Prediction of coal mine gas emission based on hybrid machine learning model

Shenghao Bi
Liaoning Technical University
Liangshan Shao
Liaoning Technical University
Zihan Qi
Liaoning Technical University
Yanbin Wang
Liaoning Technical University
Wenzhe Lai (472010072@stu.lntu.edu.cn)
Liaoning Technical University

Research Article

Keywords: Gas emission prediction, Hybrid machine learning, Random forest, Improved grey wolf optimiza, Support vector regression, Nonlinear convergence, DLH search strategy

Posted Date: September 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2080112/v1

Prediction of coal mine gas emission based on hybrid machine learning model

Shenghao Bi[1], Liangshan Shao[1], Zihan Qi[1], Yanbin Wang[2], Wenzhe Lai[2]

1. Institute of Systems Engineering, Liaoning Technical University, Huludao Liaoning, 125105, China

2. School of Business Administration, Liaoning Technical University, Huludao Liaoning, 125105, China

Shenghao Bi: 472021231@stu.lntu.edu.cn

Liangshan Shao: lntushao@163.com

Zihan Qi: 18147663263@163.com

Yanbin Wang: wyb_2000@163.com

Wenzhe Lai: 472010072@stu.lntu.edu.cn

Corresponding author: Wenzhe Lai Orcid: 0000-0002-1000-5599

Abstract: Coal mine gas accident is one of the most significant threats to the safe mining process in coal mines, so it is very important to accurately predict coal mine gas emission. To improve the accuracy of coal mine gas emission prediction, a hybrid machine learning prediction model combining random forest (RF) algorithm, improved gray wolf optimizer (IGWO) algorithm and support vector regression (SVR) algorithm is proposed. Thirty groups of actual measured gas emission data from a coal mine are selected as samples, and the latter five groups are used as test sets. Firstly, the RF algorithm is used to screen 13 influencing factors of coal mine gas emission, and finally six influencing factors are selected as the input variables of the prediction model; Secondly, the IGWO algorithm is obtained by improving the GWO algorithm with a nonlinear convergence factor and a DLH search strategy; Finally, the IGWO algorithm is used to optimize the SVR algorithm to establish the RF-IGWO-SVR model, and the model is compared with other models to verify its superiority. The results show that the average relative error of the RF-GWO-SVR model is 1.55%, and this result is better than other comparative models, which indicates that the model can effectively improve the prediction accuracy of coal mine gas emission and provide a new model for coal mine gas emission prediction.

Keywords: Gas emission prediction; Hybrid machine learning; Random forest; Improved grey wolf optimiza; Support vector regression; Nonlinear convergence; DLH search strategy

1. Introduction

Coal is the main consumed energy source in China and has played an important role in the past economic development [1]. Coal consumption in China grew at a rate of 4.6% in 2021, with a share of 56.0% in total energy consumption. The process of coal mining is often accompanied by gas emis-sions, a phenomenon that affects the safe production of coal mines and the personal health of mine workers and is a major safety hazard in coal production [2-4]. As the depth of coal mining in China increases, geological conditions are becoming more complex, leading to a gradual increase in gas emissions, which in turn leads to frequent coal mine gas accidents [5]. The occurrence of coal mine gas accidents is mainly due to abnormal gas emissions, therefore, the accurate prediction of gas emis-sions is of great importance for coal mine safety production.

In recent years, many scholars have proposed many gas emission prediction models, such as the SVR model[6], PCA-ELM model[7], BP neural network model[8], and SAPSO-ELM model [9], PCA-GA-BP model[10], etc. Although using the above prediction models in the gas emission prediction improves the accuracy, the ELM algorithm, and BP algorithm are not suitable for small sample prediction problems such as gas emission, as they are easy to cause inadequate learning phenomenon, the accuracy, and stability of gas emission prediction models need to be further improved. The SVR algorithm is suitable for small sample data prediction[11] and for solving the gas emission prediction problem, but the penalty factor and kernel function parameters in the algorithm need to be selected reasonably, otherwise, it will affect the model prediction effect[12]. The grey wolf optimiza (GWO) algorithm has the advantages of simple structure, few setting parameters, and strong global search capability[13,14]. The GWO algorithm can be used to find the optimum of the SVR algorithm parameters, but it may fall into the local optimum while finding the optimum, and it needs to be improved. Numerous influencing factors affect the amount of gas emission from coal mines, and the high latitude nature of the influencing factors affects the accuracy of the prediction of gas emission to some extent. Principal component analysis (PCA)[7] is now commonly used to re-dimension the influencing factors and thus improve the overall accuracy, but the data processed through the PCA algorithm is ambiguous and not conducive to the later analysis of the factors. If the RF algorithm is used to filter the influencing factors to achieve dimensionality reduction, it will not only be more effective than the PCA algorithm but also retain the original data of the influencing factors and facilitate later analysis of individual influencing factors.

In view of this, this study proposes a hybrid machine learning model combining random forest (RF), improved grey wolf optimiza (IGWO), and support vector regression (SVR). The contribution of this study is as follows: Firstly, the RF algorithm filters out the main factors influencing the amount of gas emission from coal mines, which reduces the computational work of the prediction model and improves the prediction accuracy. Secondly, the nonlinear convergence factor and DLH search strategy are used to optimize the GWO algorithm to avoid the algorithm from falling into the local optimum. The nonlinear convergence factor can improve the later searchability of the population, and the DLH search strategy can enrich the diversity of the population. Finally, the parameters of the SVR algorithm are optimized using the IGWO algorithm, which will improve the stability and prediction accuracy of the prediction model.

The remainder of this study is as follows: The second part introduces the basic theory of algorithms and the process of building RF-IGWO-SVR model. The third part examines the accuracy of the model and compares it with other models to check its superiority. The last part gives the conclusion.

2. Methodology

2.1. Random forests (RF)

A combinatorial classifier technique called random forest consists of many decision tree models[15-16]. The main goal of its influencing factors selection is to first determine the value that each influence factor contributed to each decision tree. This value is quantified by the out-of-bag data error, where out-of-bag data refers to the data that is not always selected when decision trees are formed. These are the precise steps:

(1) Suppose the initialization sets the number of decision trees to N. For each decision tree, the corresponding out-of-bag data is selected, and its data error is calculated, with the error for each decision tree noted as Err_{OOB1}, Err_{OOB2}, ..., Err_{OOBN}.

(2) Add noise interference to each influence M for out-of-bag data, recalculate the out-of-bag data error, and note as Err_{OOBM1}, Err_{OOBM2}, ..., Err_{OOBMN}.

