



Landslide spatial prediction based on cascade forest and stacking ensemble learning algorithm

Sijing Chen, Yutong Pan, Chengda Lu, Yawu Wang, Min Wu & Witold Pedrycz

To cite this article: Sijing Chen, Yutong Pan, Chengda Lu, Yawu Wang, Min Wu & Witold Pedrycz (2025) Landslide spatial prediction based on cascade forest and stacking ensemble learning algorithm, International Journal of Systems Science, 56:3, 658-670, DOI: [10.1080/00207721.2024.2408551](https://doi.org/10.1080/00207721.2024.2408551)

To link to this article: <https://doi.org/10.1080/00207721.2024.2408551>



Published online: 30 Sep 2024.



[Submit your article to this journal](#)



Article views: 224



[View related articles](#)



[View Crossmark data](#)



Citing articles: 5 [View citing articles](#)



Landslide spatial prediction based on cascade forest and stacking ensemble learning algorithm

Sijing Chen^a, Yutong Pan^{b,c,d}, Chengda Lu^{b,c,d}, Yawu Wang^{b,c,d}, Min Wu^{a,b,c,d} and Witold Pedrycz^e

^aSchool of Future Technology, China University of Geosciences, Wuhan, People's Republic of China; ^bSchool of Automation, China University of Geosciences, Wuhan, People's Republic of China; ^cHubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan, People's Republic of China; ^dEngineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan, People's Republic of China; ^eDepartment of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

ABSTRACT

Landslides are a major threat to the safety of human life and property. The purpose of landslide spatial prediction is to establish the relationship between the location of landslides and each landslide evaluation factor, and to spatially identify high landslide risk areas using data mining and geographic information science. In this paper, a landslide spatial prediction model is put forward based on cascade forest (CF) and Stacking ensemble learning algorithm. Firstly, the landslide spatial prediction scheme is designed. Then, the improved CF is established by combining random forest (RF) and extreme gradient boosting (XGBoost). The Stacking ensemble learning algorithm is introduced to establish CF-Stacking model combined with the improved CF. Finally, experiments are conducted using geospatial data of the actual study area. 12 landslide disaster-inducing factors are extracted from the study area, and the CF-Stacking model is applied to the spatial prediction of landslides. The result shows that CF-Stacking outperforms comparative models in terms of the area under curve and brier score, demonstrating its effectiveness in predicting landslide spatial patterns. The CF-Stacking model is used to generate a landslide susceptibility map for Fengjie, which provides valuable guidance for geological hazard early warning.

ARTICLE HISTORY

Received 12 June 2024
Accepted 20 September 2024

KEYWORDS

Landslide spatial prediction;
cascade forest; stacking;
ensemble learning; remote
sensing

1. Introduction

Landslide spatial prediction is to predict the probability of landslide occurrence in a large area to obtain the spatial location and possibility of landslide occurrence, which solves the problem of spatial location and susceptibility of regional landslides. It exhibits great scientific significance and engineering value for disaster prevention and mitigation in high-risk landslide areas (Sk et al., 2021).

There are two kinds of landslide spatial prediction methods, including knowledge-driven methods and data-driven methods. The knowledge-driven methods use the expertise of experts and scholars in the geological field to predict landslides in space. Representative knowledge-driven methods are fuzzy set methods (Akin & Hüseyin, 2023; Haoyuan et al., 2018; Talha et al., 2019) and analytical hierarchy process (Hariklia et al., 2019; Sandeep & Amit, 2022).

However, the knowledge-driven method based on the experience of experts and scholars is increasingly unable to cope with the information processing of landslides with complicated acquisition methods and huge data sources. The empirical knowledge of experts and scholars cannot include all aspects of landslide data, which is highly subjective.

The data-driven methods determine the weight of regional landslide disaster-inducing factors according to complex multi-source data by analyzing various evaluation factors of landslides. The data-driven models mainly include information method (Li et al., 2024; Xia & Jia, 2022; Yuan et al., 2022; Zhao et al., 2020), evidence weight method (Chen et al., 2017; M. L. Hussain et al., 2021; Wang et al., 2020) and machine learning method (Madruga et al., 2022; Xiong et al., 2021; Xu et al., 2022; W. Zhang et al., 2021). Among machine learning methods, ensemble learning is successfully

CONTACT Min Wu ✉ wumin@cug.edu.cn 📧 School of Future Technology, China University of Geosciences, Wuhan 430074, People's Republic of China; 📧 School of Automation, China University of Geosciences, Wuhan 430074, People's Republic of China; 📧 Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, People's Republic of China; 📧 Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, People's Republic of China

applied to landslide spatial prediction, which completes learning tasks through several learners.

Commonly used ensemble learning methods are mainly divided into Bagging (Wang, 2017), Boosting (H. Zhang et al., 2020) and Stacking (Akhtar et al., 2020). Bagging in ensemble learning improves model stability by building multiple base learners and averaging or voting. Boosting serially trains multiple weak learners, and each learner adjusts according to the performance of the previous learner to reduce the error. Stacking uses the predictions of multiple base learners as new training data, and integrates these results through meta-learners. These methods effectively improve the model performance and generalisation ability, and provide powerful tools for solving complex problems.

The Bagging methods divide the dataset by sampling with replacement and obtain the final result training of several learners in parallel. Six decision tree classifiers, including random forest (RF), Random Tree, and Representative Tree, are used to detect abnormal network patterns (Siva et al., 2011). A dynamic classifier system is proposed by searching for the minimum-sufficient ensemble and applying bagging at the ensemble level (Lei & Mohamed, 2015). Shirvani (2020) uses RF algorithm to evaluate the performance of protected and unprotected forests in northeastern Iran in terms of landslide susceptibility. Bagging usually reduces the variance of the model, but it is less effective at reducing bias.

The Boosting methods train several learners serially through an adaptive method, each of which relies on the previous model, finally combining all the learners according to a certain strategy (Hong et al., 2023). A method integrating extreme gradient boosting (XGBoost) and Bayesian optimisation algorithms for predicting the range of overbursts is proposed (W. Zhang et al., 2020). XGBoost and RF models are used to compare and predict the correlation between pathogenic parameters and landslides, demonstrating that the XGBoost method produces superior results (M. A. Hussain et al., 2021). Wu et al. (2020) consider 11 landslide conditions and propose the ADTree-AdaBoost landslide sensitivity analysis model. There is a dependency between the base learners of the Boosting method that cannot be parallelised as easily as Bagging, which affects efficiency on large datasets.

