fire

MDPI

Article

Prediction of Coal Spontaneous Combustion Hazard Grades Based on Fuzzy Clustered Case-Based Reasoning

Qiuyan Pei, Zhichao Jia *, Jia Liu, Yi Wang *, Junhui Wang and Yanqi Zhang

College of Safety and Emergency Management Engineering, Taiyuan University of Technology, Taiyuan 030024, China; peiqiuyan@tyut.edu.cn (Q.P.); wangjunhui@tyut.edu.cn (J.W.); zhangyanqi1612@link.tyut.edu (Y.Z.)

* Correspondence: jiazhichao1560@link.tyut.edu.cn (Z.J.); wangyi@tyut.edu.cn (Y.W.)

Abstract: Accurate prediction of the coal spontaneous combustion hazard grades is of great significance to ensure the safe production of coal mines. However, traditional coal temperature prediction models have low accuracy and do not predict the coal spontaneous combustion hazard grades. In order to accurately predict coal spontaneous combustion hazard grades, a prediction model of coal spontaneous combustion based on principal component analysis (PCA), case-based reasoning (CBR), fuzzy clustering (FM), and the snake optimization (SO) algorithm was proposed in this manuscript. Firstly, based on the change rule of the concentration of signature gases in the process of coal warming, a new method of classifying the risk of spontaneous combustion of coal was established. Secondly, MeanRadius-SMOTE was adopted to balance the data structure. The weights of the prediction indicators were calculated through PCA to enhance the prediction precision of the CBR model. Then, by employing FM in the case base, the computational cost of CBR was reduced and its computational efficiency was improved. The SO algorithm was used to determine the hyperparameters in the PCA-FM-CBR model. In addition, multiple comparative experiments were conducted to verify the superiority of the model proposed in this manuscript. The results indicated that SO-PCA-FM-CBR possesses good prediction performance and also improves computational efficiency. Finally, the authors of this manuscript adopted the Random Balance Designs—Fourier Amplitude Sensitivity Test (RBD-FAST) to explain the output of the model and analyzed the global importance of input variables. The results demonstrated that CO is the most important variable affecting the coal spontaneous combustion hazard grades.

Keywords: coal spontaneous combustion; mine fire prevention; signature gases; intensity classification; case-based reasoning; data balancing

Citation: Pei, Q.; Jia, Z.; Liu, J.; Wang, Y.; Wang, J.; Zhang, Y. Prediction of Coal Spontaneous Combustion Hazard Grades Based on Fuzzy Clustered Case-Based Reasoning. *Fire* **2024**, *7*, 107. https://doi.org/10.3390/fire7040107

Academic Editor: Thomas H. Fletcher

Received: 2 February 2024
Revised: 19 March 2024
Accepted: 22 March 2024
Published: 24 March 2024

1. Introduction

Coal spontaneous combustion, as a common coal mine accident, seriously threatens the lives of coal mine workers and the property safety of mining equipment [1–3]. In addition, coal spontaneous combustion also pollutes the soil and destroys the ecological environment [4–6]. In recent years, with the depletion of shallow mineral resources, more and more underground coal mine projects are going deeper underground at an unprecedented speed [7–9], causing coal spontaneous combustion to become a serious threat to many projects worldwide [10–13]. To prevent and control coal spontaneous combustion disasters, it is necessary to study the effective prediction of the coal spontaneous combustion method.

The gas analysis method, as a commonly used method for predicting spontaneous coal combustion [14], has the advantage of strong operability [15], and is widely used in the prediction of coal spontaneous combustion. This method mainly tests the signature gases generated during the coal heating process and the concentration and finds the variation relationship between it and the coal temperature, thereby indirectly predicting

the actual temperature [16]. However, it was found that the relationship between the signature gases' concentration and coal temperature is non-linear [17], and it is very difficult to describe this relationship through the most commonly used mathematical methods. To solve this problem, scholars have applied machine learning to the prediction of spontaneous coal combustion, which, as a branch of artificial intelligence, can better mine the nonlinear relationship between indicators and samples [18]. Zhang [19] proposed a prediction model based on RF and MLP, which can accurately predict coal temperature. However, the model is greatly affected by the value of hyperparameters. Guo [20] and Wang [21] used PSO to calculate the hyperparameters in the GRU and BPNN algorithms, respectively, and established PSO-GRU and PSO-BPNN temperature prediction models. The results show that the models have good prediction ability. Li [22] improved the optimization ability of the GA algorithm and combined it with a neural network to establish a temperature prediction model. The results show that the improved GA algorithm can improve the prediction accuracy of the model. Nonetheless, the aforementioned models have the following limitations: (1) a large number of training samples are required, and the calculation complexity of the model is high, leading to prolonged computational time; (2) most machine learning models do not have self-learning abilities; and (3) they are prone to overfitting during modeling. Therefore, further research is needed to explore new prediction methods. In addition, all of the above studies are quantitative, focusing on predicting coal temperatures, and there are fewer studies on predicting the coal spontaneous combustion hazard grades.

As a mature branch of artificial intelligence, case-based reasoning (CBR) has been widely applied in other fields [23]. CBR has greater classification performance compared with traditional data mining methods [24] and it has also shown excellent performance in fields like fault diagnosis [25–27], risk assessment [28,29], and forest fire prediction [30–32]. It should be noted that the weights of case characteristic attributes in CBR have a significant impact on the prediction performance of the model. However, the calculation of weights lacks a solid foundation and is strongly influenced by subjective factors. In addition, as the number of cases in case-based reasoning becomes larger, the computational cost of CBR gradually increases and the computational efficiency gradually decreases. To address this issue, scholars have applied clustering to the CBR case library [33–35], dividing the case library into several different clusters and limiting the CBR process to specific clusters, which reduces the comparison times and lowers the computational cost. However, this method may lead to the loss of cluster boundary information and result in a lower accuracy of model prediction [36]. In addition, concerning the determination of the coal spontaneous combustion hazard grades, scholars have proposed various methods for determining the coal spontaneous combustion hazard grades [37–39]. However, due to the different geological structures, mining environments, coal quality composition, and other factors in different coal mines, there are great differences in the numerical values of the same gas indexes, resulting in the absence of a fixed value for the threshold of the coal spontaneous combustion hazard grades, which also poses a challenge to the prediction of the coal spontaneous combustion hazard grades.

Although the CBR model possesses good prediction performance in other fields, it has been seldom applied to coal spontaneous combustion hazard grade prediction. Furthermore, there are efficiency bottlenecks and a lack of basis for the weights of case characteristic attributes in the CBR model. In addition, there are some limitations in the current method of classifying the risk level of spontaneous coal combustion. There are also limitations in the current methods of classifying the coal spontaneous combustion hazard grades. Therefore, to address the shortcomings of existing research, the following studies were conducted by the authors of this manuscript: (1) Through analyzing the change rule of signature gases in the process of coal warming, the method of coal spontaneous combustion hazard grade classification was proposed; (2) adaption of PCA to calculate the weights of case characteristic attributes; (3) application of fuzzy clustering to the CBR model to

construct PCA-FM-CBR models; (4) use of SO algorithm to optimize the hyperparameters in PCA-FM-CBR models; (5) use of RBD-FAST to analyze the sensitivity of input variables.

2. Dataset Preparation

2.1. Dataset Collection

The data in this paper are from experimental data on the warming of the coal autogenous combustion program [40], which was published publicly by Jiang during his research. The coal samples used in this experiment were from the Dongtan mine coal in Shandong Province, China and the experimental steps are as follows: (1) The coal sample is crushed, 200 g coal sample is selected, including different particle sizes, and mixed to obtain the mixed coal sample; (2) a 1000 g mixed coal sample is placed into the programmed heating device for heating, where the heating rate is 0.3 °C/min, and the air supply is 120 mL/min; (3) the gas product is determined, and heating is stopped when the temperature rises to a predetermined temperature.

A device for coal spontaneous combustion temperature programming was used to heat coal samples with different particle sizes and test the gas products produced by different coal samples. A total of 337 sets of coal spontaneous combustion data were obtained from this experiment, where each set of data included CO, CH_4, and CO_2 characteristic indicators. However, the selected index gas values should not only change with the change in temperature but also have an accurate relationship with the coal temperature, due to the presence of CO_2 in the tunnel and the respiratory gases of the workers, and because CH_4 is originally stored in the coal seam, these two indicators are subject to large external influences, and there is no accurate relationship between them and the coal temperature; therefore, we did not choose the CO_2-related indicators or CH_4-related indicators. In addition, when we selected the indicators, we reviewed a large number of studies, and many scholars chose CO, $CO/\Delta O_2$, C_2H_4/C_2H_6, C_2H_4, and O_2 when selecting the indicators [15,41,42], and they all thought that these five indicators could reflect the danger level of spontaneous combustion of coal.

2.2. Classification of Coal Spontaneous Combustion Hazard Grades

This paper is based on the classification standard of the coal spontaneous combustion hazard class proposed by scholars, combined with the law of collecting the signature gas in the data of coal spontaneous combustion. Article 261 of the Coal Mine Safety Regulations requires the determination of the signifying gas of natural ignition of coal seams as well as the critical value. Article 265 stipulates that the operation must be stopped when there is a sign of ignition, and Article 275 stipulates that when there is a fire, it should be extinguished immediately according to the nature of the fire and other circumstances. These three articles qualitatively describe the signature gases of spontaneous coal combustion, signs of ignition, and fire, but lack a quantitative method of judgment. Pan [43] divided the stage of spontaneous coal combustion into four stages, but the division criteria are vague and not detailed enough; therefore, we took Pan's criteria as the basis, combined with the division criteria established in the papers published by Duo [41] and Fei [44], and obtained the division criteria in Table 1, which divide the stage of spontaneous coal combustion into six stages, with a higher concentration of O_2 and a lower concentration of CO as the first stage. When CO appears, it is the second stage, when C_2H_4 gas appears, it is the third stage, when $CO/\Delta O_2$ shows an increasing state, it is the fourth stage, a constant C_2H_4/C_2H_6 value, and a decreasing O_2 concentration is the fifth stage, and a maximum C_2H_4/C_2H_6 value is the sixth stage.

Figure 1a shows that the O_2 concentration with the increase in temperature shows an overall decreasing trend and when the temperature is less than 50 degrees Celsius, the O_2 concentration is higher, which indicates that the spontaneous combustion of coal at this time is in the first stage. Figure 1b shows that with the increase in temperature, the CO concentration shows an overall increasing trend and when the temperature is greater than 50 degrees Celsius, the CO concentration begins to fluctuate, which indicates

that the spontaneous combustion of coal is in the second stage. Figure 1c shows that the C_2H_4 concentration increases with temperature, and when the temperature is greater than 60 °C, the C_2H_4 concentration fluctuates slightly, which indicates that the spontaneous combustion of coal is in the third stage. Figure 1d shows that $CO/\Delta O_2$ increases with temperature, and when the temperature is less than 100 °C, the $CO/\Delta O_2$ value increases with temperature. When the temperature is greater than 100 °C, $CO/\Delta O_2$ shows a rapid increase with the increase in temperature, which indicates that the spontaneous combustion of coal is in the fourth stage. Figure 1e shows that the C_2H_4/C_2H_6 concentration with the increased temperature shows an overall increasing trend and when the temperature is between 120 °C and 230 °C, C_2H_4/C_2H_6 shows an increasing trend with the increase in temperature, which indicates that the spontaneous combustion of coal is in the fifth stage, and when the temperature is greater than 230 °C, the highest value appears, which indicates that the spontaneous combustion of coal is in the sixth stage.

Table 1. Corresponding table of coal spontaneous combustion stage classification and gas characteristics.

Coal Spontaneous Combustion Stage Name	Gas Characteristics
The first stage	Higher O_2 concentrations and no CO
The second stage	CO starts to appear
The third stage	C_2H_4 starts to appear
The fourth stage	$CO/\Delta O_2$ values show an increasing trend
The fifth stage	C_2H_4/C_2H_6 values show an increasing trend
The sixth stage	Maximum C_2H_4/C_2H_6 value

According to the non-linear relationship between gas and temperature shown in Figure 1, combined with the classification criteria in Table 1, the six stages of coal spontaneous combustion correspond to six warning grades, thus selecting 50 °C, 60 °C, 100 °C, 120 °C, 230 °C as the temperature thresholds and classifying combustion hazards. The six stages of coal spontaneous combustion correspond to six warning grades, so that 50 °C, 60 °C, 100 °C, 120 °C and 230 °C are selected as the temperature thresholds, and the combustion hazard grades are divided into six levels: green, blue, purple, yellow, orange and red. The coal spontaneous combustion hazard grades are shown in Table 2.

