Available online at www.sciencedirect.com
ScienceDirect
Trans. Nonferrous Met. Soc. China 33(2023) 2799−2814

Transactions of Nonferrous Metals Society of China
www.tnmsc.cn

Bedrock mapping based on terrain weighted directed graph convolutional network using stream sediment geochemical samplings

Bao-yi ZHANG[1], Man-yi LI[2], Yu-ke HUAN[1], Umair KHAN[1], Li-fang WANG[3], Fan-yun WANG[1]

1. Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring (Ministry of Education), School of Geosciences & Info-Physics, Central South University, Changsha 410083, China;
2. PowerChina Zhongnan Engineering Corporation Limited, Changsha 410014, China;
3. Department of Surveying and Mapping Geography, Hunan Vocational College of Engineering, Changsha 410151, China

Received 10 July 2022; accepted 5 May 2023

Abstract: To explore an efficient strategy for intelligent bedrock mapping that can be applied in the areas with coexisting Quaternary coverages and bedrock outcrops, a graph convolutional network (GCN) was implemented for bedrock classification using stream sediment geochemical samplings in the Chahanwusu River area, Qinghai Province, China. The sampling points were organized into a terrain weighted directed graph (TWDG) using Delaunay triangulation to capture the upstream−downstream relationships among the geochemical sampling points. The experimental results indicate that the semi-supervised GCN models, only using 20% of the labeled sampling points, achieved accuracies of 68.20% and 78.31% in ten-type and five-type bedrock discrimination, respectively. In conclusion, it is feasible to map the bedrock type through the concentrations of elements on the stream sediment geochemical sampling points. The proposed data-driven GCN bedrock classification method not only improves the efficiency of bedrock mapping but also may be applied in a large area.

Key words: graph convolutional network; deep learning; stream sediment geochemical samplings; bedrock mapping; quaternary coverage

1 Introduction

Bedrock mapping is essential in geological surveys, whose results are important basic geological data for hydrogeological, engineering geological, environment geological, oil and gas, and mineral explorations. As important geological information, geochemical characteristics are good indicators for studying lithology, alteration, and mineral resources [1−3]. The geochemical elements in the bedrock migrated from subsurface to surface during complex geological processes [4,5], and the distribution patterns of geochemical concentrations are also complexly associated to bedrocks. Multi-element geochemical data, an important part of regional geoscience data, contain potential information about the lithology of bedrocks [6−8]. Beside the traditional spatial statistical and multifractal methods [9−11], shallow machine learning methods, e.g., support vector machine [12], decision tree [13], random forest [14], and boosting ensemble algorithms [15−17], to a certain extent, have solved the problem of mining complex non-linear relationships between element concentrations and bedrocks [18−20]. However, geochemical data are usually high-dimensional, non-linear, and uncertainly observed data, which are

Corresponding author: Li-fang WANG, Tel: +86-731-88877676, E-mail: csuwlf@139.com;
Fan-yun WANG, E-mail: wfy8397641@163.com

DOI: 10.1016/S1003-6326(23)66299-5

subject to multiple geological factors such as weathering denudation intensity, mineralization extent, climate, vegetation, topography, difference in subsurface media, geological structure, and the distribution of streams. It is necessary to automatically extract more abstract and deeper representations from the geochemical data through deep learning of multi-layer neural networks [21]. Convolutional neural networks (CNNs) have been widely applied in lithology discrimination using rock images [22−24]. GUAN et al [25] proposed a feature fusion convolutional autoencoder (FCAE) to extract and fuse the spatial structural features and compositional relationships of multivariate geochemicals for identifying the geochemical anomalies. ZHANG et al [26] used unsupervised convolutional autoencoder network (CAE) to support CNN modeling for synthesis of multi-geoinformation in data-driven mineral prospectivity mapping. The formation mechanism of stream sediments determines the strong spatial association among geochemical sampling points. If a convolutional neural network is used to identify bedrocks underlying stream sediments, the concentration of each element must first be interpolated into an image. However, the stream sediment geochemical sampling data do not always satisfy the premises and assumptions of various spatial interpolation algorithms, and new uncertainties will be introduced into the interpolated images.

The graph data structure has a powerful ability for expressing topological associations among data. Once the connections among data are established, a graph can be formed. GORI et al [27] introduced the concept of applying neural networks to graphs, extending neural networks to graph-structured data in non-Euclidean spaces. KIPF and WELLING [28] proposed a graph convolutional network (GCN) to convolve on graphs by introducing a simple hierarchical propagation rule and applied it to the semi-supervised classification of graph nodes. Subsequently, the GCN has been widely applied in slope deformation prediction [29], buildings vector cognition [30,31], and land cover mapping [32]. GUAN et al [33] transformed geochemical sampled point data into graphs and introduced graph learning to extract the geochemical patterns.

Since stream sediment geochemical sampling data are closely related point-set and composite systems controlled by streams, gullies, and other topographical features, we chose a graph structure to express the complex spatial correlation among stream sediment geochemical data and applied a GCN to map underlying bedrocks using stream sediment geochemical sampling points in the Chahanwusu River area of Qinghai Province. The correlation among sampling points was expressed as a terrain weighted directed graph (TWDG) based on a Delaunay triangulated network. The magnetic anomalies, faults, streams, ore occurrences, and other geological information were extracted as additional features of GCN-based bedrock mapping. The better-performing GCN-based bedrock discrimination models were applied to mapping the underlying bedrock of the Quaternary coverage, which can provide a supplement to the existing regional geological mapping.

2 Study area and dataset

2.1 Study area

The study area is located in the Chahanwusu River area, Dulan County, Qinghai Province, with an area of 893 km^2, a longitudinal range of 98°15'−98°45' E, and a latitudinal range of 35°50'−36°00'N. The outcropped strata are dominated by clasolites and volcanic rocks with well-developed magmatic rocks, mainly including the Indosinian intermediate-acid intrusive rocks and the Triassic continental intermediate-acid volcanic rocks. The study area is located in the eastern part of the east Kunlun tectonic belt with a relatively complex geological structure and prosperous mineral resources, and its geological sketch map is shown in Fig. 1. The study area is an essential metallogenic belt of precious metals, non-ferrous metals, and ferrous metals in China.

2.2 Dataset

2.2.1 Stream sediment geochemical data

The 1:50000 stream sediment geochemical data in the study area were surveyed using irregular grid sampling by Qinghai Geological Survey Institute in 2008, including 15 kinds of chemical elements, including Au, Sn, Ag, As, Sb, Bi, Co, Cu, La, Pb, Zn, W, Mo, Nb and Cd, whose sampling density is 5−6 km^{-2}. The samples were obtained through multi-pit combination sampling and were mainly collected from the debris materials of the bedrock composition in the catchment area, as well as medium- and coarse-grained sand in the stream sediments. After cleaning up and eliminating missing and outlier geochemical data, 4955 geochemical sampling points were finally obtained, as shown in Fig. 2. The methods used to analyze the concentration of heavy metals include atomic

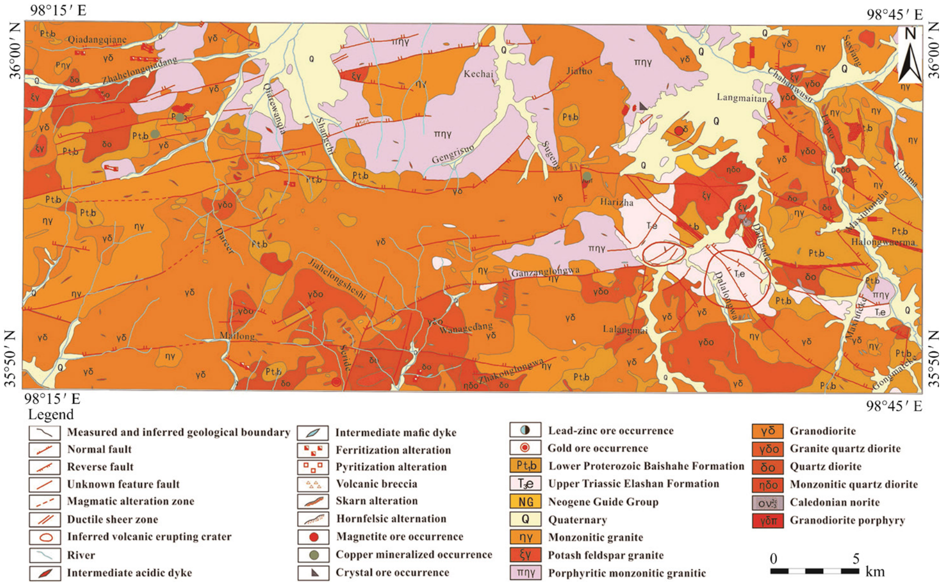

Fig. 1 Geological sketch map of study area (modified from ZHANG et al [34])

