Mineral prospectivity mapping of tungsten polymetallic deposits using machine learning algorithms and comparison of their performance in the Gannan region, China

Yonghang Lou, Yue Liu*

Faculty of Earth Resources; State Key Laboratory of Geological Process and Mineral Resources, China University of Geosciences, Wuhan 430074, China

Corresponding author: Yue Liu (liuyue7870@cug.edu.cn)

Key Points:

- Five machine learning models were used and compared for tungsten prospectivity analysis in the Gannan region, China
- Performances of the machine learning models were evaluated through ROC, P-A plot and K-fold cross validation
- CNN and LGBM models are more efficient in mapping favorable areas for tungsten exploration.

Abstract

The current study aimed at assessing the capabilities of five machine learning models in term of mapping tungsten polymetallic prospectivity in the Gannan region of China. The five models include logistic regression (LR), support vector machine (SVM), random forest (RF), convolutional neural network (CNN), and light gradient boosting machine (LGBM) models. Geochemical, lithostratigraphic, and structural datasets were used to generate 16 evidential maps, which were integrated into the machine learning models. Tungsten polymetallic deposits were randomly separated into two parts: 80% for training and 20% for validating. Performances of the models were evaluated through receiver operating characteristic (ROC) and K-fold cross validation, with an emphasis on the variable influence within different machine learning methods. The results show that the models are especially sensitive to the chemical elements: Be, Bi, Pb and Cd, implying that these are closely related to tungsten polymetallic mineralization. Compared to other models, the LGBM and CNN models performed best, while the LR model was the most stable. The results also indicated that the CNN model can predict maximum known deposits within a minimum area, based on the prediction-area plot analysis of the five models, while the RF model can capture the most well-known deposits within the smallest study area. Finally, eighteen prospective areas were delineated according to the predicting results of the machine learning models, which will provide important guidance for further tungsten polymetallic exploration and associated studies.

1 Introduction

Mineral prospectivity mapping (MPM) is concerned with quantifying and mapping the likelihood that mineral deposits are present at a certain location, which require the application of diverse methods and techniques to integrate multi-sources spatial geoscience datasets (Carranza 2009, 2017). The application of machine learning methods for MPM improves efficiency and accuracy of mineral exploration and has become a major trend (Li et al., 2020; Wang et al., 2020a; Liu et al. 2022). Machine learning is an important branch of artificial intelligence, and has a strong ability for mineral prediction. With the introduction of big data technology and artificial intelligence into geosciences, various machine learning and deep learning methods have been developed for quantitative assessment and modeling of mineral resources, such as logistic regression (LR) (Xiong and Zuo, 2018), support vector machine (SVM) (Maepa et al., 2021; Ghezelbash et al., 2021), decision tree (Sun et al., 2022), random forest (RF) (Carranza & Laborte 2015a, b; Sun et al., 2019), convolutional neural network (CNN) (Li et al., 2022), maximum entropy (Liu et al., 2018), Boltzmann machine (Chen, 2015), and extreme learning machine regression (Chen & Wu, 2017). These methods provide

powerful in modeling of mineral prospectivity by handling complex and nonlinear relationships among multisource geospatial datasets. Many successful case studies have made the use of machine learning the most typical data mining paradigm in mineral prospectivity analysis.

In the present study, five advanced machine learning methods including LR, SVM, RF, CNN and the optical gradient propulsion machine (LGBM), were introduced for tungsten polymetallic prospecitivity mapping in the Gannan region. The LR is a traditional machine learning method that can deal with linear relationships between evidential variable, while the relationships between geological layers are nonlinear. When the LR model is combined with other machine learning algorithms, its performance can be improved as demonstrated by previous studies (Chen & Zhao, 2021; Kost et al., 2021; Sinaice et al., 2021). The SVM method based on Vapnik-Chervonenkis (VC) dimension theory and the principle of structural risk minimization, aims to find the global optimal solution, balance the complexity and fitting accuracy of the model, and avoid over-fitting problems (Zhou et al., 2017; Meng et al., 2019; Ghezelbash et al., 2019; Wang et al., 2020a). The RF is a popular machine leaning method that is frequently used frequently for mineral prospectivity analysis mainly due to the following advantages: it is relatively robust to outliers; it combines continuous and categorical data as variable inputs; and it can overcome black-box limitations of artificial neural networks (Rodriguez et al., 2015; Youssef et al., 2016; Wang et al., 2020b; Xiang et al., 2020; Parsa & Maghsoudi, 2021). The CNN is an efficient deep learning method, which can process complex data through its high capacity of feature learning and convolutional filtering (Lecun et al., 1998; Li et al., 2021). The LGBM is a new machine learning method that has a fast running speed, low memory consumption, and is widely used in biology, medicine, chemistry, social science (Sun et al., 2020b; Zafari et al., 2020; Ou et al., 2022; Vaulet et al., 2022), and lithofacies classification. However, it is rarely used to integrate multisource geospatial data for MPM.

The Gannan region is situated at the southern part of the Jiangxi Province and the eastern part of the Nanling structural belt, where endows with advantageous metallogenic geological conditions and have discovered a large number of tungsten polymetallic resources (Yu et al., 2010; Chen et al., 2018; Legros et al., 2020; Zhang et al., 2021). Mesozoic granites have been studied extensively in the Gannan region owing to their close relationships with tungsten polymetallic mineralization (Yu et al., 2010; Shu et al., 2011; Liu et al., 2013, 2015; Wang et al., 2017; Zhang et al., 2017a; Guo et al., 2020; Wang et al., 2020c; Cao et al., 2021; Wu et al., 2021; Li, T., 2022). The study aimed to compare the performance and robustness of the five machine learning algorithms, and further, to determine the most appropriate model in terms of mapping favorable areas for tungsten polymetallic exploration in the Gannan region.

The performances of the individual machine learning models were evaluated by multiple performance metrics including receiver operating characteristic (ROC) curve, K-fold cross validation, and prediction–area (P-A) plot analysis. Additionally, the influence of the variables was compared to determine the optimum ore-controlling variables. Three types of datasets were considered in this study (lithostratigraphic, fault, and geochemistry data), from which 16 variables were derived to construct machine learning models. In addition, these predicting maps obtained from the five machine learning models were used to identify and delineate favorable areas in the study area, so that decision makers can grasp the basic information of regional mineral resource potential and make a better plan for future exploration and development.

2 Study Area

2.1 Geological Setting

Gannan region is located in the eastern part of the Nanling structural belt has undergone complex tectonic-magmatic activities, and is dominated by large folds and faults that occurred mainly in the early Paleozoic, Triassic, and late Jurassic eras (Figure 1). In the early Paleozoic and Triassic eras, the region was affected by the Tethys domain and Hercynian Indosinian orogeny, resulting in the development of NW-SE and S-N trending faults. In the Late Triassic, the continental collision between the North and South China blocks led to the formation of a new subduction system, and the formation of larger-scale W-Sn deposits in the Gannan metallogenic belt (Mao et al., 2013a, Yuan et al., 2019). In the late Jurassic, the region was affected by the Yanshan movement, and the fold changed to NE trending. Concurrently, the region was subducted and compressed by the Pacific plate and intracontinental orogeny, and finally formed a basin-mountain system where the base alternates with granitic volcanic intrusive rocks.

The strata of the Gannan region are composed of Sinian to Quaternary strata. The deepest sedimentary strata consist of feldspar quartz sandstone, siltstone, silty slate, phyllite, and shale ranging from Sinian to Silurian, characterized by a super thick sub-deep sea flysch clastic rock structure. Metamorphic rock turbidite sediments exist from the Ordovician period. The strata cover a large area of Devonian~Triassic carbonate rocks and sandstones, and develop coastal facies, shallow sea facies, and marine continental facies. The stratum is thinner than bedrock as a caprock. The Sinian to Ordovician and Devonian strata are considered source beds of Yanshanian granite tungsten deposits owing to high tungsten content (Chen et al., 2018). Above these strata, Jurassic~Cretaceous clastic and volcanic rocks and red beds were deposited in rift basins (Xu, X. S., 2005; Cao et al., 2018; Jiang et al., 2018). However,

because of Caledonian uplift strata, the Silurian strata in Gannan were denuded and has therefore been removed over time.

During the Caledonian, Hercynian, Indosinian, and Yanshanian periods, three magmatic intrusion cycles occurred in Gannan. The Yanshanian magmatic intrusion cycle was the most frequent, forming granite widely in Gannan. This mineralization is related to Mesozoic granitoid magmatism (e.g., Chen et al., 2013; Mao et al., 2013b; Cai et al., 2017; Xie et al., 2018). Scholars have found that the Mesozoic granite is closely related to mineral resources such as tungsten, tin, molybdenum, bismuth, copper, lead, zinc, silver, rare earth, rare metals, and uranium (Wang et al., 2020b). In the Gannan region, granites are mainly acid granites, with few intermediate acid and basic rocks (mainly monzogranite, and porphyritic, biotite, and migmatite granites). The main minerals are quartz, potassium feldspar and plagioclase (Chen et al., 2018). The granites in the Nanling area show multi-stage characteristics. In the Caledonian period, the study area predominantly exposed the Ordovician and Silurian monzogranite, biotite and migmatite granite. The study area exposes Indosinian Middle Triassic monzogranite, in a small distribution area. In the Yanshanian period, the area exposed Cretaceous and Jurassic biotite granite and monzogranite, with the largest distribution area.

2.2 Tungsten polymetallic mineralization

The Nanling belt is one of the most important tungsten polymetallic metallogenic belts in the world, and forms many ore deposits, such as Dachang, Shizhuyuan, Qitianling, Dajishan, and Xihuashan deposits (Hu & Zhou, 2012; Liu et al., 2013, 2014). Tungsten polymetallic mineralization in the region was mainly controlled by fold, fault, and multi-stage magmatic activities. Ore-controlling faults are generally NE and EW trending. Tungsten mineralized bodies are mainly skarn type (skarn froze complex type) and quartz vein type (Mao et al., 2011; Hu & Zhou, 2012; Chen et al., 2013). Spatially, tungsten polymetallic deposits are mainly distributed in the contacts of the late Jurassic granite intrusions and its surrounding rocks (such as siliciclastic rock and carbonate rock) (Liu et al. 2013, 2014, 2015). Quartz vein-type tungsten polymetallic deposits which occur in the Gannan region are closely related to Yanshanian granites (Chen et al., 2013; Xiong et al., 2020). Metal minerals are mainly predominantly wolframite, cassiterite, molybdenite, bismuth pyroxene, scheelite, pyrite, cinnabar, and chalcopyrite (Cao et al., 2021). Ore minerals are predominantly quartz, fluorite, feldspar, wollastonite, topaz, sericite, muscovite, and calcite. Associated beneficial components usually contain tin, molybdenum, bismuth, plumbum, and silver. The skarn type tungsten polymetallic ore body is strictly controlled by the contact zone of the rock mass, and usually occurs intermittently along the direction of the contact zone. Large-scale wide bodies are often developed in the concave of the rock mass. The main ore

minerals are scheelite, galena, and sphalerite, and the secondary metal minerals are molybdenite, bismuth pyroxene, pyrite, magnetite, and chalcopyrite.

The Yanshanian and Indosinian periods were the main stage of Tungsten polymetallic mineralization in the Nanling belt (Liu et al, 2014). The Jurassic era is the most important period of formation for the tungsten polymetallic deposits (Guo et al., 2020). The Gannan region has widespread Jurassic batholiths including granodiorite, biotite granite, two-mica granite, A-type granite, minor gabbro, and syenite (Zhang et al., 2017b). The Yanshanian ore-forming event has typical characteristics of a short duration in Gannan. Therefore, in this area, mineralization processes show unexpected, unique, and complex characteristics.

3 Materials and Data Preprocessing

The datasets used in this study came from the Chinese Geological Survey (CGS), including stream sediment geochemical data, rock mass, faults, and tungsten polymetallic deposits. An overview of the MPM procedures is shown in Figure 2.

3.1 Geochemical criteria

In terms of geochemical elements, there are many geochemical elements related to tungsten polymetallic deposits. In this study, the factor load matrix of geochemical elements was obtained by factor analysis (FA) of the geochemical data and can obtain the biplot of component 1 (F1) vs. component 2 (F2) (Figure 3).

Figure 3 shows that elements with large component 2 values have a high correlation with mineralization. A total of 11 chemical elements (Be, Bi, F, Li, Pb, Sn, Zn, W, Y, Cd and U) with values greater than 0.5 (considered to be more important in the same composition) were extracted to construct the dataset. Most of the extracted geochemical elements were similar to previous research results (Liu et al., 2013; Wang et al., 2020c), demonstrating that the geochemical element data were reliable and realistic. The 11 selected elements were mapped by means of an inverse distance weighted (IDW) interpolation method, on a 1x1 km cell resolution (Figure 4).