Table 2. Standard for rockburst intensity classification.

Coal Spontaneous Combustion Stage Name	Temperature Range/°C	Coal Spontaneous Combustion Hazard Grades
The first stage	$T < 50$	Green warning (0)
The second stage	$T \in [50, 60)$	Blue warning (1)
The third stage	$T \in [60, 100)$	Purple warning (2)
The fourth stage	$T \in [100, 120)$	Yellow warning (3)
The fifth stage	$T \in [120, 230)$	Orange warning (4)
The sixth stage	$T \geq 230$	Red warning (5)

Figure 2 shows that the data in the coal spontaneous combustion hazard grades database exhibit obvious unbalancedness. The proportions for Grades 0–5 are 3.2% (11 cases), 9.2% (31 cases), 30.3% (102 cases), 29.1% (98 cases), 18.9% (64 cases) and 9.2% (31 cases), respectively.

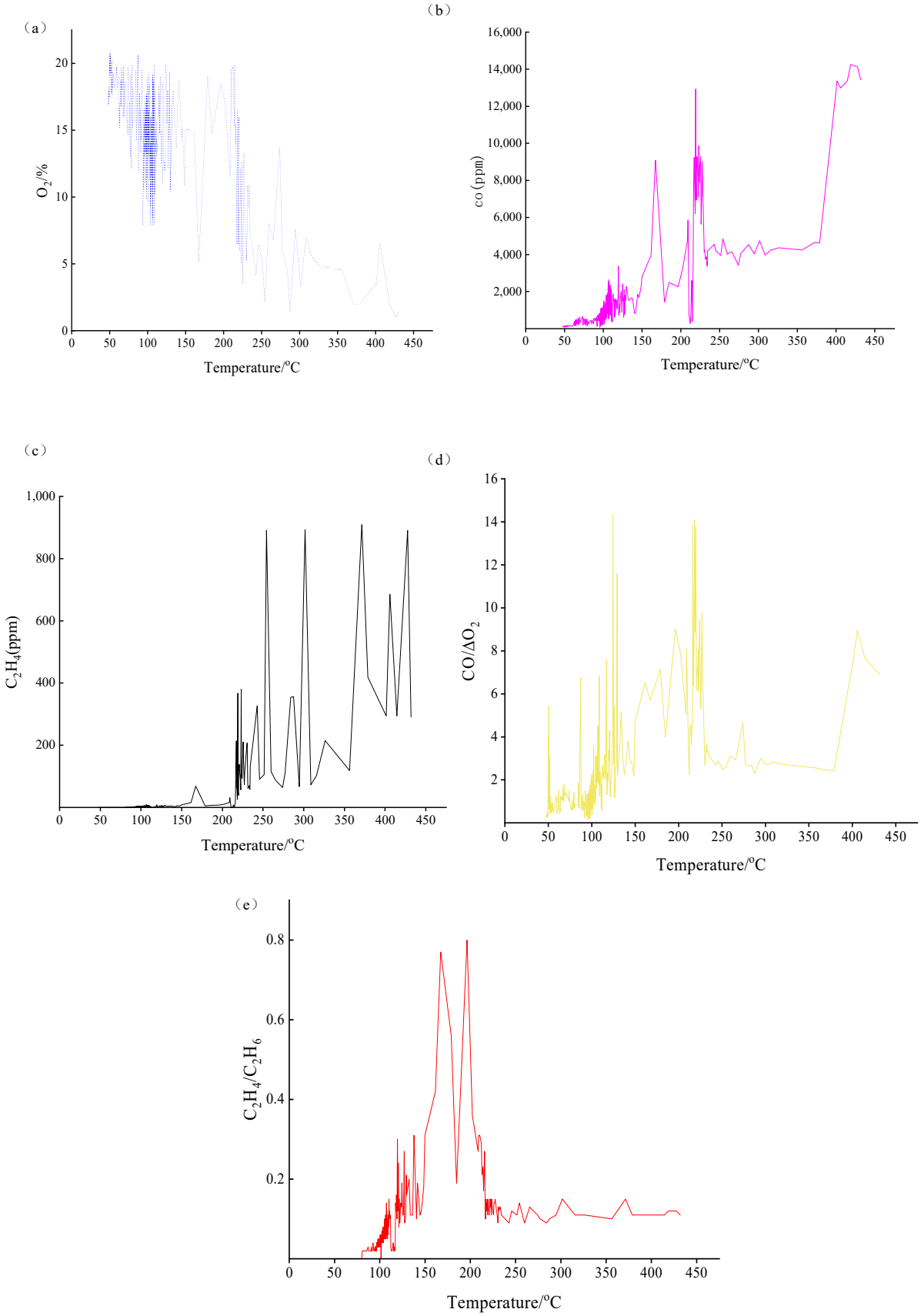

Figure 1. Signature gas change curve. (**a**) O_2 concentration variation curve; (**b**) co concentration variation curve; (**c**) C_2H_4 concentration variation curve; (**d**) $CO/\Delta O_2$ value variation curve; (**e**) C_2H_4/C_2H_6 value variation curve.

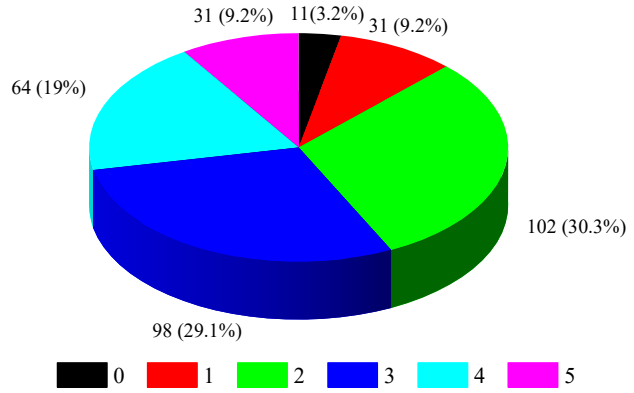

Figure 2. Pie chart for the proportion of each coal spontaneous combustion label in the coal spontaneous combustion database.

2.3. Data PreProcessing

2.3.1. Making Data Dimensionless

In the hope of eliminating the impact of dimension differences between characteristic variables on the prediction results, the coal spontaneous combustion database was made dimensionless through an averaging method using Equation (1).

$$x_{ij}^{*} = \frac{x_{ij}}{\overline{x}_j} \tag{1}$$

where x_{ij} represents the original data of the jth index of the ith sample; x_{ij}^{*} represents the data after being made dimensionless; and $\overline{x}_j$ represents the mean value of the jth index.

2.3.2. MeanRadius-SMOTE

As shown in Section 2.2, the initial coal spontaneous combustion dataset is imbalanced, which may lead machine learning models to misclassify minority class samples as majority class samples, thereby affecting their prediction performance. Hence, it is necessary to perform over-sampling with the initial coal spontaneous combustion database. In this manuscript, the MeanRadius-SMOTE algorithm was used to generate new data to achieve balance between various coal spontaneous combustion data [45].

The MeanRadius-SMOTE algorithm modifies the generation rule of the SMOTE algorithm by considering the radius and geometric center when generating new data. As a result, the new samples are more likely to be distributed around the average radius of the minority class samples. This algorithm is not only efficient for datasets of any shape dataset, but also the generated new data are more likely to be distributed near the average radius of the minority class samples, which can improve the ability of machine learning models to identify the decision boundary.

The steps of MeanRadius-SMOTE for generating new data are as follows:

(1) Calculate the geometric center of each class of minority class samples and represent it as x_c.
(2) Calculate the Euclidean distance from each minority class sample to the sample center and then calculate the average distance, represented as d_m.
(3) Randomly select k minority class samples and obtain k vectors v_i from the sample center to the samples. Calculate the composite vector of the k vector.
(4) Determine the distance between the new sample and the sample center according to the average value d_m and the parameter θ. Generate new samples based on Equation (2).

$$x_{new} = x_c + r * \sum_{i=0}^{k} v_i \quad r \sim \left(\frac{d_m}{\theta}, d_m\right) \tag{2}$$

(5) Repeat Steps 3 and 4 until the sample size of the majority and minority class is balanced.

In order to ensure that the generated data are valid, we assigned k as 3 and r as 2 (which is obtained through multiple experiments). After balancing the dataset, 91 new green warning data, 71 blue warning data, 4 yellow warning data, 38 orange warning data, and 71 red warning data were newly generated. The new coal spontaneous combustion database has a total of 510 coal spontaneous combustion data, and the quantity ratio for coal spontaneous combustion of Grades 0–5 is 1:1:1:1.

3. Machine Learning Modelling

3.1. Overview of the Machine Learning Models

3.1.1. Case-Based Reasoning

Case-based reasoning (CBR) is a machine learning algorithm proposed by Aamodt and Plaza et al. in 1994 that mimics the analogical reasoning in the human brain [46]. CBR consists of four basic processes: case representation, case retrieval, case reuse and case retain [30,47]. The schematic diagram of CBR is depicted in Figure 3, and the specific steps are as follows:

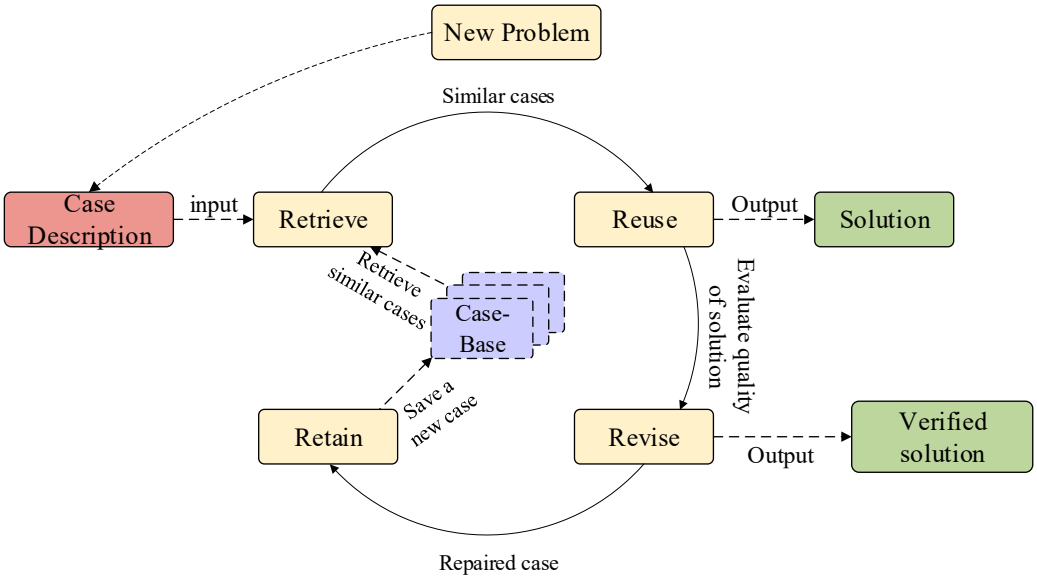

Figure 3. Schematic diagram of CBR.

Step 1: Assume that the source cases in the historical database are represented in the following binary format:

$$C_k = \{X_k, Y_k\}, k = 1, 2, 3, \cdots, m \tag{3}$$

$$X_k = (x_{k1}, x_{k2}, x_{k3}, \cdots x_{ki}) \quad i \in \{1, 2, \ldots, o\} \tag{4}$$

where m denotes the sum of historical cases; X_k denotes the case characteristic attribute; x_k is the characteristic data from case descriptions; Y_k is the case category.

Step 2: Calculate the similarity between the new case and the cases in the case base using Equation (5).

$$SIM(X_a, X_k) = \sum_{i=1}^{o} \sqrt{\theta_i (x_{ai} - x_{ki})^2} \tag{5}$$

where θ_i represents the weights of case characteristic attributes and it is typically set as $\frac{1}{o}$. The value of θ_i indicates the contribution degree of a specific characteristic attribute to the overall case.

