ISSN: 2634 - 8845

Journal of Earth and Environmental Sciences Research

SCIENTIFIC Research and Community

Review Article

Open Access

Evaluation of Landslide Susceptibility by Optimization Integrated Machine Learning Algorithm Based on Gradient Boosting-Take Both Banks of Yarlung Zangbo River and Niyang River as Examples

LIN Qin, GUO Yonggang*, Wu Shengjie, Zang Yeqi and Huang Yongfang

Water Conservancy Project & Civil Engineering College, Tibet Agriculture & Animal Husbandry University, Nyingchi Tibet, 860000, China

ABSTRACT

The geological structures on both banks of the Yarlung Zangbo River and the Niyang River are active, and landslides occur frequently. The landslide susceptibility assessment can effectively reduce the damage to human life and property caused by disasters. This paper studies the performances of Weighted Random Forests, XGBoost and LightGBM algorithms based on Gini coefficient in landslide susceptibility. 188 landslide samples and 7 influencing factors are selected. In the process of model training, taking into account of Feature Selection Algorithm, the hyperparameters are optimized by the using of Bayes' Theorem, Grid Search and Five-fold Cross Validation method. Precision, recall, F1 and Accuracy are used to analyze the prediction results of each level. The results show that landslide is most likely to occur within the elevation of 32-1544m and 2722-3752m, the gradient of 30-40°, and the distance of 200m from the fault zone, river and road. The extremely high and high landslide prone areas account for 12.14% and 12.41% respectively, and the low and extremely low landslide prone areas account for 26.47% and 29.55% respectively. More than half of the areas in Nyingchi Prefecture are not prone to landslide disasters. Among all models, LightGBM model performs best, with AUC value of 0.8432, accuracy of 0.8531, and F1 score of 0.8345. Damu Township and Bangxin Township in Motuo County, Danniang, Lilong, Zhaxi Raodeng Township in Linzhi County, Long Village in Lang County, and Jiangda Township in Gongbujiangda County are positioned in extraordinarily high-risk areas, with a excessive likelihood of landslides. Corresponding prevention and control measures should be taken in these areas.

***Corresponding author**

GUO Yonggang, Water Conservancy Project & Civil Engineering College, Tibet Agriculture & Animal Husbandry University, Nyingchi Tibet, 860000, China.

Received: May 10, 2023; **Accepted:** May 15, 2023; **Published:** May 20, 2023

Keywords: Gradient Lifting, Xgboos, Lightgbm, Machine Learning, Landslide Susceptibility

Introduction

The Yarlung Tsangpo River and the Niyang River are located in the southeastern part of the Tibetan Plateau, and the mountains in the basin are undulating, forming a large number of gullies, canyons and rivers. In the interior of the crust, for active dynamic actions, release of initial high- pressure stress and looseness of rock structure in the basin, natural disasters, such as landslides, debris flows occur frequently [1,2]. Landslides are soil damage caused by natural and human activities [3]. It is a natural disaster characterized by the movement of large amounts of rock, debris or soil towards the slope surface. Landslides, whether caused by natural or human activities, cause significant economic losses every year [4]. Therefore, using efficient and stable landslide disaster assessment technology to quickly and accurately identify the disasters in the landslide prone areas and predict the occurrence of landslide disasters can effectively improve the efficiency of disaster prediction, reduce the losses and provide reference for disaster prevention and reduction.

Landslide susceptibility zoning is an effective method for landslide prediction by predicting the probability of landslide occurrence through the attributes of impact factors after landslide occurrence [5]. Traditional qualitative and quantitative methods are usually used to evaluate landslide susceptibility. Qualitative methods rely on the experience and opinions of experts in historical data and landslide inventories, such as weighted linear combination and analytic hierarchy process, but the calculation results are influenced by human factors. Quantitative methods include data models and deterministic models [6]. Deterministic models can provide accurate analysis results, but require large amounts of data, which are difficult to obtain especially in practice large-scale regional practice [7]. In recent years, data-driven models with machine learning and statistics have made significant progress in geological hazard research, such as the weight-of-evidence (WoE) model, the frequency ratio (FR), and the certainty factor method (CF), which are computationally simple and applicable even in some large areas, but they are overly dependent on sample quality and cannot effectively handle the relationships between complex landslides and their influencing factors. Random forests, decision trees, BP neural networks, and gradient boosting in machine learning have also been widely used in landslide identification, which can better solve the problem of nonlinear relationship expression and improve the accuracy of landslide identification [8-13]. However, these models usually rely on a single learner with numerous influencing factors involved in

landslide susceptibility, and they are usually difficult to obtain ideal prediction results and prone to over-fitting. Therefore, this paper uses ensemble learning to combine multiple single-learners together for regional landslide susceptibility assessment, aiming to prove that the method proposed in this paper is more superior and efficient than traditional ones.

In recent years, a large number of methods based on machine learning have been successfully applied to the study of geological disasters. While, Gradient Boosting models, including XGBoost and LightGBM, have rarely been studied and compared in terms of landslide susceptibility. Coupled with that imbalanced class distribution may affect feature selection. Based on the above-mentioned factors, taking the banks of the Yarlung Zangbo River and the Niyang River as examples, this paper introduces the Weighted Random Forest based on Gini coefficient as the feature selection process for the first time, and analyzes and compares the landslide susceptibility of the study areas with XGBoost and LightGBM models based on Boosting algorithm.

Study Area and Data

Study Area

This paper selects the lower reaches of the Yarlung Zangbo River and both banks of the Niyang River as the research object (Figure 1). The research area is located in the west of Nyingchi City, Tibet Autonomous Region, 92° 09′ - 95° 51′ E, 27° 55′ - 30° 36′ N, with a total area of about 68000 km2, including Gongbujiangda, Bomi, Milin, Lang and Motuo County. The study area belongs to typical plateau hills, alpine and gorge landform, and is the world's biggest fall in land vertical topography. The terrain in the area fluctuates greatly, showing a trend of high in the north and low in the south. The mountains are mostly east-west oriented, most of which are high altitude and large fluctuation, followed by high altitude and medium high altitude and large fluctuation mountains. The highest altitude is 7782m, which is located at the junction of Milin and Motuo County. The study area is located in the plateau temperate humid and semi humid monsoon region from the cold zone to the tropics. The water vapor content in the region is high and the rainy season starts early and ends late, with a long duration. The average annual precipitation is about 650mm, and the average annual temperature is 9.1°C. There are many fault zones in the study area, with complex geological structures and rich rock strata. It is prone to landslides due to high rainfall and dynamic activity within the soil and plate.

