

Original Research

Identification of Mine Water Inflow Source Based on Deep Learning Approaches

Junqing Sun^{1, 2, 3, *}, Hao Wang^{1, 2, 3}, Hongbo Shang^{1, 2, 3}, Xiaodong Wang^{1, 2, 3},
Tiantian Wang^{1, 2, 3}, Wei Qiao^{1, 2, 3}

¹CCTEG Xi'an Research Institute (Group) Co., Ltd., Xi'an 710077, P.R. China

²Shaanxi Key Laboratory of Prevention and Control Technology for Coal Mine Water Hazard, Xi'an 710077, P.R. China

³Shaanxi Engineering Research Center of Mine Ecological Environment Protection and Restoration in the Middle
of the Yellow River Basin, Xi'an 710077, P.R. China

Received: 25 May 2025

Accepted: 7 September 2025

Abstract

Mine water poses serious threats to mine safety and the sustainable exploitation of resources. Efficient identification of the water source of roof water inflow in coal seams is crucial. Hydrochemical analysis is key for mine water source identification, yet it is challenged by large, complex datasets. Deep learning offers an effective solution with its powerful data representation capabilities. This study develops a water source identification model based on a deep neural network (DNN), using hydrochemical and organic composition indicators including Na^+ , K^+ , Ca^{2+} , Mg^{2+} , SO_4^{2-} , HCO_3^- , Cl^- , TDS, TOC, UV_{254} , and dissolved organic matter. Bayesian optimization is applied to tune key hyperparameters of the DNN, such as learning rate, number of neurons in each layer, and training epochs, to achieve an optimal network architecture. The model is validated using 197 water samples collected from three representative coal mines located on the border between Inner Mongolia and Shaanxi Province, China. The proposed model achieves an identification accuracy of 96.31%, outperforming traditional classifiers such as support vector machines and random forests. The results indicate that this method has high accuracy and reliability, and can provide new ideas for quickly and accurately identifying the water source of coal seam roof water inflow.

Keywords: water source discrimination, deep neural network, fluorescence spectrum, bayesian optimization

Introduction

As a fundamental energy source, coal holds an irreplaceable position in ensuring energy supply, making efficient coal extraction essential [1, 2]. High-intensity coal mining is highly prone to causing mine

water disasters, posing a threat to the safety of coal mine production [3, 4]. Rapid and accurate identification of the source of water gushing is the prerequisite and key to the management of water disasters in coal mines. The primary methods for water source identification include water quality-level-temperature (QLT) analysis [5], geostatistical approaches [6], hydrochemical analysis [7, 8], and tracer simulation [9], among which hydrochemical analysis is the most direct and cost-effective approach.

*e-mail: 17851146658@163.com

°ORCID iD: 0009-0001-6196-0377

Currently, methods for discriminating the water source of water inflow based on hydrochemical characteristics primarily include multivariate statistics [10], nonlinear analysis [11], and machine learning algorithms [12-14]. Discrimination using hydrochemical indicators generates multidimensional, nonlinear water sample datasets. Conventional multivariate statistical methods and nonlinear analytical approaches are constrained by linear assumptions [15] and the curse of dimensionality [16], making them inadequate for fully capturing the intrinsic patterns within such complex, high-dimensional data [17]. The performance of methods such as Support Vector Machines (SVM), Random Forests (RF), and single-layer neural networks has improved compared to traditional approaches. However, these are considered shallow learning methods with limited capacity for data representation. Additionally, most machine learning algorithms require manual feature engineering, and the quality of feature selection directly affects the discrimination performance [18]. Deep learning, primarily based on neural networks, autonomously learns and extracts data features through its multi-layered structure, offering stronger generalization capabilities [6, 19]. Yang C. et al. [20, 21] established Fisher and BP (Back Propagation) neural network discrimination models for the study area, and the results showed that the BP neural network achieved higher accuracy. Jiang C.L. et al. [22] developed a water inrush discrimination model for mines based on a deep feedforward network, which demonstrated high discrimination performance. Jiang Q.L. et al. [23] developed a backpropagation neural network (BPNN) discrimination model optimized using factor analysis and particle swarm optimization (PSO), which demonstrated higher classification accuracy compared to clustering analysis. Fang B. [24] employed an improved shuffled frog leaping algorithm to optimize a convolutional neural network (CNN), achieving superior performance over the support vector machine (SVM)-based water source identification model. Yan P.C. et al. [25] compared CNN and residual neural networks (ResNets), and selected a ResNet-based approach incorporating centroid variation and PCA for water source identification. Machine learning often faces the challenge of finding the global optimum, and intelligent optimization algorithms are commonly used to adjust model parameters for optimal structure. Bayesian Optimization (BO) algorithms identify the optimal solution by incorporating prior knowledge [26]. With a straightforward principle and high accuracy, they are widely used for solving optimization problems [27]. Casanova R.H. et al. [28] combined bidirectional long short-term memory (LSTM) networks with the BO algorithm, enhancing the prediction accuracy of photovoltaic power. Wang M. et al. [29] improved the CNN-LSTM network using BO algorithms and accurately predicted the service life of artillery barrels.

Overall, machine learning algorithms, represented by deep neural networks, offer more powerful analytical

capabilities and show potential advantages in the field of coal seam roof water inflow source discrimination.

China has abundant coal reserves, with more than 70% of the country's coal resources located in the western mining areas [30, 31]. However, the high-intensity coal mining in these areas has increased the possibility of roof water inrush accidents, threatening the safe production of the mines. [32, 33]. Integrating advanced machine learning methods with various hydrochemical indicators for water inrush source identification is an important approach to ensuring the safe mining of coal. However, there is little research about the application of machine learning methods to study water sources using inorganic and organic water chemical indicators. This research aims to (1) analyze the inorganic and organic hydrochemical characteristics of groundwater in the mining areas of China based on data from three typical mines using graphical methods and parallel factor analysis (PARAFAC). (2) Construct a source identification dataset by extracting features from hydrochemical indicators using principal component analysis (PCA). (3) Develop a deep learning model for source identification, optimize its structure using the BO algorithm, and train and validate the model using the constructed dataset.

Materials and Methods

Study Area

The Mongolian-Shanxi border area is a key coal production region in western China (Fig. 1). This paper uses the region as a case study to test and validate the approach for water source identification. Due to intense mining, fractures often connect multiple overlying aquifers, increasing the risk of roof water hazards. Three representative mines along the northeast-southwest trend in this area were selected for the study. The mines are situated in the Maowusu Desert, within the northern Shaanxi slope and Yishan anticline of the Ordos Basin. The coal-bearing strata in the study area are of the Yan'an Formation, with similar stratigraphic structures. The main aquifers (or aquitards) from top to bottom are (Fig. 2): Quaternary porous aquifer, Cretaceous Luohe Formation porous-fracture aquifer, Anding Formation relative aquitard, and ZhiLuo and Yan'an Formation fracture aquifers [34]. The Quaternary sandstone layer, being uniform and loose, facilitates atmospheric precipitation infiltration, resulting in strong water abundance. The Luohe Formation aquifer, with well-developed pores and fractures, has good connectivity and recharge capacity. The Zhiluo Formation, with higher lithification and poor fracture development, exhibits weaker water abundance. The Yan'an Formation, with increased mud content and poor recharge conditions, has the lowest water abundance. During the construction of the mines in the study area, there is an issue of water-rich coal seams and their immediate roof. In the future

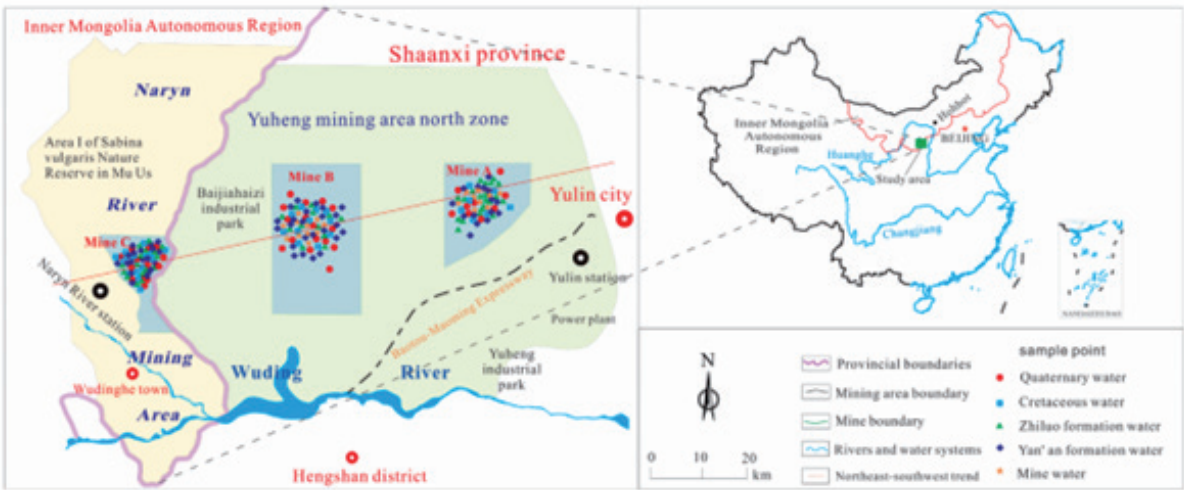


Fig. 1. Location of the study area.