Bagging and Boosting select learners as being of the same type of individual learners, such as RF with several decision trees. We call this learning method homogeneous ensemble, which is difficult to obtain the differences between different learners.

A deep forest based on non-differentiable modules such as decision tree, combining the characteristics of several granularity scanning modules and Cascade forest (CF), is explored (Zhou & Feng, 2017). Dai et al. (2023) propose a model, stacking feature deep forest, to determine the limit wheel profile for hunting instability. Unfortunately, the basic learners of the deep forest are all based on Bagging by parallel training several learners, without combining Boosting by serially training learners to build models, which may not adequately learn the spatial characteristics of the landslide. This motivates us to improve the deep forest.

In ensemble learning methods, besides Bagging and Boosting, there is also the Stacking ensemble learning algorithm. The Stacking generates a new prediction model by combining the advantages of several heterogeneous models. The ensemble learning model of multi-model fusion of Stacking framework is introduced into landslide spatial prediction, effectively improving the generalisation ability and prediction effect of the prediction model (Akhtar et al., 2020). Jia and Zhang (2020) propose a deterministic strategy for dependency modelling via stacked dependency exploitation. Experimental results show stacked dependency exploitation leads to superior performance.

Although the existing ensemble learning methods have achieved some results in improving model performance, there are still some limitations and challenges, especially in dealing with complex data and improving model generalisation capabilities (Marcio et al., 2021). In the current study, the spatial prediction methods of landslides cannot effectively capture complex spatial features, especially when it involves the formation and evolution of landslides. CF in the deep forest effectively improves the problem of classification models learning complex data features. By combining the prediction results of multiple classifiers, CF can improve the overall prediction performance and reduce the risk of overfitting. Then, a question arises that is it possible to combine Stacking and CF, and fully learn the spatially predicted probabilistic features of landslide output from CF? Answering this question is another

motivation of this article. The challenging task of integrating CF and Stacking methods is to consider the overall structure design of CF-Stacking, and how to give full play to the advantages of each model in the integration process.

In order to conduct the landslide spatial prediction, a landslide space prediction scheme is designed. An improved CF landslide spatial prediction model combined with Stacking is established, and the proposed model is verified through experiments. The innovation of this paper has three aspects.

- (1) An improved CF model based on RF and XGBoost is designed combining the advantages of Bagging and Boosting ensemble learning methods.
- (2) A landslide spatial prediction model combining CF and Stacking is proposed. The CF-Stacking model is established by using RF and XGBoost in CF as base learner and logistic regression as the meta learner.
- (3) We validate the effectiveness of CF-Stacking using remote sensing data, digital elevation model (DEM) and geological information of the actual study area. The CF-Stacking model is utilised to generate a landslide susceptibility map for Fengjie, a county town in Chongqing City, where geological disasters occur frequently. The experiments in this paper provide important guidance for the early warning of geological disasters.

2. Methodology

In this section, the landslide spatial prediction method is analysed. Firstly, the landslide spatial prediction scheme is designed; then, the CF structure is designed by using RF and XGBoost as the learners of each CF level; finally, the CF structure is used as the base learner and logistic regression is used as the meta-learner in combination with the Stacking ensemble learning method to establish the CF-Stacking model.

2.1. Landslide spatial prediction scheme

The spatial prediction of landslides is to extract landslide disaster-inducing factors from geospatial data, establish the spatial prediction model of landslides, and obtain the spatial prediction probability of landslides and landslide susceptibility mapping. Considering the redundancy of landslide susceptibility data and

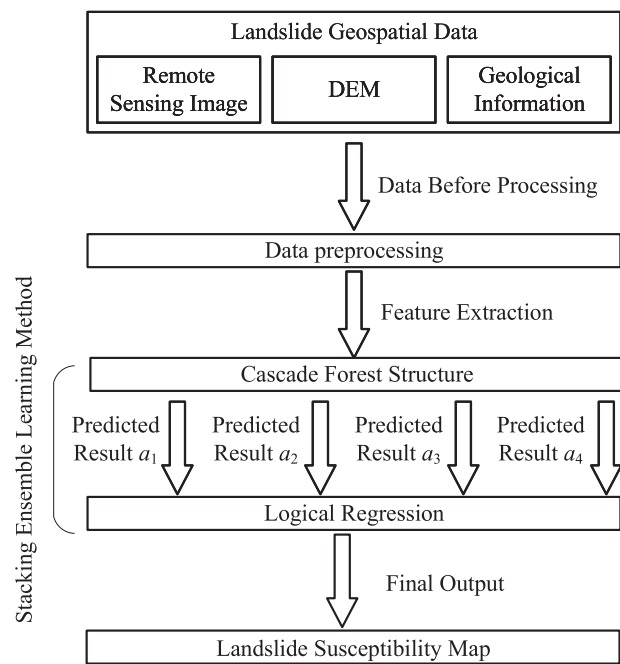


Figure 1. Spatial prediction scheme for landslide prediction.

its multi-source nature, this section designs a spatial prediction scheme for landslide prediction based on CF-Stacking model as shown in Figure 1.

Firstly, the multi-source landslide spatial data such as remote sensing images, DEM and geological data are obtained. The multi-source landslide spatial data are preprocessed. Pearson correlation is used to remove the redundant factors and extract to the downscaled features. Then, RF and XGBoost are used as the sub-modules of CF to establish the CF structure, and produce the probabilistic prediction results of the four learners a_1 , a_2 , a_3 , a_4 . Finally, the Stacking is introduced ensemble learning method, the CF structure is used as the base learner and logistic regression is used as the meta-learner to build the CF-Stacking model, which is used to make landslide spatial prediction and obtain landslide susceptibility map. This scheme achieves the purpose of evaluating landslide susceptibility by acquiring multi-source data features through heterogeneous learners by CF structure and Stacking.

2.2. Improved cascade forest

According to the research status of landslide spatial prediction, we can see that the ensemble learning method is widely used in landslide spatial prediction. Some scholars have used deep forests to carry

out research on the spatial prediction of landslides. The deep forest consists of two parts: CF and multi-granularity scanning. CF is an important part of the deep forest model, which has the structure of an ensemble learning model composed of level and level, and each level contains several ensemble learning classifiers. Multi-granularity scanning uses multiple sliding windows of different sizes to sliding values on the original data features to generate multi-scale input features.

In the spatial prediction of landslides, the data features extracted from remote sensing images, DEM and geological data are preprocessed to a certain knowledge-driven, and each evaluation factor has an important impact on the development of landslides. Therefore, we removed the multi-granularity feature scanning of the deep forest, retained only the CF structure, and input all the landslide features into the model for feature-level data fusion.

