water

MDPI

Article

Identification of Mine Mixed Water Inrush Source Based on Genetic Algorithm and XGBoost Algorithm: A Case Study of Huangyuchuan Mine

Xiang Li [1], Donglin Dong [1], Kun Liu [2,*], Yi Zhao [1,*] and Minmin Li [2]

[1] College of Geoscience and Surveying Engineering, China University of Mining & Technology-Beijing, Beijing 100083, China; lx@student.cumtb.edu.cn (X.L.); ddl@cumtb.edu.cn (D.D.)
[2] Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China; liminmin@lasac.cn
* Correspondence: liukun@lasac.cn (K.L.); zy@student.cumtb.edu.cn (Y.Z.); Tel.: +86-10-6488-9221 (K.L.)

Abstract: Mine water inrush disaster seriously threatens the production of coal mine. Rapid and accurate identification of mine water inrush sources is a key premise for mine water disaster prevention. The conventional research on the identification of water inrush source has focused on a single source, and the identification of mixed water samples from multi-source aquifers in deep coal mining environment is not yet fully explored. In this study, absorption spectrum technology was introduced into the identification of water inrush sources. The absorption spectra of the water samples with different mixing ratios were prepared using the ultraviolet and visible spectrophotometry (UV–Vis) spectrophotometer. In addition, spectral data preprocessing such as scattering correction, baseline correction, smoothing and denoising, and data enhancement were conducted to reduce the influence of experimental error, environment, radiation, molecular interaction, and other factors on the spectral data. Furthermore, a genetic algorithm (GA) was used to improve the seven parameters of the extreme gradient boosting (XGBoost) algorithm, such as learning rate, base model selection, tree parameters, regularization parameters, and iteration times. The deep-learning classifier of mine mixed water sources based on GA-XGBoost was established and used to identify 66 groups of mixed water sources in the Huangyuchuan Mine. The simulation results show that spectral preprocessing and normalization enhancement effectively improved the accuracy of the discriminant model. After 100 cross-validations, the average recognition accuracy of the GA-XGBoost model was 94%, and the results were accurate and reliable. This study provides a new direction and method for the identification of water inrush sources, particularly for mixed water inrush sources. It may also serve as a technical reference for decision-makers to formulate effective coal mine water inrush prevention and control programs and for mine water disaster prevention in similar coalfields in North China.

Keywords: absorption spectrum; mixed water source; genetic algorithm; XGBoost algorithm

Citation: Li, X.; Dong, D.; Liu, K.; Zhao, Y.; Li, M. Identification of Mine Mixed Water Inrush Source Based on Genetic Algorithm and XGBoost Algorithm: A Case Study of Huangyuchuan Mine. *Water* **2022**, *14*, 2150. https://doi.org/10.3390/w14142150

Academic Editor: Maurizio Barbieri

Received: 3 June 2022
Accepted: 4 July 2022
Published: 6 July 2022

1. Introduction

China is poor in oil and gas resources, but it has rich coal resources that have become the cornerstone of its national economic development, and its dominant position in China's energy industry will be difficult to change in the near future [1]. However, with the deepening expansion of mining depths and scope, hydrogeological conditions in mines grow more complex and complicated, and the difficulties such as the surrounding rock stress and groundwater erosion caused by local exploitation increase in severity [2,3]. Sun et al. [4] showed that 1162 coal mine water accidents occurred in China from 2000 to 2015, resulting in 4676 deaths. Water disasters in coal mine accidents have become the second largest natural disaster after coal mine gas disaster, and they endanger human life as well as the coal mine and equipment [5–7].

Accurately locating the water sources during a water inrush crisis requires rapid, timely measurements to mitigate the disaster and ensure the safety of personnel and property [4,6,8]. At present, there are many significant research efforts focused on the identification of mine water inrush source, and the applied methods can be roughly categorized as hydrological analysis methods, mathematical theoretical analysis methods, and the artificial intelligence analysis method, among others [8–20]. By combining data from water chemical characteristics and physical characteristics, such as temperature and water level, researchers have been able to determine the water sources in mine water inrush disasters [21–27]. In recent years, more inroads have been made for increasing the speed and accuracy when identifying these sources by using a variety of mathematical model calculation methods based on the characteristics of water chemical physics, as well as other factors [28–32]. Based on hydrochemical data and self-organizing feature maps (SOM), Zhao et al. [28] analyzed the water inrush source of Ningtiao mine. Fisher discriminant analysis, water temperature, and traditional hydrogeochemical discrimination methods were employed as auxiliary indicators to verify and analyze the results of the SOM, and the source of all water samples was confirmed to be surface water. Based on the hydrochemical data of 56 water sampling points, Panagopoulos et al. [29] used unsupervised and supervised statistical methods to group groundwater samples so as to identify the source of unknown water samples, which provided a simple multivariate statistical method for the identification of groundwater aquifer sources. Based on the measured water hydrochemical data, Lin et al. [11] identified the water inrush source of Zhaogezhuang mine through the coupling model of the improved genetic algorithm and extreme learning machine, and solved the complex non-linear problem encountered in identifying the water source. The research results have been conclusive and provide a scientific basis to formulate effective control measures after an inrush disaster has occurred.

However, hydrochemical methods require sampling experiments to determine the various concentrations of ions and molecules in the solution and eliminate any interference that could affect the rapid identification of water sources [33]. At the same time, many water inrush disasters have shown that the water inrush mode and the inrush points may involve various water sources [34]. Their ion content after merging may change significantly due to chemical and physical reactions, and this can impact the accurate determination of the water sources. For the identification of mixed water sources, there is a paucity of research. A high degree of groundwater mixing has become a difficult challenge to overcome, and mapping the relationship between the evaluation system and the category of the water source for identification purposes is uncertain. The accurate and rapid identification of multiple water sources in coal mine disasters could be an effective means of disaster prevention and control, but most studies have focused on only single water source identification. Without research studying its practical application, its effectiveness remains theoretical.

As a hybrid discipline of chemistry and physics, spectroscopy provides streamlined, rapid water source identification. Since the beginning of the 21st century, hyperspectral imaging has developed rapidly due to its multi-band characteristics and high resolution. Different substances absorb different spectral band energies and reflect the various internal elements in water samples on physical, chemical, and geometric scales. As compared to the traditional multispectral sensor, it can obtain more detailed, accurate spectral information. Through the analysis and calculation of spectral characteristics, it can achieve the identification of aquifer composition characteristics so as to achieve the purpose of identifying water sources. As compared to traditional methods, spectral identification of water sources does not require a complex laboratory and in situ analyses such as ion detection and water temperature/level monitoring, which improves its efficiency.

This study comprehensively considered the requirements of rapid identification and water inrush source mixing ratios and introduced spectral technology to identify water sources. By analyzing the characteristics of the spectral curves obtained by mixing different water sources of different proportions, the adaptability of deep-learning classifier

based on the genetic algorithm (GA) and extreme gradient boosting (XGBoost) algorithm to a variety of mixed mine water sources was established. The innovations of this study included the following:

(1) The application of the absorption spectrum in water source identification. Without single ion detection, the laboratory measurement stage could be simplified, and the spectral value was directly trained and greatly improved the efficiency of water source identification.
(2) The GA was used to optimize the parameters of XGBoost, and the XGBoost set algorithm was applied for water source analysis. This classifier reduced the shortcomings of the single tree model and traversed the segmentation points to achieve an optimal solution. GA simulates the principle of natural selection and optimizes the parameters while avoiding local optimal solutions, thereby obtaining the relative optimal parameters. The combination of the two could determine the proportion of water sources.
(3) The normalization algorithm and evaluation index were changed. To obtain the proportion of mixed water samples, the normalization and evaluation indexes were improved to evaluate the quality of the model.

