




OPEN

## Application of an improved naive Bayesian analysis for the identification of air leaks in boreholes in coal mines

Hong-yu Pan, Sui-nan He , Tian-jun Zhang, Shuang Song & Kang Wang

Borehole extraction is the basic method used for control of gases in coal mines. The quality of borehole sealing determines the effectiveness of gas extraction, and many influential factors result in different types of borehole leaks. To accurately identify the types of leaks from boreholes, characteristic parameters, such as gas concentration, flow rate and negative pressure, were selected, and new indexes were established to identify leaks. A model based on an improved naive Bayes framework was constructed for the first time in this study, and it was applied to analyse and identify boreholes in the 229 working face of the Xiashijie Coal Mine. Eight features related to single hole sealing sections were taken as parameters, and 144 training samples from 18 groups of real-time monitoring time series data and 96 test samples from 12 groups were selected to verify the accuracy and speed of the model. The results showed that the model eliminated strong correlations between the original characteristic parameters, and it successfully identified the leakage conditions and categories of 12 boreholes. The identification rate of the new model was 98.9%, and its response time was 0.0020 s. Compared with the single naive Bayes algorithm model, the identification rate was 31.8% better, and performance was 55% faster. The model developed in this study fills a gap in the use of algorithms to identify types of leaks in boreholes, provides a theoretical basis and accurate guidance for the evaluation of the quality of the sealing of boreholes and borehole repairs, and supports the improved use of boreholes to extract gases from coal mines.

One of the most common and dangerous natural risks associated with coal mining is methane, which can mix with air and cause disasters. Extraction of gases from coal mines is a fundamental measure taken to prevent and control disasters and accidents<sup>1,2</sup>. Drainage boreholes are used to extract gas from coal seams<sup>3</sup>. However, the concentration of gas extracted from coal seams by boreholes in China is generally low because of leaks. Air flows through a channel into a borehole and reduces the negative pressure to enable gas extraction. As a result, low concentrations of gas are extracted by boreholes<sup>4,5</sup>. The effective identification of the presence and types of gas leaks is vital to improve the efficiency of gas extraction.

Studies of the mechanism of borehole leakage have led to the development of physical models. Zhang T<sup>6</sup> explained that air leakage was caused by a local change in the strain<sup>7,8</sup> around a borehole. Zhang C<sup>9</sup> studied the mechanism of air leakage in the cracks around a borehole and concluded that the leakage mechanisms of fractures around boreholes differed depending on the extraction stage. This insight provided a theoretical basis for the classification and identification of leaks in boreholes. To further analyse the flow state and characteristic changes of air leaks in boreholes, some scholars constructed a physical model to determine the mechanism of leakage. Zhang J<sup>10,11</sup> combined numerical simulations of the leakage mechanism around a borehole in coal with the rheological and viscoelastic–plastic characteristics of coal to build a dynamic leakage model of the borehole. Based on an analysis of flow coupling between methane and air in borehole fractures, Fan J<sup>12</sup> constructed a flow model of air leakage coupling components in boreholes by using the finite difference method (FDM). Zhang Y<sup>13</sup> constructed a physical model of air leakage in boreholes and classified three types of leaks according to their source, i.e., roadway fissure zones, borehole fissure zones, and materials used in sealing sections of boreholes. Wang Z<sup>14</sup> analysed the mechanism of air leakage from boreholes by numerical simulation and established a dynamic leakage model of drainage boreholes. Wang H<sup>15</sup> and Zhang Y<sup>16</sup> discussed the influence of air leakage on gas concentration by studying the influence of factors around roadways and boreholes and constructed an air-gas

College of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an 710054, People's Republic of China.  email: he917272990@163.com

	Type	Distance	Reason
I	Air leakage from the orifice of the gas drainage borehole	0–2 m	The pipe wall broke and leaks developed
II	Air leakage from the mid-end seal section of the gas drainage borehole	2–9 m	Cracks developed in the seal
III	Air leakage from the deep coal of the gas drainage borehole	9–12 m	The strength of the seal was insufficient

**Table 1.** Types of borehole leaks.

mixed-flow coupling model. Their physical model explained the mechanisms of gas extraction and air leakage in boreholes and provided a theoretical basis for the classification of air leaks from boreholes. However, the construction of a physical model of air leakage in a drilling hole is complicated and cannot be applied quickly to guide field practice. Therefore, there is still a need for an efficient mathematical model for identification of leaks.

Advances in computer science, applied mathematics and artificial intelligence have promoted in-depth research on identification models for use in coal mining<sup>17–22</sup>. However, the algorithms used to construct these models are subject to limitations. The hierarchical cluster analysis method cannot redistribute existing data and has a small number of iterations. The chaotic immune particle swarm optimization-probabilistic neural network (CIPSO-PNN) optimizes the PNN, but the process of finding the best solution is long, and the model is complex. In Fisher's discriminant analysis, the number, representativeness and correctness of the learned samples directly impact the recognition accuracy of the model. In addition, the algorithms of existing discrimination classification models cannot adapt to differences in the relationships of various data characteristics with multiparameter nonlinearity, which is important for discrimination of leaks in extraction boreholes. naïve Bayes classification, a classification method based on the Bayes principle and independent assumption of feature conditions, has stable classification efficiency<sup>23</sup>. Compared with the above classification algorithms, decision trees and artificial neural networks perform better on small amounts of sample data and have the minimum error rate, and they have been widely used in coal mines<sup>24,25</sup>. Therefore, they have been applied to identify in air leaks in gas boreholes. However, because of low sensitivity to linear data, improvements are needed.

In summary, research is now relatively mature for the development of models for leaks in boreholes for gas extraction based on studies of the mechanism of air leakage, and models are widely used in the field of coal mining. However, research is limited on using machine learning methods to analyse multisource characteristic information about air leakage and establish a mathematical model for the recognition of leaks from boreholes. In this study, we collected data for leaks from boreholes and applied multisource data fusion theory (MDF) and principal component analysis (PCA). We also improved the traditional naïve Bayesian classification (NBC) system and established mathematical models to identify types of air leaks from boreholes. In this study, this model fills a gap by supporting an algorithm to identify types of leaks in boreholes used to extract gases from coal mines, provides a theoretical basis and accurate guidance for the evaluation of the quality of the sealing of boreholes and borehole repairs, and supports the improvement of the application of boreholes to extract gases from coal mines.

## Construction of an improved naïve Bayesian model for the identification of air leaks from gas drainage boreholes

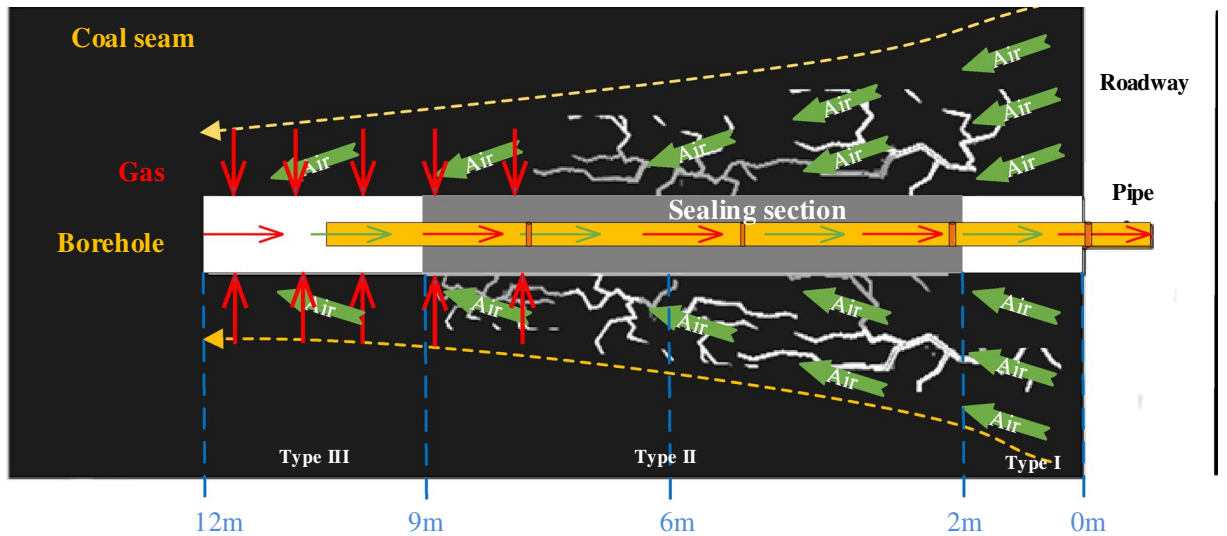
**Feature information selection.** According to previous studies<sup>4,15,26–31</sup>, air leaks from gas drainage boreholes can be divided into the three types shown in Table 1.