Stratum		Histogram	rock character	Water-containing (water-isolating)
System	Formation			
Quaternary			clayey silt, loam	pore aquifer, the water-rich property is medium to strong
Cretaceous			massive coarse-grained sandstone, glutenite	pore-fissure aquifer with medium water abundance
System	Luohe K_1l		rhythm layer of mudstone and fine sandstone	relative to the aquiclude, The water-proof function is relatively stable
Jurassic	Anding J_2a		medium coarse grained sandstone	clastic fractured aquifer with weak water abundance
	Zhiluo J_2z		clayey silt, loam	clastic fractured aquifer with weak water abundance
	Yanan J_2y		coarse-grained feldspathic sandstone, fine sandstone, siltstone, silty mudstone, mudstone, with carbonaceous mudstone and coal seam	pore aquifer, the water-rich property is medium to strong
Triassic			Medium-fine grained feldspathic sandstone, grayish white siltstone and grayish black mudstone.	

Fig. 2. Stratum histogram of the study area.

mining process, the fissure water in the roof sandstone will directly serve as the water source.

Water Sampling and Analysis

Water Sampling

A total of 197 water samples were collected, including 177 groundwater samples and 20 mine water samples. The groundwater samples are from the following formations: 38 from the Quaternary, 32 from

the Luohe Formation, 49 from the Zhiluo Formation, and 58 from the Yan'an Formation. A total of 177 groundwater samples were used for model fitting and validation in the water source model training process, while 20 mine water samples were used for testing the water source discrimination model.

Water samples were collected using 2.5 L polyethylene sampling bottles, with strict adherence to procedures for filling, sealing, and labeling. Analytical methods were as follows. K^+ , Na^+ , Ca^{2+} , and Mg^{2+} were determined using an inductively coupled plasma optical

emission spectrometer (ICP-OES, Optima 8300DV, PerkinElmer). SO_4^{2-} and Cl^- were analyzed using an ion chromatograph (ICS-600, Thermo Scientific, USA). Bicarbonate (HCO_3^-) concentrations were obtained via acid-base titration. Total dissolved solids (TDS) were measured using the gravimetric (dry weight) method. Total organic carbon (TOC) was analyzed using a multi N/C 2100 TOC/TN analyzer (Analytik Jena, Germany) after filtration through a $0.45 \mu\text{m}$ membrane. Ultraviolet absorbance at 254 nm (UV_{254}) was measured with a UV-Vis spectrophotometer (Evolution 60, Thermo Scientific) using a 1 cm quartz cuvette, with corresponding blanks measured simultaneously. Three-dimensional excitation-emission matrix (EEM) fluorescence spectra of dissolved organic matter (DOM) were obtained using a fluorescence spectrophotometer (F-7000, HITACHI) at a scanning speed of 1200 nm/min. Excitation wavelengths ranged from 200 to 420 nm at 5 nm intervals, and emission wavelengths from 240 to 600 nm at 2 nm intervals. Ultrapure water was used as the blank to correct for Raman scattering.

Parallel Factor Analysis

Water source discrimination was carried out using 10 indicators: $\text{K}^+ + \text{Na}^+$, Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-} , HCO_3^- , TDS, TOC, UV_{254} and DOM three-dimensional fluorescence spectra. Due to the difference in DOM concentration and chemical composition, each water sample has its unique fluorescence spectrum [35]. The reliability of water source identification using the DOM three-dimensional fluorescence spectrum has also been proved in the team's previous research work [36, 37]. The three-dimensional fluorescence spectra (EEM) of each sample are matrix data. To fully utilize the EEM data, Parallel Factor Analysis (PARAFAC) was used to quantify the fluorescence components in the EEM [38, 39]. PARAFAC decomposes the EEMs data into several latent fluorescent components, and its mathematical expression is given as [40]:

$$X_{ijk} = \sum_{n=1}^N a_{in} b_{jn} c_{kn} + e_{ijk}, i = 1, 2, \dots, I$$

$$j = 1, 2, \dots, J; K = 1, 2, \dots, K \quad (1)$$

Where X_{ijk} is the fluorescence intensity for sample i , at excitation wavelength j , and emission wavelength k ; a_{in} , b_{jn} , c_{kn} are the loading matrices representing the contribution of component n to sample i , excitation wavelength j , and emission wavelength k , respectively. e_{ijk} is the residual error, and N is the number of components.

Principal Component Analysis

This paper utilized the quantified fluorescence components as substitutes for DOM fluorescence spectral data, combined with $\text{Na}^+ + \text{K}^+$, Ca^{2+} , Mg^{2+} , SO_4^{2-} , HCO_3^- , Cl^- , TDS, TOC, and UV_{254} as discriminant

indicators, to construct a dataset for water source discrimination. Principal component analysis (PCA) was employed to analyze the correlations among source identification indicators and reduce redundancy by linearly transforming the original data and extracting the principal features. Its mathematical expression is given as [41, 42]:

$$\begin{cases} F_1 = a_{11}x_1' + a_{12}x_2' + \dots + a_{1q}x_q' \\ F_2 = a_{21}x_1' + a_{22}x_2' + \dots + a_{2q}x_q' \\ \dots \dots \dots \\ F_q = a_{q1}x_1' + a_{q2}x_2' + \dots + a_{q q}x_q' \end{cases} \quad (2)$$

Where x_p' is the standardized value of the p -th hydrochemical parameter for a given sample, $F_1, F_2, \dots, F_m (m \leq q)$ are the transformed variables, mutually uncorrelated. F_j is the linear combination of $X_1, X_2, \dots, X_q$ that captures the largest variance.

Construction of DNN Water Source Discrimination Model

Since neural networks effectively handle the issue of contribution distribution, neural network models are primarily employed. Any neural network with more than one layer can be considered a Deep Neural Network (DNN) [43]. A DNN consists of at least two hidden layers. Taking k inputs, m outputs, and two hidden layers as an example, the basic structure of a DNN is shown in Fig. 3. The left side of Fig. 3 represents the input layer, which contains multiple input nodes $x_1, x_2, \dots, x_k$. These nodes represent the features of the input data, with each feature serving as an input to the network. The input layer is followed by multiple hidden layers, with two layers (L_1 and L_2) shown in Fig. 3. Each node (neuron) in a hidden layer receives the output from all nodes in the previous layer and performs a weighted sum with a set of weights (connections) and biases, which is then passed through an activation function to generate the output of that layer. The right side of Fig. 3. represents the output

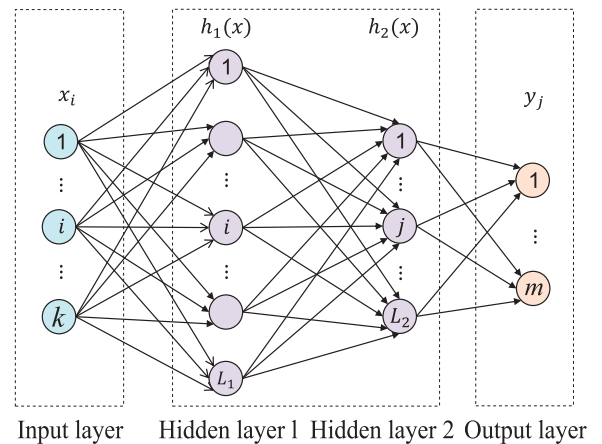


Fig. 3. Basic principle of a deep feedforward neural network.

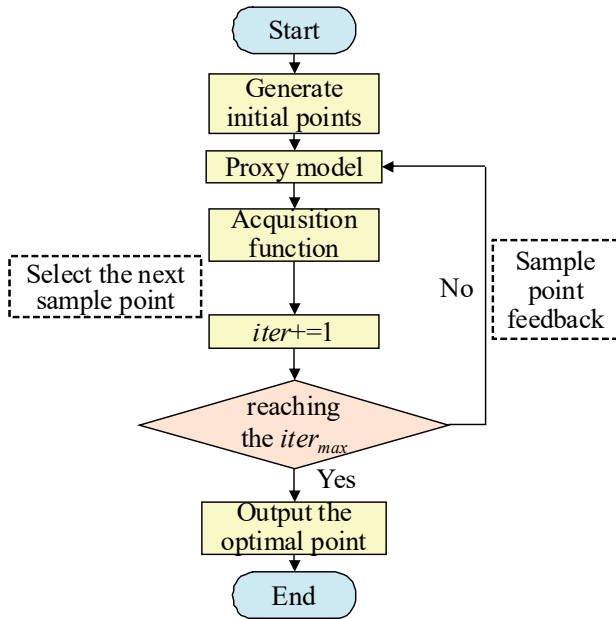


Fig. 4. Bayesian optimization algorithm flow.

layer, denoted as $y_1, y_2, \dots, y_m$. The nodes in the output layer take the output from the last hidden layer, apply a weighted sum and activation function, and produce the network's final output, which typically corresponds to the predicted results.