Step 3: Sort the similarity values in descending order, and select the top σ cases T_1, T_2, T_3, …, T_σ as similar cases. According to the reuse principle, Equation (6) was used to obtain the result.

$$Y_a = \begin{cases} Y(H_1) & \sigma = 1 \\ \mathrm{m}ajority(Y(H_i)) & \sigma > 1 \end{cases} \quad (6)$$

Step 4: Store the corresponding target cases and results in the historical case base to complete the knowledge storage and experience learning of CBR.

3.1.2. FCM

The fuzzy c-means clustering (FCM) algorithm is a clustering algorithm based on objective functions [48]. This algorithm introduces membership functions on the basis of the K-Means algorithm, which can better indicate the similarity between a sample and a certain cluster. The FCM objective function is shown in Equation (7).

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d^2(x_j, v_i) \quad (7)$$

where c is the total number of all categories; m is the weighting fuzziness parameter that is set as 2 in this manuscript; n represents the number of cases in the case base; u_{ik} denotes the membership degree of the kth case with respect to the ith category; $d^2(x_j, v_i)$ is the distance between the kth case and the ith cluster center. Using the Lagrange multiplier method, the iterative formulas for u_{ik} and v_i were obtained as follows.

$$V_i = \frac{\sum_{j=1}^{n} (u_{ij})^{\mathrm{m}} x_j}{\sum_{j=1}^{n} u_{ij}^m} \quad j = 1, 2, \cdots, N \quad (8)$$

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_l)} \right)^{\frac{2}{m-1}}} \quad (9)$$

The authors set the maximum iterations. When the relevant parameters were input, u_{ik} and v_i were continuously updated. When the maximum number of iterations was reached or the set conditions were met, the iterations were stopped. Then, the membership matrix, cluster centers and individual clusters were output.

3.1.3. SO

The Snake Optimization (SO) algorithm is a new metaheuristic algorithm proposed by Fatma A. Hashim et al. in 2022 [49]. This algorithm is inspired by the mating behavior of snakes in nature. Compared with other algorithms, SO has higher precision and faster iteration speed [50]. The algorithm includes the following four stages: initialization stage, selection stage, exploration stage and development stage (see Figure 4). The specific process is as follows.

Step 1: Initialization stage. Generate an initial population using Equation (10), and then divide the population into female and male two swarms using Equations (11) and (12).

$$x_i = x_{\min} + r \times (x_{\max} - x_{\min}) \quad (10)$$

$$N_m \approx \frac{N}{2} \quad (11)$$

$$N_f = N - N_m \quad (12)$$

where x_i represents the location of the ith individual; r is the random number between 0 and 1; $x_{\max}$ and $x_{\min}$ are the upper and lower limits for x; N represents the number of individuals; N_m denotes the number of individuals in the male population, while N_f denotes the number of individuals in the female population.

Step 2: Selection stage. By calculating the temperature T_{emp} and food quantity Q, the search stage is selected. T_{emp} and Q are calculated using Equations (13) and (14), respectively.

$$T_{emp} = \exp(\frac{-t}{T}) \tag{13}$$

$$Q = C_1 * \exp(\frac{t-T}{T}) \tag{14}$$

where t denotes the number of present iterations; T is the maximum number of iterations; C_1 is a constant of 0.5.

Step 3: Exploration stage. When the food quantity Q is below the threshold, male snakes update their positions using Equations (15) and (16), while female snakes update their positions using Equations (17) and (18).

$$X_{i,m} = X_{rand,m}(t) \pm c_2 \times A_m \times [(X_{\max} - X_{\min}) \times rand + X_{\min}] \tag{15}$$

$$A_m = \exp(\frac{-f_{rand,m}}{f_{i,m}}) \tag{16}$$

$$X_{i,f} = X_{rand,f}(t) \pm c_2 \times A_f \times [(X_{\max} - X_{\min}) \times rand + X_{\min}] \tag{17}$$

$$A_f = \exp(\frac{-f_{rand,f}}{f_{i,f}}) \tag{18}$$

where $X_{i,m}$ and $X_{i,f}$ are the locations of the ith male and female snakes, respectively; $X_{rand,m}$ and $X_{rand,f}$ are the randomly selected locations of male and female snakes. c_2 is a constant that is set as 0.05; A_m and A_f represent the ability of male and female snakes to search for food; $rand$ is a random number between 0 and 1. $f_{rand,m}$ and $f_{rand,f}$ are the fitness of $X_{rand,m}$ and $X_{rand,f}$; $f_{i,m}$ and $f_{i,f}$ represent the fitness of the ith individuals of male and female snakes.

Step 4: Development stage. When the food quantity Q is greater than the threshold, the development stage is divided into two parts according to the temperature. When the temperature is greater than the threshold, the snake is in a thermal state and it only searches for food. The position update equation is shown as follows:

$$X_{i,j}(t+1) = X_{food} \pm c_3 \times T_{emp} \times rand \times [X_{food} - X_{i,j}(t)] \tag{19}$$

where $X_{i,j}(t+1)$ represents the location of the female or male snake; X_{food} is the best individual location and c_3 is a constant that is set as 0.05.

As the number of iterations increases, the temperature decreases gradually. When the temperature is lower than the threshold, the snake is in a cold state. In this state, the snake updates its position through fighting or mating.

The equation for updating the position in fight mode is shown in Equation (20).

$$\begin{cases} X_{i,m}(t+1) = X_{i,m}(t) + c_3 \times FM \times rand \times [Q \times X_{best,f} - X_{i,m}(t)] \\ X_{i,f}(t+1) = X_{i,f}(t) + c_3 \times FF \times rand \times [Q \times X_{best,m} - X_{i,F}(t)] \end{cases} \tag{20}$$

where $X_{i,m}(t+1)$ and $X_{i,f}(t+1)$ denote the locations of the ith male and female snakes. $X_{best,m}$ and $X_{best,f}$ denote the best location of male and female snakes. FM and FF represent the fighting capability of the male and female snakes, which are calculated via Equation (21).

$$\begin{cases} FM = \exp(\frac{-f_{best,f}}{f_i}) \\ FF = \exp(\frac{-f_{best,m}}{f_i}) \end{cases} \tag{21}$$

where $f_{best,m}$ and $f_{best,f}$ represent the fitness of the best male and female snakes; f_i represents the fitness of individual i.

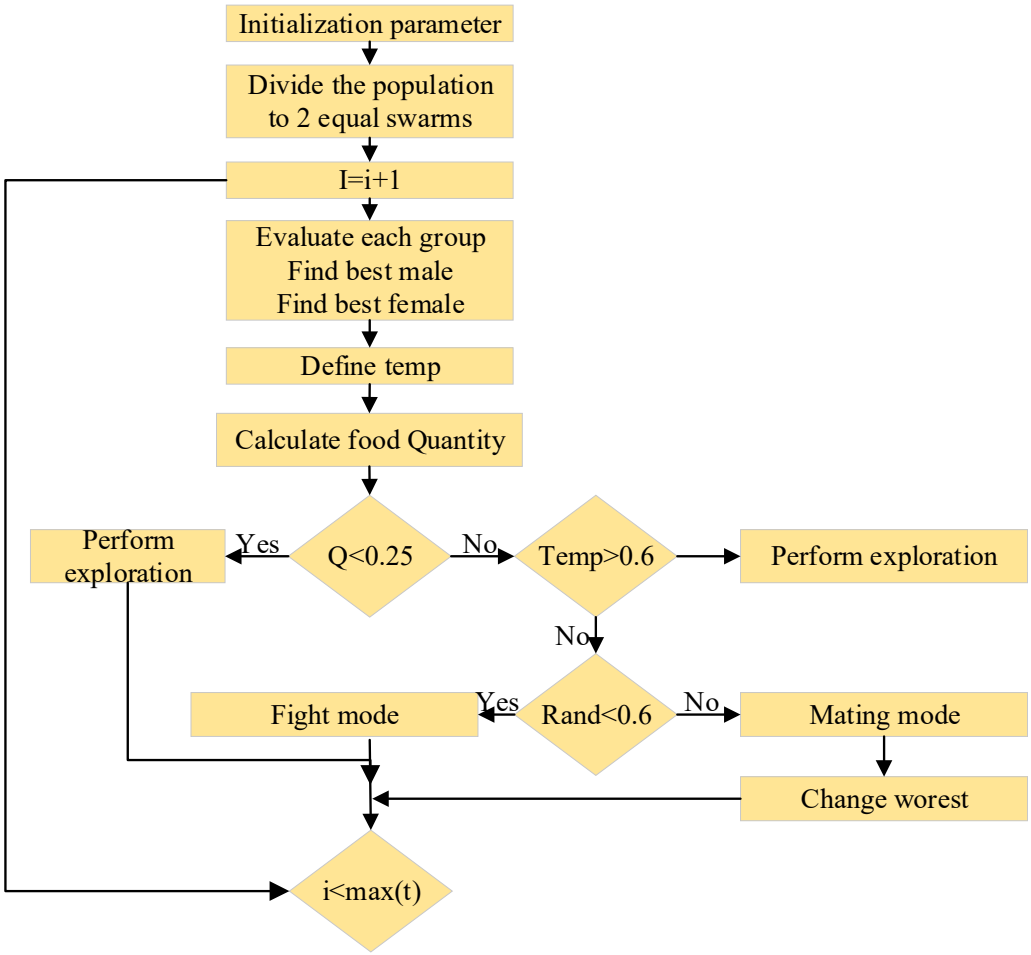

Figure 4. Flowchart of SO algorithm.

The updated position equation for the mating mode is depicted as follows:

$$\begin{cases} X_{i,m}(t+1) = X_{i,m}(t) + c_3 \times M_m \times rand \times [Q \times X_{i,f}(t) - X_{i,m}(t)] \\ X_{i,f}(t+1) = X_{i,f}(t) + c_3 \times M_f \times rand \times [Q \times X_{i,m}(t) - X_{i,f}(t)] \end{cases} \quad (22)$$

where $X_{i,m}(t)$ and $X_{i,f}(t)$ are the locations of the ith snakes in the male and female swarms, respectively; M_m and M_f are the mating competence values of male and female snakes, which are calculated as follows:

$$\begin{cases} M_m = \exp(\frac{-f_{i,f}}{f_{i,m}}) \\ M_f = \exp(\frac{-f_{i,m}}{f_{i,f}}) \end{cases} \quad (23)$$

If the snake eggs hatch, the worst individuals in the male swarm and female swarm are exchanged using the following equation:

$$\begin{cases} X_{worst,m} = X_{\min} + rand \times (X_{\max} - X_{\min}) \\ X_{worst,f} = X_{\min} + rand \times (X_{\max} - X_{\min}) \end{cases} \quad (24)$$

where $X_{worst,m}$ and $X_{worst,f}$ represent the worst individuals in male and female swarms.

3.1.4. Weight Value Calculation Based on PCA

As a commonly used dimensionality reduction algorithm, principal component analysis (PCA) can also calculate the characteristic attribute weights of the data and has achieved

good results in multiple fields [51]. Therefore, PCA was used to calculate the characteristic attribute weights of the data in this manuscript. The specific steps are as follows:

Step 1: Standardize each case characteristic attribute using Equation (25).

$$x_{ki}^{*} = \frac{x_{ki} - \overline{x_i}}{\sqrt{\mathrm{var}(x_i)}} \tag{25}$$

where x_{ki}^{*} represents the characteristic description data of the standardized case and x_{ki} represents that of the original case. $\overline{x_l}$ and $\mathrm{var}(x_i)$ are calculated using Equations (26) and (27).

$$\overline{x_i} = \sum_{k=1}^{m} \frac{x_{ki}}{m} \tag{26}$$

$$\mathrm{var}(x_i) = \frac{1}{m-1} \sum_{k=1}^{m} \left(x_{ki} - \overline{x_i}\right)^2 \tag{27}$$

Step 2: Calculate the Pearson correlation coefficient matrix R between case characteristic attributes based on Equations (28) and (29).