In Fig. 1, there are many cracks in the coal seam. Due to the poor sealing effect, air from the roadway enters the borehole through cracks in the coal seam, which leads to the leakage of the borehole. In addition, the connections between the extraction pipes are not close, which results in low extraction concentrations. In this paper, according to the actual situation of the gas drainage borehole in the 229 working face of the Xiashijie Coal Mine, eight characteristics can reflect the gas drainage effect of the borehole, including  $A_1$ : extraction flow,  $A_2$ : gas concentration at 0 m,  $A_3$ : gas concentration at 2 m,  $A_4$ : gas concentration at 6 m,  $A_5$ : gas concentration at 9 m,  $A_6$ : gas concentration at 12 m,  $A_7$ : negative pressure at the orifice and  $A_8$ : negative pressure at the extraction, and they are used in the model for the identification of leaks in boreholes for gas extraction.

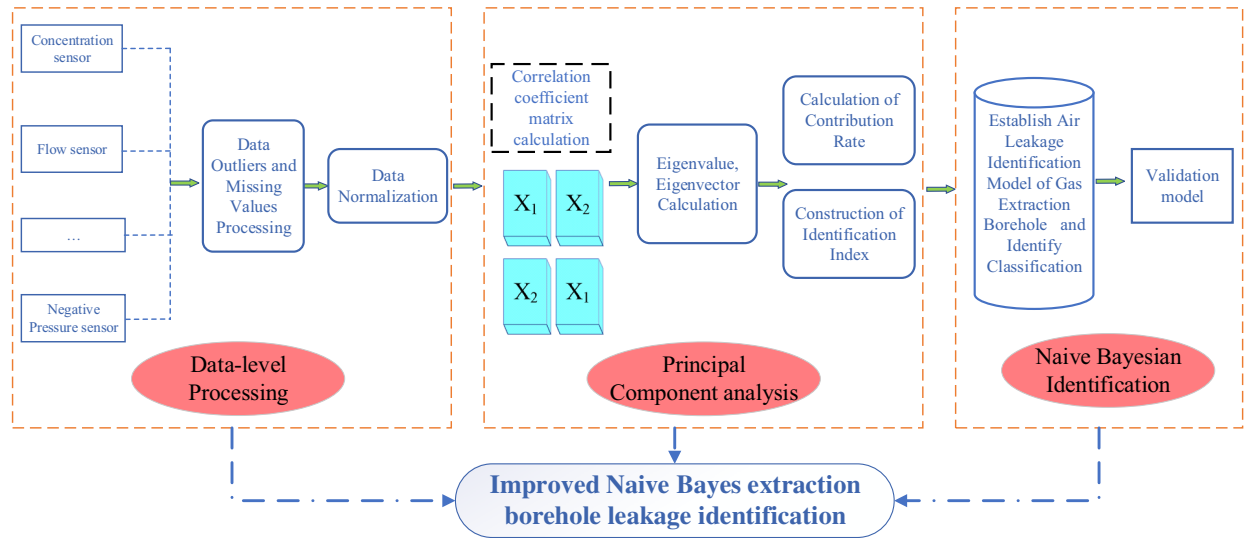
**Identification model construction.** A naïve Bayes classifier (NBC) and the eight characteristics (above) were used as the main theory for model construction. Since the NBC could not accommodate the missing data for air leakage in gas extraction boreholes, and since the identification and classification accuracy of information with strong correlations is not high, some data easily have a greater impact on the overall model<sup>32</sup>. As shown in Fig. 2, by using MDF and principal component analysis (PCA) to improve the traditional NBC, a model for the identification of air leaks from a borehole for gas extraction was established as follows:

**Data preprocessing.** The existing  $m$ -dimensional sample data of gas drainage borehole leakage  $\mathbf{x} = (x_1, x_2, \dots, x_m)$  with  $n$  independent observations,  $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n^T)$ , is used as the observation sample to build the gas drainage borehole leakage data matrix:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & \dots & \dots & \dots \\ \dots & \dots & x_{ij} & \dots \\ x_{n1} & \dots & \dots & x_{nm} \end{bmatrix} \quad (1)$$



**Figure 1.** Air leakage characteristics of a borehole for extraction of gas.



**Figure 2.** Flow chart of building the model.

$x_i = (x_{i1}, x_{i2} \dots x_{im})$  represents the observation sample of group  $i$ ,  $i = 1, 2, \dots, n$ , and  $x_{ij}$  represents the  $j$ th variable of the  $i$ th group of observation samples, where  $j = 1, 2, \dots, m$ .

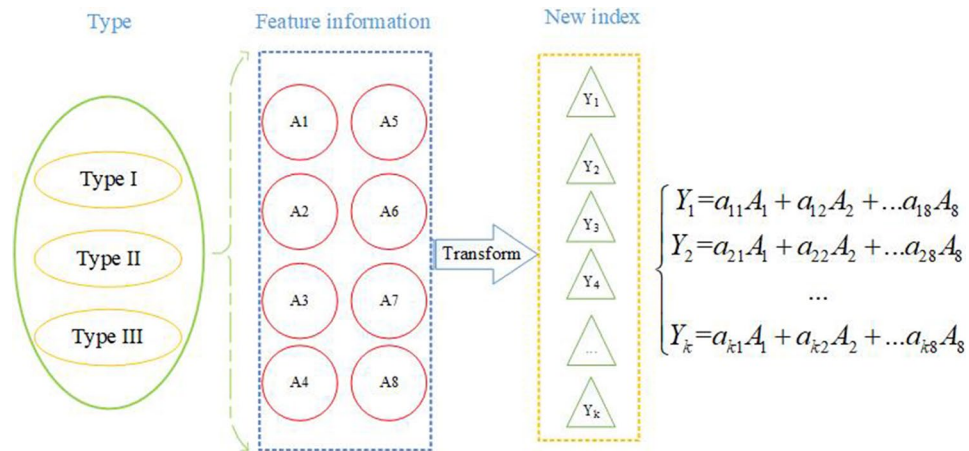
Following MDF theory, the training samples for gas drainage borehole leaks are processed at the data level<sup>33</sup>, and the processed data are standardized with Eqs. (2)–(4).

$$x_{ij}^1 = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}} \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, m \quad (2)$$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (3)$$

$$s_{ij} = \frac{1}{m-1} \sum_{j=1}^m (x_{ij} - \bar{x}_j)^2 \quad (4)$$

In Eqs. (2)–(4),  $x_{ij}^1$  represents the standardized single sample data,  $\bar{x}_j$  represents the sample mean for the same characteristic information, and  $s_{jj}$  represents the covariance of single sample data. Equations (2)–(4) can eliminate the influence of the data dimension. The standardized gas drainage borehole leakage data are still expressed in  $X$ .



**Figure 3.** Construction of the new indexes for air leakage from a gas drainage borehole.

*Principal component selection.* The correlation coefficient matrix of the gas drainage borehole leakage sample data after standardization is calculated as follows:

$$R = [r_{ij}]_{n \times n} = \frac{1}{m} XX^T \tag{5}$$

where

$$r_{ij} = \frac{1}{m-1} \sum_{i=1}^m x_{il}x_{lj} \quad i, j = 1, 2, \dots, n \tag{6}$$

The characteristic equation of the sample correlation matrix **R** is obtained with **k** eigenvalues and the corresponding **k** unit eigenvectors:

$$\begin{aligned} |\mathbf{R} - \lambda \mathbf{I}| &= 0 \\ \lambda_1 &\geq \lambda_2 \geq \lambda_3 \geq \dots \lambda_m \end{aligned} \tag{7}$$

In Eq. (7),  $\lambda$  is the characteristic value of the characteristic equation corresponding to the characteristic information, and the values are sorted according to the size of the characteristic value, from large to small.

The cumulative contribution rate and cumulative variance contribution rate are calculated as follows:

$$z = \frac{\lambda_k}{\sum_i \lambda_m} \tag{8}$$

$$z_i = \sum_{j=1}^k \left( \frac{\lambda_k}{\sum_i \lambda_m} \right) \tag{9}$$

The principal component  $z_i \geq 85\%$  is determined to reduce the dimensionality and eliminate information overlap.

*Construction of new indexes.* The unit eigenvector corresponding to the first **k** principal components is obtained:

$$a_i = (a_{1i}, a_{2i} \dots a_{ni})^T, \quad i = 1, 2, 3 \dots k \tag{10}$$

Linear transformation with **k** unit eigenvectors as coefficients yields:

$$Y_i = a_i^T \mathbf{x} \quad i = 1, 2, 3 \dots k \tag{11}$$

That is, after orthogonal transformation, potentially correlated variables or influencing factors in the gas drainage borehole leakage data are linearly combined to obtain a set of new linear irrelevant variables, simplify the data structure, extract the data characteristics, and construct a new improved naive Bayes identification index.