3.2 Granite mass

Mineralization related to magmatic activity in the Nanling region is of a long-term nature and inheritance (Wang et al., 2020b). Tungsten polymetallic deposits are closely related to the distribution of magmatic rocks in the study area. The widely distributed granite in the study area is mainly from the Yanshanian period. Genetically, most of the Nanling region W-Sn-polymetallic deposits have a relationship with Mesozoic granitoids and can be divided into 10

types (Wang et al., 2020b). Using a physically constrained variational automatic encoder (VAE), Xiong et al. (2021) identified geochemical patterns associated with tungsten polymetallic mineralization in the Nanling Mountain Range. The results also indicated that the formation of deposits is closely related to granite intrusions. Using singularity analysis, Chen et al. (2015) identified a weak gravity anomaly caused by buried granites in the Nanling region, which is closely related to the spatial distribution of tungsten polymetallic deposits. These granitic bodies form a system in the Earth's crust that supplies sufficient material and energy for ore deposit formation. The closer the contact zone of rock mass, the higher the possibility of mineral formation. According to the distributions of granite rock mass during the Caledonian, Yanshanian and Indosinian geological movements, granitic rocks were separately buffered outside into eight buffer bands with 1 km interval.

3.3 Faults

Owing to strong tectonic folding in the Yanshanian period, the NE trending faults and fault systems have developed in the study area. The regional distribution, occurrence, shape, and scale of tungsten polymetallic deposits were strictly controlled by the fault structures of this period and to the tectonic activities in the region. The mineral resources are enriched in the areas where the structures developed along the direction of the faults. Chen et al. (2015) found that linear structures, including bends and intersections, are closely related to the spatial distribution of deposits, because most deposits occur along these linear structures. Structure is the main influencing factor on the mineralization of tungsten polymetallic deposits. Moreover, at the fault junctions, structural stress is complex and diverse, and likely to form minerals. High density zones of fault intersections and lineaments have favorable physical characteristics for hosting ore-forming materials. Thus, fault lineaments were buffered into eight buffer bands with 1 km interval (Figure 5a) and fault intersection density was produced using ArcGIS software (Figure 5b).

4 Methods

4.1 Support vector machine

To meet the criterion of structural risk minimization, the SVM method maps data from low to high dimensions, and then locates the line, curve, or hyperplane that can optimally distinguish distinct categories in the data.

There are two types of SVMs: linear support and nonlinear support vector machines. Linear SVMs look for the optimal line that can divide data features in the feature space (Figure 6a). The premise of this method is that the data is linearly separable, as follows:

$$f(x) = (\vec{w}, x) + b \quad (1)$$

where $\vec{w}$ is the normal vector of the classification line, b is the constant term of the classification line, and $h=2/\|\vec{w}\|$ is the classification interval. The points on the upper and lower sides on Figure 6a represents positive and negative samples. The black bold line is the optimal classification line found in theory, and the points passed by the two dotted lines on the upper and lower sides are support vectors. By solving the relationship between y and $(\vec{w}, x) + b$, and maximizing the classification interval h, the accuracy of the data feature division by the optimal classification line can be high. However, most data used in classification problems do not conform to linear separability in practical scenarios. Hence, there is a nonlinear SVM. Through nonlinear transformation, the input variable x is mapped into the high-dimensional space to satisfy the linear segmentation condition and calculate the optimal classification surface (Figure 6b). The kernel function can implement this data mapping transformation, through the widely used Polynomial Kernel Function (Poly), Radial Basis Function (RBF), Sigmoid Kernel Function (Sigmoid), and radial basis function.

4.2 Logistic regression

The method of logistic regression uses continuous and discrete data, and qualitative dependent variables. The principle of this method is that the result of linear regression (-∞, ∞) is mapped to (0,1) by a logical function. The linear regression function is y, where θ represents the different contributions of each variable x, to the dependent variable y. However, the linear regression function can only be used with the quantitative dependent variable. When the dependent variable is qualitative, it is no longer applicable, and the logical regression method overcomes this, using the following formula:

$$y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots + \theta_n X_n = \theta^T x \quad (2)$$

The Sigmoid function is its logical function, which has an "S curve," and has a value range of 0–1. When the independent variable X approaches positive infinity, the dependent variable approaches 1, and X approaches negative infinity, the dependent variable approaches 0. In binary classification, the sigmoid function is used to output the event probability, P is the positive class probability of an event, and 1-P is its negative class probability. This is written as $g(x)$, through logit transformation:

$$g(x) = \frac{1}{1+e^{-x}} \quad (3)$$

Therefore, the functional expression of logistic regression (formula 2) into formula 3 to obtain formula 4, constitutes the logistic regression function:

$$\begin{cases} y = \theta^T x \\ g(x) = \frac{1}{1+e^{-x}} \end{cases} \quad (4)$$

4.3 Random Forest

A bagging algorithm, which was proposed by Beriman (1996), is a method of sample calculations with return. Through *m* random samples of a dataset, *m* new training sets are formed, *m* weak learners are established, and finally set into a strong learner (Figure 7). This ensures that each data sample in the initial data exists in the new dataset and model. Concurrently, the unselected data samples contain *m* out of bag data (OOB), which is used to evaluate the constructed model.

The predecessor of RF is decision tree (DT) model. It selects the data features according to certain conditions to achieve the goal, and it is easy to produce the problem of over-fitting. The RF model can be considered as a set of decision trees which are established randomly, the results of each decision tree are finally synthesized into a single result by means of averaging, voting and majority results (Figure 8). The relationships between the trees are independent with each other. It randomly takes samples from the dataset, each having its own characteristics. The samples are diverse and, therefore, the results are different. It only uses some data features from the randomly selected samples. The method adopts the highly accurate bagging integration algorithm that does not cause over-fitting, and strongly resists the interference of outliers. When the feature dimension is large, because of the unique nature between samples, the RF method can process data efficiently and maintain high accuracy.

4.4 Convolutional neural network

Lecun (1998) proposed the convolution neural network (CNN), LeNet-5, and achieved superior results. A convolutional neural network has the basic characteristics of local connection; the nodes of the convolution layer are only connected with some nodes of the previous layer, and are only used to learn local features (Figure 9a). This increases learning rate speed and avoids overfitting problems. Additionally, the extracted data feature is independent of the data position. It also has aspects of weight sharing, which means that a parameter in the convolution kernel is shared in the same convolution layer (Figure 9b). Thus, the parameters on the convolution layer can be reduced.

This CNN includes input, convolution, pooling, full connection, and output layers (Figure 10). It performs a convolution operation on the convolution layer, which is the core operation of a CNN. Using a fixed size convolution kernel (such as 3×3) to enhance input image data features (horizontally and vertically), the feature map is obtained,

and *n* data maps can be obtained by scanning and recalculating (Figure 9c). Next the activation function is introduced for map operation, which solves nonlinear problems and increases nonlinear expression abilities (also known as the incentive layer). The appropriate activation function can avoid gradient disappearance or explosion problems. An increase in convolution layers increases network speed and accuracy. Common nonlinear activation functions include the Sigmoid, Tanh, and Reluactivation functions. Additionally, pooling operations can reduce model parameters and complexity. There are two common pooling operations (Figure 9d). The first is maximum pooling, which is the maximum value in the local range of the characteristic map as the pooling result. The second is average pooling, which outputs the average characteristics of the local region. Average pooling can collect global features, while maximum pooling can extract the most significant local features, suppress background interference, and improve the invariance of the network to the translation and geometric deformation of the input feature map. When fully connected, every layer is connected to all nodes in the previous layer, and learnt data features from the convolution and pooled layers are integrated, and then mapped to the sample marker location.

4.5 Light Gradient Boosting Machine

The LGBM is an open source and efficient Gradient Boosting Decision Tree (GBDT) algorithm released by Microsoft in 2017. It is an optimization of the Extreme Gradient Boosting (XGBoost) algorithm. The LGBM optimizes the optimal split point search strategy, leaf growth strategy, gradient sampling method, feature attribute binding method, support class feature computation, and parallel learning of the decision tree. This reduces the training time and memory consumption of the model, while ensuring prediction accuracy. This algorithm is iterative and after several iterations, each weak classifier is weighted to become a stronger classifier (Figure 11).

The LGBM is an improvement on the XGBoost and GBDT as follows: (1) The LGBM uses histogram methods to process data, converting continuous values into discrete values, and increase processing speed. It divides the eigenvalues into “bins,” and converts them into *n* integers, to obtain a histogram with a width of *n*, and uses it as index. Then, histograms accumulate the required statistical data, and identify the optimal segmentation point. (2) Generally, the growth of the decision tree is based on a level-wise strategy, which divides all the leaf nodes in the same layer, while traversing the data (Figure 12a). However, when dealing with a large number of complex datasets, this algorithm cannot consider leaf nodes with very little information. The LGBM adopts the leaf growth method with depth limits. This selects the leaf node with the most information after splitting each time, and divides the leaf node further (Figure 12b). This improves accuracy by reducing the splitting of the nodes with little information, thereby reducing

model error. (3) The LGBM model uses a Gradient-based One-Side Sampling (Goss) algorithm in the sampling process. The considerable information of large gradient samples plays an important role in histogram construction. The Goss algorithm randomly samples only the small gradient samples proportionally, only if the distribution structure of the data is invariable. (4) Based on the Exclusive Feature Bundling (EFB) algorithm, the LGBM model can reduce the attribute dimensions and feature number used to construct the histogram. The complexity of the calculation is also simplified.

5 Results and discussion

5.1 Model prediction and mineral mapping

In this study, the obtained datasets were used in the model for training, and the models were debugged by means of hyperparameter optimization and grid search in the Python software. A total of 93 tungsten polymetallic deposit and 71, 179 non-deposit locations were used in the study. The datasets were randomly divided into 80% (for training) and 20% (for tests). Five trained machine models were used to predict the actual data in the study area. The probability of the models predicted identification of tungsten polymetallic deposits was imported into ArcGIS software for prospectivity mapping (Figure 13). The potential targets show a clear spatial relationship with the distribution of known tungsten polymetallic deposits. However, there are still some regions where this do not occur. The five models are fairly similar to the predicting areas of tungsten polymetallic deposits in the study area. The central and western parts of the study area were predicted as highly favorable areas for tungsten polymetallic mineralization almost simultaneously by the five models. However, in the north, there were clear differences among the five models.

5.2 Model Assessment

Receiver Operating Characteristic (ROC) is a curve with false positive rate (FPR) as abscissa and true positive rate (TPR) as ordinate. It is frequently used to evaluate the performance of models in mineral prospectivity mapping (Liu et al. 2014a; Chen & Wu, 2016). Figure 14 indicates that overall, the AUC values of the five models performed well, with all AUC values >0.85 indicated by the training data and validating data. Specifically, the CNN and LGBM had superior classification and prediction performance than the other three models. This is indicated by the ROC curves (AUC values >0.9), while the LR model had the lowest performance (as indicated by the AUC value of 0.85).

The K-fold cross validation is the most common method used in machine learning to evaluate and optimize model accuracy. In this study, the training data were inputted into the five models for K-fold cross validation. The prediction

performance of the model was quantified and compared with the original AUC values (Figure 15). The K-fold AUC and original AUC values ranged from 0.84 to 0.93, which supports the effectiveness and rationality of the constructed data. The results showed that the K-fold AUC of the LGBM model was the highest (0.89), indicating the most reliable results; however, the SVM model had the lowest reliability (0.84). This method showed that the K-fold AUC values of all models were reduced compared with the original AUC values. The AUC value of the CNN model decreased the most, by 0.0811, showing poor stability and implying that it is the most sensitive model among the five models. The LR model was the most robust, with an AUC difference of only 0.0053.

5.3 Variables contribution analysis

The importance of the variables to each model differs. The model assigns different weightings to different variables, which also represents the degree of learning and variable sensitivity. The "Permutation Importance" method can change the data in one column in the test dataset and other columns remain unchanged. It then quantifies the change of the model prediction ability after the data is disrupted, keeping the parameters unchanged after the model training. This method was applied to the five models, and the degree of influence of 16 variables on the prediction results was obtained, which showed the sensitivity of the models to each variable (Figure 16).

All models were highly sensitive to the variable W, which is consistent with other studies. In the LR and SVM models, W, Sn, Cd, Bi, Li, Be and Pb had similar high influences on the predicted results, while Zn was relatively low. Similarly, the two models have similar recognition characteristics and sensitivity of the five models in granite rock mass data and fault data. Conversely, Caledonian granite, Yanshanian granite, and fault distribution had little or no influence in the RF model. The learning results of the five models for these variables are consistent with the previous research conclusion that the anomalies of W, Sn, Cd, Bi, Be, Pb, and other elements are closely related to the tungsten polymetallic deposits in southern Jiangxi (Liu et al., 2014a, 2019; Xiong et al., 2021). Additionally, these elements, specifically W and Sn, have similar geochemical properties and closely relate to the residual magma, resulting from partial melting of the continental crust (Liu et al, 2014a). Table 1 shows the calculated cumulative contribution rate of each type of variable for each model.