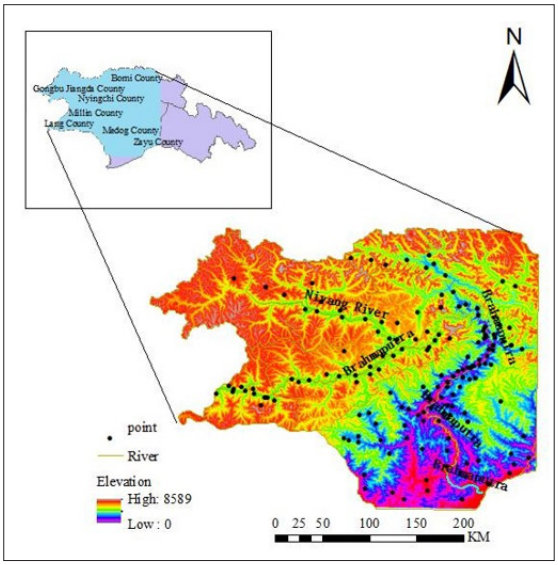

Figure 1: Geographical Location and Landslide Distribution of the Study Area

Sources of Data and Processing

The main data sources include: (1) ASTER GDEM 30m resolution digital elevation data of geospatial data cloud, and slope information is extracted based on ArcGIS software; (2) The 1:50000 geological map is derived from the China Geological Survey to extract the lithological properties of the stratum; (3) Landsat8 images are derived from the general survey of geographical conditions and used for the extraction of land use data; (4) Landslide data are obtained from the Resource and Environmental Science Data Center of the Chinese Academy of Sciences; (5) The fault zones are obtained from the seismic active fault exploration data center. Based on the existing research methods, in this paper, 30m×30m grid size is set as the basic evaluation unit, and the study area is divided into 123156296 grids. Meanwhile, in order to solve the problem of sample imbalance, the text adopts down-sampling method to select the same number of landslide points from the non-landslide area to form 188 sample points [14,15]. The landslide unit is set as 1, and the non landslide unit is set as 0. 70% (131) of the data is randomly selected as the training samples, and the remaining 30% (57) is taken as the test samples. Specific flow chart of landslide point as shown in the following:

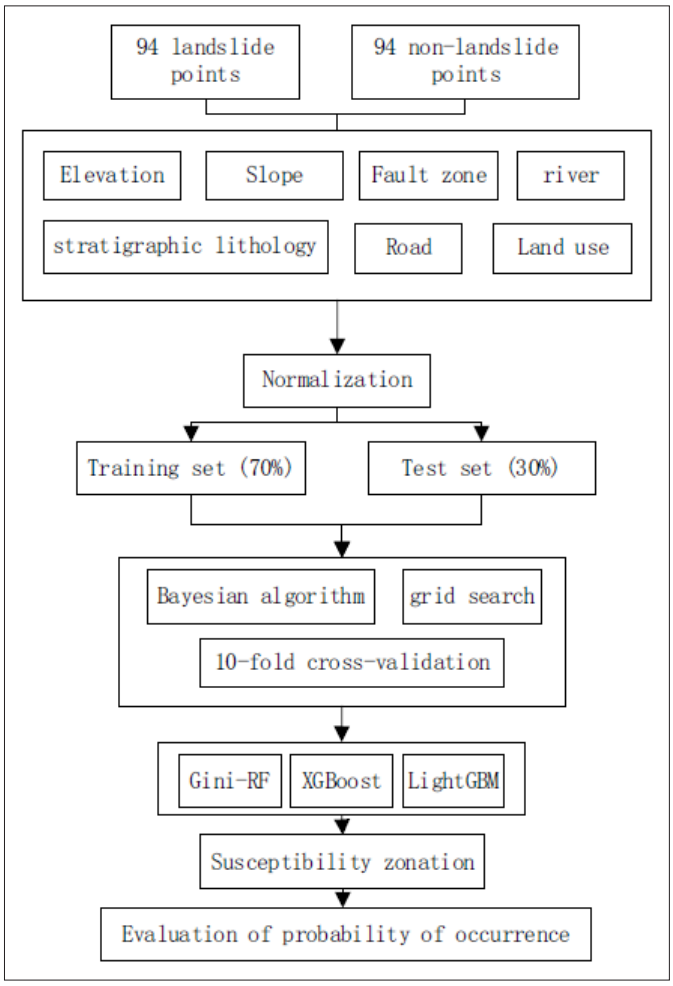

Figure 2: Flow Chart

Selection of Evaluation Factors and Independence Test

Selection of Evaluation Factors

Existing research results and field survey of the Yarlung Zangbo River basin show that the continuous erosion of water to the valley and the weathering of rocks in the landslide area aggravated by freeze-thaw in the high-altitude and alpine areas make the Yarlung Zangbo River basin prone to landslides [16]. Formation lithology is an important factor for landslide generation [17]. And slope is the main controlling factor of landslide occurrence [18]. Then based on the research and analysis of the formation conditions of geological disasters and geological environment background in the study area, seven evaluation factors of elevation, slope, fracture zone and fault, river, road, stratigraphic lithology, and land use are selected in this paper. Using ArcGIS software, combined with distribution norms, the Fisher Jenks algorithm is used to divide the study area of the four continuous-type factors of elevation, slope, stratum lithology, and land use into five grades (Figure 3(a-d)), and for discrete-type factors such as fracture zones and faults, rivers and roads, the multi- ring buffer tool is used to establish 5 grade ranges: 0-200, 200-400, 400-600, 600-800 and >800m (Figure 3(e-g)).