As shown in Fig. 3, for the characteristic information obtained in the lower section, the original eight-dimensional sample characteristic information ( $A_1, A_2, \dots, A_8$ ) is converted into a new **p**-dimensional identification index ( $Y_1, Y_2, \dots, Y_k$ )  $k < 8$ . The associated characteristic information is combined and retains most of the information of the original variables<sup>34</sup> while eliminating overlapping information.

**Modelling.** The data matrix  $\mathbf{Y} = [y_1, y_2 \dots y_k]$  is constructed according to the new leakage index of the drainage borehole. Among them,  $y_i = [y_1, y_2 \dots y_n]^T$ ,  $y_i^{(j)}$  is the  $j$ th feature of sample  $i$ ,  $y_i^{(j)} \in \{a_{j1}, a_{j2}, \dots, a_{jns}\}$ , and  $a_{jl}$  is the possible value of the  $j$ th feature,  $i = 1, 2, 3 \dots n$ ;  $j = 1, 2, 3 \dots k$ ;  $l = 1, 2, 3 \dots s_n$ . The sample category is  $G = \{g_1, g_2 \dots g_T\}$ ,  $y_i \in \{g_1, g_2 \dots g_T\}$ .

The prior probability and conditional probability of the air leakage category of the extraction borehole are calculated. Because the characteristic information data optimized by the principal component are normally distributed, the Gaussian function is used to determine the conditional probability, as shown in Eqs. (12)–(13).

$$P(Y = g_t) = \frac{\sum_{i=1}^n I(y_i = g_t)}{n} \quad t = 1, 2, 3 \dots T \quad (12)$$

$$P\left(y_i^{(j)} = a_{jl} \mid Y = g_t\right) = \frac{1}{\sqrt{2\sigma_{Y=g_t}^2}} e^{-\frac{(y_i^{(j)} - a_{jl})^2 - u_{y=g_t}}{2\sigma_{Y=g_t}^2}} \quad (13)$$

In Eq. (13),  $u_{y=g_t}$  is the normalized expected value of the sample data of category  $g_t$ ;  $\sigma_{Y=g_t}$  is the normalized variance of the sample data of category  $g_t$ . The posterior probability is calculated for the given leakage sample data  $y_i = [y_1, y_2 \dots y_n]^T$  of the extraction borehole.

$$P(Y = g_t) \prod_{j=1}^k P(Y^{(j)} = y^{(j)} \mid Y = g_t) \quad (14)$$

The category of an actual case is determined, and the probability model of gas leakage identification of the extraction borehole is built as shown in Eq. (15):

$$G_{y_i} = \arg \max_{g_t} P(Y = g_t) \prod_{j=1}^k P(Y^{(j)} = y^{(j)} \mid Y = g_t) \quad (15)$$

where  $G_{y_i}$  is the maximum posterior probability value of the corresponding category of the leakage of the extraction borehole.

In the actual extraction process, there are gas drainage boreholes with good drainage effects. When the sealing effect is good, the difference in gas concentration at various positions is small. Combined with the air leakage characteristics of the drainage borehole, the gas concentration at different positions in the drainage borehole is defined as  $C_{y_i}^{(b)}$ ,  $i = 1, 2 \dots n$ , for borehole gas concentration positions  $b = 0, 1, 2 \dots, b$ .

$$\frac{C_{y_i}^{(b-1)}}{C_{y_i}^{(b)}} \geq \frac{C_{y_i}^{(0)}}{C_{y_i}^{(b)}} \geq 90\% \quad (16)$$

The corresponding borehole is a borehole with a good gas drainage effect, and there is no need to evaluate the type of leakage. Incorporating Eq. (15), the gas drainage borehole leakage identification model can be constructed as follows:

$$\begin{cases} \frac{C_{y_i}^{(b-1)}}{C_{y_i}^{(b)}} > \frac{C_{y_i}^{(0)}}{C_{y_i}^{(b)}} \geq 90\% \\ G_{y_i} = \arg \max_{g_t} P(Y = g_t) \prod_{j=1}^k P(Y^{(j)} = y^{(j)} \mid Y = g_t) \end{cases} \quad (17)$$

This identification model can realize the identification of leakage and leakage type.

## Model application

**Data acquisition and preprocessing.** The model was applied to the gas drainage borehole of the 229 working face in the Xiashijie Coal Mine of Tongchuan, as shown in Fig. 4. Mainly No. 4 coal is mined in the working face, the thickness of the coal seam is 0 ~ 34.28 m, the original gas content of the coal seam is 3.48 m<sup>3</sup>/t, and the gas pressure is 0.4 MPa, which classifies the mine as a high-gas mine. The gas in the coal seam is extracted by a parallel borehole arrangement.

For the purpose of this study, the characteristic information of gas concentration, flow rate and negative pressure at different depths of the borehole were effectively measured. We designed a detection device to connect each borehole and collect data; the device is shown in Fig. 5. By changing the length of the probe, we monitored the gas concentration and extraction flow at different positions in the borehole. The collected data were used to establish the model discussed in this paper.

According to the actual layout of the test boreholes, 30 groups of 240 monitoring data of gas flow, concentration and negative pressure sensors were selected, and they were divided into 18 groups of training samples and 12 groups of test samples according to the ratio of 6:4. The 18 groups of training samples were preprocessed.

Gas drainage borehole leakage data are multidimensional and multivariate<sup>35,36</sup>, with complex correlations. MDF theory was used to preprocess the data<sup>37</sup>. The Newton interpolation method was used to eliminate abnormal



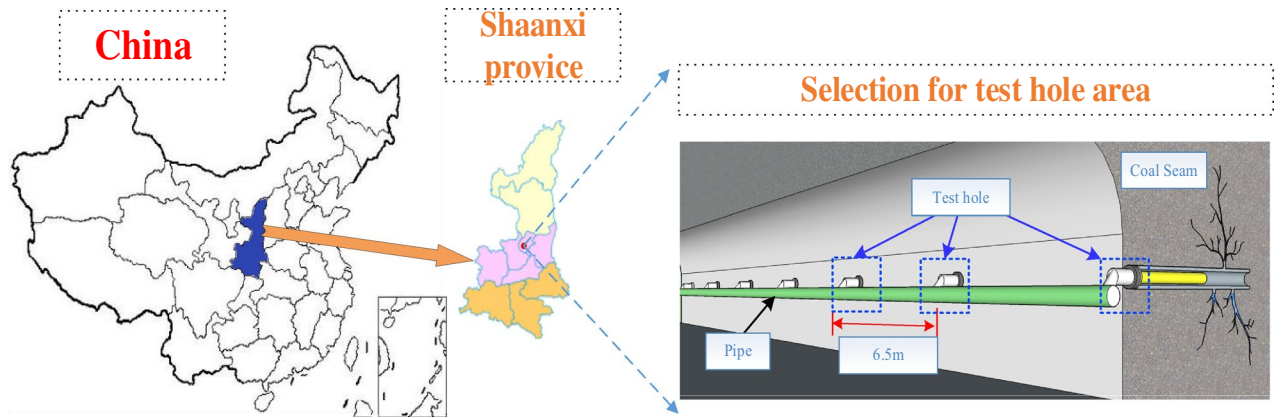


Figure 4. Study area.

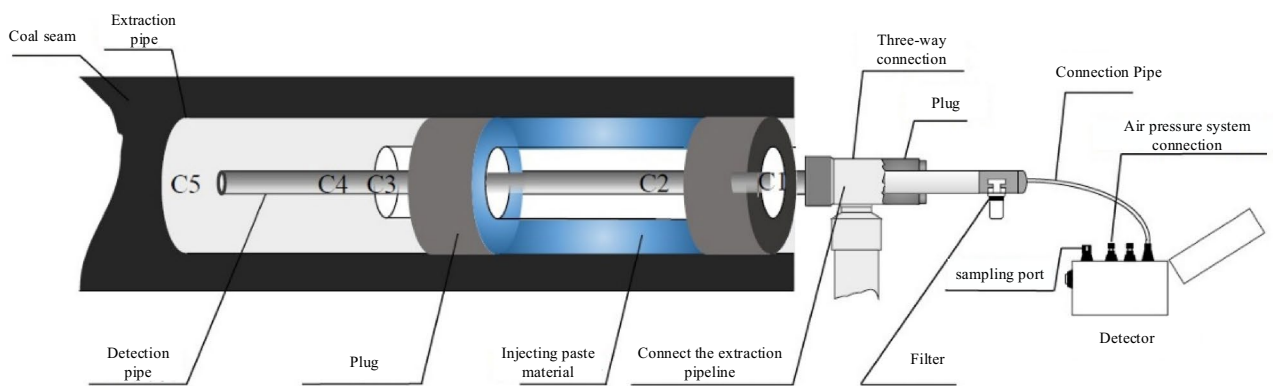


Figure 5. Gas extraction and detection device.

values and fill the missing values of the training sample data of the gas leakage borehole<sup>38</sup> according to Eqs. (18) and (19):

$$f(x_n, x_{n-1}, \dots, x_1, x) = \frac{f[x_{n-1}, \dots, x_1, x] - f[x_n, x_{n-1}, \dots, x_1, x]}{x - x_n} \tag{18}$$

$$f(x) = f(x_1) + (x - x_1)f[x_2, x_1] + (x - x_1)(x - x_2)f[x_3, x_2, x_1] + (x - x_1)(x - x_2) \dots (x - x_{n-1})f[x_n, x_{n-1}, \dots, x_1] \tag{19}$$

The missing value corresponding to the  $x$ -sequence value was substituted into the calculated value  $f(x)$  to eliminate some abnormal values affecting the overall analysis, fill missing values caused by sensor problems, human operation and other factors, and provide perfect and accurate data for the identification air leaks in boreholes. The complete sample data are shown in Table 2.