The geological chemical elements is the most dominant factor in all models, with a contribution rate >78% and, therefore, this factor has the largest impact on the prediction performance of each model. Among the five models, the RF model is the most sensitive to element data (cumulative contribution rate >97%), indicating that this data is imperative in this model. However, the model may have overlooked other information that has a large impact on the

results, in the process of learning the other two types of data (granite mass and faults). For all models, the influence of granite rock mass data and fault data was different. The combined proportion of these two in the RF model was the lowest (combined total 2.86%). The granite rock mass data and fault data had the highest influence in the SVM model, at 12.83% and 8.61%, respectively (combined total 21.44%). The distribution of these two types of data was similar in the CNN and LGBM models, and both types of data account for approximately 5%. In the LR model, granite data (8.07%) is slightly higher than fault data (4.98%). The contribution rates to the models from these two types of data depends on the models overall learning and data feature extraction. The RF model may reduce the proportion of these two variables excessively, which makes the result less reliable. The proportion distribution of the SVM model was logical.

However, Li et al. (2020b) used the "Synthetic Minority Over Sampling Technology" (SMOTE) method to deal with the imbalance of data classes in the Nanling Mountain Range. Subsequently, the RF model was used to predict the mineralization and achieved good results. However, Sun et al. (2020a) generated three datasets to test the impact of negative sample data using the RF model. It captured 66.95% of the known ore within 9% of the study area, indicating that an imbalance in class can have an extremely large impact. This shows that the RF model may have great disadvantages in dealing with class unbalanced data sets.

5.4 Prediction–area plot

The P–A plot analysis was accepted to determine suitable threshold that can evaluate prediction rate of the tungsten polymetallic deposits with smaller area including more deposits. Many researchers have evaluated the model performance and quality through prediction–area (P-A) plots (e.g., Nezhad et al., 2017; Zhang et al. 2017a; Roshanravan et al., 2018; Liu et al., 2019). Yousefi et al. (2012, 2013) plotted and compared distribution against predicted percentage of known ore deposit. These plots are used to identify mineral deposits based on area size. The smaller the area, the higher the probability of accuracy (Yousefi & Carranza, 2015). The results show the intersection point of the two curves from the five models (Figure 17). The performances of the SVM, RF, CNN and LGBM models are similar; however, the LR model intersection point is closer to the center of the curve.

Using these plots, the proportion of study area corresponding to each intersection point, and the percentage of known ore occurrences were calculated (Table 2). The CNN model achieved the best results, with 81% of the known ore deposit occurrences identified, in approximately 19% of the study area, and 0.57% of known deposits in high

potential areas. The SVM model produced the worst results, identifying 73% of the known ore deposits, in approximately 27% of the study area, and only 0.36% of known deposits in high potential areas.

5.5 Prospectivity analysis

Based on the prediction results of five models, 18 high potential areas for tungsten polymetallic exploration were delineated as shown in the Figure 18. The results show that the areas 12 and 13 are the most important in the study area as they have the widest area and the clearest directivity among the high potential areas. Additionally, they have a high coincidence with the distribution of known deposits. This shows that the metallogenic potential of the two areas has strong reliability. The five models have high potential prediction effects, and most areas capture known deposits in the areas 3, 5 ~7, 9, 11, and 16 ~18. This shows that the potential effect of these areas has a certain reliability. For the areas 1, 2, 8, 10 and 14, the LR, SVM and RF models have similar predicting results. It was concluded that these areas are linearly distributed along the fault, showing a high spatial correlation with the faults. For area 15, in addition to CNN model, other four models predicted high potential metallogenic areas, which corresponds with known deposit areas. The LR, RF, and LGBM models predict that the area 4 is a high potential area for the formation of tungsten deposits.

6 Conclusions

(1) According to ore-controlling factors of tungsten polymetallic deposits in the Gannan region, sixteen evidential variables were produced from three types of datasets including geochemical, fault, and lithostratigraphic data, which were integrated into the LR, SVM, RF, CNN and LGBM models. Our studies indicate that the five models have strong sensitivity to geochemical elements Be, Bi, Pb and Cd in addition to W and Sn, which can be identified by the contribution rates of the variables, implying that these elements are closely related to tungsten polymetallic mineralization and have a symbiotic association system.

(2) The CNN and LGBM models are more efficient in mapping favorable areas for tungsten exploration in the study area based on comparative studies. The RF and CNN models performed well according to the P-A plot. However, the performance of the P-A plot for the LGBM model was poor, with a clear over-fitting phenomenon. The RF model assigns an overly high weighting to geochemical data in the learning variables process, while the fault data may be ignored.

(3) Based on the prediction results of the five models, the study area was divided into 18 favorable areas for tungsten exploration, of which 10 areas have higher reliability and show strong spatial relationships with known deposits. however, the remaining areas need to be investigated and validated further, because there were no clear, unified conclusions from the models.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (42172325) and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUG2106202).

Author Contributions:

Conceptualization: Yue Liu, Yonghang Lou

Methodology: Yonghang Lou, Yue Liu

Writing – original draft: Yonghang Lou

Writing – review & editing: Yue Liu

Supervision: Yue Liu

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The training and testing data sets for tungsten polymetallic prospectivity analysis based on machine learning methods are provided in the online data set https://figshare.com/articles/dataset/Data_and_Python_code_ rar/21120022.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Cai, Y., Feng, Z., Shao, T., Hu, R., Zhou, Y., & Xu, J. (2017). New precise zircon U-Pb and muscovite 40 Ar- 39 Ar geochronology of the Late Cretaceous W-Sn mineralization in the Shanhu orefield, South China. *Ore Geology Reviews, 84*, 338-346.

392 Carranza, E. J. M. (2009). Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and
393 spatial association with geological features. Ore Geology Reviews, 35, 383–400.
394 Carranza, E. J. M. (2017). Natural resources research publications on geochemical anomaly and mineral potential
395 mapping, and introduction to the special issue of papers in these fields. Natural Resources Research, 26(4), 379-
396 410.
397 Carranza, E. J. M., & Laborte, A. G. (2015a). Random forest predictive modeling of mineral prospectivity with small
398 number of prospects and data with missing values in Abra (Philippines). Computers & Geosciences, 74, 60-70.
399 Carranza, E. J. M., & Laborte, A. G. (2015b). Data-driven predictive mapping of gold prospectivity, Baguio district,
400 Philippines: Application of Random Forests algorithm. Ore Geology Reviews, 71, 777-787.
401 Cao, J., Wu, Q., Yang, X., Kong, H., Li, H., Xi, X., Huang, Q., Liu, B. (2018). Geochronology and Genesis of the
402 Xitian W-Sn Polymetallic Deposit in Eastern Hunan Province, South China: Evidence from Zircon U-Pb and
403 Muscovite Ar-Ar Dating, Petrochemistry, and Wolframite Sr-Nd-Pb Isotopes. *Minerals, 8*(3).
404 Cao, J. Y., Li, H., Algeo, T. J., Yang, L. Z., & Tamehe, L. S. (2021). Two-stage magmatism and tungsten mineralization
405 in the Nanling Range, South China: Evidence from the Jurassic Helukou deposit. *American Mineralogist, 106*(9),
406 1488-1502.
407 Chen, G. X., Huang, N., Wu, G. P., Luo, L., Wang, D. T., & Cheng, Q. M. (2022). Mineral prospectivity mapping
408 based on wavelet neural network and Monte Carlo simulations in the Nanling W-Sn metallogenic province. *Ore*
409 *Geology Reviews, 143*.
410 Chen, G. X., Liu, T. Y., Sun, J. S., Cheng, Q. M., Sahoo, B., Zhang, Z. J., & Zhang, H. L. (2015). Gravity method for
411 investigating the geological structures associated with W-Sn polymetallic deposits in the Nanling Range, China.
412 *Journal of Applied Geophysics, 120*, 14-25.
413 Chen, J., Wang, R. C., Zhu, J. C., Lu, J. J., & Ma, D. S. (2013). Multiple-aged granitoids and related tungsten-tin
414 mineralization in the Nanling Range, South China. *Science China-Earth Sciences, 56*(12), 2045-2055.
415 Chen, L.-L., Ni, P., Li, W.-S., Ding, J.-Y., Pan, J.-Y., Wang, G.-G., & Yang, Y.-L. (2018). The link between fluid
416 evolution and vertical zonation at the Maoping tungsten deposit, Southern Jiangxi, China: Fluid inclusion and stable
417 isotope evidence. *Journal of Geochemical Exploration, 192*, 18-32.
418 Chen, Y. L. (2015). Mineral potential mapping with a restricted Boltzmann machine. Ore Geology Reviews, 71, 749–
419 760.

Chen, Y. L., & Wu, W. (2016). A prospecting cost-benefit strategy for mineral potential mapping based on ROC curve analysis. *Ore Geology Reviews, 74*, 26-38.

Chen, Y. L., & Wu, W. (2017). Mapping mineral prospectivity using an extreme learning machine regression. *Ore Geology Reviews, 80*, 200-213.

Chen, Y. L., & Zhao, Q. Y. (2021). Mineral exploration targeting by combination of recursive indicator elimination with the l(2-)regularization logistic regression based on geochemical data. *Ore Geology Reviews, 135*.

Fan, J. L., Ma, X., Wu, L. F., Zhang, F. C., Yu, X., & Zeng, W. Z. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management, 225*.

Ghezelbash, R., Maghsoudi, A., & Carranza, E. J. M. (2019). Performance evaluation of RBF-and SVM-based machine learning algorithms for predictive mineral prospectivity modeling: integration of SA multifractal model and mineralization controls. Earth Science Informatics, 12(3), 277-293.

Ghezelbash, R., Maghsoudi, A., Bigdeli, A., & Carranza, E. J. M. (2021). Regional-scale mineral prospectivity mapping: Support vector machines and an improved data-driven multi-criteria decision-making technique. Natural Resources Research, 30(3), 1977–2005。

Guo, X. F., Wu, K. X., Tian, Y., & Zhang, X. B. (2020). U-Pb geochronology and Raman spectroscopy of zircons from the granites in the Xihuashan and Tieshanlong Deposit: Implications for W-Sn mineralization in Southern Jiangxi Province, South China. *Resource Geology, 70*(2), 141-156.

Hu, R. Z., & Zhou, M. F. (2012). Multiple Mesozoic mineralization events in South China-an introduction to the thematic issue. *Mineralium Deposita, 47*(6), 579-588.

Jiang, W. C., Li, H., Evans, N. J., Wu, J. H., & Cao, J. Y. (2018). Metal Sources of World-Class Polymetallic W-Sn Skarns in the Nanling Range, South China: Granites versus Sedimentary Rocks? *Minerals, 8*(7).

Kost, S., Rheinbach, O., & Schaeben, H. (2021). Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling. *Geochemistry, 81*(4).

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the Ieee, 86*(11), 2278-2324.

Legros, H., Harlaux, M., Mercadier, J., Romer, R. L., Poujol, M., Camacho, A., Marignac, C., Cuney, M., Wang, R. C., Charles, N., & Lespinasse, M. Y. (2020). The world-class Nanling metallogenic belt (Jiangxi, China): W and

Sn deposition at 160 Ma followed by 30 m.y. of hydrothermal metal redistribution. *Ore Geology Reviews, 117*.

Li, S., Chen, J. P., Liu, C., & Wang, Y. (2021). Mineral Prospectivity Prediction via Convolutional Neural Networks Based on Geological Big Data. *Journal of Earth Science, 32*(2), 327-347.

Li, T., Zuo, R. G., Zhao, X. F., & Zhao, K. D. (2022). Mapping prospectivity for regolith-hosted REE deposits via convolutional neural network with generative adversarial network augmented data. *Ore Geology Reviews, 142*.

Li, T. F., Xia, Q. L., Zhao, M. Y., Gui, Z., & Leng, S. (2020). Prospectivity Mapping for Tungsten Polymetallic Mineral Resources, Nanling Metallogenic Belt, South China: Use of Random Forest Algorithm from a Perspective of Data Imbalance. *Natural Resources Research, 29*(1), 203-227.

Liu, L. M., Cao, W., Liu, H. S., Ord, A., Qin, Y. Z., Zhou, F. H., & Bi, C. X. (2022a). Applying benefits and avoiding pitfalls of 3D computational modeling-based machine learning prediction for exploration targeting: Lessons from two mines in the Tongling-Anqing district, eastern China. *Ore Geology Reviews, 142, 104712*.

Liu, X. X., Zhang, J., Huang, F., Cheng, J. W., Lu, K. X., Yang, J. F., Wang, M., Wang, Y. X., Qiu, J. W., & Zhang, X. Y. (2022b). Tungsten deposits in southern Jiangxi Province: Constraints on the origin of wolframite from in-situ U-Pb isotope dating. *Ore Geology Reviews, 143*.

Liu, Y., Carranza, E. J. M., & Xia, Q. (2022). Developments in quantitative assessment and modeling of mineral resource potential: an overview. Natural Resources Research, 31(4),1825-1840.