However, the base learner in the traditional deep forest CF structure consists of two completely random forest (CRF) and two RF per level. The main difference between the two types of forests is the candidate feature space, which essentially belongs to the Bagging ensemble learning method. The learner selected in this method is the individual learner of the same type. We call this learning method homogeneous ensemble, which is difficult to obtain the differences of different learners. Homogeneous ensemble cannot fully learn the characteristics of landslide multi-source data. Therefore, an improved CF model based on RF and XGBoost is established.

RF is one of the most commonly considered algorithms in Bagging ensemble method, which is widely used in data classification and regression problems (Breiman, 2001). RF consists of multiple decision trees. In landslide spatial prediction, each decision tree produces a landslide spatial prediction result, and the final landslide spatial prediction probability is obtained through voting or averaging strategy.

RF uses classification and regression trees (CART) as the base learner. CART fits data by recursively partitioning the dataset into subsets based on different feature thresholds to construct a decision tree. By selecting the best features and thresholds to minimise impurity in each subset, CART aims to fit the data such that each leaf node contains predominantly samples of the same class, enabling accurate predictions for classification or regression tasks.

CART uses the Gini index to measure the impurity of the spatial sample features of the segmented landslide, so as to select the best landslide evaluation factor for segmentation. The Gini index is computed as follows:

$$\text{Gini}(D) = 1 - \sum_{k=1}^y p_k^2 \quad (1)$$

where $p_k^2 = \frac{C^k}{D}$. D is the size of landslide spatial sample. C^k is the number of class k landslide spatial sample.

The greater the Gini index, the higher the uncertainty of the landslide spatial dataset. The split *Gini* index is computed as follows:

$$\text{Gini}(D, A) = \frac{D_L}{D} \text{Gini}(D_L) + \frac{D_R}{D} \text{Gini}(D_R) \quad (2)$$

where the dataset of parent node is D . CART splits into two child nodes after selecting features, and the dataset are D_L and D_R .

The essence of the decision tree fitting process in the landslide spatial prediction model is that the Gini coefficient of the sample decreases, so the landslide spatial feature splitting with the smallest Gini coefficient should be selected. When the dataset's Gini index threshold is reached, the split is stopped. Finally, according to the feature splitting results, the samples are classified or regressed and the spatial prediction results of landslides are obtained.

XGBoost is a Boosting algorithm to improve the Gradient Boosting Decision Tree (GBDT) (Chen & Guestrin, 2016). In landslide spatial prediction, the core idea of the XGBoost algorithm is to introduce the residuals trained by the weak landslide spatial prediction model into the new weak model, improve the prediction accuracy of the weak model step by step through the gradient direction of the loss function, and finally generate a strong spatial prediction model to obtain a more landslide spatial prediction probability.

The XGBoost prediction for the i th landslide spatial sample can be expressed as follows:

$$\hat{y}_i^{(t)} = \sum_{j=1}^t f_j(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (3)$$

where t is the number of landslide spatial prediction model iterations. $f_j(x_i)$ is the weight of each leaf node decision tree.

The XGBoost objective function is composed of the loss function of the model and the regularisation term

that suppresses the complexity of the model, which is defined as follows:

$$\begin{aligned} \text{OBJ}^{(t)} &= \sum_{i=1}^M l(y_i, \hat{y}_i^{(t)}) + \Omega(f_t) \\ &= \sum_{i=1}^M l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \end{aligned} \quad (4)$$

where M is the landslide spatial sample number. $l(y_i, \hat{y}_i^{(t)})$ is the loss function of the model. $\Omega(f_t)$ is the regularisation term.

The objective function is second-order expanded by Taylor formula, which is computed approximately as follows:

$$\begin{aligned} \text{OBJ}^{(t)} &= \sum_{i=1}^M [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] \\ &\quad + \Omega(f_t) \end{aligned} \quad (5)$$

where $\hat{y}_i^{(t-1)}$ is the predicted value of the i th landslide spatial sample in the first $t-1$ decision tree. g_i is the first derivative of the loss function. h_i is the second derivative of the loss function. The complexity of a single decision tree is computed as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^t f_j(x_i)^2 \quad (6)$$

where T is the number of leaf nodes. γ is the penalty for the number of leaf nodes. λ is the L2 regularisation penalty. $l(y_i, \hat{y}_i^{(t-1)})$ is a fixed value, and do not affect the optimisation process. The objective function $\text{OBJ}^{(t)}$ after t iterations is obtained as follows:

$$\begin{aligned} \text{OBJ}^{(t)} &= \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] \\ &\quad + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^t f_j(x_i)^2 \end{aligned} \quad (7)$$

CF enhances the expressiveness and prediction performance of the landslide spatial model by organising and connecting forests at different levels. Each forest is referred to as a level in a cascading forest, and each level contains multiple RF models. The whole cascade forest can be composed of multiple levels. The output of each level is the spatial prediction result of the landslide. The input of the secondary level is the

fusion of the prediction output of the previous level and the original landslide features. Cross-validation is used to automatically determine the optimal number of CF levels. After each level is trained, the performance of the current model is evaluated through cross-validation. If the model's performance no longer improves significantly, or if it starts to degrade, stop adding new levels.

The structure of the improved CF model is shown in Figure 2, the model combines the parallel learning performance of Bagging and the serial learning performance of Boosting. The improved CF uses RF and XGBoost as the base learners, and the final prediction result can more fully learn the spatial characteristics of the landslide.

2.3. CF-Stacking landslide spatial prediction model

Stacking is one of the ensemble learning methods. It combines several base learners and uses a meta-learner to retrain the results of base learner training to obtain better results than a single base learner.

The CF-Stacking is a landslide spatial prediction model combining the CF and the Stacking ensemble learning method. The combined CF and Stacking ensemble learning algorithm can learn the prediction results of the base learner output by CF as the second-level features again to obtain better prediction results.

The size of training sets for landslide spatial prediction samples is set to k_1 . The size of validation sets is set to k_2 . The size of test sets is set to k_3 . The training, validation, and test sets are divided into 8:1:1 ratios.