2. Study Area and Data

2.1. Study Area

Huangyuchuan coal mine is located in the midwest of Zhungeer coalfield (Figure 1). The administrative division is under the jurisdiction of Changtan and Xuejiawan in Zhungeer. The coal mine is 20 km north of Xuejiawan, approximately 120 km from Hohhot, and approximately 150 km west of Ordos. The east of the coal mine is adjacent to Shiyangou and Qingchunta coal mines, the north is adjacent to Suancigou coal mine, the south is adjacent to Changtan coal mine, and the west is bounded by the southern detailed investigation area. The terrain is high in the northwest and low in the southeast. The elevation of the upper reaches of the Tahara River in the northwest is 1366 m, and the elevation of the southern Haomigetuo is 870 m, with a maximum elevation difference of 496 m. The elevation is generally between 1050 m and 1250 m. The Yellow River flows from north to south through the eastern part of the coalfield, and the flow of major valleys in the coalfield finally reach the Yellow River. The larger valleys, from north to south, include Kongduigou, Longwanggou, Heidaigou, Haerwusugou, Guanzigou, and Shilichangchuan, and the extension direction is oblique or vertical to the strata. The branch ditches are mostly distributed in branches, and the source erosion is the main. The cross-section is mostly a V-type, belonging to the erosive loess plateau landform. Most of the upper reaches of the major valleys have springs that form streams in the middle and lower reaches. Mountain floods erupt in the rainy season with large flows and short duration. The major valleys are also the main channel to discharge atmospheric precipitation and groundwater in the coalfield. Huangyuchuan mine aquifer can be divided into three categories: loose-layer pore phreatic aquifer, clastic-rock-pore fissure aquifer, and limestone karst fissure aquifer.

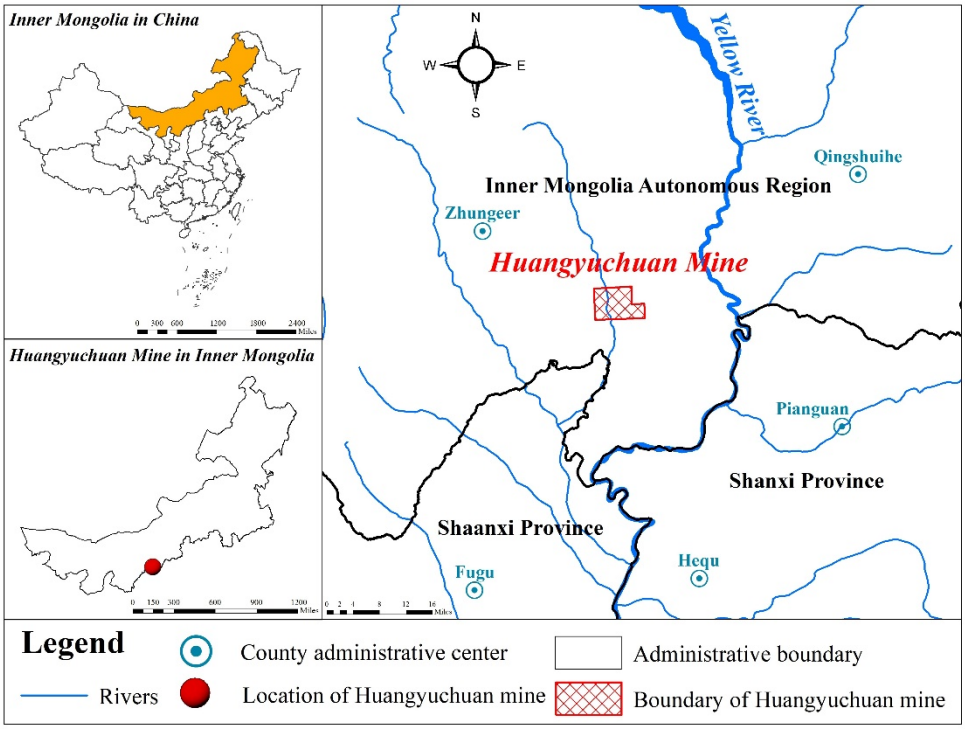

Figure 1. Study Area.

The water-filling sources in the mining area are mainly atmospheric precipitation and surface water, coal-bearing sandstone fissure water, Ordovician limestone water, and aged goaf water.

Atmospheric precipitation and surface water: there is no large surface water in the mining area except Shilichangchuan in Sanpan District, and there are some ponds scattered throughout the valley. Mine wellhead elevation is higher than the local flood level; during the rainy season, it is not threatened by surface water. Overall, atmospheric precipitation and surface water in mine water filling is limited.

Coal-measure-sandstone fissure water: Coal-measure-sandstone aquifer is water-rich with poor uniformity. In addition, the local fold-and-fault structures in the mining area are more developed, and when the water gathers at low depressions, there can be static reserves. In the actual underground production process, the bedrock fissure aquifer of coal-measure strata and the overlying quaternary loose-pore aquifer are weak water-rich aquifers with limited supply and have little influence on coal mining as they are easily drained with underground drainage. The roof of each working face showed no significant water leaching.

Ordovician limestone water: Ordovician limestone in coal-measure basements lacks peak–peak groups within the scope of the mine field. The karst development in the geological strata in the first and second panel areas is uneven with corrosion cracks and small dissolved pores. The two groups of corrosion cracks developed vertically, and the water content varies by location. It is separated from the Taiyuan Formation coal-measure strata by the Benxi Formation clay rock and mudstone aquifer. Under normal circumstances, the Ordovician karst water does not easily recharge the coal-measure strata. However, faults and collapsed columns are well developed in the field. Suspected karst collapse columns have been circled through three-dimensional earthquakes in the northern part of the first and second panels. When the Ordovician limestone water level is higher than the coal seam floor, there is a risk that Ordovician limestone water could enter the mine along the water-conducting fault, fracture zone, or collapsed column.

Aged empty water: According to the current exploration and production data, there are no aged kiln and production mines around and in the mine field and no aged kiln

water. The goaf water formed by mining the upper 4# coal seam in this mine could have affected the normal production of the lower 6# coal seam. According to the mining and excavation succession plan of the mine during the last three years, the influence areas of goaf water on mine safety production were 216 upper 05, 226 upper 01, and 226 upper 02 working faces. Therefore, the water in the goaf has been discharged in the detection and production process with little effect on water filling.

Overall, coal-measures sandstone fissure water and Ordovician limestone water are the greatest threats to mining production in the experimental area.

2.2. Data Collection

To develop a practical test for sampling according to the water source influence, water intake point should be relatively close, and we used a simple principle of water via hydrological observation hole sampling, drainage hole sampling, aquifer sampling, and water hole sampling of four selected sampling points in the Huangyuchuan coal mine, two sandstone fissure water, and two Ordovician limestone water. Sampling location information is shown in Table 1.

Table 1. Water sample point information.

Number	Water Sample Type	Sampling Point Location		Water Quality Type
		Longitude	Latitude	
H-1	Coal-measure sandstone fissure water	111.2047	39.6788	HCO_3-Ca •Mg
H-2	Ordovician limestone water	111.2049	39.6734	HCO_3-Na •Ca •Mg
H-3	Coal-measure sandstone fissure water	111.2080	39.6790	Cl •HCO_3-Mg (Ca)
H-4	Ordovician limestone water	111.1861	39.6785	Cl •HCO_3-Na

3. Methods

3.1. Research Framework

In this study, the discrimination of water inrush source in the Huangyuchuan Mine was conducted according to the following steps:

(1) Preparation of water sample spectral data. The water samples in the study area were collected and mixed according to the ratios of 0:10, 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1, 10:0, and 0:0 to prepare the absorption spectra data of the water samples with different mixing ratios.

(2) Spectral data preprocessing. The spectral data of the water samples were preprocessed by data filtering, scattering correction, baseline correction, smoothing and denoising, and data enhancement to correct the spectral dataset and improve the recognition accuracy.