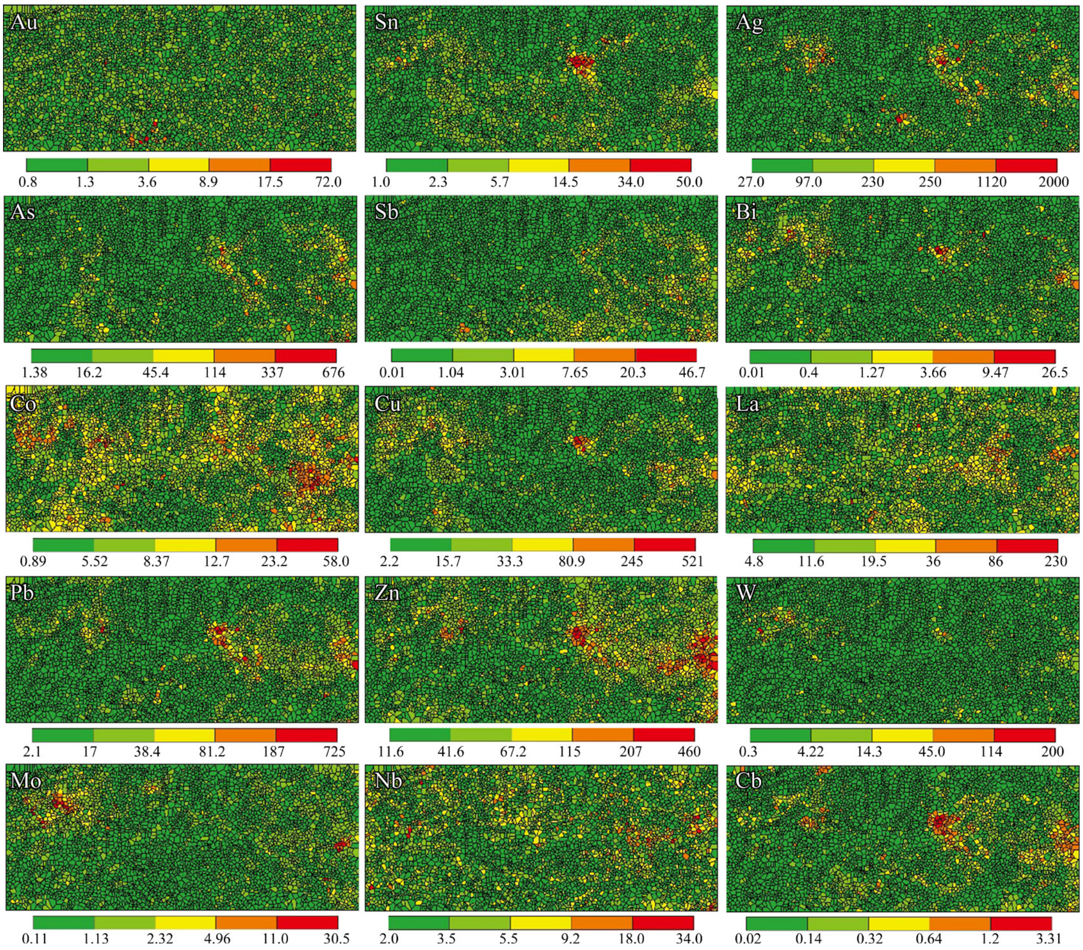

Fig. 2 Voronoi concentration maps of 15 elements (The concentrations of Au and Ag elements are in milligram per tonnage, and those of other elements are in gram per tonnage. The latitude and longitude range of each subfigure is consistent with that in Fig. 1)

emission spectrometry (AES) for Au, Ag and Sn; atomic fluorescence spectrometry (AFS) for As, Sb, and Bi; atomic absorption spectrometry (AAS) for Cu, Pb, Zn, Co, and Ni; polarography (POL) for W and Mo.

The underlying bedrocks of the stream sediment geochemical sampling points were identified according to the geological map of the study area. In geological mapping, areas where bedrocks are not outcropped and bedrock types cannot be determined will be delineated as the quaternary, which are regarded as unknown bedrocks. A total of 849 of the 4955 sampling points were marked as the unknown bedrocks underlying the quaternary coverages. The remaining 4106 sampling points were delineated to 10 bedrock types or 5 merged bedrock types. The specific codes and sampling point number of each bedrock are shown in Table 1.

2.2.2 Topographic features

Topography affects the migration and accumulation of elements in stream sediments. We extracted three topographic features from the digital elevation model (DEM) of the study area, including elevation, slope, and slope of aspect (SOA) (Fig. 3), and incorporated them into the GCN training models. The DEM is derived from global digital elevation public data with a spatial resolution of 30 m from 2009 (http://www.gscloud.cn/). Elevation reflects the elevation above sea level of a sampling point (Fig. 3(a)). Slope is the tangent value of the slope angle, reflecting the degree of the steepness of the ground surface (Fig. 3(b)). SOA refers to the change in the aspect of the ground surface (Fig. 3(c)). The calculation of the slope and the SOA involves the sampling point itself and its neighborhood, which, instead of treating the sampling point as an independent point in space, can reflect the influence of the spatial neighborhood to some extent.

2.2.3 Multivariate geological features

Due to the complexity of the geological system, the bedrock and multiple geological factors lead to the nonlinear and fuzzy geochemical spatial characteristics of stream sediments. Therefore, the additional multivariate geological features, e.g., magnetic anomalies, faults, streams, and mineral occurrences, were extracted and incorporated into the GCN training models. The magnetic anomalies (ΔT) of the 1:50000 high-precision magnetic measurement data in the study area were surveyed by the Qinghai Geological Survey Institute in 2008 (Fig. 4). The Euclidean distance fields of faults (D_{F}), streams (D_{S}), and mineral occurrences (D_{O}) were also constructed (Fig. 5).

2.2.4 Multicollinearity test

Multicollinearity exists in certain relevant or highly correlated feature variables and is often measured by the variance inflation factor (VIF). Generally, there is obvious multicollinearity among variables when the VIF value is greater than 5.0, which will decrease the performance and accuracy

Table 1 Original and merged litho-stratigraphic types in study area

Original			Merged		
Code	Litho-stratigraphic type	Number of sampling points	Code	Litho-stratigraphic type	Number of sampling points
Pt$_1$b	Lower Proterozoic Baishahe formation	435	Pt$_1$b	Lower Proterozoic Baishahe formation	435
T$_3$e	Upper Triassic Elashan formation	157	T$_3$e	Upper Triassic Elashan formation	157
NG	Neogene Guide group	11	NG	Neogene Guide group	11
ηγ	Monzonitic granite	695	γ	Granite	1387
ξγ	Potash feldspar granite	104			
πηγ	Porphyritic monzonitic granite	588			
γδ	Granodiorite	1598	δ	Diorite	2116
γδo	Granite quartz diorite	327			
δo	Quartz diorite	167			
ηδo	Monzonite quartz diorite	24			

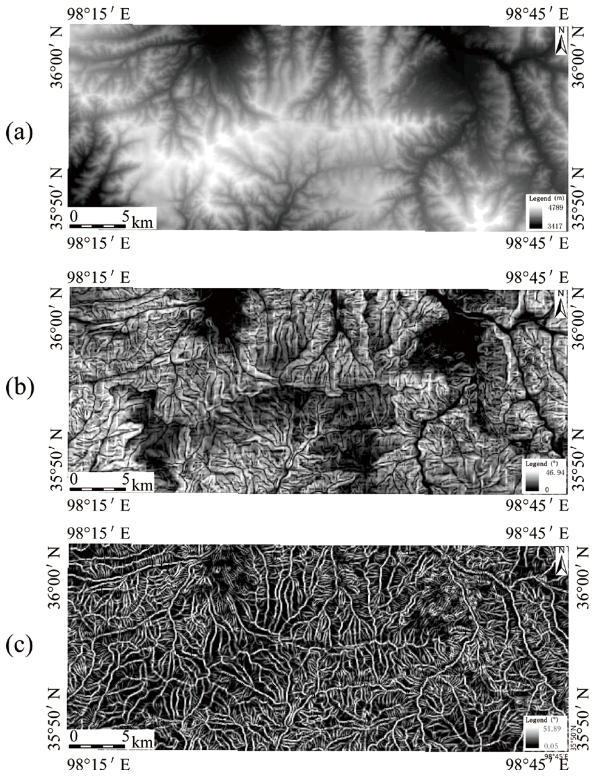

Fig. 3 Three topographic features: (a) Elevation; (b) Slope; (c) SOA

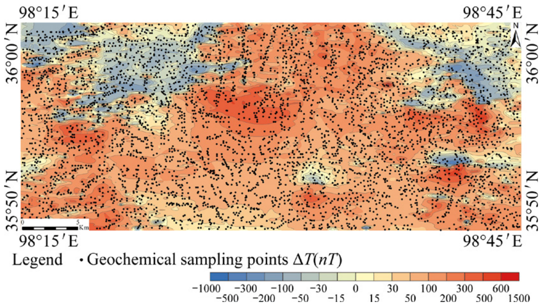

Fig. 4 Distribution of magnetic anomalies (ΔT) in study area

of the model. All feature variables passed the multicollinearity test of the VIF coefficients, which are all less than 5.0 (Table 2).