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & r_{pp} \end{bmatrix} \tag{28}$$

$$r_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} x_{ki}^{*} x_{kj}^{*} \tag{29}$$

where r_{ij} is the correlation coefficient between the ith and the jth indicators.

Step 3: Calculate the eigenvalues λ_i and eigenvectors μ_{ij} of the correlation coefficient matrix R using the Jacobi method.

Step 4: Calculate the number of principal components n using Equation (30).

$$n = \min\left\{ l \middle| \sum_{i=1}^{l} \lambda_i / \sum_{j=1}^{o} \lambda_j \geq \delta, o \geq l \geq 1 \right\} \tag{30}$$

where δ is the contribution threshold, which is set as 92%; $\sum_{i=1}^{l} \lambda_i / \sum_{j=1}^{o} \lambda_j$ is the cumulative contribution rate of the top l principal components.

Step 5: Calculate the load matrix of principal component factor using Equations (31) and (32).

$$a = \begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{bmatrix} \tag{31}$$

$$a_{ij} = \mu_{ij} \sqrt{\lambda_j} \quad i = 1, 2, \cdots, o \quad j = 1, 2, \cdots, n \tag{32}$$

where a_{ij} is the correlation coefficient between the ith indicator and the jth principal components.

a_{ij} represents the importance of the ith characteristic attribute of the case for the jth principal component. The smaller the $|a_{ij}|$, the less influence of the ith characteristic attribute of the case on the jth principal component; on the contrary, a larger $|a_{ij}|$ represents a greater influence.

Step 6: Calculate the weight of each indicator on each principal component according to Equation (33), and calculate the weights of case characteristic attributes through Equation (34).

$$\theta_{ij}^{1} = \frac{|a_{ij}|}{\sum_{l=1}^{o}|a_{lj}|} \quad i = 1, 2, \cdots, o \quad j = 1, 2, \cdots, n \tag{33}$$

$$\theta_{i} = \frac{\sum_{j=1}^{n}\theta_{ij}^{1}}{\sum_{j=1}^{o}\theta_{ij}^{1}} \quad i = 1, 2, \cdots, o \tag{34}$$

where θ_{ij}^{1} denotes the weight of indicator i on principal component j, and θ_i is the weight of indicator i.

3.2. PCA-FC-CBR

The traditional PCA–clustering–CBR model is a hybrid model that applies PCA and clustering to CBR. To some extent, this model solves the problem that the weights of case characteristic attributes are difficult to determine in the traditional CBR model and the case retrieval efficiency is reduced as the case increases in the case base.

The PCA–clustering–CBR model first adopts the PCA algorithm to calculate the basic weights of case characteristic attributes and applies them to the calculation of similarity, which increases the calculation accuracy and improves the prediction ability. In order to improve the efficiency of case retrieval, a clustering algorithm is used to divide n cases in the case base into k clusters with cluster centers serving as representative cases. The mean value of each data point in each cluster is taken. When a new problem arises, it is first compared with cluster centers, and then assigned to the most relevant cluster, where the entire CBR process is conducted. However, this model may cause the loss of cluster boundary information, reducing the prediction ability of the machine learning model and affecting its generalization ability.

To find a solution to the aforementioned problem, we introduced fuzzy clustering into CBR to construct the PCA-FC-CBR model. The case base was divided into sets of fuzzy clusters with overlapping boundaries, allowing any case to belong to multiple clusters simultaneously. During the prediction process, cases are screened based on their membership degree with respect to the cluster centers, thereby reducing the loss of boundary information.

3.3. Modeling Building and Hyperparameter Tuning

3.3.1. Modeling Building

The flowchart of model construction is shown in Figure 5 and the steps of model construction are shown as follows:

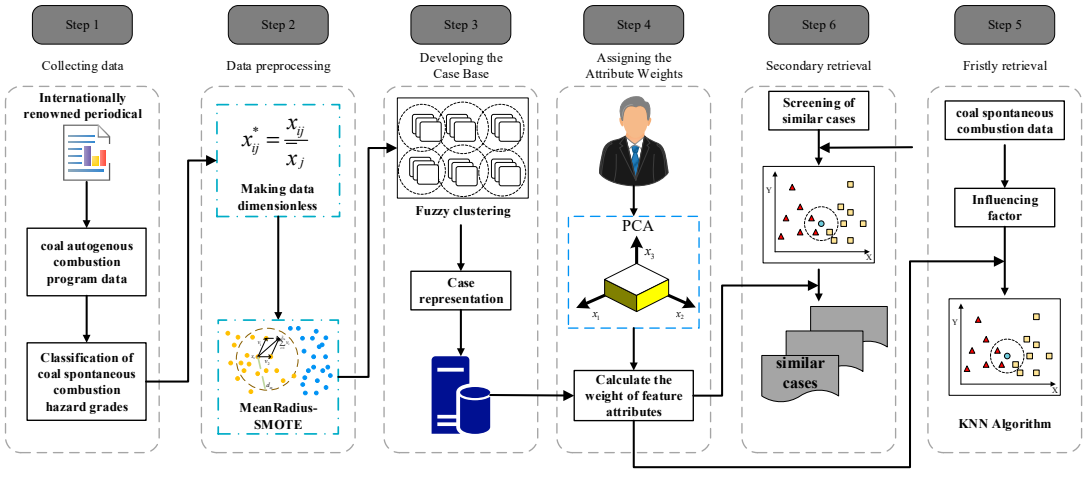

Figure 5. Flowchart of PCA-FC-CBR algorithm.

Step 1: Collect coal spontaneous combustion data, screen prediction indicators for coal spontaneous combustion, and analyze the data to build a raw coal spontaneous combustion data base.

Step 2: Data preprocessing. Initially, stratify the hazard grades associated with spontaneous coal combustion. Subsequently, apply dimensionless normalization to the dataset. Finally, equalize the dataset distribution through the implementation of the MeanRadius-SMOTE algorithm.

Step 3: Divide the coal spontaneous combustion data into a training set and a test set, using the training set data to construct a coal spontaneous combustion database. The test set data are used to verify the model's performance. Apply the FCM algorithm to perform fuzzy clustering on the coal spontaneous combustion database, obtaining cluster centers and membership functions. Construct a fuzzy clustering CBR model.

Step 4: Set the cumulative contribution rate and calculate the weights of each case characteristic attribute using PCA.

Step 5: Primary case retrieval. Firstly, screen out the cluster centers whose similarity with the new case is no less than γ_1 (similarity threshold). Secondly, screen out all cases whose membership degree with cluster centers is greater than γ_2 (similarity threshold), form a new case library with the filtered cases.

Step 6: Secondary case retrieval. Sort the similarity degrees from largest to smallest and take the top σ cases as similar cases. Determine the coal spontaneous combustion hazard grades on the basis of the majority rule principle. If there is a case with a similarity of 1 to X in D, take the case as the matching one for X.

3.3.2. Hyperparameter Tuning

The effectiveness of machine learning models largely depends on parameter selection. If parameters are selected based on empirical selection or grid search, deep learning models may experience overfitting or underfitting. Furthermore, since there are many hyperparameters, it is often difficult to set their values through experience alone. Therefore, the SO algorithm was adopted to identify the hyperparameters in the PCA–clustering–CBR hybrid model in this research. Table 3 displays the hyperparameters in PCA–clustering–CBR that require adjustment and their respective ranges of values.

Table 3. Hyperparameters in PCA–clustering–CBR model.

Parameters	γ_1	γ_2	σ
Ranges	[0, 1]	[0, 1]	[0, 10]

The number of iterations was set as 100 with 10 individuals in each generation. All parameters in the SO algorithm were set through experimental testing. During hyperparameter tuning, each set of hyperparameters is represented by a snake in the SO algorithm. The SO algorithm was adopted to update the position of each snake by using different optimization formulas at different stages. Hence, the fitness value was minimized. When meeting the termination condition, the optimal hyperparameters were selected.

To ensure both the computational efficiency and prediction precision of the model, the authors set the accuracy rate and the number of comparisons as the objective function. For multiple objective functions, the optimization process was set as follows: (1) initialize the parameters of the SO algorithm. (2) Use the accuracy rate as the objective function to find the optimal parameter for PCA–clustering–CBR. (3) Find the optimal parameter for PCA–clustering–CBR by setting the highest accuracy as the limiting condition and the number of comparisons as the objective function.

4. Experiments

4.1. Dataset Division and Evaluation Indicators

The data was divided into two sets, with the first 70% used as the training set to create and train the SO–PCA–clustering–CBR model, and the remaining 30% used as the test set to evaluate the model performance.

Accuracy and recall are commonly used indicators to evaluate the predictive ability of classification models. They are calculated through a confusion matrix, as demonstrated in Figure 6. The confusion matrix is widely used to evaluate the prediction precision of classification models in binary classification. For the confusion matrix of multi-class classification problems, each category is successively considered as positive, while other categories are considered as negative, thus converting the multi-class classification problem into multiple binary classification problems [52]. The specific schematic diagram is shown in Figure 6.

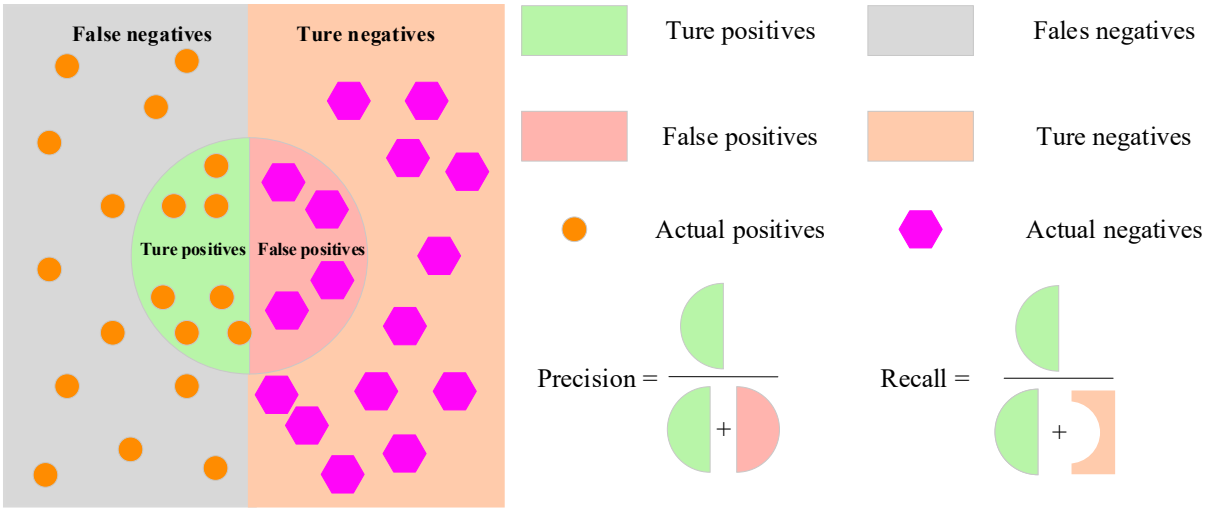

Figure 6. Schematic diagram of confusion matrix.

Accuracy and *F*1 are also cited in this study to evaluate the prediction performance of machine learning models, and the calculation formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{35}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{36}$$

In addition, to quantify the improvement ratio of different models, the study introduced $IR(i)$, which is calculated according to Equation (37).

$$IR(i) = \begin{cases} (A(i) - B(i))/B(i) & \text{The bigger the i indicator, the better} \\ (B(i) - A(i))/A(i) & \text{The smaller the i indicator, the better} \end{cases} \tag{37}$$

where i represents the different evaluation indicators. $IR(i)$ represents the improvement ratio. $A(i)$ and $B(i)$ represent the value of i for model A and model B, respectively.