To enhance the accuracy of the DNN-based classification model and accelerate its training process, Bayesian Optimization (BO) is employed to optimize key hyperparameters, including the learning rate (η), the number of neurons in each hidden layer ($units1$, $units2$, $units3$), and the number of training epochs ($epoch$). In each iteration, BO first uses a surrogate model to predict the performance of each candidate point (η , $units1$, $units2$, $units3$, $epoch$). Then, based on the acquisition function, it selects the most promising point for evaluation and updates the surrogate model using the observed data from this new candidate [44]. This process is repeated iteratively until the maximum number of iterations specified by the BO is reached, as shown in Fig. 4.

Fig. 5 depicts the structure and specific training process of the DNN-based water source discrimination model. The process begins with the DNN on the left, where the first step involves data preparation, including collection, cleaning, normalization, and principal component extraction. Only after these preprocessing steps are completed can the processed data be input into the model for training. Upon completion of data preparation, the model construction phase commences. The network architecture consists of an input layer, three hidden layers, and an output layer. Each hidden layer is comprised of a fully connected layer, an activation function layer, and a Dropout layer. The Dropout layer

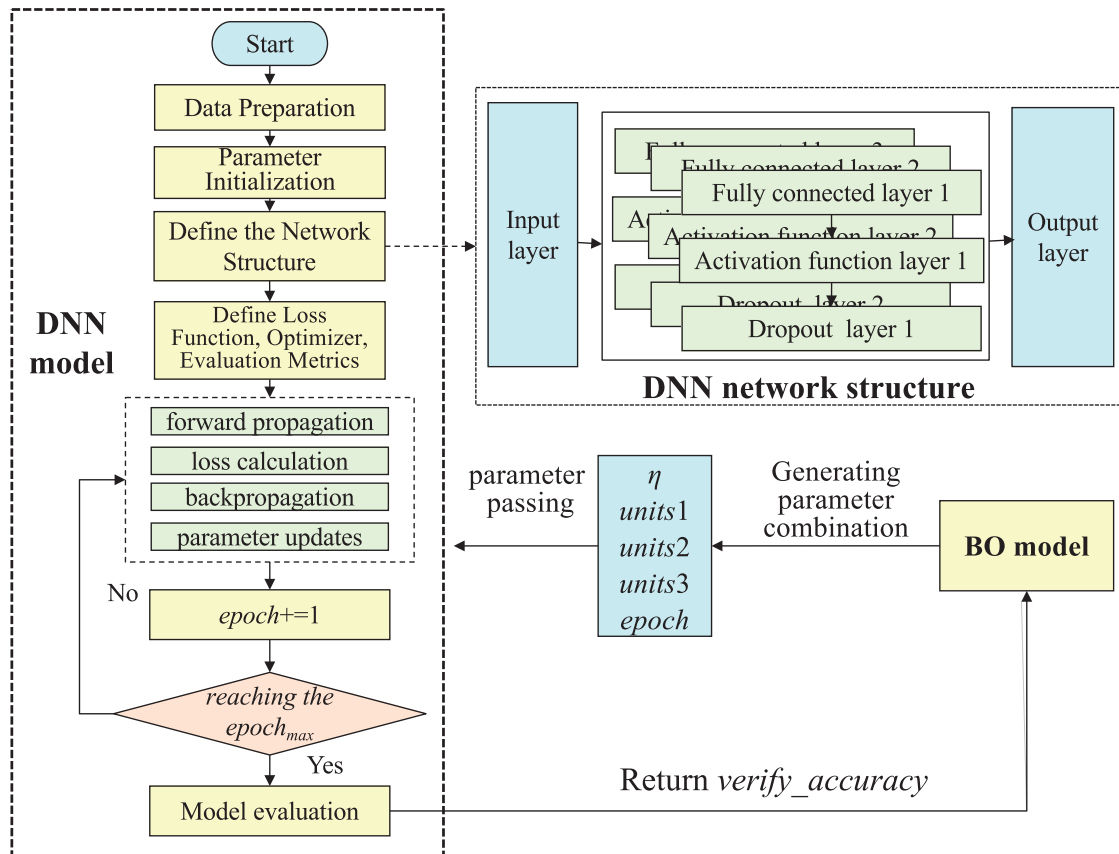


Fig. 5. Flow chart of the DNN water source discrimination model.

is introduced to enhance the network's generalization ability.

The activation function layer employs the Leaky ReLU function to address the vanishing gradient problem in the neural network. The number of neurons in each hidden layer (*units1*, *units2*, *units3*) and the number of training epochs (*epoch*) are hyperparameters to be optimized using the BO. An epoch represents one complete pass of the entire training dataset through both forward and backward propagation operations. The output layer employs a softmax classifier. The model's loss function is the cross-entropy loss function. The optimizer is set to Adam, with the learning rate (η) being another hyperparameter to be optimized. The evaluation metric for the model is the average classification accuracy obtained from 5-fold cross-validation. The model training phase consists of four processes: forward propagation, loss calculation, backpropagation, and parameter updates. The training of the neural network stops once the set number of epochs is reached. The final performance evaluation is performed using 5-fold cross-validation.

The Bayesian model initially generates a random set of hyperparameters (η , *units1*, *units2*, *units3*, *epoch*) and passes them to the DNN. After training the DNN for the specified number of epochs, the model returns the performance (*verify_accuracy*) of the current hyperparameter set to the Bayesian model. Based on this feedback, the Bayesian model generates the next set of hyperparameters. This process is repeated iteratively, with the Bayesian model intelligently adjusting the

hyperparameters based on all previous iterations, until the predefined number of Bayesian iterations (*iter*) is reached.

After determining the optimal hyperparameter combination using the Bayesian Optimization (BO) algorithm, the parameters (weights and biases) of the neural network are fixed. The performance of the model is then evaluated using the validation set, which is partitioned through cross-validation. The accuracy of the validation set (*verify_accuracy*) serves as a direct quantitative measure of the model's generalization ability. This process ensures the establishment of a DNN model that is well-suited to the data while avoiding overfitting.

Results and Discussion

Hydrochemical Characteristics

Table 1 presents the statistical values of the water sample test results. A balance test for anions and cations was conducted on the test results, and the relative error E ranged from 0.0382% to 4.81%, which was less than 5% and thus can be used for subsequent research.

A Piper diagram was plotted using the measured hydrochemical data, as shown in Fig. 6. In the Quaternary aquifer samples, Ca^{2+} is the dominant cation, while HCO_3^- is the dominant anion. The water type is $\text{HCO}_3\text{-Ca}$, which is similar to that of atmospheric precipitation. Water samples from the Luohe Formation

Table 1. Test result statistics values.

Source	Statistical values	Index mass concentration/(mg·L ⁻¹)								UV ₂₅₄ /(cm·L ⁻¹)
		K ⁺ +Na ⁺	Ca ²⁺	Mg ²⁺	Cl ⁻	SO ₄ ²⁻	HCO ₃ ⁻	TDS	TOC	
Quaternary	Minimum	7.56	27.81	6.05	1.74	3.76	133.02	184.28	0.28	0.001
	Maximum	19.56	88.43	15.52	12.47	50.63	265.00	410.613	3.55	0.071
	Average	13.52	51.50	10.77	5.90	18.44	196.45	305.50	0.94	0.016
Luohe Formation	Minimum	13.22	21.15	5.00	4.79	7.98	123.70	186.33	0.06	0.003
	Maximum	57.66	42.08	13.83	20.35	44.03	305.42	439.35	1.95	0.044
	Average	32.83	32.54	8.39	11.14	26.04	190.64	290.78	0.86	0.007
Zhiluo Formation	Minimum	267.54	103.32	16.66	22.35	943.57	34.05	1499	0.11	0.001
	Maximum	1571.27	507.12	59.23	48.00	5001.84	174.96	7765.61	2.01	0.046
	Average	920.22	404.22	40.19	33.62	2956.58	95.88	4458.08	0.81	0.009
Yan'an Formation	Minimum	568.37	167.04	19.86	30.71	1386.42	68.00	1156.46	0.25	0.001
	Maximum	3260.33	477.00	100.00	901.23	5889.45	433.00	9602.22	6.01	0.748
	Average	1499.04	360.32	52.29	155.62	3610.49	178.17	5823.84	1.68	0.05
Mine water	Minimum	552.47	182.76	29.44	28.32	1958.37	61.26	1386.79	0.57	0.002
	Maximum	1428.36	506.23	81.83	184.38	3875.92	208.37	6094.85	4.33	0.112
	Average	996.32	387.51	46.63	52.35	2989.67	104.65	4530.39	1.28	0.024