There are large differences in the values of element concentrations and feature variables of the sampling points, which will lead to inconsistent gradients in the data transmitted by the neural network when using the gradient descent method in model training. Therefore, it will affect the choice of learning rate, making it difficult to optimize. Therefore, in order to unify the dimensions of the data, we performed min–max normalization to transform all the data into values between 0 and 1.

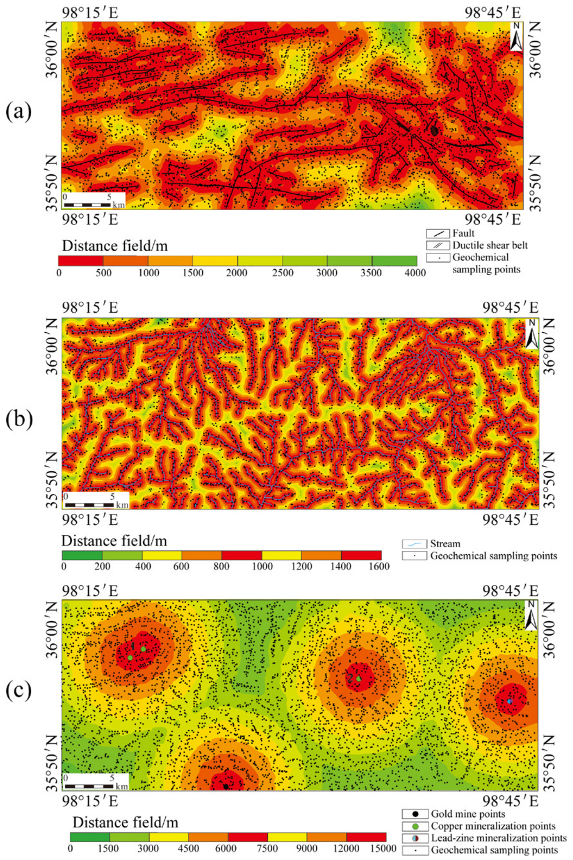

Fig. 5 Distance fields of faults (D_F) (a), streams (D_S) (b), and ore occurrences (D_O) (c)

Table 2 Variance inflation factors (VIFs) of feature variables

No.	Variable	VIF
1	Au	1.016364
2	Sn	2.191803
3	Ag	1.801263
4	As	1.687495
5	Sb	1.282906
6	Bi	1.801990
7	Co	1.561821
8	Cu	2.288033
9	La	1.139138
10	Pb	2.638745
11	Zn	3.305668
12	W	1.372772
13	Mo	1.279554
14	Nb	1.074042
15	Cd	3.238984
16	Slope	1.261176
17	SOA	1.214702
18	Elevation	1.306327
19	ΔT	1.150862
20	D_F	1.366760
21	D_O	1.130898
22	D_S	1.162263

2.2.5 Information gain ratio analysis

Information gain ratio (IGR) is the ratio of node information gain to node split information metric. It is a factor ranking technique that independently assesses the relevance of different factors and detects irrelevant or redundant factors. The method calculates the average merit (AM) of the feature variables, with AM values greater than 0 for feature variables indicating high factor importance and negative or values equal to 0 indicating negative or very low significance. The importance of the feature variables is shown in Table 3, where they are ranked by decreasing average merit. The results show that the AM values of all variables are greater than 0; therefore, all factors can be added to the modeling process.

Table 3 Importance of feature variables

Rank	Variable	AM
1	D_S	0.630504
2	Au	0.228391
3	Ag	0.111122
4	Elevation	0.104490
5	ΔT	0.104490
6	D_O	0.104490
7	D_F	0.104490
8	Slope	0.103844
9	Nb	0.099472
10	La	0.091498
11	Sn	0.082914
12	Cd	0.077532
13	Bi	0.052952
14	W	0.029667
15	Sb	0.025489
16	Mo	0.025051
17	Co	0.024562
18	Pb	0.021454
19	SOA	0.018552
20	Cu	0.017891
21	As	0.016148
22	Zn	0.015450

3 Methods

3.1 Workflow

Combined with topological features and additional features (e.g., magnetic anomalies, faults, streams, and ore occurrences), we constructed a TWDG based on a Delaunay triangulation network of stream sediment geochemical sampling points to achieve GCN-based bedrock discrimination models, as shown in Fig. 6. The edge connections of the Delaunay triangulation network were terrain weighted according to the elevation difference and distance between the node pairs, and different GCN models were trained on this graph structure. We used the precision, accuracy, recall, and F1-score, and a confusion matrix as the evaluation indicators of model performance and generalization abilities to unknown data. Two GCN-based bedrock discrimination models with the highest accuracies in 5 and 10 types of bedrock classification were applied to the mapping of the underlying bedrocks of the Quaternary coverage, respectively.

3.2 Terrain weighted directed graph

Delaunay triangulation is recognized as an optimal triangulation solution and is widely used in discrete data analysis. It uses an undirected graph with a series of connected triangle edges to represent a complex terrain surface, in which scattered sampling points are deemed as the vertexes of triangles. The stream sediment geochemical sampling points are highly correlated, constituting a composite system on the ground surface. Therefore, we used the Delaunay triangulation network with terrain weighted edges to represent the scattered geochemical sampling points as a directed graph. The sampling points are regarded as the vertexes of the graph, and the edges of the triangles are regarded as the links of the graph. The edges of the Delaunay triangulation network are used to express the spatial correlation and its intensity between sampling point pairs. These edge weights are integrated into the GCN model learning as valuable information.

According to the formation mechanism of stream sediments, in mountainous areas with uneven terrain, the solid phase materials falling down the slope either accumulate near the slope foot as a colluvium or directly enter the rivers with the ground surface water. Therefore, the materials are transported from higher locations to lower locations in this process. The link intensity between sampling point pairs depends on the relative length and the angle of slope. When constructing the graph, the DEM of the study area was used to obtain the height of the vertex and the relative height difference between adjacent vertex pairs. To form a directed graph, different edge weights were set according to the height difference and distance between different vertex pairs. As shown in Fig. 7, taking the edge between vertexes *a* and *b* as an example, the height of vertex *a* is less than that of vertex *b*, and the line from vertex *a* to vertex *b* is an upward slope. It is difficult to migrate materials from vertex *a* to vertex *b*, so we assign a small weight for edge *ab*; however, it is downhill from vertex *b* to vertex *a*, so we assign a large weight for edge *ba*. Given that l_{ab} is the distance between vertexes *a* and *b* on the ground surface, d_{ab} is the distance in the two-dimensional planar, and h_{ab} is the height difference, then the edge weights w_{ab} and w_{ba} of edges *ab* and *ba* are calculated by Eq. (1), respectively.

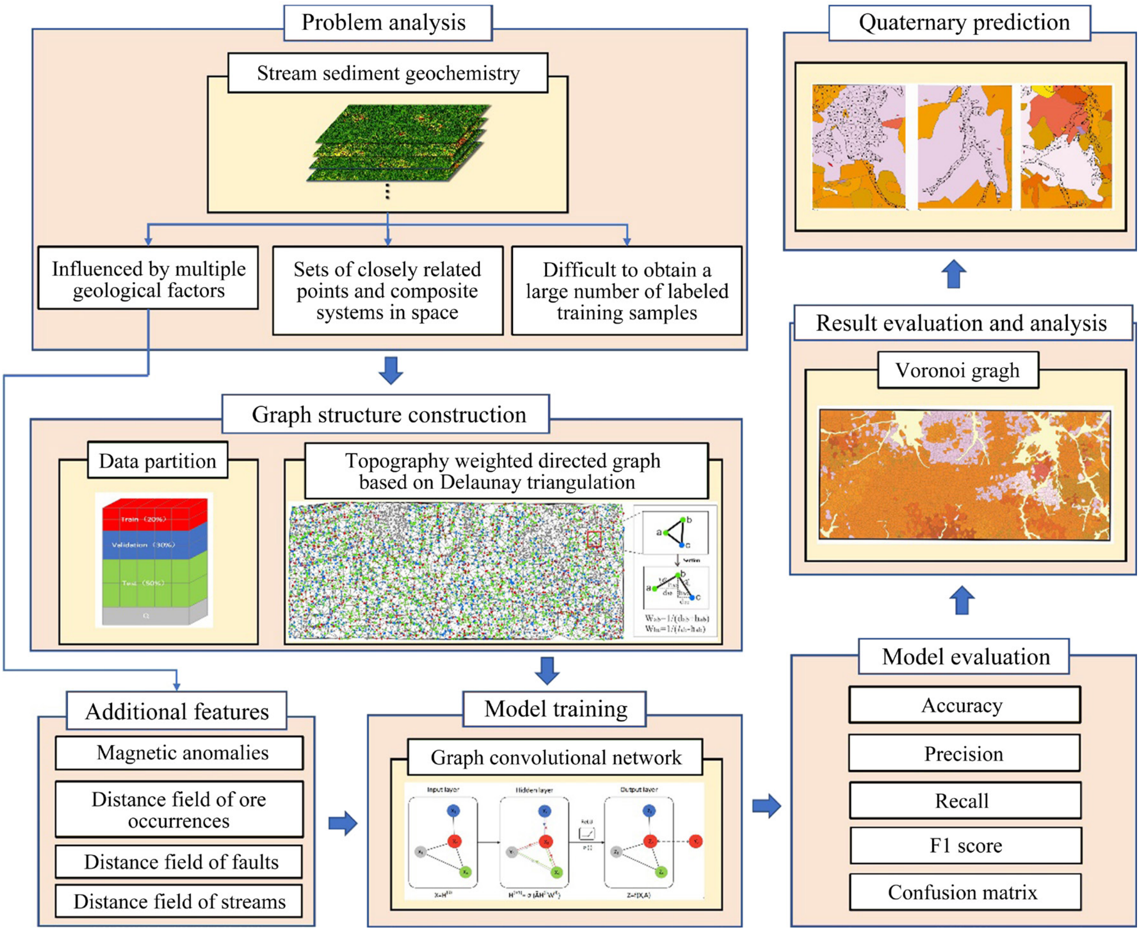
Problem analysis
Stream sediment geochemistry
Influenced by multiple geological factors
Sets of closely related points and composite systems in space
Difficult to obtain a large number of labeled training samples
Graph structure construction
Data partition
Train (20%)
Validation (30%)
Test (50%)
Topography weighted directed graph based on Delaunay triangulation
Additional features
Magnetic anomalies
Distance field of ore occurrences
Distance field of faults
Distance field of streams
Model training
Graph convolutional network
Input layer
Hidden layer
Output layer
Model evaluation
Accuracy
Precision
Recall
F1 score
Confusion matrix
Result evaluation and analysis
Voronoi gragh
Quaternary prediction

