Geochemistry xxx (xxxx) xxx

Contents lists available at ScienceDirect

Geochemistry

journal homepage: www.elsevier.com/locate/chemer

# A SMOTified-GAN-augmented bagging ensemble model of extreme learning machines for detecting geochemical anomalies associated with mineralization

Min Guo, Yongliang Chen *

*College of Earth Sciences, Jilin University, Changchun, Jilin 130061, China*

ARTICLE INFO

Handling Editor: Francky Fouedjio

*Keywords:*
SMOTified-GAN
Bagging ensemble model
Extreme learning machine
Detecting geochemical anomalies
Polymetallic mineralization
Geochemical exploration

ABSTRACT

The use of supervised machine learning and deep learning techniques to automatically detect geochemical anomalies associated with mineralization has become a current research hotspot. However, due to the scarcity of known mineral deposits in the study area, the establishment of supervised machine learning and deep learning models faces the challenge of highly imbalanced data classification. To address this challenge, the SMOTified-GAN oversampling technique and bagging strategy were combined with extreme learning machines (ELMs) to construct a robust high-performance ensemble classification model for detecting geochemical anomalies associated with mineralization. In this ensemble model, SMOTified-GAN is used to balance the ratio of positive (mineralized) to negative (background) samples in geochemical exploration data set, while keeping the sample distribution pattern of positive samples unchanged. The bagging strategy is used to construct a robust ensemble model from the simple ELM classifiers to improve the robustness of the supervised anomaly detection model. Taking the Helong area (Jilin Province, China) as the case study area, three bagging ensemble models of the simple ELM classifiers with SMOTified-GAN, GAN and SMOTE augmentations were established on the 1: 50,000 stream sediment survey data, and used to automatically detect geochemical anomalies associated with polymetallic mineralization. The receiver operating characteristic (ROC) curve of the three ensemble models are very close to the upper left corner the of the ROC space, with the SMOTified-GAN augmented bagging ensemble model dominating the other two; and the area under the ROC curves (AUCs) of the three ensemble models are very close to 1.0 (0.99998, 0.99681, and 0.96803, respectively). Therefore, in terms of ROC curves and AUCs, the SMOTified-GAN augmented bagging ensemble model has the best performance in detecting geochemical anomalies associated with polymetallic mineralization. In addition, the geochemical anomalies associated with polymetallic mineralization detected by the SMOTified-GAN augmented bagging ensemble model have the close spatial correlation with the ore-forming control factors in the study area, and are mainly distributed around known polymetallic deposits. In other words, a bagging ensemble model with high performance can be constructed from the simple ELM classifiers with SMOTified-GAN augmentation in detecting geochemical anomalies associated with mineralization.

## 1. Introduction

The anomalous concentrations of indicator elements in mineral exploration area are often spatially correlated with the presence of mineralization and mineral deposits (Zuo, 2018). By identifying and analyzing these anomalies, it is possible to guide the mineral exploration and exploitation of mineral resources. Therefore, geochemical anomaly detection holds significant importance in mineral resource exploration. In recent years, supervised machine learning and deep learning methods have been employed to detect anomalies associated with mineralization from geochemical exploration data, and the detected geochemical anomalies are then used to guide the mineral prospecting process. For example, geographically weighted regression algorithms (Tian et al., 2018), maximum margin learning algorithm (Wang et al., 2019a), geographical weighted Lasso algorithms (Wang and Zuo, 2020), joint utilization of deep convolutional neural networks and pixel pair feature methods (Zhang et al., 2021), Bayesian convolutional neural networks (Huang et al., 2022) and a random forest model optimized by the

* Corresponding author.
*E-mail address:* chenyongliang2009@hotmail.com (Y. Chen).

https://doi.org/10.1016/j.chemer.2024.126156
Received 20 March 2024; Received in revised form 14 May 2024; Accepted 10 June 2024
Available online 14 June 2024

competitive mechanism and beetle antennae search (Cao et al., 2023). These supervised machine learning and deep learning methods need known mineral deposits to define positive samples in detecting geochemical anomalies associated with mineralization, and the detected geochemical anomalies associated with mineralization usually have a strong spatial correlation with the spatial locations of known mineral deposits. Therefore, they are effective methods for the detection of geochemical anomalies associated with mineralization (Zuo, 2017; Wang et al., 2019a, 2019b; Li et al., 2020).

The supervised detection of geochemical anomalies associated with mineralization needs to establish binary classification models to classify geochemical anomalies associated with mineralization and the background (Zhang et al., 2021). In geochemical exploration data set, traditionally, the mineralized samples are deemed as positive, while background samples are deemed as negative. However, a distinctive feature of geochemical exploration data is that the number of background samples is significantly larger than the number of mineralized samples (Cheng, 2007; Li et al., 2021; Parsa, 2021), resulting in a severe class imbalance between the positive and negative samples (Xiong and Zuo, 2018). In the supervised detection of geochemical anomalies associated with mineralization, an imbalanced training data set causes the classification scores of the binary classification model to be biased towards the background while deviate from geochemical anomalies associated with mineralization. Some semi-supervised learning methods have applied to detect mineralized geochemical anomalies for overcoming the issue of insufficient positive training samples. The examples include the convolutional neural network and transfer learning (Li et al., 2020) and the self-paced ensemble model (Chen et al., 2023).

To solve the class-imbalance problem in the supervised detection of geochemical anomalies associated with mineralization, Alina and Chen (2024) proposed a SMOTified ELM model that combines the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) with extreme learning machines (ELMs) (Huang and Chen, 2007, 2008; Huang et al., 2004, 2006, 2012; Ding et al., 2015; Chen and Wu, 2017; Wang et al., 2022). Although the SMOTified ELM model has a high-performance in the detection of geochemical anomalies associated with mineralization, its robustness is poor. In this study, to improve the robustness of the SMOTified ELM model in the detection of geochemical anomalies associated with mineralization, the bagging (Breiman, 1996) ensemble classification model of ELMs with SMOTE was constructed for the detection of geochemical anomalies associated with mineralization. To further overcome the limitation of SMOTE in data augmentation, the SMOTified-GAN (Sharma et al., 2022) was used to replace SMOTE to combine with ELM, and a bagging ensemble model of ELMs with SMOTified-GAN augmentation was finally constructed for the detection of geochemical anomalies associated with mineralization. The SMOTified-GAN technique is a hybrid oversampling technique that integrates SMOTE with the generative adversarial network (GAN) (Goodfellow et al., 2014) so as to overcome the limitations of both SMOTE and GAN in a data oversampling process. The SMOTified-GAN technique can be employed to solve the problem of class imbalance of the input data set by oversampling positive samples and simultaneously enhance the diversity of positive samples in the data set, while keeping the sample distribution pattern of positive samples unchanged. It is expected that the bagging ensemble model of ELMs with SMOTified-GAN augmentation have better performance and robustness than the SMOTified ELM model in the detection of geochemical anomalies associated with mineralization.

To test the performance and robustness of the bagging ensemble model of ELMs with SMOTified-GAN augmentation in the detection of geochemical anomalies associated with mineralization, a case study was carried out in the Helong area of Jilin Province, China. The bagging ensemble models of ELMs with SMOTified-GAN, GAN and SMOTE augmentations were constructed and compared in the supervised detection of geochemical anomalies associated with polymetallic mineralization. In the establishment of the three bagging ensemble models, the bagging algorithm was used as the ensemble constructing technique and ELMs were used as the base classifiers. The main difference among the three ensemble classification models is their training data sets. The training data of the bagging ensemble models of ELMs with SMOTified-GAN, GAN and SMOTE were obtained by oversampling the positive samples of geochemical exploration data using the SMOTified-GAN oversampling technique, GAN oversampling technique and SMOTE oversampling technique, respectively. The GAN oversampling technique ensures that the population distribution of the oversampled and original positive samples is the same. The SOMTE oversampling technique can adjust the ratio of the positive to negative samples of geochemical exploration data by generating synthesized positive samples. The performances of the three bagging ensemble classification models were evaluated using receiver operating characteristic (ROC) curves (Hanley and Mcneil, 1983; Fawcett, 2006) and the area under the ROC curve (AUC) (Chen, 2015; Nykänen et al., 2015, 2017; Chen and Wu, 2016, 2017; Parsa et al., 2017; Zuo, 2018; Xiong and Zuo, 2020). The C-A multifractal model (Cheng, 2007) was used to determine the thresholds for classifying geochemical anomalies associated with polymetallic mineralization.

The main contribution of this study is combining the SMOTified-GAN sampling technique with ELMs to build high-performance base classifiers, and the bagging algorithm is used to construct the robust high-performance ensemble model from the high-performance base classifiers. A case study is conducted to show the robustness and high performance of the SMOTified-GAN augmented bagging ensemble model in the supervised detection of geochemical anomalies associated with polymetallic mineralization.

## 2. Methods

### 2.1. SMOTified-GAN oversampling algorithm

The SMOTified-GAN algorithm (Sharma et al., 2022) is a hybrid oversampling technique, which ingeniously combines SMOTE with GAN. In this context, “positive samples” refer to minority class samples, and “negative samples” refer to majority class samples. The algorithm aims to balance the minority-to-majority ratio in original data set through a combination of SMOTE and GAN techniques. The oversampling process of SMOTified-GAN algorithm involves (a) generating a required number of synthetic positive samples using SMOTE and (b) preserving the same population distribution of the synthetic positive samples and original positive samples using the GAN technique.

GAN is an unsupervised generative algorithm constructed on the basis of the concept of training two competing neural networks: a generator (G), which aims at generating fake sample that closely mimicking real samples, and a discriminator (D), which aims at accurately distinguishing between real and G-generated fake samples (Goodfellow et al., 2014). The cooperative interaction of G and D enables GAN to generate fake samples that follow the same sample population distribution pattern as the real samples (Fig. 1) (Zhang et al., 2018; Ali-Gombe and Elyan, 2019; Suh et al., 2021; Sharma et al., 2022). Suppose the probability distribution of the set of real samples is $P_D$. The task of a GAN model is to generate a set of fake samples that obey the distribution $P_D$. First, a set of fake samples are generated by the G model, and it can be assumed that these generated fake samples obey the distribution $P_G(x; \theta)$, $\theta$ is the parameter set of the distribution. Suppose that $n$ samples, $\{x_1, x_2, \ldots, x_n\}$, are taken from the distribution $P_D$, the likelihood value $P_G(x_i; \theta)$ can be calculated and maximized by adjusting the parameter set $\theta$. This is equivalent to maximizing the probability that the samples generated by the $G$ model are from the distribution $P_D$. In other words, adjusting $\theta$ makes $P_G$ more similar to $P_D$. This problem can be translated into the KL divergence form, which describes the difference between two probability distributions. When the GAN model converges, $P_G$ approximates $P_D$.

Traditional GAN training requires a large number of real samples to

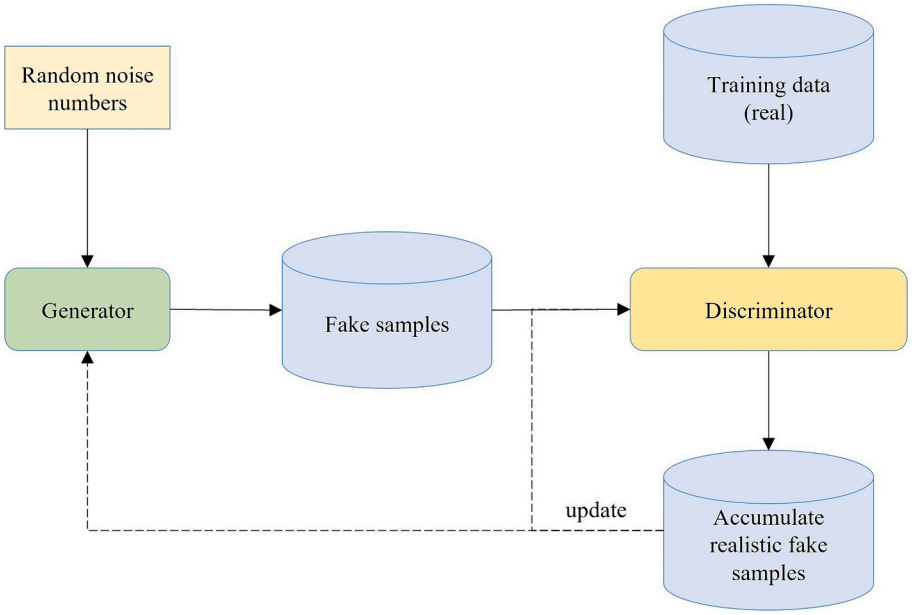


**Fig. 1.** GAN Framework for generating fake samples (modified from Sharma et al., 2022).

generate high-quality fake samples (Sharma et al., 2022). However, in practical applications such as geochemical exploration, it is difficult to obtain enough positive samples. SMOTified-GAN algorithm can be used to improve traditional GANs, using the synthetic positive samples generated by SMOTE as input of the generator, and the original positive samples as the training samples of the discriminator to compare with the output of the generator (Fig. 2). The SOMTE oversampling technique is used to adjust the ratio of the positive to negative samples of original data set to 1. The synthetic positive samples are used as the input samples of the generator in the GAN architecture and concurrently transfer distribution pattern knowledge learned from the original positive samples to the training process of GAN (Sharma et al., 2022). This hybrid oversampling method, known as SMOTified-GAN, emulates GAN-regenerated positive samples with the synthetic positive samples generated by SMOTE, helping to improve the quality of GAN regenerated positive samples (Sharma et al., 2022). From another perspective, this hybrid oversampling method uses GAN to improve the synthetic positive samples generated by SMOTE, which enhances the consistency of the distribution pattern between the synthetic positive samples and the original positive samples, and at the same time strengthens the diversity of the synthetic positive samples generated by SMOTE (Sharma et al., 2022).