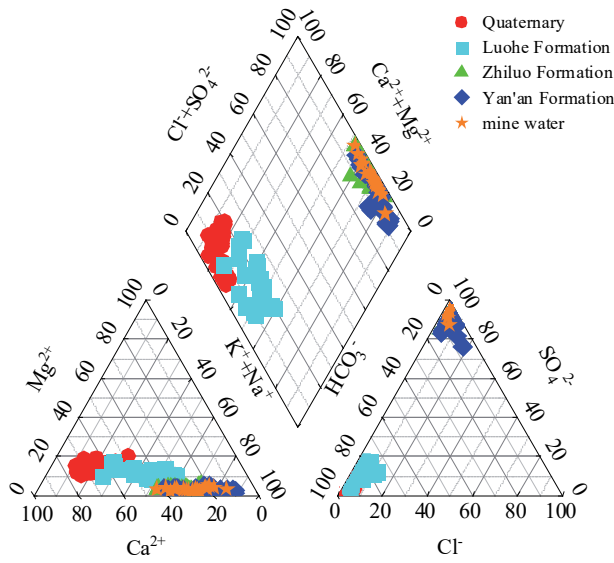


Fig. 6. Three-line diagram of water chemistry Piper.

are mainly of the $\text{HCO}_3\text{-Ca}\cdot\text{Na}$ and $\text{HCO}_3\text{-Na}\cdot\text{Ca}$ types, and are relatively similar to those from the Quaternary. Samples from the Zhiluo and Yan'an formations plot in the upper right of the diamond field, and the main hydrochemical types are $\text{SO}_4\text{-Na}$ and $\text{SO}_4\text{-Na}\cdot\text{Ca}$. The mine water is also dominated by $\text{SO}_4\text{-Na}$ and $\text{SO}_4\text{-Na}\cdot\text{Ca}$ types, showing strong similarity to the Zhiluo and Yan'an formations.

Boxplots of TOC and UV_{254} concentrations for each water sample are shown in Fig. 7. The Quaternary samples exhibited slightly higher organic matter content than those from the Luohe and Zhiluo Formations, likely due to direct recharge from surface water and greater exposure to organic-rich sediments and anthropogenic pollution. The lower TOC and UV_{254} levels in the Luohe and Zhiluo samples indicate minimal human impact.

Samples from the Yan'an Formation showed higher TOC and UV_{254} concentrations, suggesting the dissolution of organic matter from coal-bearing strata. TOC and UV_{254} levels in mine water samples fall between those of the Zhiluo and Yan'an Formations, indicating the presence of some dissolved organic matter as well.

PARAFAC identified three components, as shown in Fig. 8. Based on the fluorescence region classification [45], Component 1 (C1) consists of hydrophobic organic acids and humic acid-like substances, Component 2 (C2) includes tyrosine and tryptophan-containing proteins, and Component 3 (C3) mainly consists of tryptophan and tryptophan-containing proteins. Table 2 shows the average fluorescence intensity of each component for five types of water samples. In this way, the fluorescence intensity of the three components identified by PARAFAC replaces the experimentally measured EEM data, providing a data foundation for the model training.

Analysis of hydrochemical characteristics among different water sample types revealed indicator differences, but these alone were insufficient to accurately identify their sources. Besides, the obtained hydrochemical data are numerous and complicated. Therefore, it is necessary to develop a deep learning-based model to learn the underlying relationships among hydrochemical indicators and further identify the sources of mine water intrusion.

PCA Analysis of Data Set

K^+Na^+ , Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-} , HCO_3^- , TDS, UV_{254} , TOC, C1, C2, C3 were denoted as $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}$, and X_{12} . A total of 177 groundwater samples were selected as training samples, resulting in a dataset X_i with dimensions of 177×12 . The results of the KMO and Bartlett's tests for dataset X_i are presented in Table 3. As shown, the KMO value, approximate chi-square,

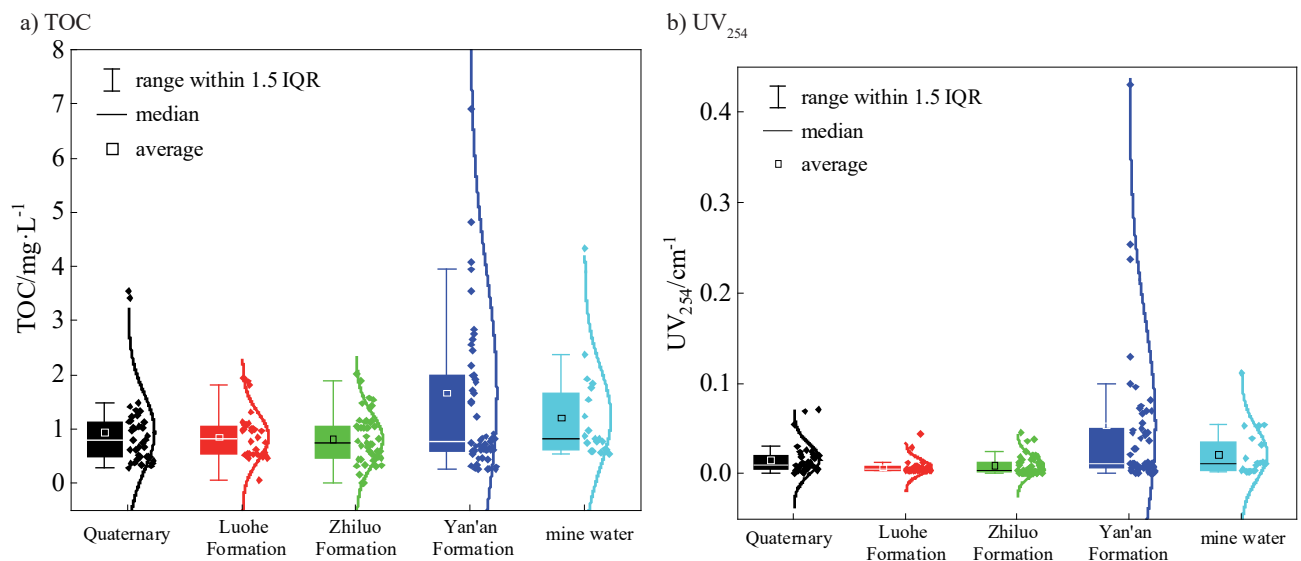


Fig. 7. Box diagram of TOC and UV_{254} concentration in each water sample.

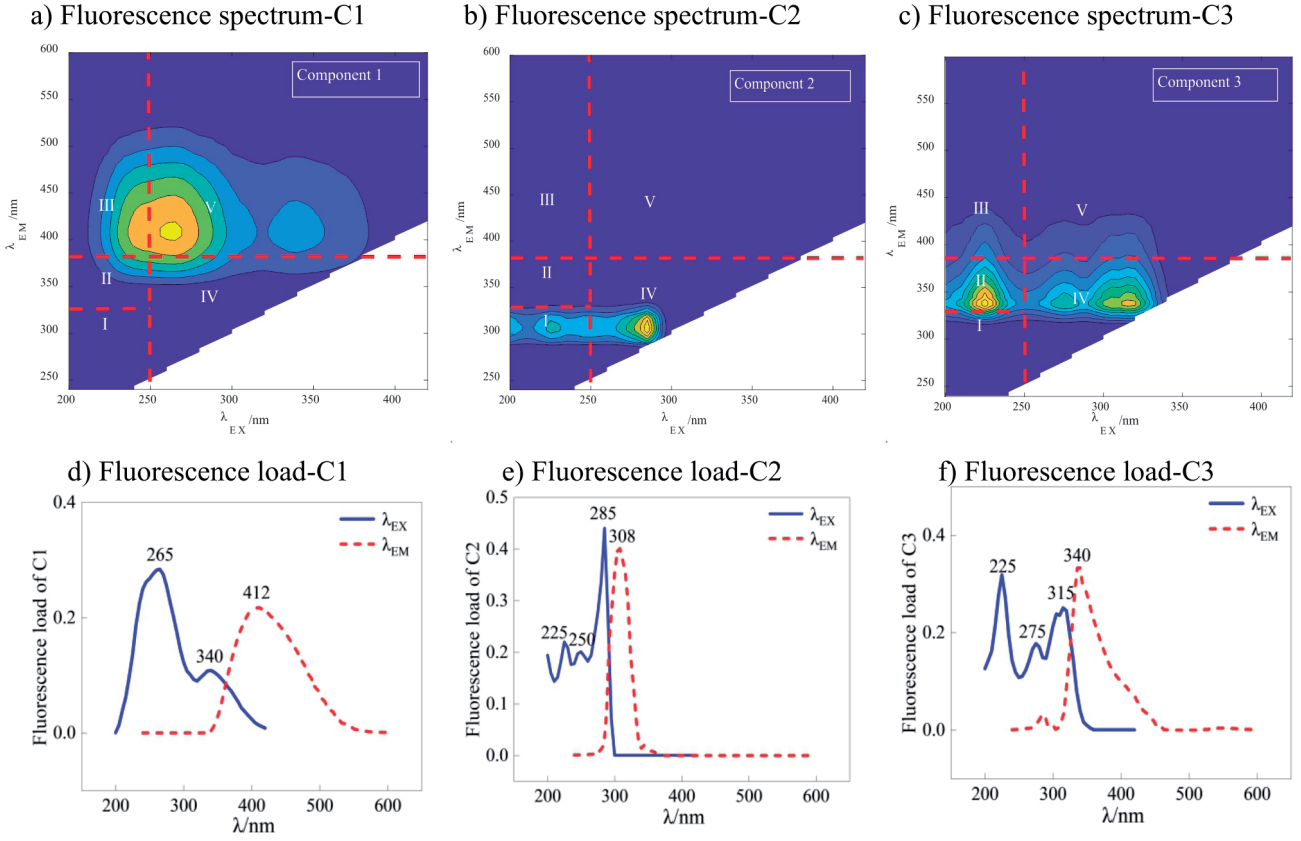


Fig. 8. Three-dimensional fluorescence spectra of DOM components.