Liu, Y., Cheng, Q. M., Xia, Q. L., & Wang, X. Q. (2013). Application of singularity analysis for mineral potential identification using geochemical data - A case study: Nanling W-Sn-Mo polymetallic metallogenic belt, South China. *Journal of Geochemical Exploration, 134*, 61-72.

Liu, Y., Cheng, Q. M., Xia, Q. L., & Wang, X. Q. (2014). Mineral potential mapping for tungsten polymetallic deposits in the Nanling metallogenic belt, South China. *Journal of Earth Science, 25*(4), 689-700.

Liu, Y., Cheng, Q. M., Xia, Q. L., & Wang, X. Q. (2015). The use of evidential belief functions for mineral potential mapping in the Nanling belt, South China. *Frontiers of Earth Science, 9*(2), 342-354.

Liu, Y., Zhou, K., & Xia, Q. (2018). A MaxEnt model for mineral prospectivity mapping. Natural Resources Research, 27(3), 299–313.

Liu, Y., Xia, Q., & Carranza, E. J. M. (2019). Integrating sequential indicator simulation and singularity analysis to analyze uncertainty of geochemical anomaly for exploration targeting of tungsten polymetallic mineralization, Nanling belt, South China. *Journal of Geochemical Exploration, 197*, 143-158.

476 Maepa, F., Smith, R. S., & Tessema, A. (2021). Support vector machine and artificial neural network modelling of
477 orogenic gold prospectivity mapping in the Swayze greenstone belt, Ontario, Canada. *Ore Geology Reviews, 130*.
478 Mao, J. W., Cheng, Y. B., Chen, M. H., & Pirajno, F. (2013a). Major types and time-space distribution of Mesozoic
479 ore deposits in South China and their geodynamic settings. *Mineralium Deposita, 48*(3), 267-294.
480 Mao, J. W., Pirajno, F., & Cook, N. (2011). Mesozoic metallogeny in East China and corresponding geodynamic
481 settings - An introduction to the special issue. *Ore Geology Reviews, 43*(1), 1-7.
482 Mao, Z. H., Cheng, Y. B., Liu, J. J., Yuan, S. D., Wu, S. H., Xiang, X. K., & Luo, X. H. (2013b). Geology and
483 molybdenite Re-Os age of the Dahutang granite-related veinlets-disseminated tungsten ore field in the Jiangxin
484 Province, China. *Ore Geology Reviews, 53*, 422-433.
485 Meng, E. H., Huang, S. Z., Huang, Q., Fang, W., Wu, L. Z., & Wang, L. (2019). A robust method for non-stationary
486 streamflow prediction based on improved EMD-SVM model. *Journal of Hydrology, 568*, 462-478.
487 Nezhad, S. G., Mokhtari, A. R., & Rodsari, P. R. (2017). The true sample catchment basin approach in the analysis of
488 stream sediment geochemical data. *Ore Geology Reviews, 83*, 127-134.
489 Ni, P., Wang, X. D., Wang, G. G., Huang, J. B., Pan, J. Y., & Wang, T. G. (2015). An infrared microthermometric study
490 of fluid inclusions in coexisting quartz and wolframite from Late Mesozoic tungsten deposits in the Gannan
491 metallogenic belt, South China. *Ore Geology Reviews, 65*, 1062-1077.
492 Ou, S. M., Lee, K. H., Tsai, M. T., Tseng, W. C., Chu, Y. C., & Tarng, D. C. (2022). Artificial Intelligence for Risk
493 Prediction of Rehospitalization with Acute Kidney Injury in Sepsis Survivors. *Journal of Personalized Medicine,*
494 *12*(1).
495 Parsa, M., & Maghsoudi, A. (2021). Assessing the effects of mineral systems-derived exploration targeting criteria for
496 random Forests-based predictive mapping of mineral prospectivity in Ahar-Arasbaran area, Iran. *Ore Geology*
497 *Reviews, 138*.
498 Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning
499 predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and
500 support vector machines. *Ore Geology Reviews, 71*, 804-818.
501 Roshanravan, B., Aghajani, H., Yousefi, M., & Kreuzer, O. (2018). An Improved Prediction-Area Plot for
502 Prospectivity Analysis of Mineral Deposits. *Natural Resources Research, 28*(3), 1089-1105.
503 Shu, X. J., Wang, X. L., Sun, T., Xu, X. S., & Dai, M. N. (2011). Trace elements, U-Pb ages and Hf isotopes of zircons

from Mesozoic granites in the western Nanling Range, South China: Implications for petrogenesis and W-Sn mineralization. *Lithos, 127*(3-4), 468-482.

Sinaice, B. B., Owada, N., Saadat, M., Toriya, H., Inagaki, F., Bagai, Z., & Kawamura, Y. (2021). Coupling NCA Dimensionality Reduction with Machine Learning in Multispectral Rock Classification Problems. *Minerals, 11*(8).

Sun, G. T., Zeng, Q. D., & Zhou, J. X. (2022). Machine learning coupled with mineral geochemistry reveals the origin of ore deposits. *Ore Geology Reviews, 142*.

Sun, T., Chen, F., Zhong, L. X., Liu, W. M., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. *Ore Geology Reviews, 109*, 26-49.

Sun, T., Li, H., Wu, K. X., Chen, F., Zhu, Z., & Hu, Z. J. (2020a). Data-Driven Predictive Modelling of Mineral Prospectivity Using Machine Learning and Deep Learning Methods: A Case Study from Southern Jiangxi Province, China. *Minerals, 10*(2).

Sun, X. L., Liu, M. X., & Sima, Z. Q. (2020b). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters, 32*.

Vaulet, T., Al-Memar, M., Fourie, H., Bobdiwala, S., Saso, S., Pipi, M., Stalder, C., Bennett, P., Timmerman, D., Bourne, T., & De Moor, B. (2022). Gradient boosted trees with individual explanations: An alternative to logistic regression for viability prediction in the first trimester of pregnancy. *Computer Methods and Programs in Biomedicine, 213*.

Wang, K. J., Zheng, X. Q., Wang, G. W., Liu, D. Y., & Cui, N. (2020a). A Multi-model ensemble approach for gold mineral prospectivity mapping: A Case Study on the Beishan Region, Western China. *Minerals, 10*(12).

Wang, C. B., Pan, Y. P., Chen, J. G., Ouyang, Y. P., Rao, J. F., & Jiang, Q. B. (2020b). Indicator element selection and geochemical anomaly mapping using recursive feature elimination and random forest methods in the Jingdezhen region of Jiangxi Province, South China. *Applied Geochemistry, 122*.

Wang, D. H., Huang, F., Wang, Y., He, H. H., Li, X. M., Liu, X. X., Sheng, J. F., & Liang, T. (2020c). Regional metallogeny of Tungsten-tin-polymetallic deposits in Nanling region, South China. *Ore Geology Reviews, 120*.

Wang, X., Ren, M. H., & Chen, J. (2017). The muscovite granites: Parental rocks to the Nanling Range tungsten mineralization in South China. *Ore Geology Reviews, 88*, 702-717.

Wu, J. H., Kong, H., Li, H., Algeo, T. J., Yonezu, K., Liu, B. A., Wu, Q. H., Zhu, D. P., & Jiang, H. (2021). Multiple

metal sources of coupled Cu-Sn deposits: Insights from the Tongshanling polymetallic deposit in the Nanling Range, South China. *Ore Geology Reviews, 139*.

Xiang, J., Xiao, K. Y., Carranza, E. J. M., Chen, J. P., & Li, S. (2020). 3D Mineral Prospectivity Mapping with Random Forests: A Case Study of Tongling, Anhui, China. *Natural Resources Research, 29*(1), 395-414.

Xie, G., Mao, J., Li, W., Fu, B., & Zhang, Z. (2018). Granite-related Yangjiashan tungsten deposit, southern China. *Mineralium Deposita, 54*(1), 67-80.

Xiong, Y., & Zuo, R. (2018). GIS-based rare events logistic regression for mineral prospectivity mapping. Computers & Geosciences, 111, 18-25.

Xiong, Y. H., Zuo, R. G., Luo, Z. J., & Wang, X. Q. (2021). A Physically Constrained Variational Autoencoder for Geochemical Pattern Recognition. *Mathematical Geosciences*.

Xiong, Y. Q., Shao, Y. J., Cheng, Y. B., & Jiang, S. Y. (2020). Discrete Jurassic and Cretaceous Mineralization Events at the Xiangdong W(-Sn) Deposit, Nanling Range, South China. *Economic Geology, 115*(2), 385-413.

Xu, X. S., O'Reilly, S. Y., Griffin, W. L., Deng, P., & Pearson, N. J. (2005). Relict proterozoic basement in the Nanling Mountains (SE China) and its tectonothermal overprinting. *Tectonics, 24*(2).

Yousefi, M., & Carranza, E. J. M. (2015). Prediction–area (P–A) plot and C–A fractal analysis to classify and evaluate evidential maps for mineral prospectivity modeling. *Computers & Geosciences, 79*, 69-81.

Yousefi, M., Carranza, E. J. M., & Kamkar-Rouhani, A. (2013). Weighted drainage catchment basin mapping of geochemical anomalies using stream sediment data for mineral potential modeling. *Journal of Geochemical Exploration, 128*, 88-96.

Yousefi, M., Kamkar-Rouhani, A., & Carranza, E. J. M. (2012). Geochemical mineralization probability index (GMPI): A new approach to generate enhanced stream sediment geochemical evidential map for increasing probability of success in mineral potential mapping. *Journal of Geochemical Exploration, 115*, 24-35.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides, 13*(5), 839-856.

Yu, X. Q., Wu, G. G., Zhao, X., Gao, J. F., Di, Y. J., Zheng, Y., Dai, Y. P., Li, C. L., & Qiu, J. T. (2010). The Early Jurassic tectono-magmatic events in southern Jiangxi and northern Guangdong provinces, SE China: Constraints from the SHRIMP zircon U–Pb dating. *Journal of Asian Earth Sciences, 39*(5), 408-422.

Yuan, S. D., Williams-Jones, A. E., Romer, R. L., Zhao, P. L., & Mao, J. W. (2019). Protoliyh-related thermal controls on the decoupling of Sn and W in Sn-W metallogenic provinces: insights from the Nanling region, China. *Economic Geology, 114*(5), 1005-1012.

Zafari, M., Kumar, D., Umer, M., & Kim, K. S. (2020). Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts. *Journal of Materials Chemistry A, 8*(10), 5209-5216.

Zhang, J., Liu, X., Li, W., Zeng, Z., Hu, H., Peng, L., Cheng, J., Lu, K., & Li, P. (2021). The metallogenic epoch and geological implications of the tungsten-tin polymetallic deposits in southern Jiangxi Province, China: Constraints from cassiterite U-Pb and molybdenite Re-Os isotopic dating. *Ore Geology Reviews, 134*.

Zhang, R. Q., Lu, J. J., Lehmann, B., Li, C. Y., Li, G. L., Zhang, L. P., Guo, J., & Sun, W. D. (2017a). Combined zircon and cassiterite U-Pb dating of the Piaotang granite-related tungsten-tin deposit, southern Jiangxi tungsten district, China. *Ore Geology Reviews, 82*, 268-284.

Zhang, Y., Yang, J. H., Chen, J. Y., Wang, H., & Xiang, Y. X. (2017b). Petrogenesis of Jurassic tungsten-bearing granites in the Nanling Range, South China: Evidence from whole-rock geochemistry and zircon U-Pb and Hf-O isotopes. *Lithos, 278*, 166-180.

Zhou, T., Wang, F. X., & Yang, Z. (2017). Comparative Analysis of ANN and SVM Models Combined with Wavelet Preprocess for Groundwater Depth Prediction. *Water, 9*(10).