Table 2 shows 18 groups of drilling test sample data, of which 15 groups correspond to boreholes with leaks and 3 groups correspond to boreholes without leaks. Because the gas concentrations in boreholes 16, 17, and 18 show little change at 0 m, 2 m, 6 m, 9 m, and 12 m, and the proportion is more than 90% according to model Eq. (16), the drainage effect was good, and there is no need to identify the type of leak. In addition, Table 2 shows that due to the air leakage of boreholes, the concentration decreased greatly from the bottom to the orifice in the boreholes in groups 1–15.

The leakage data of the first 15 groups of gas extraction boreholes were standardized by Eqs. (2)–(4), as shown in Table 3. The original data were compared with the box diagram of the standardized data. (Box plots can also be used to detect outliers.) Table 2 and Fig. 6a show that the extraction flow rates in the 15 groups of training samples were very similar, approximately 2.0 m<sup>3</sup>/min, which was relatively low. From the negative pressure of extraction to the negative pressure of the orifice, the pressure loss was obvious. Due to the air leakage in the gas drainage borehole, the differences in gas concentrations between samples in each group at 0 m, 2 m, 6 m, 9 m and 12 m were large, and the distribution of gas concentration in the borehole was not uniform. The specific positions of the different types of air leaks in the gas drainage borehole differed. Figure 6b shows that the range

Borehole number	A1: Extraction flow m <sup>3</sup> /min	A2: Gas concentration at 0 m %	A3: Gas concentration at 2 m %	A4: Gas concentration at 6 m %	A5: Gas concentration at 9 m %	A6: Gas concentration at 12 m %	A7: Orifice negative pressure/kPa	A8: Extraction negative pressure/kPa	Type of air leak
1	2.06	5.56	15.40	15.40	15.80	16.42	1.50	20.60	I
2	2.19	5.98	14.56	15.24	15.74	16.08	1.60	20.40	I
3	1.85	6.65	15.24	15.65	15.67	16.22	1.90	21.30	I
4	2.03	5.84	9.84	10.97	11.56	13.25	1.50	21.40	I
5	1.92	2.45	6.85	7.54	7.68	7.93	0.60	20.60	I
6	1.88	6.11	8.66	11.14	13.25	16.40	1.80	20.90	II
7	2.12	7.56	9.25	10.56	16.58	17.65	2.50	22.40	II
8	1.85	4.54	4.56	11.68	11.98	12.38	1.40	19.60	II
9	2.06	6.85	6.84	6.94	15.28	16.34	1.80	20.70	II
10	1.94	5.98	5.84	14.67	15.06	16.57	1.80	21.60	II
11	1.69	8.65	9.64	10.21	10.28	17.68	2.10	19.50	III
12	2.64	7.79	7.81	7.85	7.92	23.21	1.90	20.70	III
13	1.68	8.19	8.21	8.23	8.31	25.80	2.00	20.10	III
14	1.96	6.15	6.15	6.25	6.40	18.21	1.60	19.30	III
15	1.68	5.85	5.88	5.92	8.94	16.01	1.50	21.60	III
16	4.56	10.35	10.56	10.60	10.97	11.35	9.60	21.40	No leakage
17	6.58	19.68	19.89	19.96	20.65	21.14	9.10	20.80	No leakage
18	5.43	13.95	14.00	14.19	14.21	14.37	9.40	21.20	No leakage

**Table 2.** Multisource data table for gas drainage boreholes.

Borehole number	A1: Drainage flow	A2: Gas concentration at 0 m	A3: Gas concentration at 2 m	A4: Gas concentration at 6 m	A5: Gas concentration at 9 m	A6: Gas concentration at 12 m	A7: Orifice negative pressure	A8: Extraction negative pressure
1	0.39583	0.50161	1	0.97431	0.92338	0.4751	0.47368	0.41935
2	0.53125	0.56935	0.92251	0.95786	0.91749	0.45607	0.52632	0.35484
3	0.17708	0.67742	0.98524	1	0.91061	0.46391	0.68421	0.64516
4	0.36458	0.54677	0.48708	0.51901	0.50688	0.29771	0.47368	0.67742
5	0.25	0	0.21125	0.1665	0.12574	0	0	0.41935
6	0.20833	0.59032	0.37823	0.53649	0.67289	0.47398	0.63158	0.51613
7	0.45833	0.82419	0.43266	0.47688	1	0.54393	1	1
8	0.17708	0.3371	0	0.59198	0.54813	0.24902	0.42105	0.09677
9	0.39583	0.70968	0.21033	0.10483	0.8723	0.47062	0.63158	0.45161
10	0.27083	0.56935	0.11808	0.89928	0.85069	0.48349	0.63158	0.74194
11	0.01042	1	0.46863	0.4409	0.38114	0.54561	0.78947	0.06452
12	1	0.86129	0.29982	0.19836	0.14931	0.85506	0.68421	0.45161
13	0	0.92581	0.33672	0.23741	0.18762	1	0.73684	0.25806
14	0.29167	0.59677	0.14668	0.03392	0	0.57527	0.52632	0
15	0	0.54839	0.12177	0	0.24951	0.45215	0.47368	0.74194

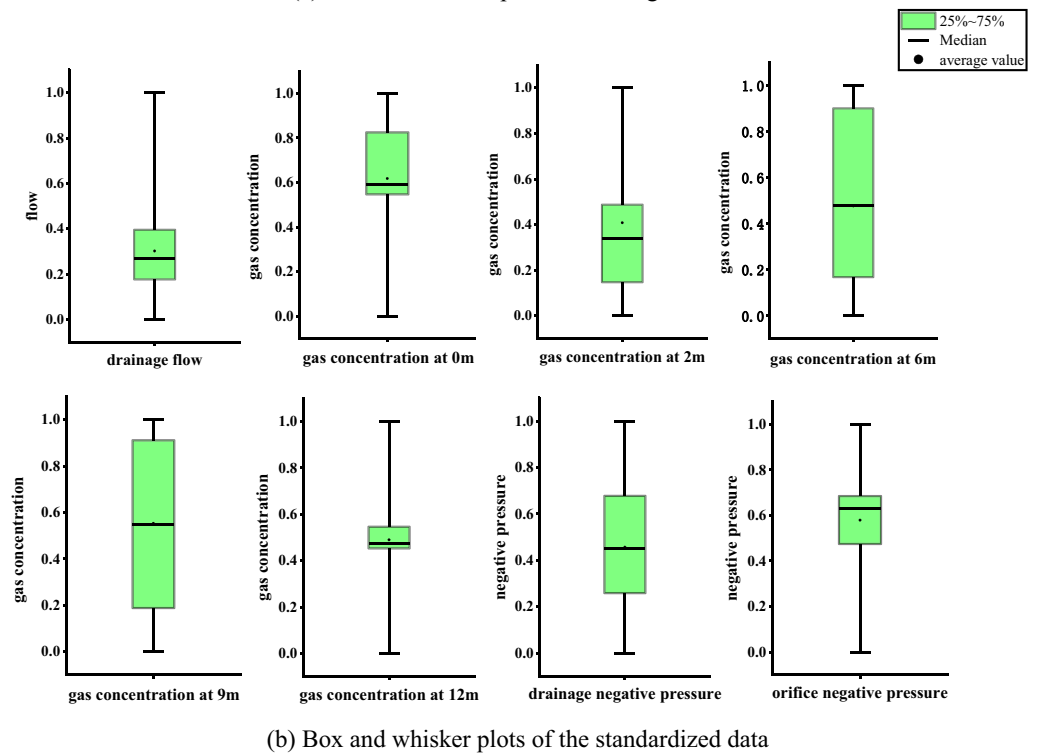
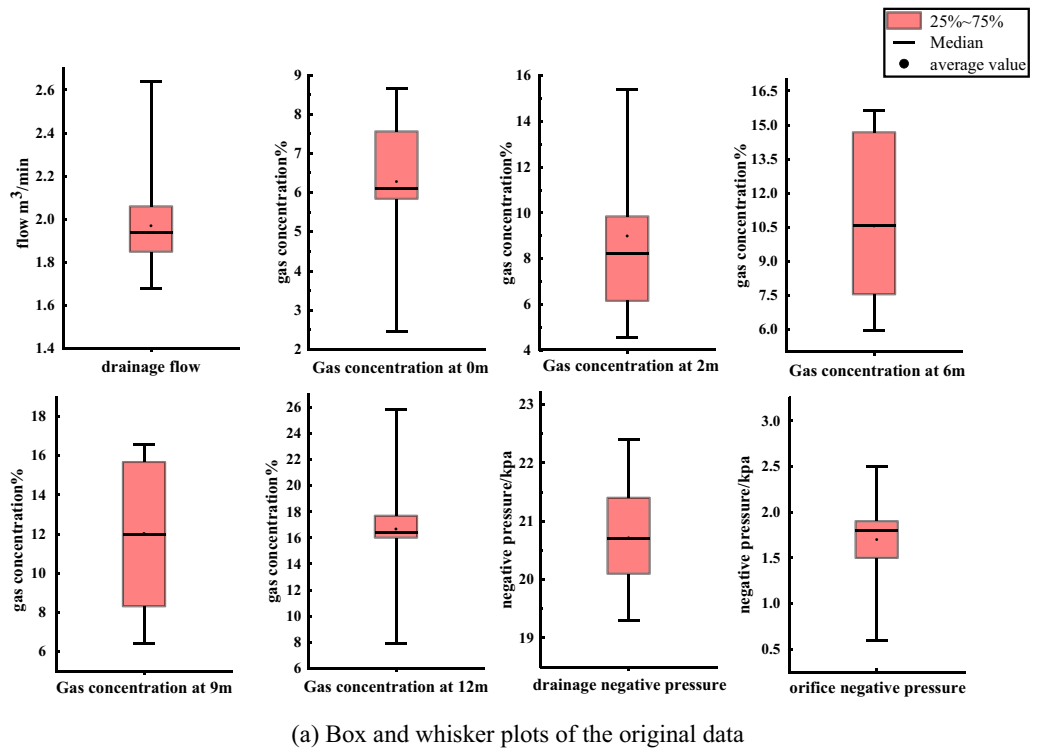
**Table 3.** Standardized data for gas extraction boreholes.

