

¹ Cover Letter

² **Combining class-weighted algorithm and machine learning models in landslide susceptibility**
³ **mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China**

⁴ Huijuan Zhang, Yingxu Song, Shiluo Xu, Yueshun He, Zhiwen Li, Xianyu Yu, Ye Liang, Weicheng Wu, Yue Wang

⁵ Dear Editors-in-Chief,

⁶
⁷ please find the enclosed manuscript "Combining class-weighted algorithm and machine learning models in landslide
⁸ susceptibility mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China" which we are submit-
⁹ ting for exclusive consideration for publication in Computers & Geosciences. We confirm that the submission follows
¹⁰ all the requirements and includes all the items of the submission checklist.

¹¹
¹² In this contribution, to solve the imbalanced landslide samples (landslides, non-landslides) in the landslide suscepti-
¹³ bility evaluation, the application of the class-weighted algorithm combined with traditional machine learning (logistic
¹⁴ regression) and ensemble machine learning models (LightGBM and random forest) have been investigated. Wanzhou
¹⁵ section of the Three Gorges Reservoir area, China, where the number of landslide samples is 19 times more than
¹⁶ non-landslide samples, is chosen as an example. The landslide inventory database was produced using field investi-
¹⁷ gation and remote sensing images provided by Google Earth. Of the 233 landslides in the inventory, 40% were used
¹⁸ for validation, and the remaining 60% were used for training purposes. Twelve environmental parameters (elevation,
¹⁹ slope, aspect, curvature, distance to river, NDVI, NDWI, rainfall, seismic intensity, land use, TRI, lithology) were
²⁰ used as inputs of the models to produce landslide susceptibility map (LSM). The AUC value, Balanced accuracy, and
²¹ Geometric mean score were used to estimate the quality of models. Research has found that the weighted models
²² (weighted logistic regression, weighted LightGBM, weighted random forest) are better than unweighted methods and
²³ the weighted random forest method has the best performance. The class-weighted algorithm turned the susceptibility
²⁴ evaluation problem into a cost-sensitive problem by setting unequal weights for different classes, which is probably to
²⁵ be applied to the landslide susceptibility evaluation in other areas.

²⁶
²⁷ We provide the source codes in a public repository with details listed in the section "Code availability".

²⁸
²⁹ Thanks for your consideration.

³⁰ Sincerely,

³¹
Yingxu Song
Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, School of Information and
Engineering, East China University of Technology, Nanchang, 330013, Jiangxi, China; yxsong@ecut.edu.cn

32 Highlights

33 **Combining class-weighted algorithm and machine learning models in landslide susceptibility**
34 **mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China**

35 Huijuan Zhang,Yingxu Song,Shiluo Xu,Yueshun He,Zhiwen Li,Xianyu Yu,Ye Liang,Weicheng Wu,Yue Wang

- 36 • The imbalanced landslide samples (landslides, non-landslides) in the landslide susceptibility evaluation is em-
37 phasized.
- 38 • The class-weighted algorithm combined with machine learning (Logistic regression) and ensemble machine
39 learning models (LightGBM and random forest) were applied to the landslide susceptibility evaluation.
- 40 • The weighted models are applicable for solving the problem of imbalanced landslide samples and have improved
41 the landslide susceptibility mapping well.

42 Combining class-weighted algorithm and machine learning models
43 in landslide susceptibility mapping: a case study of Wanzhou
44 section of the Three Gorges Reservoir, China

45 Huijuan Zhang^{a,b}, Yingxu Song^c, Shiluo Xu^d, Yueshun He^e, Zhiwen Li^f, Xianyu Yu^g,
46 Ye Liang^h, Weicheng Wu^a and Yue Wang^b

47 ^aKey Lab of Digital Land and Resources and Faculty of Earth Sciences, East China University of Technology, Nanchang, 330013, Jiangxi, China

48 ^bSchool of Earth Sciences, East China University of Technology, Nanchang, Jiangxi Province 330013, China