(3) Establish a deep-learning classifier. The GA-XGBoost deep-learning classifier was constructed, and seven parameters of XGBoost were optimized by GA to classify and learn the absorption spectrum data of the water samples.

3.2. Preparation of Spectral Data of Mixed Water Samples

Ultraviolet-visible-near infrared short-wave spectroscopy was used to detect a variety of water quality parameters based on the absorption properties of organic and inorganic matter in water. It has obvious advantages in water quality monitoring technology [35,36]. The analysis method was divided into qualitative and quantitative analyses based on the curve shape of the absorption spectrum, such as the corresponding band, absorption peak height, and absorption peak number. The principle of quantitative analysis was according to the Lambert–Beer law, that is, the intensity of the absorption spectrum is proportional to the concentration of the measured substance when the absorption path is fixed [37]. The formula is:

$$A = \lg I/I\_0 = kbc$$

where A is absorbance, I_0 is the intensity of incident light, I is the transmission light intensity, k is the molar absorption coefficient, b is the liquid layer thickness (absorption path), and c is the concentration of light absorbing material.

The total absorbance of the measured object is equal to the sum of the absorbance of each substance in the medium when there are many substances that absorb light from the measured object. The formula is:

$$A = A\_1 + A\_2 + A\_3 + \ldots + A\_n$$

According to the additional properties of absorbance and the absorption spectrum characteristic curve of the measured object, multi-material analysis and multi-parameter correlational analysis can be done [38].

In this study, an ultraviolet and visible spectrophotometry (UV–Vis) was used to prepare the spectral data of mixed water samples taken at the study area. In the experiment, it was necessary to control the influence of environmental changes on the experiment, so the environmental parameters of the instrument were set including temperature of 19 °C with a relative humidity of 22%; the instrument was not affected by direct sunlight, strong electric or magnetic fields, or strong vibration interference, corrosive gas, and strong airflow, and it had an adequate AC power supply.

In the collected water samples, H-1 and H-2 were selected as a group, and H-3 and H-4 were selected as a group. The samples were mixed according to the ratios of 0:10, 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1, 10:0, and 0:0. Three water samples were collected for each ratio, and 66 water sample data were collected. The images were plotted after the absorbance values were obtained using a UV–Vis. The images were expressed in dark red, red, red-orange, orange-yellow, yellow-green, green, navy blue, blue, purple, and black lines, in proportion.

3.3. Preprocessing of the Absorption Spectroscopic Data

The spectral pretreatment work was divided into the following steps [39–44]:

(1) Filter data. Due to the band selective absorption of light, the peaks of spectral data were concentrated in the ultraviolet region. The absorbance of most water samples was distributed in the band 190–250 nm, and some water samples were more than 250 nm. The absorbance values of the water sample spectral data between 300 and 1100 nm were basically 0. To reduce the amount of spectral data, facilitate data preprocessing, improve the modeling efficiency of water inrush source identification model, and maximize the retention of spectral information, only the spectral data between 190 and 300 nm were selected.

(2) Baseline correction. In the process of spectral data acquisition, the error caused by instrument background or other factors could not be avoided and would cause a baseline drift in the collected spectral data. Furthermore, the interference among different components of the sample could also lead to overlapping in the absorption spectrum. To solve this problem, the necessary baseline correction method was used to process and optimize the spectrum.

(3) Scattering correction. Light scattering correction accounted for the spectral differences caused by the uneven distribution of sample particles and the scattering of particle sizes in the process of diffuse reflection data acquisition.

(4) Smoothing and denoising. Spectral signals contain not only useful signals, but also useless signals, which are called noise. Smoothing and denoising reduces the random noise in the spectral signal to improve the signal noise ratio of the overall spectral signal.

(5) Data enhancement. To effectively improve the robustness and prediction ability of the model, the original spectral data had to be processed using a data enhancement

algorithm to eliminate repetitive information, increase the difference between samples, and normalize the absorbance of samples to the same order of magnitude; this would enhance the internal comparison of the data and eliminate errors due to large differences in values.

Through the analysis and spectral preprocessing methods, the original data were significantly adjusted according to the standard normal variable transformation, and the multivariate scattering correction was adopted. The baseline correction considered that the noise of the second-order derivative was too large, which rendered the original relative gathered data more discrete and adopted the first-order derivative method; smoothing and denoising applied various evaluation indexes to calculate the results of each smoothing method. After comparing the evaluation indexes, the wavelet transform method with db19 as the wavelet transform function was selected as the smoothing method. As compared to the original data and other smoothing methods, this smoothing method was closer to the real optimal value. The data enhancement selected the commonly used min–max normalization as the basic algorithm, and on this basis, it was improved. The specific principles of spectral data preprocessing methods can be found in Supplementary Materials.

3.4. XGBoost Algorithm

XGBoost was proposed by Chen et al. [45]. Due to its outstanding efficiency and high prediction accuracy, it has drawn widespread attention. As compared to the conventional boost algorithm, the XGBoost algorithm accounts for the generation of weak learners by optimizing the structured loss function (the loss function with the regularization term was added, which reduced the risk of overfitting). Moreover, the XGBoost algorithm does not use a search method but directly uses the first-order and second-order derivatives of the loss function, and improves the performance of the algorithm by pre-sorting, weighted quantile, and other technologies.

The XGBoost algorithm adopted the step-by-step forward additive model, and the basic model was the tree model. XGBoost continuously splits the features to grow a tree. When the t-th tree is constructed, the residuals generated by the regression prediction of the training samples of the previous t-1 tree are fitted. When each fitting generates a new tree, the tree structure is traversed to obtain the tree structure that minimizes the objective function value. Finally, when the training is completed, k trees are obtained. The corresponding scores of each tree are calculated, where each leaf node corresponds to a score. Finally, the predicted value of the sample is obtained by adding the corresponding scores of each tree [46,47]. The model is as follows. The detailed principle of the model is shown in the supplementary material:

$$F_T(X) = \sum_{m=1}^{T} f_m(X)$$

Among them, T is the number of trees; $f_m(X)$ is the expression of the m-th tree.

XGBoost can be used for both classification and ranking problems as well as regression problems. In this study, the regularization method was used to measure the complexity of the tree, which controlled the complexity of the model and avoided overfitting. In addition, in the process of model optimization, the second-order derivative was introduced by the second-order Taylor expansion of the loss function, so that the model algorithm could be convex-optimized, and the convergence rate in the training process was improved along with the adaptability of the model. Referring to random forest algorithm, the algorithm supported row sampling and column sampling, which not only reduced the risk of overfitting but also the amount of calculation.

3.5. Genetic Algorithm

A genetic algorithm is a stochastic global search optimization method that simulates natural selection and genetics. Starting from any initial population, a group of individuals more suitable for the environment is generated by random selection, crossover, and mutation operations, and then the population evolves into an increasingly appropriate area for the search environment. This generation continues to multiply and evolve until it finally converges upon a group of individuals most suitable for the environment, so as to obtain high-quality solutions to the problem. For iterative enumeration, the local minimum trap is a typical error in processing that then cannot solve the global maximum. A genetic algorithm uses random probability mechanism to iterate, avoiding the problem of local maximum. The enumeration search process does not consider the intrinsic properties of the model; the optimization problem does not need to improve the algorithm, only to use the evaluation function to inspire, so there are no mathematical requirements for the algorithm problem. The process is simple; it is flexible and scalable for special problems. Hybrid construction ensures the effectiveness of the algorithm and is easy to combine with other algorithms [48].

The parameter tuning of XGBoost was more complex, including learning rate, base model selection, tree parameters (e.g., maximum depth, minimum weight, gamma, subsample ratio, tree column sample), adjusting regularization parameters (lambda and alpha), iterations, and so on. In this study, a genetic algorithm was selected to optimize the parameter tuning process of XGBoost. The specific operation steps were as follows:

The first step, encode. Selected the appropriate encoding scheme to convert variable encoding to chromosome string, usually binary encoding.