Tables

Table 1. Statistics of cumulative contribution rate of three variable types

	LR	SVM	RF	CNN	LGBM
Geochemical element	86.96%	78.56%	97.14%	90.24%	88.86%
Granite mass	8.07%	12.83%	1.35%	5.09%	5.16%
Fault	4.98%	8.61%	1.51%	4.67%	5.99%

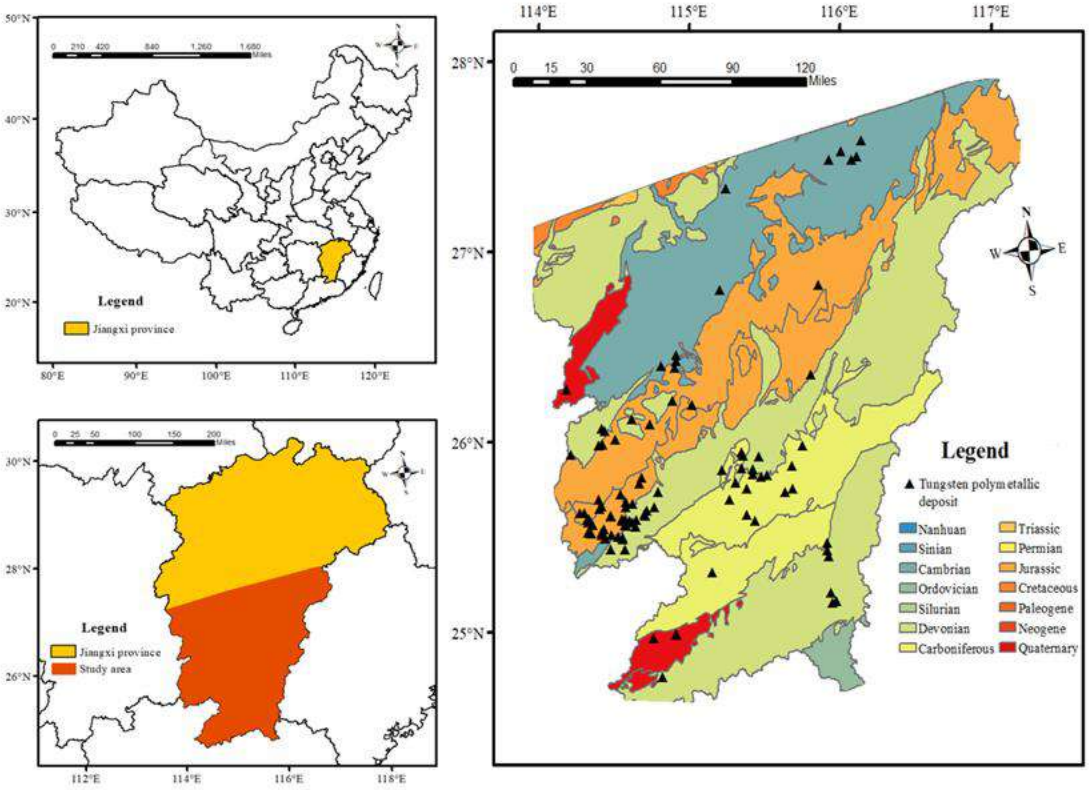
Table 2. Model intersection data of the P-A plots

	LR	SVM	RF	CNN	LGBM
Percentage of study area	22%	27%	22%	19%	25%
Percentage of tungsten polymetallic deposit	78%	73%	78%	81%	75%
Likelihood of deposit occurrence	0.47%	0.36%	0.47%	0.57%	0.40%

Figure captions

612 **Figures**

613



614

615 Figure 1. Geographical location of the study area

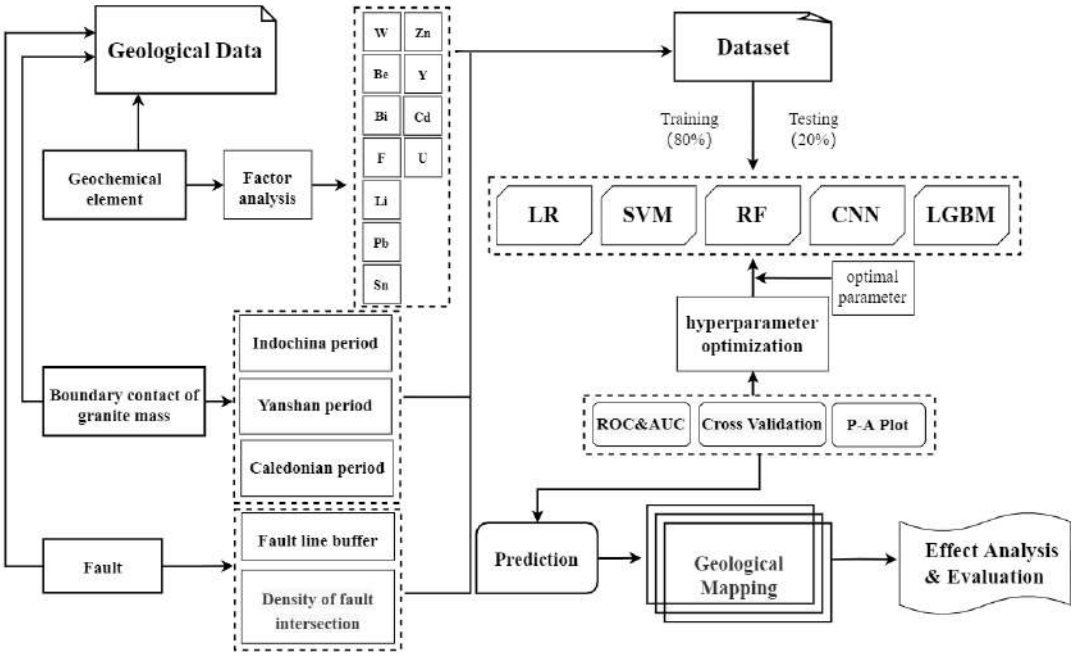


Figure 2. Flow chart of the study showing modelling strategy

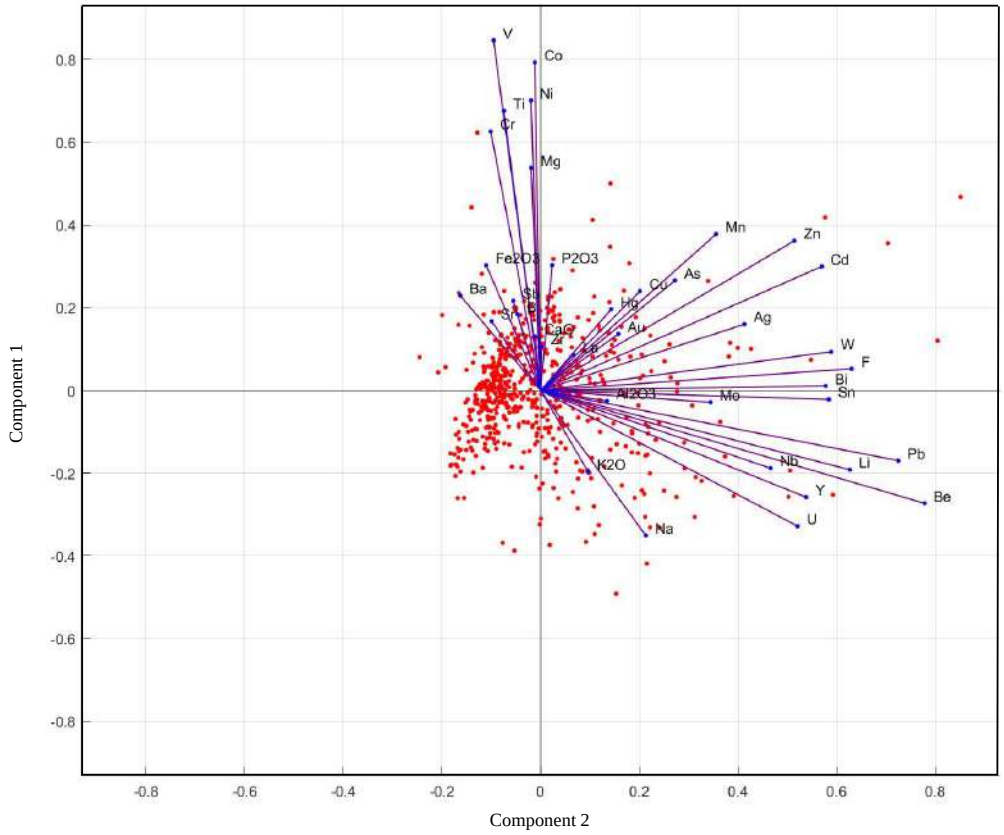


Figure 3. Biplot of F1 vs. F2 elements

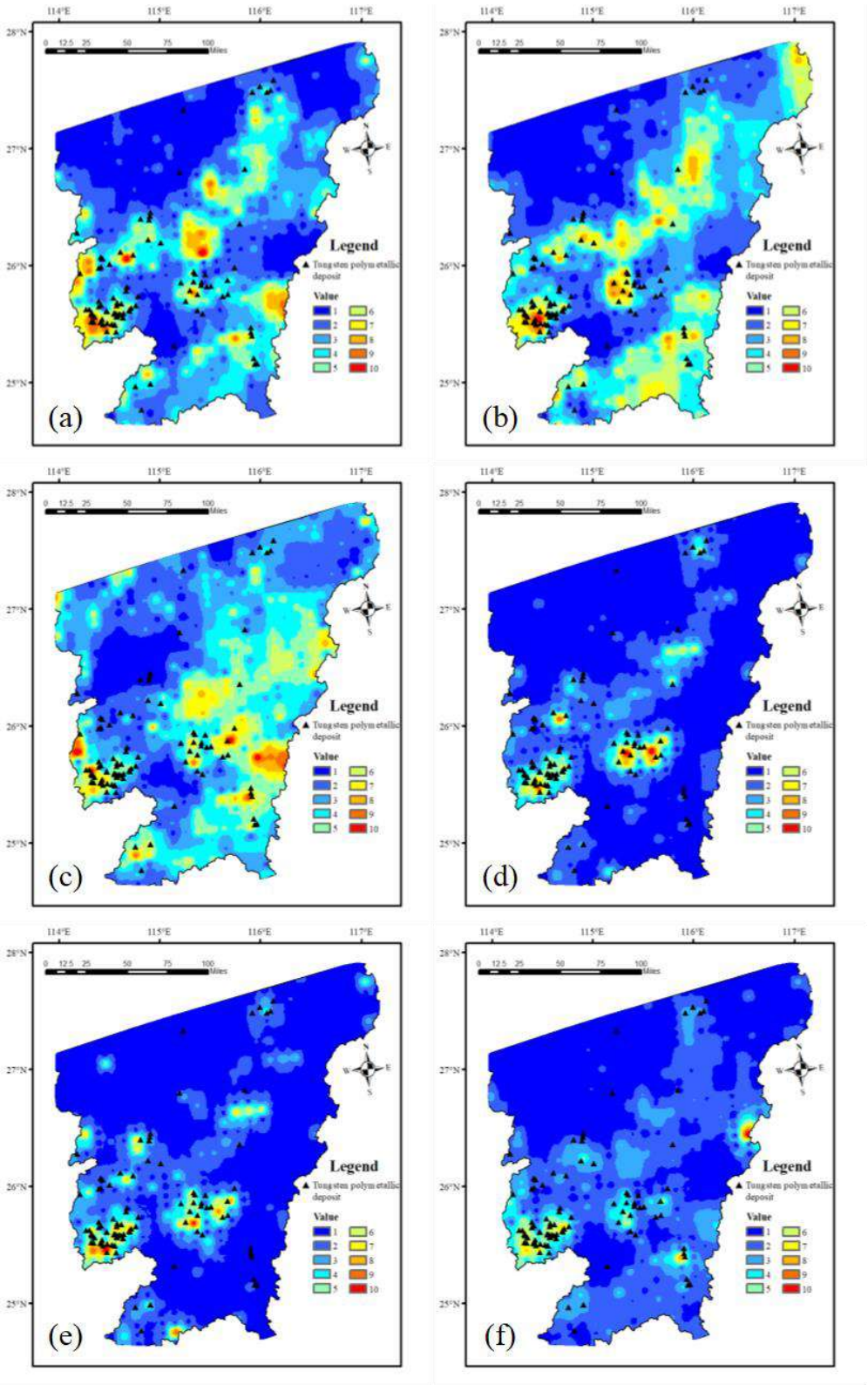
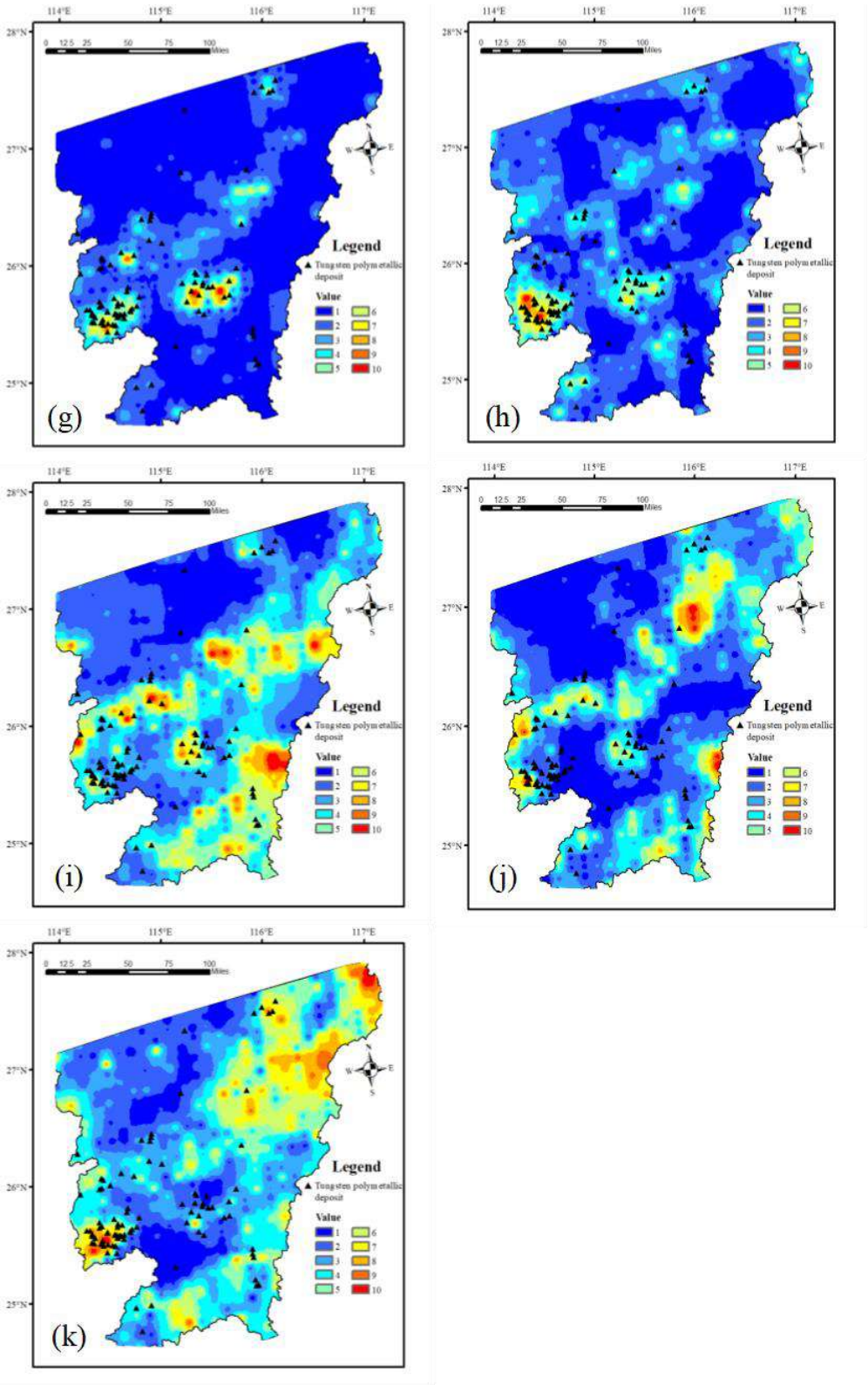
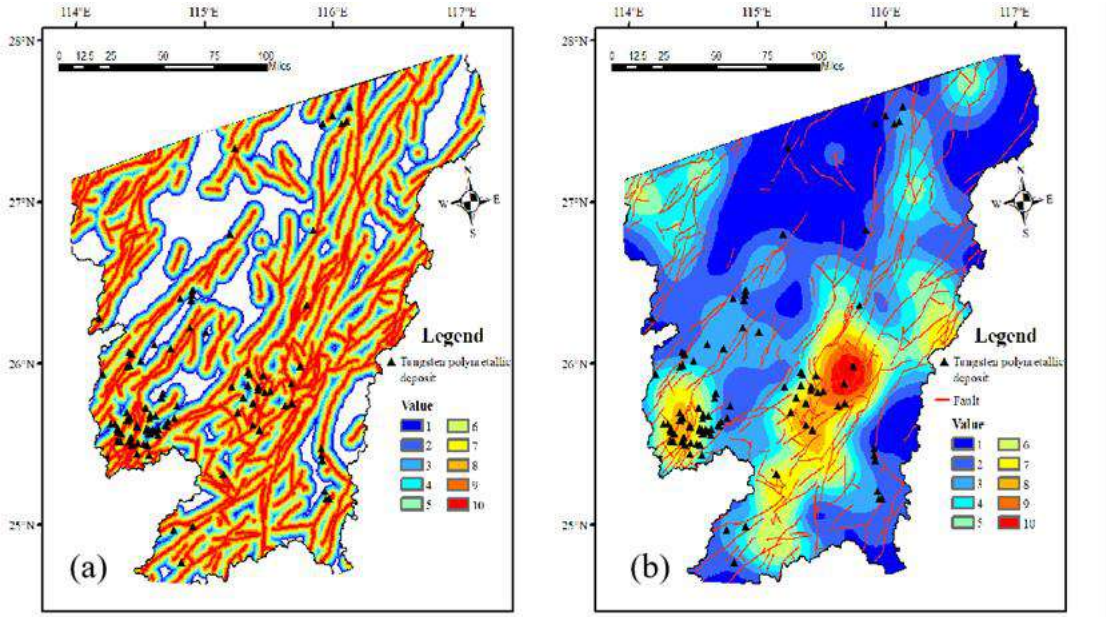