In the initial phase of the SMOTified-GAN oversampling process, for each positive sample $x_p$ in the original data set, the SMOTE technique synthesizes new positive samples by selecting $K$-nearest neighbors within the same class and interpolating along the line segment in the feature space connecting $x_p$ to its $K$-neighbors, as shown in Fig. 3. It should be pointed out that the SOMTE algorithm only considers the similarity between positive samples in the geochemical feature space, but does not consider the relationship between positive samples from the perspective of geological feature similarity. Therefore, when the intrinsic relationship between the positive samples is not well described

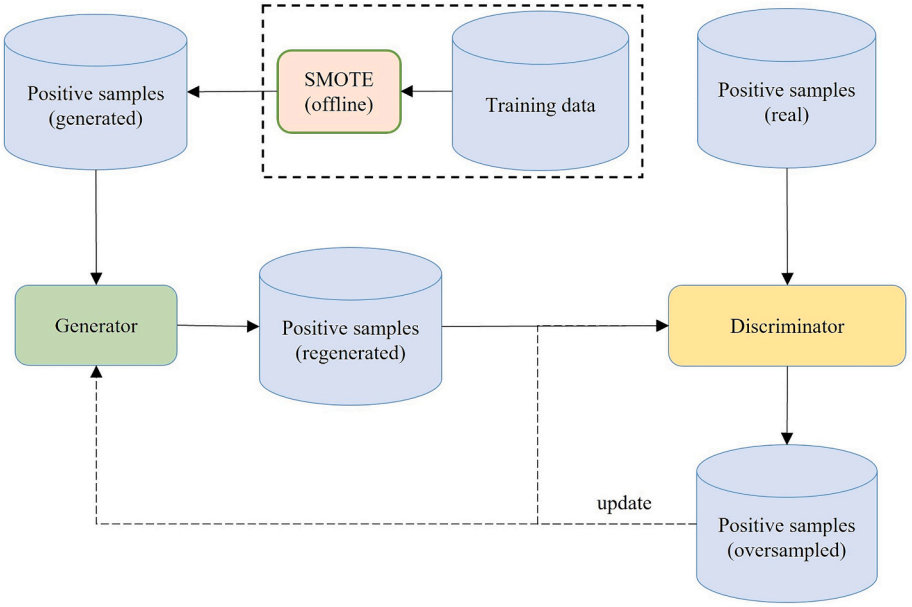


**Fig. 2.** SMOTified-GAN framework for generating fake samples (modified from Sharma et al., 2022).

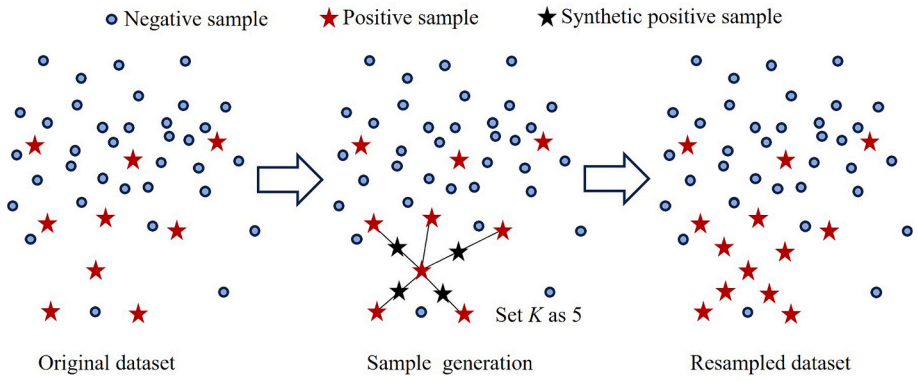


**Fig. 3.** SMOTE framework for generating synthetic samples in the feature space (modified from Farahbakhsh et al., 2023).

by the geochemical characteristics of the positive samples, the synthetic samples generated by the algorithm model may deviate from the positive sample population.

The generation of a synthetic positive sample $x_{SMOTE}$ for $x_p$ can be mathematically expressed as

$$x_{SMOTE} = x_p + Rand(0,1) \times |x_p - x_k| \quad (1)$$

where $x_k$ is a randomly selected sample from the $K$-nearest neighbors of sample $x_p$; and the subscripts $p$ and $k$ indicate positive sample and its $K$-nearest neighbors, respectively.

In the GAN regeneration phase of the SMOTified-GAN algorithm, both the generator ($G_s$) and the discriminator ($D_s$) are conceptualized as a multi-layer perceptron. The primary task of $G_s$ is to produce fake positive samples indistinguishable from real positive samples, while maximizing the loss of $D_s$ (i.e., classifying fake positive samples as real positive samples). Conversely, $D_s$ is tasked with accurately distinguishing between real and fake positive samples, while minimizing its misclassification loss. This adversarial interaction between $G_s$ and $D_s$ is mathematically encapsulated by Goodfellow et al. (2014) as

$$\min_{G_s}\max_{D_s} V(G_s, D_s) = E_{x^*}[logD_s(x^*|x)] + E_{x_z}[log(1 - D_s(G_s(z)))] \quad (2)$$

where $x^*$ is a positive sample from the real positive sample population $x$, and $x_z$ is a synthetic positive sample from the synthetic positive sample population $z$ generated by SMOTE. $D_s(x^*|x)$ is the output probability from $D_s$ for $x^*$ being a real sample and $D_s$ ($G_s(z)$) is the output probability from $D_s$ for $x_z$ being a real positive sample.

The adversarial training iterates until both $G_s$ and $D_s$ are well-trained (Goodfellow et al., 2014). In this way, SMOTified-GAN generates diverse fake positive samples consistent with the population distribution pattern observed in the original positive sample population.

### 2.2. Extreme learning machine algorithm (ELM)

The extreme learning machine (ELM) algorithm is a novel training strategy of single-hidden layer feedforward neural networks (SLFNs) (Huang and Chen, 2007, 2008; Huang et al., 2012). During the training of a SLFN model, the ELM algorithm avoids the iterative weight adjustments used in traditional backpropagation algorithms. In practice, an ELM model can be used as either a high-performance multiclass classifier or a high-performance nonlinear regressor. Chen and Wu (2017) established an ELM regression model for mapping mineral prospectivity based on the data of binary evidence layers. In geochemical exploration, an ELM model can be used as a binary classifier for supervised detection of geochemical anomalies associated with mineralization.

Fig. 4 shows the network structure of the ELM model proposed by Huang et al. (2012). Given $M$ arbitrarily distinct training samples$(x_i, t_i)$, $(i = 1, \cdots, M)$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{iN})^T \in R^N$ represents the $i$th input vector with $N$-dimensional features, and $t_i = (t_{i1}, t_{i2}, \ldots, t_{ic})^T \in R^c$ represents the $c$-dimensional target output layer vector of sample $x_i$. Given a standard SLFNs with $P$ hidden neurons and activation function $g(w, b, x)$, the predicted output of SLFNs for an input vector can be mathematically

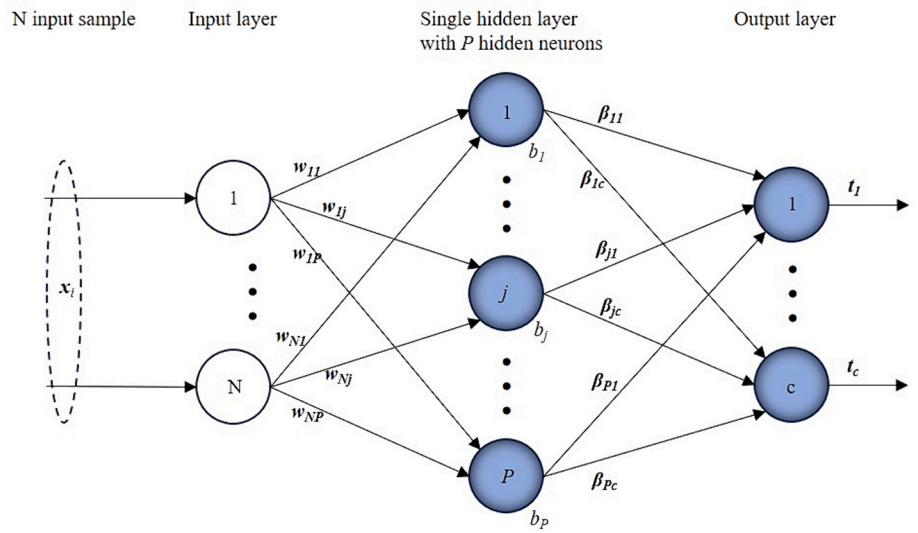


**Fig. 4.** The network structure of ELM (modified from Raghuwanshi and Shukla, 2018).

represented as

$$\sum_{j=1}^{P}\beta_j g\left(w_j \bullet x_i + b_j\right) = o_i, i = 1, 2, \ldots, M \tag{3}$$

where $w_j = \left(w_{1j}, w_{2j}, \ldots, w_{Nj}\right)^T$ is the input weight vector connecting the $N$ input neurons to the $j$th hidden neuron, $\beta_j = \left(\beta_{j1}, \beta_{j2}, \ldots, \beta_{jc}\right)^T$ is the output weight vector connecting the $j$th hidden neuron from the hidden layer to the $c$ output neurons, $b_j$ is the biases of the $j$th hidden neuron, and $g\left(w_j \bullet x_i + b_j\right)$ is an activation function for the $j$th hidden neuron.

According to Huang et al. (2004), in case of that the dimension of distinct input vectors $N$ is equal to the number of the distinct input vectors $M$, the standard SLFNs with $P$ hidden neurons and activation function $g(w, b, x)$ is able to approximate the input $M$ distinct training samples with zero error, i.e., $\sum_{i=1}^{N}\|o_i - t_i\| = 0$. In this case, Eq. (3) can be expressed as

$$\sum_{j=1}^{P}\beta_j g\left(w_j \bullet x_i + b_j\right) = t_i, i = 1, 2, \ldots, M. \tag{4}$$

A matrix form of the $M$ equations in Eq. (4) can be expressed as

$$H\beta = T \tag{5}$$

where

$$H = \begin{bmatrix} g(b_1, w_1, x_1) & \cdots & g(b_P, w_P, x_1) \\ \vdots & \cdots & \vdots \\ g(b_1, w_1, x_M) & \cdots & g(b_P, w_P, x_M) \end{bmatrix}_{M\times P}, \beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_P^T \end{bmatrix}_{P\times c} \text{ and } T = \begin{bmatrix} t_1^T \\ \vdots \\ t_M^T \end{bmatrix}_{M\times c} \tag{6}$$

where $H$ represents the hidden layer output matrix of the neural network, $\beta$ represents the output weight matrix, and $T$ represents the target output matrix.

To train a SLFN model, the weights $w_j$ and biases $b_j$, ($j = 1, 2, \ldots, P$), are initialized randomly. To obtain a strong generalization performance such that it can perform well in predicting new inputs, the ELM algorithm aims to find out an output weight matrix $\widehat{\beta}$ to achieve minimal training error and minimize the norm of output weights (Huang et al., 2012). Since ELM is a general linear system $H\beta = T$, the optimization objective function of the ELM algorithm can be expressed as

$$min\|H\beta - T\|^2 + \lambda\|\beta\| \tag{7}$$

where λ is the regularization parameter to prevent the model from overfitting, $H$ is the hidden layer output matrix and $T$ is the target vector matrix of inputs $x_i (i = 1, \cdots, M)$.

The output weight matrix $\widehat{\beta}$ can be computed by

$$\widehat{\beta} = H^{\dagger}T \tag{8}$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse matrix (Moore, 1920) of the hidden layer output matrix $H$. $H^{\dagger}$ is commonly computed using singular value decomposition (SVD) (Serre, 2002), and the $\widehat{\beta}$ calculated using Eq. (8) is the unique minimum norm least-squares solution of Eq. (5).