(3) Equation (1) is used to find the importance I_M of each influencing factor and the influencing factors are ranked by their importance magnitude.

$$I_M = \frac{1}{N}\sum_{i=1}^{N}\left(Err_{OOBM} - Err_{OOB}\right) \tag{1}$$

2.2. Support vector regression (SVR)

Support vector regression (SVR) is a support vector machine (SVM) regression prediction model [17]. The fundamental idea is to execute a non-linear mapping of the predicted data before solving the regression issue in the high dimensional space. This transfers the data from low latitude space to high dimensional space[18-19]. The SVR algorithm is widely used in various fields of forecasting and is suitable for predicting small sample data. The SVR algorithm steps are as follows：

Suppose the training sample set is $S = \{(x_i, y_i)\}, i = 1,2,\cdots,m$. Among them， x_i is the input vector and y_i is the output vector, the decision function expression of the SVR model is as follows.

$$F(x) = \omega^T\varphi(x) + b \tag{2}$$

In Equation (2), ω^T is the weighting factor, $\varphi(x)$ is a nonlinear mapping function and b is the bias amount.

The process of solving the decision function can be seen as solving the minimization process， the expressions are as follows.

$$\min Q(\omega) = \frac{1}{2}|\omega|^2 + c\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) \tag{3}$$

$$\omega^T\varphi(x) + b - y_i \le \varepsilon + \xi_i \qquad \xi_i \ge 0 \tag{4}$$

$$\omega^T\varphi(x) + b - y_i \ge \varepsilon + \xi_i^* \qquad \xi_i^* \ge 0 \tag{5}$$

Among them, c is the penalty factor, ξ_i and ξ_i^* is the slack variable and ε is an insensitive parameter.

Applying Lagrange's equation and pairwise theory, expression (3) can be converted into a pairwise problem for the SVR algorithm， the expressions are as follows.

$$\begin{aligned}\max Q(\alpha) = {} & \frac{1}{2}\sum_{i,j=1}^{m}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)K\left(x_i, x_j\right) + \varepsilon\sum_{i=1}^{m}\left(\alpha_i^* + \alpha_i\right) \\ & -\sum_{i=1}^{m} y_i\left(\alpha_i - \alpha_i^*\right)\end{aligned} \tag{6}$$

$$\sum_{i=1}^{m}\left(\alpha_i^* - \alpha_i\right) = 0 \qquad \alpha_i^*, \alpha_i \in [0, c] \tag{7}$$

$$K\left(x_i, x_j\right) = \varphi\left(x_i\right)\varphi\left(x_j\right) \tag{8}$$

In Equation (6)-(8)， α_i、α_i^*、α_j and α_j^* is the Lagrangian equation multiplier and $K(x_i, x_j)$ is the kernel function. The Radical Basis Function (RBF) kernel function is used as the kernel function of the SVR algorithm. The RBF kernel function is a typical local kernel function with excellent local interpolation capability, which is conducive to improving the computational power of the algorithm. The RBF kernel function is expressed as follows.

$$K\left(x_i, x_j\right) = \exp\frac{-\left\|x_i - x_j\right\|^2}{g^2} \tag{9}$$

In Equation (9)， g is the kernel function parameter. The decision function of the SVR algorithm is transformed as follows.

$$F\left(x\right) = \sum_{i=1}^{m}\left(\alpha_i^* - \alpha_i\right)K\left(x_i, x_j\right) + b \tag{10}$$

The values of the penalty factor c and the kernel function parameter g of the SVR algorithm are chosen to directly affect the overall performance of the algorithm, so the two parameters of the penalty factor c and the kernel function parameter g need to be optimized for optimization.

2.3. Grey wolf optimiza (GWO)

The grey wolf optimiza (GWO) algorithm is an algorithm proposed by Mirjalili in 2014 [20]. The grey wolf algorithm classifies wolves in a pack into a total of four classes, from highest to lowest α、β、δ and ω. Its hunting process is mainly divided into tracks, encircling, and hunting[21-22].

The formula for the tracking process is as follows:

$$D = \left|C \cdot X_P\left(t\right) - X\left(t\right)\right| \tag{11}$$

$$X\left(t+1\right) = X_p\left(t\right) - A \cdot D \tag{12}$$

$$A = 2ar_1 - a \qquad C = 2r_2 \tag{13}$$

$$a = 2 - t/Maxiter \tag{14}$$

In Equation (11)-(14), t represents the tth iteration, $X_p(t)$ and $X(t)$represent the tth iteration prey position and gray wolf position, A and C are the convergence factor and the coefficient constants, $Maxiter$ is the maximum number of iterations, r_1 and r_2 is a random number belonging from 0 to 1.

The encircling process formula is expressed as follows.

$$\begin{cases} D_\alpha = C_1 \cdot X_\alpha\left(t\right) - X\left(t\right) \\ D_\beta = C_2 \cdot X_\beta\left(t\right) - X\left(t\right) \\ D_\delta = C_3 \cdot X_\delta\left(t\right) - X\left(t\right) \end{cases} \tag{15}$$

In Equation (15), D_α、D_β and D_δ denote the distances between wolves α、β、δ and wolf ω, respectively.$X_\alpha(t)$、$X_\beta(t)$、$X_\delta(t)$ represent the location of wolves α、β and δ, $X(t)$ is the location of wolves.

The hunting process formula is expressed as follows.

$$\begin{cases} X_1 = X_\alpha\left(t\right) - A_1 \cdot D_\alpha \\ X_2 = X_\beta\left(t\right) - A_2 \cdot D_\beta \\ X_3 = X_\delta\left(t\right) - A_3 \cdot D_\delta \end{cases} \tag{16}$$

$$X\left(t+1\right) = \frac{X_1 + X_2 + X_3}{3} \tag{17}$$

2.4. Improve grey wolf optimiza (IGWO)

2.4.1. Nonlinear convergence

In the original GWO algorithm, The convergence factor a decreases linearly in the interval [0, 2], this will affect the local search capability of the model later. After researching, transforming the convergence factor a from linear convergence to nonlinear convergence helps to improve the overall merit-seeking ability of the GWO algorithm. A cosine function is used to convert the convergence factor a to a nonlinear convergence factor, and the specific mathematical expression is as follows.

$$a = 2\cos\left(\frac{\pi}{2}\cdot\frac{t}{Maxiter}\right) \tag{18}$$

By using a maximum iteration value of 100, the nonlinear convergence factor is compared to the original convergence factor, and the detailed findings are displayed in Figure 1.

