processes

MDPI

Article

A Gas Emission Prediction Model Based on Feature Selection and Improved Machine Learning

Liangshan Shao [1] and Kun Zhang [2,*]

[1] Liaoning Institute of Science and Engineering, Jinzhou 121000, China
[2] College of Business Administration, Liaoning Technical University, Huludao 125105, China
* Correspondence: 472021099@stu.intu.edu.cn

Abstract: This paper proposed a gas emission prediction method based on feature selection and improved machine learning, as traditional gas emission prediction models are neither accurate nor universally applicable. Through analysis, this paper identified 12 factors that affected gas emissions. A total of 30 groups of typical data for gas outflow were standardized, after which a full subset regression feature selection method was used to categorize 12 influencing factors into different regular patterns and select 18 feature parameter sets. Meanwhile, based on nuclear principal component analysis (KPCA), an optimized gas emission prediction model was constructed where the dimensionality of the original data was reduced. An optimized algorithm set was constructed based on the hybrid kernel extreme learning machine (HKELM) and the least squares support vector machine (LSSVM). The performance of feature parameters adopted in the prediction algorithm was evaluated according to certain metrics. By comparing the results of different sets, the final prediction sequence could be obtained, and a model that was composed of the optimal feature parameters was applied to the optimal machine learning algorithm. The results showed that the HKELM outperformed LSSVM in prediction accuracy, running speed, and stability. The root meant square error (RMSE) for the final prediction sequence was 0.22865, the determination coefficient (R2) was 0.99395, the mean absolute error (MAE) was 0.20306, and the mean absolute percentage error (MAPE) was 1.0595%. Every index of accuracy evaluation performed well and the constructed prediction model had high-prediction accuracy and a wide application.

Keywords: gas emission; regression forecasting; feature selection; machine learning; hybrid kernel extreme learning machine

check for updates

Citation: Shao, L.; Zhang, K. A Gas Emission Prediction Model Based on Feature Selection and Improved Machine Learning. *Processes* **2023**, *11*, 883. https://doi.org/10.3390/pr11030883

Academic Editors: Xiao Feng and Minbo Yang

Received: 7 February 2023
Revised: 7 March 2023
Accepted: 10 March 2023
Published: 15 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Since carbon peaking and carbon neutrality goals were put forward, China has witnessed a declining coal demand. However, coal will still play a major role in the energy sector as this sector is thought to be the lifeblood of the national economy [1,2]. At present, most coal mines in China still function underground; coal producers are faced with a complex geological environment that gives rise to security risks. The deeper the coal mining goes down, the more gas is emitted, resulting in gas-related disasters. One of the major causes of such disasters is the excessive amount of gas emissions. Therefore, in order to prevent gas-related disasters, it is imperative to foster gas storage laws and regulations, predict accurately the amount of gas emitted, and adopt preventive measures beforehand.

Gas emission is complex and dynamic. There are many influencing factors, such as information overlap, but these factors do not affect gas emissions proportionally [3]. Therefore, in order to predict gas emissions accurately and in time, new gas emission prediction techniques and methods are called upon [4]. Scholars at home and abroad have done a lot of research in this regard [5–8]. Traditional prediction methods mainly focused on mine statistics based on mining seam depth data, sub-source prediction based on gusher sources, and mathematical modeling based on geological data. Although they are convenient to

predict gas emissions in shallow coal seams, influencing factors usually generate combining effects on gas emission due to the complexity of the mine's geological structure and as coal seam mining goes deeper. Old methods fail to address this issue properly. Therefore, the traditional methods are gradually losing their advantages in predicting coal seam gas emissions and giving way to artificial intelligence algorithms that benefit from the wide application of computer technology. New methods can not only improve prediction accuracy significantly but also reduce the time for prediction. Wang Yuhong et al. [9] proposed a gas emission prediction model based on variational mode decomposition and depth integration, the original data is decomposed into high frequency and low frequency components, and the results are predicted separately and added linearly. Dai Wei et al. [10] referred to variational mode decomposition (VMD) and adopted the differential evolution (DE) algorithm and correlation vector machine (RVM) to predict the absolute gas emission in stope face. Li Bing et al. [11] combined principal component analysis (PCA) and extreme learning machine (ELM) neural network to establish a mine gas emission prediction model based on PCA-ELM. Xiao et al. [12] used BP neural network improved by a compressed mapping genetic algorithm to construct a CMGANN-coupled algorithm on the basis of data dimensionality reduction through kernel principal component analysis. Wen et al. [13] constructed a PSO-BP-Adaboost combined prediction model based on gas emission source prediction. Moreover, Chen Weihua et al. [14] proposed an improved model by using Chaos immune genetic optimization algorithm (CIGOA) to improve Elman neural network. Peng Xiaohua et al. [15] first introduced the wavelet packet neural network model to achieve the same purpose. Xu et al. and Ma and Li [16,17] improved the weights and threshold parameters of BP neural network with different optimization algorithms. Wang Yuanbin et al. [18] applied principal component dimensionality reduction to the data and introduced them to the XGBoost model of Bayesian optimization (BOA) hyperparameters, thus improving the prediction accuracy. Chen Qian et al. [19] achieved dimensionality reduction through the LARS algorithm by proposing irrelevant and redundant features and adopting the LASSO penalty regression prediction model for simulation.

The above-mentioned scholars introduced better and more stable models to predict gas emissions in an efficient and time-saving way. Their methods improved the prediction accuracy. However, the inherent law of gas flow is yet to be discovered. Some of these scholars failed to address the nonlinear and non-stationary features of gas emission data, and some reduced the dimensionality of the data blindly. As a result, the original features of the data [20] were lost. They also failed to find out the main control factors that caused gas emissions. Moreover, how their findings could affect the data awaited an explanation. This paper proposed a gas emission prediction model based on feature selection and an improved machine learning algorithm. Using the full subset regression method, this paper combined feature parameters of influencing factors to become different sets. These sets were then applied to different optimized prediction algorithms for comparison [21]. An evaluation of the results was conducted to find out the optimal combination of the parameter set and algorithm. This method can be used to predict gas emissions under many circumstances and makes gas emission prediction more accurate.