Given new inputs $X = (x_1, x_2, \ldots, x_k)^T \in R^{N\times k}$ and the output weight matrix $\widehat{\beta}$ calculated using Eq. (8), the prediction results of the ELM for new inputs $X$ can be calculated with the simplified form of the ELM output function given by

$$F(X) = h(X) \bullet \widehat{\beta} \tag{9}$$

where $F(X) = [F(x_1), \cdots, F(x_k)]$ is the prediction result of ELM for new inputs $X$, $h(X) = [g(X, w_1, b_1), \cdots, g(X, w_P, b_P)]$ is the output matrix of the hidden layer for $X$. For binary classification task, the decision function of ELM is applied to map $F(X)$ to the following binary form:

$$f(X) = sign(F(X)) \tag{10}$$

where $f(X)$ is the binary class prediction of the ELM model for the new inputs $X$, and $sign(\bullet)$ is a symbolic function which is to map the prediction result of a new input $x_i$ to class label 1 (positive class) if $F(x_i) \geq 0$, otherwise map the prediction result of $x_i$ to label 0 (negative class), $i = 1, 2, \ldots, k$.

### 2.3. Bagging strategy

The bagging strategy (Breiman, 1996) is an ensemble learning technique designed to enhance the accuracy and stability of predictions in classification and regression tasks by combining a set of base classifiers to reduce the variance of the outcome, thereby mitigating the risk of overfitting. The core concept of bagging strategy includes bootstrap sampling technique (Efron and Tibshirani, 1993) and the parallel training of a set of base classifiers.

The bagging strategy employs bootstrap sampling technique to conduct multiple rounds of random sampling from the original training set. Consequently, each round of bootstrap sampling generates a subset (bootstrap samples) of the original training set. These subsets are distinct and independent of one another. Each of these subsets is used to train a base classifier. These base classifiers can be either different classifiers or regressors, or identical ones, with no dependencies among them, allowing for parallel training. When it comes time to make predictions, the prediction results from all the base classifiers are combined. For classification tasks, a simple voting method is commonly used, while for regression tasks, a simple averaging method is applied. The simple voting method entails combining the prediction results of each base classifier based on majority rule, whereas the simple averaging method involves calculating the arithmetic mean of all the base classifiers' predictions.

In each round of bootstrapping sampling, the out-of-bag (OOB) estimate is used to evaluate the generalization performance of the base classifiers. The OOB estimate involves using samples from the original training set that were not included in the bootstrap samples for assessing the performance of the base classifiers. These samples are referred to as out-of-bag samples, as they have not been used to train any of the base classifiers, thus serving as an unbiased measure to evaluate the generalization capability of the base classifiers. The OOB estimate provides a convenient way to cross-validate the model without the need for a separate validation set, making it a cost-effective solution for estimating model performance.

In the supervised detection of geochemical anomalies associated with mineralization, the bagging strategy needs to be slightly modified due to the extreme class-imbalance of geochemical exploration data. The class-imbalanced geochemical exploration data set is divided into positive sample set and negative sample set, the size of positive sample set is much smaller than that of the negative sample set. In this case, the bootstrap sampling technique is used to randomly sample the negative sample set, and then the bootstrap negative sample and the positive sample are combined to form the training subset, so as to ensure that the training subset must contain the positive samples. Fig. 5 shows the bagging ensemble of ELMs constructed on the geochemical exploration data.

Assume that $T$ is the number of iterations for constructing a bagging ensemble classification model. In $t$th training iteration ($1 \leq t \leq T$), the base classifier $f_t$ is trained and used to predict the OOB negative samples. The likelihood (denoted as $l$) of each negative sample being predicted is cumulatively tracked across iterations. The prediction of base classifier $f_t$ for a negative sample $u$ is denoted as $f_t(u)$ ($t = 1, 2, \cdots, T$). The ensemble

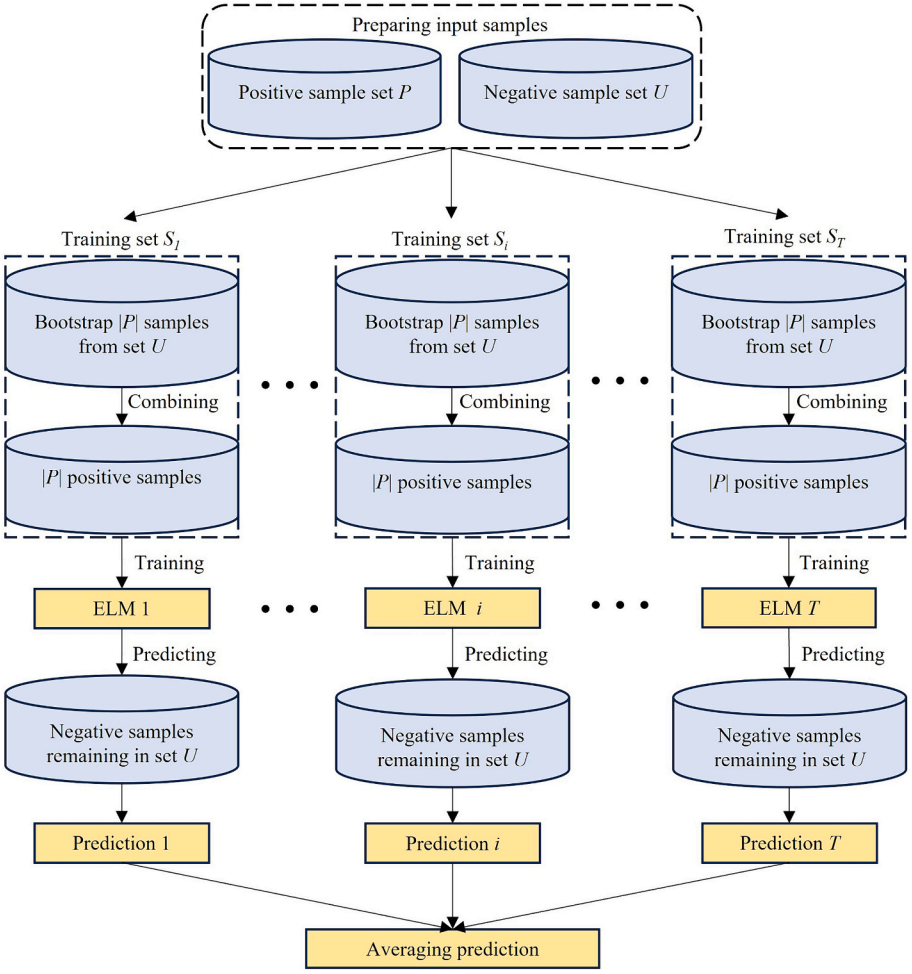


**Fig. 5.** Flowchart of the bagging ensemble model of ELMs.

prediction for sample $u$ is calculated by

$$\bar{f}(u) = \frac{\sum_{t=1}^{T} f_t(u)}{l} \tag{11}$$

where $\bar{f}(u)$ represents the ensemble prediction for negative sample $u$, and $T$ is the number of iterations.

### 2.4. Bagging ensemble of ELMs with SMOTified-GAN augmentation

A bagging ensemble of ELMs (Huang and Chen, 2007, 2008; Huang et al., 2012) with SMOTified-GAN augmentation is constructed from ELMs on the training data augmented by the SMOTified GAN algorithm proposed by Sharma et al. (2022). In this ensemble, ELMs are used as the base classifiers, the SMOTified GAN algorithm is used to generate class-balanced augmented training data from original class-imbalanced input data, and bagging strategy is used to build the ensemble model from ELMs. In geochemical exploration, the bagging ensemble classification model can be constructed from binary ELM classifiers for the supervised detection of geochemical anomalies associated with mineralization. The SMOTified-GAN oversampling technique is used to generate the class-balanced training data from extremely class-imbalanced geochemical exploration data, and the bagging strategy (Breiman, 1996) is used to build the ensemble classification model from binary ELM classifiers on the augmented geochemical exploration data.

Given a geochemical exploration data set $S$, in which all known mineralized samples in $S$ constitute the positive sample set $P$, and the remaining samples constitute the negative sample set $U$. The number of samples in $P$ and $U$ satisfies $|P| << |U|$. The number of mineralized samples $|P|$ is usually very small ($|P| = 14$) in the study area. In this case, the SMOTE oversampling technique (Chawla et al., 2002) is first used to increase $|P|$ by generating $|U|$ - $|P|$ synthetic positive samples based on the 14 mineralized samples. All the synthetic positive samples and all the mineralized samples are then combined to form the SMOTified positive sample set $P_s$. The GAN algorithm (Goodfellow et al., 2014) is then used to generate fake positive samples based on the $|U|$ - $|P|$ synthetic positive samples in data set $P_s$. The $|U|$ - $|P|$ fake positive samples generated by the converged GAN model, and the population distribution of the $|U|$ - $|P|$ fake positive samples is the same as that of the original positive sample set $P$. All the fake positive samples and all the original positive samples in $P$ are combined to form the SMOTified-GAN positive sample set $P_{sg}$. Finally, set $P_{sg}$ and set $U$ are combined to form the class-balanced training data set for constructing the bagging ensemble model from ELMs for supervised detection of geochemical anomalies associated with mineralization.

Given the maximum number of iterations $T$, the size of training subsets $Q$, and exploration data sets $P$ and $U$, the algorithm for constructing the bagging ensemble model from ELMs on the SMOTified-GAN augmented geochemical exploration data is outlined as follows:

Step 1 Use the SMOTE oversampling technique to oversample positive samples in set $P$ to generate the SMOTE-augmented positive sample set $P_s$.

Step 2 Use the GAN oversampling technique to oversample positive samples in $P_s$ to generate fake positive samples to form the SMOTified-GAN-augmented positive sample set $P_{sg}$.
Step 3 Combine the positive data set $P_{sg}$ and negative data set $U$ into $PU$ as the training data for bagging process.
Step 4 For $t = 1$ to $T$ do:
Step 4.1 Randomly select $Q$ samples from the training data set $PU$ to form a bootstrapped sample set $S_t$ of size $Q$.
Step 4.2 Train the ELM classification model on the bootstrapped data set $S_t$.
Step 4.3 Test the ELM classification model on the out-of-bagging samples ($PU \setminus S_t$).
Step 5 The $T$ ELM classification models built in Step 4 are combined into a bagging ensemble model.
Step 6 The bagging ensemble model is used to classify all the samples in the geochemical exploration data set $S$.

In geochemical exploration, geologists care only about whether a sample is mineralized or not. In this case, the output of the ELM model can be specified as the probability that the sample belongs to the mineralized anomaly and the background. Geologists can delineate potential mineralized anomaly areas based on the probability that each sample belongs to the mineralized anomaly.

### 2.5. Model performance evaluation methods

The ROC curve (Chen, 2015; Nykänen et al., 2015, 2017; Chen and Wu, 2016, 2017; Parsa et al., 2017; Zuo, 2018; Xiong and Zuo, 2020) of a binary classification model is a visual tool to evaluate the performance of the classification model by drawing a performance curve in the ROC space. In practice, to compare the performance of different binary classification models, the ROC curves of all the binary classification models are plotted in the same ROC space. Those binary classification models with better classification performance have the ROC curves closer to the upper left corner of the ROC space. The AUC value of a binary classification model represents the area between the ROC curve of the binary classification model and the bottom horizontal coordinate of the ROC space. A binary classification model with better classification performance has the AUC value closer to 1.0 than 0.5. In this study, the ROC curves and AUCs are used to evaluate the performance of the three bagging ensemble classification models in the detection of geochemical anomalies associated with polymetallic mineralization.

According to Chen (2015), AUC is a random variable depending on the output of the corresponding binary classification model. It can be