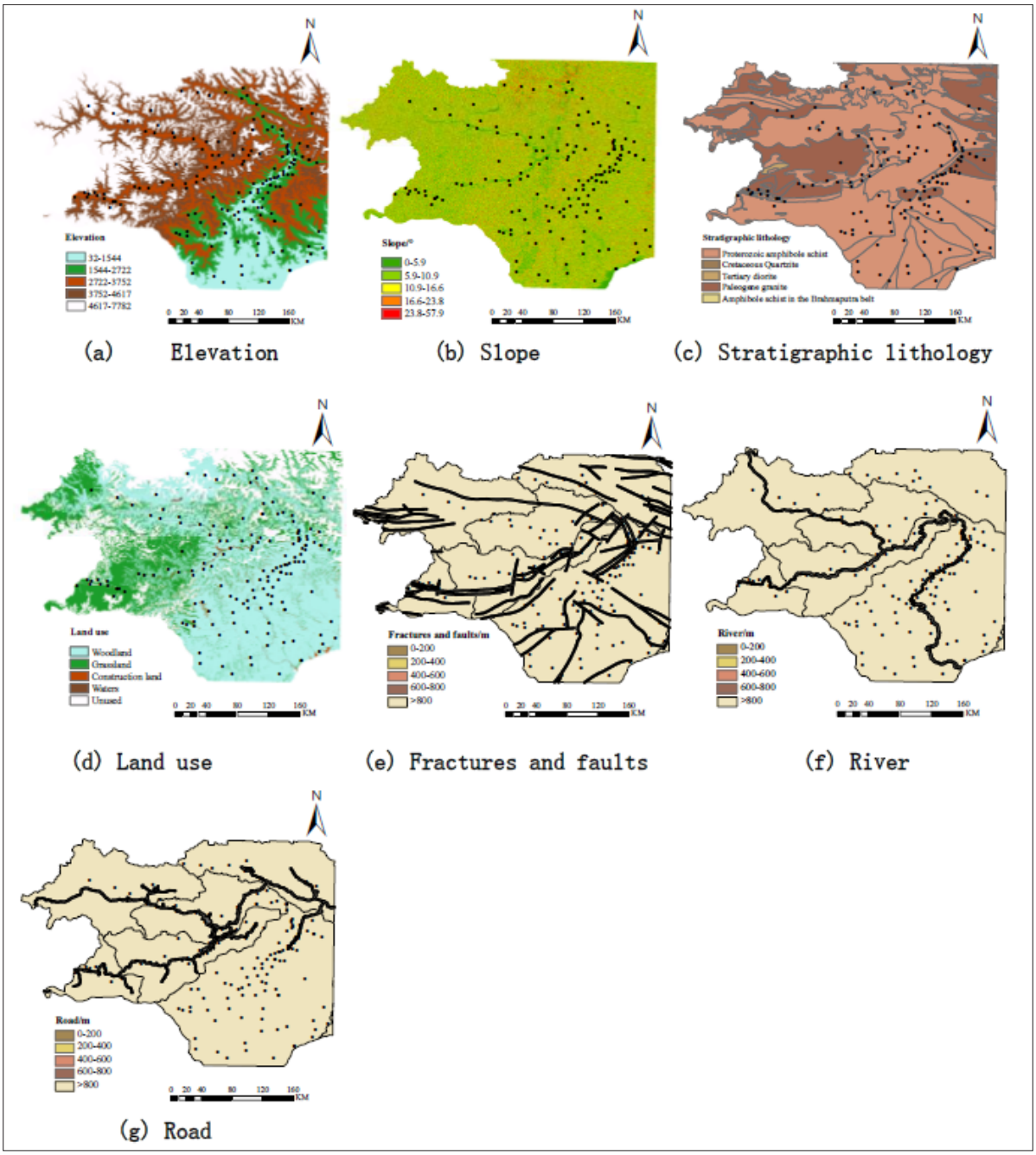

Figure 3: Grading Chart of Evaluation Factors

The number of landslide points in the grading range of each evaluation factor is counted and clustered column charts are drawn (Figure 4). The results show that when the altitude is between 32-1544m, the landslide occurs the most, accounting for 30.9% of the total number, followed by those occurring in the range of 2722-3752m. The reason is that when the altitude is lower than 1544m, human activities such as excavating the foot of the slope are frequent. As the altitude rises, the slope increases, which intensifies the

occurrence of landslides; as the slope rises, the number of landslides also increases until the slope reaches the threshold of 40°, and the probability of disaster decreases, from the original 41.5% to 16.0% gradually. When the lithology of the local stratum is diorite schist, compared with other lithology, landslides occur most frequently; Grassland soil erosion is serious, which is an important cause for shallow landslide. In this paper, a large number of landslide points are distributed on the grassland with a slope of 10-20°. The fracture zone and fault will reduce the strength and integrity of the rock stratum, which is the key to increase the landslide susceptibility. Landslides are prone to occur within 200m from the fault zone, and the landslide points account for 41.5% of the total number. The farther away from the fault zone, the less landslide disasters will occur. The river bank is constantly scoured by water, and the soil and rock are more unstable under the action of groundwater and gravity. Therefore, the closer to the river, the more likely landslide will occur. The landslide occurs within 200m away from the river, with the highest frequency of 40.4%. Due to vigorous blasting and forced excavation in the construction of railways and highways, the lower part of the slope often loses its support and slides. The number of landslides within 200m from the road accounts for more than half of the total, reaching 52.1%, the farther away from the road, the less landslide activities. The conclusions in this paper are consistent with relevant studies [19,20].