2. Data Processing for Gas Emission Prediction

2.1. Initial Index System of Gas Emission Prediction

This paper conducted a case study of coal seams in the southern part of the Yan'an Huangling mine and constructed a prediction index system of gas emission from the perspective of natural geology and mining technology. The prediction data, excluding irrelevant factors, are shown in Table 1 (only five groups of data are listed due to limited space), including original gas content (X1, m^3/t), the gas content of adjacent seam (X2, m^3/t), coal seam thickness (X3, m), coal seam buried depth (X4, m), coal seam dip angle (X5, °), coal seam spacing (X6, m), interlayer lithology (X7, m), floor elevation (X8, m), advance speed (X9, m/d), gas pressure (X10, MPa), gas extraction pure quantity (X11, m^3/min), and roof management mode (X12). Among them, 24 groups were used as training sets, while the

last 6 groups were used as test sets. It can be seen from Figure 1 that there is no obvious linear correlation between mine gas emission and influencing factors(Black square is the fitting equation and related parameters, red line is the fitting curve).

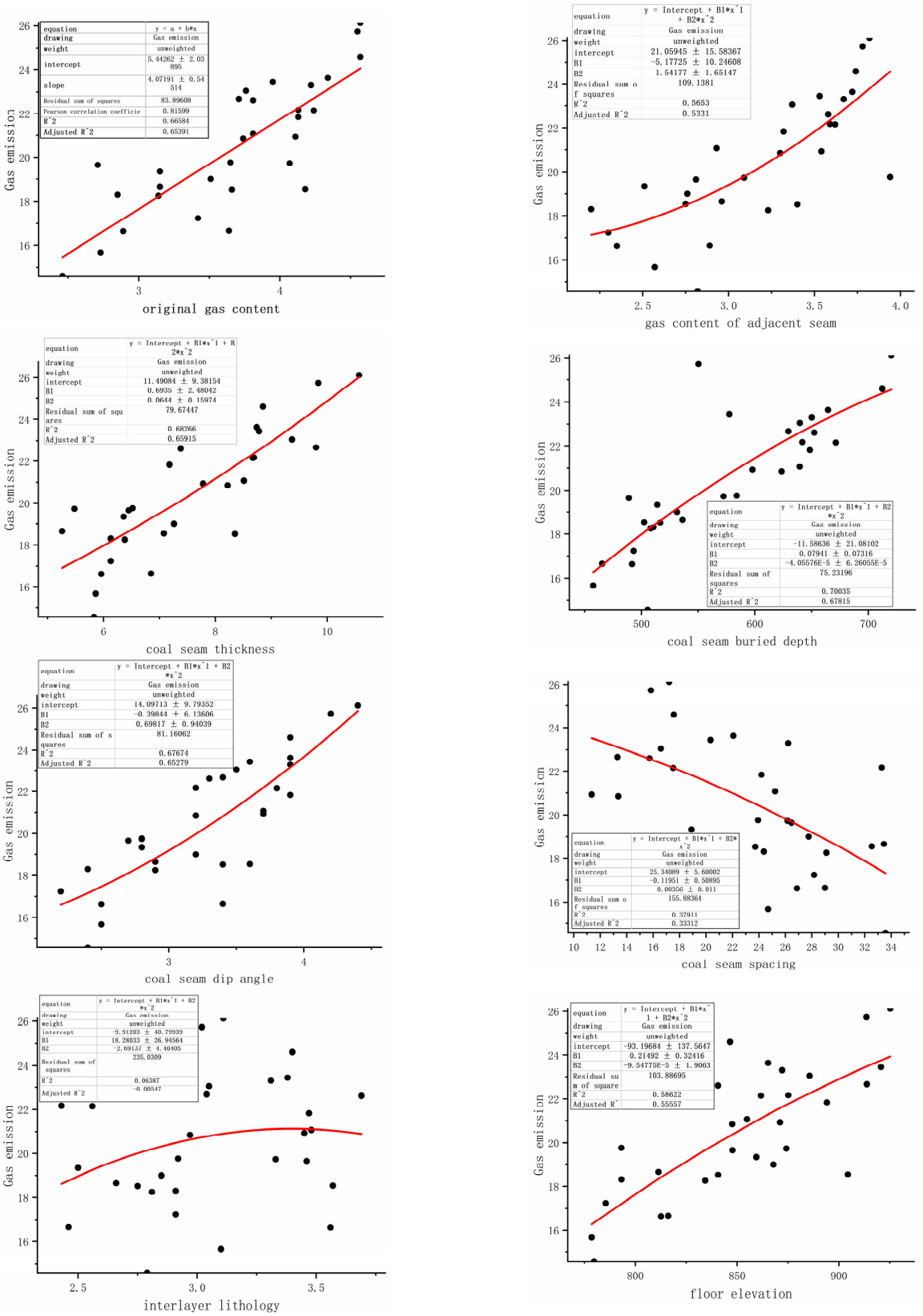


Figure 1. *Cont.*

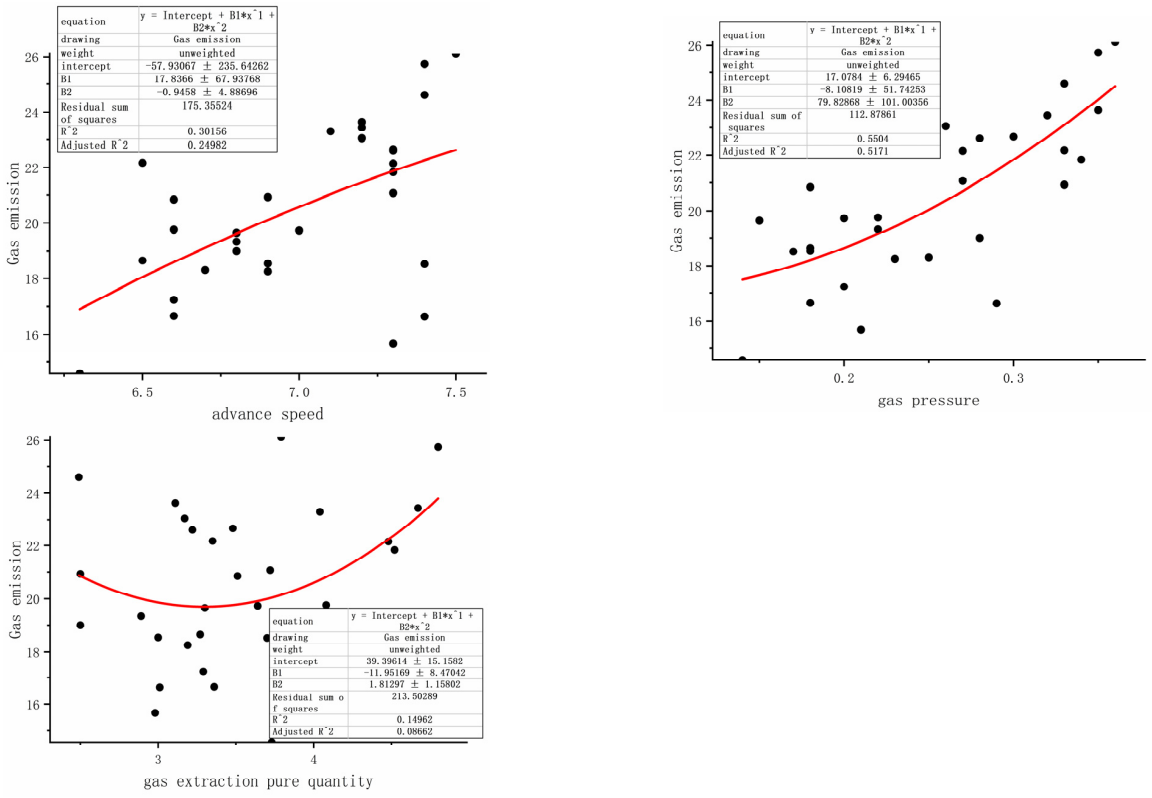


Figure 1. Linear fitting diagram of gas emission and each quantitative index.

Table 1. Mine gas emission and influencing factors.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	Y
1	4.55	3.78	9.84	550.32	4.20	15.83	3.02	913.54	7.40	0.35	4.80	1	25.73
2	3.95	3.53	8.78	577.5	3.60	20.34	3.38	920.65	7.20	0.32	4.67	1	23.44
3	2.85	2.2	6.13	510.58	2.40	24.37	2.91	793.20	6.70	0.25	4.50	1	18.30
4	3.81	2.93	8.51	639.53	3.70	25.22	3.48	854.84	7.30	0.27	3.72	1	21.07
5	4.22	3.67	7.54	650.12	3.90	26.21	3.31	872.09	7.10	0.23	4.04	1	23.30
6	4.13	3.59	8.69	641.82	3.20	33.28	2.43	875.15	6.50	0.33	3.35	1	22.17
7	4.34	3.72	8.74	664.48	3.90	22.06	2.90	865.18	7.20	0.35	3.11	1	23.63
8	4.57	3.82	10.57	720.22	4.40	17.21	3.11	925.24	7.50	0.36	3.79	1	26.12
9	3.81	3.58	7.38	652.35	3.30	15.73	3.69	840.59	7.30	0.28	3.22	1	22.61
10	2.89	2.35	5.96	491.75	2.50	26.87	3.56	812.59	7.40	0.29	3.01	1	16.63
11	3.14	3.23	6.38	508.17	2.90	29.10	2.81	834.33	6.90	0.23	3.19	1	18.25
12	4.57	3.74	8.85	712.25	3.90	17.56	3.40	846.53	7.40	0.33	2.49	1	24.60
13	3.51	2.76	7.26	531.35	3.20	27.76	2.85	867.83	6.80	0.28	2.50	1	19.00
14	3.71	2.84	9.8	629.55	3.40	13.30	3.04	913.71	7.30	0.30	3.48	1	22.67
15	3.76	3.37	9.37	639.67	3.50	16.58	3.05	885.61	7.20	0.26	3.17	1	23.05
16	3.15	2.51	6.36	514.03	2.80	18.90	2.50	859.43	6.80	0.22	2.89	1	19.34
17	4.11	3.54	7.78	597.87	3.70	11.35	3.45	871.07	6.90	0.33	2.50	1	20.93
18	4.18	2.75	7.08	502.45	3.60	32.53	3.57	904.41	6.90	0.18	3.00	1	18.54