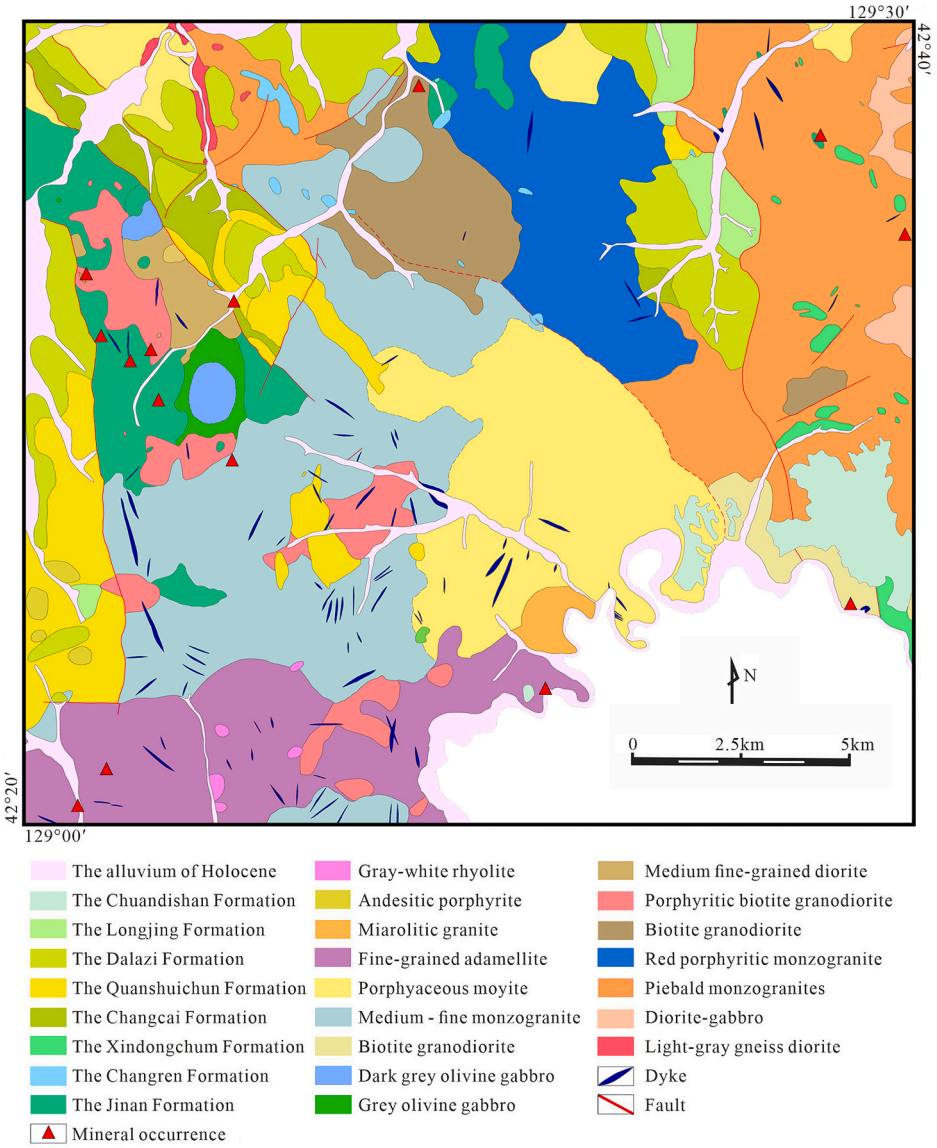


**Fig. 6.** Simplified geological map of the geochemical exploration area (modified from Chen et al., 2021).

used to formulate a standard normally distributed random variable, $Z_{AUC}$, and then use $Z_{AUC}$ to conduct a statistical hypothesis test. For how to formulate the $Z_{AUC}$ from AUC, see Chen (2015). If the $Z_{AUC}$ value calculated from the AUC value is greater than the threshold value 1.96 listed in the normal distribution table, it can be inferred that the AUC value is significantly different from 0.5, indicating that the binary classification model is significantly better than the random guess model. In geochemical exploration, according to whether or not the $Z_{AUC}$ value of a certain geochemical element is >1.96, one can judge whether the geochemical element can be used as an indicator element for mineral exploration targeting. If the $Z_{AUC}$ value of an element is >1.96, it indicates that there is a significant statistical correlation between the element concentration and known mineral deposits (Chen, 2015; Chen and Wu, 2017, 2019; Chen et al., 2021).

## 3. Study area and data

### 3.1. Geological background

The Helong area in Jilin Province (China) is taken as the case study area. The area spans from approximately 129°00′ to 129°30′ east longitude and from about 42°20′ to 42°40′ north latitude (Fig. 6). It tectonically lies between the northeastern margin of the North China Craton and the Xingmeng Orogenic Belt (Chai and Liu, 2015). The area has experienced the three primary tectonic evolutionary stages including the Archean evolution stage that formed the paleo-continental nucleus, the evolution stage of the paleo-Asian Ocean tectonic domain, and the evolution stage of the coastal Pacific tectonic domain (Yu et al., 2012; Wu et al., 2005; Zhang et al., 2004). The study area exhibits a well-developed fault system, primarily distributed along the boundaries of Mesozoic rift basins or parallel to the Gudonghe fault zone. There are three main groups of faults in the study area which are the NW-trending faults represented by the Gudonghe fault (the most significant fault zone), the nearly N-S to NNW-trending faults, and the NNE to NE-trending faults (the NE-trending faults are within the study area). The Gudonghe fault zone is the northern margin fault of the North China Craton, extending northwestward through the entire study area, controlling the main structural framework of the area (Chai and Liu, 2015; Chen et al., 2021). Since the Phanerozoic, the study area has been affected by the superposition of the Paleo-Asian Oceanic tectonic domain and the Pan-Pacific tectonic domain. This geological interaction has led to frequent magmatic activities in the study area, resulting in the formation of widely distributed intrusive rocks and a small number of volcanic rocks. Specifically, intrusive rocks cover 69.58 % of the study area and mainly include granite, granodiorite, diorite, gabbro, etc. These intrusive rocks form a wide variety of exposed bedrock and rock groups. Wu et al. (2013) conducted LA-ICP-MS zircon U—Pb isotope analysis on two samples of amphibolite from the study area, and the zircon U—Pb dating indicates that the amphibolite in the study area was formed in the Middle Jurassic between 1.73–1.75 Ma. The igneous rocks in the study area belong to the Xiaohinggan Mountains-Zhangguangcai Mountains igneous rock belt, and the exposed area accounts for 77.55 % of the study area. The exposed stratigraphic area in the study area is small, accounting for about 22.45 % of the study area (Chai and Liu, 2015; Chen et al., 2021), and it is mainly distributed in the eastern side of the Helong city, in the Mesozoic basins in the study area, as well as in the Quaternary alluvial plains (Chai and Liu, 2015).

The main mineralization periods in the study area are the Hercynian and Yanshanian tectonic active periods, and the tectonomagmatic activities in these two periods provided the inherent conditions for the formation of endogenic metal deposits such as Au, Ag, Mo, Cu, Pb, Zn, and Ni (Chai and Liu, 2015; Yan et al., 2015). A total of 14 polymetallic deposits have been found in the study area (Liu and Zhang, 1999), the majority of which are hydrothermal deposits (e.g. Au, Ag, Mo deposits), with a few being hydrothermal skarns (e.g. Fe, Mn deposits). The Au and Ag deposits in the study area are mainly middle-low to medium-temperature hydrothermal quartz vein type; the Ni—Co deposits are mainly produced in the ultramafic and ferromagnesic rock mass of the late Early Jurassic, whose lithology is olivine gabbro; the shallow rock mass related to Mo deposits are all calc-alkaline series granitoids with high silicon and rich potassium, and the rock types are mainly porphyritic monzonitic granite (Chai and Liu, 2015). The formation of these deposits is closely related to multi-phase magmatic activities (Yan et al., 2015; Pan et al., 2016). The main controlling factors for polymetallic mineralization in the study area are regional deep faults, Archean tectonics, and Yanshanian magmatic rocks.

### 3.2. Geochemical survey data

The geochemical exploration data used in this study were obtained from stream sediment measurements conducted by the Geological Survey Institute of Jilin University in the study area at a scale of 1:50,000 (Chai and Liu, 2015). The sampling process for these stream sediment measurements was conducted in accordance with Chinese Geochemical Survey Criteria (No.DZ/T0011–91) (Chai and Liu, 2015). A total of 6999 stream sediment survey samples were collected from the study area (Fig. 7). The sampling coverage spanned an area of 1320 $km^2$, with an average of 1–2 samples taken per 0.25 $km^2$. Inner Mongolia Mineral Experiment Institute of China conducted testing and analysis of the concentrations of the 13 elements (Au, Hg, As, Ag, Sb, Mo, W, Cu, Pb, Zn, Bi, Ni, and Co). The concentration of Au was determined using atomic absorption spectroscopy (AAN), the concentrations of Hg and As were determined using Atomic Fluorescence Spectrometry (AFS), and the concentrations of Ag, Sb, Mo, W, Cu, Pb, Zn, Bi, Ni, and Co elements were determined using inductively coupled plasma mass spectrometry (ICP-MS) (Chai and Liu, 2015; Chen et al., 2021).

The irregularly distributed data points in the study area need to be converted into regular grid data for further geochemical exploration. A series of interpolation techniques, such as the inverse distance weighting (IDW) and Kriging, can be used for the interpolation of stream sediment data. The IDW method is simple and easy to implement in practical application. Therefore, it is a commonly used interpolation method in the preprocess of geochemical exploration data. Compared with the IDW method, Kriging methods are more complicated in implementation. The methods require a complex process of experimental variogram estimation and theoretical variogram modeling to estimate a series of parameters for the Kriging estimation, such as range, sill, nugget constant, and so on. The advantage of the Kriging methods is that they can provide both the interpolation result and the uncertainty of the interpolation result. Stream sediment survey data is a special kind of exploration geochemical data. Stream sediment samples were distributed in the smallest catchment areas of the study area. In each of the smallest catchments, streams flow only from the upper reaches of the catchment to the lower reaches. Therefore, stream sediment samples can only be used for grid point estimation in the upstream region, not in the downstream region. The best way for the interpolation of stream sediment samples is that the stream catchments are used to constrain the interpolation algorithms. However, the 1:50,000 digital elevation data for the study area is a state secret and cannot be used publicly. In this case, only IDW method can be used in this study to interpolate stream sediment samples.

Golden Software Surfer was used to convert the concentration values of each geochemical element in the 6999 irregular spatial drainage sediment samples into regular grid point values, and 200 × 134 grid point values were obtained. Each grid point value was estimated from 8 to 64 drainage sediment samples around the grid point. Each grid point represents a grid cell with a size of 0.2282 × 0.2296 $km^2$. Such a small grid cell size guarantees that one grid cell contains no more than one known polymetallic deposit. As a result, only 14 of the 26,800 grid cells contain known polymetallic deposits. These 14 grid cells were designated as positive samples for the ROC curve analysis. Of the remaining grid cells, 2370 grid cells are located in empty regions where there are

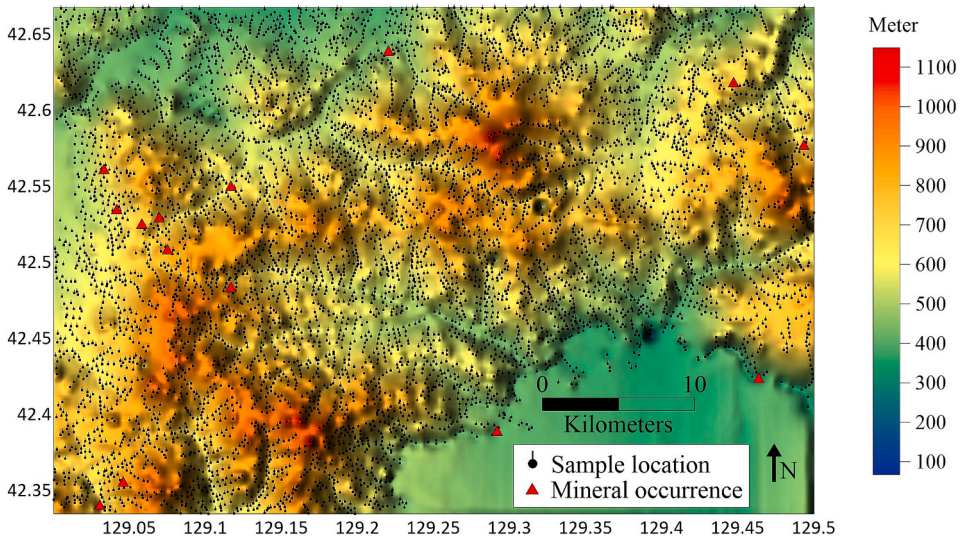


**Fig. 7.** Stream sediment sampling locations of the study area.

neither known polymetallic deposits nor element concentration values. A total of 24,416 grid cells with element concentration values but not containing known polymetallic deposits were designated as negative samples for ROC curve analysis (Chen et al., 2021). In this case, the 24,416 negative samples definitely contain a few cells that contain undiscovered polymetallic deposits. These misdefined cells make up the noise in the negative samples. These noises only account for a very small percentage of negative samples, so they do not significantly affect model training and ROC curve analysis results.

## 4. Results

### 4.1. Indicator element selection and training data set definition

In this study, indicator elements were selected from the 13 trace elements with assayed concentrations according to whether there was a significant spatial correlation between the concentration of elements and the locations of known polymetallic deposits. If there is a significant spatial correlation between the concentration of an element and the locations of the known polymetallic deposits, the $Z_{AUC}$ value of the element must be greater than the threshold of 1.96 in the standard normal distribution table (Chen and Wu, 2016, 2017, 2019). Table 1 shows the AUCs and $Z_{AUC}$s of the 13 elements estimated based on the grid points data in Section 3.2. It can be seen from Table 1 that the $Z_{AUC}$ values of Au, Co, Cu, Mo, Ni, and W are greater than the threshold value of 1.96. Therefore, Au, Co, Cu, Mo, Ni, and W were selected as indicator elements for the detection of geochemical anomalies associated with polymetallic mineralization in the study area.

According to the geological and polymetallic mineralization characteristics of the study area discussed in Section 3.1, Au, Co, Cu, Mo are polymetallic metallogenic elements, while Ni and W are primary halo elements (Wan et al., 2010; Yan et al., 2015; Pan et al., 2016). Therefore, the indicator elements selected by the statistical testing method have geological and metallogenic significancy, and the selected elements of Au, Co, Cu, Mo, Ni, and W can be used as indicator elements for geochemical exploration in the study area. Fig. 8 is the color relief maps of the 6 indicator elements.