According to Figure 1, the value of the improved nonlinear convergence factor has been higher than the value of the original convergence factor throughout the iterative process, making it more advantageous to broaden the population's search space and prevent it from settling on local optimal solutions.

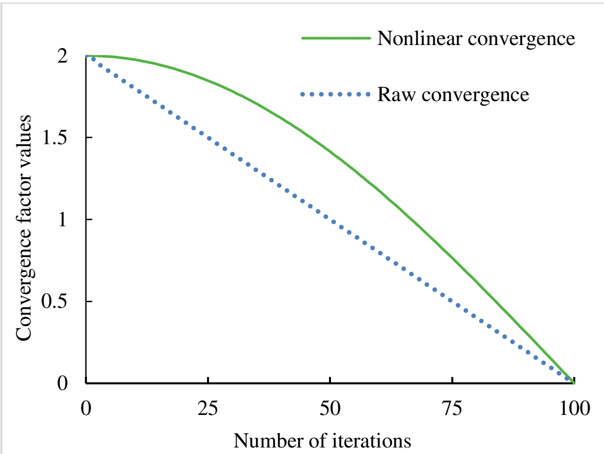


Fig.1 Comparison of convergence factors

2.4.2. DLH search strategy

The DLH search strategy is an improved idea proposed by Nadimi-Shahraki in 2020 [23]. The original GWO algorithm leads to a reduction in population diversity at a later stage and is prone to fall into local optimal solutions. The introduction of the DLH search strategy in the population increases the diversity of the population and improves the optimal solution-finding ability.

The expression for the gray wolf population location update using the DLH search strategy is as follows.

$$X_{DLH}(t+1) = X(t) + rand\left[X_n(t) - X_r(t)\right] \tag{19}$$

In Equation (19), $X_{DLH}(t+1)$ is the new result generated after the wolf adopted the DLH search strategy,$X_r(t)$ is a random selection from α、β and δ wolves, $X_n(t)$ is a randomly selected proximity wolf fromN_t .The mathematical equation of N_t is expressed as.

$$\begin{cases} N(t) = \left\{ X_j(t) \middle| D\left(X(t), X_j(t) \right) \le R(t) \right\} \\ R(t) = \left\| X(t) - X_{GWO}(t+1) \right\| \end{cases} \tag{20}$$

In Equation (20), $X_{j(t)} \in \alpha, \beta, \delta$, D is the distance between $X(t)$ and $X_j(t)$, $X_{GWO}(t+1)$is the location update of the original GWO. Finally, comparing $X_{GWO}(t+1)$ and $X_{DLH}(t+1)$, and the best one are selected to perform the location update.

2.5. RF-IGWO-SVR model building steps

Step 1: The RF algorithm is used to find the importance of each influencing factor and to screen for influencing factors.

Step 2: Set the initial parameters of the Grey Wolf Optimisation (GWO) algorithm, the number of grey wolves, and the maximum number of iterations.

Step 3: Find the fitness value of each wolf in the pack and classify the pack into four types:α、β、δ and ω.

Step 4: Update the convergence factor a according to equation (18) and update the value of A and C according to equation (13).

Step 5: Prey encirclement and hunting processes in grey wolf populations according to equations (6) and (7).

Step 6: Grey wolf population location updates were carried out according to equation (17).

Step 7: The update of grey wolf population locations using the DLH search strategy was carried out according to equation (19) and compared with the grey wolf population locations obtained by process step 5 to retain the optimal locations.

Step 8: Processes step 4 to step 7 repeated until the maximum number of iterations is reached, and the run is stopped to find the optimal position of the grey wolf population. The coordinate components of the optimal position correspond to the values of the penalty factor c and the kernel function parameter g. The values of c and g are brought into the SVR algorithm to construct the RF-IGWO-SVR model.

In conclusion, the RF algorithm is first modified to include the gas outflow influencing variables for the selection of influencing factors; Second, the IGWO algorithm is employed to optimize the SVR algorithm while the GWO method is improved. Figure 2 illustrates the process of the data is divided into the training set and test set substituted into the optimized SVR algorithm.