19	2.71	2.81	6.45	488.96	2.70	26.46	3.46	847.72	6.80	0.15	3.30	1	19.65
20	3.64	2.89	6.85	465.42	3.40	28.99	2.46	816.14	6.60	0.18	3.36	1	16.65
21	3.66	3.40	8.35	516.57	3.40	23.72	2.75	840.67	7.40	0.17	3.70	1	18.52
22	4.07	3.09	5.48	572.34	2.80	26.16	3.33	874.15	7.00	0.20	3.64	1	19.73
23	3.74	3.30	8.22	623.52	3.20	13.36	2.97	847.57	6.60	0.18	3.51	1	20.85
24	2.73	2.57	5.86	457.53	2.50	24.69	3.10	778.53	7.30	0.21	2.98	1	15.67
25	3.42	2.30	6.13	493.20	2.20	28.17	2.91	785.41	6.60	0.20	3.29	1	17.24
26	3.65	3.94	6.52	584.00	2.80	23.93	2.92	793.10	6.60	0.22	4.08	1	19.76
27	3.15	2.96	5.26	536.24	2.90	33.45	2.66	811.35	6.50	0.18	3.27	1	18.65
28	4.13	3.32	7.18	648.45	3.90	24.19	3.47	894.11	7.30	0.34	4.52	1	21.83
29	4.24	3.62	8.67	671.30	3.80	17.52	2.56	861.71	7.30	0.27	4.48	1	22.15
30	2.46	2.82	5.83	505.57	2.40	33.57	2.79	779.70	6.30	0.14	3.73	1	14.57

Note: There are three main roof management methods: full caving method, filling method, and coal pillar support method, which are represented by 1, 2, and 3 in the data set. Interlayer lithology: the weighted average of surrounding rock hardness is quantified by formula conversion.

2.2. Data Standardization Processing

If the original data were directly applied to the analysis, the prediction results would be biased. This is because these indices have different magnitude orders. To improve prediction accuracy, it is necessary to standardize the data before constructing the model [22]. Therefore, zero-mean normalization (z-scor) was used to serve this purpose. After going through Equations (1)–(3), a new set of sequences were obtained, which are able to effectively enhance prediction accuracy:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{2}$$

$$h_i = \frac{x_i - \bar{x}}{s} \tag{3}$$

In these formulas, x_i is the original sequence, $\overline{x}$ is the average value of the sequence, s is the standard deviation, and h_i represents the new sequence after transformation, $i \in [1, n]$.

3. Construction of the Mine Gas Emission Prediction Model

3.1. Determination of Characteristic Parameter Sets for Gas Emission Prediction

Results show that selecting variables in a scientific-based way may enhance the accuracy of the model (Figure 2), especially when gas emission is affected by different factors whose coupling effect may pose a greater impact. To determine characteristic parameter sets, we need to reduce the dimensionality of the data, so as to reduce the complexity of the model operation and avoid overfitting caused by excessive dimensions [23]. Meanwhile, in order to ensure the accuracy of the prediction model, it is also necessary to make sure that the selected characteristic parameter sets retain as many data features as possible.