The grid point data of Au, Co, Cu, Mo, Ni and W were used as the original training data for constructing the bagging ensemble classification models for the detection of geochemical anomalies associated with polymetallic mineralization in this study. Of the 26,800 grid cells, there are 2370 grid cells located in empty regions. These grid cells have neither known polymetallic mineralization information nor geochemical element concentrations. Therefore, both in the construction of the bagging ensemble classification models and in the detection of geochemical anomalies related to polymetallic mineralization, they are ignored in this study. As a result, the original training data set consists of 14 positive grid cells (contain polymetallic deposits) and 24,416 negative grid cells (do not contain polymetallic deposits but have element concentrations). Each grid cell in the training data set is represented as a vector with six components (element concentration values). The known polymetallic deposits were used to determine the labels of grid cells. The 14 positive cells have label value of 1 and the 24,416 negative grid cells have label value of 0. Therefore, the original training data set $S$ consists of positive data set $P$ with $|P| = 14$ and negative data set $U$ with $|U| = 24{,}416$. The three bagging ensemble classification models were constructed on the original training data set $S$.

### 4.2. Bagging ensemble classification model construction

In this study, the Python code for the SMOTE algorithm (Chawla et al., 2002) is from the imblearn library (Lemaître et al., 2017), the Python code for the GAN algorithm (Goodfellow et al., 2014) and SMOTified-GAN algorithm (Sharma et al., 2022) is modified from the Python code at https://github.com/sydney-machine- learning/GAN-classimbalanced, and the Python code for the ELM algorithm (Huang et al., 2012) is modified from the Python code at http://www.extreme-learning-machines.org.

**Table 1**
AUCs and $Z_{AUC}$s for 13 elements.

| Element | AUC | $Z_{AUC}$ | Element | AUC | $Z_{AUC}$ | Element | AUC | $Z_{AUC}$ |
|---|---|---|---|---|---|---|---|---|
| Ag | 0.52 | 0.24 | Cu | 0.73 | 2.92 | Sb | 0.59 | 1.13 |
| As | 0.61 | 1.31 | Hg | 0.47 | −0.40 | W | 0.67 | 2.18 |
| Au | 0.73 | 3.01 | Mo | 0.69 | 2.36 | Zn | 0.63 | 1.57 |
| Bi | 0.62 | 1.52 | Ni | 0.73 | 3.04 | | | |
| Co | 0.67 | 2.14 | Pb | 0.40 | −1.48 | | | |

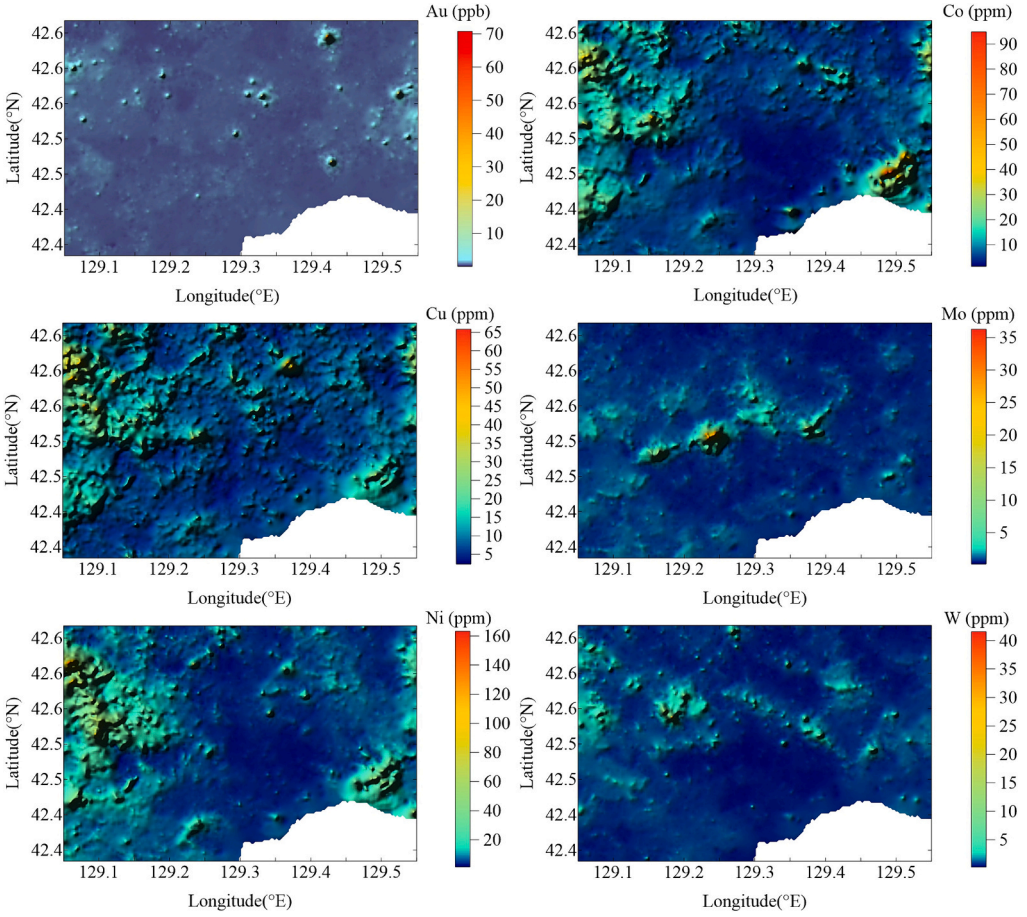


**Fig. 8.** Color relief maps of the 6 indicator elements.

The hyperparameters for SMOTE (Chawla et al., 2002), GAN (Goodfellow et al., 2014), bagging (Breiman, 1996) and ELM algorithms (Huang et al., 2012) need to be defined for controlling the construction process of the bagging ensemble models. SMOTE was used to produce enough number of positive samples for training the GAN model in the bagging ensemble model, so the hyperparameter 'sampling_strategy' of the SMOTE model was set to 0.05, meaning that 1206 positive samples were produced by the SMOTE algorithm. The optimal value of the hyperparameter "k_neighbors" of the SMOTified ELM model determined by the trial-and-error method in this study is 3. The remaining hyperparameters of the SMOTE algorithm were set to default values. In the reference with Sharma et al. (2022), the hyperparameters for the GAN algorithm were set as "learning_rate = 0.0002", "epochs = 200", "batch_size = 128", and remaining hyperparameters equal to default values. The only one hyperparameter for the bagging algorithm was empirically set as "Maxiter = 1000", indicating that 1000 base classifiers were trained and then used to form the bagging ensemble model. The ELM algorithm provides several modules. In this study, the "Simple ELM Classifier" was selected as the base classifiers in the construction of bagging ensemble classification models. In the reference with Huang et al. (2006), the hyperparameters for the "Simple ELM Classifier" module were set as "n_hidden = 20", "activation_func = '*tanh*' ", and the remaining hyperparameters equal to default values.

When completing the above initialization process, the data augmentation process was conducted. In this process, the SMOTified-GAN algorithm (Sharma et al., 2022) was used to augment the training data set *S*. The SMOTE algorithm was first used to oversample the positive data set *P* containing only 14 positive grid cells, and the SMOTE augmented positive data set $P_s$ containing 1220 positive samples (containing 1206 synthetic grid cells and 14 original positive grid cells) was generated. The SMOTE-augmented positive data set $P_s$ together with the original negative data set *U* (containing 24,416 negative grid cells) were used to train the GAN model, and the SMOTified-GAN-augmented training data set (containing 1220 positive grid cells and 24,416 negative grid cells) was finally generated. In addition, the original training data set *S* (containing 14 positive grid cells and 24,416 negative grid cells) was also used to train the GAN model to generate the GAN-augmented training data set (containing 1220 positive grid cells and 24,416 negative grid cells). Fig. 9 shows T-SNE plots of the original positive samples and the pseudo-positive samples produced by SMOTE, GAN, and SMOTified-GAN techniques, respectively. Fig. 9a shows the synthetic positive samples produced by SMOTE around the original positive samples in T-SNE space. Fig. 9b shows that the distribution of false positive samples generated by GAN is basically the same as that of original positive samples in T-SNE space. Fig. 9c shows that fake samples generated by SMOTified-GAN in T-SNE space.

The original training data set augmented by using the SMOTE, GAN, and SMOTified-GAN techniques was used for the construction of the bagging ensemble model of ELM classifiers. Accordingly, the three bagging ensemble models of ELM classifiers were constructed on the augmented geochemical exploration data for the detection of geochemical anomalies associated with polymetallic mineralization. Each of the three bagging integration models consists of 1000 simple ELM classifiers, each of which is built in one iteration of the bagging process. The outputs of 1000 simple ELM classifiers constituting each bagging ensemble model is finally synthesized into the output of the bagging ensemble model by the voting method (Chen and Chen, 2023).

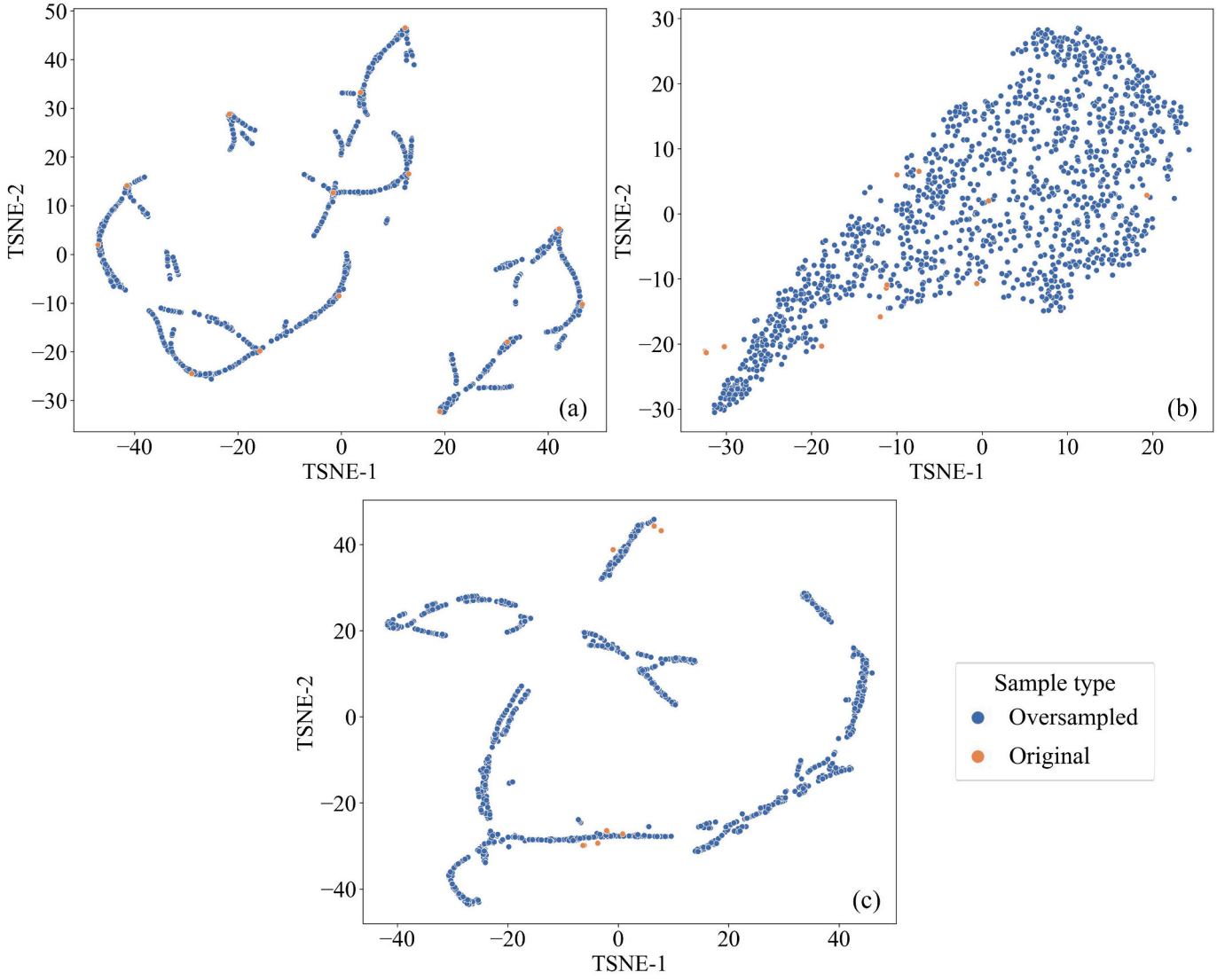


**Fig. 9.** T-SNE visualization of the original samples and positive samples oversampled by (a) SMOTE, (b) GAN and (c) SMOTified-GAN.

### 4.3. Delineation of polymetallic mineralization anomalies

The output of each ensemble model is the probability that the input grid cell belongs to a geochemical anomaly associated with polymetallic mineralization, and the probability that the input grid cell belongs to the background, the sum of which equals 1.0. The probability of the input grid cell belonging to a geochemical anomaly associated with polymetallic mineralization reflects the degree of the input grid cell belonging to the geochemical anomaly associated with polymetallic mineralization. Therefore, it is used as the basis for identifying geochemical anomalies associated with polymetallic mineralization in this study. Fig. 10 shows the probability that each grid cell belongs to a geochemical anomaly associated with polymetallic mineralization. In Fig. 10, the color gradient represents the probability values, with red indicating higher probability values and blue indicating the background.