The second step, generate populations. Randomly generated N different individuals to form an initial group according to the parameters selected by the model.

The third step, calculate the adaptation values. Enumerate each individual in the early generation and calculate the fitness value of each individual according to the fitness function.

The fourth step, selection. Referred to the individual's fitness value, according to specific rules to select the best individual retention as the parent.

The fifth step, crossover and mutation. Two individuals were selected from the parent generation, and two gene points were randomly selected. The eigenvalues of the gene points of the individuals were exchanged to obtain the new individuals.

The sixth step, judgment loop. If the genetic algebra satisfied the termination condition, the calculation was terminated; otherwise, the calculation was transferred to the second step.

3.6. Model Accuracy Evaluation

Root-mean-square error (RMSE) represents the error between the measured value and the predicted value, which is a common evaluation index to measure model prediction ability and algorithm performance [49]. In this study, since the model would predict the proportional relationship of mixed water samples, the *RMSE* was improved, and a score evaluation method based on *RMSE* was established:

$$score = \frac{A}{RMSE}$$

Among them, *RMSE* was calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}$$

Among them, y_i is the measured value; $\hat{y}_i$ is the predicted value.

A represents the absolute value of the difference between the measured value and the predicted value, the ratio of the number of items less than 0.8 to the total number:

$$A = \frac{len(|y_i - \hat{y}_i| \leq 0.8)}{len(|y_i - \hat{y}_i|)}$$

In this study, the larger the score value, the better the performance of the model.

4. Results and Analysis

4.1. Raw Spectral Data Analysis

The band of this study was 190–1100 nm, and the spectral resolution was set at 1 nm. To reduce the amount of spectral data, improve the modeling efficiency of the water inrush source identification model, and maximize the retention of the spectral information, the original spectral data were screened, and only 190–250 nm data were used as experimental data for model processing. The spectral data of the filtered mixed water samples are shown in Figure 2.

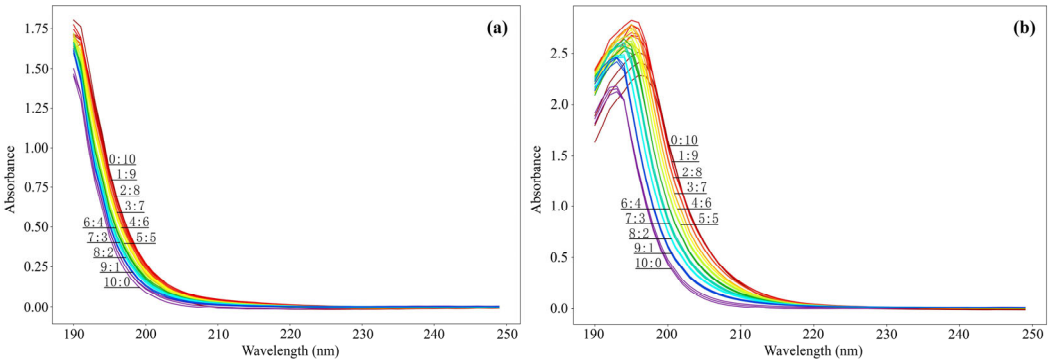

Figure 2. (**a**) H-1 and H-2 mixed water sample spectra, (**b**) H-3 and H-4 mixed water sample spectra.

As shown in Figure 2, the mixed water samples of H-1 and H-2 showed a downward trend between 190 nm and 220 nm and then gradually slowed down, and the 220 nm tended to be 0; the spectral curve ascended as the proportion of the OM7 water sample increased; the fluctuation of the spectral values of each proportion is small. The mixed water samples of H-3 and H-4 first increased between 190 nm and 195 nm, began to decrease and gradually slowed down after 195 nm, and then tended towards 0 after 220 nm, but the size relationship of the spectral curve was reversed. With the increase in the proportion of the H-3 water samples, the whole spectral curve ascended further, and the peak was near the ultraviolet direction, but only the peak of the H-3 water samples and the H-3/H-4 ratio of 9:1 had a low curve. The stability of the water samples mixed between 190 nm and 197 nm was better, while the volatility of the pure water samples was larger. Overall, the difference between the two groups of mixed samples was small, and the connectivity of the sandstone aquifer and the Ordovician limestone aquifer in Huangyuchuan Mine was better.

Due to differences in the types of substances, their concentrations, their absorption and scattering intensities, the turbidity of the water samples, and the PH values, the spectral curves of different aquifers were relatively different. The results shown in Figure 2 indicated that the spectral absorbance of aquifers of different mines was different. The aquifer water samples of the same mine had common characteristics, but the curves of characteristic peaks and spectral curve inclination were also relatively different. Therefore, the correlation of the original water samples for each aquifer was calculated and the Pearson's correlation coefficient between each aquifer and between each band and proportion was obtained (Figure 3).

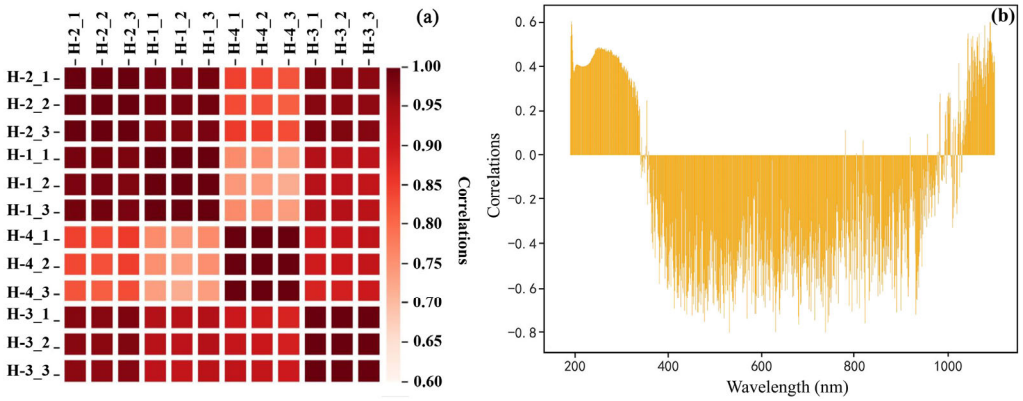

Figure 3. (**a**) Pearson's correlation coefficient between each aquifer; (**b**) Pearson's correlation coefficient between each band and proportion.

From the correlation of the water samples of various water layers, the H-3 sample was approximately 0.7. The correlation of the water samples of several other water-containing layers was relatively strong. According to the graphs of the correlation coefficients of each band, the correlation between 190–300 nm and 1030–1100 nm was relatively large, but most were approximately 0.4, and the maximum was not 0.6. It also showed that the water samples needed to be processed further to obtain a correlation in the data analysis.

4.2. Spectral Data Preprocessing Analysis

In this study, the spectral data of the original water sample were simply preprocessed, including reducing the absorbance of the blank water sample of distilled water, the influence of measurement background, and water sample interference. On this basis, the portion with the largest fluctuation of 190–250 nm of the two mixed water samples was preprocessed, including selecting multivariate scattering correction to eliminate the particle scattering in the original data, using the first-order derivative to eliminate the baseline translation in the data, and using wavelet transform to eliminate the spectral noise.

Figure 4 shows the processing results of the scattering correction. As shown in Figure 4, the baseline translation and offset of each spectrum were corrected according to the standard spectrum, which not only improved the signal-to-noise ratio of the spectrum but also corrected the fluctuation of the rising band of the mixed water sample and increased the spectral curve.