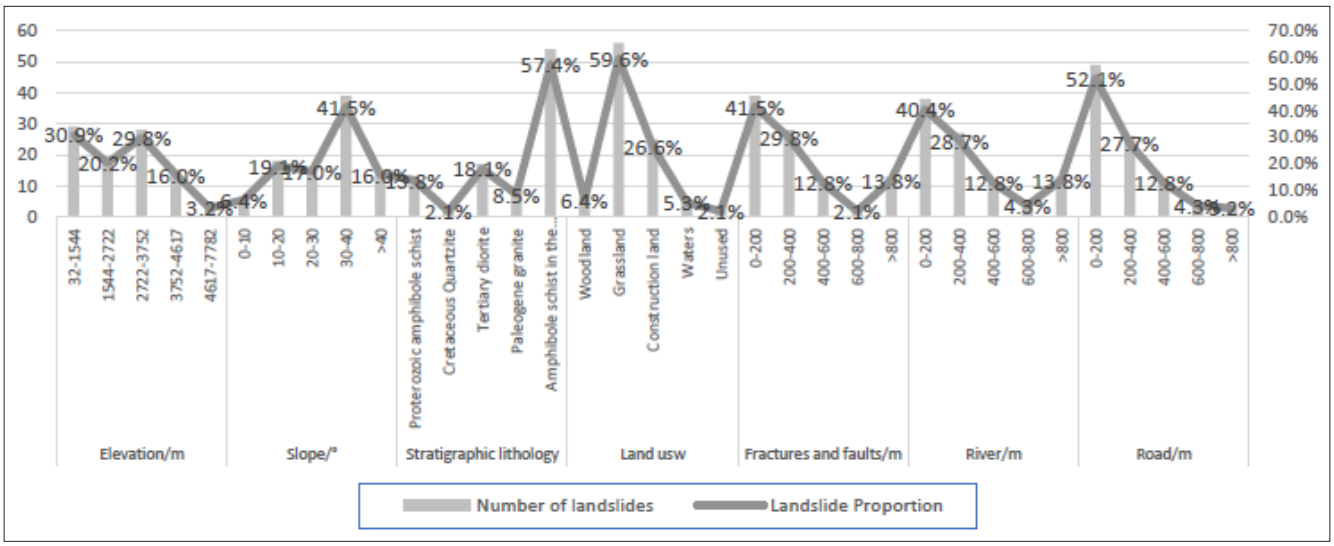

Figure 4: Relationship between assessment factors and disaster points

Independence Test of Evaluation Factors

In order to study the relative independence of each evaluation factor and the accuracy and reliability of the evaluation model, Pearson correlation coefficient is used to calculate the correlation of the influencing evaluation factors. Pearson correlation coefficient is used to measure the linear relationship between two variables, which is calculated by using the covariance and standard deviation of two variables [21].

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}} \quad (1)$$

X and Y are variables, and N is the number of values.

When the correlation between variables is very weak, the correlation coefficient is between 0.0-0.2. And 0.2-0.4 indicates weak correlation between variables. The 7 attribute values of the evaluation factors are substituted into Formula (1) for calculation and the results are shown in Table

1. It is found that the highest correlation is between slope and road (R=0.3493), and the correlation between other variables is less than 0.4. In general, the collinearity of variables is not strong.

Table 1: Pearson correlation coefficient between factors

	Elevation	Road	River	Slope	Fractures and faults	Stratigraphic lithology	Land use
Elevation	1	-0.1624	0.1554	-0.1708	0.2317	-0.2564	-0.0298
Road	-0.1624	1	0.1405	0.3493	-0.2076	-0.093	0.0025
River	0.1554	0.1405	1	0.1269	-0.0672	0.3011	0.0122
Slope	-0.1708	0.3493	0.1269	1	-0.2371	-0.051	-0.0649
Fractures and faults	0.2317	-0.2076	-0.0672	-0.2371	1	-0.196	-0.2654
Stratigraphic	-0.2564	-0.093	0.3011	-0.051	-0.196	1	0.0725
Lithology Land use	-0.0298	0.0025	0.0122	-0.0649	-0.2654	0.0725	1

Assessment of Landslide Susceptibility along the Banks of Yarlung Zangbo River and Niyang River

Evaluation of Landslide Susceptibility based on Gini-RF

Random Forest is an optimized version of Bagging (Bootstrap Aggregation) based on Decision Tree. Because of its strong robustness, applicability to high-dimensional dense data, parallel integration, automatic error adjustment for unbalanced data sets, fine tuning of super parameters and other advantages, accurate results can be obtained, which is often used in various classification and regression tasks [22]. Its basic unit is the Decision Tree, but its essence is ensemble learning method, which is a branch of machine learning. Its core idea is always Bagging. For some special improvements have been made, random forest uses the CART Decision Tree as the basic learner.

Random forest based on Gini coefficients is built on many decision trees and supports various feature weighting measures. One of them is the correlation between features and imbalanced data output. Once the Gini coefficient is measured by the classifier, this feature selection technique adopts the weight adjustment technique in RF. Gini index has the ability to divide binary classes in specific nodes [23]. For attributes with more than two different values, a subset of attributes is considered and the feature importance score is calculated by adjusting the weights in the random forest algorithm with unbalanced class distribution and splitting the tree using the Gini coefficient criterion. The higher the GI value is, the average contribution of the feature to the model prediction is greater and the explanatory ability of the model is more better. The sum of all GI characteristics is 1. The specific formula is as follows:

$$GI_{\mathrm{m}} = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \quad (2)$$

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \quad (3)$$

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \quad (4)$$

GI_{m} is Gini coefficient, K stands for K categories, p_{mk} Represents the proportion of k in node M; $VIM_{ij}^{(Gini)}$ Represents the weight of feature I in the J[th] tree; Equation (4) indicates that all the calculated important degree scores are normalized.

In this paper, 94 landslide points are noted as 1 and an equal number of non-landslide points are noted as 0. The attributes of 7 evaluation index factors are extracted to the training set, and a random forest binary classification model is constructed. Random Forest Classifier method is called from SkLearn library to substitute the training set into the RF model for training. Meanwhile, to ensure the reliability and accuracy of the results, Bayesian optimization algorithm is used to search for the optimal parameter values based on the original parameter settings. In the optimization results, the best results are obtained when the step size is taken as 0.1, max_depth is taken as 4, and num_round is taken as 30 when the weights are updated after each iteration is completed. Finally, the RF model is predicted with the test set, and the weights of each evaluation factor are normalized and imported into the raster calculator in ArcGIS to generate the landslide susceptibility map. using the Fisher Jenks algorithm to divide the partition map into five grades: extremely high, high, medium, low and very low (Figure 5). The higher the susceptibility, the more likely landslides will occur.