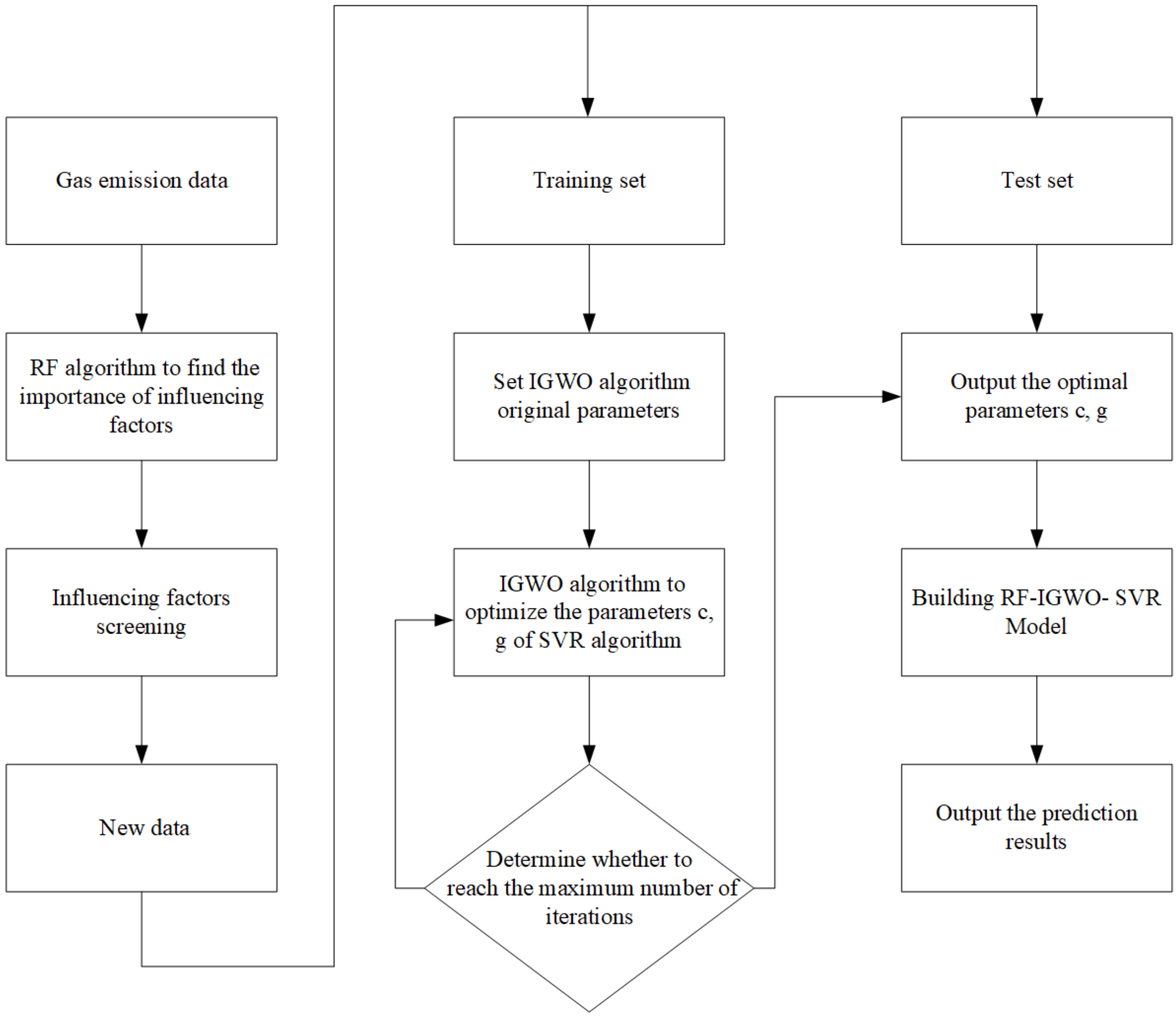


Fig.2 RF-IGWO-SVR model process

3. Model application and analysis

3.1. Data source

A coal mine in Shenyang, China, is selected for the analysis and prediction of 30 sets of historical gas emission detection data. The first 25 sets of data are used as the training set, and the last 5 sets of data are used to test the accuracy of the prediction model, and the relevant parameters are shown in Table 1[24]. Among them, the influencing factors of gas emission include coal seam depth X_1 (m), coal seam thickness X_2 (m), coal seam dip angle X_3 (°), the original gas content of mining seam X_4 ($m^3 \cdot t^{-1}$), coal seam spacing X_5 (m) , mining height X_6 (m), adjacent layer gas content X_7 ($m^3 \cdot t^{-1}$), adjacent layer thickness X8 (m), interlayer lithology X9, working face length X_{10} (m), advancing speed X_{11} ($m \cdot d^{-1}$), extraction rate X_{12} (%), daily production X_{13} ($m \cdot d^{-1}$), gas emission expressed as Y ($m^3 \cdot min^{-1}$). The average relative error is used to measure the prediction result of gas emission, and its mathematical formula is as follows.

$$M = \frac{100\%}{N} \sum_{n=1}^{N} \frac{\left| y_n - y_n^* \right|}{y_n} \tag{21}$$

In Equation (21), N is the total number of test sets, y_n is the actual value of gas emission, y_n^* is the predicted value of gas emission.

Table 1 Statistics of influencing factors for gas emission

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	Y
1	412	2.5	8	2.12	24	2	2.1	1.53	4.78	140	4.16	0.96	1528	2.91
2	423	1.5	11	2.14	17	1.4	2.55	1.62	4.75	180	4.14	0.95	1751	3.52
3	436	2.3	10	2.53	14	2.2	2.4	1.48	4.91	145	4.67	0.945	2074	3.62
4	459	2.4	15	2.45	24	2.3	2.42	1.78	4.75	155	4.57	0.944	2104	4.13
5	511	2.8	13	3.24	14	2.4	2.21	1.72	4.78	180	3.45	0.93	2241	4.6
6	515	2.3	17	2.85	17	2.5	2.77	1.87	4.51	170	3.25	0.94	1973	4.94
7	556	2.7	9	3.37	13	2.5	1.88	1.42	4.85	165	3.68	0.932	2287	4.78
8	550	3.1	12	3.67	15	2.9	2.32	1.65	4.83	155	4.01	0.92	2352	5.25
9	590	3	11	3.68	12	3.6	3.11	1.46	4.53	175	3.53	0.94	2410	5.56
10	581	5.2	8	4.31	17	5.9	3.47	1.57	4.76	170	2.8	0.797	3131	7.26
11	611	6.7	9	4.05	16	6.7	3.15	1.8	4.7	175	2.64	0.812	3354	7.8
12	408	2	10	1.92	20	2	2.02	1.5	5.03	155	4.42	0.96	1825	3.34
13	411	2	8	2.15	22	2	2.1	1.21	4.87	140	4.16	0.95	1527	2.97
14	420	1.8	11	2.14	19	1.8	2.64	1.62	4.75	175	4.13	0.95	1751	3.56
15	432	2.3	10	2.58	17	2.3	2.4	1.48	4.91	145	4.67	0.95	2078	3.62
16	456	2.2	15	2.4	20	2.2	2.55	1.75	4.63	160	4.51	0.94	2104	4.17
17	516	2.8	13	3.22	12	2.8	2.21	1.72	4.78	180	3.45	0.93	2242	4.6
18	527	2.5	17	2.8	11	2.5	2.81	1.81	4.51	180	3.28	0.94	1979	4.92
19	531	2.9	9	3.35	13	2.9	1.88	1.42	4.82	165	3.68	0.93	2288	4.78
20	550	2.9	12	3.61	14	2.9	2.12	1.6	4.83	155	4.02	0.92	2352	5.23
21	563	3	11	3.68	12	3	3.11	1.46	4.53	175	3.53	0.94	2410	5.56
22	590	5.9	8	4.21	18	5.9	3.4	1.5	4.77	170	2.85	0.795	3139	7.24
23	604	6.2	9	4.03	16	6.2	3.15	1.8	4.7	180	2.64	0.812	3354	7.8
24	607	6.1	9	4.34	17	6.1	3.02	1.74	4.62	165	2.77	0.785	3087	7.68

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	Y
25	634	6.5	12	4.8	15	6.5	2.98	1.92	4.55	175	2.92	0.773	3620	8.51
26	640	6.3	11	4.67	15	6.3	2.56	1.75	4.6	175	2.75	0.802	3412	7.95
27	450	2.2	12	2.43	16	2.2	2	1.7	4.84	160	4.32	0.95	1996	4.06
28	544	2.7	11	3.16	13	2.7	2.3	1.8	4.9	165	3.81	0.93	2207	4.92
29	629	6.4	13	4.62	19	6.4	3.35	1.61	4.63	170	2.8	0.803	3456	8.04
30	401	2	10	1.87	25	2.4	2.14	1.78	5.12	150	4.52	0.95	1855	3.38

3.2. RF algorithm influencing factors screening

The influencing factors of gas emission have high-dimensional complexity, which will affect the prediction accuracy. To improve the prediction accuracy of the model, it is necessary to screen the influencing factors of gas emission in advance. The RF algorithm is used to select the influencing factors and obtain the importance of the influencing factors. The detailed results of the importance ranking of the influencing factors are shown in Table 2.