and distribution trend of the standardized data were consistent with the original data. After data standardization, the range of the original data was reduced to [0, 1], and the influence of each data dimension was eliminated, thereby optimizing the data for subsequent PCA to determine the new index for the identification of leaks.

**New indexes of the model for the identification of leaks.** In this study, PCA was used to linearly combine several representative new indexes for identification of leaks. The correlation between the original feature information should be considered to determine whether the PCA is applicable<sup>39,40</sup>. The Kaiser–Meyer–Olkin (KMO) and Butterley sphericity tests were applied in SPSS, as shown in Table 4.

As shown in Table 4, the value of Bartlett's test statistic was 65.343, and the significance level was approximately 0, which was less than the statistical significance level ( $\alpha = 0.05$ ) specified by SPSS. Thus, the original hypothesis was rejected. That is, the variables in the original data had a statistically significant influence, and the KMO test value was greater than 0.5, which indicated that the air leakage data of the gas drainage borehole were suitable for PCA.

According to the standardized data for air leakage of the gas drainage borehole in Table 3, the correlation coefficient matrix of air leakage characteristics was calculated, as shown in Table 5. The closer the correlation



**Figure 6.** Data comparison.

coefficient is to 1, the greater the degree of correlation of the corresponding two groups of characteristics; e.g., the correlation coefficient of gas concentrations at 6 m and 9 m is 0.7447, which indicates a strong correlation. The closer the correlation coefficient is to 0, the smaller the degree of correlation of the corresponding two groups of characteristics; e.g., the correlation coefficient of the gas concentrations at 2 m and 12 m is 0.0593, which indicates a weak correlation. A negative correlation coefficient indicates that the two groups are inversely correlated. For example, the correlation coefficient between the gas concentrations at 9 m and 12 m is  $-0.1529$ .



KMO test values for sampling adequacy	0.664
<b>Bartlett sphericity test</b>	
Test value	65.343
Degrees of freedom	28
Significance level	0

**Table 4.** KMO and Bartlett tests.

Index	Extraction flow	Gas concentration at 0 m	Gas concentration at 2 m	Gas concentration at 6 m	Gas concentration at 9 m	Gas concentration at 12 m	Orifice negative pressure	Extraction of negative pressure
Extraction flow	1							
Gas concentration at 0 m	0.0746	1						
Gas concentration at 2 m	0.1594	0.1562	1					
Gas concentration at 6 m	0.0634	-0.0516	0.6955	1				
Gas concentration at 9 m	0.1202	0.0743	0.5351	0.7447	1			
Gas concentration at 12 m	0.1725	0.8366	0.0593	-0.1431	-0.1529	1		
Orifice negative pressure	0.0887	0.9019	0.1599	0.1281	0.3572	0.6947	1	
Extraction of negative pressure	0.1716	0.0028	0.1306	0.1927	0.4870	-0.0959	0.2453	1

**Table 5.** Correlation coefficient matrix for air leakage characteristics of gas drainage boreholes.

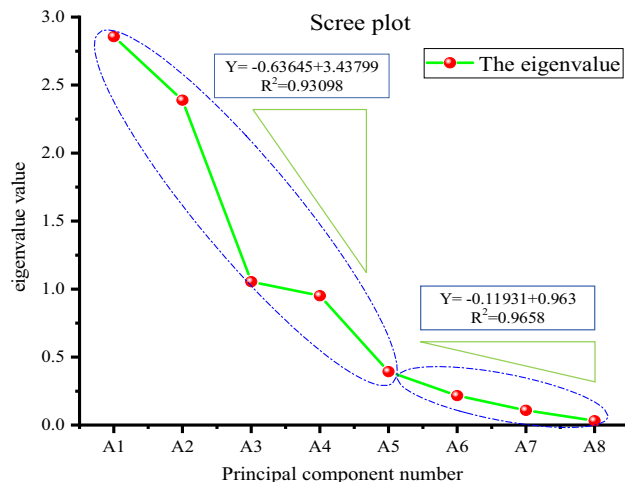
Principal component number	Index	Eigenvalue	Variance proportion	Cumulative contribution rate
A1	Extraction flow	2.85632	35.70%	35.70%
A2	Gas concentration at 0 m	2.38875	29.86%	65.56%
A3	Gas concentration at 2 m	1.05432	13.18%	78.74%
A4	Gas concentration at 6 m	0.95071	11.88%	90.63%
A5	Gas concentration at 9 m	0.39309	4.91%	95.54%
A6	Gas concentration at 12 m	0.21658	2.71%	98.25%
A7	Orifice negative pressure	0.10898	1.36%	99.61%
A8	Extraction of negative pressure	0.03125	0.39%	100.00%

**Table 6.** Eigenvalues of correlation coefficients.

As shown in Table 5, some of the 8 selected gas drainage borehole leakage characteristics are strongly correlated. Using these 8 kinds of characteristic data to identify gas drainage borehole leakage will lead to an incorrect decision, thus affecting the accuracy of the identification model. Therefore, it is necessary to analyse the training sample data via PCA to obtain the eigenvalue and contribution of each feature and select the appropriate principal component to eliminate the strong correlations from the feature data.

The eigenvalues and cumulative contribution rates of the eight types of feature information were obtained through calculations and analysis, as shown in Table 6. The characteristic value of  $A_1$ : extraction flow was the largest, and the contribution of its variance contribution was also the largest. The characteristic value and contributions of  $A_2$ – $A_8$  decreased in turn, and the contributions were small; the 6th–8th principal component,  $A_6$ – $A_8$ , were ignored. The cumulative contribution rate of extraction flow and gas concentrations at 0 m, 2 m, 6 m and 9 m reached 95.54%. According to Eq. (9), the contributions of these variables were more than 85%, and they were preliminarily considered as the main identification indexes of the improved naive Bayesian extraction leakage identification model. In a practical sense, the flow rate and concentration are the main variables of gas extraction in the borehole. The concentrations at 0 m, 2 m, 6 m and 9 m can reflect concentration changes in the borehole. The negative pressure has a linear relationship with the concentration and flow rate, and a change in negative pressure affects the concentration and flow rate; thus, it is advisable to select 5 principal components.