Figure 5 shows the results of the baseline correction. In Figure 5, the curve of the first derivative changed significantly, and the general trend was to descend and then rise. The curve above 0 fluctuated more and was more discrete while the part of the 0 axis was separated; the part of the curve that originally tended towards the 0 axis had a size relationship; now, it was on the 0 axis. Although the spectrum was more conducive to the identification and purity testing of the analyte, the noise increase was too great, and the amount of information was too large. This would result in an increase in the proportional discrimination error of the water source, and better results would be obtained after the first derivative of the baseline treatment.

Figure 6 shows the results of smoothing and denoising. As shown in Figure 6, as compared to the original curve, the curve after the wavelet inverse transformation had a significant change, the smoothing effect of the spectral curve was obvious, the extreme value was optimized, the noise was reduced, and the general fluctuation was retained.

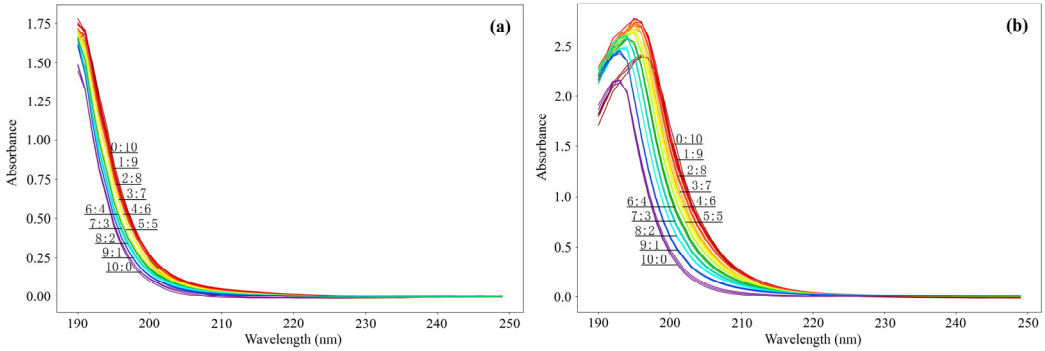

Figure 4. Scattering correction diagram. (**a**) Spectra of H-1 and H-2 mixed water samples, (**b**) Spectra of H-3 and H-4 mixed water samples.

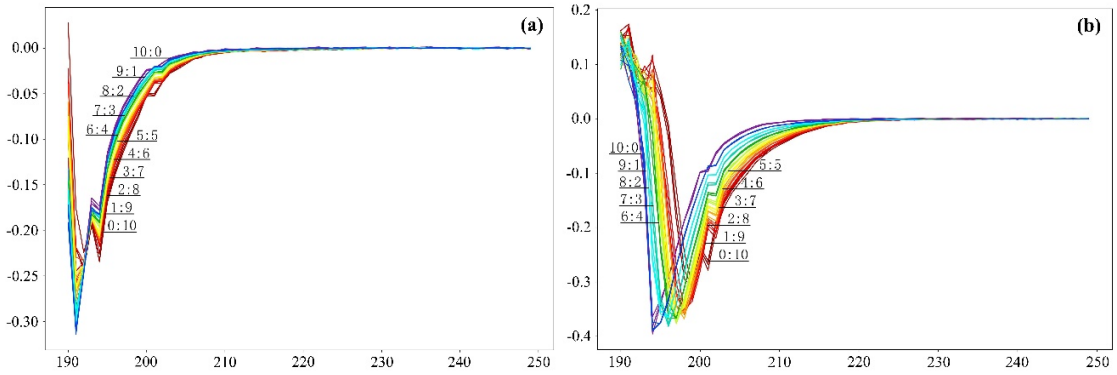

Figure 5. Baseline correction diagram. (**a**) Spectra of H-1 and H-2 mixed water samples, (**b**) Spectra of H-3 and H-4 mixed water samples.

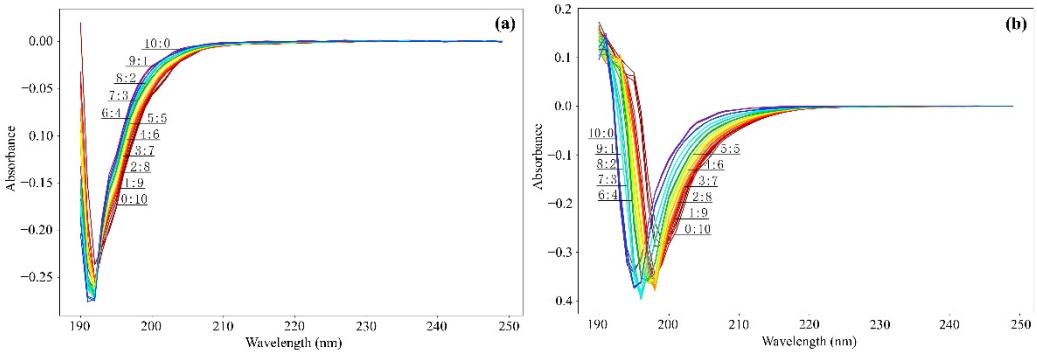

Figure 6. Smooth denoising diagram. (**a**) H-1 and H-2 mixed water sample spectrum, (**b**) H-3 and H-4 mixed water sample spectrum.

Overall, after spectral processing, the image had been significantly improved for both the two aquifers with the same mixing ratio of the spectral curve dispersion and the two aquifers with different mixing ratio of spectral curve discrimination. The correlation of the original water samples of each aquifer was further calculated, and Pearson's correlation coefficient of each aquifer and of each band are shown in Figure 7. As shown in Figure 7, the correlation between the transformed and the water samples before the transformation was significantly reduced, which ensured that the difference between the samples was larger.

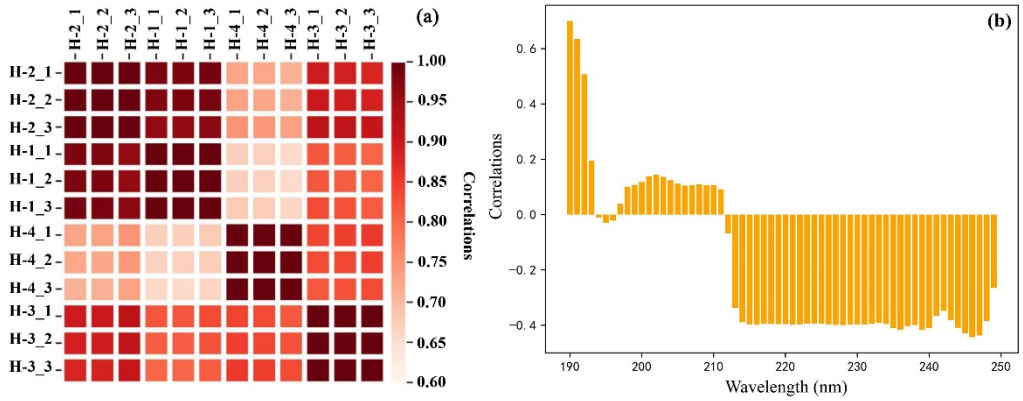

Figure 7. (**a**) Pearson's correlation coefficient between each aquifer after treatment, (**b**) Pearson's correlation coefficient between each band and proportion after treatment.

Finally, to normalize all the spectral data and facilitate model analysis, the data enhancement processing was improved for each spectral dataset.

4.3. Model Result Analysis

In this study, the water-like spectrum data were randomly divided into training and test sets at a ratio of 8:2. During the operation, the optimal solution was not uniform due to the differences in the breeding algebra and the differences in the random initial population. The results of the calculation would not be large enough. After several experiments, the initial population was 20, and the population breeding algebra was determined 300 times. The optimization parameters of the genetic algorithm are shown in Figure 8.