3.1.1. Total Subset Regression

Correlation analysis was applied to each influencing parameter, and the correlation coefficient of each factor with gas emission was calculated according to the Spearman correlation coefficient. The purpose of doing so was to determine how important each influencing factor was in the gas emission index system. As shown in Figure 3, the correlation coefficients of 0.8–1.0, 0.6–0.8, 0.4–0.6, 0.2–0.4, and 0–0.2 indicates an extremely strong correlation, strong correlation, moderate correlation, weak correlation, and moderate irrelevancy or strong irrelevancy, respectively.

Twelve factors affecting gas emission were randomly combined based on the full subset regression method, and a subset of gas emission characteristic parameters was constructed after the importance of characteristic parameters and the correlation between each factor were considered. In order to streamline calculation, R^2 was used. R^2 is the determination coefficient that reflects the accuracy of model fitting data and ranges from 0 to 1. The closer the value is to 1, the more the variable in the equation explains y and the better the model fits the data. The least square fitting was applied to all variable combinations, and a total of 18 characteristic parameter sets with a sound fitting effect of $R^2 \geq 0.90$ were selected, which are expressed as F-1, ... At the same time, the original set was expressed as F-0, and the selection of each influencing factor in the set is shown in Table 2 with "☆".

Table 2. Gas emission characteristic parameter set.

Influencing Factor	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-0	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆
F-1	☆		☆	☆	☆		☆		☆	☆		
F-2	☆			☆	☆	☆	☆	☆	☆			
F-3	☆	☆			☆	☆		☆	☆	☆		
F-4	☆			☆	☆	☆		☆	☆	☆		
F-5	☆		☆	☆	☆			☆	☆	☆		
F-6	☆			☆	☆		☆		☆	☆		
F-7	☆			☆	☆		☆	☆		☆		
F-8	☆	☆		☆			☆		☆	☆		
F-9	☆			☆	☆		☆	☆		☆		
F-10	☆			☆				☆	☆	☆		
F-11	☆			☆	☆			☆	☆	☆		
F-12	☆				☆	☆		☆	☆	☆		
F-13	☆			☆	☆		☆	☆	☆			
F-14	☆		☆	☆			☆	☆		☆		
F-15	☆			☆	☆	☆		☆		☆		
F-16	☆		☆	☆	☆		☆	☆	☆			
F-17	☆	☆		☆	☆			☆	☆	☆		
F-18	☆		☆	☆	☆			☆		☆		

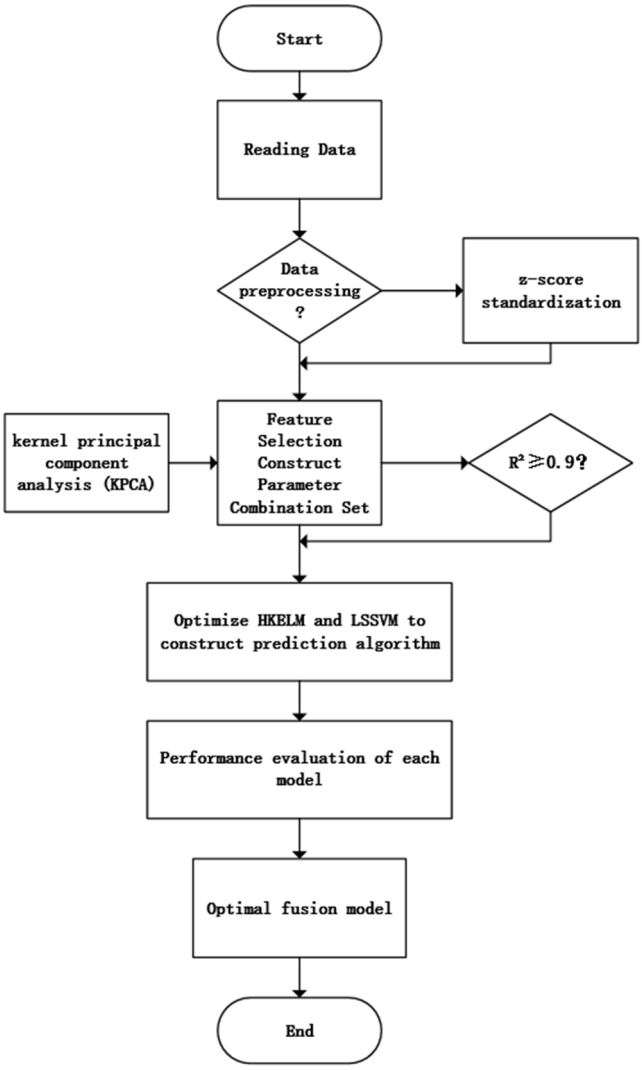


Figure 2. The flow chart for constructing the prediction model.

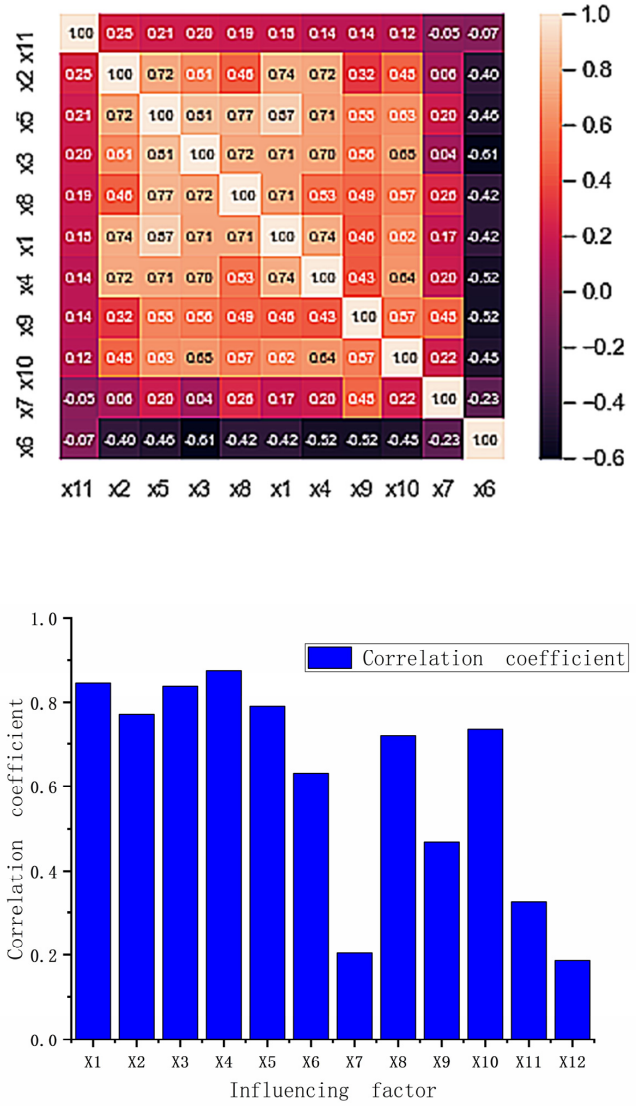


Figure 3. Correlation intensity between characteristic parameters and gas emission.