The feature set of the landslide evaluation factor input can be expressed as:

$$A_1 = \{a_{11}, a_{12}, \dots, a_{1t}\} \quad (8)$$

The training set of landslide spatial prediction samples can be expressed as:

$$X_m = \{(x_{m1}, y_{m1}), (x_{m2}, y_{m2}), \dots, (x_{mk_1}, y_{mk_1})\} \quad (9)$$

The verification set of landslide spatial prediction samples can be expressed as

$$X_n = \{(x_{n1}, y_{n1}), (x_{n2}, y_{n2}), \dots, (x_{nk_2}, y_{nk_2})\} \quad (10)$$

The test set of landslide spatial prediction samples can be expressed as

$$X_l = \{(x_{l1}, y_{l1}), (x_{l2}, y_{l2}), \dots, (x_{lk_3}, y_{lk_3})\} \quad (11)$$

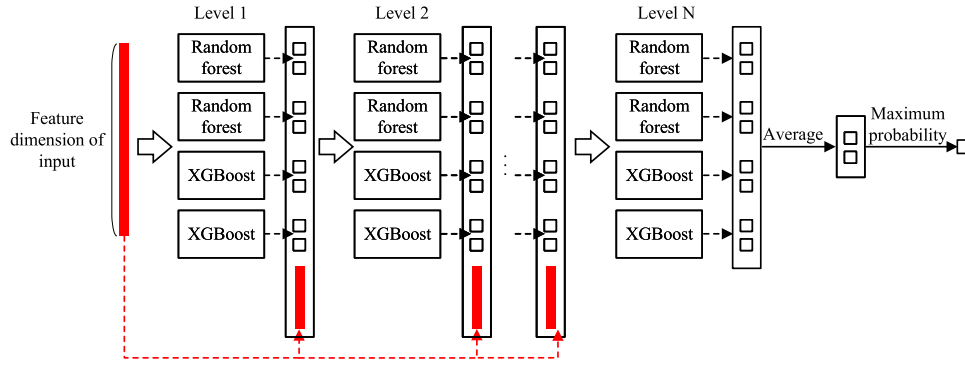


Figure 2. Improved CF model.

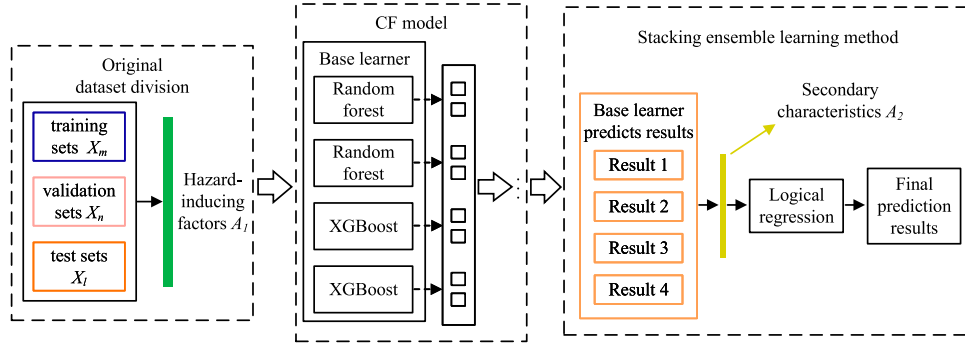


Figure 3. CF-Stacking landslide spatial prediction model.

In the training process of landslide spatial prediction model, the training and validation sets are input into each base learner of the CF for k fold cross validation. Each base learner fits step by step from training set, and the validation set is entered into the fitted base learner to extract secondary features, which is

$$A_2 = \{a_{21}, a_{22}, a_{23}, a_{24}\} \quad (12)$$

In the Stacking, a simple structure model is generally chosen as the meta-learner in order to avoid overfitting. In this paper, logistic regression model is used as a meta-learner. The secondary feature A_2 is input into the meta-learner for fitting to obtain the final prediction probability. The formula for calculating the logistic regression meta-learner is as follows:

$$z = w_1 \cdot a_{21} + w_2 \cdot a_{22} + w_3 \cdot a_{23} + w_4 \cdot a_{24} + b \quad (13)$$

where z is the result of a linear combination of the weighted summations of A_2 . b is a constant term. w_1 , w_2 , w_3 and w_4 are the weights of the prediction results of the base learner.

The results of the linear function are transformed by the *Sigmoid* function to obtain the final spatial

prediction result of the landslide:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (14)$$

The structure of the CF-Stacking model is shown in Figure 3, and the steps for calculating the model are as follows:

Step 1: The original data for landslide spatial prediction is divided into training data and a test set X_l .

Step 2: The training data is further divided into k folds.

Step 2.1: The current i th fold is used as the validation set X_n , while the remaining $k-1$ folds are used as the training set X_m .

Step 2.2: The base learner CF is trained using X_m .

Step 2.3: X_n is used to predict landslide spatial probabilities using the trained CF, obtaining the landslide spatial prediction probabilities for the current fold, referred to as secondary features A_2 .

Step 3: The secondary features A_2 from each fold in the k -fold cross-validation are concatenated to obtain the complete training data for the meta-learner for landslide spatial prediction.

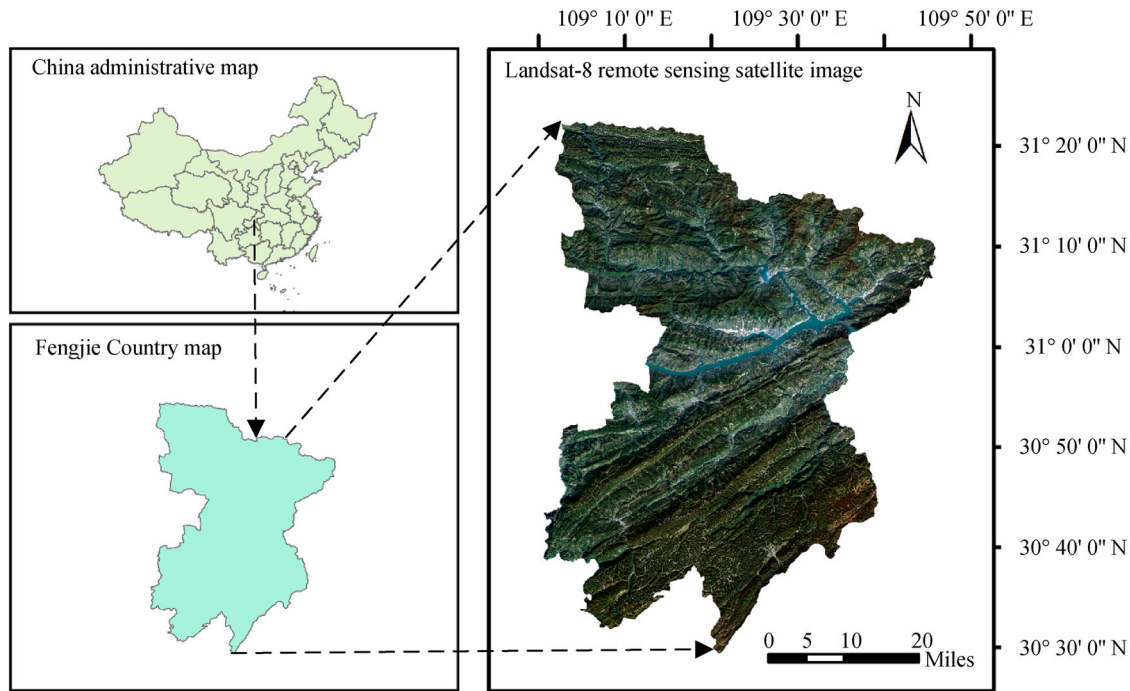


Figure 4. Landsat-8 remote sensing satellite image of Fengjie County.