49 ^cJiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, School of Information and Engineering, East China
50 University of Technology, Nanchang, 330013, Jiangxi, China; yxsong@ecut.edu.cn

51 ^dSchool of Information Engineering, Huzhou University, Huzhou 313000, China; xushiluo@163.com

52 ^eEast China University of Technology, Nanchang, 330013, Jiangxi, China; heys@ecut.edu.cn

53 ^fSchool of Environmental and Chemical Engineering, Foshan University, Foshan, 528000, China; lizw1982@163.com

54 ^gSchool of Civil Engineering, Architecture and Environment, Hubei University of Technology, Wuhan, Hubei Province 430074, China;
55 yuxianyu@hbust.edu.cn

56 ^hJiangxi Engineering Technology Research Center of Nuclear Geoscience Data Science and System, East China University of Technology,
57 Nanchang, 330013, Jiangxi, China; liangye@ecut.edu.cn

58 ^aKey Lab of Digital Land and Resources and Faculty of Earth Sciences, East China University of Technology, Nanchang, 330013, Jiangxi, China;
59 wuwch@ecut.edu.cn/wuwc030903@sina.com

60 ^bSchool of Earth Sciences, East China University of Technology, Nanchang, Jiangxi Province 330013, China; 2020210058@ecut.edu.cn

61

ARTICLE INFO

62
63 **Keywords:**
landslide susceptibility mapping
64 class-weighted algorithm
imbalanced landslide data
65 machine learning model
Three Gorges Reservoir area

ABSTRACT

This study aims to investigate the application of the class-weighted algorithm combined with traditional machine learning (logistic regression) and ensemble machine learning models (LightGBM and random forest) to the landslide susceptibility evaluation. Wanzhou section of the Three Gorges Reservoir area, China, which have numerous landslides and the number of landslide samples is 19 times more than non-landslide samples, is chosen as an example. The class-weighted algorithm focuses on the class-imbalanced problem of landslide and non-landslide samples in the assessment of landslide susceptibility and can turn the class-imbalanced issue into a cost-sensitive problem by setting unequal weights for different classes, which contribute to improving landslide susceptibility evaluation accuracy. The landslide inventory database was produced by field investigation and remote sensing images derived from Google Earth. Of the 233 landslides in the inventory, 40% were used for validation, and the remaining 60% were used for training purposes. Twelve environmental parameters (elevation, slope, aspect, curvature, distance to river, NDVI, NDWI, rainfall, seismic intensity, land use, TRL, lithology) were treated as inputs of the models to produce landslide susceptibility map (LSM). The AUC value, Balanced accuracy, and Geometric mean score were utilized to estimate the quality of models. The results showed that the weighted models (weighted logistic regression, weighted LightGBM, weighted random forest) have higher AUC values, Balanced accuracy, and Geometric mean scores than those of unweighted methods, which demonstrated that the weighted models exhibit better than unweighted methods, with the weighted random forest method having the best performance. The landslide susceptibility map of the Wanzhou section display that the high and very high landslide susceptibility are mainly distributed on both sides of the river. The insights from this research will be useful for ameliorating the landslide susceptibility mapping and the development of prevention and mitigation Wanzhou section.

88

89 CRediT authorship contribution statement

90 **Huijuan Zhang:** Conceptualization, Validation, Writing-original draft preparation, Writing-review and editing.
91 **Yingxu Song:** Conceptualization, Methodology, Software, Validation, Investigation, Resources, Funding acquisition.

ORCID(s): 0000-0002-9273-2019 (Y. Song)

92 **Shiluo Xu:** Software, Resources. **Yueshun He:** Project administration, Funding acquisition. **Zhiwen Li:** Conceptualization. **Xianyu Yu:** Resources, Funding acquisition. **Ye Liang:** Funding acquisition. **Weicheng Wu:** Writing-review
93 and editing. **Yue Wang:** Software.