To delineating geochemical anomalies associated with polymetallic mineralization in the study area, the C-A multifractal model (Cheng et al., 1996; Cheng, 2007) was adopted to determine thresholds for categorizing the geochemical anomalies associated with polymetallic mineralization. In the C-A plot (Fig. 11), four straight lines were fitted with three thresholds, which can classify the probabilities produced by the bagging ensemble model of the simple ELM classifiers with SMOTE, GAN, and SMOTified-GAN augmentation into four distinct probability levels: background, weak anomaly, high anomaly, and strong anomaly, respectively. The $R^2$ values of the fitting lines shown in Fig. 11 are all >0.80, indicating a good fit of these lines.

Fig. 12 shows the geochemical anomalies associated with polymetallic mineralization classified by the three thresholds determined by the C-A multifractal model. In Fig. 12a, the three thresholds determined by the C-A multifractal model of the SMOTE-augmented bagging ensemble model of ELMs were 0.002, 0.126 and 0.897, respectively. In Fig. 12b, the three thresholds determined by the C-A multifractal model of the GAN-augmented bagging ensemble model of ELMs were 0.001, 0.105 and 0.794, respectively. In Fig. 12c, the three thresholds determined by the C-A multifractal model of the SMOTified-GAN-augmented bagging ensemble model of ELMs were 0.004, 0.079 and 0.841, respectively. As can be seen from Fig. 12, the high and strong level geochemical anomalies associated with polymetallic mineralization identified by the bagging ensemble model of the simple ELM classifiers with SMOTE, GAN, and SMOTified-GAN augmentation are overall consistent with the distribution of known polymetallic occurrences and contain all the known polymetallic occurrences. However, it is obvious that the high and strong geochemical anomalies associated with polymetallic mineralization predicted by the bagging ensemble model of the simple ELM classifiers with SMOTified-GAN augmentation is the smallest and is mainly distributed in the areas where the known

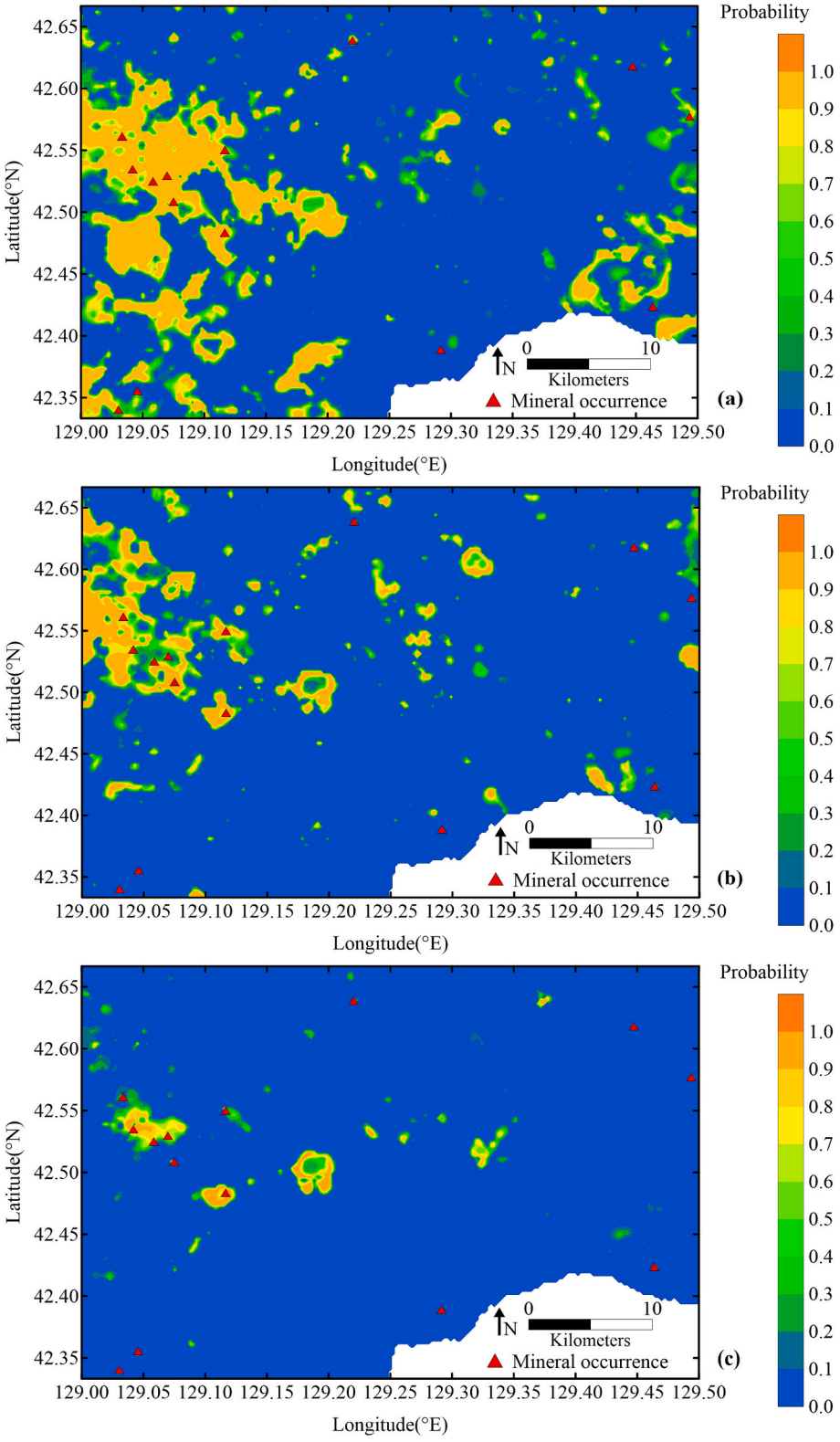


**Fig. 10.** The geochemical anomalies associated with polymetallic mineralization detected by (a) the SMOTE-augmented bagging ensemble model of ELMs, (b) the GAN-augmented bagging ensemble model of ELMs and (c) the SMOTified-GAN-augmented bagging ensemble model of ELMs.

polymetallic occurrences are concentrated.

## 5. Discussion

### 5.1. Model performance valuation

The ROC curve, AUC, and data processing time of a classification

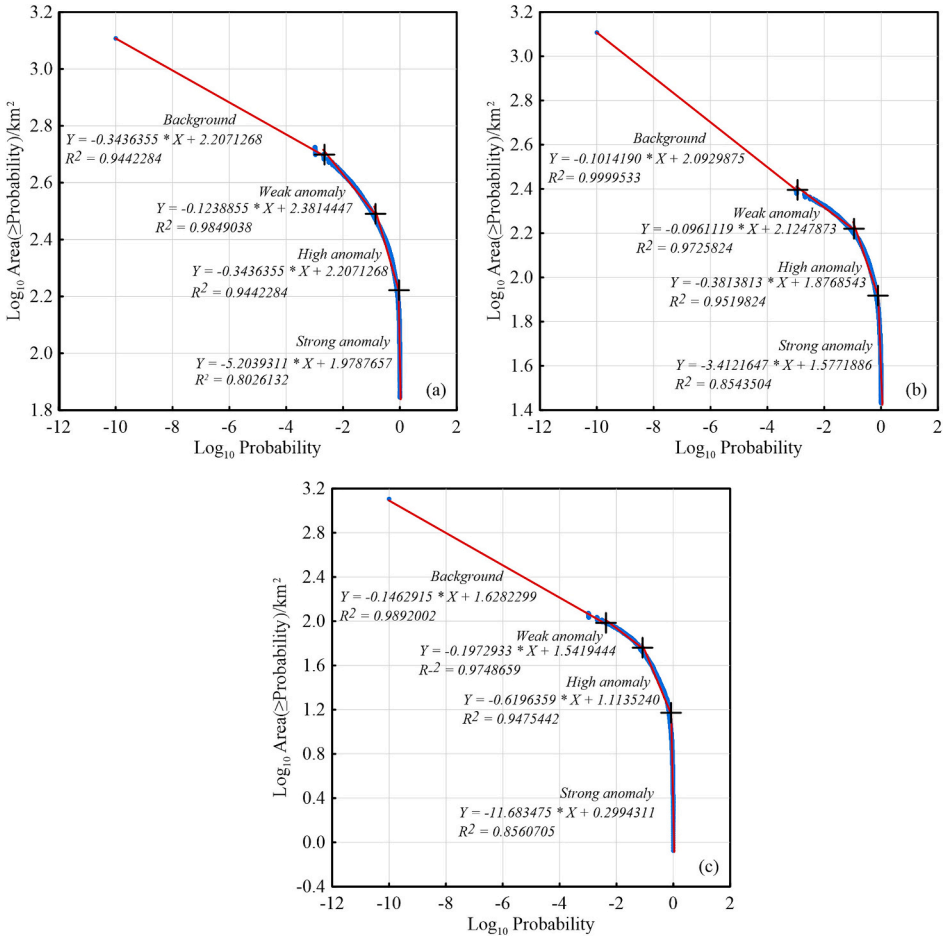


**Fig. 11.** Log-log plot of the anomaly probabilities and the accumulative area for (a) the SMOTE-augmented bagging ensemble model of ELMs, (b) the GAN-augmented bagging ensemble model of ELMs and (c) the SMOTified-GAN-augmented bagging ensemble model of ELMs.

model are usually used to assess the performance of the model in mineral prospectivity mapping. Previous studies (Chen, 2015; Chen and Wu, 2016, 2017; Chen et al., 2021) have demonstrated the effectiveness of ROC curves and AUCs in evaluating geochemical anomaly detection models. In geochemical anomaly detection, a specific threshold is used to classify grid cells into positive grid cells (geochemical anomalies) and negative grid cells (the background). Consequently, the true positive rate (TPR) and false positive rate (FPR) are then calculated according to the classification result. When changing the threshold used in the classification, the corresponding TPR and FPR also change. The ROC curve is used to describe the change of TPR versus FPR at various threshold settings. Besides the ROC curve and AUC value, the accumulative gain curve and area under the accumulative gain curve (AUL) (Wu and Chen, 2017) were also used to evaluate the performance of the ensemble models in this study.

Fig. 13 shows the ROC curves (a) and accumulative gain curves (b) for the bagging ensemble models of simple ELM classifiers constructed on the training data augmented by the SMOTified GAN, GAN and SMOTE algorithms. The three ROC curves in Fig. 13a reveal that the bagging ensemble model constructed on the SMOTified-GAN-augmented data set is superior to that constructed on the GAN-augmented data set, and the bagging ensemble model constructed on the GAN-augmented data set is superior to that constructed on the SMOTE-augmented data set. The three accumulative gain curves in Fig. 13b reveal that the high-level anomalies identified by the bagging ensemble models augmented by SMOTified-GAN, GAN, and SMOTE account for 4.22 %, 12.48 % and 23.64 %, respectively. All the known polymetallic occurrences are distributed in the high-level anomalies identified by the three ensemble models. The AUL values of the three ensemble models are respectively 0.99967, 0.98944 and 0.97268, indicating that the ensemble model augmented by SMOTified GAN has the best performance among the three ensemble models.

Table 2 shows the AUCs, $Z_{UAC}$s and data processing time of the bagging ensemble models constructed on the training data augmented by the SMOTified-GAN, GAN and SMOTE algorithms. The AUCs of the three ensemble models are respectively 0.99996, 0.98972, and 0.97295, indicating that the SMOTified-GAN augmentation is better than the GAN augmentation, and the GAN augmentation is better than the SMOTE augmentation. The $Z_{UAC}$s of the three ensemble models are respectively 413.38541, 25.75636, and 15.53214, which are all much greater than the threshold value of 1.96 listed in the standard normal distribution table. Therefore, the three bagging ensemble models can effectively detect geochemical anomalies associated with polymetallic mineralization in the study area. The data processing time of the three bagging ensemble models are respectively 31.08673 s, 31.33171 s, and 34.53458 s, indicating that the three bagging ensemble models have similar data processing efficiency in the detection of geochemical anomalies associated with polymetallic mineralization.