Table 2 Ranking of influencing factors importance

Influencing factor	Factor importance	Influencing factor	Factor importance
X_{11}	0.3143	X_3	0.0094
X_{12}	0.2268	X_8	0.0037
X_6	0.1118	X_9	0.0030
X_4	0.1095	X_{10}	0.0021
X_2	0.0876	X_7	0.0019
X_1	0.0727	X_5	0.0015
X_{13}	0.0557		

From Table 2, the ranking of factor importance in descending order is X_{11}、X_{12}、X_6、X_4、X_2、X_1、X_{13}、X_3、X_8、X_9、X_{10}、X_7 and X_5, with the sum of factor importance of the 13 influencing factors being 1. To select the best combination of factors, 13 different sets of influencing factors are generated by increasing one by one from X_{11} to X_5 in order of the importance of the factor values, and the sets are substituted into the IGWO-SVR model in turn, with the average relative error as the measure, and the detailed results of the relationship between the number of factors and the average relative error are shown in Figure 3.

From Figure 3, the average relative error value drops to the lowest when the number of factors is 6. Therefore, the first 6 influence factors are selected as the combination of input influence factors for the prediction model, which is advancing speed X_{11}, extraction rate X_{12}, mining height X_6, the original gas content of mining seam X_4, and coal seam thickness X_2, and coal seam depth X_1. The sum of the cumulative factor importance of the 6 screened impact factors is 0.9227.

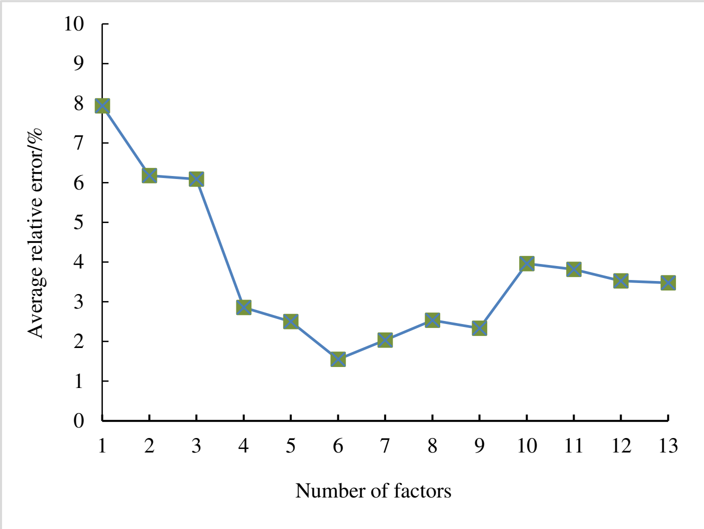
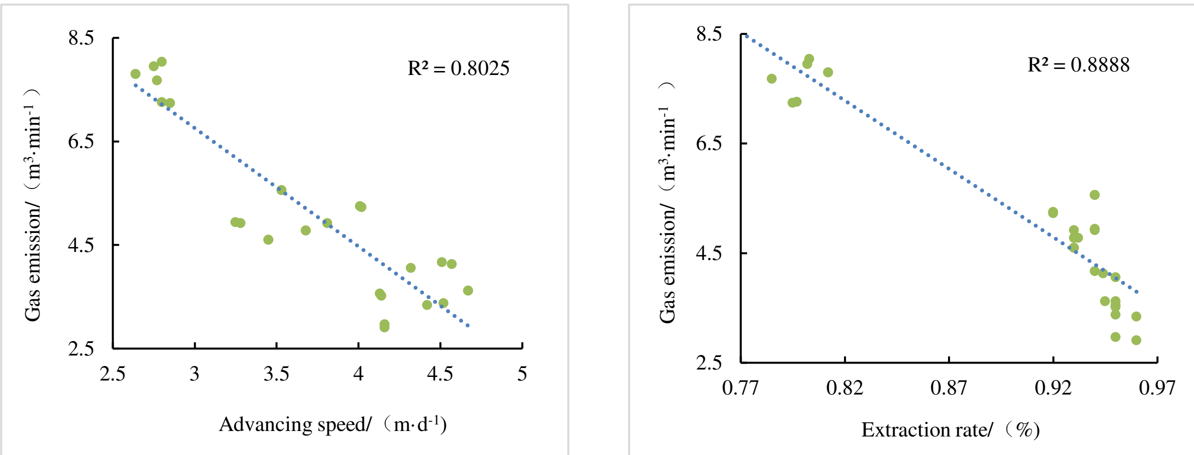


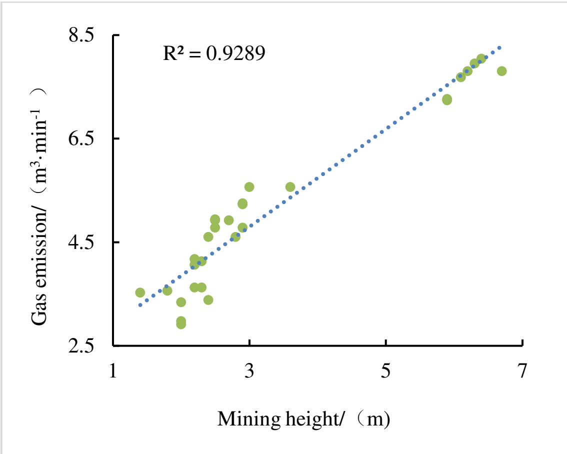
Fig.3 The relationship between the number of factors and the average relative error

To further explore the correlation between the six influencing factors screened for factors and gas emissions, the curve fitting toolbox cftool in MATLAB was used to analyze the relationship, and Figure 4 displays the complete outcomes.