95 **1. Introduction**

96 Landslide refers to a natural phenomenon in which the soil or rock mass on the slope slides downwards along the soft
97 surface under the action of gravity or other external forces. Landslide is a common geological disaster, causing many
98 economic losses and unfortunate casualties, such as devastating soil, vegetation, and dwellings, as well as critically
99 blocking transportation lines and waterways (Abuzied et al., 2016; Chen et al., 2017). The China Geological Survey
100 reported that there were 6181 geological disasters in 2019, including landslides, collapses, mudrock flows, the ground
101 collapses, ground fissures, and land subsidence, resulting in 211 deaths, 13 missings, 75 injured and direct economic
102 losses of 2.77 billion Yuan. Among them, 4020 landslides occurred, mainly distributed in Southwestern China, and
103 brought about a large number of missing persons and severe economic losses. Various factors, such as natural factors
104 (e.g., heavy rainfall, earthquake, loose lithology, and low vegetation coverage, etc.) and human-made factors (e.g.,
105 infrastructures construction and road irrigation, etc.) can trigger landslides (Wilde et al., 2018). Especially in recent
106 years, the rapid urbanization and industrialization have increased the likelihood of landslide occurrence (Kocaman
107 et al., 2020), which led to higher number of human casualties and more enormous loss of property. It is therefore of
108 significant necessity to develop landslide susceptibility map, which represents the probability of the spatial distribution
109 of landslides in a specific region based on historical landslides and related factors (Yu et al., 2016; Song et al., 2018).
110 Government agencies have attempted to take various measures to reduce the casualties and financial losses caused by
111 landslides. This process generally involves carrying out LSM, representing the probability of the spatial distribution
112 of landslides in a specific region based on historical landslides and related factors (Yu et al., 2016; Song et al., 2018).
113 Landslide susceptibility map can help government agencies to take preventable measures for reducing the casualties
114 and financial losses caused by landslides.

115 Various methods and techniques, which can be defined as qualitative or quantitative, have been implemented in the
116 landslide susceptibility assessment and have achieved notable progress (Fang et al., 2020; Guzzetti et al., 1999; Bui
117 et al., 2020). Qualitative methods are based on expert knowledge to identify the main triggering factors, determine
118 the weights of natural and human-made factors and acquire landslide susceptible zones (Aditian et al., 2018), such as
119 analytic hierarchy process (AHP) (Barredo et al., 2000; Yalcin, 2008; Feizizadeh et al., 2014)(Barredo et al., 2000;
120 Yalcin, 2008), interval pairwise comparison matrix (IPCM)(Ghorbanzadeh et al., 2019), and fuzzy logic models(Aksoy
121 and Ercanoglu, 2012; Anbalagan et al., 2015; Shahabi et al., 2015; Roy and Saha, 2019). Whereas quantitative methods
122 rely on mathematical models including the statistical and deterministic models(Abuzied et al., 2016; Reichenbach
123 et al., 2018; Fang et al., 2020). With the rapid advancement of computer technology and the improvement of remote

sensing (RS) and geographic information system (GIS) technology, the quantitative methods develop swiftly. Many studies have demonstrated that the quantitative approaches are more precise than qualitative methods because the qualitative methods have much subjectivity concerning the prediction of landslides(Aditian et al., 2018; Bui et al., 2020). Machine learning model which is one of the qualitative methods has the capability of handling non-linear data with different scales and from different type of sources(Bui et al., 2020). Different machine learning algorithms together with GIS and RS techniques have been widely applied to assess landslide susceptibility and perform well, such as LR (logistic regression), which were most widely used and often found successful in the landslide susceptibility evaluation (Ayalew and Yamagishi, 2005; Eeckhaut et al., 2006; Bai et al., 2010; Akgun, 2012; Sevgen et al., 2019; Dağ et al., 2020). Additionally, the ensemble learning methods acting as an improvement of traditional machine learning models arise and show more robust performance in many real-world tasks, widely used in landslide susceptibility evaluation (Althuwaynee et al., 2014; Napoli et al., 2020; Hong et al., 2020; Saha et al., 2021). Random forest (RF) (Breiman, 2001), which is an extended variant of the bagging method, has a simple implementation and low computational overhead (Youssef et al., 2015; Kim et al., 2017). LightGBM is a new member of the boosting ensemble models, having faster training efficiency, higher accuracy, and more robust ability to handle large-scale data (Song et al., 2018).