3.1.2. Kernel Principal Component Analysis (KPCA)

A linear dimensionality reduction method is usually used to deal with the phenomenon data. However, in practice, most data are so nonlinear that they are difficult to be accurately described by linear dimensionality reduction methods. Hence, a nonlinear dimension reduction method is needed to retain local features. Kernel principal component analysis (KPCA) is such an analysis method. It is suitable for conducting dimensionality reduction to influencing factors, given that there are too many of them and they present obvious nonlinear features. According to the KPCA algorithm, the sample data set is assumed to be X_k (k = 0, 1, 2, . . . , 12; there are 12 factors affecting mine gas emission), $X_k \in R$. By introducing a nonlinear function $\Phi(X_k)$, the data sample might be converted into a high-dimensional space, and the covariance matrix C is as follows:

$$C = \frac{1}{m}\sum_{k=1}^{m} \varphi(x_k)^T \quad (4)$$

The eigenvalues of the matrix are able to be solved by functions $\varphi(x_k)$ to solve the corresponding eigenvectors.

$$\phi(x_k)V - \lambda\phi(x_k)V = 0 \quad (5)$$

where V is represented by $\varphi(x_i)$, namely:

$$V = \sum_{i=1}^{m} a_i \varphi(x_i) \tag{6}$$

a_i is the Lagrange multiplier. After combining Formulas (2) and (3) and adding kernel function K, we can get:

$$m\lambda a - Ka = 0 \tag{7}$$

where a refers to the eigenvector of kernel function K.

Under the condition of $\sum_{k=1}^{m} \varphi(x_k) = 0$, the above formula is deduced. However, since many data cannot meet the above conditions, it is necessary to transform the kernel function K:

$$\widetilde{K} = K - L_n K - K L_n + L_n K L_n \tag{8}$$

where L_n represents a matrix of x_n and can represent that the elements in the matrix are $\frac{1}{n}$ (n∈R). According to the formula $\sum_{k=1}^{s} \lambda_k / \sum_{k=1}^{m} \lambda_k$, the contribution rate of each factor affecting the amount of gas emission can be calculated. Influencing factors with a cumulative contribution rate of more than 85% (Table 3) can be added to the index.

Table 3. Variance contribution rate.

Kernel Principal Component	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
variance contribution rate %	53.11	11.41	8.31	6.57	5.63	4.30	3.38	2.14	1.75	1.35	0.90	0.52

The cumulative contribution of the core main component index can be calculated. As depicted in Figure 4, the cumulative contribution is 85.04%, and five core main component indices are extracted. Since the dimensionality reduction condition is met, the number of extracted core principal component indices can be determined, which is 5. The data combination of dimensionality reduction index obtained through kernel principal component analysis is kept as the gas emission characteristic parameter sets obtained through F-K, with the original data added. A total of 20 characteristic parameter sets are used as model inputs.

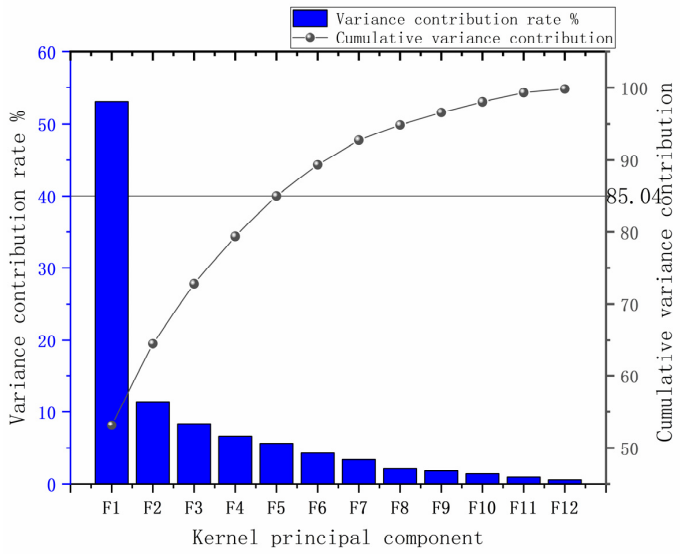


Figure 4. Cumulative contribution rate.

3.2. A Selection of Gas Emission Prediction Algorithms

This paper did a literature review and an experimental comparison and selected two algorithms, the least square support vector machine (LSSVM) and the hybrid kernel Extreme Learning Machine (HKELM) to predict gas emission. The newest optimal algorithms include the sparrow search algorithm (SSA), the genetic algorithm (GA), particle swarm optimization (PSO), the whale optimization algorithm (WOA), the moth flame optimization algorithm (MFO), and the slime mold optimization algorithm (SMA), all of which can optimize the least squares support vector machine (LSSVM) and the hybrid kernel extreme learning machine (HKELM). The prediction results from different parameter sets applied to different algorithms were compared to choose from the best, as described in Figure 5.