Step 4: The meta-learner logistic regression model is trained using the complete secondary features A_2 .

Step 5: The trained CF-Stacking model is used to test the test set X_I , obtaining the distribution range of landslide susceptibility.

3. Experiments

In this section, we conduct a comparative experiment to verify the effectiveness of CF-Stacking for a specific study area, Fengjie County. Firstly, the geospatial characteristics of the study area are introduced and the data are preprocessed; then, the experimental environment and model parameters for landslide spatial prediction are introduced, and the landslide spatial prediction data are predicted using CF-Stacking, RF-Stacking, X-Stacking, RF and XGBoost; finally, the experimental results are analysed to compare CF-Stacking with other models and landslide sensitivity mapping is performed.

3.1. Landslide spatial prediction procedure

Fengjie is located in Chongqing City, China, as shown in Figure 4. It includes 4 streets, 18 towns and 11 townships, including Yong'an Street, Baidi Town and Yanwan Township, with a total area of 4098 km². Fengjie is located in the eastern edge of the Sichuan Basin,

the front edge of the Daba Mountain and the western mountain of Hubei Province. Its topography and geomorphology are controlled by the regional geological structure and lithology. The trend of the major mountains in the study area is consistent with the direction of the tectonic line. The mountains to the north of Qiyao Mountain are nearly east-west, and the mountains to the south are northeast-southwest. The highest standard elevation of Taishan Temple is 810.8 m and the lowest standard elevation of Yangtze River is 80.0 m. The relative elevation difference is 730.8 m.

The data sources in this paper include three parts:

- (1) Remote sensing satellite images from the Landsat 8 satellite digital product, published by geospatial data cloud, 7 February 2020 with a spatial resolution of 30 m. Landsat8 has 11 spectral bands ranging from visible to thermal infrared wavelengths. These bands capture different aspects of the Earth's surface, such as vegetation health, land cover types, water quality and more. Landsat8 images are provided in geographic tagged image file format (GeoTIFF).
- (2) DEM derived from advanced spaceborne thermal emission and reflection radiometer global digital elevation model, published by geospatial data cloud with a spatial resolution of 30 m, covering

Fengjie and surrounding areas. The DEM covers the terrain information of Fengjie and its surrounding areas, suitable for applications such as terrain analysis, hydrological modelling and resource management. The DEM is provided in GeoTIFF.

- (3) The geological map from the geological science data publishing system. Geological maps include topographic information such as contour lines, which help in understanding the elevation and relief of the area. Geological maps are provided in GIS formats like shapefiles, GeoTIFFs and geodatabases, enabling integration with GIS software for spatial analysis.

The original remote sensing images of the geospatial data cloud is a regional sub-images, which needs to form a remote sensing images covering the study area completely through geometric mosaic. After geometric mosaic, atmospheric correction and radiometric calibration of remote sensing images are performed respectively.

Atmospheric correction is a method to eliminate the radiation error caused by the atmosphere in geomatics to obtain the true reflectivity of the ground object. The total radiance of the ground object measured by the satellite sensor is not a reflection of the true reflectivity of the surface, which includes the radiation error caused by atmospheric absorption, especially scattering. During the extraction of disaster-inducing factors, factors such as the nomalied difference vegetation index (NDVI) need to be extracted from remote sensing images closely related to the real reflectance of ground objects.

Radiometric calibration is to convert the grey value of the images into absolute radiance while calculating the spectral reflectance or spectral radiance. In this paper, the atmospheric correction and radiometric calibration are processed by the ArcGIS software of the American Institute of Environmental Systems.

The disaster-inducing factors of landslide are divided into control factors and influencing factors. The control factors refer to the factors that case the landslide by the characteristics of the hidden points, and the influencing factors refer to the factors that case the landslide by the external environment. In the control factors, this paper extracts 11 disaster-inducing factors such as geological structure, engineering rock group and slope. In the influencing

factors, 6 disaster-inducing factors such as basin area and topographic water index are extracted.

Geological structure, engineering rock group and other factors are extracted directly by ArcGIS software tools. NDVI and Normalized Difference Water Index (NDWI) are calculated by the environment for visualising images software for different bands of remote sensing images, representing the coverage of surface vegetation and the distribution of water bodies respectively. The corresponding calculations are completed as follows:

$$a_{NDVI} = \frac{a_{NIR} - a_{Red}}{a_{NIR} + a_{Red}} \quad (15)$$

$$a_{NDWI} = \frac{a_{Green} - a_{NIR}}{a_{Green} + a_{NIR}} \quad (16)$$

where a_{NIR} is the near-infrared band of remote sensing images. a_{Red} is the red band and a_{Green} is the green band.

Missing data is generated during the process of cutting disaster-inducing factors level. Imputation methods of missing data are interpolation, before and after filling method, etc. In the factors level, the data of missing values are often close to the surrounding pixel values, so we use the mean value method to fill the missing values. Due to the different scales between different factors, the data need to be normalised linearly, as shown below.

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (17)$$

where x is the data before normalisation. y is the data after normalisation. x_{min} is the minimum value of the original data. x_{max} is the maximum value of the original data.

After obtaining the linear normalised data, the strong correlation factors are eliminated in disaster-inducing factors by Pearson correlation analysis to reduce the calculation of spatial prediction in order to avoid data redundancy caused by strong correlation between data.

Pearson correlation coefficient is widely used to measure the correlation between two variables, $\bar{x}$ and $\bar{y}$. The calculations are completed as follows:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (18)$$

Table 1. Table of stratigraphic lithology characteristics in Fengjie County.