The choice of samples seriously affects the accuracy of the machine learning models. Some researchers have paid attention to the sample selection in the evaluation of landslide susceptibility, polygon-based random sampling (PBRs) (San, 2014), two-level random sampling (2LRS) (Ada and San, 2017; Aktas and San, 2019) were used to produce more realistic landslide susceptibility maps.

However, the area of the landslide area is often much smaller than that of the non-landslide area. Selecting the same amount of samples under different categories will often result in underrepresentation of non-landslide samples, waste of non-landslide samples and loss of important information, lead to poor performance in landslide susceptibility evaluation models.

The class-weighted algorithm treats the susceptibility assessment as a cost-sensitive issue and sets different misclassification weights for different categories (landslides, non-landslides). This method has been widely used to solve the unbalanced variety, but the application to landslide susceptibility assessment is still relatively few.

Wanzhou district of Chongqing is in the Three Gorges Reservoir area's hinterland, playing a significant role in the prevention and domination of geological disasters in the Three Gorges Reservoir area. In recent decades, because of the abundant precipitation and cyclical fluctuation of water level in the Yangtze River, landslides and other geological disasters in this area have increased significantly, seriously destroying the ecological environment and socially sustainable development. In this study, the Wanzhou section of Three Gorges Reservoir was selected as the research area, and the class-weighted algorithm combined with traditional machine learning model (Logistic regression) and ensemble machine learning models (LightGBM and random forest) were applied to the landslide susceptibility evaluation. The

purpose of this research attempts to achieve the relatively optimal method in which the impact of unbalanced landslide samples can be minimized, and the accuracy of the landslide susceptibility map is improved, providing essential introductory information for mitigating the landslide hazard by governmental subdivisions or decision-makers. Different from previous work, the novelty of this paper are 1) the class-weighted algorithm is firstly applied to landslide susceptibility mapping; 2) the advantages and disadvantages of traditional machine learning model (Logistic regression) and ensemble machine learning models (LightGBM and random forest) combined with class-weighted algorithm were compared in the Wanzhou section.

2. Study area and data used

Wanzhou District belonging to Chongqing Municipality, is in the hinterland of the Three Gorges Reservoir area. The terrain of Wanzhou District is mostly mountains and hills, with large topographic fluctuations which is largely attributed to its location at the eastern margin of East Sichuan Fold belt. Additionally, the study area is located in the Yangtze River Valley, and the floodplain landform is widely developed, forming a typical river terrace landform. The existence of river terraces and low mountain hills makes the area widely developed with various slopes, which is more conducive to the occurrence of landslide disasters. The study area with 223 historical landslides (1a) is the bank section of Wanzhou District, having many rivers and streams of the Yangtze River system (Yu et al., 2016; Song et al., 2018).

Wanzhou District is in the subtropical monsoon region with plentiful precipitation. The rainfall is mainly concentrated from May to September, which accounts for about 60% of the annual rainfall, triggering abundant landslides. The rivers and streams in Wanzhou District have deep cuts, large drops, and branch-like distribution, all of which belong to the Yangtze River system. The rivers in the territory with a drainage area of more than 100 km^2 include the Zhuxi River, Duhe River, Shiqiao River, Ruxi River, and Puli River in northern of the Yangtze River, and Nixi River, Wuqiao River and Xintian River in southern of the Yangtze River. Wanzhou District is located in the northwest edge of the Sichuan-Hubei-Hunan uplift fold belt of the first-class structure of the Neocathaysian system, mainly including Changliangzi anticline and its syncline, Yushan anticline, Qiyashan anticline and Hengshixi anticline in the East. A number of tectonic fissures are distributed in NNE or NE direction. There are Triassic, Jurassic and Quaternary strata (including alluvial deposits and slope deposits, etc.) in the study area (Song et al., 2018).