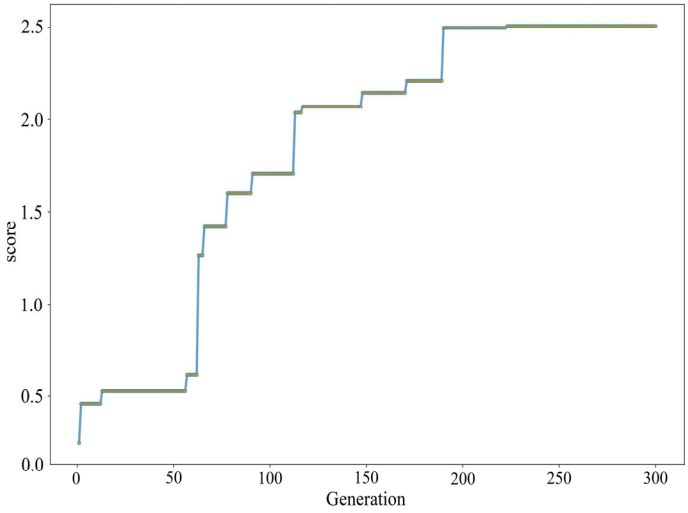

Figure 8. Optimization parameter process of genetic algorithm.

The optimal evaluation index in the genetic algorithm continued to increase. At iteration 222, the maximum value of the construction evaluation index did not change, and the maximum value was 2.51. The optimal parameters were:

learning_rate: 0.10
n_estimators: 100
max_depth: 1
min_child_weight: 1.99
gamma: 0.01
subsample: 0.95
colsample_bytree: 0.32

Since the results obtained by different training and prediction set samples could be different, to ensure that the model constructed by this parameter had good adaptability in a specific environment, a new segmentation and cross-validation was conducted on the data in the ratio of 8:2. After 19 min of training and 26 s of prediction, the GA-XGBoost model yielded the results shown in Figure 9.

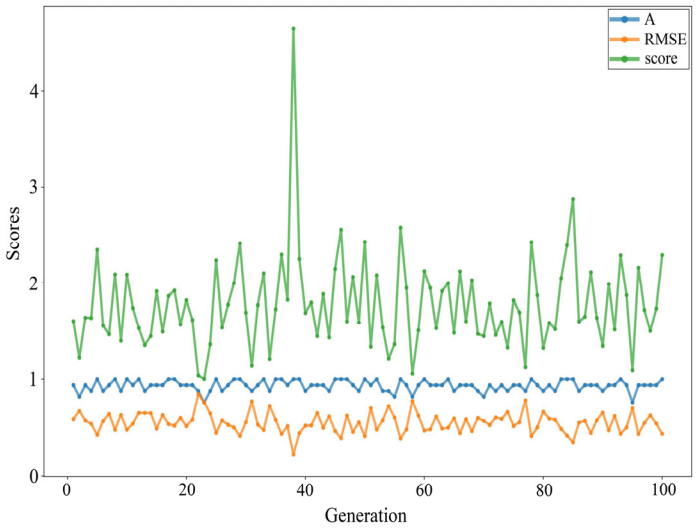

Figure 9. Cross-validation of evaluation scores.

As shown in Figure 9, the average value of A (blue line) was approximately 0.94, which was a proportion within the threshold; in other words, 16 of the 17 predicted samples were accurate. The RMSE value (average value of yellow line) was approximately 0.55, which was much smaller than the value of the adjacent label reduction. This indicated that the overall predictive value and the experimental value level were relatively small. The score value (green line) reached approximately 4.7 and only 1, and the fluctuations were relatively large. It also showed that the selection of samples of A and RMSE was relatively large. Overall, the score value was distant from the A and RMSE values, which also indicated that the prediction was explained. Therefore, the effect was better.

To show the necessity of data preprocessing and enhancement, GA was used to calculate the corresponding parameters, predict, and cross-validate the original samples and the samples without improved normalization. The results are shown in Figure 10.

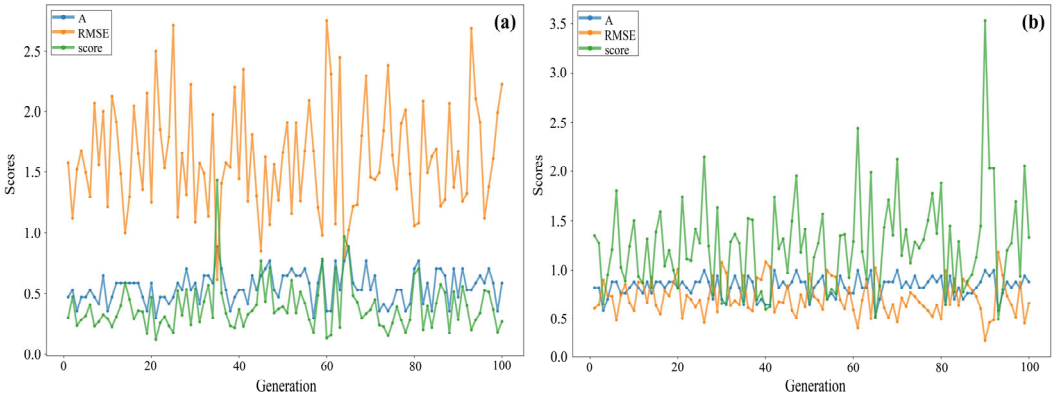

Figure 10. (**a**) Prediction results without pretreatment, (**b**) Prediction results without data enhancement.

As shown in Figure 10, the RMSE values of the prediction results without pretreatment were significantly higher than those of A, which resulted in a smaller score value; therefore, the prediction results were insufficient. Although the average score value of the

prediction results without data enhancement reached the level of 1.55, it was not separate from the RMSE and A values, and the lines overlapped. Moreover, the predicted value of A was less than that of the data with improved normalization, and the average value was approximately 0.83. In other words, 14 of the 17 met the requirements, which also indicated that preprocessing and normalization were necessary.

To further highlight the advanced nature of the research algorithm and the necessity of spectral data preprocessing, a box diagram of A, RMSE, and score after different data processing methods was constructed (Figure 11). As shown in Figure 11, the tail length and interquartile spacing of the untreated and unnormalized box map were the box map with the obtained A values, which were unstable. The median line of the processed A and score value was higher than the other two, while the median line of RMSE value was much lower, and the accuracy of the processed data was greatly improved.

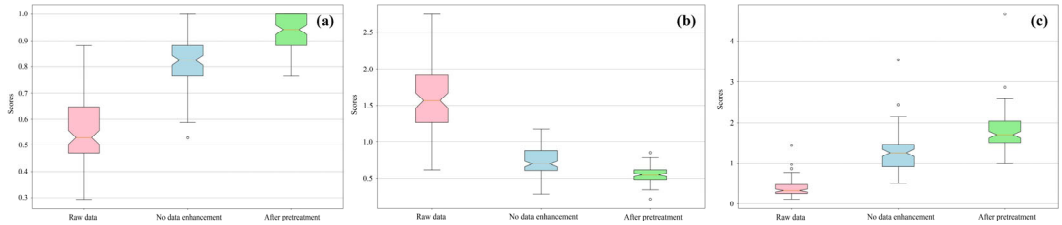

Figure 11. (**a**) A, (**b**) RMSE, (**c**) score.

5. Discussion

In this study, absorption spectrum technology was introduced to identify mine water inrush sources, and 66 spectral curves of the water samples in Huangyuchuan Mine were prepared using a UV–Vis. To address the spectral data redundancy, baseline drift, uneven particle distribution, random noise, and many other factors that could have affected the prediction results of the model, a water sample spectral preprocessing method system, including data screening, scattering correction, baseline correction, smoothing and denoising, and data enhancement was established. To address the complex parameter tuning of the XGBoost algorithm, a deep-learning classifier of water inrush sources based on GA-XGBoost was established. GA selected seven optimal XGBoost algorithm parameters through random global search optimization; the XGBoost algorithm measured the complexity of the tree by using the regularization method and introduced the second-order derivative to control the complexity of the model and avoiding over-fitting, so that the model algorithm could be convex-optimized, and the convergence speed and adaptability in the training processes were improved. In addition, considering the particularity of the study, a new model accuracy evaluation method was defined. The model operation results showed that the training time of GA-XGBoost model was 19 min, the prediction time was 26 s, and the average evaluation accuracy was 94%, which indicated that the GA-XGBoost model was reliable in the rapid identification on a mixed water inrush source spectrum. In addition, the prediction accuracy of the model before and after spectral data preprocessing was further compared. The results showed that the accuracy of the processed data had been greatly improved, and the spectral preprocessing was proved.