4.2. Experiments and Comparison

To verify the superior performance of the proposed model in this manuscript, multiple sets of comparative experiments were designed as depicted in Figure 7. The specific steps are as follows:

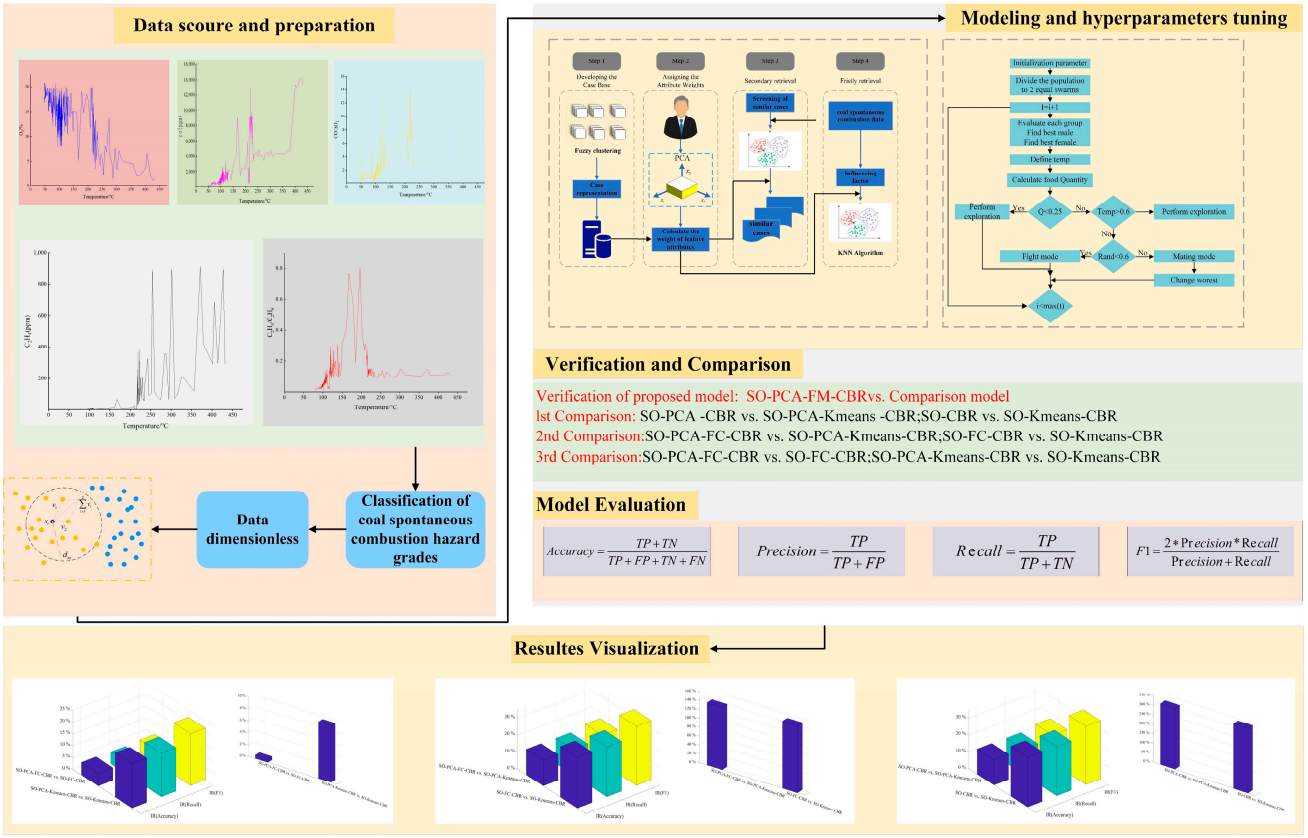

Figure 7. Framework of proposed model.

Step 1: Data Preprocessing. Firstly, the data were collected and analyzed. Secondly, based on the analysis of the data, the hazard classification criteria for the spontaneous coal combustion data were established. Thirdly, the data were dimensionless and balanced by the MeanRadius-SMOTE method. Finally, the dataset was divided into a training set and test set in the ratio of 7:3.

Step 2: Model construction. Firstly, the SO-PCA-FC-CBR model and other comparative models were built, and the training samples were inputted to train the prediction models. Accuracy was used as the objective function, while the SO algorithm was used to determine the hyperparameters of the models. Secondly, five single models (SVM, RF, Bayesian, GBDT and CBR) and five hybrid models (PCA-CBR, SO-PCA-Kmeans-CBR, SO-Kmeans-CBR, SO-PCA-FC-CBR and SO-FC-CBR) were constructed.

Step 3: Verification, comparison and visualization of results. To verify the performance of the proposed SO-PCA-FC-CBR model in this study, the above-mentioned six single models and three hybrid models were compared with each other when predicting the hazard grades of coal spontaneous combustion. All models were based on the same dataset. The accuracy, precision, recall, and F1 index were selected as evaluation indicators to verify the prediction effect of the models on the test set.

5. Results and Discussions

5.1. Verification of Data Preprocessing Effect

To verify the effect of data preprocessing method in this study, firstly, four classic machine learning models were constructed, including one ensemble algorithm, namely RF, and three individual classic algorithms, namely SVM, Bayesian and GBDT. Secondly, the raw database was preprocessed using SMOTE [53,54], Kmeans–SMOTE [55], and Mean-radiusSMOTE. Finally, the prediction effects of these four machine learning models on both the raw and preprocessed coal spontaneous combustion databases were compared using accuracy as the evaluation indicator. The test results are shown in Table 4. After

data preprocessing, the accuracy of all algorithms improved to varying degrees and the method proposed in this manuscript was superior to other over-sampling methods. For the dataset processed via the method proposed in this manuscript, GBDT possesses the highest accuracy. Compared with the raw dataset and datasets processed via SMOTE and Kmeans–SMOTE, the accuracy of the method proposed in this manuscript has increased by 49.1%, 23.9% and 10%, respectively. The accuracy of Bayesian is the lowest. However, compared with the raw dataset and datasets processed via SMOTE and Kmeans–SMOTE, the accuracy of the method proposed in this manuscript has increased by 41.5%, 27.1% and 17.2%.

Table 4. Prediction accuracy of machine learning models under different data preprocessing.

Model	Raw Data Base	Data Base Processed via SMOTE	Data Base Processed via Kmeans–SMOTE	Data Base Processed via MeanradiusSMOTE
GBDT	0.59	0.71	0.8	0.88
SVM	0.52	0.64	0.79	0.826
RF	0.74	0.79	0.83	0.86
Bayesian	0.53	0.59	0.64	0.75

5.2. Parameter Tuning

Figure 8 shows the iterative process of the SO algorithm searching for the maximum accuracy. It can be seen that as the SO algorithm iterates, the accuracy gradually increases, indicating that the SO algorithm is effective at optimizing the hyperparameters of the SO-PCA-FC-CBR hybrid model. The accuracy is the lowest (0.71) at the first iteration, and it increases to 0.95 at the 36th iteration.

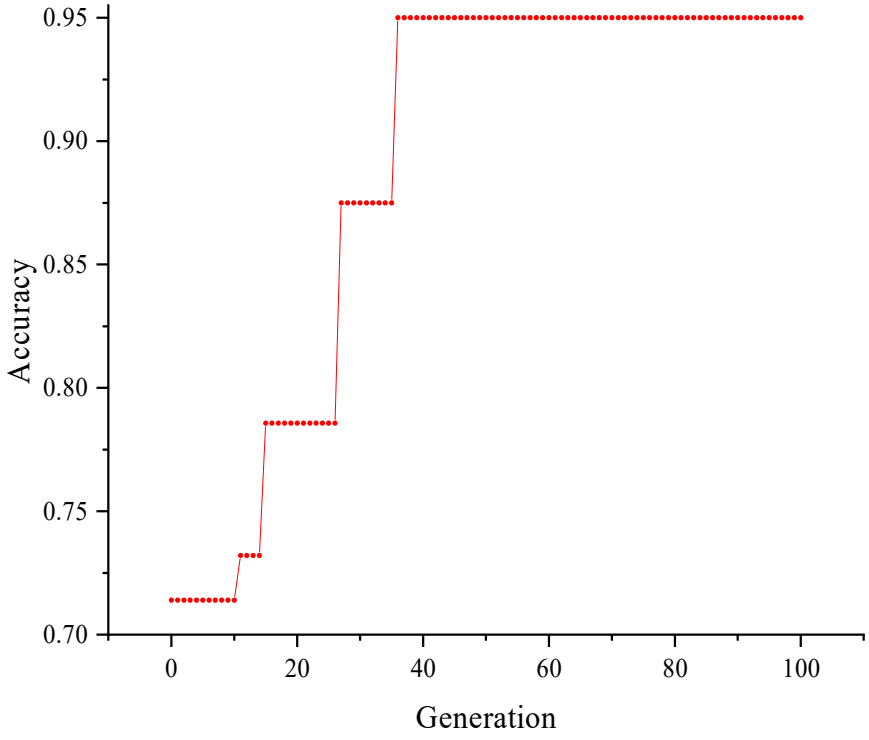

Figure 8. Iterative graph of accuracy.

Figure 9 demonstrates the iterative process of the SO algorithm searching for the minimum number of comparisons. The figure illustrates that as the SO algorithm iterates, the number of comparisons gradually decreases, indicating that the SO algorithm is also

effective at improving the computational efficiency of the SO-PCA-FC-CBR hybrid model. The number of comparisons is the highest at the first iteration, which is 78,752, and then decreases to 45,000 at the 45th iteration.

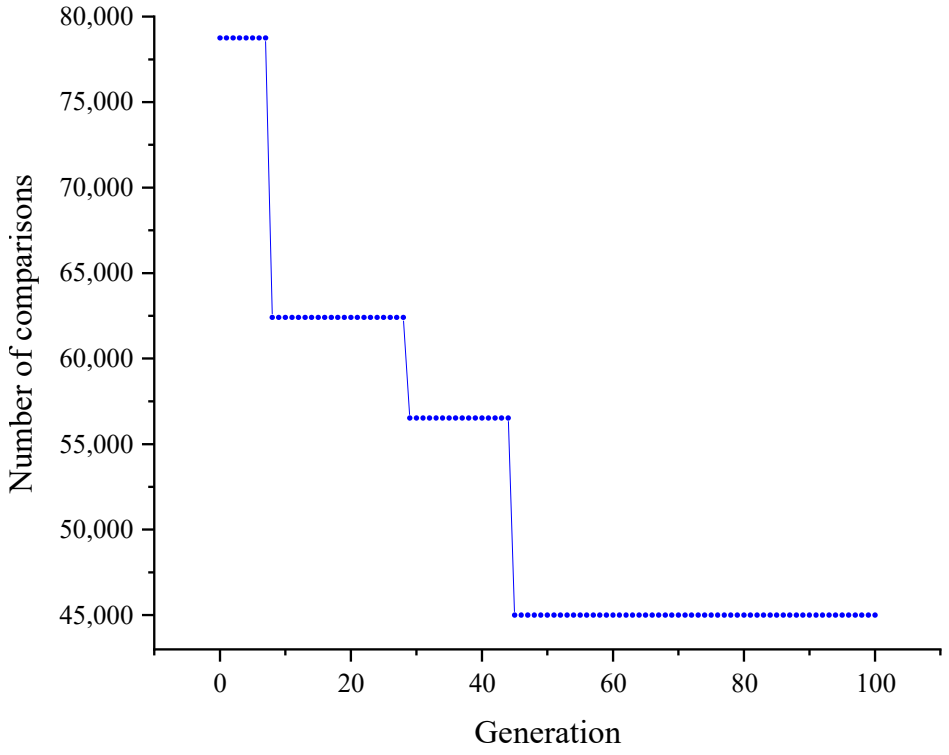

Figure 9. Iterative graph of number of comparisons.

5.3. Model Comparison Analysis

5.3.1. Comparison between SO–PCA–Clustering–CBR and Other Models

In order to verify the performance of the model proposed in this manuscript, five single models (SVM, RF, Bayesian, GBDT and CBR) and two hybrid models (SO-PCA-Kmeans-CBR and SO-Kmeans-CBR) were compared with each other by using them to predict coal spontaneous combustion hazard grades. Table 5 shows the F1, recall, and accuracy of SO-PCA-FC-CBR and other models at each intensity grade (0–5 levels) of coal spontaneous combustion. In the prediction of initial values for warning, SO-PCA-FC-CBR has the highest F1-score, recall, and accuracy of 1, 1, and 1, respectively. In the prediction of gray warning, SO-PCA-FC-CBR has the highest F1-score, recall, and accuracy of 0.98, 1, and 0.97, respectively. In the prediction of blue warning, CBR and RF have the highest accuracy of 1, while SO-PCA-FC-CBR has the highest F1-score and recall of 0.97 and 0.97, respectively. In the prediction of yellow warning, RF has the highest recall of 1, while SO-PCA-FC-CBR has the highest F1-score and accuracy of 1 and 0.98, respectively. In the prediction of orange warning, SVM has the highest F1 of 1, SO-PCA-FC-CBR has the highest recall of 0.81 and RF has the highest accuracy of 1. In the prediction of red warning, SVM has the highest F1-score, recall, and accuracy of 1, 1, and 1, respectively. Overall, SO-PCA-FC-CBR shows the best predictive performance for all of the six coal spontaneous combustion hazard grades.