### 5.2. *Geological interpretation of geochemical anomalies delineated in the study area*

The bagging ensemble model of simple ELM classifiers with the SMOTified-GAN augmentation identified three high anomaly zones in the study area (Fig. 14). The first high anomaly zone is distributed in the Late Archean Ji'nan Formation and near to porphyritic biotite

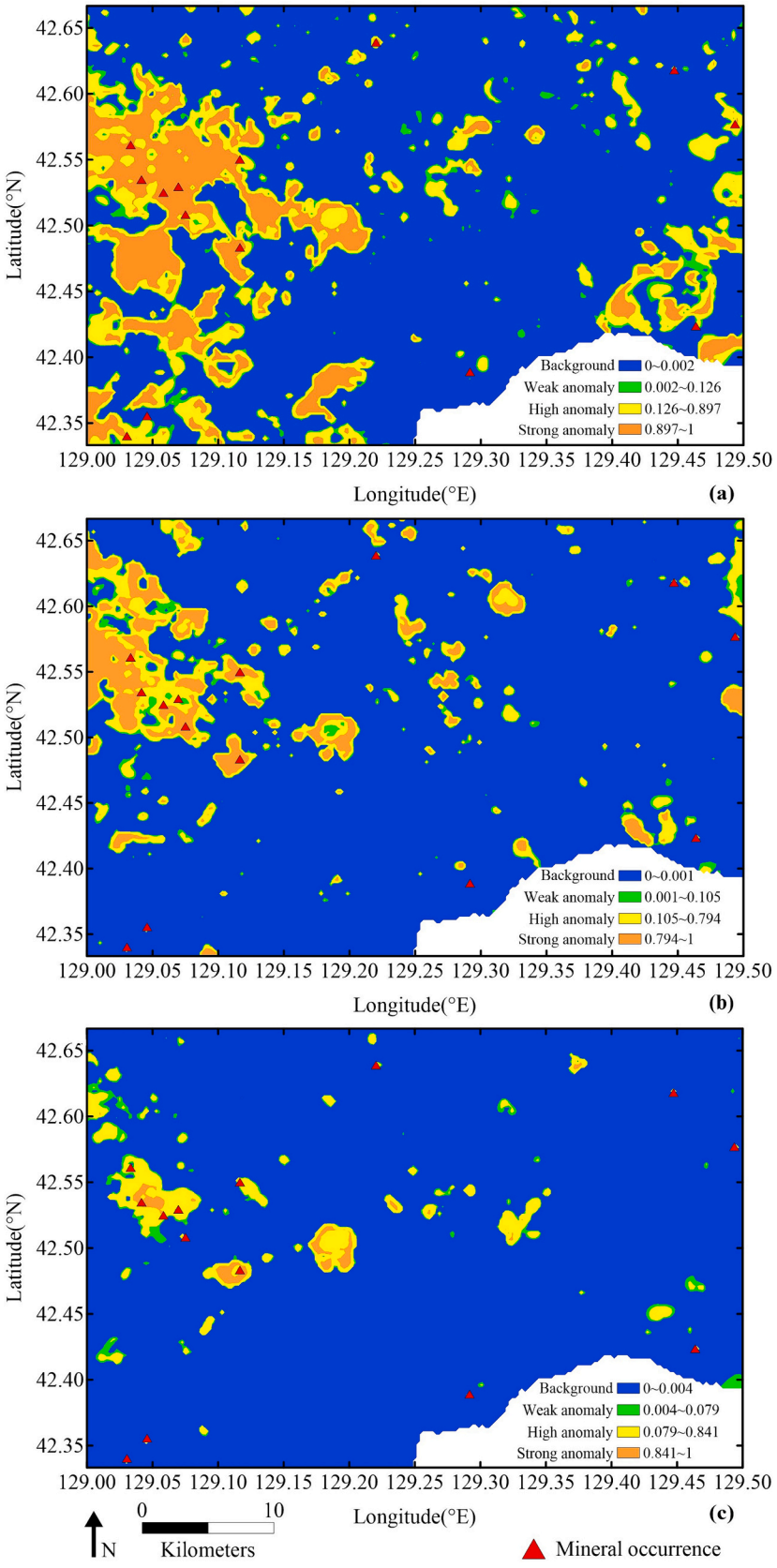


**Fig. 12.** Geochemical anomalies associated with polymetallic mineralization classified by the three thresholds determined by the fractal model: (a) the SMOTE-augmented bagging ensemble model of ELMs, (b) the GAN-augmented bagging ensemble model of ELMs and (c) the SMOTified-GAN-augmented bagging ensemble model of ELMs.

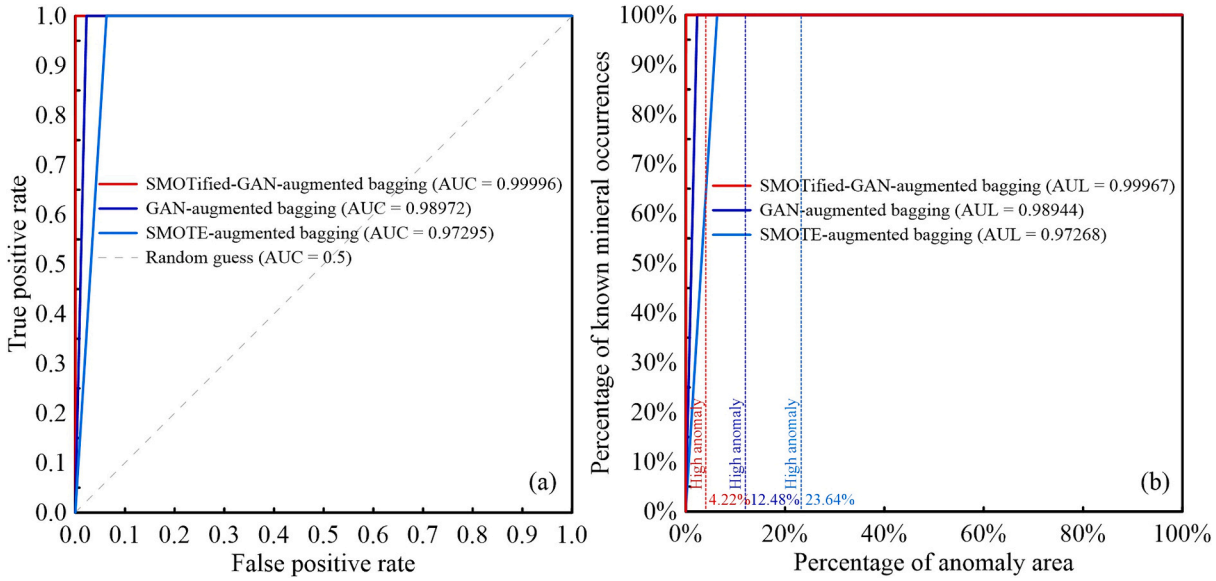


**Fig. 13.** The ROC curves (a) and accumulative gain curves (b) of the bagging ensemble models augmented by SMOTified-GAN, GAN and SMOTE.

**Table 2**
Statistical indices of the three bagging ensemble classification models.

| Model | AUC | $Z_{AUC}$ | Time(s) |
|---|---|---|---|
| SGABE | 0.99996 | 413.38541 | 31.08673 |
| GABE | 0.98972 | 25.75636 | 31.33171 |
| SABE | 0.97295 | 15.53214 | 34.53458 |

Note: SGABE is the abbreviation of SMOTified-GAN-augmented bagging ensemble; GABE is the abbreviation of GAN-augmented bagging ensemble; and SABE is the abbreviation of SMOTE-augmented bagging ensemble.

granodiorite rocks in the western part of the study area. This high anomaly zone contains four known polymetallic deposits, indicating that the high anomaly zone has a strong spatial correlation with the known polymetallic deposits. This anomaly zone is controlled by a NW-trending fault and spreads along the NW-direction. There are a quartzite vein and a gabbroic vein in the anomaly zone. According to Chai and Liu (2015), quartzite vein is closely related to Au—Ag mineralization, and gabbro vein is closely related to Ni—Co mineralization. The second high anomaly zone is located in the sporadic Late Archean Ji'nan Formation in the southwestern part of the study area. In the vicinity of the high anomaly zone, porphyry biotite granodiorite and meso-fine monzonite granite are developed, which are closely related to Mo mineralization. A gabbroic vein is situated at the boundary of this anomaly zone. There is a known polymetallic deposit in the anomaly zone, indicating that this anomaly zone is spatially correlated with known polymetallic deposits. The third high anomaly zone is located in the center of the study area, east of the Lower Mesozoic Cretaceous Quanshuicum Formation. There

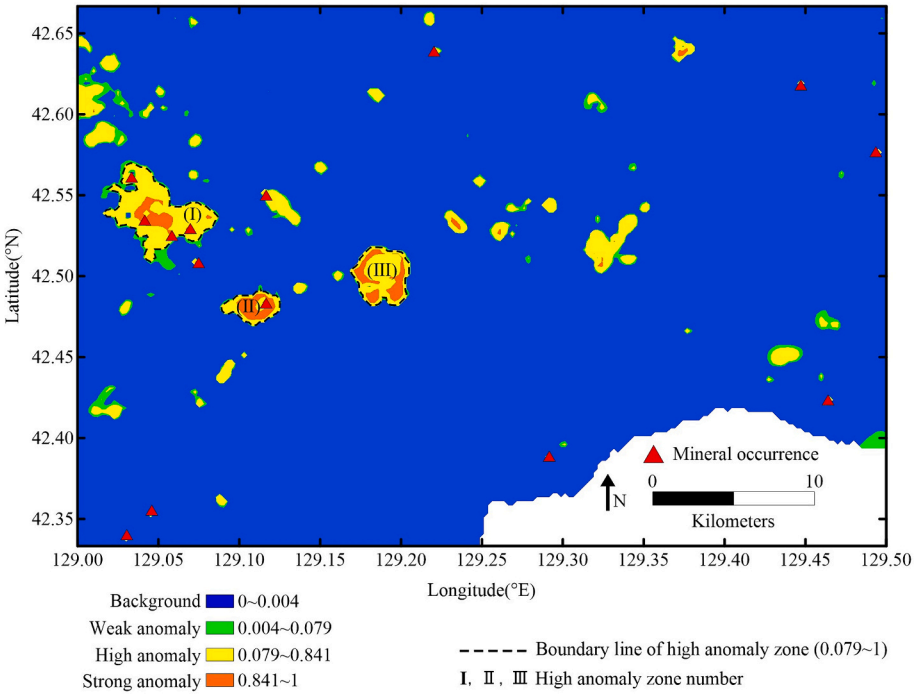


**Fig. 14.** Three high anomaly zones identified by the SMOTified-GAN-augmented bagging ensemble model of ELMs.

are medium-fine monzogranite and porphyaceous moyite rocks near to the anomaly zone. The anomaly zone does not contain any known polymetallic deposit, but contains a quartzite vein, a diorite-porphyrite vein, and a granite pegmatite vein. In addition, the anomaly zone lies on a regional fault.

The main controlling factors for polymetallic mineralization in the study area are regional deep faults, Archean tectonics, and Yanshanian magmatic rocks, as mentioned in Section 3.1. All the three high anomaly zones contain intrusive rocks, indicating that there are certain magmatic activities in these three high anomaly zones, which is conducive to the transportation and enrichment of mineral-bearing materials. Therefore, the three high anomaly zones have a good consistency with Yanshanian magmatic rocks, which are one of the basic controlling factors of polymetallic mineralization. In addition, the first two anomaly zones are located in the Late Archean Ji'nan Formation, which is one of the controlling factors for polymetallic mineralization. The second and third high anomaly zones lie on regional faults, indicating that the two high anomaly zones are in good agreement with the regional deep faults, and the faults may play the roles of ore conducting and hosting. In conclusion, the geochemical anomalies associated with polymetallic mineralization identified in the study area are highly consistent with the regional geological and polymetallic metallogenic characteristics of the study area, and are spatially associated with the three polymetallic mineralization controlling factors in the study area.

### *5.3. Advantages and limitations of SMOTified-GAN augmented bagging ensemble*

The SMOTified-GAN-augmented bagging ensemble is a hybrid model combining the SMOTified-GAN oversampling technique, bagging strategy and simple ELM classifiers, which fully exploits the advantages of the three individual components and efficiently identifies geochemical anomalies associated with polymetallic mineralization in the study area. The hybrid model has several advantages: (a) the SMOTified-GAN method can generate a class-balanced training data set to prevent the trained classification model from bias to the majority class, thus improving the classification performance of the model; (b) simple ELM classifiers have fast learning speed, no need for iterative optimization, and good generalization performance; and (c) bagging strategy can build a high-performance ensemble model from a set of low-performance base classifiers, thus improving both the classification performance and robustness of classification model. In summary, the SMOTified-GAN-augmented bagging ensemble model has a strong generalization, high efficiency and high classification performance in the detection of geochemical anomalies associated with mineralization.

To test the robustness of the SMOTified-GAN-augmented bagging ensemble, five experiments were conducted independently with the same initialization conditions for calculating the AUCs and $Z_{AUC}$s of the SMOTified-GAN-augmented bagging ensemble, and the results of the five experiments are listed in Table 3. It can be seen from Table 3 that the AUCs of the SMOTified-GAN-augmented bagging ensemble in the five experiments are different. However, the difference in AUC values is very small, the difference between the maximum and minimum AUC values is only 0.00741. According to Alina and Chen (2024), the difference between the maximum and minimum AUC values is 0.1 for the ELM model in five repetitions, and the difference between the maximum and minimum AUC values is 0.014 for the SMOTified ELM model in five repetitions. Therefore, compared with the ELM and SMOTified ELM models, the SMOTified-GAN-augmented bagging ensemble model has better robustness. However, the robustness of the SMOTified-GAN-augmented bagging ensemble model is still a question that has not been well resolved and will require further investigation in the future. Another limitation of the SMOTified-GAN-augmented bagging ensemble model is that it can only be built in study areas where some known deposits have been found; Otherwise, it is impossible to define positive samples to build a supervised classification model.