Factor category	Disaster-inducing factors
Controlling factor	Geologic structure
	Engineering rock group
	DEM
	Slope gradient
	Aspect
	Terrain surface convexity
	Terrain surface texture
	Terrain convergence
Influencing factor	Valley depth
	River discharge
	Distance to drainages
	NDVI

where $\bar{x}$ and $\bar{y}$ are the mean values of the corresponding variables.

After correlation analysis, the correlation coefficient matrix of the disaster-inducing factors is obtained, and the factor with the correlation coefficient less than 0.5 is selected as the disaster-inducing factors of the landslide. Finally, 12 factors such as engineering rock group and geological structure are selected as the input of the model, as shown in Table 1. The correlation coefficient matrix of disaster-inducing factors obtained is shown in Table 2.

The 12 disaster-inducing factor features of the sample are used as the first-level feature A_1 input of the CF-Stacking model, and the RF and XGBoost of the CF structure are used as the base learners to regress the probability of landslide occurrence, and the preliminary prediction results are obtained. The prediction result of the base learner is used as the secondary feature A_2 , and the logistic regression model is used as the meta learner of the Stacking ensemble learning method. The secondary feature A_2 is regressed again to obtain the final prediction result.

Table 3. Table of stratigraphic lithology characteristics in Fengjie County.

Model name	Parameter 1	Parameter 2	Set value
Proposed	RF	Number of submodels	50
		Maximum depth	6
RF-Stacking	XGBoost	Number of submodels	50
		Maximum depth	6
	RF1	Number of submodels	50
		Maximum depth	6
X-Stacking	RF2	Number of submodels	100
		Maximum depth	10
	XGBoost1	Number of submodels	50
		Maximum depth	6
CF	XGBoost2	Number of submodels	100
		Maximum depth	10
	RF	Number of submodels	50
		Maximum depth	6
RF	CRF	Number of submodels	100
		Maximum depth	10
XGBoost	Number of submodels	Maximum depth	50
		Maximum depth	6

3.2. Experimental details

The experiment is programmed in Python using the PyCharm integrated development environment. The parameters set by each model are shown in Table 3.

In order to verify the effectiveness of the improved CF, the comparison experiments used a combination of RF and Stacking algorithms called RF-Stacking, including two kinds of RF with different parameters. The comparison experiments also used a combination of XGBoost and Stacking algorithms, called X-Stacking, including two kinds of XGBoost with different parameters. In addition to this, CF, RF and XGBoost are used as comparison models for the comparison experiments. In the improved CF, we use random forest and XGboost as the base model, so that the spatial features of the landslide can be fused through

Table 2. Correlation coefficient matrix of disaster-inducing factors.

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
E1	1.000	−0.229	−0.463	−0.063	−0.051	−0.064	−0.494	−0.053	0.078	0.001	0.005	0.145
E2		1.000	0.411	0.005	0.002	0.038	0.344	−0.027	−0.046	0.005	−0.035	0.021
E3			1.000	0.011	0.014	0.271	0.417	0.040	−0.169	−0.014	0.288	0.221
E4				1.000	0.012	0.184	0.173	0.003	0.082	0.081	−0.055	0.229
E5					1.000	−0.009	−0.010	−0.457	0.004	0.142	0.007	−0.027
E6						1.000	0.234	0.021	0.012	0.034	0.404	0.169
E7							1.000	0.044	−0.011	0.015	0.018	0.194
E8								1.000	−0.059	−0.136	0.004	0.006
E9									1.000	−0.031	−0.099	−0.011
E10										1.000	0.011	0.024
E11											1.000	0.090
E12												1.000

E1 = Structure, E2 = Rock, E3 = DEM, E4 = Slope, E5 = Aspect, E6 = Convexity, E7 = Texture, E8 = Convergence, E9 = Valley, E10 = River, E11 = Drainages, E12 = NDVI

the differences of Bagging and Boosting. In order to simulate these differences of Bagging and Boosting, we set different parameters in RF-Stacking, X-Stacking and the traditional CF base model. Under the set parameters, each model's performance is at its optimal state, so as to ensure the fairness of the experiment.

In the experiment, the pixel grid of the hazard factor level is used as a sample, and 10,000 sample points are extracted, including 5000 landslide sample points and 5000 non-landslide sample points. The training set and verification set accounted for 8000, and the test set accounted for 2000. The training set and the verification set is subjected to 5-fold cross-validation, that is the training set and the verification set are divided into 5 parts on average, and 4-fold samples are randomly sampled as the training set and 1-fold samples are randomly sampled as the verification set.

3.3. Result analysis

To verify the validity of the CF-Stacking model, the testing set is used to verify the model and the validity of the model is checked by the receiver operating characteristic(ROC) curve and the brier score.

The ROC curve is an effective method to evaluate the classification or regression effect of prediction models (Akobeng, 2007). The Area Under the Curve (AUC) takes into account the ability of the classification area to classify positive and negative cases to evaluate the effectiveness of the prediction model. The brier score is an evaluation metric used to measure the accuracy of probability prediction models, commonly employed in assessing the probability predictions of binary or multiclass models. Unlike AUC, the brier score directly measures the discrepancy between probability predictions and the true labels. When comparing the performance of different probability prediction models, a lower brier score indicates more accurate and reliable predictions by the model. In this section, both AUC and brier score are utilised as evaluation metrics to assess the model's prediction results.

The ROC curve of the prediction results of the CF-Stacking model is calculated as shown in Figure 5. The CF-Stacking AUC value reached 0.8562 and the brier score reached 0.1596. According to the results of comparative experiments, the AUC values of six models such as RF-Stacking and X-Stacking are calculated, as shown in Table 4.

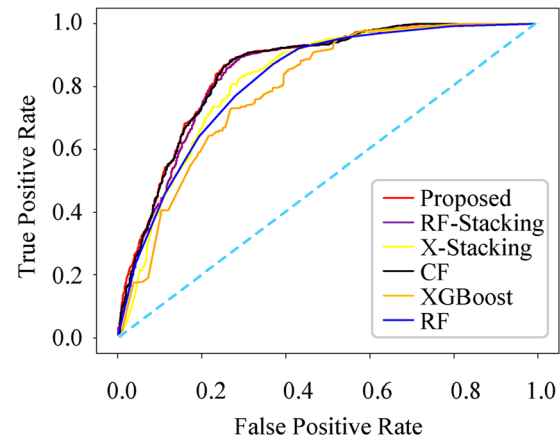


Figure 5. ROC curve of contrast experiment.

Table 4. AUC of contrast experiment.