To further confirm the rationality of this selection, a scree plot was used. A scree plot is a trend map that reflects changes in data characteristics. The steepness of the decrease in eigenvalues shows whether the selected



**Figure 7.** Principal component scree plot.

Index	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>	Y <sub>4</sub>	Y <sub>5</sub>
A <sub>1</sub>	0.15881	0.03039	0.57660	0.76498	-0.20849
A <sub>2</sub>	0.44685	-0.39790	-0.10068	-0.06039	0.00492
A <sub>3</sub>	0.34194	0.33001	-0.34096	0.32301	0.62830
A <sub>4</sub>	0.30134	0.46679	-0.30251	0.08705	-0.21162
A <sub>5</sub>	0.37329	0.42654	0.04862	-0.19592	-0.43937
A <sub>6</sub>	0.35162	-0.46037	-0.06492	0.15732	0.19106
A <sub>7</sub>	0.50706	-0.25446	0.04217	-0.23946	-0.24950
A <sub>8</sub>	0.21746	0.23521	0.66429	-0.42281	0.47451

**Table 7.** Component analysis matrix.

features are correct and reasonable. The scree plot in Fig. 7 shows that the slope  $k_1$  of principal components A<sub>1</sub>–A<sub>5</sub> is  $-0.63645$ , and the trend is steep. The slope  $k_2$  of principal components A<sub>5</sub>–A<sub>8</sub> is  $-0.11931$ , and the trend is relatively flat. Principal component A<sub>5</sub> is an inflection point, and thus, it is reasonable to select these five variables as the principal components.

The original A<sub>1</sub>–A<sub>8</sub> feature information with strong correlations was reconstructed into the selected principal component features Y<sub>1</sub>–Y<sub>5</sub>, and the component analysis matrix table of Y<sub>1</sub>–Y<sub>5</sub> was established according to the PCA (Table 7) to establish the new feature information index of gas drainage borehole leakage.

Among the new indexes Y<sub>1</sub>–Y<sub>5</sub>, the higher the load coefficient corresponding to the original feature is, the closer the relationship between the feature information and the new indicator, which is the main influence quantity in the new index. According to the component analysis matrix in Table 7, the new index coefficient expression of gas drainage borehole leakage is:

$$\begin{aligned}
 Y_1 &= 0.15881A_1 + 0.44685A_2 + 0.34194A_3 + 0.30134A_4 + 0.37329A_5 + 0.35162A_6 + 0.50706A_7 + 0.21746A_8 \\
 Y_2 &= 0.03039A_1 - 0.39790A_2 + 0.33001A_3 + 0.46679A_4 + 0.42654A_5 - 0.46037A_6 - 0.25446A_7 + 0.23521A_8 \\
 Y_3 &= 0.57660A_1 - 0.10068A_2 - 0.34096A_3 - 0.30251A_4 + 0.04862A_5 - 0.06492A_6 - 0.04217A_7 + 0.66429A_8 \\
 Y_4 &= 0.76498A_1 - 0.06039A_2 - 0.32301A_3 - 0.08075A_4 - 0.19592A_5 - 0.15732A_6 - 0.23946A_7 - 0.42281A_8 \\
 Y_5 &= -0.20849A_1 + 0.00492A_2 + 0.62830A_3 - 0.21162A_4 - 0.43937A_5 + 0.19106A_6 - 0.24950A_7 + 0.47451A_8
 \end{aligned}$$

According to the expression for the characteristic information and the PCA matrix (Table 7), the gas concentrations A<sub>2</sub>–A<sub>6</sub> in the first principal component Y<sub>1</sub> at different depths and the load coefficient of negative pressure at orifice A<sub>7</sub> were greater than those of the other indexes, and this was the main characteristic influence of the first principal component index Y<sub>1</sub>. Therefore, the principal component index of Y<sub>1</sub> was interpreted as the influencing factor of negative pressure-concentration hole leakage. The load coefficient of A<sub>3</sub>–A<sub>5</sub> in the second principal component index Y<sub>2</sub> was higher, so the second principal component index Y<sub>2</sub> was interpreted as the influencing factor of hole depth-concentration borehole leakage. By analogy, Y<sub>3</sub> was the influencing factor of negative pressure-flow borehole leakage, Y<sub>4</sub> is the influencing factor of single flow borehole leakage, and Y<sub>5</sub> was the negative pressure-hole gas concentration borehole leakage factor.

**Analysis of test sample results.** Y<sub>1</sub>, Y<sub>2</sub>, Y<sub>3</sub>, Y<sub>4</sub> and Y<sub>5</sub> were used as the new identification indexes of the improved naive Bayesian extraction borehole leakage model, and the prior probability under the new index

Borehole number	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	Type of air leak
1	1.78760	0.75049	-0.18532	-0.52607	0.10793	I
2	1.78703	0.67208	-0.12678	-0.38738	-0.00997	I
3	1.93963	0.68025	-0.19066	-0.84648	0.19778	I
4	1.29884	0.31448	0.26743	-0.40568	0.16031	I
5	0.34467	0.30730	0.30644	-0.09238	0.18911	I
6	1.39628	0.17619	0.08761	-0.46850	-0.03418	II
7	1.91326	0.20825	0.52494	-0.65004	-0.03102	II
8	0.83946	0.18236	-0.05390	-0.14265	-0.41292	II
9	1.30982	-0.05108	0.33858	-0.25551	-0.20570	II
10	1.57869	0.39450	0.26274	-0.49350	-0.25669	II
11	1.37513	-0.31145	-0.39515	-0.44456	-0.02569	III
12	1.49119	-0.51862	0.55055	0.35058	0.08350	III
13	1.42849	-0.65358	-0.19528	-0.34884	0.21311	III
14	0.78195	-0.56312	-0.01172	0.10143	0.00570	III
15	0.93027	-0.22577	0.35894	-0.47733	0.28984	III

**Table 8.** Improved naive Bayesian training samples.

Borehole number	A1: Extraction flow	A2: Gas concentration at 0 m	A3: Gas concentration at 2 m	A4: Gas concentration at 6 m	A5: Gas concentration at 9 m	A6: Gas concentration at 12 m	A7: Orifice negative pressure	A8: Extraction negative pressure	Type of air leak
1	1.98	6.52	15.10	15.20	15.90	16.12	1.70	20.80	I
2	2.05	5.26	13.58	14.27	14.85	15.56	1.70	20.60	I
3	1.89	6.23	14.88	14.94	15.64	15.76	1.80	20.90	I
4	1.95	6.34	6.95	14.56	14.72	15.24	1.90	20.90	II
5	2.04	7.79	8.68	8.75	15.31	16.05	1.80	21.20	II
6	2.12	5.46	5.65	12.05	12.34	12.79	1.50	19.70	II
7	2.05	6.26	6.33	6.52	14.72	15.29	1.80	20.80	II
8	1.78	9.59	10.81	10.93	11.94	17.25	2.20	21.50	III
9	1.87	5.24	5.26	5.34	5.64	16.52	1.70	19.70	III
10	1.79	5.73	6.18	6.27	9.46	15.26	1.60	19.60	III
11	5.67	24.68	24.89	25.35	25.48	25.64	9.70	20.10	No leakage
12	5.14	17.98	18.12	18.34	18.42	18.56	9.80	19.80	No leakage

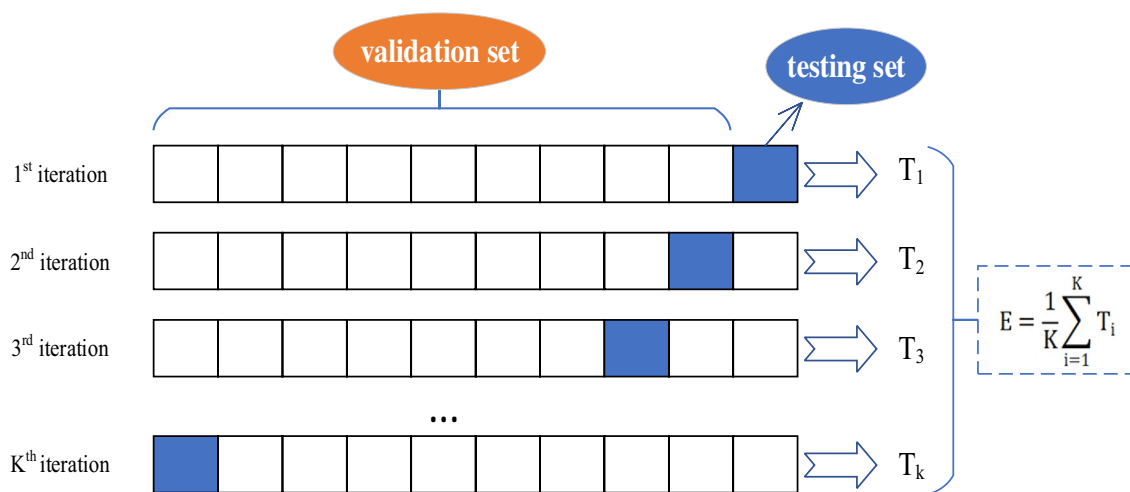
**Table 9.** Test sample data.

identification was calculated by Eqs. (12)–(13). According to the numerical relationship between the score of the new index and the leakage type of the above 15 groups of samples, the training was performed by MATLAB platform programming, as shown in Table 8. The final score of each index was taken as the training sample of the improved NBC.