Fig. 6 Workflow of GCN-based bedrock mapping underlying stream sediment geochemical samples

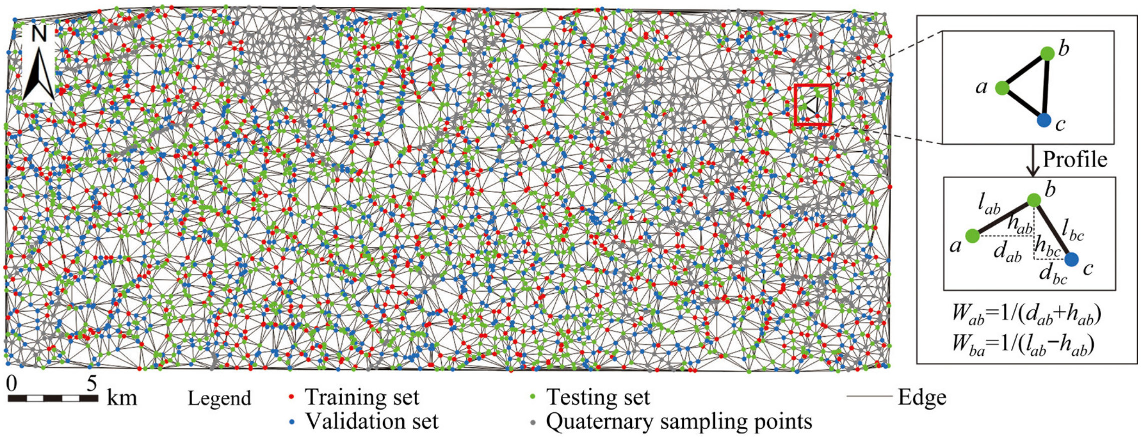

Fig. 7 Terrain weighted directed graph based on Delaunay triangulation of stream sediment geochemical sample points

$$\begin{cases} w_{ab}=\dfrac{1}{d_{ab}+h_{ab}} \\ w_{ba}=\dfrac{1}{l_{ab}+h_{ab}} \end{cases} \tag{1}$$

3.3 Graph convolutional network (GCN)

A graph convolutional network (GCN), proposed by KIPF and WELLING [28], introduces a simple hierarchical propagation rule that can convolve on graphs and the semi-supervised classification of graph structured data. The goal of the GCN model is to obtain the function $f=(\boldsymbol{X},\boldsymbol{A})$ as the output of the vertex set by learning the characteristic signals on a graph $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$, where $\boldsymbol{V}$ and $\boldsymbol{E}$ are the vertex set and the edge set of graph $\boldsymbol{G}$, respectively. Given that the number of nodes in the graph is N, the number of features is C, and the number of label types is F, then the input information of the GCN model includes: (1) feature matrix $\boldsymbol{X}$, which describes the feature attribute of the vertex set, with a dimension of N×C, and (2) adjacency matrix $\boldsymbol{A}$, which describes the adjacent information of vertex pairs in the graph, with the dimension of N×N. Taking a simple two-layer GCN as an example (Fig. 8), to perform semi-supervised vertex classification on a graph with adjacency matrix $\boldsymbol{A}$ (binary or weighted), we first calculated $\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1/2}$ during data preprocessing to symmetrically normalize the adjacency matrix, where $\tilde{\boldsymbol{D}}$ is the degree matrix of the graph, and then obtain the output in the following form:

$$\boldsymbol{Z}=f(\boldsymbol{X},\boldsymbol{A})= \\ \mathrm{softmax}[\hat{\boldsymbol{A}}\mathrm{ReLU}(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^{(0)})\boldsymbol{W}^{(1)}] \tag{2}$$

where $\boldsymbol{W}^{(0)}$ is the weight matrix from the input layer to the hidden layer, with the dimensions C×H, here H is the feature dimension of the hidden layer; $\boldsymbol{W}^{(1)}$ is the weight matrix from the hidden layer to the output layer, with the dimensions of H×F; ReLU is the rectified liner unit activation function; softmax is the normalized exponential function defined as

$$\mathrm{softmax}(x_i)=\frac{\exp(x_i)}{\sum_i \exp(x_i)}.$$

The number of vertexes at each layer is constant in the GCN, and so is the dimension of the adjacency matrix. We can only change the feature expression of each layer vertex. The feature dimension can be adjusted by setting different hidden layer vertexes. During GCN training, the features of adjacent vertexes are continuously convolved into the feature expression of the current vertex. After the first layer convolution, each vertex will contain the features of its directly adjacent vertexes; after the second layer convolution, each vertex will contain the features of its secondary adjacent vertexes, and so on. Therefore, the more the layers of GCN, the wider the receptive domain.

The GCN is a semi-supervised learning network. When classifying the vertexes in the graph, only the labels of a small part of vertexes (e.g., Y_2 in Fig. 8) are needed. The labeled or unlabeled vertexes are indexed, and the feature representation of each vertex can be learned during training. The weights can then be updated by back-propagating the cross-entropy loss of the labeled vertexes. The cross-entropy loss calculation formula is

$$L=-\sum_{l\in y_L}\sum_{f=1}^{F}\boldsymbol{Y}_{lf}\ln\boldsymbol{Z}_{lf} \tag{3}$$

where y_L represents the indexes of the labeled vertexes, F is the feature dimension of the output layer, and $\boldsymbol{Y}_{lf}$ and $\boldsymbol{Z}_{lf}$ represent the label and mode output of the vertex and label type f, respectively.

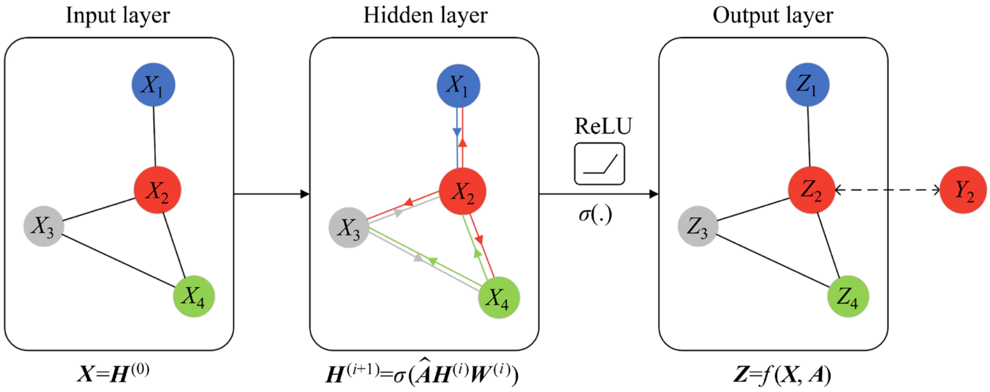
Input layer
Hidden layer
Output layer
X_1
X_2
X_3
X_4
ReLU
$\sigma(.)$
Z_1
Z_2
Z_3
Z_4
Y_2
$\boldsymbol{X}=\boldsymbol{H}^{(0)}$
$\boldsymbol{H}^{(i+1)}=\sigma(\hat{\boldsymbol{A}}\boldsymbol{H}^{(i)}\boldsymbol{W}^{(i)})$
$\boldsymbol{Z}=f(\boldsymbol{X}, \boldsymbol{A})$

Fig. 8 Graph convolutional network (GCN) with two layers ($\boldsymbol{H}^{(i)}$ represents the eigenmatrix of the ith layer)

4 Results

4.1 Model training

We used Nvidia Quadro P2000 GPU and CUDA 11.2 API to train the proposed models as shown in Table 4, and the training and predicting were implemented using the Python package Keras with the Tensorflow as backend.