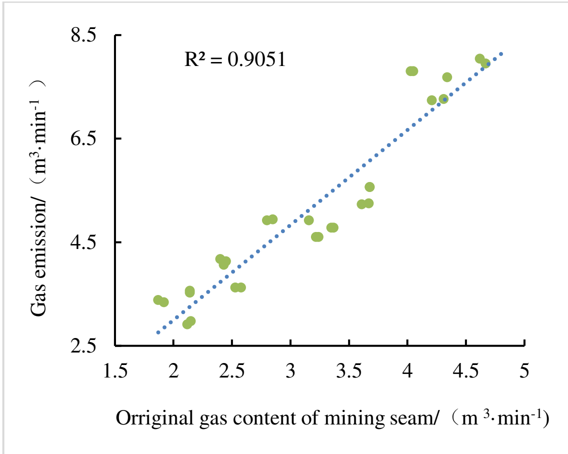
From Figure 4, the linear regression coefficients of determination (R^2) for advancing speed, extraction rate, mining height, the original gas content of the mined seam, coal seam depth, and coal seam thickness are 0.8025, 0.8888, 0.9289, 0.9051, 0.9164 and 0.8817 respectively.There is a significant association between the 6 influencing factors and gas emission, as shown by the R2 of the factors being larger than 0.8, among which are the mining height, the original gas content of the mining seam, coal seam depth, and coal seam thickness show a positive correlation with gas emission, and advancing speed and extraction rate show a negative correlation with gas emission.



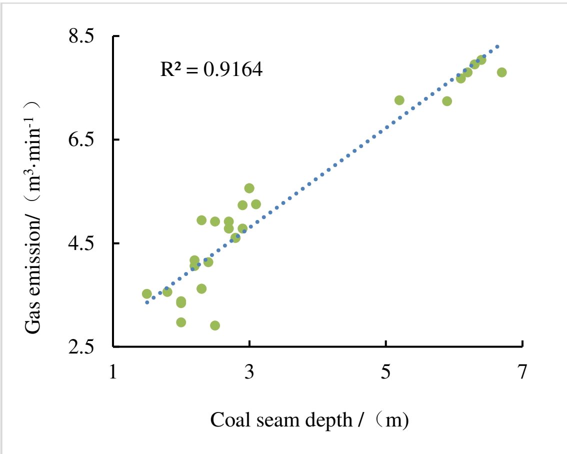
(a) Advancing speed (b) Extraction rate



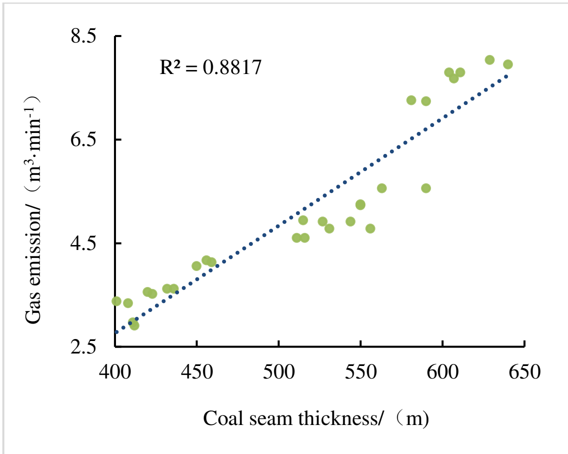
(c) mining height



(d) Original gas content of mining seam



(e) Coal seam depth



(f) Coal seam thickness

Fig.4 The relationship between influencing factors and gas emission

3.3. IGWO-SVR parameter selection

The initial setup parameters required for the IGWO-SVR model were the number of populations and the number of iterations. The set of influential elements filtered by the RF model was split into a training set and a test set for incorporation into the prediction model, with the first 25 data sets being the training set and the last 5 being the test set.The number of iterations is set to 50.

The number of wolves is set to increase from 2 to 50 one by one, and the average relative error is used as the measurement index, and the detailed results of the relationship between the number of wolves and the average relative error are shown in Fig. 5. When the number of wolves is from 2 to 35, the average relative error is still in an unstable state; when the number of wolves is from 36 to 50, the average relative error starts to maintain a stable state at the lowest point, so the number of wolves is set to 36 during initialization.

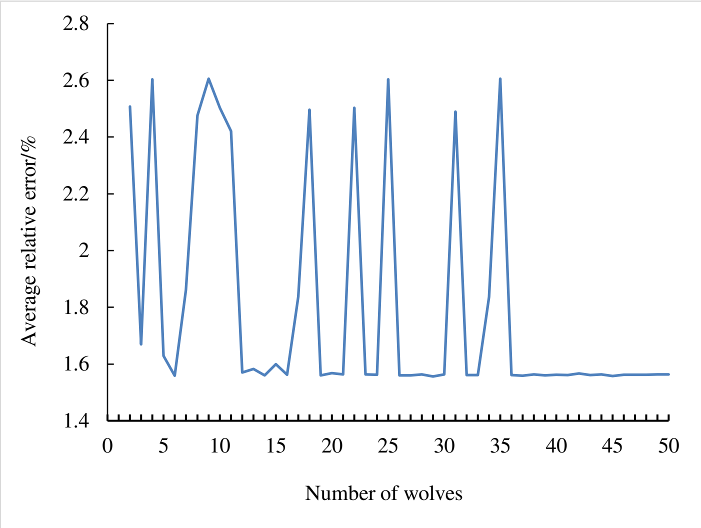


Fig.5 The relationship between the number of wolves and the average relative error

3.4. Comparison of predicted results

The influencing factor dataset screened by the RF algorithm was selected as the input variable, and the first 25 groups and the last 5 groups of the input variable were split into training and test sets, respectively. The RF-IGWO-SVR model was built with 36 wolves and 50 iterations, respectively.

The PCA-IGWO-SVR model and the IGWO-SVR model were built for comparison with the RF method to assess its efficacy in screening eigenvalues, while the PCA algorithm followed the process of the paper.[10]. Figure 6 illustrates a comparison of the predicted and actual values of gas emissions, and the results of the absolute error of the predicted values are shown in Figure 7.

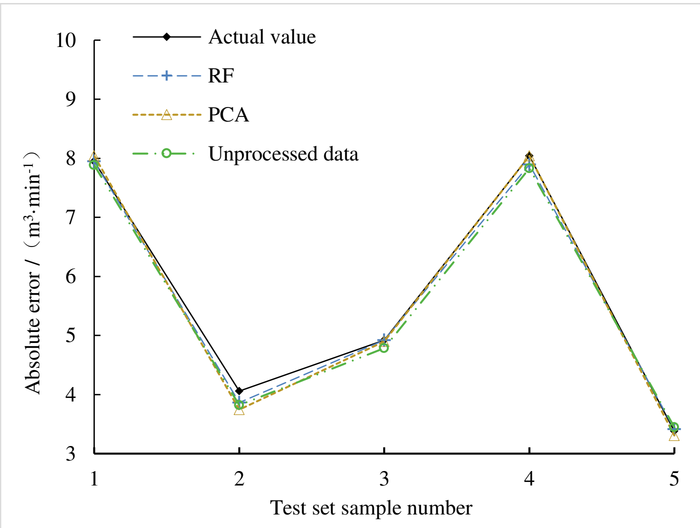


Fig.6 Comparison of predicted value and actual value

Figures 6 and 7 show that the RF-IGWO-SVR model has a smaller error than the PCA algorithm processed data for test set samples 1, 2, 3, and 5 test points, except for the 4th test sample where the error is higher than the PCA-IGWO-SVR model, and the overall stability of the RF-IGWO-SVR model is better than the PCA-IGWO-SVR model. The RF-IGWO-SVR model outperformed the IGWO-SVR model at all five test points when compared to the unprocessed raw data. In summary, the data set processed with the RF algorithm is more conducive to improving the accuracy and stability of the model.