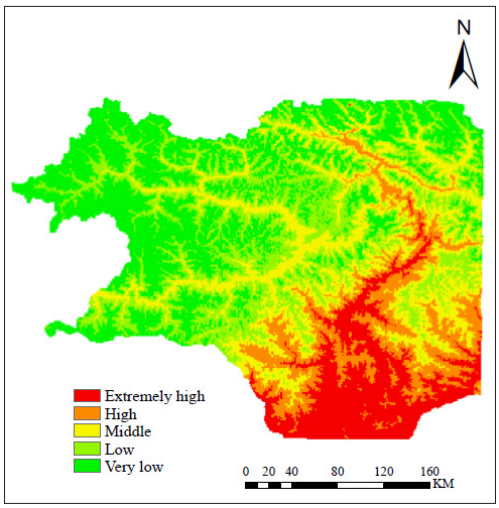

Figure 5: Susceptibility Zoning Map of Gini-RF

Evaluation of Susceptibility of XGBoost

XGBoost is an integrated machine learning algorithm based on decision tree and gradient boosting. In order to control the complexity of the model, it adds a regularization term which includes the sum of squares of the weights of each leaf node and the number of nodes, to the loss function. XGBoost processes the missing values and selects the best default segmentation direction for missing values by the learning model [24] .

After the preprocessing process of the data described in 4.1, Scikit-Learn is used to build the XGBoost multi-split landslide susceptibility model based on Python3.6 and R language [25]. In

order to test the subsequence on an independent validation data set to reduce the contingency, the optimal subtree is selected, several sets of hyperparameter combinations are preset by grid search, and the average value of each model evaluation metric is obtained by using the Five-fold cross validation. The average index of all test sets is considered as the final result. The prediction results are imported into ArcGIS to draw a landslide susceptibility map (Figure: 6) Figure: 7 shows the cross-validation accuracy results for the sample set on the selected parameter values. After the fifth five-fold crossover, the AUC values of the training and testing sets reach the maximum value and tend to be stable.

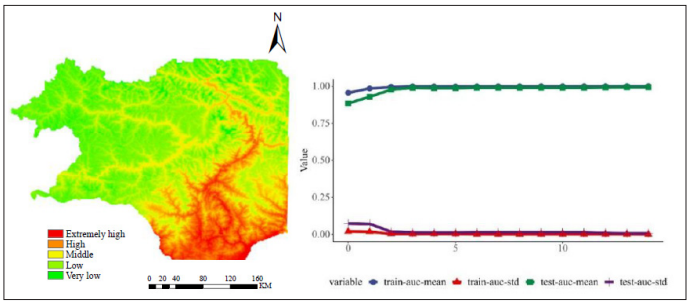

Figure 6: Susceptibility Zoning Map of XGBoost Figure: 7 XGBoost 50% Cross Validation Results

Evaluation of susceptibility of LightGBM

Light Gradient Boosting Machine (LightGBM) is a high performance, open source and fast method for classification, regression and ranking. It is also a Gradient Boosting algorithm based on decision tree algorithm. LightGBM adopts histogram algorithm to reduce memory consumption and make data segmentation easier. Continuous features of floating point are discretized into discrete values in the formula, and a histogram with width of k is constructed. The data is trained through traversal, and the cumulative statistical information of each discrete value in the histogram is calculated. In feature selection, it is only necessary to search for the best segmentation point according to the discrete values of the histogram [26].

On the basis of 4.2 usage method, the 123156296 grids in the study area are extracted from the attribute values of each evaluation factor to the point, and the 123156296X7 table is generated. The table is imported into the trained machine learning model to predict the probability of landslides in each grid. All the points are generated into grid data using the Point to Raster tool. Then the natural discontinuity method is used to divide the landslide prone areas in the study area into five categories: extremely high, high, medium, low and very low (Figure: 8).

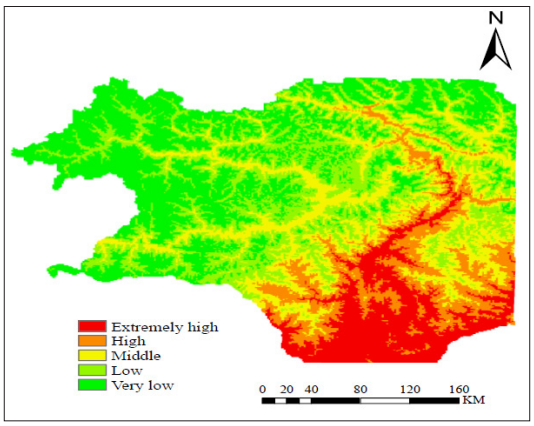

Figure 8: Susceptibility zoning map of Gini-RF

Validation of landslide susceptibility evaluation results

Vulnerability Partition Results and Comparison

Based on ArcGIS, the number of grids and landslide points of three different machine learning models in each vulnerability zone are counted respectively (Table 2). The landslide vulnerability results of the three models show some differences, but tend to be the same as a whole. Gini RF, XGBoost and LightGBM models all have the highest percentage values in the very low category. For Gini RF model, the area ratios of extremely high to very low susceptibility are 11.99%, 12.63%, 19.58%, 26.77% and 29.03% respectively. The percentage ratio of XGBoost model shows that the extremely high, high, medium, low and very low risk areas account for 12.05%, 12.50%, 19.62%, 26.78% and 29.05% respectively. For the LightGBM model, the very low, low, medium, high and extremely high risk areas account for 12.14%, 12.41%, 19.43%, 26.47% and 29.55% respectively. It can be seen from the distribution of landslide locations that most historical landslide records are located in highly prone areas, as predicted by Gini RF, XGBoost and LightGBM models. The results show that LightGBM model has the highest performance, followed by XGBoost and Gini RF.