Table 4 Overview on GPUs and host systems used for training and predicting

GPU	NVIDIA Quadro P2000
fp32 peak [TFLOPS]	3.0
mem-bandw. [GB/s]	140
Bus	PCIe 3.0
CPU	Intel CoffeLake 1×i7-9700F
Cores (Threads)	1×8 (×1)

The lack of labeled data is the main problem when applying machine learning methods to geosciences, and it is usually not possible to obtain sufficient labeled geoscience data. However, as a semi-supervised learning network, GCN only uses labels of a small number of vertexes to classify the vertexes of the graph, which achieves high enough discrimination accuracy. In this study, the main steps of training a GCN discrimination model of bedrocks underlying stream sediment geochemical samplings were as follows.

Firstly, the dataset was divided into four parts, i.e., training set, validation set, testing set, and Quaternary coverage dataset (Q). Among the 4106 sampling points (excluding Q), 20% of the sampling data of each bedrock type were chosen for the training set, which was used in model training, while 30% of the sampling data of each bedrock type were extracted for the validation set for parameter fine-tuning. In contrast, 50% of the sampling data of each bedrock type were extracted for the testing set to evaluate model performance. The optimal classification result was achieved using only 20% of the labelled sampling data, and discriminative accuracy of model was no longer improved even using more labelled data.

Secondly, three input matrices of the GCN model were prepared, i.e., feature matrix, label matrix, and adjacency matrix. We constructed a TWDG as the adjacency matrix based on a Delaunay triangulation network from the stream sediment geochemical sampling points. Multiple geoscience data (i.e., magnetic anomalies, faults, ore occurrences, and streams) were extracted as additional features. Therefore, the feature matrix contains the concentrations of 15 elements, elevation, slope, SOA, magnetic anomalies, the distance field of faults, the distance field of streams, and the distance field of ore occurrences. The label matrix contains 10 or 5 bedrock types. Four models were designed according to different features and label combinations, as shown in Table 5.

Finally, the feature matrix, label matrix and adjacency matrix were input into the two-layer GCN model to obtain the output bedrock labels and to evaluate the models' accuracies. The constructed graph contains the sampling points of the Quaternary coverage. However, this part of the vertexes was only used to obtain the output labels and was not used for the accuracy evaluation. All models were iterated 1000 times (epochs) with a learning rate of 0.01. A Chebyshev polynomial with a depth of 3 was used as the approximate function of the graph signal. There are 34, 38, 41 and 42 hidden layer units in the four models, respectively. The specific parameter settings are shown in Table 6.

Table 5 Design of GCN bedrock discrimination models

Model code	Feature matrix	Adjacency matrix	Label vector
GCN10_19	Element concentrations + topography +ΔT	Weighted triangulation	10 types
GCN5_19	Element concentrations + topography +ΔT	Weighted triangulation	5 types
GCN10_22	Element concentrations + topography + ΔT + geological distance fields	Weighted triangulation	10 types
GCN5_22	Element concentrations + topography + ΔT + geological distance fields	Weighted triangulation	5 types

Table 6 Hyperparameter of GCN models

Parameter	Description	GCN10_19	GCN5_19	GCN10_22	GCN5_22
Learning_rate	Learning rate	0.01	0.01	0.01	0.01
Epoch	Number of training iterations	1000	1000	1000	1000
Hidden1	Number of hidden layer units	34	38	41	42
Dropout	Dropout rate	0.5	0.5	0.5	0.5
Weight_decay	L2 regularization	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
Max_degree	Chebyshev polynomial depth	3	3	3	3

4.2 Model evaluation

4.2.1 10-type bedrock classification

(1) Accuracy

Two trained 10-type bedrock discrimination GCN models were evaluated using accuracy, precision, recall and F1-score. The accuracies of models on the training, validation and testing datasets are presented in Table 5. The highest accuracy of the 10-type GCN bedrock discrimination models is 68.20% in the testing set. In addition, the accuracy of the GCN bedrock discrimination model with a feature dimension of 22, where distance fields (e.g., the distance fields of ore occurrences, faults, and streams) were added to the feature matrix, is higher than that of the model with a feature dimension of 19. The experimental results also show that the ore occurrences, streams, and faults are all essential multiple geological features and have significantly improved the accuracy of the GCN bedrock discrimination model. ZHANG et al [34] adopted four machine learning methods, i.e., decision tree (DT), random forest (RF), extreme gradient boosting (XGB), and light gradient boosting machine (LGBM), to discriminate bedrocks underlying stream sediment geochemical data. We compared the accuracies of GCN models with those of DT, RF, XGB, and LGBM, meanwhile, shallow machine learning methods are all less accurate than GCN10_22 model as shown in Table 7, indicating the higher performance of GCN models.

(2) Confusion matrix

To further explore the classification abilities of the GCN models on different bedrocks, we analyzed the confusion matrix of the GCN10_22 model on the testing dataset (Table 8). In the 10-type bedrock discrimination model GCN10_22, the upper Triassic Elashan formation (T_3e), granodiorite ($\gamma\delta$), and porphyritic monzonitic granite ($\pi\eta\gamma$) obtained higher precisions and recall values and were easy to be distinguished, and their F1-scores, higher than the other types, are 73%, 75%, and 76%, respectively. The low recall values of the Neogene guide group (NG), potassium feldspar granite ($\xi\gamma$), and monzonitic diorite ($\eta\delta o$) are due to the small number of samples and their discontinuous spatial distribution, where the spatial distances between samples were not low enough to predict them correctly.

Table 7 Accuracies of 10-type GCN bedrock classification models

Model code	Training set accuracy/%	Validation set accuracy/%	Testing set accuracy/%
DT10_22	52.80	−	50.97
RF10_22	65.32	−	65.45
XGB10_22	66.08	−	68.00
LGBM10_22	67.13	−	67.40
GCN10_19	65.36	64.07	64.66
GCN10_22	70.75	68.50	68.20

4.2.2 5-type bedrock classification

The highest accuracy of the 5-type GCN bedrock discrimination models is 78.31% in the testing set. The merged 5-type bedrock discrimination models were better than the 10-type bedrock discrimination models. In general, when dealing with multiple classification problems, the more classes that the learner needs to be distinguished, the more difficult it is to classify them. The accuracies of DT, RF, XGB, and LGBM [34] are all less than that of GCN5_22 model as shown in Table 9.

The confusion matrix of the GCN5_22 model on the testing dataset is shown in Table 10. In the

Table 8 Confusion matrix of testing set of GCN10_22 model

Code		Pt_1b	T_3e	NG	γδ	γδο	ηγ	ξγ	πηγ	δο	ηδο
Real class	Pt_1b	86	9	0	90	1	7	2	4	10	0
	T_3e	7	57	0	10	1	1	0	0	0	0
	NG	1	3	2	0	0	0	0	0	0	0
	γδ	10	6	0	655	8	43	2	40	4	0
	γδο	2	1	0	86	62	2	0	4	1	0
	ηγ	13	0	0	67	0	215	1	35	3	0
	ξγ	1	4	2	7	0	12	18	5	1	1
	πηγ	2	1	0	35	1	18	0	226	0	0
	δο	13	0	0	20	12	6	4	0	26	0
	ηδο	0	0	0	4	0	5	1	0	0	2
Total real class		209	76	6	768	158	334	51	283	81	12
Total prediction class		135	81	4	974	85	309	28	314	45	3
Precision		0.64	0.7	0.5	0.67	0.73	0.7	0.64	0.72	0.58	0.67
Recall		0.41	0.75	0.33	0.85	0.39	0.64	0.35	0.8	0.32	0.17
F1-score		0.5	0.73	0.4	0.75	0.51	0.67	0.46	0.76	0.41	0.27
Accuracy		0.68200									

Table 9 Accuracies of 5-type GCN bedrock classification models

Model code	Training set accuracy/%	Validation set accuracy/%	Testing set accuracy/%
DT5_22	67.08	−	63.50
RF5_22	74.26	−	73.36
XGB5_22	78.44	−	77.86
LGBM5_22	78.16	−	77.86
GCN5_19	77.72	76.13	76.74
GCN5_22	80.78	79.18	78.31

Table 10 Confusion matrix of testing set of GCN5_22 model

Code		Pt_1b	T_3e	NG	γ	δ
Real class	Pt_1b	79	10	0	28	92
	T_3e	3	57	0	9	7
	NG	0	4	0	2	0
	γ	20	6	0	565	77
	δ	16	8	0	147	848
Total real class		209	76	6	668	1019
Total prediction class		118	85	0	751	1024
Precision		0.67	0.67	0	0.75	0.83
Recall		0.38	0.75	0	0.85	0.83
F1-score		0.48	0.71	0	0.8	0.83
Accuracy		0.78311				

5-type bedrock discrimination model GCN5_22, the classification accuracies of granite (γ) and diorite (δ) are significantly higher than those of the other types due to the imbalance of the dataset.