Model name	AUC	Brier score
Proposed	0.8562	0.1596
RF-Stacking	0.8477	0.1675
X-Stacking	0.8331	0.2500
CF	0.8466	0.1667
RF (Ohata et al.)	0.8126	0.1733
XGBoost (Hong et al.)	0.7932	0.2494

The results indicate that, based on the evaluation metrics of AUC and brier score, the CF-Stacking model consistently exhibits better predictive performance. Compared to the RF-Stacking and X-Stacking models, CF-Stacking achieved an improvement of 1.00 and 2.77% in AUC values, along with a relative reduction of 0.0079 and 0.0904 in brier scores. These results demonstrate the effectiveness of the improved CF model. The integration of Bagging and Boosting ensemble learning methods can effectively learn the spatial features of complex landslides. Compared with the random forest-based model in the traditional CF model, the improved CF model has stronger fusion ability in the face of complex data features. In comparison to CF, RF and XGBoost, CF-Stacking has an increase of 1.13, 5.37 and 7.94% in AUC values, and a relative decrease of 0.0071, 0.0137 and 0.0898 in brier scores. Through the calculation results, it can be seen that the AUC of the proposed method is higher than that of the other five methods, and the Brier score is lower than that of the other five methods. This shows that the introduction of the stacking method integrates the differences of different base models in the improved CF, and has better classification effect than single CF, RF and XGboost. The comparative simulation verified the effectiveness of CF-Stacking, enabling

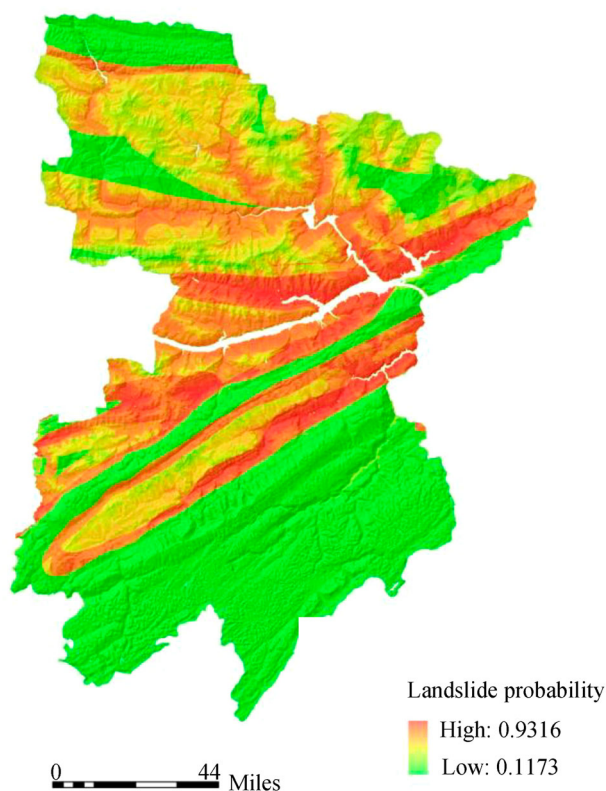


Figure 6. Landslide susceptibility mapping in Fengjie County.

more accurate prediction of the spatial distribution of landslides.

Finally, the complete data set is used to fit the model, and the landslide susceptibility map of Fengjie is obtained, as shown in Figure 6. The colours in different spaces indicate the spatial probability of landslide occurrence.

4. Conclusion and future work

In this paper, we have analyzed the topography and geomorphology, geological conditions and mechanism of Fengjie, preprocessed remote sensing image data and extracted 12 disaster-inducing factors. On this basis, an improved CF combined with RF and XGBoost has been proposed, and the Stacking has been introduced to establish the CF-Stacking model. Finally, we have compared the CF-Stacking model proposed in this paper with RF-Stacking, X-Stacking and other algorithms in Fengjie County, which uses the data set collected from the open source geospatial data cloud and the geological science data publishing system. The experimental results have shown that the proposed method can predict the landslide space

more effectively than the selected comparison methods. The obtained landslide susceptibility map can provide guidance for geological disaster warning.

In future work, a broader range of machine learning models can be explored to enhance the selection of base learners through optimisation techniques, thus obtaining a more rational combination of base learners. After obtaining the area with a high probability of landslide, a small-scale landslide early warning study can be carried out for this area.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the Fundamental Research Funds for the Central Universities, China University of Geosciences; in part by the National Natural Science Foundation of China (grant number 62273317), the 111 project (grant number B17040).

Data availability statement

The data that support the findings of this study are openly available in Geospatial Data Cloud at <https://www.gscloud.cn/>.

References

- Akhtar, M. S., Ekbal, A., & Cambria, E. (2020). How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1), 64–75. <https://doi.org/10.1109/MCI.10207>
- Akın, O. A., & Hüseyin, K. (2023). Strait fuzzy sets, strait fuzzy rough sets and their similarity measures-based decision making systems. *International Journal of Systems Science*, 54(12), 2519–2535. <https://doi.org/10.1080/00207721.2023.2233971>
- Akobeng, A. K. (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Materialia*, 96(5), 644–647.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery. <https://dl.acm.org/doi/10.1145/2939672.2939785>
- Chen, X., Li, J., Zhang, Y., Jiang, W., Tao, L., & Shen, W. (2017). Evidential fusion based technique for detecting landslide barrier lakes from cloud-covered remote sensing images. *IEEE Journal of Selected Topics in Applied*