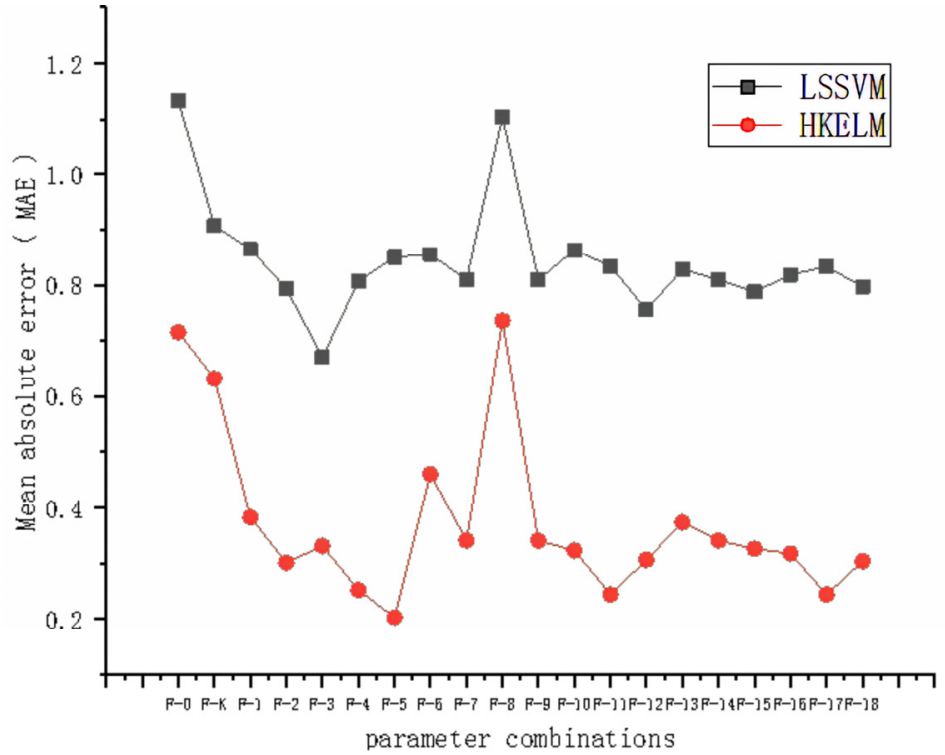


Figure 5. Comparison of the optimization results of different parameter sets under different algorithms.

The kernel extreme learning machine is a single hidden layer feedforward neural network. The kernel function is introduced based on ELM. The traditional gradient descent training algorithm has a large number of iterations and often falls into the local optimum, whereas kernel functions do better in this regard. Meanwhile, it maintains excellent generalization characteristics and fast learning speed. Based on this, HKELM increases the number of kernel functions to further improve the general performance of the model.

The least squares support vector machine is developed on the basis of the support vector machine, with the error sum of squares loss function being the training set [24]. The inequality constraint is changed into the equality constraint. So, the question is shifted from solving the quadratic plan problem to solving a linear equation set. As a result, the accuracy and relevance of the solved data can be improved.

The Whale Watching Algorithm (WOA) is a meta-heuristic optimization algorithm that simulates the hunting behavior and spiral simulation attack of humpback whales to search for the optimal agent [25]. It is easy to operate, requires few parameters, and can achieve better optimization results.

It is clear from Figure 5 that HKELM demonstrates better predictive performance over LSSVM, and the optimized hybrid kernel extreme learning machine algorithm shows advantages in running speed and stability during the algorithm operation.

4. Optimal Fusion Model Selection

4.1. Determination of the Optimal Parameter Set

To verify the performance of the model, RMSE, MAE, MAPE, R2, and NSE were selected as evaluation indices to test the HKELM model with different algorithms [26].

The root mean square error (*RMSE*) is used to measure the deviation between the predicted value and the true value, i.e., the number of root mean square errors:

$$RMSE(X,\ h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2} \tag{9}$$

The mean absolute error (*MAE*) is the average of the absolute error and is better able to reflect the actual situation of the predicted value error:

$$MAE(X,\ h) = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i| \tag{10}$$

The mean absolute percentage error (*MAPE*) is employed as a statistical indicator to measure the prediction accuracy:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{11}$$

The Nash coefficient (NSE) is an evaluation indicator to evaluate the quality of the model:

$$E = 1 - \frac{\sum_{t=1}^{T}(Q_0^t - Q_m^t)^2}{\sum_{t=1}^{T}(Q_0^t - \bar{Q}_0)^2} \tag{12}$$

In this formula, Q_o refers to the observed value, Q_m refers to the simulated value, Q_t represents a value at time t, and $\bar{Q}_0$ represents the overall average observed value. The value of E is negative infinity to 1. Additionally, the closer E is to 1, the higher the quality and credibility of the model.

R^2: Determination coefficient, which reflects the accuracy of the model fitting data and ranges from 0 to 1. The closer the value is to 1, the more the variable in the equation explains y and the better the model fits to the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{13}$$

The mean absolute percentage error (*MAPE*) of 108 fusion model results is less than or equal to 2%, and 42 fusion models are obtained. F-5 is the optimal parameter set for gas emission prediction. The distribution of each result from different parameter sets is demonstrated in Figure 6.

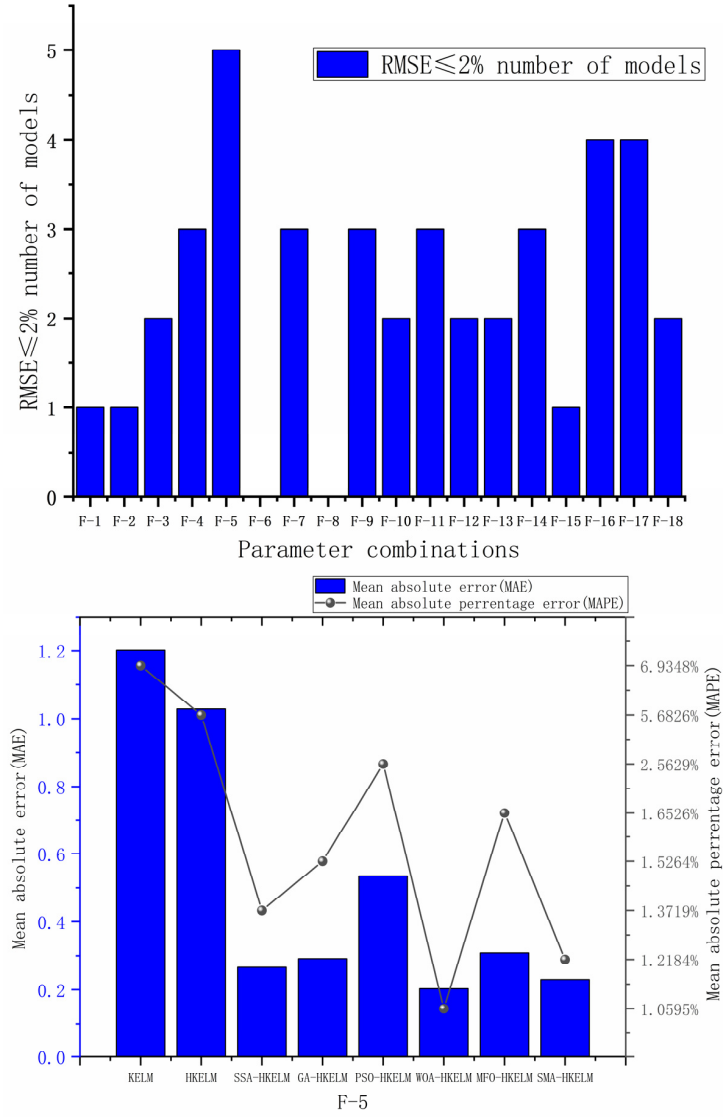


Figure 6. Results of each model.