4.3 Model validation

We further visualized the bedrock classification results of the well-trained GCN models through a Voronoi diagram. Figures 9 and 10 show the classification visualization results of 10-type and 5-type bedrock discrimination models, respectively. The white and black Voronoi polygons represent correct and incorrect classifications, respectively. The classification results of the GCN models show a certain degree of spatial continuity with clear boundaries between different bedrock types, indicating that the GCN models depend more on the spatial connection between sampling points during classification. Therefore, the adjacent sampling points tend to be predicted as the same type of bedrock. However, the models have weak abilities to distinguish the fragmented distribution patterns of bedrocks. For example, the southwestern part of the study area is dominated by granodiorite (γδ) or diorite (δ) in the 10 or 5 types of bedrocks, respectively, so some fragmented distributed Lower Proterozoic Baishahe formations (Pt_1b) in this area were misclassified.

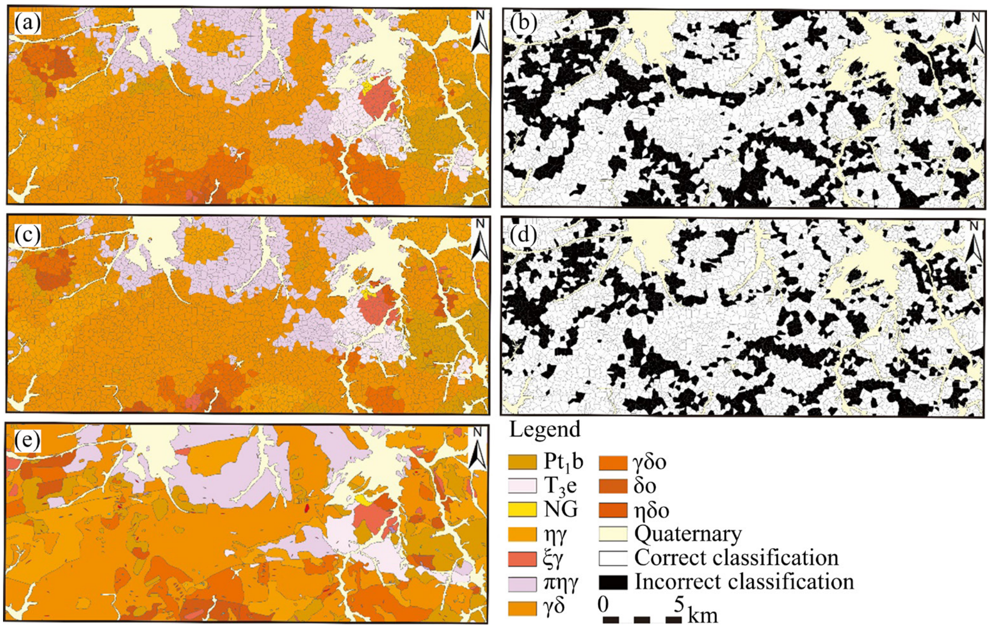

Fig. 9 Classification of 10-type bedrock discrimination models: (a) GCN10_19 classification results; (b) GCN10_19 correct or incorrect classification; (c) GCN10_22 classification result; (d) GCN10_22 correct or incorrect classification; (e) Geologic map

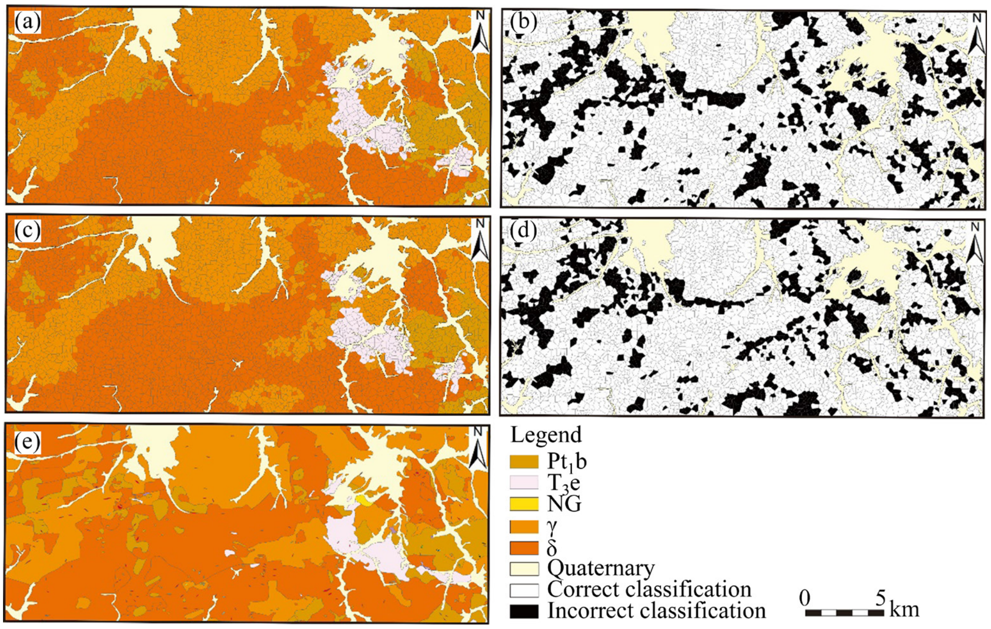

Fig. 10 Classification of 5-type bedrock discrimination models: (a) GCN5_19 classification result; (b) GCN5_19 correct or incorrect classification; (c) GCN5_22 classification result; (d) GCN5_22 correct or incorrect classification; (e) Geologic map

5 Discussion

5.1 Predicting bedrocks underlying quaternary

Traditionally, bedrocks underlying the loose sediments of the Cenozoic Quaternary are not outcropped, and their types need to be ascertained using exploration engineering methods in order to penetrate the Quaternary sediments. In this study, the better-performing GCN bedrock discrimination models (i.e., GCN10_22 and GCN5_22) were used to predict bedrocks underlying the Quaternary coverage. To recognize these bedrocks, we used 4955 sampling points of the study area, including the Quaternary sampling points, to construct a TWDG based on the Delaunay triangulation.

However, these Quaternary sampling points were not involved in the training and testing of the GCN models and only obtained the output labels to predict bedrocks. Figures 11 and 12 are the bedrock prediction results of the GCN10_22 and GCN5_22 models, respectively. The rest of the real bedrocks were compared with the prediction results of bedrock underlying the Quaternary, for which three local areas were selected for detailed demonstration. The prediction results of the GCN models are consistent with the distribution of bedrocks in their neighborhood, and there is a clear boundary between different bedrocks; therefore, the GCN models can correctly classify the bedrocks underlying the Quaternary coverage in the study area.

5.2 Limitation

Instead of considering each sampling point in isolation, we organized the sampling points as a TWDG based on the Delaunay triangulation to express the upstream and downstream relationship between the geochemical sampling points of stream sediments. However, many kinds of graphs can express a spatial correlation of sampling points, and other graphs could be explored in the future. Therefore, it is necessary to deeply analyze the spatial correlation of sampling points and find a more reasonable and effective graph of sampling points to improve the accuracy of model.

The dataset used in this study has a certain imbalance caused by the uneven distribution of bedrocks in the study area, which is also the main reason why many samples were misclassified. Most machine learning algorithms are based on the assumption that the training dataset is balanced. However, it is difficult to obtain a balanced dataset in actual geoscience problems, which is also an objective limitation of geoscience datasets. Therefore, it is necessary to introduce new methods to solve the problem of classification errors caused by dataset imbalances.