In addition, the research also had some limitations. These included the amount of data in the research being relatively small and the generalizability of the model being poor. The absorbance curves of the water samples in different mining areas had significant differences and diversity. The correlation between the spectral curves from a single water source and different water sources was strong, and it was not easy to distinguish in the algorithm, so a larger volume of data could improve the discriminant ability of the model. Secondly, the study only examined the mixing of two types of the water sources. For scenarios with many types of water samples, the treatment method would need to be further refined, and the model parameters would need to be further optimized. Future research

should consider using a larger dataset and improving data processing and model discrimination in multidimensional situations.

6. Conclusions

This objective of this study was to establish a new method for identifying mixed water inrush sources using absorption spectroscopy. Therefore, a mixed water source identification method was established, including spectral data preprocessing, deep-learning classification for spectral data identification of mixed water samples, and model accuracy evaluation. The results showed that the spectral pretreatment of the water samples effectively resolved redundancies in the original spectral data, baseline drifting, uneven particle distributions, and random noise. The GA-XGBoost deep-learning classifier was used to identify and predict the water sample data. The prediction accuracy of the final model was 94%, which was 11% higher than that of the prediction before the spectral data were preprocessed. This further illustrated the reliability of the GA-XGBoost model for the spectral identification of mixed water sources, as well as the necessity and importance of spectral data preprocessing. The method and research results established in this study provided a new direction for the study of mine water inrush source discrimination, as well as a technical reference for decision-makers when developing effective mine water disaster prevention programs. In addition, due to the lack of experimental data, the study only conducted a mixed simulation experiment with two types of water sources. Future research should involve a more in-depth experiment and simulation for the mixing of multi-aquifer water sources.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/article/10.3390/w14142150/s1.

Author Contributions: Conceptualization, X.L. and D.D.; methodology, K.L.; software, Y.Z.; validation, Y.Z. and X.L.; formal analysis, Y.Z.; investigation, X.L.; resources, D.D.; data curation, Y.Z.; writing—original draft preparation, X.L.; writing—review and editing, K.L. and M.L.; visualization, Y.Z. and X.L.; supervision, K.L.; project administration, X.L.; funding acquisition, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the State Key Laboratory of Resources and Environmental Information System, the National Natural Science Foundation (41972255, U1710258).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the author.

Acknowledgments: We like to thank the anonymous reviewers for their helpful remarks.

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

References

1. Wang, S.M. Thoughts about the main energy status of coal and green mining in China. *China Coal* **2020**, *46*, 11–16.
2. Gu, Q.X.; Huang, Z.; Li, S.J.; Zeng, W.; Wu, Y.; Zhao, K. An approach for water-inrush risk assessment of deep coal seam mining: A case study in Xinlongzhuang coal mine. *Environ. Sci. Pollut. Res.* **2020**, *27*, 43163–43176.
3. Chen, Y.; Zhu, S.Y.; Wang, Z.G.; Li, F.L. Deformation and failure of floor in mine with soft coal, soft floor, hard roof and varying thicknesses of coal seam. *Eng. Fail. Anal.* **2020**, *115*, 104653.
4. Sun, W.J.; Han, Q.; Yang, H.; Yang, W.K.; Li, S.Q. Analysis on water inrush accidents in China's coal mines 2000–2015. *Coal Eng.* **2017**, *49*, 95–98.
5. Wu, Q.; Guo, X.M.; Shen, J.J.; Xu, S.; Liu, S.Q.; Zeng, Y.F. Risk Assessment of Water Inrush from Aquifers Underlying the Gushuyuan Coal Mine, China. *Mine Water Environ.* **2017**, *1*, 96–103.
6. Wu, Q.; Zhao, S.Q.; Dong, S.N.; Li, J.S. *Coal Industry Press Coal Mine Prevention and Control Volume*; China Coal Industry Publishing House: Beijing, China, 2013.
7. Zhang, H.T.; Xu, G.Q.; Chen, X.Q.; Wei, J.; Yu, S.T.; Yang, T.T. Hydrogeochemical Characteristics and Groundwater Inrush Source Identification for a Multi-aquifer System in a Coal Mine. *Acta Geol. Sin. Engl. Ed.* **2019**, *93*, 1922–1932.