Figure 4. IDW results of 11 geochemical elements: (a) Be; (b) Pb; (c) F; (d) Li; (e) W; (f) Sn; (g) Bi; (h) Cd; (I) Y; (j) U; (k) Zn



629

630 Figure 4 continued



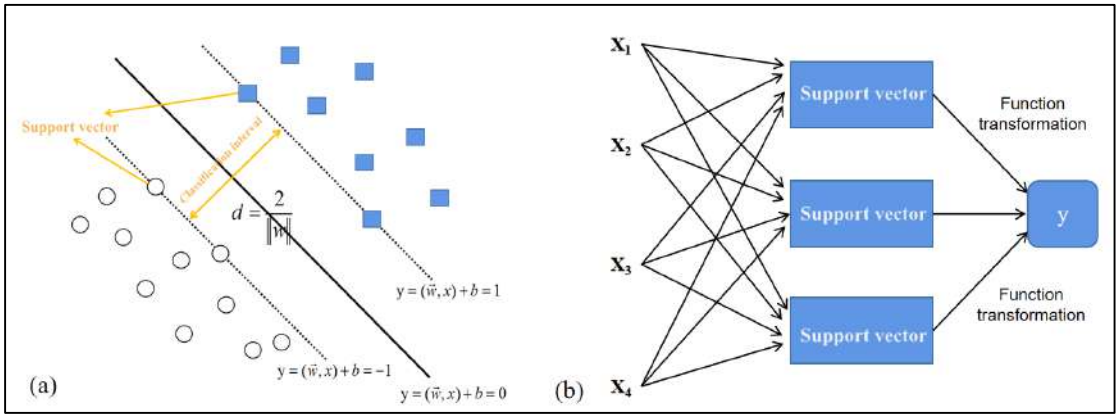
631

632 Figure 5. Fault-related elements: (a)buffering analysis of faults; (b) density of fault intersections

633

634

635



636

637 Figure 6. Schematic diagram of SVM hyperplane division principle

638

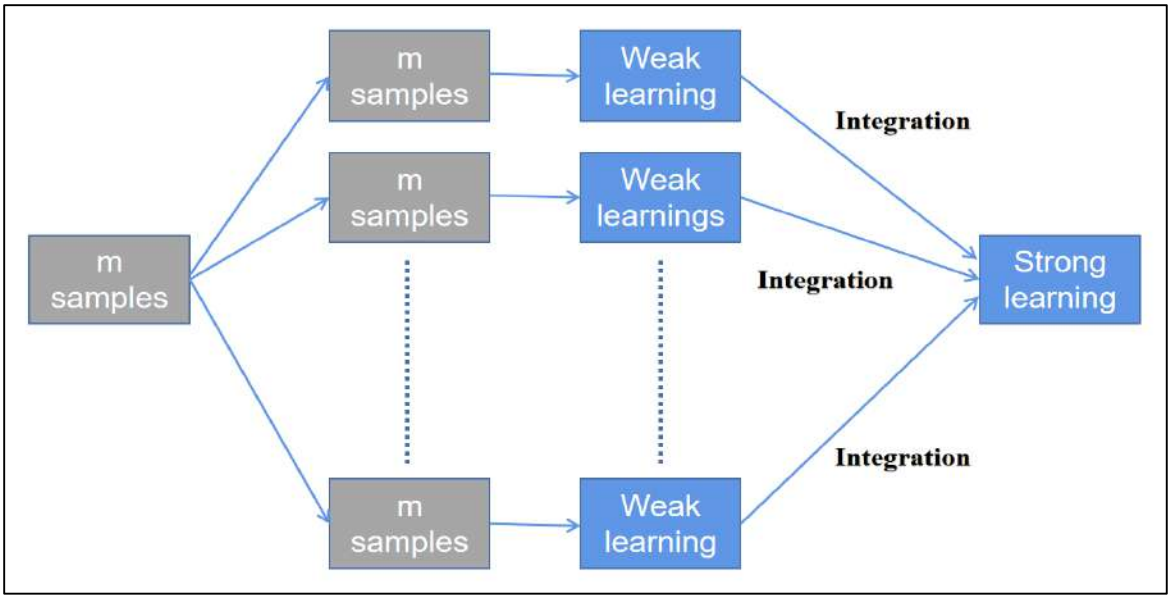


Figure 7. Schematic diagram of bagging algorithm

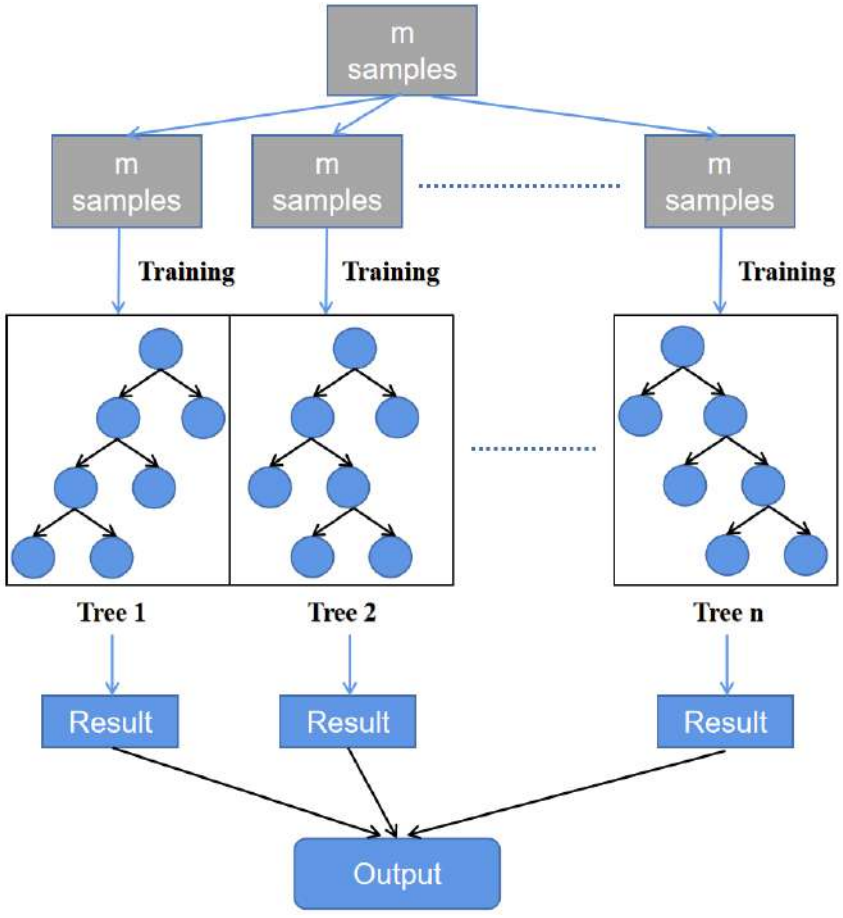


Figure 8. Schematic diagram of random forest algorithm

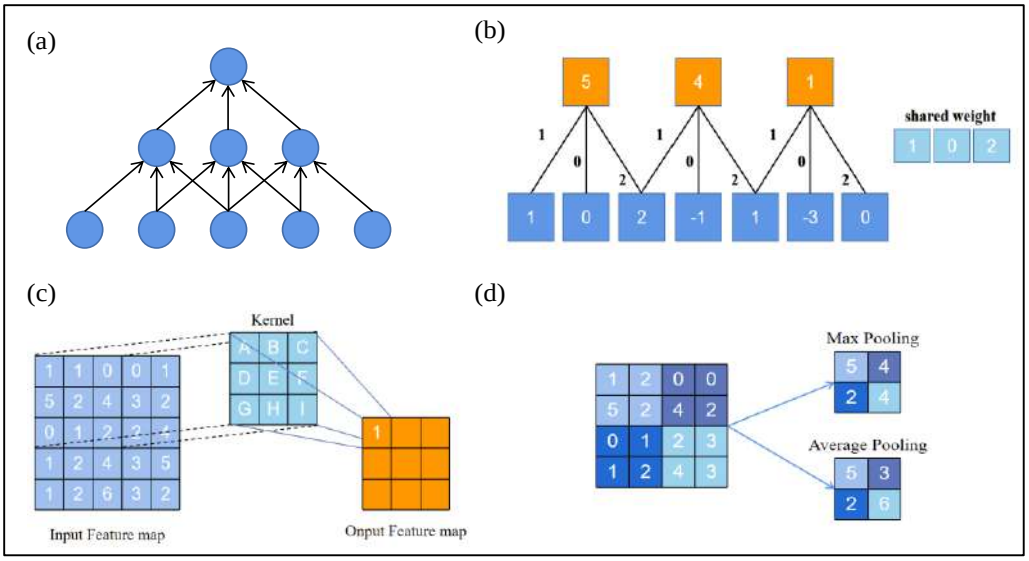


Figure 9. Local connection diagram: (a) Local connection diagram; (b) weight sharing schematic diagram; (c) Principle of convolution operation; (d) Principle of pool operation