Table 2. Average fluorescence intensity of each component.

Source	C1	C2	C3
Quaternary	254.98	981.783	320.54
Luohe	250.80	306.34	242.33
Zhiluo	224.61	265.59	252.23
Yanan	321.54	350.46	331.27
Mine water	308.16	375.32	383.92

degrees of freedom, and significance level are 0.729, 2477.966, 66, and 0, respectively. Therefore, it is concluded that there is sufficient correlation among the variables in dataset X_I , making it suitable for principal component analysis. The total variance explanation of X_I is shown in Table 4. As can be seen, the cumulative contribution rate of the first six principal components (F_1 to F_6) exceeds 90%, encompassing the majority of the information in X_I . Therefore, six principal components

were selected, resulting in a dataset X_2 with dimensions of 177×6 . Dataset X_2 was then used to train the DNN water source discrimination model.

DNN Water Source Discrimination Model

Model Training

The dimensionality-reduced dataset X_2 was employed to train the DNN model for water source discrimination. For the DNN model utilized in this paper, the hyperparameters requiring optimization are (η , $units1$, $units2$, $units3$, $epoch$). These parameters define the domain space of the DNN model, and their respective value ranges are presented in Table 5. The BO algorithm was employed as the search algorithm, with the average accuracy of 5-fold cross-validation during DNN model training serving as the evaluation metric. Additionally, the dropout rate and batch size were manually specified as hyperparameters, set to 0.5 and 32, respectively. These values are commonly adopted in practice and were not

Table 3. KMO and Bartlett test results.

KMO	Bartlett		
	Approximate chi-square	Degree of freedom	Significance
0.729	2477.966	66	0.000

Table 4. Total variance interpretation.

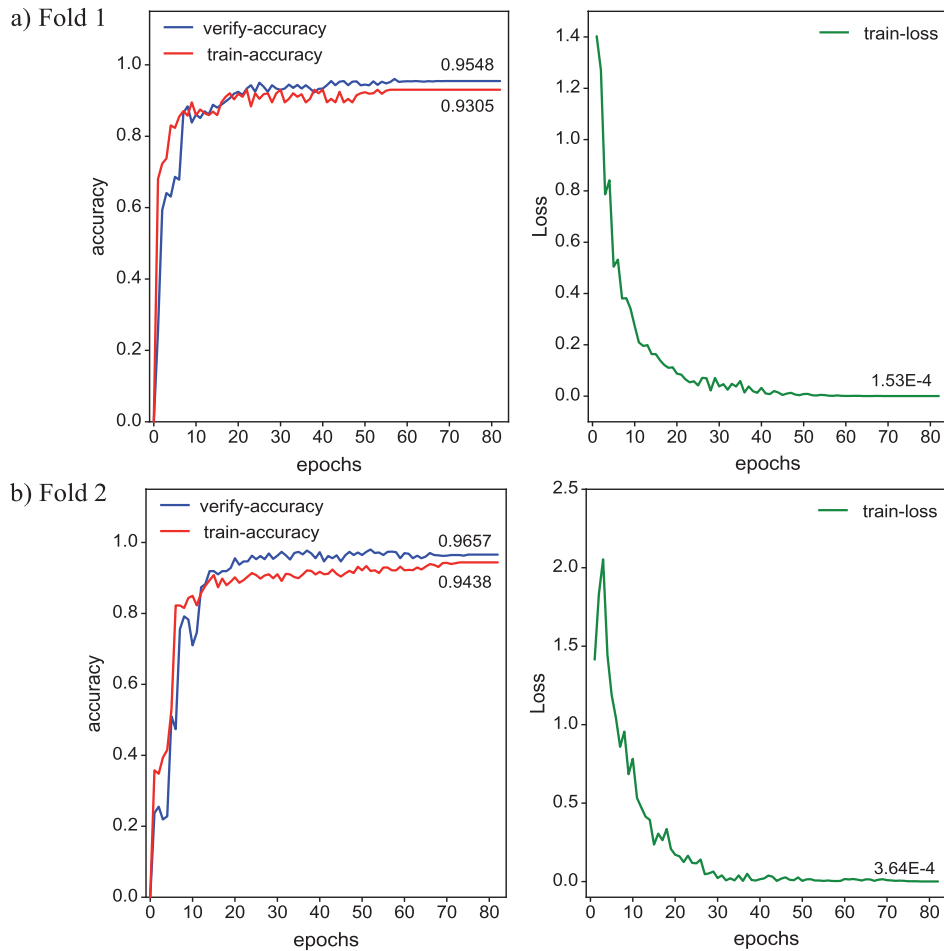
Component	Eigenvalue	Variance percentage	Cumulative contribution rate %
F_1	5.091	42.424	42.424
F_2	2.091	17.426	59.849
F_3	1.553	12.938	72.787
F_4	1.264	10.532	83.319
F_5	0.741	6.177	89.496
F_6	0.504	4.202	93.698

Table 5. Hyperparametric domain space of DNN.

Hyperparameter	Domain space
Learning rate (η)	(0.001,0.1)
Number of neurons 1 (<i>units1</i>)	(64,256)
Number of neurons 2 (<i>units2</i>)	(32,64)
Number of neurons 3 (<i>units3</i>)	(16,32)
Number of training rounds (<i>epoch</i>)	(50,120)
Dropout rate (<i>dropout</i>)	0.5
Batch size (<i>batch size</i>)	32
Number of Bayesian iterations (<i>iter</i>)	50

subjected to further adjustment. The number of iterations for the Bayesian optimization algorithm was set to 50.

After 50 iterations, the Bayesian optimization algorithm identified the optimal parameter combination as (η , *units1*, *units2*, *units3*, *epoch*) = (0.0844, 142, 64, 19, 82). Under this parameter combination, the model achieved an average accuracy of 96.31%, surpassing the expected threshold of 95%. Therefore, it is concluded that 50 iterations are sufficient to meet the training requirements of the model. Fig. 9 illustrates the training accuracy, validation accuracy, and training loss for each fold of the 5-fold cross-validation under this parameter combination, plotted as a function of epochs for the DNN model.