The lithology is relatively complicated, and the particles can be divided into shale and sand-mudstone interbedded, mudstone, siltstone, sandstone, red clastic rock according to the material composition. The lithology is characterized by soft and hard phases, low mechanical strength, and obvious differential weathering, which provides favorable materials for the landslides. Wanzhou District is subordinate to the weak seismic zone in southern China, and thus lacks any notable threat of earthquakes to local geo-hazards. The combination of the above natural environmental characteristics

Table 1

Data used, formats and sources

Data used	Format	Source
Historic Landslide	Shapefile	geological survey and Google Earth software
DEM	Tiff	Aster GDEM (https://earthdata.nasa.gov/)
Landsat 8 OLI	Tiff	USGS (https://earthexplorer.usgs.gov/)
Lithology	Shapefile	local Land and Resources Bureau
Meteorological data	Shapefile	Meteorological Bureau (http://www.cma.gov.cn/)

Table 2

The names, types, and class of landslide impact factors.

Variables	Name	Variable type	Class
Y	Landslide	Binary	Landslide
X1	Elevation	Continuous	Topography
X2	Slope	Continuous	Topography
X3	Aspect	Discrete	Topography
X4	Curvature	Continuous	Topography
X5	Distance to river	Continuous	Hydrology
X6	NDVI	Continuous	Land cover
X7	NDWI	Continuous	Land cover
X8	Rainfall	Discrete	Triggered
X9	Seismic intensity	Discrete	Triggered
X10	Land use	Discrete	Triggered
X11	TRI	Continuous	Topography
X12	Lithology	Continuous	Topography

and human influences (such as accelerating engineering construction and increasing population) leads to some geo-hazards in Wanzhou District, especially landslides. The landslide data mainly come from landslide geological surveys and the remote sensing images provided by Google Earth. The DEM data with 30×30 m resolution derived from Aster GDEM. A Landsat-8 satellite image which was acquired on 2013-08-12 were utilized as primary remote sensing data. Table 1 shows the formats and sources of data in this study.

A total of 12 landslide contributing factors and the types of data were shown in Table 2.

Figure 3 shows the distributions of twelve landslide factors. Elevation, slope, aspect, curvature, and topographic roughness index (TRI) were derived from the DEM data using the ArcGIS and QGIS. The lithological data and the distance to the river were vectorized from the geological and topographic maps. The NDVI/NDWI data were acquired from the Landslide 8 OLI images. The rainfall data were provided by the Meteorological Bureau. The land-use data came from the geological survey and the Landslide 8 OLI images.

3. Methodology

The flowchart of landslide susceptibility mapping for the study area is shown as in Fig. 4.

Firstly, twelve landslide contributing factors and landslide samples were selected as independent variables and de-

201 pendent variables, respectively, to form an initial decision table for training the models. Not all the landslide contribut-
 202 ing factors are indispensable for the landslide susceptibility assessment (Dou et al., 2015). Therefore, multicollinearity
 203 analysis of landslide contributing factors is essential for improving the robustness of the models. In this study, the vari-
 204 ance inflation factor method (VIF) was utilized to carry out multicollinearity analysis of landslide conditioning factors.
 205 Secondly, a so-called "Pipeline" strategy was used to connect data processing and classifiers. The disposing of data
 206 includes factor-normalization and factor-reduction in which the StandardScaler function and PCA method provided by
 207 Sklearn were implemented (Pedregosa et al., 2011). The purpose of employing "Pipeline" is to ensure the consistency
 208 of the data preprocessing in the training set and test set. Thirdly, the traditional machine learning (logistic regression)
 209 and ensemble machine learning models (LightGBM and random forest) were applied to achieve the landslide suscepti-
 210 bility mapping. Finally, several evaluation indicators (e.g., AUC value, balanced accuracy, and geometric mean score)
 211 were implemented to evaluate the LSM models.