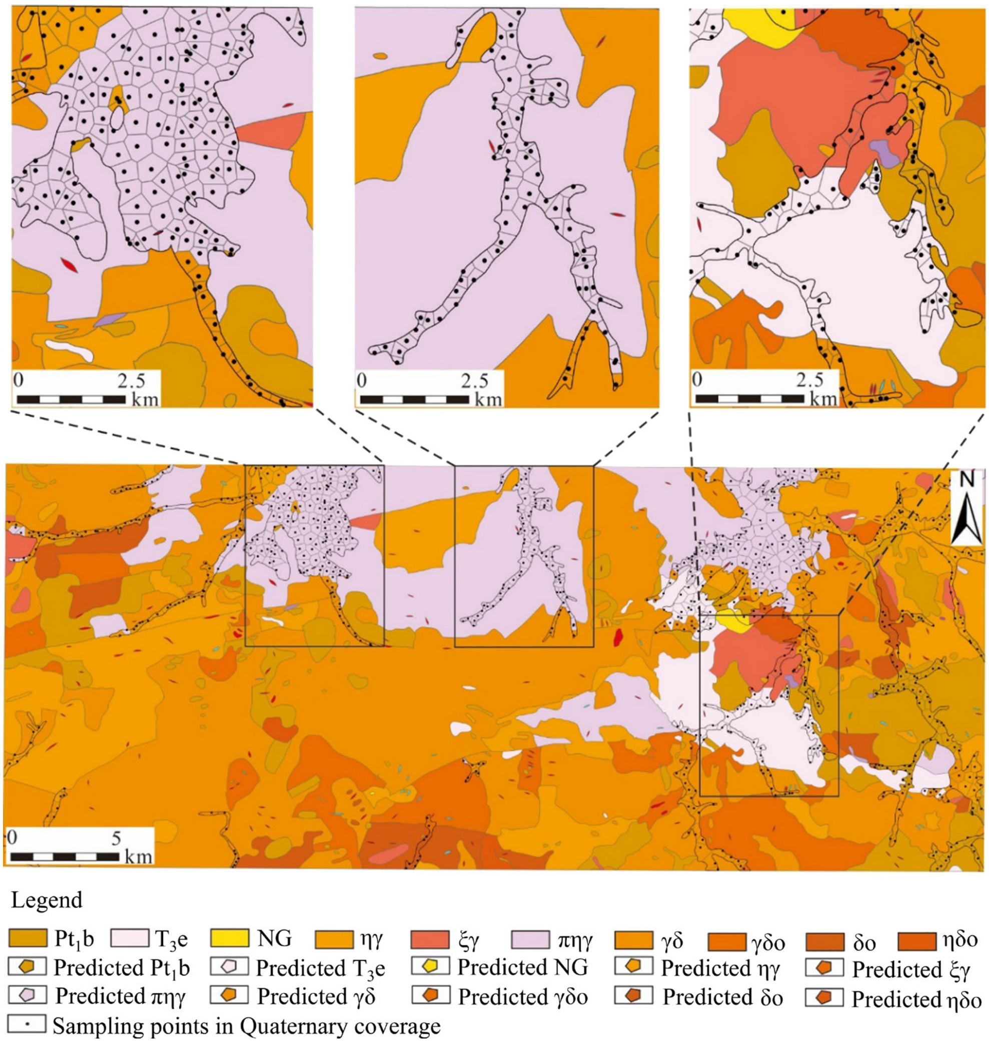

Fig. 11 Predicted bedrocks underlying Quaternary by GCN10_22 model

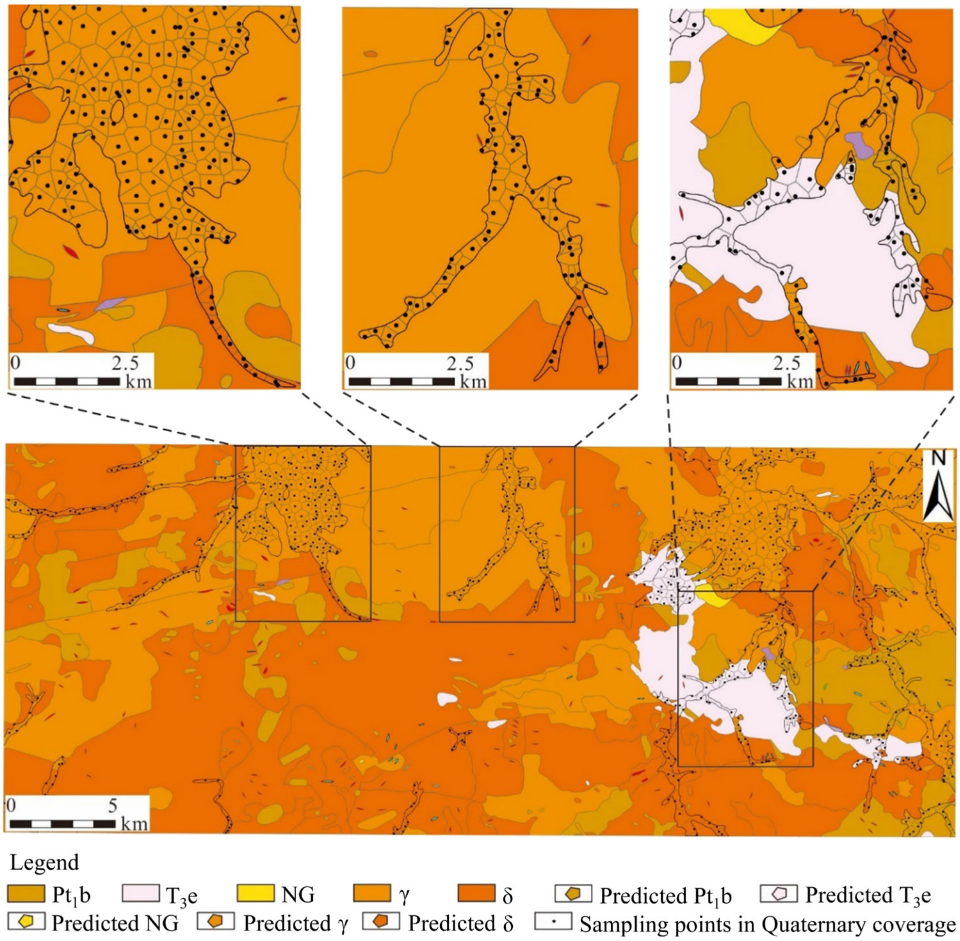

Fig. 12 Predicted bedrocks underlying Quaternary by GCN5_22 model

We compared the accuracies of GCN models with those of DT, RF, XGB and LGBM from ZHANG et al [34], meanwhile, shallow machine learning methods are all less accurate than GCN10_22 or GCN5_22 models, indicating the higher performance of GCN models. ZUO and XU [35] employed graph deep learning algorithms, including graph convolutional networks and graph attention networks, to produce mineral potential maps. In the future, our GCN model should be compared with graph attention networks (GAT) [36], both belonging to graph neural networks (GNN) methods, to further improve the accuracy of bedrock discrimination.

A complete spatial distribution of the bedrocks underlying the Quaternary coverages was predicted through two GCN models with highest accuracies in 5 and 10 types of bedrock classification, respectively. However, the true bedrocks underlying the Quaternary coverage are unknown, so drilling boreholes or geophysical exploration should be arranged to verify the prediction results of the GCN bedrock discrimination models in the future.

6 Conclusions

(1) It is feasible to map bedrocks using the concentrations of elements combined with multiple geoscience features. The semi-supervised GCN bedrock discrimination models only need 20% of the labeled geochemical sampling points to reach accuracies of 68.20% (10 types of bedrocks) and 78.31% (5 types of bedrocks). The intelligent data-driven bedrock discrimination methods can improve efficiency and can be applied in a large area.

(2) The Delaunay triangulation network is an effective tool for processing scattered geochemical sampling points, and scattered points can be considered from a spatial correlation perspective. We constructed a TWDG graph based on Delaunay triangulation to express the upstream and downstream relationship between the geochemical sampling points of stream sediments, where the

sampling points are no longer regarded as independent points.

(3) The GCN bedrock discrimination model can be used to predict the bedrocks underlying Quaternary coverage and explore the complete distribution of bedrocks in the adjacent area. The experimental results show that the predicted bedrocks are consistent with the surrounding bedrocks, and there is a clear boundary between different bedrocks.

Acknowledgments

The authors would like to thank the MapGIS Laboratory Co-constructed by the National Engineering Research Center for the Geographic Information System of China and Central South University for providing the MapGIS® software (Wuhan Zondy Cyber-Tech Co., Ltd., Wuhan, China). We also thank Senior Engineer Professor Shao-ning ZHANG (The 8th Team of Qinghai Provincial Bureau of Nonferrous Metals and Geological Exploration) and Professor Jian-qing LAI (Central South University) for their kind assistance in the area of data collection.

The authors are grateful for the financial supports from the National Natural Science Foundation of China (Nos. 42072326, 41772348), and the National Key Research and Development Program, China (No. 2019YFC1805905).

References

[1] WANG Fan-yun, MAO Xian-cheng, DENG Hao, ZHANG Bao-yi. Manganese potential mapping in western Guangxi-southeastern Yunnan (China) via spatial analysis and modal-adaptive prospectivity modeling [J]. Transactions of Nonferrous Metals Society of China, 2020, 30(4): 1058−1070.

[2] LIU Zhan-kun, MAO Xian-cheng, WANG Fan-yun, TANG Lei, CHEN Guang-huan, CHEN Jin, DENG Hao. Deciphering anomalous Ag enrichment recorded by galena in Dayingezhuang Au(-Ag) deposit, Jiaodong Peninsula, Eastern China [J]. Transactions of Nonferrous Metals Society of China, 2021, 31(12): 3831−3846.

[3] ZHU Da-peng, LI Huan, JIANG Wei-cheng, WANG Chong, HU Xiao-Jun, KONG Hua. Ore-forming environment of Pb-Zn mineralization related to granite porphyry at Huangshaping skarn deposit, Nanling Range, South China [J]. Transactions of Nonferrous Metals Society of China, 2022, 32(9): 3015−3035.