- Earth Observations and Remote Sensing*, 10(5), 1742–1757. <https://doi.org/10.1109/JSTARS.4609443>
- Dai, X. L., Qu, S., Huang, C. H., & Wu, P. B. (2023). The determination of limit wheel profile for hunting instability of railway vehicles using stacking feature deep forest. *Engineering Applications of Artificial Intelligence*, 125, 106732. <https://doi.org/10.1016/j.engappai.2023.106732>
- Haoyuan, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.-X., & Chen, W. (2018). Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. *Science of the Total Environment*, 625, 575–588. <https://doi.org/10.1016/j.scitotenv.2017.12.256>
- Hariklia, D. S., Bathrellos, G. D., Chousianitis, K., Youssef, A. M., & Pradhan, B. (2019). Multi-hazard assessment modeling via multi-criteria analysis and GIS: A case study. *Environmental Earth Sciences*, 78(2), 47. <https://doi.org/10.1007/s12665-018-8003-4>
- Hong, Z. X., Tao, M., Liu, L. L., Zhao, M. S., & Wu, C. Q. (2023). An intelligent approach for predicting overbreak in underground blasting operation based on an optimized XGBoost model. *Engineering Applications of Artificial Intelligence*, 126, 2023.
- Hussain, M. A., Chen, Z., Wang, R., & Shoaib, M. (2021). PS-InSAR-based validated landslide susceptibility mapping along Karakorum Highway, Pakistan. *Remote Sensing*, 13(20), 4129. <https://doi.org/10.3390/rs13204129>
- Hussain, M. L., Shafique, M., Bacha, A. S., Chen, X., & Chen, H. (2021). Landslide inventory and susceptibility assessment using multiple statistical approaches along the Karakoram highway, northern Pakistan. *Journal of Mountain Science*, 18(3), 583–598. <https://doi.org/10.1007/s11629-020-6145-9>
- Jia, B. B., & Zhang, M. L. (2020). Multi-dimensional classification via stacked dependency exploitation. *Science China Information Sciences*, 63(12), 222102. <https://doi.org/10.1007/s11432-019-2905-3>
- Lei, C., & Mohamed, S. K. (2015). MSEBAG: A dynamic classifier ensemble generation based on ‘minimum-sufficient ensemble’ and bagging. *International Journal of Systems Science*, 47(2), 406–419.
- Li, P. L., Wang, Y., Si, T. Z., Ullah, K., Han, W., & Wang, L. Z. (2024). MFFSP: Multi-scale feature fusion scene parsing network for landslides detection based on high-resolution satellite images. *Engineering Applications of Artificial Intelligence*, 127, 107337. <https://doi.org/10.1016/j.engappai.2023.107337>
- Madrugá, S. P., Tavares, A. H. B. M., Luiz, S. O. D., T. P. do Nascimento, & Lima, A. M. N. (2022). Aerodynamic effects compensation on multi-rotor UAVs based on a neural network control allocation approach. *IEEE/CAA Journal of Automatica Sinica*, 9(2), 295–312. <https://doi.org/10.1109/JAS.2021.1004266>
- Marcio, H. R. S., Sytze, D. B., Carlos, S., & Martin, H. (2021). Land use and land cover area estimates from class membership probability of a random forest classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 4402711.
- Sandeep, P., & Amit, S. K. (2022). Landslide hazard assessment using analytic hierarchy process (AHP): A case study of national highway 5 in India. *Ain Shams Engineering Journal*, 13(3), 101626. <https://doi.org/10.1016/j.asej.2021.10.021>
- Shirvani, Z. (2020). A holistic analysis for landslide susceptibility mapping applying geographic object-based random forest: A comparison between protected and non-protected forests. *Remote Sensing*, 12(3), 434. <https://doi.org/10.3390/rs12030434>
- Siva, S. S. S., Geetha, S., & Kannan, A. (2011). Evolving optimised decision rules for intrusion detection using particle swarm paradigm. *International Journal of Systems Science*, 43(12), 2334–2350.
- Sk, A. A., Farhana, P., Jana, V., Romulus, C., Nguyen, T. T. L., Quoc, B. P., Matej, V., Ljubomir, G., Ateeque, A., & Mohammad, A. G. (2021). GIS-based landslide susceptibility modeling: A comparison between fuzzy multi-criteria and machine learning algorithms. *Geosciences*, 12(2), 857–876.
- Talha, A., Ahmed, W., Haseeb, A., Ahmad, R., Gul, S. T., & Zheng, F.-C. (2019). A low complexity online controller using fuzzy logic in energy harvesting WSNs. *Science China Information Sciences*, 62(4), 42305. <https://doi.org/10.1007/s11432-018-9751-5>
- Wang, X. (2017). Ladle furnace temperature prediction model based on large-scale data with random forest. *IEEE/CAA Journal of Automatica Sinica*, 4(4), 770–774. <https://doi.org/10.1109/JAS.6570654>
- Wang, Y., Zhang, J., & Qiao, J. (2020). An information entropy-based method of evidential source separation and refusion. *IEEE Sensors Journal*, 20(1), 77–84. <https://doi.org/10.1109/JSEN.7361>
- Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H., & Hong, H. (2020). Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *Catena*, 187, 104396. <https://doi.org/10.1016/j.catena.2019.104396>
- Xia, M., & Jia, K. (2022). Reconstructing missing information of remote sensing data contaminated by large and thick clouds based on an improved multitemporal dictionary learning method. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14.
- Xiong, Y., Zhou, Y., Wang, F., Wang, S., Wang, J., Ji, J., & Wang, Z. (2021). Landslide susceptibility mapping using ant colony optimization strategy and deep belief network in Jiuzhaigou region. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 11042–11057. <https://doi.org/10.1109/JSTARS.2021.3122825>
- Xu, G., Wang, Y., Wang, L., Soares, L. P., & Grohmann, C. H. (2022). Feature-based constraint deep CNN method for mapping rainfall-induced landslides in remote regions with mountainous terrain: An application to Brazil. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 2644–2659. <https://doi.org/10.1109/JSTARS.2022.3161383>
- Yuan, Z., Zhang, W., Tian, C., Rong, X., Zhang, Z., Wang, H., Fu, K., & Sun, X. (2022). Remote sensing cross-modal text-image retrieval based on global and local information.

- IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16.
- Zhang, W., Jiang, J., Shao, Y., & Cui, B. (2020). Snapshot boosting: A fast ensemble framework for deep neural networks. *Science China Information Sciences*, 63(1), 112102. <https://doi.org/10.1007/s11432-018-9944-x>
- Zhang, H., Li, Y., Lv, Z., Sangaiah, A. K., & Huang, T. (2020). A real-time and ubiquitous network attack detection based on deep belief network and support vector machine. *IEEE/CAA Journal of Automatica Sinica*, 7(3), 790–799. <https://doi.org/10.1109/JAS.6570654>
- Zhang, W., Wang, J., & Lan, F. (2021). Dynamic hand gesture recognition based on short-term sampling neural networks. *IEEE/CAA Journal of Automatica Sinica*, 8(1), 110–120. <https://doi.org/10.1109/JAS.2020.1003465>
- Zhao, D., Jiao, Y., Wang, J., Ding, Y., Liu, Z., Liu, C., Qiu, Y., Zhang, J., Xu, Q., & Wu, C. (2020). Comparative performance assessment of landslide susceptibility models with presence-only, presence-absence, and pseudo-absence data. *Journal of Mountain Science*, 17(12), 2961–2981. <https://doi.org/10.1007/s11629-020-6277-y>
- Zhou, Z. H., & Feng, J. (2017). Deep forest: Towards an alternative to deep neural networks. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 3553–3559). International Joint Conferences on Artificial Intelligence. <https://dl.acm.org/doi/10.5555/3172077.3172386>