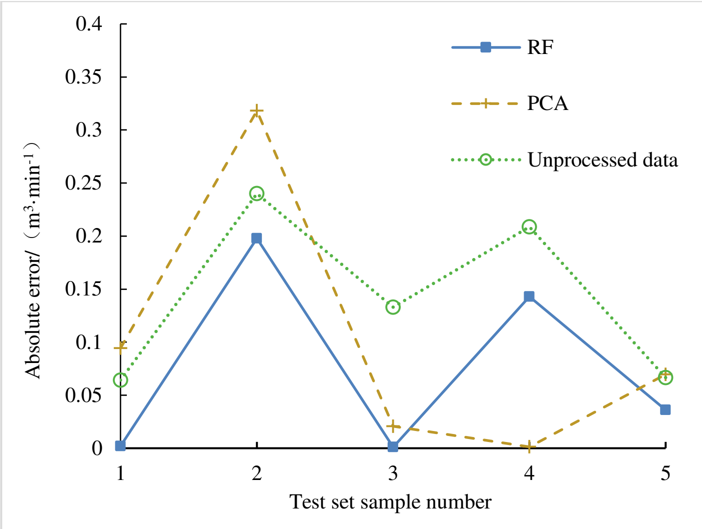


Fig.7 Absolute error of predicted value of gas emission

To further test the superiority of the predictive capability of the RF-IGWO-SVR model, the RF-GWO-SVR, RF-SVR, PCA-IGWO-SVR, PCA-GWO-SVR, PCA-SVR, GWO-SVR and SVR models were selected for comparison, where the initialization of the IGWO and GWO algorithms was extended using the prior initialization settings, and the specific outcomes of the comparison of the anticipated outcomes from the gas emission are displayed in Table 3.

Table 3 Comparison of prediction results of gas emission

No.		1	2	3	4	5	Average relative error/100%
Actual value		7.95	4.06	4.92	8.04	3.38	
RF	IGWO-SVR	7.9480	3.8621	4.9190	7.8972	3.4160	1.55
	GWO-SVR	7.9295	3.8583	4.9098	7.8796	3.4423	1.85
	SVR	8.0835	3.6355	4.9125	8.0774	3.5278	3.43
PCA	IGWO-SVR	8.0444	3.7419	4.8992	8.0413	3.3103	2.3
	GWO-SVR	8.0441	3.7425	4.8993	8.0411	3.3073	2.32
	SVR	8.0167	3.6913	4.8876	8.0330	3.6788	3.9
Unprocessed data	IGWO-SVR	7.8835	3.8181	4.7863	7.8324	3.4379	2.76
	GWO-SVR	7.8858	3.8201	4.7872	7.8314	3.4466	2.8
	SVR	7.5337	4.0211	4.8897	6.9489	4.4767	10.57

Table 3 shows that the RF-IGWO-SVR model has better predictive power than the other eight comparison models. Compared with those with PCA algorithm processed data as input variables, the average relative errors of the IGWO-SVR, GWO-SVR, and SVR models with RF algorithm data processing improved by 0.75%, 0.47%, and 0.47% respectively; compared with those with unprocessed influence characteristics as input variables, the average relative errors of the IGWO-SVR, GWO- SVR, and SVR models with the RF algorithm data processing improved the average relative errors by 1.21%, 0.95%, and 7.14% respectively, again demonstrating that the use of the RF algorithm to pre-process the data is beneficial to improving the overall accuracy of the models.

According to Table 3, the IGWO-SVR model performs better than the GWO-SVR model and the SVR model under the identical circumstances. When pre-processed with the RF algorithm, the IGWO-SVR model improved the average relative error by 0.29% and 1.88% over the GWO-SVR and SVR models respectively. When pre-processed with the PCA algorithm, the average relative error of the IGWO-SVR model improved by 0.02% and 1.6% compared to the GWO-SVR and SVR models respectively. When using the unprocessed data, the average relative error of the IGWO-SVR model improved by 0.04% and 7.81% compared to the GWO-SVR and SVR models respectively.

To make the results more intuitive, the average relative errors of the nine different models in Table 3 are plotted as bar charts, as shown in Figure 8, numbering the RF-GWO-SVR model to the SVR model as 1, 2, 3, 4, 5, 6, 7, 8 and 9 in that order.

Figure 8 allows to conclude more intuitively that the average relative error of the RF-GWO-SVR model is lower than that of the other comparison models and is 6.82 times higher than the average relative error of the unprocessed SVR model. According to the data pre-processing process, the histograms are divided into three groups, the first being 1, 2, and 3, the second being 4, ,5, and 6, and the third being 7, 8, and 9. Within each group, the average relative error values tend to increase from low to high, demonstrating that the prediction accuracy of the IGWO-SVR model is better than that of the GWO-SVR model and the unoptimized SVR model. The mean relative error values in the first group are significantly lower than those in the second and third groups, demonstrating the superiority of using the RF algorithm for data processing.