[4] CHENG Qiu-ming. Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas [J]. Journal of Geochemical Exploration, 2012, 122(7): 55−70.

[5] SHAHI H, GHAVAMI R, ROUHANI A K. Detection of deep and blind mineral deposits using new proposed frequency coefficients method in frequency domain of geochemical data [J]. Journal of Geochemical Exploration, 2016, 162(3): 29−39.

[6] ZHANG Bao-yi, CHEN Yi-ru, HUANG An-shuo, LU Hao, CHENG Qiu-ming. Geochemical field and its roles on the 3D prediction of concealed ore-bodies [J]. Acta Petrologica Sinica, 2018, 34(2): 352−362. (in Chinese)

[7] WANG Li-fang, WU Xiang-bin, ZHANG Bao-yi, LI Xue-feng, HUANG An-shuo, MENG Fei, DAI Peng-yao. Recognition of significant surface soil geochemical anomalies via weighted 3D shortest-distance field of subsurface orebodies: A case study in the Hongtoushan copper mine, NE China [J]. Natural Resources Research, 2019, 28(3): 587−607.

[8] ZHANG Bao-yi, JIANG Zheng-wen, CHEN Yi-ru, CHENG Nan-wei, KHAN U, DENG Ji-qiu. Geochemical association rules of elements mined using clustered events of spatial autocorrelation: A case study in the Chahanwusu river area, Qinghai Province, China [J]. Applied Sciences, 2022, 12(4): 2247.

[9] XIE Shu-yun, CHENG Qiu-ming, XING Xi-tao, BAO Zheng-yu, CHEN Zhi-jun. Geochemical multifractal distribution patterns in sediments from ordered streams [J]. Geoderma, 2010, 160(1): 36−46.

[10] SHAHRESTANI S, MOKHTARI A R, ALIPOUR-ASLL M. Assessment of estimated bedrock and stream sediment geochemical backgrounds in catchment basin analysis [J]. Natural Resources Research, 2019, 28(3): 1071−1087.

[11] WU Guo-peng, CHEN Guo-xiong, CHENG Qiu-ming, ZHANG Zhen-jie, YANG Jie. Unsupervised machine learning for lithological mapping using geochemical data in covered areas of Jining, China [J]. Natural Resources Research, 2021, 30(2): 1053−1068.

[12] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273−297.

[13] BREIMAN L, FRIEDMAN J, OLSHEN R A, STONE C J. Classification and regression trees [M]. Routledge, 2017.

[14] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5−32.

[15] CHEN Tian-qi, GUESTRIN C. XGBoost: A scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, UAS: ACM, 2016: 785−794.

[16] KE Guo-lin, MENG Qi , FINLEY T, WANG Tai-feng , CHEN Wei , MA Wei-dong , LIU Tie-Yan. Lightgbm: A highly efficient gradient boosting decision tree [C]// Advances in Neural Information Processing Systems. 2017, 30: 3147−3155.

[17] HUAN Yu-ke, SONG Lei, KHAN U, ZHANG Bao-yi. Stacking ensemble of machine learning methods for landslide susceptibility mapping in Zhangjiajie City, Hunan Province, China [J]. Environmental Earth Sciences, 2023, 82(1): 35.

[18] HARRIS J R, GRUNSKY E C. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data [J]. Computers and Geosciences, 2015, 80(2015): 9−25.

[19] KODIKARA G R L, WOLDAI T. Spectral indices derived,

non-parametric Decision Tree Classification approach to lithological mapping in the Lake Magadi area, Kenya [J]. International Journal of Digital Earth, 2018, 11(10): 1020−1038.

[20] ASANTE-OKYERE S, SHEN C B, ZIGGAH Y Y, RULEGEYA M M, ZHU Xiang-feng. A novel hybrid technique of integrating gradient-boosted machine and clustering algorithms for lithology classification [J]. Natural Resources Research, 2020, 29(4): 2257−2273.

[21] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436−444.

[22] ZHANG Ye, LI Ming-Chao, HAN Shuai. Automatic identification and classification in lithology based on deep learning in rock images [J]. Yanshi Xuebao/Acta Petrologica Sinica, 2018, 34(2): 333−342. (in Chinese)

[23] ALZUBAIDI F, MOSTAGHIMI P, SWIETOJANSKI P, CLARK S R., ARMSTRONG, RT. Automated lithology classification from drill core images using convolutional neural networks [J]. Journal of Petroleum Science and Engineering, 2021, 197: 107933.

[24] DOS ANJOS C E M., AVILA M R V, VASCONCELOS A G P, PEREIRA NETA A M, MEDEIROS L C, EVSUKOFF A G, SURMAS R, LANDAU L. Deep learning for lithological classification of carbonate rock micro-CT images [J]. Computational Geosciences, 2021, 25(3): 971−983.

[25] GUAN Qing-feng, REN Shu-liang, CHEN Li-rong, FENG Bin, YAO Yao. A spatial-compositional feature fusion convolutional autoencoder for multivariate geochemical anomaly recognition [J]. Computers & Geosciences, 2021, 156: 104890.

[26] ZHANG Shuai, CARRANZA E J M., WEI Han-tao, XIAO Ke-yan, YANG Fan, XIANG Jie, XU Yang. Data-driven mineral prospectivity mapping by joint application of unsupervised convolutional auto-encoder network and supervised convolutional neural network [J]. Natural Resources Research, 2021, 30(2): 1011−1031.

[27] GORI M, MONFARDINI G, SCARSELLI F. A new model for learning in graph domains [C]//Proceedings of 2005 IEEE International Joint Conference on Neural Networks. Montreal, Canada: IEEE, 2005: 729−734.

[28] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks [EB/OL]. http://arxiv.org/abs/1609.02907.pdf

[29] MA Zheng-jing, MEI Gang, PREZIOSO E, ZHANG Zhong-jian, XU Neng-xiong. A deep learning approach using graph convolutional networks for slope deformation prediction based on time-series displacement data [J]. Neural Computing and Applications, 2021, 33(21): 14441−14457.

[30] YAN Xiong-feng, AI Ting-hua, YANG Min, TONG Xiao-hua. Graph convolutional autoencoder model for the shape coding and cognition of buildings in maps [J]. International Journal of Geographical Information Science, 2021, 35(3): 490−512.

[31] WEI Shi-qing, JI Shun-ping. Graph convolutional networks for the automated production of building vector maps from aerial images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1−11.

[32] ZHANG Xi-ning, GE Yong, LING Feng, CHEN Jin, CHEN Yue-hong, JIA Yuan-xin. Graph convolutional networks-based super-resolution land cover mapping [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 7667−7681.

[33] GUAN Qing-feng, REN Shu-liang, CHEN Li-rong, YAO Yao, HU Ying, WANG Rui-fan, CHEN Wen-hui. Recognizing multivariate geochemical anomalies related to mineralization by using deep unsupervised graph learning [J]. Natural Resources Research, 2022, 31(5): 2225−2245.

[34] ZHANG Bao-yi, LI Man-yi, LI Wei-xia, JIANG Zheng-wen, KHAN U, WANG Li-fang, and WANG Fan-yun. Machine learning strategies for lithostratigraphic classification based on geochemical sampling data: A case study in area of Chahanwusu River, Qinghai Province, China [J]. Journal of Central South University, 2021, 28(5): 1422−1447.

[35] ZUO Ren-guang, XU Ying. Graph deep learning model for mapping mineral prospectivity [J]. Mathematical Geosciences, 2023, 55(1): 1−21.

[36] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, ROMERO A, LIÒ P, BENGIO, Y. Graph attention networks [EB/OL]. http://arxiv.org/abs/1710.10903.pdf.

基于水系沉积物地球化学采样的
地形加权图卷积网络基岩填图方法

张宝一 [1]，李曼懿 [2]，浣雨柯 [1]，Umair KHAN [1]，王丽芳 [3]，汪凡云 [1]

1. 中南大学 地球科学与信息物理学院 有色金属成矿预测与地质环境监测教育部重点实验室，长沙 410083；

2. 中国电建集团 中南勘测设计研究院有限公司，长沙 410014；

3. 湖南工程职业技术学院 测绘地理学院，长沙 410151

摘　要：为了探索高效的第四系覆盖及露头较少区域的基岩智能填图方法，应用图卷积网络(GCN)对青海省察汗乌苏河地区水系沉积物地球化学采样的下伏基岩进行分类。基于 Delaunay 三角化采样点被组织为一个地形加权的有向图来表达水系沉积物地球化学采样点之间的河流上下游关系。实验结果表明：半监督的 GCN 模型仅使用了 20%的采样点标签，分类精度达到 68.20%(10 类基岩)和 78.31%(5 类基岩)。该方法能有效利用水系沉积物地球化学采样中的元素含量进行基岩填图，且能提高基岩填图的效率并能进行大面积应用。

关键词：图卷积网络；深度学习；水系沉积物地球化学采样；基岩填图；第四系覆盖物

(Edited by Xiang-qun LI)