Figure 10 shows the overall accuracy of five single models, two hybrid models and the model proposed in this manuscript. Evidently, the SO-PCA-FC-CBR model possesses the highest overall accuracy of 95%, demonstrating the superiority of the proposed algorithm in predicting coal spontaneous combustion hazard grades.

Table 5. The performance comparison of SO–PCA–Clustering–CBR and other ML models.

Model	Metrics	0	1	2	3	4	5
SO-PCA-FC-CBR	F1	1.00	0.98	0.97	0.98	0.88	0.89
	Recall	1.00	1.00	0.97	0.97	0.81	0.97
	Accuracy	1.00	0.97	0.97	1.00	0.96	0.83
SO-PCA-Kmeans-CBR	F1	0.97	0.87	0.77	0.77	0.61	0.91
	Recall	1.00	0.94	0.67	0.90	0.47	1.00
	Accuracy	0.94	0.81	0.91	0.68	0.88	0.83
SO-Kmeans-CBR	F1	0.77	0.55	0.72	0.73	0.38	0.86
	Recall	1.00	0.48	0.60	0.93	0.25	0.93
	Accuracy	0.63	0.65	0.90	0.60	0.80	0.80
CBR	F1	0.94	0.94	0.95	0.97	0.79	0.84
	Recall	0.94	0.97	0.90	0.97	0.69	0.97
	Accuracy	0.94	0.91	1.00	0.97	0.92	0.74
SVM	F1	0.79	0.61	0.84	0.84	0.90	1.00
	Recall	1.00	0.48	0.77	0.93	0.81	1.00
	Accuracy	0.65	0.83	0.92	0.76	1.00	1.00
RF	F1	0.85	0.79	0.95	0.97	0.80	0.86
	Recall	1.00	0.68	0.90	1.00	0.66	1.00
	Accuracy	0.74	0.95	1.00	0.94	1.00	0.75
Bayesian	F1	0.82	0.71	0.77	0.76	0.47	0.87
	Recall	0.87	0.74	0.67	0.93	0.31	1.00
	Accuracy	0.77	0.68	0.91	0.64	1.00	0.77
GBDT	F1	0.97	0.94	0.88	0.75	0.75	0.91
	Recall	1.00	0.94	0.90	0.83	0.62	1.00
	Accuracy	0.94	0.94	0.87	0.69	0.95	0.84

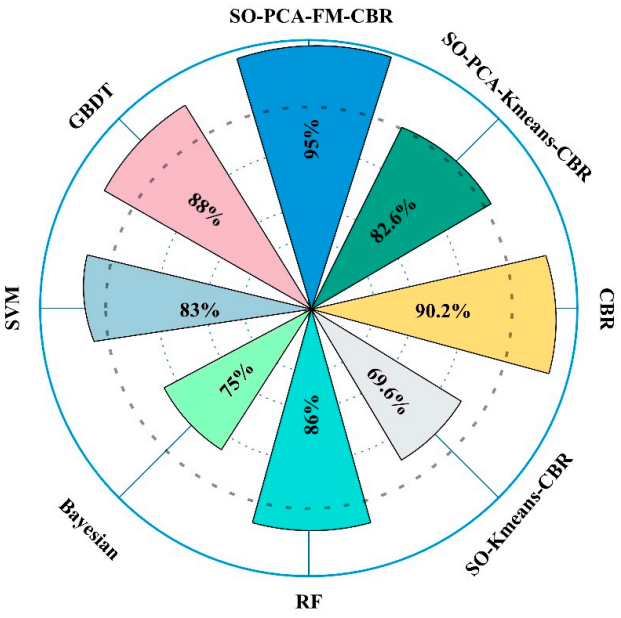

Figure 10. The accuracy of SO-PCA-FC-CBR and other ML models.

5.3.2. Comparison of Subgroups

To further analyze the prediction results, three sets of comparisons were conducted in this manuscript. Firstly, to verify that Kmeans can cut the computational cost of the model at the expense of lower prediction accuracy, the first set of comparisons was conducted. Secondly, to demonstrate the superiority of FC over Kmeans, a second set of comparisons was made. Finally, the third set of comparisons was conducted to verify the effectiveness of the weight calculated via PCA. The settings of these three sets of comparisons are shown in Table 6.

Table 6. Model comparison.

Comparison Groups	Description
1st set of comparison	SO-PCA -CBR vs. SO-PCA-Kmeans-CBR SO-CBR vs. SO-Kmeans-CBR
2nd set of comparison	SO-PCA-FC-CBR vs. SO-PCA-Kmeans-CBR SO-FC-CBR vs. SO-Kmeans-CBR
3rd set of comparison	SO-PCA-FC-CBR vs. SO-FC-CBR SO-PCA-Kmeans-CBR vs. SO-Kmeans-CBR

To explore the impact of clustering on the performance of the CBR model, the following comparison groups were introduced in this manuscript: PCA-CBR vs. SO-PCA-Kmeans-CBR and CBR vs. SO-Kmeans-CBR. The comparison results are shown in Table 7 and Figure 11. Compared with SO-PCA-Kmeans-CBR, the accuracy, recall, and F1 of PCA-CBR increased by 15.01%, 14.73% and 16.42%, respectively, but the number of comparisons increased by 320.01%. Compared with SO-Kmeans-CBR, the accuracy, recall, and F1 of CBR increased by 29.59%, 29.14% and 34.63%, respectively, but the number of comparisons increased by 340%. To sum up, although the application of clustering to the CBR model can reduce the running time of the model and improve its computational efficiency, it also causes information loss, resulting in the lower prediction accuracy of the CBR model.

Table 7. Results of the 1st set of comparisons.

	SO-PCA-CBR vs. SO-PCA-Kmeans-CBR	SO-CBR vs. SO-Kmeans-CBR
IR(Accuracy)	15.01%	29.59%
IR (Recall)	14.73%	29.14%
IR (F1)	16.42%	34.63%
IR (Number of comparisons)	320.01%	340%

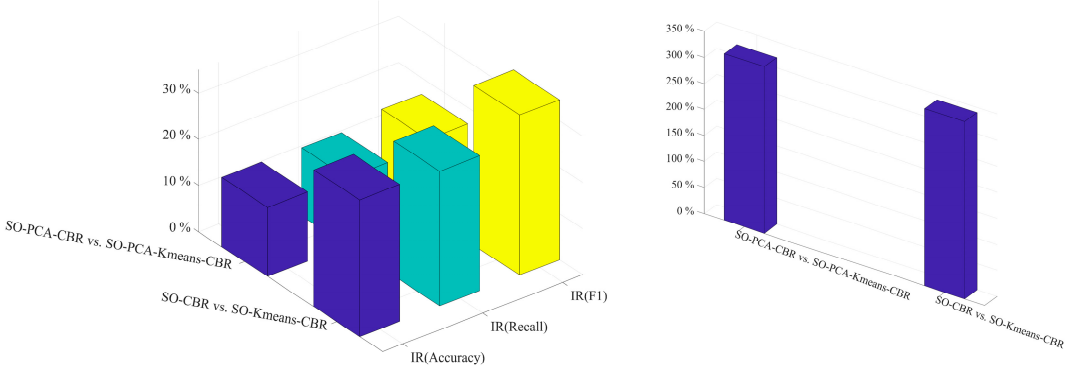

Figure 11. Results of the 1st set of comparisons.

To verify the superiority of FC over Kmeans, the following control groups were introduced in this study: SO-PCA-FC-CBR vs. SO-PCA-Kmeans-CBR and SO-FC-CBR vs. SO-Kmeans-CBR. The comparison results are displayed in Table 8 and Figure 12. Compared with SO-PCA-Kmeans-CBR, the accuracy, recall and F1 of SO-PCA-FC-CBR increased by 15.01%, 14.73% and 16.42%, respectively, but the number of comparisons increased by 140.32%. Compared with SO-Kmeans-CBR, the accuracy, recall and F1 of SO-FC-CBR increased by 29.59%, 29.41% and 34.63%, respectively, but the number of comparisons increased by 150.83%. In summary, compared with ordinary clustering, although fuzzy clustering increases the computational cost, it avoids the loss of boundary information, preventing the prediction accuracy of the CBR model from declining.

Table 8. Results of the 2nd set of comparisons.

	SO-PCA-FC-CBR vs. SO-PCA-Kmeans-CBR	SO-FC-CBR vs. SO-Kmeans-CBR
IR (Accuracy)	15.01%	29.59%
IR (Recall)	14.73%	29.14%
IR (F1)	16.42%	34.63%
IR (Number of comparisons)	140.32%	150.83%

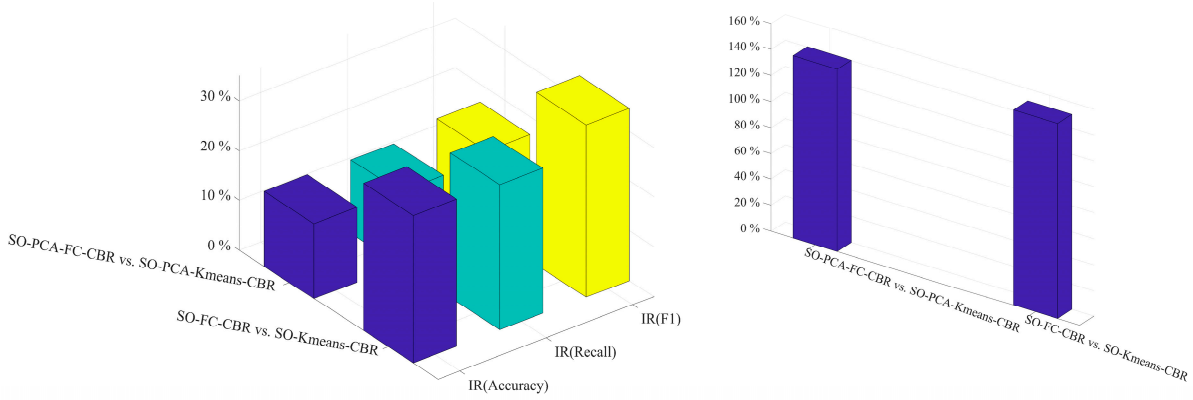

Figure 12. Results of the 2nd set of comparisons.

Finally, to verify the effectiveness of the weight calculated via PCA, this study introduced the following two control groups: SO-PCA-FC-CBR vs. SO-FC-CBR and SO-PCA-Kmeans-CBR vs. SO-Kmeans-CBR. The comparison results are shown in Table 9 and Figure 13. Compared with SO-FC-CBR, the accuracy, recall and F1 of SO-PCA-FC-CBR increased by 5.32%, 5.01% and 5.32% respectively, and the number of comparisons increased by 0.57%. Compared with SO-Kmeans-CBR, the accuracy, recall and F1 of SO-PCA-Kmeans-CBR increased by 18.68%%, 18.29% and 21.79%, respectively, but the number of comparisons increased by 9.48%. To sum up, the weight calculated via PCA can effectively enhance the prediction performance of CBR and has almost no impact on the computational efficiency of the model.

Table 9. Results of the 3rd set of comparisons.