212 3.1. Logistic Regression (LR)

213 Logistic regression (LR) is a classic machine learning model with the capacity to settle classification problems
 214 (Ayalew and Yamagishi, 2005; Bai et al., 2010; Song et al., 2018). It is widely used in landslide susceptibility evaluation
 215 because of its simplicity, parallelization, and strong interpretability. Logistic regression can be treated as a variant of
 216 linear regression, and the variables of the LR model could be continuous or discrete (Ayalew and Yamagishi, 2005;
 217 Bai et al., 2010). The core concept of logistic regression is to map the domain's value from $(-\infty, +\infty)$ to $(0,1)$. 0
 218 and 1 represent different categories, respectively. They represent non-landslides (0) and landslides (1) in the landslide
 219 susceptibility evaluation. A Sigmoid function is employed to express this mapping relationship, as shown below
 220 (Equation 1).

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

221 3.2. LightGBM

222 LightGBM is a new gradient boosting framework proposed by Microsoft (Friedman, 2002). LightGBM belongs to
 223 the Boosting family in ensemble learning and relies on decision tree algorithms. LightGBM is widely used for classi-
 224 fication tasks and machine learning competitions because of its higher efficiency and lower memory usage than other
 225 gradient boosting frameworks (e.g., Adaboost, GBDT, etc.). The application of LightGBM addresses the problems
 226 encountered by GBDT in massive data and en-sures the better performance of GBDT in industrial practice.

227 **3.3. Random Forests (RF)**

228 The RF method belongs to the Bootstrap aggregation, a basic ensemble learning model (Breiman, 2001). Random
 229 forests have a simple implementation, low computational overhead, and robust performance in many machine learning
 230 tasks. The diversity of Bagging basic learners comes from sample perturbations and attributes perturbations, further
 231 improving the generalization performance of the final integration (Youssef et al., 2015).

232 **3.4. Class-weighted machine learning models**

233 When the samples of landslide and non-landslide are equal or similar, the machine learning will have excellent
 234 performance. Otherwise, the process of machine learning will be seriously affected by imbalanced samples. The
 235 imbalance of categories may cause the predictive results to be biased towards the side with more sample categories:
 236 the non-landslide area. If the landslide area is predicted as a non-landslide area, the accuracy and practicability of
 237 the landslide sensitivity evaluation result will be low. For example, there are 98 negative examples (non-landslides)
 238 but only 2 positive examples (landslides). The learning model only requires returning a learner that always predicts
 239 new samples as negative examples, which can achieve 98% accuracy. However, such learners are worthless because
 240 they cannot predict any positive cases. The class-imbalanced problem can be solved by oversampling positive samples
 241 (landslides), undersampling negative samples (assuming the non-landslide is the majority class) or treating the machine
 242 learning process as a cost-sensitive learning problem. The representative oversampling methods are the SMOTE and
 243 Borderline-SMOTE, while the representative undersampling technique is the EasyEnsemble method (Verbiest et al.,
 244 2014). The oversampling method's time overhead is usually more than that of the undersampling method because the
 245 former method adds many positive examples and makes the classifier training set much larger than the initial training
 246 set. Moreover, the oversampling method cannot simply repeat the initial sampling of the initial positive samples,
 247 leading to serious overfitting. Although the undersampling method can reduce time overhead by randomly discarding
 248 the negative examples, some critical information might be lost during this process. When viewed as a cost-sensitive
 249 issue, the class-imbalanced problem could be well solved because a so-called cost matrix used in the machine learning
 250 process can set the weights corresponding to different categories for improving the accuracy of classification. The
 251 class-weighted machine learning methods used in this article belong to this category. In this study, the entire study
 252 area was resampled into 553,172 non-landslide samples and 29,313 landslide samples. The ratio of non-landslide
 253 samples to landslide samples was approximately 19:1. Therefore, the LSM process in this study should be regarded as
 254 a typical class-imbalanced problem. Table 3 shows the cost matrix used in this study.