4.2. Determination of the Optimal Improved Machine Learning Algorithm

In order to obtain the optimal fusion model for gas emission prediction with the most suitable characteristic parameter set and the most effective algorithm, the calculation results from 18 parameter sets applied to different prediction algorithms are compared, and the fusion model with MAPE $\leq$ 1.50% stands out. These results are shown in Table 4.

Table 4. Parameter combination and prediction algorithm fusion model results.

Parameter Combinations	Improved Algorithm	*RMSE*	*MAE*	*MAPE*	R^2	*NSE*
F-0	GA-HKELM	0.93431	0.63198	3.8216%	0.88091	0.87333
F-K	SSA-HKELM	1.02890	0.71577	4.2977%	0.88348	0.84639
F-4	WOA-HKELM	0.28456	0.25234	1.3347%	0.98987	0.98825
F-5	SSA-HKELM	0.37306	0.26626	1.3719%	0.99184	0.97980
F-5	SMA-HKELM	0.25932	0.23025	1.2184%	0.99194	0.99024
F-5	WOA-HKELM	0.22865	0.20306	1.0595%	0.99395	0.99241
F-11	SSA-HKELM	0.37306	0.26626	1.3719%	0.99184	0.97980
F-11	MFO-HKELM	0.31620	0.24417	1.2260%	0.99592	0.98549
F-11	WOA-HKELM	0.31637	0.24134	1.2068%	0.99594	0.98548
F-17	MFO-HKELM	0.31620	0.24417	1.2260%	0.99592	0.98549
F-17	WOA-HKELM	0.31637	0.24134	1.2068%	0.99594	0.98548

The comparison shows that for different parameter sets, the WOA-HKELM prediction algorithm achieves the best performance. The result of the optimal fusion prediction model is shown in Figure 7.

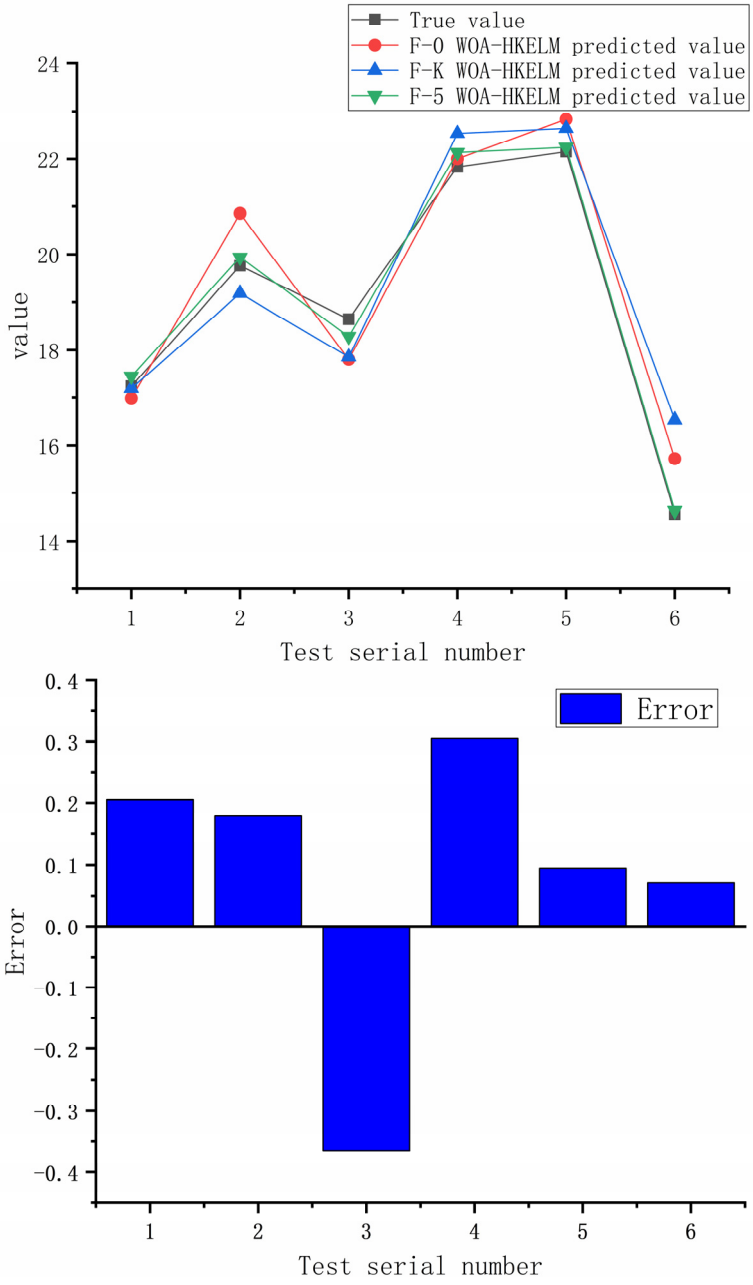


Figure 7. Prediction performance of the best fusion model.

5. Conclusions

(1) Based on the full subset regression, 18 characteristic parameter sets, including 12 influencing factors were constructed. The parameter sets considered the correlation among characteristic parameters and the influence of the parameter on gas emission. Kernel principal component analysis (KPCA) was used to reduce the dimensionality of the data. Five factors with a cumulative contribution rate of 85.04% were expressed as F-K, and a total of 20 parameter sets were constructed based on the original data, 18 of which were selected to predict gas emission.

(2) SSA, GA, PSO, WOA, MFO, and SMA were employed to optimize the Hybrid Kernel Extreme Learning Machine (HKELM) and the Least Squares Support Vector Machine

(LSSVM), as well as input variables. Results showed that the HKELM outperformed the LSSVM in prediction accuracy, stability, and running speed.

(3) The optimal fusion prediction model was the integration of F-5 and the WOA-HKELM. The evaluation index of each model performed well, with the root mean square error (RMSE) being 0.22865, the determination coefficient (R^2) being 0.99395, the Nash coefficient (NSE) being 0.99241, the mean absolute error (MAPE) being 0.20306, and the mean absolute percentage error (MAPE) being 1.0595%. Results showed that the model proposed by this paper was better than the original index system and the one undergone KPCA dimensionality reduction by a large margin.

(4) This paper selected characteristic parameter sets through full subset regression and applied different parameter sets to different algorithms to predict the amount of gas emission. It reduced the complexity of the prediction process caused by numerous influencing factors and their complex correlations and improved the compatibility between the parameter sets and the prediction algorithm. This model can be widely applied, generating much better prediction results. It is more practical and easier to operate than models in previous research.