	SO-PCA-FC-CBR vs. SO-FC-CBR	SO-PCA-Kmeans-CBR vs. SO-Kmeans-CBR
IR (Accuracy)	5.32%	18.68%
IR (Recall)	5.01%	18.29%
IR (F1)	5.32%	21.79%
IR (Number of comparisons)	0.57%	9.48%

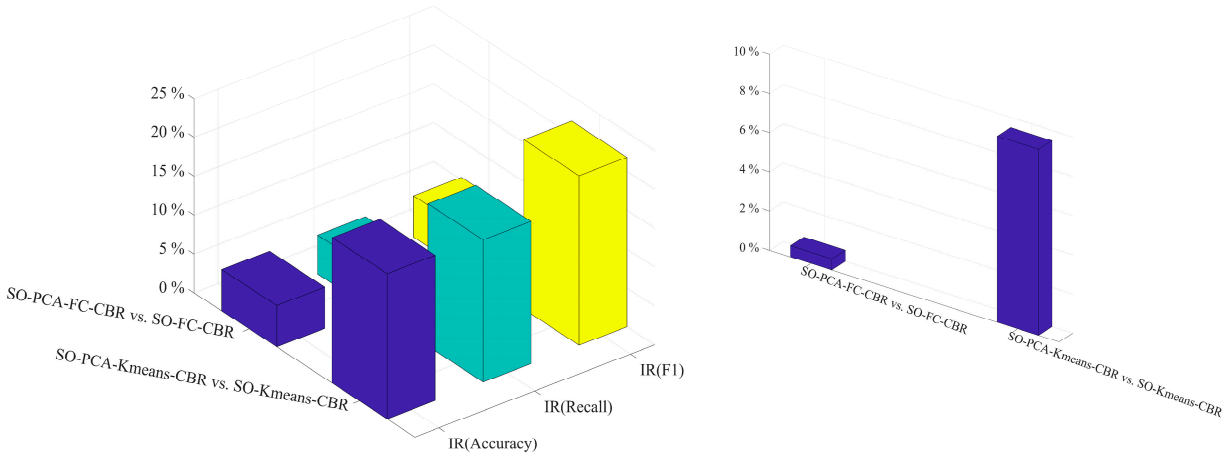

Figure 13. Results of the 3rd set of comparisons.

5.4. Variable Importance

To calculate the relative importance of coal spontaneous combustion characteristic variables, this manuscript took SO-PCA-FC-CBR as the target function and adopted the random balance design Fourier amplitude sensitivity test (RBD-FAST) method to carry out sensitivity analysis on characteristic variables. RBD-FAST is a method used to reduce computational costs by implementing the latest developed Fourier amplitude sensitivity test (FAST) using random balance design (RBD) technology [56]. All parameters were set to the same frequency and then reorganized after sampling. Fast Fourier Transform (FFT) was used for the model output based on the previous reorganization sequence. The first-order sensitivity analysis results of corresponding parameters were recorded [57].

In this method, the changes in the results can be simplified as follows:

$$S_i = \frac{V_{x_i}}{V(Y)} \tag{38}$$

where V_{x_i} is the first-order influence of the input factor x_i based on the method. $V(Y)$ represents the total variance of SO–PCA–Clustering–CBR.

The relative importance of input variables is shown in Figure 14. It can be seen that CO is the most important input variable with a relative importance score of 0.28, followed by $CO/\Delta O_2$ (0.23), C_2H_4/C_2H_6 (0.19) and C_2H_4 (0.17). O_2 (0.13) is the least sensitive predictive factor.

The figure illustrates that CO is the most important factor affecting the coal spontaneous combustion hazard grades and this is consistent with the research results of many scholars, but at present, most of the methods used to reach this conclusion have been obtained by observing the chemical reaction of coal molecules and the change law of gas in the temperature program [58–60]. However, in this paper, the relative importance of CO is quantified by the RBD-FAST method, and this conclusion is justified by specific values. This method can also be utilized to determine the gases with the greatest relative importance in each stage of coal autogenous combustion, and thus to select the signature gases for each stage of coal autogenous combustion. With the continuous deepening of research in this field, scholars have found more and more factors that influence the spontaneous combustion hazard class of coal [61,62]. However, most of the methods have also been obtained by analyzing the experimental results, which cannot illustrate the importance of the gas in a quantitative manner, while the RBD-FAST method can make up for this shortcoming, providing scholars with a quantitative method to illustrate the newly found importance of the discovered gases through numerical values, which makes the obtained conclusions more convincing. In addition, scholars have established different coal spontaneous combustion stage division systems [41,44], and selected different signature gases in different stages as the basis of discrimination; however, most of the characteristic gas selection methods are

based on the change law of different gases in the programmed temperature experiment of coal and the possible chemical reaction of coal molecular groups, but this method does not have specific values, resulting in a lack of persuasion. The RBF-FAST method can quantify the relative importance of each gas at each stage of coal spontaneous combustion, and provide data support for the establishment of coal spontaneous combustion early warning systems.

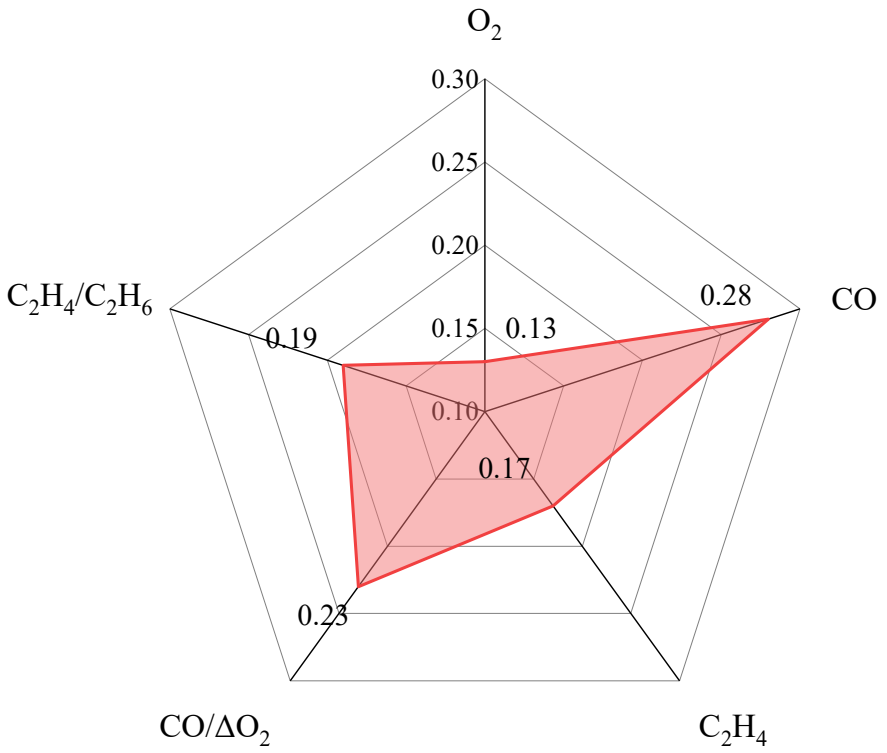

Figure 14. Relative importance of characteristic variables.

6. Conclusions

In this study, a new method used to obtain coal spontaneous combustion hazard grades is proposed based on the changing law of the concentration of various signature gases in the process of coal spontaneous combustion, and a prediction model of coal spontaneous combustion hazard grades is established. The findings can be summarized as follows:

(1) By analyzing the change rule of the experimental data of heating up coal in the spontaneous combustion procedure, six characteristic temperatures and their thresholds were determined, and the hazard classes of coal were classified into six classes: green (0), blue (1), purple (2), yellow (3), orange (4), and red (5).

(2) MeanRadius-SMOTE can be adopted to address the imbalance of the dataset. By comparing the predictive ability of four prediction models on different datasets, it was found that the proposed method performs the best when compared to the SMOTE and Kmeans–SMOTE methods.

(3) Three sets of comparative experiments were conducted in this research to compare the performance of different machine learning models in predicting coal spontaneous combustion hazard grades. The experimental results indicate that (1) the traditional PCA–Clustering–CBR model reduces the computational cost but also causes boundary information loss, resulting in lower prediction accuracy of machine learning models; (2) compared with the traditional PCA–Clustering–CBR, fuzzy clustering avoids the loss of boundary information and improves the computational efficiency of the model without affecting the prediction accuracy; and (3) PCA can improve the prediction

accuracy of machine learning models by calculating characteristic attribute weights based on the cumulative contribution rate.

(4) Aiming at the multi-objective optimization problem of the PCA-FM-CBR model, this manuscript adopted the SO algorithm to optimize γ_1, γ_2 and σ of the PCA-FM-CBR model step by step. The optimization shows that the model demonstrates optimal performance when the values of γ_1, γ_2 and σ are 0.71, 0.39, and 2, respectively. The calculation cost is reduced to the greatest extent.

(5) RBD-FAST was used to conduct sensitivity analysis for input variables, and the results demonstrated that CO is the most important input variable with a relative importance score of 0.28. Therefore, attention should be paid to CO in practical underground engineering.

Author Contributions: Data curation, Y.Z. and J.W.; writing—original draft preparation, Q.P. and Z.J.; writing—review and editing, J.L. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Basic Research Program (Free Exploration) Project of Shanxi Province, grant number 202303021222021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Y.M.; Liu, R.J.; Chen, X.Y.; Zou, X.Y.; Li, D.R.; Wang, S.S. Experimental Study on the Microstructural Characterization of Retardation Capacity of Microbial Inhibitors to Spontaneous Lignite Combustion. *Fire* **2023**, *6*, 20. [CrossRef]
2. Wei, D.Y.; Du, C.F.; Lei, B.; Lin, Y.F. Prediction and prevention of spontaneous combustion of coal from goafs in workface: A case study. *Case Stud. Therm. Eng.* **2020**, *21*, 9. [CrossRef]
3. Kursunoglu, N.; Gogebakan, M. Prediction of spontaneous coal combustion tendency using multinomial logistic regression. *Int. J. Occup. Saf. Ergon.* **2022**, *28*, 2000–2009. [CrossRef]
4. Guo, J.; Yan, H.; Liu, Y.; Li, S.S. Preventing spontaneous combustion of coal from damaging ecological environment based on thermogravimetric analysis. *Appl. Ecol. Environ. Res.* **2019**, *17*, 9051–9064. [CrossRef]
5. Lu, X.X.; Wang, M.Y.; Xue, X.; Xing, Y.; Shi, G.Y.; Shen, C.; Yang, Y.C.; Li, Y.B. An novel experimental study on the thermorunaway behavior and kinetic characteristics of oxidation coal in a low temperature reoxidation process. *Fuel* **2022**, *310*, 12. [CrossRef]
6. Kong, B.; Li, Z.H.; Yang, Y.L.; Liu, Z.; Yan, D.C. A review on the mechanism, risk evaluation, and prevention of coal spontaneous combustion in China. *Environ. Sci. Pollut. Res.* **2017**, *24*, 23453–23470. [CrossRef]
7. Wang, C.P.; Du, Y.X.; Deng, Y.; Zhang, Y.; Deng, J.; Zhao, X.Y.; Duan, X.D. Study on Spontaneous Combustion Characteristics and Early Warning of Coal in a Deep Mine. *Fire* **2023**, *6*, 18. [CrossRef]
8. Wang, C.P.; Chen, L.J.; Bai, Z.J.; Deng, J.; Liu, L.; Xiao, Y. Study on the dynamic evolution law of spontaneous coal combustion in high-temperature regions. *Fuel* **2022**, *314*, 12. [CrossRef]
9. Zhang, X.Q.; Zhou, F.Y.; Zou, J.X. Numerical Simulation of Gas Extraction in Coal Seam Strengthened by Static Blasting. *Sustainability* **2022**, *14*, 17. [CrossRef]
10. Xu, G.; Li, K.G.; Li, M.L.; Qin, Q.C.; Yue, R. Rockburst Intensity Level Prediction Method Based on FA-SSA-PNN Model. *Energies* **2022**, *15*, 19. [CrossRef]
11. Wang, J.F.; Liu, F.S.; Zhao, W.B.; Cai, H.L.; Zhao, J.; Liu, Y. Study on coal spontaneous combustion at low-medium temperature in the same coal seam with different buried depths and protolith temperatures. *Int. J. Coal Prep. Util.* **2022**, *42*, 3451–3463. [CrossRef]
12. Liu, C.D.; Zhang, R.; Wang, Z.X.; Zhang, X.Q. Research on the fire extinguishing performance of new gel foam for preventing and controlling the spontaneous combustion of coal gangue. *Environ. Sci. Pollut. Res.* **2023**, *30*, 88548–88562. [CrossRef]
13. Zhang, X.Q.; Pan, Y.Y. Preparation, Properties and Application of Gel Materials for Coal Gangue Control. *Energies* **2022**, *15*, 15. [CrossRef]
14. Guo, Q.; Ren, W.X.; Lu, W. Risk evaluation of coal spontaneous combustion from the statistical characteristics of index gases. *Thermochim. Acta* **2022**, *715*, 10. [CrossRef]
15. Li, S.; Xu, K.; Xue, G.Z.; Liu, J.; Xu, Z.Q. Prediction of coal spontaneous combustion temperature based on improved grey wolf optimizer algorithm and support vector regression. *Fuel* **2022**, *324*, 11. [CrossRef]