255 The reasons for choosing 1:19 are as follows: First, the ratio of landslide samples to non-landslide samples in the
 256 study area is approximately 1:19, and the choice of cost matrix is usually related to the sample ratio. Secondly, set the
 257 misclassification cost of landslide samples to 1, and increase the misclassification cost of non-landslide samples from

Table 3

Cost matrix used in this study.

True Label	Predicted Label	
	Non-landslide	Landslide
Non-landslide	0	1
Landslide	19	0

258 1 to 30. Use G-mean as the evaluation index to evaluate the three weighted models, and the best misclassification cost
 259 is 19 for WLighGBM (Figure 5).

260 **3.5. Model elevation**

261 **3.5.1. Confusion matrix and ROC curve**

262 The confusion matrix is comprised of the following four indexes: true positive (TP), false positive (FP), true
 263 negative (TN), and false negative (FN). Various statistical indicators, including accuracy (Equation 2), TPR/recall
 264 (Equation 3), TNR (Equation 4), ROC curve (Receiver Operating Characteristic), and AUC (area under ROC curve),
 265 could be calculated through the above four indexes. These indicators are usually employed to evaluate the performance
 266 of machine learning tasks, consisting of land-use classification (Jr and Si, 2014), LSM, etc.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$TPR = \text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$TNR = \text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

267 **3.5.2. Balanced accuracy and G-mean score**

268 In the cost sensitivity problem, the ROC curve cannot directly reflect the models' pros and cons. Thus, we used
 269 balanced accuracy and G-mean score provided by Sklearn Pedregosa et al. (2011) as the evaluation indexes. The
 270 balanced accuracy (Equation 5) in classification problems is defined as the average recall (TPR) obtained under each

Table 4

Multicollinearity analysis, tolerance and variance inflation factor (VIF).

Variables	Name	Collinearity Statistics	
		Variance Inflation Factors (VIF)	Tolerance
X1	Elevation	2.405	0.416
X2	Slope	9.667	0.103
X3	Aspect	1.029	0.972
X4	Curvature	1.016	0.985
X5	Distance to river	1.917	0.522
X6	NDVI	2.168	0.461
X7	NDWI	1.432	0.698
X8	Rainfall	1.446	0.691
X9	Seismic intensity	1.691	0.591
X10	Land use	1.634	0.612
X11	TRI	9.39	0.106
X12	Lithology	1.636	0.611

271 class, and the G-mean (Equation 6) is the root of the product of TPR and TNR.

$$\text{Balanced Accuracy} = \frac{\text{TPR} + \text{TNR}}{2} \quad (5)$$

$$G - \text{mean} = \sqrt{\text{TPR} * \text{TNR}} \quad (6)$$

272

4. Results and discussions

273

4.1. Multicollinearity Analysis of Landslide Factors

274 It is of great significance to employ multicollinearity analysis before landslide susceptibility modeling. Identifying
 275 and selecting appropriate landslide factors is the prerequisite for ensuring the robustness of these models. In this study,
 276 the variance inflation factor (VIF) was utilized to develop the multicollinearity analysis with the Python programming
 277 language (Table 4).

278 If the value of VIF exceeds 10, meaning that there are multiple collinearities among variables. Results display that
 279 all the VIF values of the twelve factors are less than 10, denoting that all the 12 landslide-related factors are appropriate
 280 for LSM.

281

4.2. Landslide susceptibility mapping results

282 LR, LightGBM, RF models, and their weighted models (WLR, WLighGB, WRF) are utilized for landslide sus-
 283 ceptibility mapping. Twelve landslide contributing factors: elevation, slope, aspect, curvature, distance to the river,

284 NDVI, NDWI, rainfall, seismic intensity, land use, and topographic roughness index (TRI), and lithology were used
285 as the input of these six models. The probability values of the six models range from 0 to 1, which are the so-called
286 landslide prediction index values (LPI). The LPI values generated by six models were reclassified to develop the land-
287 slide susceptibility map with the Natural Breaks method and the ArcGIS software. The landslide susceptibility maps
288 (LR & WLR, LightGBM & WLightGBM, RF & WRF) derived from the six models are shown in Figure 6 af. These
289 landslide susceptibility maps (LSMs) are classified into very low, low, medium, high, and very high susceptibility to
290 landslides.