Table 2: Comparison of Machine Learning Model Vulnerability Zones

	Machine learning model											
	Gini-RF				XGBoost				LightGBM			
	Number of grids	Grid ratio	Number of landslide points	Landslide Proportion	Number of grids	Grid ratio	Number of landslide points	Landslide Proportion	Number of grids	Grid ratio	Number of landslide points	Landslide Proportion
Extremely high	14766439	11.99%	44	23.40%	14840333	12.05%	52	27.66%	14951174	12.14%	56	29.79%
High	15554640	12.63%	68	36.17%	15394537	12.50%	72	38.30%	15283696	12.41%	75	39.89%
Middle	24114003	19.58%	38	20.21%	24163265	19.62%	40	21.28%	23929268	19.43%	42	22.34%
Low	32968940	26.77%	22	11.70%	32981256	26.78%	10	5.32%	32599471	26.47%	8	4.26%
Very low	35752274	29.03%	16	8.51%	35776905	29.05%	14	7.45%	36392714	29.55%	7	3.72%

According to the selection of assessment factors in 2.1 and the vulnerability assessment zoning map, it can be seen that the high and extremely high landslide prone areas are mostly located in Damu and Bangxin Township of Motuo County, Danniang, Lilong and Zhaxi Raodeng Township of Nyingchi County, Long Village of Lang County and Jiangda Township of Gongbujiangda. Appropriate geological disaster prevention and control measures should be taken in these areas, especially the areas located on the banks of Yarlung Tsangpo River and Niyang River with low elevation, slope between 30°and 40°, and within 200m from rivers, roads and fracture zones.

The reason is that this kind of area is located in the south of both banks of the Yarlung Zangbo River and the Niyang River, which is at the border of India plate and Eurasian plate. The crustal movement is intense, which breeds a series of regional faults. The fault zones and faults reduce the integrity and strength of the rock stratum. The elevation is among 200-1000 meters, and most of the slopes are less than 40°. In this range, intensive activities such as slope cutting, building and road construction are carried out manually, resulting in a large number of exposed slopes. In addition, the long-term water action causes serious erosion and scouring on both banks of the river, resulting in sediment saturation, thus reducing the integrity of the slope and causing slope movement or mass movement. And the closer to the road, the destructiveness caused by road construction will have a negative impact on the stability of the slope, so landslide disasters occur frequently.

On the contrary, the low landslide prone areas are mainly distributed in Cuogao and Zhula districts of Gongbujiangda County, Chongguoye and Gangaru districts of Nyingchi City, and Sulu fat area of Milin County, which are characterized by relatively slow slopes, less human activities, and far away from roads, rivers and fault zones.

Comparison of Model Accuracy

In machine learning, performance metrics are often used to predict the correct number of test sets in binary classification. In this paper, six indexes including Accuracy, Precision, Recall, F1 score, ROC curve and AUC value are used to evaluate the accuracy of different machine learning models. Accuracy score is the most commonly used index to evaluate the performance of a model in binary classification problems. It represents the probability of being correctly identified among all samples. Accuracy is a measure to evaluate the performance of a model by calculating the frequency of positive examples when the model prediction is true. Recall rate is a measure of the model to detect true positive instances correctly. F1 score is the trade-off index between recall and accuracy, and both FP and FN are taken into account to make the model more accurate overall. The specific formula is as follows:

$$准确度 = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$精确度 = \frac{TP}{TP+FP} \quad (6)$$

$$召回率 = \frac{TP}{TP+FN} \quad (7)$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \quad (8)$$

TP and TN are true positives and true negatives respectively, representing the number of correctly classified pixels, and FP and FN are false positives and false negatives respectively, representing the number of incorrectly classified pixels.

In order to obtain the prediction accuracy of different machine learning algorithms on the test data set, the generalization ability and accuracy of the models are derived based on the above method using Eqs. (5)-(8) to calculate the precision, accuracy, recall and F1 index, and 30% samples are randomly selected as test samples to obtain the generalization ability and accuracy of the model (Table 2). It can be seen that the prediction performance based on different framework algorithms is not the same. Among the three machine learning models, the AUC (0.8432), ACC (0.8531), F1 score (0.8345) and Precesion(0.8251) of LightGBM model under hyperparameter optimization are higher than those of the other two machine learning models.

Table 3: Accuracy of each machine learning model

Machine learning model	Gini-RF	XGBoost	LightGBM
AUC	0.7524	0.8035	0.8256
5-fold	0.8225	0.8358	0.8432
ACC	0.7234	0.8148	0.8256
5-fold	0.7534	0.8359	0.8531
F1-score	0.7752	0.7867	0.8021
5-fold	0.8026	0.8256	0.8345
Precesion	0.7834	0.7968	0.8045
5-fold	0.8026	0.8132	0.8251

In machine learning, ROC curve is widely used in binary classification problems to evaluate the reliability of classifiers [27]. AUC is the area under the ROC curve. AUC = 1 indicates that there is at least one threshold for perfect prediction of the curve. The vertical axis of the curve is the true positive rate TPR, and the horizontal axis is the false positive rate FPR. The closer it is to the upper left corner, the better the predictive ability of this indicator is. It can be seen from this ROC curve that the blue curve LightGBM model after grid search and 5-fold cross-validation is closer to the upper left corner, with AUC value of 0.8432, which is significantly improved compared with 0.8225 of Gini RF model, and the accuracy is higher than 0.9358 of XGBoost model (Figure: 9). Compared with GINI-RF, XGBoost improves the loss function of the model and adds the regularization term of model complexity. LightGBM optimizes the training speed of the model on the basis of XGBoost. Therefore, LightGBM has the best generalization ability and high reliability of susceptibility partitioning.

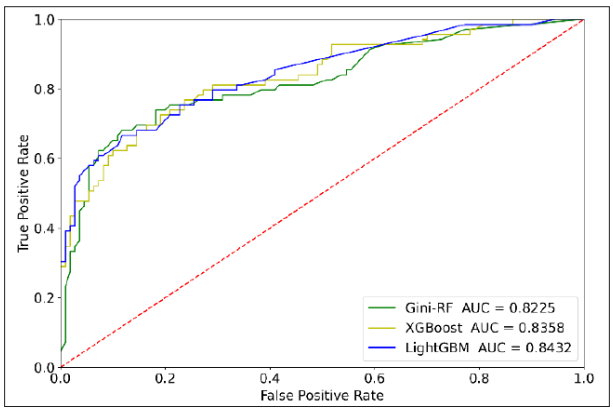

Figure 9: AUC curve of machine learning model

Typical landslide Verification

Table 4 shows the landslide events on both banks of Yarlung Zangbo River and Niyang River in recent years. The landslide susceptibility map generated by importing 9 landslide information shows that 3 landslide points are located in the medium-prone area, 3 are located in the high-prone area, and the rest are all located in the extremely prone area.