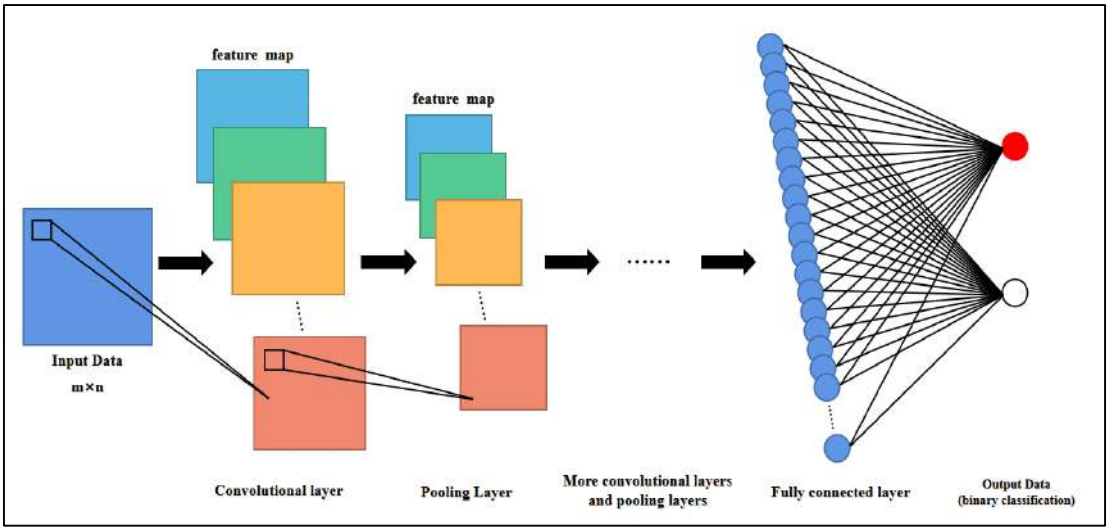


Figure 10. Convolutional neural network structure diagram

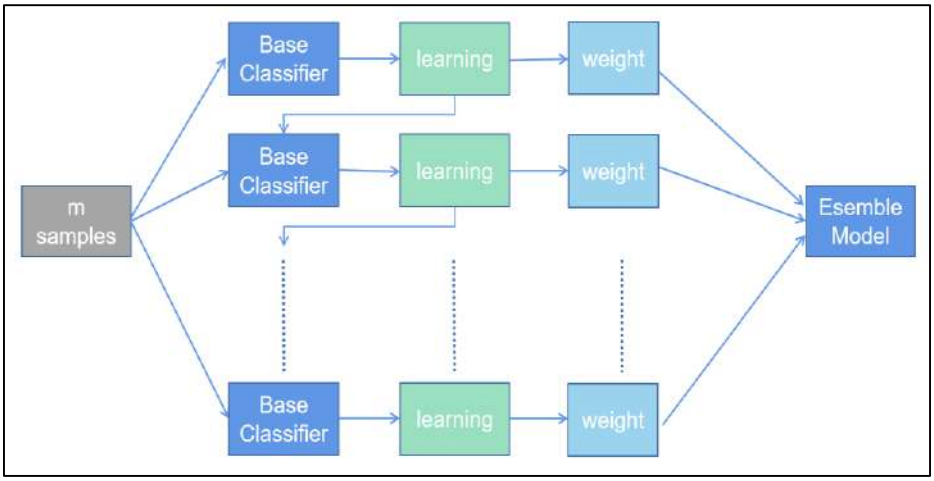


Figure 11. Boosting algorithm schematics

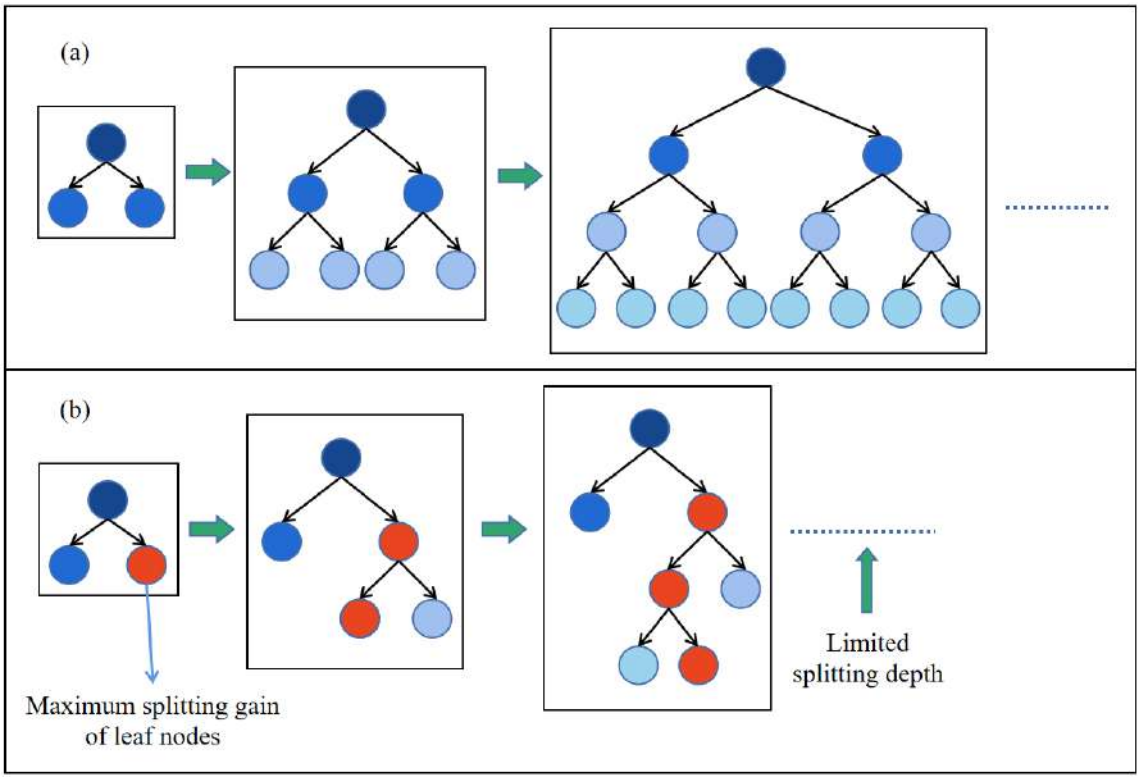
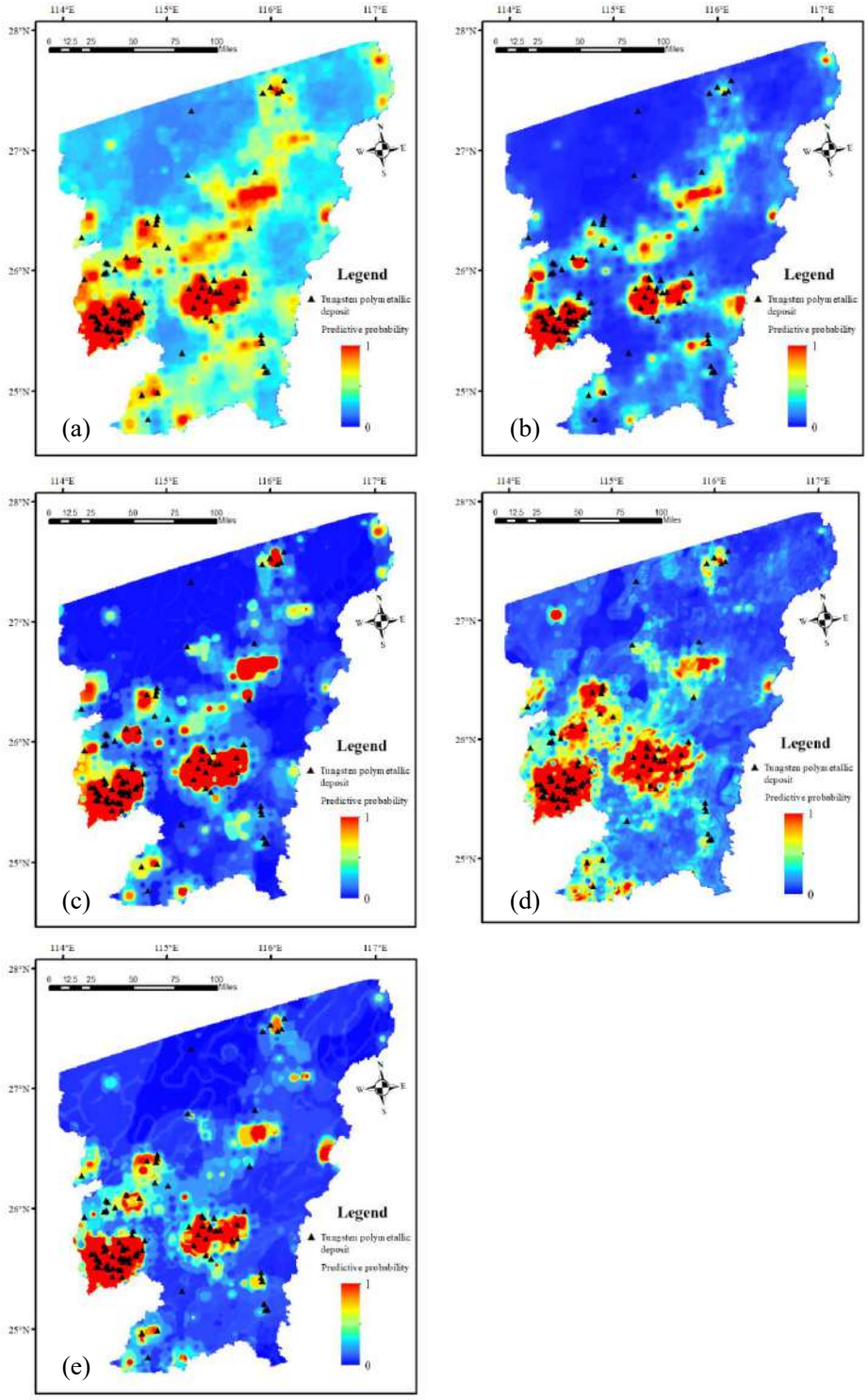