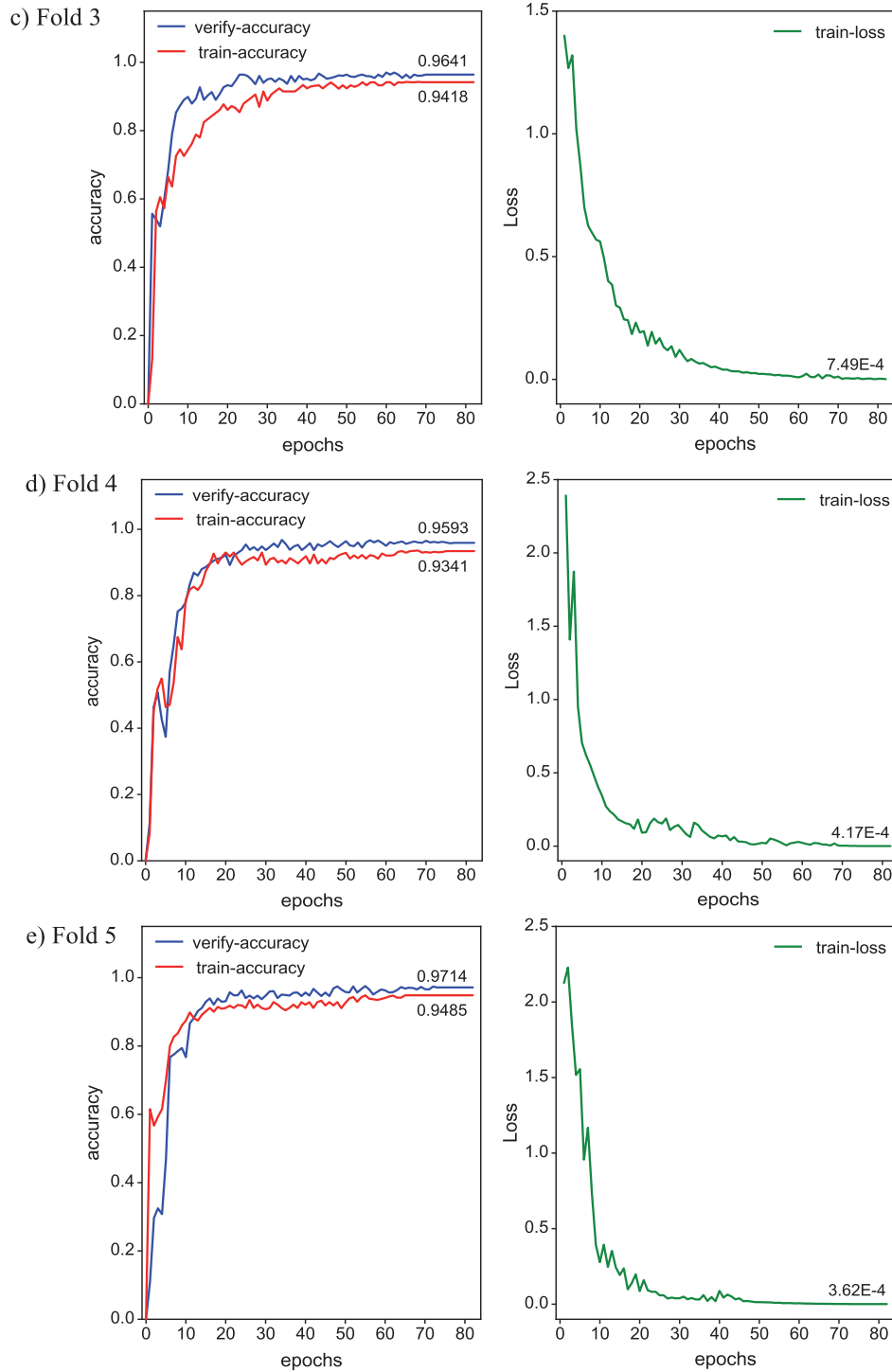


Fig. 9. Epochs trajectory of the deep neural network.

The trends observed in Fig. 9a) to Fig. 9e) are similar. Here, Fig. 9a), corresponding to the first fold, is taken as an example for analysis.

The left half of Fig. 9a) depicts the changes in the accuracy of the DNN model on the training set (red line) and the validation set (blue line). Both accuracy curves exhibited a sharp initial increase, followed by a gradual slowdown in growth as training progressed, eventually stabilizing at 93.05% for the training set and 95.48% for the validation set. This trend indicates

strong generalization capabilities of the model. The right half depicts the training loss over epochs, showing a rapid initial decline followed by a gradual stabilization around 0.000153. This low loss indicates near-optimal performance on the training set.

By synthesizing the epoch trajectories in Figs. 9a) to 9e), it is evident that the model exhibits strong learning and generalization capabilities, with no signs of overfitting.

Table 6. 5-fold cross-validation performance.

Folds	Accuracy	Precision	Recall	f1_score
Fold 1	95.48%	95.02%	95.31%	95.34%
Fold 2	96.57%	96.24%	96.25%	96.19%
Fold 3	96.41%	96.12%	96.21%	95.98%
Fold 4	95.93%	95.87%	95.72%	95.82%
Fold 5	97.14%	97.47%	97.14%	97.08%
average	96.31%	96.14%	96.13%	96.08%

Table 6 presents the performance of the validation set for each fold during cross-validation. As shown in Table 6, the average values of accuracy, precision, recall, and F-measure for the validation set across the 5-fold cross-validation were 96.31%, 96.14%, 96.13%, and 96.08%, respectively. The model exhibited consistent performance across all folds, demonstrating robust stability. An accuracy of 96.31% indicates that the model maintains highly accurate predictions across different data subsets. The high precision, recall, and F-measure values indicate the balanced performance of the model across all categories. These results demonstrate the model's excellent performance during cross-validation, suggesting its reliability for discriminating water sources in coal seam roof water inflow.

Model Testing

To intuitively analyze the fitting performance of the DNN water source discrimination model on the training set, the model was applied to classify 177 water samples used in training. Only one misclassification occurred, indicating that the model has effectively learned the features of the training data.

To verify the generalization ability of the DNN water source discrimination model on new samples, the model

was used to classify the mine water from the 20 test samples mentioned earlier (Table 7), with comparisons made to SVM, LR, RF, and AFSA-RF models from earlier work [34]. In Table 7, the mine water samples numbered 1-5 were respectively taken from the water exploration boreholes HF2-1, HF7-2, YS5-2, DQ2-1, and DH1-3 in Mine A, with the final borehole positions located in the middle-lower section of the Zhiluo Formation aquifer (ZL). Mine water samples numbered 15-20 were respectively taken from the water exploration boreholes ZJ2, ZJ3, ZJ6, ZJ8, and ZJ9 in Mine B. The boreholes penetrated the No. 2 coal seam at a depth of 2-3 m, with the final borehole positions located in the Yanan Formation aquifer (YA). Mine water samples numbered 6-8 were collected from the drainage points in the auxiliary vertical shaft tunnels of Mine B. Samples numbered 9-10 were taken from the water inflow points in the excavation tunnels of Mine B. Samples numbered 11-12 were collected from the water inflow points in the excavation tunnels of Mine A. Samples numbered 13-14 were obtained from the drainage points in the auxiliary vertical shaft tunnels of Mine A.

During the construction of the exploration and drainage boreholes, casing pipes were installed to ensure effective isolation between the borehole walls and the surrounding aquifer, thereby guaranteeing

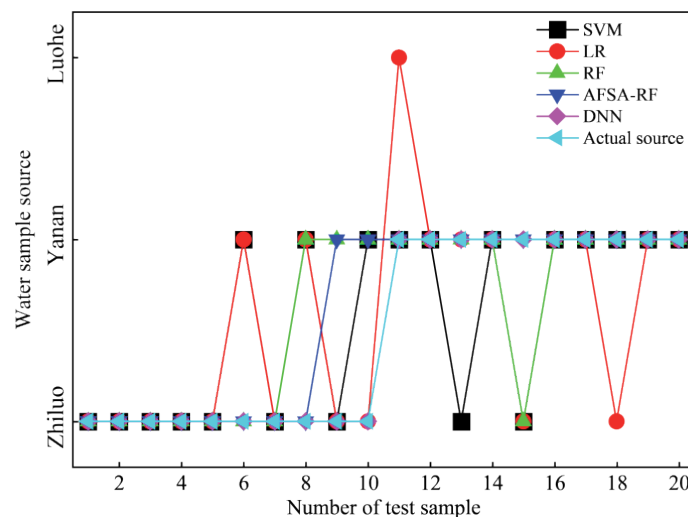


Fig. 10. Discriminant results of mine water sample.

the singularity of water samples at the terminal layer. Mine B primarily mines the No. 2 coal seam. In the previously mined areas, there is essentially no effective aquitard between the roof of the No. 2 coal seam and the Zhiluo Formation aquifer. The mining-induced water-conducting fracture zone has developed and vertically connected the aquifer, allowing water from the Zhiluo Formation to flow into the tunnels through the mining-induced fractures. During this process, a small amount of water from the Yan'an Formation was mixed in, but since the aquifer of the Yan'an Formation has poor water yield, its influence can be neglected. During the production process at Mine A, coal seam mining induced roof fracture zones that directly connect the fissure water-bearing layers from the 2nd coal seam roof to the 3-1 coal seam floor. This aquifer section serves as the direct water source for Mine A. Based on the previous hydrogeological exploration data, on-site construction conditions, and the sampling point locations, it is concluded that water samples 1-10 are from the Zhiluo Formation aquifer, while samples 11-20 are from the Yan'an Formation aquifer.

The discrimination results are shown in Fig. 10. The SVM, LR, RF, and AFSA-RF models misclassified

5, 5, 4, and 2 samples, respectively, while the DNN water source discrimination model's predictions were fully consistent with field observations. This demonstrates its high reliability in practical applications.

To evaluate the overall performance of different discrimination models, the cross-validation performance, misclassifications in the 177-sample training set, and mine water misclassifications of the five discrimination models-DNN, AFSA-RF, RF, LR, and SVM are listed in Table 8, ranked from highest to lowest accuracy.

As shown in Table 8, the DNN model outperformed the others, with its average cross-validation accuracy, precision, recall, and F1 score significantly higher than those of the three shallow machine learning discrimination models. Additionally, the DNN model has no misclassifications in the mine water (testing set). In comparison, the AFSA-RF, RF, LR, and SVM models exhibit slightly lower accuracy and performance metrics, with a notable number of misclassifications observed in both the training and test sets. While AFSA-RF enhanced the performance of RF to some extent, its accuracy still requires further enhancement. This demonstrates the robustness of DNN models, providing a novel methodology for efficient and accurate

Table 7. Test results and sources of mine water samples (test set).