Table 4: Landslide events in recent years

Area	Location	Time of occurrence	Source	Predisposition Partition
Gala Village, Nyingchi City	29°41′45″N 94°54′04″E	2018.10.29	Xinhua News Agency	Middle
7 kilometers downstream of Gala Village, Nyingchi City	29°41′27″N 94°54′24″	2022.01.22	China Youth Network	Middle
Qiangna Natural Village, Suotong Village, Guxiang, Bomi County, Nyingchi City	30°00′21″N 95°27′41″	2017.8.24	China Military Television Network	Middle
K80, National Highway 560, Langxian District, Nyingchi City	29°04′03″N 92°49′24″E	2022.7.22	Lang County Public Security Bureau	High
Gala Village, Pai Town, Milin County, Nyingchi City	29°41′45″N 94°54′04″E	2018.10.17	Voice of Tibet	High
Lang County, Nyingchi City	29°04′42″N 93°00′48″E	2022.7.23	Lang County Housing and Urban-rural Development Bureau China Natural Resou	High
Damu Township, Medog County, Nyingchi City	29°29′35″N 95°27′46″E	2021.7.4	Department of Transportation of Tibet Autonomous Region rces News	Extremely high
National Highway 559 from Bomi to Medog	29°19′14″N 97°02′03″E	2019.5.16	西藏自治区交通运输厅	Extremely high
Damuloba Minority Township Primary School, Medog County, Nyingchi City	29°29′46″N 95°27′52″E	2020.8.26	Beijing News	Extremely high

In order to further verify the reliability of the analysis method in this paper, two landslide site investigations, the Qiangna Baga landslide and the Murdoch County road landslide, are selected for comparison and verification (Figure: 10).

The landslide in Baga Village, Qiangna Township, Milin County, Nyingchi City, Tibet Autonomous Region is located at 29°20′16 ″N, 94°24′34″E, where is high mountain valley landform. The underlying bed is slate. The slope structure is rock soil composite with a gradient of 30°. The vegetation cover is average and the land use is lower. The front edge of the landslide to the road below the slope, the back edge to the ridge of the slope, the landslide body is mainly debris and rock and the sliding bed is slate. The main deformation feature of the landslide is the slope foot excavated by the road ahead, which leads to the slope instability.

The road of Motuo County in Nyingchi City is located at 29°08 ‘28 “N and 93°38’ 10” E. The landslide is 30m long, 40m wide, 2m thick, with an area of 1200m² and a volume of 2400m³. Slope is 35°and slope direction is 260°. The side boundary and leading edge of the landslide are clearly discernible. The micro landform of the landslide is a steep slope. The formation lithology is mudstone (OTJ), and it is located near Bailong fault. The slope structure is an soil slope with a convex shape. Under the landslide, there are few human activities, only a short stretch of highway, low vegetation coverage and low shrubs, and the landslide is located on the right convex bank of the river. The current condition is unstable.

The two landslides are located in the high landslide susceptibility areas, which verifies the

accuracy of the machine learning model zoning proposed in this paper again. The research results can be used as the reference for regional landslide prevention and control departments.

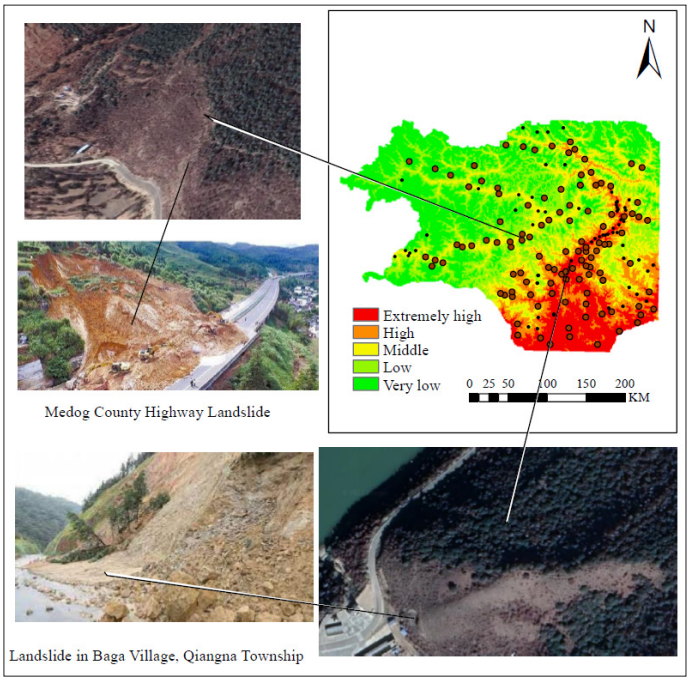

Figure 10: Verification of Typical Landslides

Conclusion

Taking both banks of the Yarlung Zangbo River and the Niyang River as examples, the same amount of 94 landslides and 94 non-landslides are divided randomly. Among them, 70% is slected as the training set, and the remaining 30% as the test set. Three integrated machine learning models, Gini RF, XGBoost and LightGBM, are used to analyze the landslide susceptibility. The conclusions are as follows:

1. The number of landslide points within the grading range of each evaluation factor is counted, and the results show that landslides occur most frequently in the areas which have the features of elevations of 32-1544m and 2722-3752m, slopes of 30-40°, amphibole formation lithology, grassland, and within 200m from the fault zone, rivers and roads.
2. The accuracy of Gini-RF model based on Bayesian optimization algorithm is improved from 0.7524 to 0.8225 after the use of Five-fold cross validation, and the accuracy of XGBoost and LightGBM models with superparameters obtained by grid search is also improved by 0.0323 and 0.0176, respectively. The three models have high accuracy for landslide zoning in the study areas, among which LightGBM model has the best performance, and the accuracy rate of AUC value, accuracy, F1 score, generalization ability and fitting degree is higher.
3. Three machine learning algorithms are used to analyze the study area, which shows that the high and extremely high landslide prone areas are mostly located in Damu and Bangxin Township of Motuo County, Danniang, Lilong, Zhaxi Raodeng Township of Linzhi County, Long Village of Lang County, and Jiangda Township of Gongbujiangda. Especially, the areas located on the banks of Yarlung Tsangpo River and Niyang River with low elevation, slope between 30°and 40°, and within 200m from rivers, roads and fracture zones. Corresponding geological disaster prevention and control measures should be taken in these areas.
4. The extremely high and high landslide prone areas account for 12.14% and 12.41% respectively, and the low and very low landslide prone areas account for 26.47% and 29.55% respectively. More than half of the areas in the region are not prone to landslide disasters. The results of landslide susceptibility zoning are in good agreement with the results of on-site landslide disaster investigation. At the same time, the landslide points that have occurred in the study area in recent years are used for verification, which shows that the reliability of the model is high, and the landslide zoning map can provide guidance for the disaster prevention and mitigation activities of relevant local departments.