**Table 3**
AUCs and $Z_{AUC}$s of the SMOTified-GAN-augmented bagging ensemble in 5 repetitions.

| Repetition | SGABE_1 | SGABE_2 | SGABE_3 | SGABE_4 | SGABE_5 |
|---|---|---|---|---|---|
| AUC | 0.99996 | 0.99904 | 0.99255 | 0.99570 | 0.99863 |
| $Z_{AUC}$ | 413.38541 | 85.17656 | 30.35674 | 40.12680 | 71.30312 |

Note: SGABE – SMOTified-GAN-augmented bagging ensemble.

## 6. Conclusion

By combining oversampling techniques and bagging strategy with extreme learning machines, a SMOTified-GAN-augmented bagging ensemble model of extreme learning machines, GAN-augmented bagging ensemble model of extreme learning machines and SMOTE-augmented bagging ensemble model of extreme learning machines were constructed on geochemical exploration data. The three bagging ensemble models were compared in detecting geochemical anomalies associated with polymetallic mineralization in the Helong area, Jilin Province, China. In term of receiver operating characteristic curves and the area under the receiver operating characteristic curve, the SMOTified-GAN-augmented bagging ensemble model is superior to the other two models, and the GAN-augmented bagging ensemble model is superior to the SMOTE-augmented bagging ensemble model. Therefore, when building supervised classification models for the detection of geochemical anomalies associated with mineralization, the SMOTified GAN oversampling technique and bagging strategy can be combined with extreme learning machines to build an efficient high-performance ensemble model.

The polymetallic mineralization anomalies detected in the study area by the SMOTified-GAN-augmented bagging ensemble model of extreme learning machines are primarily distributed in the regional deep faults, Archean metamorphic rocks and Hercynian-Yanshan magmatic complex. In addition, the geochemical anomalies associated with polymetallic mineralization detected in the study area contain most known polymetallic deposits found in the study area. It can be concluded that the geochemical anomalies associated with polymetallic mineralization detected in the study area are highly consistent with the regional geological and metallogenic characteristics. In future mineral exploration, attention should be paid to the areas where the geochemical anomalies associated with polymetallic mineralization coexist with regional metallogenic control factors, such as regional deep faults, Archean metamorphic rocks and Hercyn-Yanshan magmatic complex.

## CRediT authorship contribution statement

**Min Guo:** Writing – original draft, Visualization, Validation, Formal analysis. **Yongliang Chen:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

## Declaration of competing interest

The authors declared that there is no conflict interest.

## Acknowledgments

The author thanks Professor Qingying Zhao from Jilin University for her helps in collecting geological and geochemical data. This research was supported by the National Natural Science Foundation of China (Grant no. 42172324).

## References

Ali-Gombe, A., Elyan, E., 2019. MFC-GAN: class-imbalanced dataset classification using multiple fake class generative adversarial network. Neurocomputing 361, 212–221.

Alina, S., Chen, Y.L., 2024. A SMOTified extreme learning machine for identifying mineralization anomalies from geochemical exploration data: a case study from the Yeniugou area, Xinjiang, China. Earth Sci. Inf. 17, 1329–1343.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.
Cao, M.X., Yin, D.M., Zhong, Y., Lv, Y., Lu, L.J., 2023. Detection of geochemical anomalies related to mineralization using the random Forest model optimized by the competitive mechanism and beetle antennae search. J. Geochem. Explor. 249, 107195.
Chai, S.L., Liu, Z.H., 2015. Experimental Demonstration on 1:50,000 Scale Mineral Geology Survey of Four Geological Maps in the Helong Area, Jilin Province (China). Mineral Geology Survey Report, Jilin University (Changchun, China). (In Chinese Without English Abstract).
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16, 321–357.
Chen, Y.L., 2015. Mineral potential mapping with a restricted Boltzmann machine. Ore Geol. Rev. 71, 749–760.
Chen, J.X., Chen, Y.L., 2023. A high-performance voting-based ensemble model of graph convolutional extreme learning machines for identifying geochemical anomalies related to mineralization. Ore Geol. Rev. 162, 105706.
Chen, Y.L., Wu, W., 2016. A prospecting cost-benefit strategy for mineral potential mapping based on ROC curve analysis. Ore Geol. Rev. 74, 26–38.
Chen, Y.L., Wu, W., 2017. Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. Geochemistry Exploration Environment Analysis 17, 231–238.
Chen, Y.L., Wu, W., 2019. Separation of geochemical anomalies from the sample data of unknown distribution population using Gaussian mixture model. Comput. Geosci. 125, 9–18.
Chen, Y.L., Sun, G.S., Zhao, Q.Y., Wang, S.C., 2021. Detection of multivariate geochemical anomalies using the bat-optimized isolation forest and bat-optimized elliptic envelope models. J. Earth Sci. 32 (2), 415–426.
Chen, Y.L., Du, X.D., Guo, M., 2023. Self-paced ensemble for constructing an efficient robust high-performance classification model for detecting mineralization anomalies from geochemical exploration data. Ore Geol. Rev. 157, 105418.
Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32, 314–324.
Cheng, Q., Agterberg, F.P., Bonham-Carter, G.F., 1996. A spatial analysis method for geochemical anomaly separation. J. Geochem. Explor. 56 (3), 183–195.
Ding, S., Zhao, H., Zhang, Y., Xu, X., Nie, R., 2015. Extreme learning machine: algorithm, theory and applications. Artif. Intell. Rev. 44, 103–115.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap.
Farahbakhsh, E., Maughan, J., Müller, R.D., 2023. Prospectivity modelling of critical mineral deposits using a generative adversarial network with oversampling and positive-unlabeled bagging. Ore Geol. Rev. 162, 105665.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 8 (27), 861–874.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets. Neural Information Processing Systems. MIT Press.
Hanley, J.A., Mcneil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.
Huang, G.B., Chen, L., 2007. Convex incremental extreme learning machine. Neurocomputing 70 (16–18), 3056–3062.
Huang, G.B., Chen, L., 2008. Enhanced random search based incremental extreme learning machine. Neurocomputing 71 (16–18), 3460–3468.
Huang, G.B., Zhu, Q.Y., Siew, C.K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. Proceedings of International Joint Conference on Neural Networks (IJCNN) 985–990.
Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1–3), 489–501.
Huang, G.B., Zhou, H.M., Ding, X.J., Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society 42 (2), 513–529.
Huang, D.Z., Zuo, R.G., Wang, J., 2022. Geochemical anomaly identification and uncertainty quantification using a Bayesian convolutional neural network model. Appl. Geochem. 146, 105450.
Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research 18, 559–563.
Li, H., Li, X., Yuan, F., Jowitt, S.M., Zhang, M., Zhou, J., Wu, B., 2020. Convolutional neural network and transfer learning based mineral prospectivity modeling for geochemical exploration of Au mineralization within the Guandian–Zhangbaling area, Anhui Province, China. Applied Geochemistry 122, 104747.
Li, T., Zuo, R.G., Xiong, Y.H., Peng, Y., 2021. Random-drop data augmentation of deep convolutional neural network for mineral prospectivity mapping. Nat. Resour. Res. 30 (1), 27–38.
Liu, F.S., Zhang, M.L., 1999. Complete quality management of the new-round land resources survey. China Geology 8, 20-21+48 (In Chinese).
Moore, E.H., 1920. On the reciprocal of the general algebraic matrix. Bull. Am. Math. Soc. 26 (9), 394–395.
Nykänen, V., Lahti, I., Niiranen, T., Korhonen, K., 2015. Receiver operating characteristics (ROC) as validation tool for prospectivity models—a magmatic Ni–Cu case study from the Central Lapland Greenstone Belt, northern Finland. Ore Geol. Rev. 71, 853–860.
Nykänen, V., Niiranen, T., Molnár, F., Lahti, I., Korhonen, K., Cook, N., Skyttä, P., 2017. Optimizing a knowledge-driven prospectivity model for gold deposits within Peräpohja Belt, northern Finland. Natural Resources Research 26, 571–584.
Pan, Y.D., Xu, B.J., Sun, Y., Hou, L., 2016. Geological features of the Jinchengdong gold deposit in Helong City, Jilin Province, China. Jilin Geology 35, 30–35 (In Chinese with English Abstract).
Parsa, M., 2021. A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted Zn-Pb mineral systems of Western Iran. J. Geochem. Explor. 106811.
Parsa, M., Maghsoudi, A., Yousefi, M., 2017. A receiver operating characteristics-based geochemical data fusion technique for targeting undiscovered mineral deposits. Natural Resources Research 26 (2), 1–14.
Raghuwanshi, B.S., Shukla, S., 2018. Underbagging based reduced kernelized weighted extreme learning machine for class imbalance learning. Eng. Appl. Artif. Intel. 74 (SEP.), 252–270.
Serre, D., 2002. Matrices: Theory and Applications. Springer-Verlag, New York.
Sharma, A., Singh, P.K., Chandra, R., 2022. SMOTified-GAN for class imbalanced pattern classification problems. IEEE Access 10, 30655–30665.
Suh, S., Lee, H., Lukowicz, P., Lee, Y.O., 2021. CEGAN: classification enhancement generative adversarial networks for unraveling data imbalance problems. Neural Networks: The Official Journal of the International Neural Network Society 133, 69–86.
Tian, M., Wang, X.Q., Nie, L.S., Zhang, C.S., 2018. Recognition of geochemical anomalies based on geographically weighted regression. J. Geochem. Explor. 190, 381–389.
Wan, W.Z., Wang, J.B., Feng, X.Y., Zhang, H., Jia, N., Zhang, Y.L., 2010. Geological features and prospecting directions of the Heanhe gold deposit in the Helong area, Jilin Province, China. Jilin Geology 29, 71–75 (In Chinese with English Abstract).
Wang, J., Zuo, R.G., 2020. Assessing geochemical anomalies using geographically weighted lasso. Appl. Geochem. 119, 104668.
Wang, Z.Y., Dong, Y.N., Zuo, R.G., 2019a. Mapping geochemical anomalies related to Fe-polymetallic mineralization using the maximum margin metric learning method. Ore Geol. Rev. 107, 258–265.
Wang, Z.Y., Zuo, R.G., Dong, Y.N., 2019b. Mapping geochemical anomalies through integrating random forest and metric learning methods. Natural Resources Research 28 (4), 1285–1298.
Wang, J., Lu, S., Wang, S.H., Zhang, Y.D., 2022. A review on extreme learning machine. Multimed. Tools Appl. 81, 41611–41660.
Wu, W., Chen, Y.L., 2017. Cumulative gain and lift charts for model performance assessment in mineral potential mapping. Global Geology 20 (2), 118–130.
Wu, F.Y., Lin, J.Q., Wilde, S.A., Zhang, X.O., Yang, J.H., 2005. Nature and significance of the early cretaceous Giant igneous event in eastern China. Earth Planet. Sci. Lett. 233 (1/2), 103–119.
Wu, P.F., Sun, D.Y., Wang, T.H., Gou, J., Li, R., Liu, W., Liu, X.M., 2013. Chronology, geochemical characteristic and Petrogenesis analysis of diorite in Helong of Yanbian area, northeastern China. Geol. J. China Univ. 19 (4), 600–610 (In Chinese with English Abstract).
Xiong, Y.H., Zuo, R.G., 2018. GIS-based rare events logistic regression for mineral prospectivity mapping. Comput. Geosci. 111, 18–25.
Xiong, Y.H., Zuo, R.G., 2020. Recognizing multivariate geochemical anomalies for mineral exploration by combining deep learning and one-class support vector machine-sciencedirect. Comput. Geosci. 140, 104484.
Yan, D., Li, N., Xu, M., Miao, M.M., 2015. Mineralization characteristics and genesis of the Bailiping silver deposit in Helong City, Jilin Province. Jilin Geology 34, 36–41 (In Chinese with English Abstract).
Yu, J.J., Wang, F., Xu, W.L., Gao, F.H., Pei, G.P., 2012. Early Jurassic mafic magmatism in the Lesser Xingan-Zhangguangcai Range, NE China, and its tectonic implications: constraints from zircon U-Pb chronology and geochemistry. Lithos 142-143, 256–266.
Zhang, Y.B., Wu, F.Y., Wilde, S.A., Zhai, M.G., Lu, X.P., Sun, D.Y., 2004. Zircon U-Pb ages and tectonic implications of 'early Paleozoic' Granitoids at Yanbian, Jilin Province, Northeast China. The Island Arc 13 (4), 484–505.
Zhang, L., Yang, H., Jiang, Z., 2018. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. Biomed. Eng. Online 17 (1), 181.
Zhang, Z.J., Zuo, R.G., Xiong, Y.H., 2021. Detection of the multivariate geochemical anomalies associated with mineralization using a deep convolutional neural network and a pixel-pair feature method. Appl. Geochem. 130, 104994.
Zuo, R.G., 2017. Machine learning of mineralization-related geochemical anomalies: a review of potential methods. Natural Resources Research 26, 457–464.
Zuo, R.G., 2018. Selection of an elemental association related to mineralization using spatial analysis. J. Geochem. Explor. 184, 150–157.