Number	Index mass concentration/(mg·L ⁻¹)								UV ₂₅₄ /(cm·L ⁻¹)	Fluorescence intensity/(R.U.)			Source
	K ⁺ +Na ⁺	Ca ²⁺	Mg ²⁺	Cl ⁻	SO ₄ ²⁻	HCO ₃ ⁻	TDS	TOC		C1	C2	C3	
1	650.23	451.98	45.48	38.21	2458.92	87.86	3832.45	0.873	0.003	354.56	163.28	174.32	ZL
2	849.05	475.46	50.42	39.04	2868.93	94.53	4288.75	1.236	0.005	6.74	264.35	213.43	ZL
3	790.82	506.23	29.56	38.43	2813.24	92.16	4235.12	0.749	0.053	325.68	150.04	170.32	ZL
4	552.47	453.21	33.83	28.76	2263.25	61.26	3397.47	0.975	0.017	41.51	344.38	483.86	ZL
5	912.65	421.26	38.53	36.21	3105.63	93.26	5021.42	0.583	0.002	302.16	243.54	156.32	ZL
6	1033.45	457.86	42.35	30.68	3387.65	86.53	5045.64	0.597	0.002	308.12	251.27	156.74	ZL
7	1237.86	463.24	41.25	39.05	3625.84	82.36	5489.43	0.816	0.002	111.24	492.31	576.27	ZL
8	1383.41	477.84	46.53	39.02	3875.92	142.64	5784.21	0.763	0.003	86.21	283.47	274.32	ZL
9	1168.27	473.27	43.52	36.05	3864.57	106.87	5432.84	1.926	0.039	904.27	186.52	206.43	ZL
10	1314.67	482.28	45.28	39.56	3803.96	145.68	5765.76	0.825	0.003	84.48	274.52	267.64	ZL
11	586.41	343.58	31.92	42.46	2039.48	108.32	1386.79	1.542	0.004	70.25	412.58	230.46	YA
12	752.75	272.45	29.44	28.32	2402.46	65.21	3603.12	0.568	0.009	367.47	662.15	580.84	YA
13	648.46	316.58	31.87	76.12	1958.37	115.89	3166.59	0.762	0.012	134.56	165.32	112.78	YA
14	1171.19	302.15	81.83	64.75	3010.18	137.85	5541.27	2.382	0.053	923.12	768.46	1074.78	YA
15	1370.72	311.58	74.48	63.84	3152.68	89.53	5142.56	4.328	0.112	336.75	264.78	305.67	YA
16	1428.36	400.27	70.55	109.23	3800.36	129.87	6094.85	0.632	0.012	75.67	257.48	351.29	YA
17	1214.57	182.76	48.95	184.38	2742.55	208.37	4531.02	1.754	0.032	254.75	302.69	351.06	YA
18	791.84	297.62	36.57	35.28	2431.25	73.18	3623.86	0.608	0.013	376.46	673.49	594.82	YA
19	1259.41	368.34	75.68	44.47	3711.23	98.58	5593.48	1.842	0.054	734.58	664.32	796.73	YA
20	1259.41	368.34	75.68	44.47	3711.23	98.58	5593.48	1.842	0.054	734.58	664.32	796.73	YA

Table 8. Performance comparison of the water source discrimination model.

Models	Average accuracy	Average precision	Average recall	Average fl_score	Training set misjudgment	Mine water misjudgment
DNN	96.31	96.14	96.13	96.08	1	0
AFSA-RF	91.29	91.38	90.59	91.05	5	2
RF	85.06	88.47	86.31	85.74	9	4
LR	83.26	87.52	85.79	84.96	21	5
SVM	73.48	75.65	79.69	75.13	27	5

identification of water sources during coal seam roof water outburst events.

Conclusions

Through comprehensive approaches including geological data analysis, in-situ water sampling, laboratory chemical experiments, and statistical analysis, we analyzed the hydrochemical characteristics of groundwater in the study area, constructed a dataset of water source indicators, and developed a deep neural network (DNN) model for source discrimination, which was subsequently validated for reliability.

(1) The hydrochemical types of water samples from the Quaternary and Luohe formations are $\text{HCO}_3\text{-Ca}(\text{Na})$ and $\text{HCO}_3\text{-Na}(\text{Ca})$, while those from the Zhiluo, Yan'an formations, and mine water are $\text{SO}_4\text{-Na}(\text{Ca})$. Three fluorescent components were extracted and quantified from the three-dimensional using parallel factor analysis (PARAFAC). Significant differences are observed among the water types in terms of UV_{254} absorbance, TOC concentrations, and fluorescent regions of dissolved organic matter, with mine water exhibiting relatively more complex organic compositions. The organic matter content in the Yan'an Formation and mine water samples is relatively high, which is related to the dissolution of organic matter from the coal seams.

(2) The initial dataset constructed from 12 water quality parameters (Na^+K^+ , Ca^{2+} , Mg^{2+} , SO_4^{2-} , HCO_3^- , Cl^- , TDS, TOC, UV_{254} , C1, C2, C3) exhibited significant information redundancy. Principal component analysis (PCA) was subsequently employed for dimensionality reduction, effectively compressing the dataset from 12 dimensions to 6 principal components. This optimized dataset served as the fundamental input for establishing the water source discrimination model.

(3) The Bayesian optimization algorithm was employed to optimize the hyperparameters of the deep neural network (DNN), yielding the optimal parameter combination: $(\eta, \text{units1}, \text{units2}, \text{units3}, \text{epoch}) = (0.0844, 142, 64, 19, 82)$. With this parameter set, the DNN model achieved a cross-validation accuracy of 96.31%, representing improvements of 4.92%, 10.95%, 12.95%, and 22.73% over the SVM, LR, RF, and AFSA-RF models from previous studies, respectively. And the

DNN model correctly identified all 20 test samples. These results demonstrate that the DNN model can provide a new approach for accurate discrimination of mine water sources.

This paper developed a water source identification model using deep learning, which effectively captured the intrinsic patterns within complex, high-dimensional hydrochemical data while minimizing the influence of human bias. Unlike previous studies that relied mainly on traditional inorganic indicators such as the six major ions or solely on DOM fluorescence spectra, this paper integrated both inorganic and organic indicators for comprehensive analysis. The model was validated using field water samples from three representative coal mines in western China, demonstrating its robustness and reliability. Although the validation was conducted at the regional scale, the proposed approach shows good generalizability. Future research will incorporate data from diverse hydrogeological settings to further expand the applicability of the model.

Acknowledgments

This research was financially supported by the National Natural Science Foundation of China (52204262, 52374255), the Shaanxi province innovation ability support plan project (2024RS-CXTD-44), and the Shaanxi Outstanding Youth Science Fund Project (2025JC-JCQN-012). The authors would like to thank the editors and anonymous reviewers for their comments and suggestions, which greatly helped us to improve this manuscript. The authors are grateful to the anonymous reviewers for their helpful comments on the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. DONG S.N., ZHOU W.F., LIU Q.S., WANG H., JI Y.D. Methods and Techniques for Preventing and Mitigating Water Hazards in Mines. Springer International Publishing: Cham, Switzerland, 2021.