References

1. Su Libin, Guo Yonggang, Wu Yue, Yang Yongtao (2020) Geomorphological Analysis of the Niyang River Basin Based on DEM [J]. Chinese Soil and Water Conservation Science 18: 12-21.
2. Wu Chenshuang (2021) GIS-based Geological Hazard Assessment of Nyingchi Section of Sichuan-Tibet Railway [D]. Tibet University
3. Taalab K, Cheng T, Zhang Y (2018) Mapping landslide susceptibility and types using Random Forest[J]. Big Earth Data 2: 159-178.
4. Tien Bui D, Shahabi H, Shirzadi A, Chapi K, Alizadeh M, et al. (2018) Landslide detection and susceptibility mapping by airsar data using support vector machine and index of entropy models in cameron highlands, malaysia[J]. Remote Sensing 10: 1527.
5. Shen LL, Liu LY, Xu Chong, WANG Jingpu (2016) Multi-models based landslide susceptibility evaluation—illustrated with landslides triggered by minxian earthquake[J]. Journal of Engineering Geology 24: 19-28.
6. Rehman A, Song J, Haq F, Mahmood S, Irfan Ahamad M, et al. (2022) Multi-Hazard Susceptibility Assessment Using the Analytical Hierarchy Process and Frequency Ratio Techniques in the Northwest Himalayas, Pakistan[J]. Remote Sensing 14: 554.
7. YANG CQ, TAO P, YANG Z (2022) Landslide susceptibility zoning based on logistic regression tree coupled entropy index model:case of landslide in Wuqi County,Yan'an City,Shaanxi Province[J]. People's Yangtze River 53: 128-134.
8. Batar AK, Watanabe T (2021) Landslide susceptibility mapping and assessment using geospatial platforms and weights of evidence (WoE) method in the Indian Himalayan region: Recent developments, gaps, and future directions[J]. ISPRS International Journal of Geo-Information 10: 114.
9. Khan H, Shafique M, Khan MA, Bacha MA, Shah SU, et al. (2019) Landslide susceptibility assessment using Frequency Ratio, a case study of northern Pakistan[J]. The Egyptian Journal of Remote Sensing and Space Science 22: 11-24.
10. QIAO De-jing, WANG Nian-qin, GUO You-Jin (2020) Landslide susceptibility assessment based on weighted certainty factor model[J]. Journal of Xi'an University of Science and Technology 40: 259-267.
11. Arabameri A, Pradhan B, Rezaei K, Wook Lee C (2019) Assessment of landslide susceptibility using statistical-and artificial intelligence-based FR–RF integrated model and multiresolution DEMs[J]. Remote Sensing 11: 999.
12. Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, et al. (2018) Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)[J]. Catena 163: 399-413.
13. ZHANG LF, WANG JY, ZHANG MS, Shebin C, Tao W (2022) Evaluation of Regional Landslide Susceptibility Assessment Based on BP Neural Network[J]. Northwest Geology 55: 260-270.
14. Tanyas H, Rossi M, Alvioli M, van Westen CJ, Marchesini I (2019) A global slope unit-based method for the near real-time prediction of earthquake-induced landslides[J]. Geomorphology 327: 126-146.
15. Polykretis C, Chalkias C (2018) Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models[J]. Natural hazards 93: 249-274.
16. ZHAO YH (2019) Deformation and Failure Model and Evolution Process of Giant Landslides in Gagong Valley in the Yarlung Zangbo River Basin[J]. Journal of Institute of Disaster Prevention 21: 1-7.
17. ZHAO YH (2021) Research on Development Characteristics and Failure Process of Highway Landslide along the Yarlung Zangbo River[J]. Highway 66: 6-10.
18. WANG RQ, WANG XL, LIU HY (2019) Identification and main controlling factor analysis of collapse and landslide based on fine dem——taking jiacha-langxian section of yarlung zangbo suture zone as an example[J]. Journal of Engineering Geology 27: 1146-1152.
19. Kouhartsiouk D, Perdikou S (2021) The application of DInSAR and Bayesian statistics for the assessment of landslide susceptibility[J]. Natural Hazards 105: 2957-2985.
20. Zweifel L, Samarin M, Meusburger K, Alewell C (2021) Investigating causal factors of shallow landslides in grassland regions of Switzerland[J]. Natural Hazards and Earth System Sciences 21: 3421-3437.
21. Lee DH, Kim YT, Lee SR (2020) Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions[J]. Remote Sensing 12: 1194.
22. Alsahaf A, Azzopardi G, Ducro B, Veerkamp RF, Petkov N (2018) Predicting Slaughter Weight in Pigs with Regression Tree Ensembles[C]//APPIS 1-9.

23. Disha RA, Waheed S (2022) Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique[J]. Cybersecurity 5: 1-22.
24. Inan MSK, Ulfath RE, Alam FI (2021) Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis[C]//2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). IEEE 1046-1050.
25. Alsahaf A, Azzopardi G, Ducro B, Veerkamp RF, Petkov N (2018) Predicting Slaughter Weight in Pigs with Regression Tree Ensembles[C]//APPIS 2018: 1-9.
26. Zeng H, Yang C, Zhang H, Wu Z, Zhang J, et al. (2019) A lightGBM-based EEG analysis method for driver mental states classification[J]. Computational intelligence and neuroscience 2019.
27. ZHANG Q K, LING S X, LI X N, et al.Comparison of landslide susceptibility mapping rapid assessment models in Jiuzhaigou County,Sichuan province,China[J]. Chinese Journal of Rock Mechanics and Engineering, 2020, 39(8): 1595-1610.