8. Mao, Z.Y.; Huang, C.J.; Lu, S.C.; Han, R.Y. KPCA-MPSO-ELM based model for discrimination of mine water inrush source. *China Saf. Sci. J.* **2018**, *28*, 111–116.
9. Wang, Y.; Zuo, W.Z.; Wang, B.H.; Cheng, Z.H. Study Progress of Discriminant Method of the Sources of Mine Water Inrush. *Mod. Min.* **2018**, *34*, 69–73.
10. Dong, D.L.; Chen, Z.Y.; Lin, G.; Li, X.; Zhang, R.M.; Ji, Y. Combining the Fisher Feature Extraction and Support Vector Machine Methods to Identify the Water Inrush Source: A Case Study of the Wuhai Mining Area. *Mine Water Environ.* **2019**, *38*, 855–862.
11. Lin, G.; Jiang, D.; Dong, D.L.; Fu, J.Y.; Li, X. A Multilevel Recognition Model of Water Inrush Sources: A Case Study of the Zhaogezhuang Mining Area. *Mine Water Environ.* **2021**, *40*, 773–782.
12. Bi, Y.S.; Wu, J.W.; Zhai, X.R.; Wang, G.T.; Shen, S.H.; Qing, X.B. Discriminant analysis of mine water inrush sources with multi-aquifer based on multivariate statistical analysis. *Environ. Earth Sci.* **2021**, *80*, 144.
13. Vincenzi, V.; Gargini, A.; Goldscheider, N. Using tracer tests and hydrological observations to evaluate effects of tunnel drainage on groundwater and surface waters in the Northern Apennines (Italy). *Hydrogeol. J.* **2009**, *17*, 135–150.
14. Zhong, N.N.; Shi, Y.L.; Wang, F.Y.; Chen, D.Y. Three-Dimensional Excitation Emission Matrix Fluorescence Spectroscopic-Characterization of Dissolved Organic Matter in Water of Coal-Mining Area. *Spectrosc. Spectr. Anal.* **2008**, *28*, 174–177.
15. Hering, J.G.; Stumm, W. Fluorescence spectroscopic evidence for surface complex formation at the mineral-water interface: Elucidation of the mechanism of ligand-promoted dissolution. *Langmuir* **1991**, *7*, 1567–1570.
16. Ebrahimi, A.; Jafari, M.T. Negative corona discharge-ion mobility spectrometry as a detection system for low density extraction solvent-based dispersive liquid–liquid microextraction. *Talanta* **2015**, *134*, 724–731.
17. Wang. Y.; Zhou, M.R.; Yan, P.C.; He, C.Y.; Liu, D. Identification of Coalmine Water Inrush Source with PCA-BP Model Based on Laser-Induced Fluorescence Technology. *Spectrosc. Spectr. Anal.* **2017**, *37*, 978–983.
18. Zhou, M.R.; Lai, W.H.; Wang, Y.; Hu, F.; Li, D.T.; Wang, R. Application of CNN in LlF Fluorescence Spectrum Image Recognition of Mine Water lnrush. *Spectrosc. Spectr. Anal.* **2018**, *38*, 276–280.
19. Ayadi, R.; Trabelsi, R.; Zouari, K.; Saibi, H.; Itoi, R.; Khanfir, H. Hydrogeological and hydrochemical investigation of groundwater using environmental isotopes (^{18}O, ^{2}H, ^{3}H, ^{14}C) and chemical tracers: A case study of the intermediate aquifer, Sfax, southeastern Tunisia. *Hydrogeol. J.* **2018**, *26*, 983–1007.
20. Bouzourra, H.; Bouhlila, R.; Elango, L.; Slama, F.; Ouslati, N. Characterization of mechanisms and processes of groundwater salinization in irrigated coastal area using statistics, Gis, and hydrogeochemical investigations. *Environ. Sci. Pollut. Res.* **2015**, *22*, 2643–2660.
21. Chen, Y.; Tang, L.S.; Zhu, S.Y. Comprehensive study on identification of water inrush sources from deep mining roadway. *Environ. Sci. Pollut. Res.* **2021**, *29*, 19608–19623.
22. Hu, Y.S.; Huang, P.H.; Gao, H.F.; Su, Q.Q. State of the Practice Worldwide: HCA-PCA-EWM Discrimination Model of Water Inrush Source in Mines. *Ground Water Monit. Remedition* **2022**, *42*, 67–76.
23. Zhao, X.M.; Xu, Z.M.; Sun, Y.J. Identification of Mine Water Source Based on AHP-Entropy and Set Pair Analysis. *Geofluids* **2022**, *2022*, 3682746.
24. Chen, Y.; Zhu, S.Y.; Yang. C.W.; Xiao, S.J. Analysis of hydrochemical evolution in main discharge aquifers under mining disturbance and water source identification. *Environ. Sci. Pollut. Res.* **2021**, *28*, 26784–26793.
25. Yan, P.C.; Shang, S.H.; Zhang, C.Y.; Zhang, X.F. Classification of Coal Mine Water Sources by Improved BP Neural Network Algorithm. *Spectrosscopy Spectr. Anal.* **2021**, *41*, 2288–2293.
26. Jiang, C.L.; An, Y.Q.; Zhang, L.G.; Huang, W.W. Water source discrimination in a multiaquifer mine using a comprehensive stepwise discriminant method. *Mine Water Environ.* **2021**, *40*, 442–455.
27. Zhang, H.; Yao, D.X. The Bayes Recognition Model for Mine Water Inrush Source Based on Multiple Logistic Regression Analysis. *Mine Water Environ.* **2020**, *39*, 888–901.
28. Zhao, D.; Zeng, Y.F.; Wu, Q.; Du, X.; Gao, S.; Mei, A.S.; Zhao, H.N.; Zhang, Z.H.; Zhang, X.H. Source Discrimination of Mine Gushing Water Using Self-Organizing Feature Maps: A Case Study in Ningtiaota Coal Mine, Shaanxi, China. *Sustainability* **2022**, *14*, 6551.
29. Panagopoulos, G.P.; Angelopoulou, D.; Tzirtzilakis, E.E.; Giannoulopoulos, P. The contribution of cluster and discriminant analysis to the classification of complex aquifer systems. *Environ. Monit. Assess.* **2016**, *188*, 591.
30. Wang, D.K.; Ju, Q.D.; Wang, Y.Q.; Hu, Y.B.; Liu, Q.M.; Chai, H.C.; Liu, Y. Source identification of mine water inrush based on the exponential whitenization function and the grey situation decision model. *Energy Explor. Exploit.* **2022**, *40*, 1217–1235.
31. Ashwani, K.T.; Prasoon, K.S.; Mukesh, K.M. Environmental geochemistry and a quality assessment of mine water in the West Bokaro Coalfield. India: A Case Study. *Mine Water Environ.* **2016**, *35*, 525–535.
32. Sakizadeh, M.; Mirzaei, R.; Ghorbani, H. Geochemical influences on the quality of groundwater in eastern part of Semnan Province, Iran. *Environ. Earth Sci.* **2016**, *75*, 917.
33. Yan, P.C.; Shang, S.H.; Zhang, C.Y.; Yin, N.N.; Zhang, X.F.; Yang, G.K.; Zhang, Z.; Sun, Q.S. Research on the Processing of Coal Mine Water Source Data by Optimizing BP Neural Network Algorithm with Sparrow Search Algorithm. *IEEE Access* **2021**, *9*, 108718–108730.
34. Roy. A.; Das, B.K.; Bhattacharya. J. Development and validation of a spectrophotometric method to measure sulfate concentrations in mine water without interference. *Mine Water Environ.* **2011**, *30*, 169–174.
35. Shi, Z.; Chow, C.W.K.; Fabris, R.; Liu, J.; Jin, B. Alternative particle compensation techniques for online water quality monitoring using UV–Vis spectrophotometer. *Chemom. Intell. Labotatory Syst.* **2020**, *204*, 104074.

36. Zhang, H. Research progress and analysis of water quality monitoring technology based on UV spectral analysis. *Technol. Innov. Appl.* **2016**, *18*, 79.
37. Wei, K.L.; Wen, Z.Y.; Wu, X.; Zhang, Z.W.; Zeng, T.L. Research Advances in Water Quality Monitoring Technology Based on UV-Vis Spectrum Analysis. *Spectrosc. Spectr. Anal.* **2011**, *31*, 1074–1077.
38. Chen, C.F. *Online Monitoring Technology of Flue Gas Emissions with DOAS*; Tianjin University: Tianjin, China, 2007.
39. Diwu, P.Y.; Bian, X.H.; Wang, Z.F.; Liu, W. Study on the Selection of Spectral Preprocessing Methods. *Spectrosc. Spectr. Anal.* **2019**, *39*, 2800–2806.
40. Chu, X.L.; Yuan, H.F.; Lu, W.Z. Progress and Application of Spectral Data Pretreatment and Wavelength Selection Methods in NIR Analytical Technique. *Prog. Chem.* **2004**, *4*, 528–542.
41. Zhang, J.Y.; Zhu, H.T. Overview of spectral pretreatment methods. *West Leather* **2017**, *39*, 14.
42. Guo, Q.Q. *Research on Prediction Model of Soil Organic-Matter Based on the Near-Infrared Spectroscopy Technology*; Henan Agricultural University: Zhengzhou, China, 2016.
43. Wang, X.M.; Zhu, B.Y.; Yin, C. Application of derivative spectrometry in pharmaceutical analysis. *Fujian Anal. Test.* **2001**, *2*, 1431–1438.
44. Zhou, F.B.; Li, C.G.; Zhu, H.Q. Research on Threshold lmproved Denoising Algorithm Based on Lifting Wavelet Transform in UV-Vis Spectrum. *Spectrosc. Spectr. Anal.* **2018**, *38*, 506–510.
45. Chen, T.Q.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13 August 2016.
46. Tao, M.Q.; Liu, J.X.; Wu, Y.; Ning, Z.Q.; Fang, Y.H. Application of XGBoost in Gas Infrared Spectral Recognition. *Acta Opt. Sin.* **2020**, *40*, 201–206.
47. Zhang, X.; Luo, A.L. XGBOOST Based Stellar Spectral Classification and Quantized Feature. *Spectrosc. Spectr. Anal.* **2019**, *39*, 3292–3296.
48. Hua, C.Z.; Zhao, L.; Song, J.J. Selection of Wavelength for UV-visible Spectroscopy Based on BLS Combined with GAS. *J. Sichuan Norm. Univ. Nat. Sci.* **2019**, *42*, 825–829.
49. Wu, X.M.; Liu, Z.Q.; Zhang, T.L.; Li, H. A Method Based on Double Models Combination to Further Reduce Root-Mean-Square Error and Relative Error of Prediction. *Chin. J. Anal. Chem.* **2015**, *43*, 754–758.