16. Wang, F.S.; Xu, Y.Y.; Song, Z.Q.; Guo, L.W. Designing system predicting coal spontaneous combustion by means of method of gas analysis. In Proceedings of the 3rd International Symposium on Modern Mining and Safety Technology, Fuxin, China, 4–6 August 2008; pp. 326–328.
17. Guo, Q.; Ren, W.X.; Lu, W. A Method for Predicting Coal Temperature Using CO with GA-SVR Model for Early Warning of the Spontaneous Combustion of Coal. *Combust. Sci. Technol.* **2022**, *194*, 523–538. [CrossRef]
18. Shukla, R.; Khandelwal, M.; Kankar, P.K. Prediction and Assessment of Rock Burst Using Various Meta-heuristic Approaches. *Min. Metall. Explor.* **2021**, *38*, 1375–1381. [CrossRef]
19. Zhang, L.D.; Song, Z.Y.; Wu, D.J.; Luo, Z.M.; Zhao, S.S.; Wang, Y.H.; Deng, J. Prediction of coal self-ignition tendency using machine learning. *Fuel* **2022**, *325*, 17. [CrossRef]
20. Guo, J.; Chen, C.M.; Wen, H.; Cai, G.B.; Liu, Y. Prediction model of goaf coal temperature based on PSO-GRU deep neural network. *Case Stud. Therm. Eng.* **2024**, *53*, 13. [CrossRef]
21. Wang, W.; Liang, R.; Qi, Y.; Cui, X.; Liu, J. Prediction model of spontaneous combustion risk of extraction borehole based on PSO-BPNN and its application. *Sci. Rep.* **2024**, *14*, 5. [CrossRef]
22. Li, X.P.; Zhang, J.; Ren, X.P.; Liu, Y.Q.; Zhou, C.H.; Li, T.Y. Study on condition analysis and temperature prediction of coal spontaneous combustion based on improved genetic algorithm. *AIP Adv.* **2022**, *12*, 10. [CrossRef]
23. Watson, I.; Marir, F. Case-based reasoning: A review. *Knowl. Eng. Rev.* **1994**, *9*, 327. [CrossRef]
24. Deng, S.G.; Li, W.S. Spatial case revision in case-based reasoning for risk assessment of geological disasters. *Geomat. Nat. Hazards Risk* **2020**, *11*, 1052–1074. [CrossRef]
25. Yang, S.K.; Bian, C.; Li, X.; Tan, L.; Tang, D.X. Optimized fault diagnosis based on FMEA-style CBR and BN for embedded software system. *Int. J. Adv. Manuf. Technol.* **2018**, *94*, 3441–3453. [CrossRef]
26. Chen, M.Q.; Xia, J.Y.; Huang, R.Y.; Fang, W.G. Case-Based Reasoning System for Aeroengine Fault Diagnosis Enhanced with Attitudinal Choquet Integral. *Appl. Sci.* **2022**, *12*, 16. [CrossRef]
27. Perez-Pons, M.E.; Parra-Dominguez, J.; Hernandez, G.; Bichindaritz, I.; Corchado, J.M. OCI-CBR: A hybrid model for decision support in preference-aware investment scenarios. *Expert Syst. Appl.* **2023**, *211*, 10. [CrossRef]
28. Guerrero, J.I.; Miró-Amarante, G.; Martín, A. Decision support system in health care building design based on case-based reasoning and reinforcement learning. *Expert Syst. Appl.* **2022**, *187*, 7. [CrossRef]
29. Zhao, Z.; Chen, J.H.; Yao, J.M.; Xu, K.H.; Liao, Y.Y.; Xie, H.W.; Gan, X.X. An improved spatial case-based reasoning considering multiple spatial drivers of geographic events and its application in landslide susceptibility mapping. *Catena* **2023**, *223*, 13. [CrossRef]
30. Dorodnykh, N.; Nikolaychuk, O.; Pestova, J.; Yurin, A. Forest Fire Risk Forecasting with the Aid of Case-Based Reasoning. *Appl. Sci.* **2022**, *12*, 24. [CrossRef]
31. Khan, M.J.; Hayat, H.; Awan, I. Hybrid case-base maintenance approach for modeling large scale case-based reasoning systems. *Hum.-Centric Comput. Inf. Sci.* **2019**, *9*, 25. [CrossRef]
32. Liang, D.C.; Fu, Y.Y.; Xu, Z.S. Time-Varying Intuitionistic Fuzzy Integral for Emergency Materials Demand Prediction With Case-Based Reasoning. *IEEE Trans. Fuzzy Syst.* **2022**, *30*, 3617–3632. [CrossRef]
33. Zhang, H.; Yang, J. A Case Retrieval Strategy for Traffic Congestion Based on Cluster Analysis. *Math. Probl. Eng.* **2022**, *2022*, 8. [CrossRef]
34. Kuo, R.J.; Cha, C.L.; Chou, S.H. Developing a diagnostic system through the integration of ant colony optimization systems and case-based reasoning. *Int. J. Adv. Manuf. Technol.* **2006**, *30*, 750–760. [CrossRef]
35. Lin, M.C.; He, D.B.; Sun, S.X. Multivariable Case Adaptation Method of Case-Based Reasoning Based on Multi-Case Clusters and Multi-Output Support Vector Machine for Equipment Maintenance Cost Prediction. *IEEE Access* **2021**, *9*, 151960–151971. [CrossRef]
36. Khan, M.J.; Khan, C. Performance evaluation of fuzzy clustered case-based reasoning. *J. Exp. Theor. Artif. Intell.* **2021**, *33*, 313–330. [CrossRef]
37. Wang, C.; Hu, P.; Sun, Y.; Yang, C. Study on CO source identification and spontaneous combustion warning concentration in the return corner of working face in shallow buried coal seam. *Environ. Sci. Pollut. Res.* **2024**, *31*, 15050–15064. [CrossRef]
38. Li, L.; Ren, T.; Zhong, X.X.; Wang, J.T. Study of the Abnormal CO-Exceedance Phenomenon in the Tailgate Corner of a Low Metamorphic Coal Seam. *Energies* **2022**, *15*, 16. [CrossRef]
39. Liang, Y.T.; Song, S.L.; Guo, B.L.; Gao, L.Y.; Liu, J.F.; Lu, W.; Wang, W.; Kong, B. Study on the Coupling Characteristics of Infrasound-Temperature-Gas in the Process of Coal Spontaneous Combustion and a New Early Warning Method. *Combust. Sci. Technol.* **2023**. [CrossRef]
40. Peng, J. Research on Prediction Model of Coal Spontaneous Combustion Temperature Based on Machine Learning. Master's Thesis, Xi'an University of Science and Technology, Xi'an, China, 2020.
41. Zhang, D.; Cen, X.X.; Wang, W.F.; Deng, J.; Wen, H.; Xiao, Y.; Shu, C.M. The graded warning method of coal spontaneous combustion in Tangjiahui Mine. *Fuel* **2021**, *288*, 7. [CrossRef]
42. Xu, X.F.; Zhang, F.J. Evaluation and Optimization of Multi-Parameter Prediction Index for Coal Spontaneous Combustion Combined with Temperature Programmed Experiment. *Fire* **2023**, *6*, 15. [CrossRef]
43. Lei, P. Study on Early warning technology of coal spontaneous combustion in goaf of Linhuan 9 Coal Seam. Master's Thesis, Anhui University of Science and Technology, Anhui, China, 2022.

44. Biao, F.J. Study on Stage Determination Theory and Classified Early Warning Method for Spontaneous Combustion of Coal. Ph.D. Thesis, Xi'an University of Science and Technology, Xi'an, China, 2019.
45. Duan, F.; Zhang, S.; Yan, Y.; Cai, Z. An Oversampling Method of Unbalanced Data for Mechanical Fault Diagnosis Based on MeanRadius-SMOTE. *Sensors* **2022**, *22*, 5166. [CrossRef]
46. Aamodt, A.; Plaza, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Commun.* **1994**, *7*, 39–59. [CrossRef]
47. Wu, H.T.; Zhong, B.T.; Medjdoub, B.; Xing, X.J.; Jiao, L. An Ontological Metro Accident Case Retrieval Using CBR and NLP. *Appl. Sci.* **2020**, *10*, 24. [CrossRef]
48. Askari, S. Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Syst. Appl.* **2021**, *165*, 27. [CrossRef]
49. Hashim, F.A.; Hussien, A.G. Snake Optimizer: A novel meta-heuristic optimization algorithm. *Knowl.-Based Syst.* **2022**, *242*, 34. [CrossRef]
50. Zheng, W.M.; Pang, S.Y.; Liu, N.; Chai, Q.W.; Xu, L.D. A Compact Snake Optimization Algorithm in the Application of WKNN Fingerprint Localization. *Sensors* **2023**, *23*, 16. [CrossRef]
51. Yan, X.; Tu, N.; Wu, S.; Zhu, Y. Dynamic Prediction of Coal and Gas Outburst Based on Clustering and Case-Based Reasoning. *Chin. J. Sens. Actuators* **2015**, *28*, 7.
52. Trajdos, P.; Kurzynski, M. Weighting scheme for a pairwise multi-label classifier based on the fuzzy confusion matrix. *Pattern Recognit. Lett.* **2018**, *103*, 60–67. [CrossRef]
53. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
54. Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
55. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [CrossRef]
56. Mara, T.A. Extension of the RBD-FAST method to the computation of global sensitivity indices. *Reliab. Eng. Syst. Saf.* **2009**, *94*, 1274–1281. [CrossRef]
57. Gao, B.; Yang, Q.; Peng, Z.J.; Xie, W.H.; Jin, H.; Meng, S.H. A direct random sampling method for the Fourier amplitude sensitivity test of nonuniformly distributed uncertainty inputs and its application in C/C nozzles. *Aerosp. Sci. Technol.* **2020**, *100*, 8. [CrossRef]
58. Xu, Q.; Yang, S.Q.; Cai, J.W.; Zhou, B.Z.; Xin, Y.A. Risk forecasting for spontaneous combustion of coals at different ranks due to free radicals and functional groups reaction. *Process Saf. Environ. Protect.* **2018**, *118*, 195–202. [CrossRef]
59. Zhang, Y.T.; Shi, X.Q.; Li, Y.Q.; Liu, Y.R. Characteristics of carbon monoxide production and oxidation kinetics during the decaying process of coal spontaneous combustion. *Can. J. Chem. Eng.* **2018**, *96*, 1752–1761. [CrossRef]
60. Yan, H.W.; Nie, B.S.; Liu, P.J.; Chen, Z.Y.; Yin, F.F.; Gong, J.; Lin, S.S.; Wang, X.T.; Kong, F.B.; Hou, Y.N. Experimental investigation and evaluation of influence of oxygen concentration on characteristic parameters of coal spontaneous combustion. *Thermochim. Acta* **2022**, *717*, 11. [CrossRef]
61. Wu, K.; Yao, Q.; Chen, Y.; Zhao, P.T.; Xi, C.Z.; Zhao, Y.; Wang, Q. Dependence evaluation of factors influencing coal spontaneous ignition. *Energy Sci. Eng.* **2023**, *11*, 3738–3750. [CrossRef]
62. Mohalik, N.K.; Lester, E.; Lowndes, I.S. Review of experimental methods to determine spontaneous combustion susceptibility of coal—Indian context. *Int. J. Min. Reclam. Environ.* **2017**, *31*, 301–332. [CrossRef]