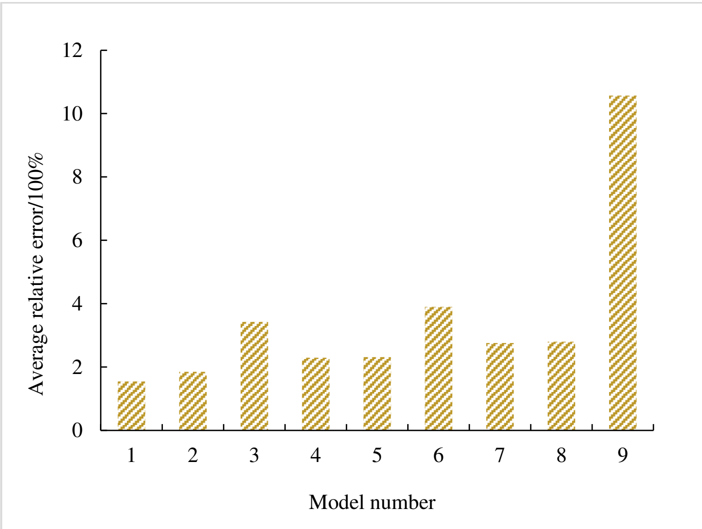


Fig.8 Comparison of average relative errors of prediction models

4. Conclusions

For the problem of gas emission prediction with small samples and high dimensional data characteristics, a novel hybrid machine learning prediction model (RF-IGWO-SVR model) is proposed in this study. In the prediction process, the RF algorithm is used to screen the influencing factors, which not only improves the prediction accuracy and stability of the model, but also retains the original information of the influencing factors. The nonlinear convergence factor and DLH search strategy are used to improve the GWO algorithm, which improves the overall search capability and avoids the model falling into local optimal solutions. The IGWO algorithm is used to optimize the penalty factor and kernel function parameters in SVR, which improves the prediction accuracy of the model. The experimental results show that the average relative error between the predicted and actual values of the RF-IGWO-SVR model is 1.55%, and this result is better than other comparative models, indicating that the RF-IGWO-SVR model can predict the actual situation of coal mine gas emissions more accurately.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

Author Contributions

Shenghao Bi: Conceptualization, methodology, data collection & Writing. Liangshan Shao: Review & supervision. Zihan Qi: Writing & language embellishment. Yanbin Wang: Language embellishment & review. Wenzhe Lai: Writing & review. All authors read and approved the fnal manuscript.

Funding Statement

Our manuscript is supported by National Natural Science Foundation of China (71771111) and Liaoning Provincial Department of Education Research Funding Project (General Project) (LJKZ0359).

References

[1] Zhang Y, Li J, Tian Y, Deng Y, and Xie K (2021) Virtual water flow associated with interprovincial coal transfer in China: Impacts and suggestions for mitigation, *Journal of Cleaner Production*, 289:125800.

[2] Liang C, Wang E, Feng J, Kong X, Zhang Z (2016) A dynamic gas emission prediction model at the heading face and its engineering application. *Journal of Natural Gas Science and Engineering*, 30, 228-236.

[3] Liu A, Liu S, Wang G, Elsworth D (2020) Predicting fugitive gas emissions from gob-to-face in longwall coal mines: coupled analytical and numerical modeling. *International Journal of Heat and Mass Transfer*, 150, 119392-.

[4] Liu C, Li S, Yang S (2018) Gas emission quantity prediction and drainage technology of steeply inclined and extremely thick coal seams. *International Journal of Mining Science and Technology*, 28(3), 415-422.

[5] Wei C, Xu M, Sun J, Xiang L, Ji C (2011) Coal mine gas emission gray dynamic prediction. *Procedia Engineering*, 26(none), 1157-1167.

[6] Li Y, Liu C (2013) Prediction of gas emission based on partial correlation analysis and svr. *Applied Mathematics & Information Sciences*, 7(5), 1671-1675.

[7] Bing L I, Zhang C, Xinjie L I, Wang X (2016) Prediction of mine gas emission based on pca-elm. *World Sci-Tech R & D*. 38 (1), pp.49-53.

[8] Guo X, Ren Z, Wang Q, Cui C (2019) Prediction of gas emission by bp neural network based on wavelet analysis. *IOP Conference Series: Earth and Environmental Science*, 252(5), 052046 (8pp).

[9] Ren H, Yan Y, Wu Q (2021) Different-source prediction of gas emission based on SAPSO-ELM and its application, *COAL GEOLOGY & EXPLORATION*, 49(02): 102-109.

[10] Qiu J W, Liu Z G, Zhou L, Qin R X (2017). Prediction model of gas quantity emitted from coal face based on pca-ga-bp neural network and its application. *Journal of Power Technologies*, 97(3):169-178.

[11] Yu P S, Chen S T, Chang I F (2006) Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, 328(3-4), 704-716.

[12] Roozbeh M, Babak M, Shahaboddin S, Kwok-Wing C (2018) Coupling a firefly algorithm with support vector regression to predict evaporation in northern iran. *Engineering Applications of Computational Fluid Mechanics*, 12(1), 584-597.

[13] Al-Betar M A, Awadallah M A, Faris H, Aljarah I, Hammouri A I (2018) Natural selection methods for grey wolf optimizer. *Expert Systems with Applications*, 113.

[14] Zhao X, Lv H, Lv S, Sang Y, Wei Y, Zhu X (2021) Enhancing robustness of monthly streamflow forecasting model using gated recurrent unit based on improved grey wolf optimizer. *Journal of Hydrology*, 601, 126607.

[15] Makungwe M, Chabala L M, Chishala B H, Lark R M (2021) Performance of linear mixed models and random forests for spatial prediction of soil ph. *Geoderma*, 397(4), 115079.

[16] Abu El-Magd S A, Ali S A, Pham Q B (2021) Spatial modeling and susceptibility zonation of landslides using random forest, nave bayes and k-nearest neighbor in a complicated terrain. *Earth Science Informatics*, 1-17.

[17] Cortes C, Vapnik V N (1995) Support vector networks. *Machine Learning*, 20(3), 273-297.

[18] Bakhtiar M S A, Zargar G, Riahi M A, Ansari H R (2020) Sand fraction prediction from seismic attributes using optimized support vector regression in an oil reservoir. *Earth Science Informatics*, 13(1).

[19] Dendi S, Channappayya S S (2020) No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Transactions on Image Processing*, PP(99), 1-1.

[20] Mirjalili S, Mirjalili S M, Lewis A D (2014) Grey wolf optimizer. *Advances in Engineering Software*, 69, 46-61.

[21] Ramadan A E, Kamel S, Khurshaid T, Oh S R, Rhee S B (2021) Parameter extraction of three diode solar photovoltaic model using improved grey wolf optimizer. *Sustainability*, 13.

[22] Adhikary J, Acharyya S (2022) Randomized Balanced Grey Wolf Optimizer (RBGWO) for solving real life optimization problems," *Applied Soft Computing*, vol. 117, 108429.

[23] Nadimi-Shahraki M H, Taghian S, Mirjalili S (2020) An improved grey wolf optimizer for solving engineering problems. Expert Systems with Applications, 166.

[24] Dong X, Jia J, Bai Y, Fan C (2016) Prediction for gas-gushing amount from the working face of stope based on the SVM coupling genetic algorithm," *Journal of Safety and Environmen*, 16(02): 114-118.