To verify the accuracy and reliability of the model, 12 sets of gas drainage borehole data corresponding to three different types of leaks in the Xiashijie Coal Mine were collected as test samples, as shown in Table 9. A total of 240 data points from 30 groups of training samples and test samples were divided into a verification set and a test set according to a ratio of 0.6, which prevented a poor model identification rate and overfitting caused by a verification set that was too large as well as inaccurate model verification caused by a test set sample that was too small. In this study, we adopted the hold-out verification method, namely, the twofold cross-validation method. The schematic diagram of k-fold cross-validation is shown in Fig. 8<sup>41,42</sup>. The data set was divided into a training set and a test set for verification, and the average and accuracy of the final verification results were calculated.

We classified the 12 test sample data using single NBC identification and an improved naive Bayes extraction borehole identification model. The results of the analysis are shown in Table 10. Single NBC identification identified 3 boreholes with type I leaks, 5 boreholes with type II leaks, and 2 boreholes with type III leaks, but it could not identify boreholes without leaks. Improved naive Bayesian identification successfully identified 2 boreholes without leaks, 3 boreholes with type I leaks, 4 boreholes with type II leaks, and 3 boreholes with type III leaks.

Table 10 indicates that the single NBC identified the leakage of the No. 8 gas drainage borehole factors, which resulted in errors; this approach could not identify whether the gas drainage borehole was leaking. The recall rate was 75%, and the training time of the identification was 0.0045 s. The identification and analysis recall rate of the improved model was type II, and its real type was type III. This was because the mutual influence between the original eight characteristic Bayesian air leakage identification models of gas extraction boreholes



**Figure 8.** k-fold cross-validation schematic diagram.

Borehole number	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>	Y <sub>4</sub>	Y <sub>5</sub>	Actual type of air leakage	Single NBC identification type	Improved naive Bayesian identification type
1	1.92280	0.85626	0.07374	-0.38545	0.22717	I	I	I
2	1.60666	0.87215	0.23606	-0.10815	0.07676	I	I	I
3	1.83057	0.85743	-0.02364	-0.63877	0.25861	I	I	I
4	1.52035	0.54786	0.35919	-0.27686	-0.29476	II	II	II
5	1.68033	0.22065	0.69849	-0.11895	0.00684	II	II	II
6	0.69447	0.63193	0.41884	0.54398	-0.58930	II	II	II
7	1.19167	0.18151	0.76912	0.09367	-0.21040	II	II	II
8	2.09311	-0.16483	0.12256	-0.91361	0.38560	III	II	III
9	0.54409	-0.43729	0.12125	0.24340	0.05829	III	III	III
10	0.56498	-0.10154	-0.07867	-0.04215	-0.06021	III	III	III
11	0	0	0	0	0	No leakage	0	No leakage
12	0	0	0	0	0	No leakage	0	No leakage

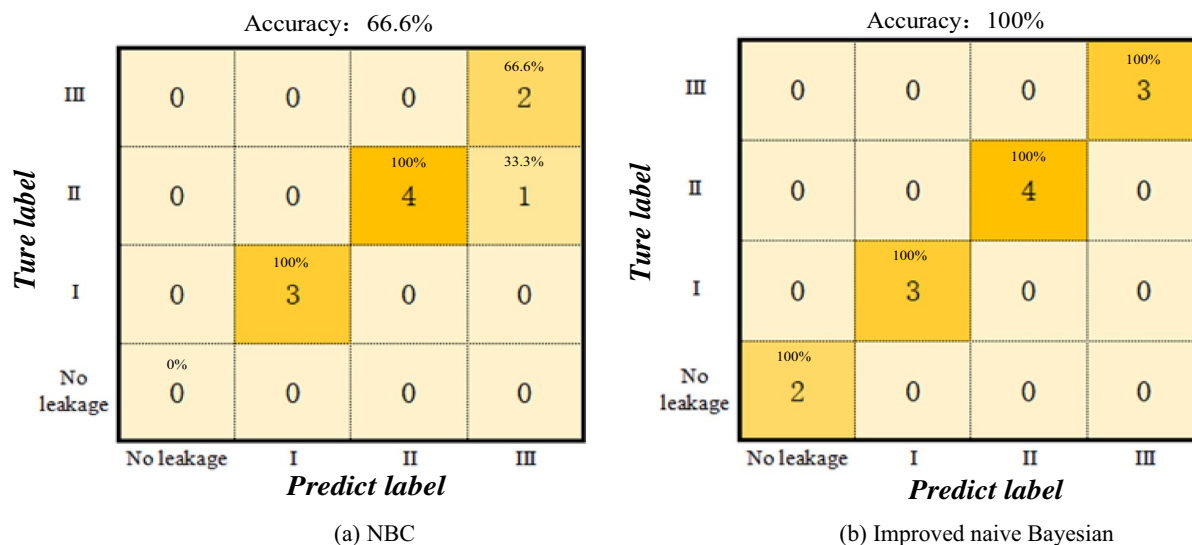
**Table 10.** Improved naive Bayesian identification results.

was 98.9%, the identification accuracy improved by 31.8%, and the training time decreased to 0.0020 s, which was an improvement of 55%. To further analyse the error rate, a confusion matrix comparison diagram was drawn<sup>43</sup>. It showed that the improved naive Bayesian gas extraction borehole leakage identification model could fully identify the type of borehole leakage in the Xiashijie coal mine, the identification accuracy was high, and the identification rate was fast.

As shown in Fig. 9, the single naive Bayesian identification failed to identify boreholes without leaks due to the inability to calculate eigenvalues. This resulted in the effective identification of only 10 out of 12 groups, and type II boreholes were mistakenly identified as type III boreholes. The improved naive Bayes model accurately identified 12 groups of boreholes, and the true value of the improved naive Bayesian model was consistent with the predicted value. Thus, the identification analysis showed that the improved naive Bayesian gas drainage borehole leakage identification model was superior to the single NBC identification analysis. Depending on its superiority, it could more accurately identify the type of air leak and provide further guidance for borehole sealing and repair to improve the efficiency of gas extraction and prevent gas disasters.

### Conclusions

- 1) Through multisource data fusion theory (MDF) and principal component analysis (PCA), the traditional naive Bayes method was improved, and an enhanced naive Bayes air leakage identification model of gas drainage boreholes was constructed. The new model overcame the shortcomings of the naive Bayes method that could not accommodate missing and nonstandard data and eliminated the misvaluation caused by the superposition of a large amount of feature information in the process of identification of leaks in gas drainage boreholes.
- 2) The model was applied to the 229 working face of the Xiashijie Coal Mine. Combined with 8 types of characteristic information of gas drainage boreholes. Thirty groups of 240 gas drainage borehole data were divided



**Figure 9.** Comparison diagram of the confusion matrix.

into training samples and test samples for analysis, and 12 groups of gas drainage borehole test sample data were successfully identified, including 2 boreholes without leaks, 3 boreholes with type I leaks, 4 boreholes with type II leaks, and 3 boreholes with type III leaks, which were consistent with the conditions of the actual gas drainage boreholes. Thus, this study provides a basis for improving gas drainage efficiency and ensuring safe mining in the Xiashijie Coal Mine.

- The feasibility of the model was verified by the hold-out method. The recall rate of model identification analysis was 98.9%, and the running time was 0.0020 s. Compared with the single naive Bayes method, the operation rate increased by 55%, and the identification accuracy increased by 31.8%. The improved model filled the gap related to the determination and identification of leaks in boreholes and provides a theoretical basis for the evaluation of the quality of sealing and borehole repairs.

### Data availability

All data generated or analysed during this study are included in this published article.