Author Contributions: A Gas Emission Prediction Model Based on Feature Selection and Improved Machine Learning: Use full subset regression and machine learning algorithm, software, Matlab; data curation, K.Z.; writing—original draft preparation, K.Z.; supervision, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation Project (71771111). Project name: Research on prediction method and application of coal and gas outburst based on big data. Discipline classification: G0104. Prediction and evaluation. Project leader: Liangshan Shao. Funding amount: 460,000 yuan.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Research data has been presented in the article.

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

References

1. Qin, Y.J.; Su, W.W.; Jiang, W.Z.; Chen, Y.P. Mine gas emission prediction technology in China. Research progress and development direction of coal mine safety, ploidy of 2020. 52–59. [CrossRef]
2. Xie, H.P.; Zhou, H.W.; Xu, D.J.; Wang, H.W.; Zhang, R.; Gao, F. Research and consideration on deep coal mining and critical mining depth. *J. China Coal Soc.* **2012**, *37*, 535–542. [CrossRef]
3. Chen, L.; Liu, Q. Weighted grey target decision method to predict coal and gas outburst dangerous study. *J. Saf. Environ. Eng.* **2021**, *28*, 61–66. [CrossRef]
4. Qin, Y.; Su, W.; Jiang, W.; Chen, Y. Research progress and development direction of mine gas emission prediction technology in China. *Coal Mine Saf.* **2020**, *51*, 52–59. [CrossRef]
5. Yang, C.; Zhang, J.; Huang, Z. Numerical study on cavitation-vortex-noise correlation mechanism and dynamic mode decomposition of a hydrofoil. *Phys. Fluids* **2022**, *34*, 125105. [CrossRef]
6. Guo, B.; Wang, Y.; Zhou, H.; Hu, F. Can environmental tax reform promote carbon abatement of resource-based cities? Evidence from a quasi-natural experiment in China. *Environ. Sci. Pollut. Res.* **2022**, 1–13. [CrossRef]
7. Lu, S.; Guo, J.; Liu, S.; Yang, B.; Liu, M.; Yin, L.; Zheng, W. An improved algorithm of drift compensation for olfactory sensors. *Appl. Sci.* **2022**, *12*, 9529. [CrossRef]
8. Qin, X.; Liu, Z.; Liu, Y.; Liu, S.; Yang, B.; Yin, L.; Liu, M.; Zheng, W. User OCEAN personality model construction method using a bp neural network. *Electronics* **2022**, *11*, 3022. [CrossRef]
9. Wang, Y.; Zhan, G.; Fu, H.; Wang, S. Gas emission prediction based on variational mode decomposition and deep integrated combination model. *Control Eng.* **2022**, 1–12. [CrossRef]
10. Dai, W.; Fu, H.; Ji, C.P.; Wang, Y.J. Prediction method of VMD-DE-RVM interval for gas emission in mining face. *China Saf. Sci. J.* **2018**, *28*, 109–115. [CrossRef]
11. Li, B.; Zhang, C.H.; Li, X.J.; Wang, X.F. Prediction of mine gas emission based on PCA-ELM. *World Sci. Res. Dev.* **2016**, *38*, 49–53. (In Chinese) [CrossRef]
12. Xiao, P.; Xie, X.; Shuang, H.; Liu, C.; Wang, H.; Xu, J. Prediction of gas emission based on KPCA-CMGANN algorithm. *China Saf. Sci. J.* **2020**, *30*, 39–47. [CrossRef]

13. Wen, T.; Sun, X.; Kong, X.; Tian, H. Sub-source prediction of gas emission based on PSOBP-AdaBoost model. *China Saf. Sci. J.* **2016**, *26*, 94–98. [CrossRef]
14. Chen, W.H.; Yan, X.H.; Fu, H. Application of improved Elman neural network in gas emission prediction. *J. Saf. Environ.* **2015**, *15*, 19–24. [CrossRef]
15. Peng, X.H.; Liu, L.Q. Wavelet packet neural network prediction method in the application of gas emission. *J. Microelectron. Comput.* **2016**, *114*, 129–133. [CrossRef]
16. Xu, Y.; Qi, C.; Feng, S. Prediction model of gas emission based on IGSA-BP network. *J. Electron. Meas. Instrum.* **2019**, *33*, 111–117. [CrossRef]
17. Ma, S.; Li, X. Improved BP neural network prediction model for coal mine gas emission. *Min. Res. Dev.* **2019**, *39*, 138–142. [CrossRef]
18. Wang, Y.; Li, Y.; Han, Q.; Li, Y.; Zhou, C. Prediction of gas emission in mining face based on PCA-BO-XGBoost. *J. Xi'an Univ. Sci. Technol.* **2022**, *42*, 371–379. [CrossRef]
19. Chen, Q.; Huang, L. Prediction of gas emission in Mining face based on LASSO-LARS. *Coal Sci. Technol.* **2022**, *50*, 171–176. [CrossRef]
20. Lin, H.; Zhou, J.; Gao, F.; Jin, H.; Yang, Z.; Liu, S. Coal seam gas content prediction based on feature selection and machine learning fusion. *Coal Sci. Technol.* **2021**, *49*, 44–51. [CrossRef]
21. Chen, J.; Wang, S.; Liu, X.; Zheng, S.; Wang, G.; Sun, L. Study on influencing factors of gas emission from horizontal sublevel mining face in steeply inclined extra-thick coal seam. *Coal Sci. Technol.* **2022**, *50*, 127–135. [CrossRef]
22. Xiong, Y.; Cheng, J.; Duan, Z. Gas emission prediction model of coal mine based on CSBP algorithm. *ITM Web Conf.* **2016**, *7*, 09006. [CrossRef]
23. Cao, L.Y.; Fan, Q.Q.; Huang, J.Y. Intraoperative hypothermia prediction based on feature selection and XGBoost optimization. *Data Acquis. Process.* **2022**, *37*, 134–146. [CrossRef]
24. Wang, L.; Liu, Y.; Liu, Z.; Qi, J. Research on gas emission prediction model based on IABC-LSSVM. *Sens. Microsyst.* **2022**, *41*, 34–38. [CrossRef]
25. Shi, Y.S.; Li, J.; Ren, J.R.; Zhang, K. Prediction of residual service life of lithiumion batteries based on WOA-XGBoost. *Energy Storage Sci. Technol.* **2022**, *11*, 3354–3363. (In Chinese) [CrossRef]
26. Jia, J.; Ke, D.; Chen, Y. Prediction of coal mine gas emission based on orthogonal test-multiple regression. *J. Saf. Environ.* **2021**, *21*, 2037–2044. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.