-training-competition-creation-production" for innovation and entrepreneurship of mechanical engineering majors in applied local universities (Grant. No.: CXKT202405), Mechanical manufacturing equipment design school-level "gold class" construction project (Grant. No.: 30120324001).

REFERENCES

[1] Wang, L. N., & Chen, L. (2017). Research on the application of support vector machine (SVM) combined with object-oriented method in information extraction of open pit mining area. *The 19th Academic Exchange Meeting of six Provinces and one City of Surveying and Mapping Society in East China and the 2017 Cross-Straits Surveying and mapping Technology Exchange and Academic Seminar*, 201-206.

[2] Hu, H. B., & Zhan, Y. L. (2018). Change characteristics of land use landscape pattern based on decision tree classification of remote sensing image. *Green Technology*, 24(072): 200-205.

[3] Wang, D. Y. (2021). Research on data reconstruction method of ground penetrating radar based on time series analysis. *Shandong Technology and Business University*.

[4] Wu, J. F. (2021). Research on anomaly detection and fault warning technology for in orbit satellites based on LSTM. *National University of Defense Technology*.

[5] Zhou,G. Y., Wan, S. P., & Chen.Y. N. (2022). Research on denoising algorithm for phase sensitive time domain reflectometer based on moving variance average algorithm. *Journal of Instruments and Meters*, 43(10): 233-240.

[6] Zhang, J. L., Guo, S. Y., & Ren, C. P. (2024). Research on personal credit rating card model based on logistic regression. *Modern Information Technology*, 8(05): 12-16.

[7] Du, B. Y., Gao, J. H., & Zhang, G. Z. (2024). Seismic prediction method and application of fracture density inversion for shale reservoir in-situ stress. *Petroleum Geophysical Exploration*, 59(2): 279-289.

[8] Zhou, Z. X., Zhen, X. J., & Liang, Y. G. (2024). Study on acoustic emission source location of ancient timber based on wavelet packet transform and cross-correlation. *Shanxi Architecture*, 50(9): 1-5.

[9] Yao, J. W., Li, Y., & Lv, Y. J. (2024). Research on the trend of water quality evolution in drinking water source areas based on water quality index method and M-K Test. *Environmental Science and Management*, 49(04): 43-48.

[10] Nadir, B., Azzeddine, D., & Ahmed, B. (2024). Exploring the effects of overvoltage unbalances on three phase induction motors: Insights from motor current spectral analysis and discrete wavelet transform energy assessment. *Computers and Electrical Engineering*, 117109242.

[11] Zhang, X. G., & Luo R. (2024). Prediction and analysis of TCM constitution based on ARIMA time series model. *Asia-pacific traditional medicine*, 20(04): 156-16.

[12] Li, Y. H., Zhou, Y., & Wu, X. F. (2024). A study on the potential supply evaluation method of regional landscape recreation services based on geodetectors and random forest models - taking the subtropical moist huaiyang low mountain landscape area as an example. *Journal of Ecology*, 13: 1-17.