Received: 11 May 2022; Accepted: 14 September 2022

Published online: 27 September 2022

### References

- Cheng, L. *et al.* A sequential approach for integrated coal and gas mining of closely-spaced outburst coal seams: Results from a case study including mine safety improvements and greenhouse gas reductions. *J. Energies*. **11**(11), 3023 (2018).
- Niu, Y. *et al.* Experimental study and field verification of stability monitoring of gas drainage borehole in mining coal seam. *J. Pet. Sci. Eng.* **189**, 106985 (2020).
- Lin, B. *et al.* Significance of gas flow in anisotropic coal seams to underground gas drainage. *J. Pet. Sci. Eng.* **180**, 808–819 (2019).
- Liu, P., Jiang, Y. & Fu, B. A novel approach to characterize gas flow behaviors and air leakage mechanisms in fracture-matrix coal around in-seam drainage borehole. *J. Nat. Gas Sci. Eng.* **77**, 103243 (2020).
- Liu, P. *et al.* Evaluation of underground coal gas drainage performance: Mine site measurements and parametric sensitivity analysis. *J. Process Saf. Environ. Prot.* **148**, 711–723 (2021).
- Zhang, T. *et al.* Strain localization characteristics of perforation failure of perforated specimens. *J. China Coal Soc.* **45**(12), 4087–4094. <https://doi.org/10.13225/j.cnkj.jccs.2019.143> (2020) ((in Chinese)).
- Wang, K., Pan, H. & Zhang, T. Experimental study of prefabricated crack propagation in coal briquettes under the action of a CO<sub>2</sub> gas explosion. *J. ACS omega*. **6**(38), 24462–24472 (2021).
- Wang, K. *et al.* Experimental study on the radial vibration characteristics of a coal briquette in each stage of its life cycle under the action of CO<sub>2</sub> gas explosion. *J. Fuel*. **320**, 123922 (2022).
- Zhang, C. *et al.* Experimental research and field application of anti-sloughing support material in gas extraction borehole sealing section. *J. Min. Saf. Eng.* **38**(1), 199–205. <https://doi.org/10.13545/j.cnki.jmse.2020.0029> (2021) ((in Chinese)).
- Junxiang, Z., Bo, L. & Yuning, S. Dynamic leakage mechanism of gas drainage borehole and engineering application. *Int. J. Min. Sci. Technol.* **28**(3), 505–512 (2018).
- Zhang, J. *et al.* A fully multifield coupling model of gas extraction and air leakage for in-seam borehole. *J. Energy Rep.* **7**, 1293–1305 (2021).
- Fan, J. *et al.* A coupled methane/air flow model for coal gas drainage: Model development and finite-difference solution. *J. Process Saf. Environ. Prot.* **141**, 288–304 (2020).
- Zhang, Y., Zou, Q. & Guo, L. Air-leakage Model and sealing technique with sealing–isolation integration for gas-drainage boreholes in coal mines. *J. Process Saf. Environ. Prot.* **140**, 258–272 (2020).
- Wang, Z. *et al.* A coupled model of air leakage in gas drainage and an active support sealing method for improving drainage performance. *J. Fuel*. **237**, 1217–1227 (2019).
- Wang, H. *et al.* Study on sealing effect of pre-drainage gas borehole in coal seam based on air-gas mixed flow coupling model. *J. Process Saf. Environ. Prot.* **136**, 15–27 (2020).

16. Zhang, Y. *et al.* A novel failure control technology of cross-measure borehole for gas drainage: A case study. *J. Process Saf. Environ. Prot.* **135**, 144–156 (2020).
17. Liu, Q. *et al.* Application of the comprehensive identification model in analyzing the source of water inrush. *Arabian J. Geosci.* **11**(9), 1–10 (2018).
18. Hui, L. & Xiaojun, Z. Predictive analysis of impact hazard level of coal rock mass based on fuzzy inference network. *J. Intell. Fuzzy Syst.* **38**(2), 1509–1518 (2020).
19. Jiang, C. *et al.* Identification model and indicator of outburst-prone coal seams. *Rock Mech. Rock Eng.* **48**(1), 409–415 (2015).
20. Wang, H. & Zhang, Q. Dynamic identification of coal-rock interface based on adaptive weight optimization and multi-sensor information fusion. *Inf. Fusion.* **51**, 114–128 (2019).
21. Li, N., Feng, X. & Jimenez, R. Predicting rock burst hazard with incomplete data using Bayesian networks. *Tunn. Undergr. Space Technol.* **61**, 61–70 (2017).
22. Li B., Wu Q., Liu Z. Identification of mine water inrush source based on PCA-FDA: Xiandewang coal mine case. *J. Geofluids.* **2020**, (2020).
23. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *C. Proceedings of the 23rd international conference on Machine learning.* 161–168(2006).
24. Dimitrios, P. & Andreas, B. Enhancing machine learning algorithms to assess rock burst phenomena. *Geotech. Geol. Eng.* **39**(8), 5787–5809 (2021).
25. Huang, P. *et al.* Research on piper-PCA-bayes-LOOCV discrimination model of water inrush source in mines. *Arabian J. Geosci.* **12**(11), 1–14 (2019).
26. Ba Q. Research on gas leakage mechanism and detection technology of coal mine gas drainage. *C. IOP Conference Series: Earth and Environmental Science. IOP Publishing.* **446**(5), 052018 (2020).
27. Zhang, X. *et al.* Study on the influence mechanism of air leakage on gas extraction in extraction boreholes. *J. Energy Explor. Exploit.* **40**(5), 1344–1359 (2022).
28. Zheng, C. *et al.* Effects of coal properties on ventilation air leakage into methane gas drainage boreholes: Application of the orthogonal design. *J. Nat. Gas Sci. Eng.* **45**, 88–95 (2017).
29. Hao, J. *et al.* Analysis of gas leakage field and location determination of gas leakage in surrounding rock of gas extraction borehole. *J. Coal Eng.* **51**(5), 143–147. <https://doi.org/10.11799/ce201905033> (2019) ((in chinese)).
30. Zhou, H., Shen, K. & Chen, B. Classification of leakage types and application of efficient holesealing technology for gas drainage drilling. *J. Min. Saf. Prot.* **46**(01), 33–3642. <https://doi.org/10.3969/j.issn.1008-4495.2019.01.008> (2019) ((in chinese)).
31. Ping, G. Study on leakage model of gas extraction borehole and optimization of sealing process. *J. Coal Technol.* **39**(06), 82–85. <https://doi.org/10.13301/j.cnki.ct.2020.06.025> (2020) ((in chinese)).
32. Zhao, Y. & Tian, S. Identification of hidden disaster causing factors in coal mine based on Naive Bayes algorithm. *J. Intell. Fuzzy Syst.* **41**(2), 2823–2831 (2021).
33. He, Y. *et al.* Rock hardness identification based on optimized PNN and multi-source data fusion. *J. Proc. Inst. Mech. Eng., Part. C-J. Mech. Eng. Sci.* **236**(7), 3701–3716 (2022).
34. Uddin, M. P., Mamun, M. A. & Hossain, M. A. Effective feature extraction through segmentation-based folded-PCA for hyperspectral image classification. *Int. J. Remote Sens.* **40**(18), 7190–7220 (2019).
35. Ju, Q. & Hu, Y. Source identification of mine water inrush based on principal component analysis and grey situation decision. *J. Environ. Earth Sci.* **80**(4), 1–14 (2021).
36. Zhou, F., Wang, X. & Liu, Y. Gas drainage efficiency: An input–output model for evaluating gas drainage projects. *J. Nat. Hazard.* **74**(2), 989–1005 (2014).
37. Cai, J. *et al.* Numerical analysis of multi-factors effects on the leakage and gas diffusion of gas drainage pipeline in underground coal mines. *J. Process Saf. Environ. Prot.* **151**, 166–181 (2021).
38. Pérez-Ortiz, J. A. *et al.* Kalman filters improve LSTM network performance in problems unsolvable by traditional recurrent nets. *J. Neural Netw.* **16**(2), 241–250 (2003).
39. Öcal, M. E. *et al.* Industry financial ratios—application of factor analysis in Turkish construction industry. *J. Build. Environ.* **42**(1), 385–392 (2007).
40. Zhang, J. *et al.* Investigation of carbon dioxide emission in China by primary component analysis. *J. Sci. Total Environ.* **472**, 239–247 (2014).
41. Ji, J. *et al.* Application of GSK-XGBOOST Model in prediction of bottom hole air temperature. *J. China Saf. Sci. Technol.* **18**(03), 131–136. <https://doi.org/10.11731/j.issn.1673-193x.2022.03.020> (2022) ((in chinese)).
42. Liu, Y. & Wang, Y. Review of various cross-validation estimation methods of generalization error. *J. Appl. Res. Comput.* **32**(5), 1287–1290, 1297. <https://doi.org/10.3969/j.issn.1001-3695.2015.05.002> (2015) ((in chinese)).
43. Yang, X. Survey for performance measure index of classification learning algorithm. *J. Compu. Sci.* **48**(8), 209–219. <https://doi.org/10.11896/jsjx.200900216> (2021) ((in chinese)).

## Acknowledgements

This research was financially supported by The National Natural Science Foundation of China (51874234, 52104215). The authors are also grateful to the anonymous reviewers for their constructive comments.

## Author contributions

The main research idea and manuscript preparation were contributed by H.P.; S.H. drafted the manuscript and verified the research; S.S. and T.Z. gave several suggestions from a professional point and supervised the manuscript; K.W. assisted on finalizing research work and manuscript. All authors have read and agree to the published version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022