2. LI Y., PAN S.Q., NING S.Z., SHAO L.Y., JING Z.H., WANG Z.S. Coal measure metallogeny: Metallogenic system and implication for resource and environment. *Science China-Earth Sciences*. **66** (7), 1211, **2022**.
3. ZENG Y.F., PANG Z.Z., WU Q., LIAN H.Q., DU X. Roof Water Disaster in Coal Mining in Ecologically Fragile Mining Areas: Formation Mechanism and Prevention and Control Measures. Springer International Publishing: Cham, Switzerland, **2023**.
4. DONG S.N., WANG H., GUO X., ZHOU Z.F. Characteristics of Water Hazards in China's Coal Mines: A Review. *Mine Water and the Environment*. **40** (2), 325, **2021**.
5. WU Q., MU W.P., XING Y. Source discrimination of mine water inrush using multiple methods: A case study from the Beiyangzhuang Mine in northern China. *Bulletin of Engineering Geology and the Environment*. **78** (1), 469, **2019**.
6. DONG D.L., ZHANG J.L. Discrimination Methods of Mine Inrush Water Source. *Water*. **15** (18), 3237, **2023**.
7. JI Y., YU L.J., WEI Z.L., DING J., DONG D.L. Research Progress on Identification of Mine Water Inrush Sources: A Visual Analysis Perspective. *Mine Water and the Environment*. **44** (1), 3, **2025**.
8. LI C., GUI H.R., CHEN J.Y., GUO Y., LI J., HAO Y., XU J.Y. Hydrogeochemical Characteristics of Limestone Water in Taiyuan Formation of Hengyuan Coal Mine and Its Response to High Pressure Grouting Project. *Polish Journal of Environmental Studies*. **31** (6), 5095, **2022**.
9. ZENG Y.F., MEI A.S., WU Q., HUA Z.L., ZHAO D., DU X., WANG L., LV Y., PAN X. Source discrimination of mine water inflow or inrush using hydrochemical field and hydrodynamic field tracer simulation coupling. *Journal of China Coal Society*. **47** (12), 4482, **2022**.
10. 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. *Environmental Earth Sciences*. **80** (4), 144, **2021**.
11. ZHENG Q.S., WANG C.F., YANG Y., LIU W.T., ZHU Y. Identification of mine water sources using a multi-dimensional ion-causative nonlinear algorithmic model. *Scientific Reports*. **14** (1), 3305, **2024**.
12. WANG M., ZHANG J.G., WANG X.Y., ZHANG B., YANG Z.W. Source Discrimination of Mine Water by Applying the Multilayer Perceptron Neural Network (MLP) Method – A Case Study in the Pingdingshan Coalfield. *Water*. **15** (19), 3398, **2023**.
13. WEI Z.L., DONG D.L., JI Y., DING J., YU L.J. Source Discrimination of Mine Water Inrush Using Multiple Combinations of an Improved Support Vector Machine Model. *Mine Water and the Environment*. **41** (4), 1106, **2022**.
14. LING J.R., FU Z.B., XUE K.L. Rapid identification model of mine water inrush source using random forest optimized by multi-strategy improved sparrow search algorithm. *Heliyon*. **10** (15), e43523, **2024**.
15. TAN P.N., STEINBACH M., KUMAR V. Introduction to Data Mining. Addison Wesley, Pearson Education, Inc: London, Britain, **2006**.
16. GRÜNE L., LARS G. Overcoming the curse of dimensionality for approximating Lyapunov functions with deep neural networks under a small-gain condition. *Ifac Papersonline*. **54** (9), 317, **2021**.
17. MAVAIÉ P., HOLDER L., SKINNER MICHAEL K. Hybrid deep learning approach to improve classification of low-volume high-dimensional data. *Bmc Bioinformatics*. **24** (1), 419, **2023**.
18. YASUHIKO S. Deep Learning from Scratch. O'Reilly Japan, Inc.: Tokyo, Japan. **2016**.
19. MA H., ZHAO W.J., LI F.H., YAN H.H., LIU Y.H. Study on Remote Sensing Image Classification of Oasis Area Based on ENVI Deep Learning. *Polish Journal of Environmental Studies*. **32** (3), 2231, **2023**.
20. YANG C., ZHU S.Y., YAO C.W., XIAO S.J. Analysis of hydrochemical evolution in main discharge aquifers under mining disturbance and water source identification. *Environmental Science and Pollution Research*. **28** (21), 26784, **2021**.
21. CHEN Y., TANG L.S., ZHU S.Y. Comprehensive study on identification of water inrush sources from deep mining roadway. *Environmental Science and Pollution Research*. **29** (13), 19608, **2022**.
22. JIANG C.L., ZHU S.J., HU H., AN S.K., SU W., CHEN X., LI C., ZHENG L.G. Deep learning model based on big data for water source discrimination in an underground multiaquifer coal mine. *Bulletin of Engineering Geology and the Environment*. **81** (1), 26, **2021**.
23. JIANG Q.L., LIU Q.M., LIU Y., CHAI H.C., ZHU J.Z. Groundwater chemical characteristic analysis and water source identification model study in Gubei coal mine, Northern Anhui Province, China. *Heliyon*. **10** (5), e26925, **2024**.
24. FANG B. Method for Quickly Identifying Mine Water Inrush Using Convolutional Neural Network in Coal Mine Safety Mining. *Wireless Personal Communications*. **127** (2), 945, **2022**.
25. YAN P.C., ZHAO Y.T., LI G.D., WANG J.B., WANG W.C. Water source identification in mines combining LIF technology and ResNet. *Journal of Mountain Science*. **20** (11), 3392, **2023**.
26. VIDNEROVÁ P., KALINA J. Multi-objective Bayesian Optimization for Neural Architecture Search. 21st International Conference on Artificial Intelligence and Soft Computing (ICAISC): Zakopane, Poland, **2023**.
27. SYLVAIN G., CÉLINE A., LAURENT B., MICHAEL V.G., MICHAEL N., PASCAL M. To Estimate or to Forecast? Lessons From a Comparative Analysis of Four Bayesian Fitting Methods Based on Nonparametric Models. *Therapeutic Drug Monitoring*. **43** (4), 461, **2021**.
28. CASANOVA R.H., CONDE A., PÉREZ C.S. Hour-Ahead Photovoltaic Power Prediction Combining BiLSTM and Bayesian Optimization Algorithm, with Bootstrap Resampling for Interval Predictions Sensors (Basel, Switzerland). **24** (3), 882, **2024**.
29. WANG M., LU X.K., ZHOU Y.D. Gun Life Prediction Model Based on Bayesian Optimization CNN-LSTM. *Integrated Ferroelectrics*. **228** (1), 107, **2022**.
30. WANG Q.M., DONG S.N., WANG H., YANG J., HUANG H., DONG X.L., YU B. Hydrogeochemical processes and groundwater quality assessment for different aquifers in the Caojiatan coal mine of Ordos Basin, northwestern China. *Environmental Earth Sciences*. **79** (9), 199, **2020**.
31. ZENG Y., PANG Z., WU Q., HUA Z., LV Y., WANG L., ZHANG Y., DU X., LIU S. Study of Water-Controlled and Environmentally Friendly Coal Mining Models in an Ecologically Fragile Area of Northwest China. *Mine Water and the Environment*. **41** (3), 802, **2022**.
32. XUE W.F., HOU E.K., ZHAO X., YE Y., TSANGARATOS P., ILIA I., CHEN W. Discriminant Analysis of Water Inrush Sources in the Weibei Coalfield, Shaanxi Province, China. *Water*. **15** (3), 453, **2023**.

33. SHANG H.B., WANG H., WANG T.T., XUE J.K., ZHOU Z.F. Infiltration Characteristics of Surface Water in Coal Seam Mining Beneath Gullies and Corresponding Preventive Measures. *Polish Journal of Environmental Studies*. **34** (4), 3779, **2025**.
34. WANG H., SUN J.Q., ZENG Y.F., SHANG H.B., WANG T.T., QIAO W. An intelligent water source discrimination method for water intrushes from coal seam roofs in the Inner Mongolia-Shaanxi border region. *Coal Geology & Exploration*. **52** (4), 76, **2024**.
35. HAO C.M., HE K.K., LIU C., WANG Y.T., GUI H.R. Contrasting Spectroscopic Characterization and Environmental Behavior of Dissolved Organic Matter in Shallow and Deep Mine Water of Shendong Mining Area, China. *Polish Journal of Environmental Studies*. **32** (5), 4011, **2023**.
36. YANG J., LIU J., JIN D.W., WANG Q.M. Method of determining mine water intrush source based on combination of organic-inorganic water chemistry. *Journal of China Coal Society*. **43** (10), 2886, **2018**.
37. SUN J.Q., WANG H., YANG J., SHANG H.B., WANG T.T., QIAO W. Application of inorganic-organic comprehensive index in identifying water intrush source of coal seam roof. *Safety in Coal Mines*. **54** (12), 182, **2023**.
38. CHEN X.D., GAO L.M., CHEN X.Q., PANG Z.D., GE J., ZHANG H.Q., ZHANG Z., MU M., QIU Y.H., ZHAO X.L. Spectral Characteristic Changes of Dissolved Organic Matter in Aquatic Systems Under the Influences of Agriculture and Coal Mining. *Polish Journal of Environmental Studies*. **31** (3), 2549, **2022**.
39. MU M., GAO L.M., ZHANG H.Q., GE J., ZHANG Z., QIU Y.H., ZHAO X.L. Characterization of Dissolved Organic Matter Components of Agricultural Waste Sources in Water Bodies Based on EEMs. *Polish Journal of Environmental Studies*. **32** (3), 2261, **2023**.
40. STEDMON C.A., BRO R. Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. *Limnology and Oceanography: Methods*. **6** (11), 572, **2008**.
41. YANG Z.W., LV H., WANG X.Y., YAN H.R., XU Z.F. Classification of Water Source in Coal Mine Based on PCA-GA-ET. *Water*. **15** (10), 1945, **2023**.
42. YU J., CAO Y.T., SHI F., SHI J.G., HOU D.B., HUANG P.J., ZHANG G.X., ZHANG H.J. Detection and Identification of Organic Pollutants in Drinking Water from Fluorescence Spectra Based on Deep Learning Using Convolutional Autoencoder. *Water*. **13** (19), 2633, **2021**.
43. QIU X.P. *Neural Networks and Deep Learning*. Beijing: China Machine Press: Beijing, China, **2020**.
44. LAI Y.C. Application and Effectiveness Evaluation of Bayesian Optimization Algorithm in Hyperparameter Tuning of Machine Learning Models: 2024 International Conference on Power, Electrical Engineering, Electronics and Control, Peec: Athens, Greece, **2024**.
45. CHEN W., WESTERHOFF P., LEENHEER J.A., BOOKSH K. Fluorescence excitation-emission matrix regional integration to quantify spectra for dissolved organic matter. *Environmental Science & Technology*. **37** (24), 5701, **2003**.