Figure 12. (a) Level-wise strategy; (b) Leaf-wise strategy with depth constraint



659

660 Figure 13. Metallogenic prediction of five models: (a) LR, (b) SVM, (c) RF, (d) CNN, (e) LGBM

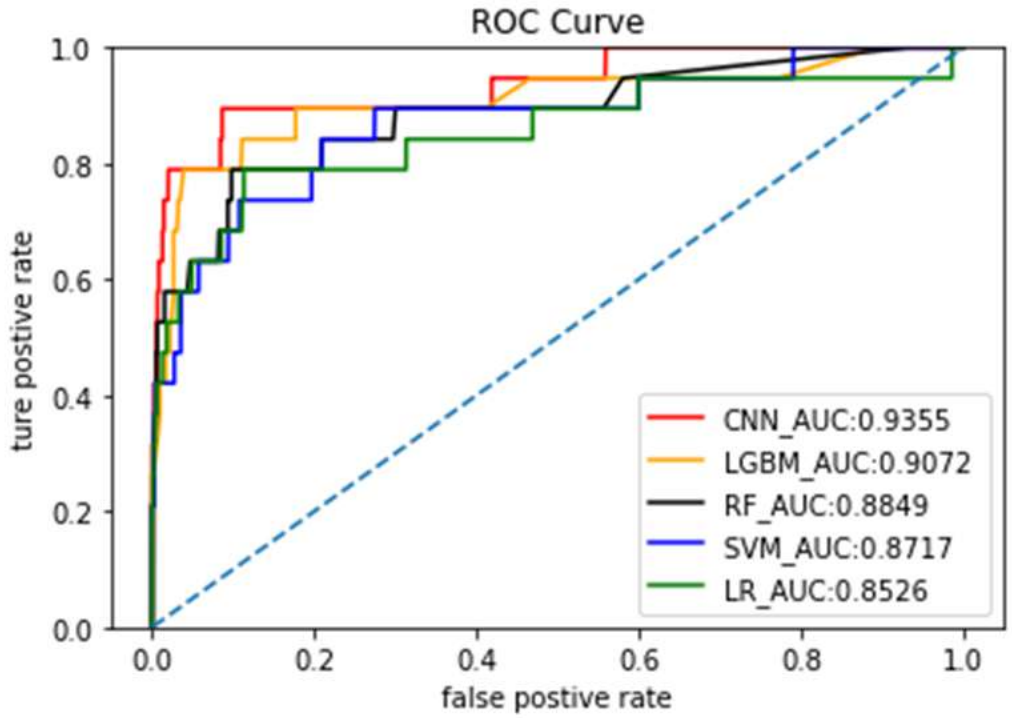


Figure 14. ROC curve of five models

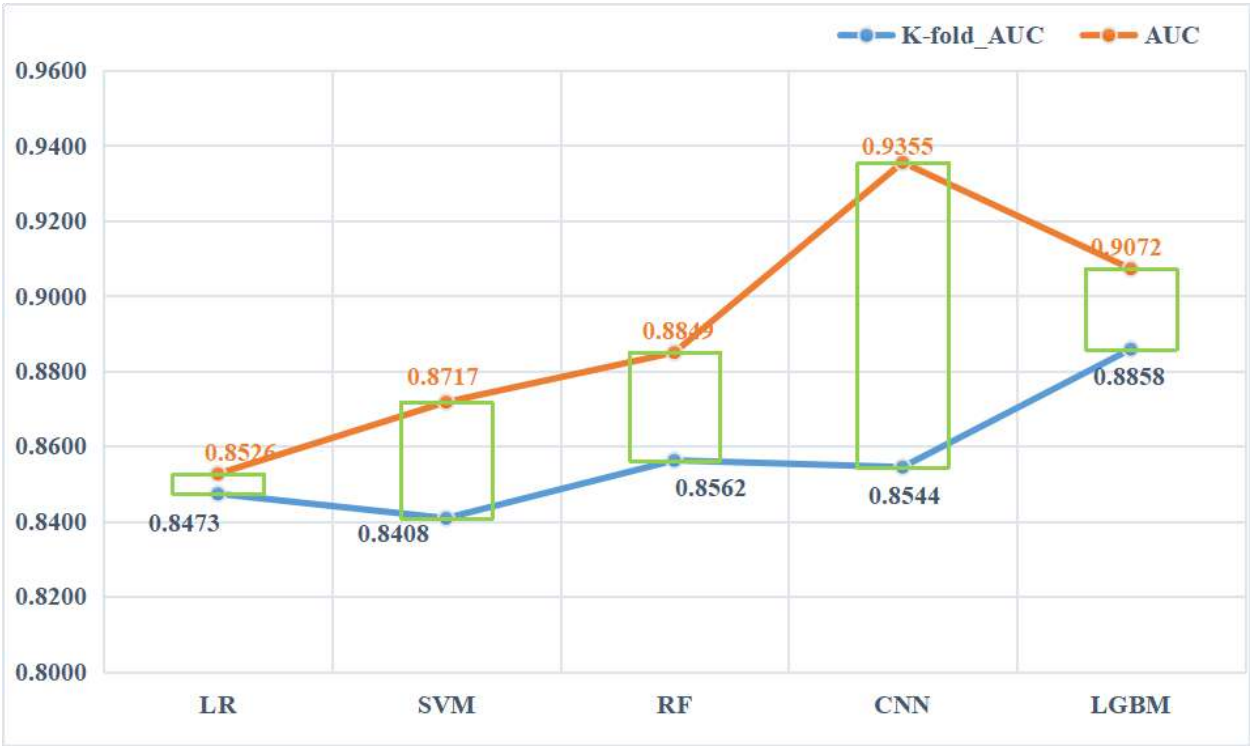
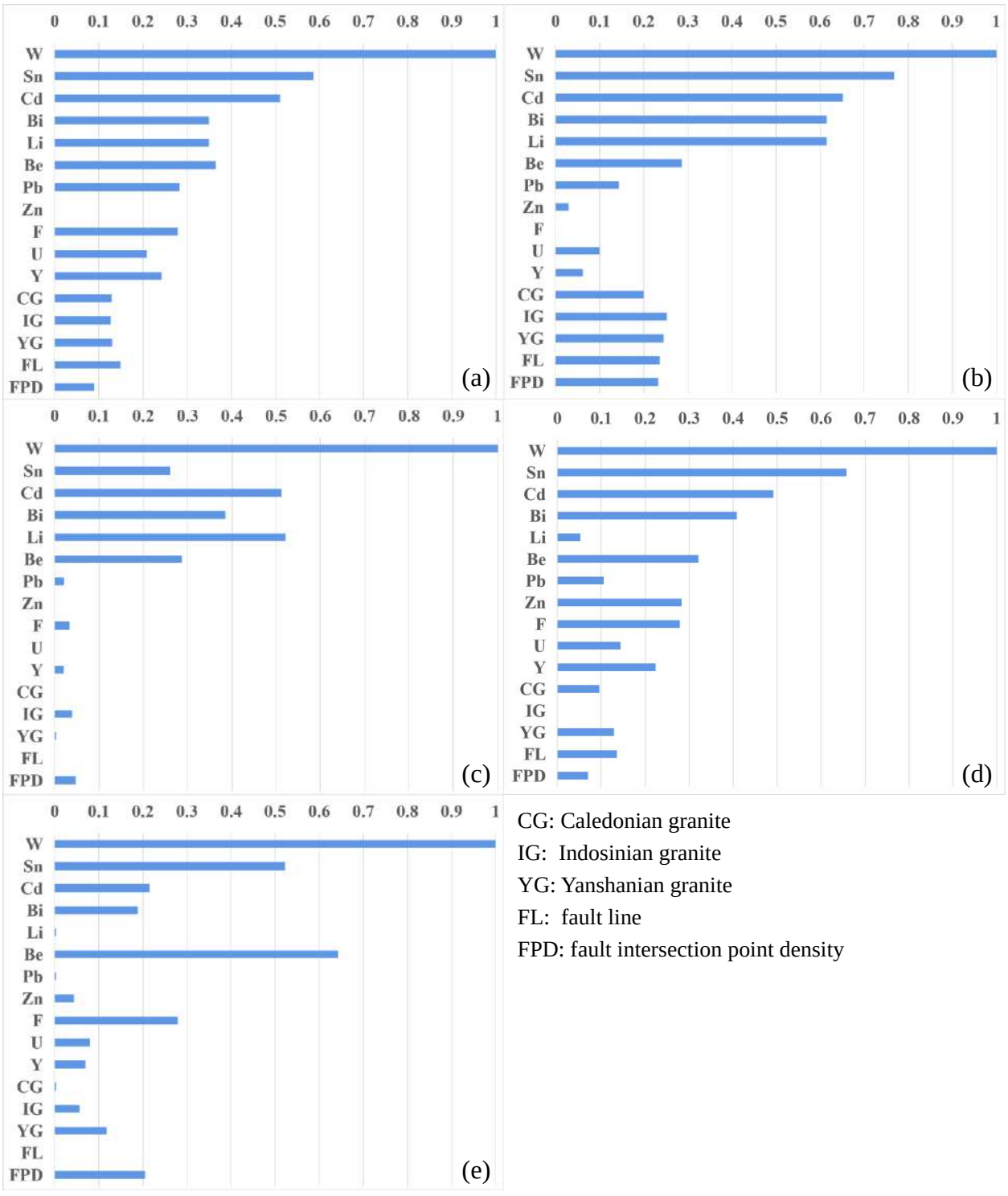


Figure 15. K-fold_AUC and original AUC value of five models



CG: Caledonian granite
IG: Indosinian granite
YG: Yanshanian granite
FL: fault line
FPD: fault intersection point density

Figure 16. Contribution rate of five model variables: (a)LR, (b)SVM, (c)RF, (d)CNN, (e)LGBM

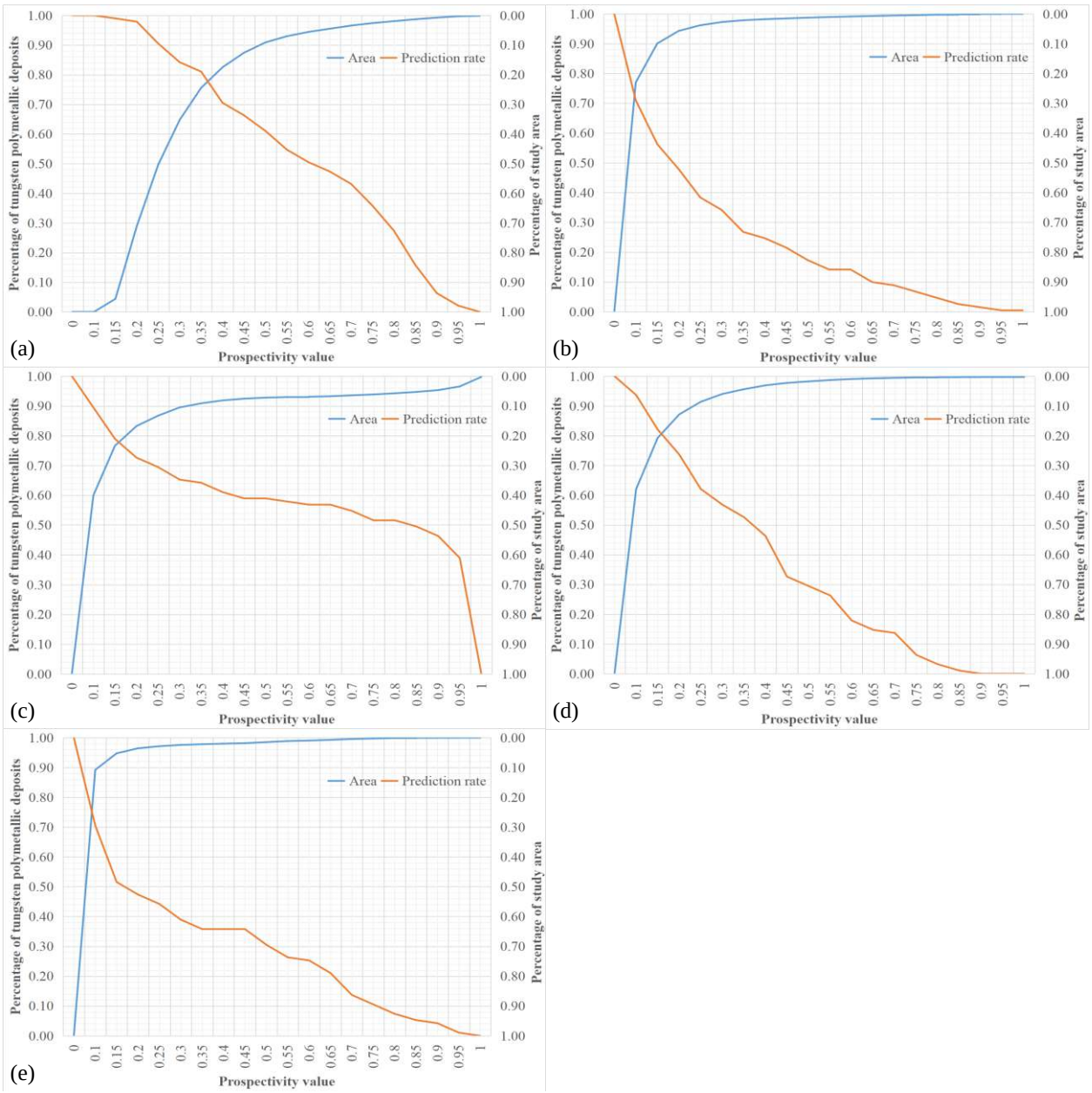
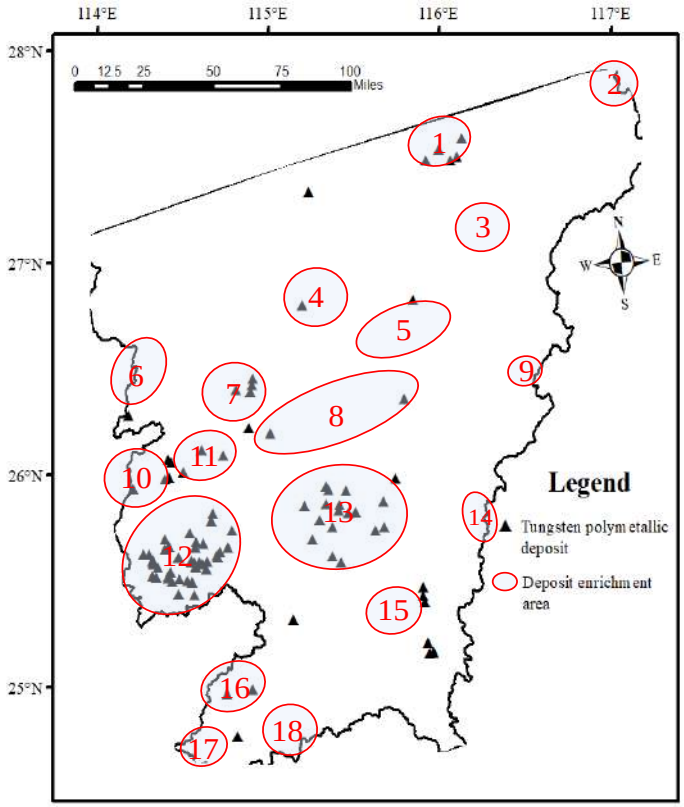


Figure 17. P-A Plot of five models: (a) LR, (b) SVM, (c) RF, (d) CNN, (e) LGBM



672

673 Figure 18. High potential areas for tungsten polymetallic exploration in the study area
674
675

676 **End**