291 The percentages of each category in the six models are illustrated in Figure 6. In the LR case, the five landslide
292 susceptibility classes of very low, low, medium, high, and very high covered 41.74%, 31.55%, 15.44%, 8.57%, and
293 2.70% area of the districts, respectively. In the LightGBM and RF case, the class of very low area is much higher
294 than those in LR case, while the class of low area is lower than those in LR case, and the classes of medium, high,
295 and very high regions are almost the same as those in LR case. The percentages of very low and low classes in LR,
296 LightGBM, and RF cases are higher than those in weighted models, but the percentage of very high and high areas in
297 LR, LightGBM, and RF cases are lower than those in weighted models.

298 4.3. Implications for landslide-prone Areas

299 The regions with the high and very high landslide susceptibility are mainly distributed on both sides of the river
300 (Figure 6), most likely related to the water level. Wanzhou reservoir area is the hinterland of the Three Gorges Reservoir
301 area with the frequently variable water level. The rising water level of the Yangtze River can lead to the decrease of
302 shear strength of the sliding body through softening and silting the slope (Wang and ping Qiao, 2013; Gui et al.,
303 2016). In contrast, the drop in the water level produces a much larger hydrodynamic pressure, which increases the
304 sliding force along the direction of underground seepage and then brings about the landslides (Wang and ping Qiao,
305 2013; Gui et al., 2016). There is the highest landslide susceptibility at the middle and lower reaches of the river
306 (Figure 6). In addition to lithology, rainfall, and vegetation, the type of land-use is also probably to account for this
307 characteristic. The strata exposed in the Wanzhou reservoir area are mainly Jurassic Shaximiao Formation (J2s) and
308 Suining Formation (J3s) (Zhu et al., 2014). The lithology is off-white feldspathic quartz sand-stone intercalated with
309 purplish-red argillaceous siltstone, purplish-red sandstone, and mudstone. It is easy to form a soft top and hard bottom
310 structural surface because of the difference in weathering speed of mudstone and sandstone, providing an effective
311 structure for the loose accumulation material sliding along the bedrock surface. Wanzhou District is the center of
312 a rainstorm in eastern Chongqing. According to the Datankou hydrological station's statistics, the average annual
313 precipitation is 1243 mm, and the maximum annual rainfall is about 1550 mm (Yu et al., 2016; Song et al., 2018).
314 The rainstorm strongly scours the landslide soil, infiltrate into cracks and potential sliding surfaces, resulting in the

Table 5

Balanced accuracy, G-mean, Recall, Accuracy, AUC of the models.

	Balanced Accuracy	G-mean	Recall	Accuracy	AUC
LR	0.500	0.000	0.000	0.950	0.835
WLR	0.776	0.774	0.819	0.736	0.839
LightGBM	0.550	0.321	0.104	0.952	0.890
WLIGHTGBM	0.844	0.842	0.900	0.793	0.888
RF	0.511	0.150	0.023	0.950	0.844
WRF	0.823	0.821	0.880	0.772	0.913

315 aggravation of landslide deformation. On the other hand, the rainfall will increase the slope's self-weight, thereby
 316 increasing the sliding force of the hill. Therefore, the combination of pore water pressure and soil softening can
 317 increase the probability of landslides (Finlay et al., 1997; Dahal et al., 2007). The plant roots have a powerful tensile
 318 effect on improving the anti-sliding ability of rock and soil, which anchor the loose weathered layer to the more stable
 319 rock and soil layer to prevent them from sliding along the slope. The plant stems and leaves, and litters can intercept
 320 and absorbing rainwater, which plays an inhibitory role in slope runoff and rain erosion (Sittadewi and Tejakusuma,
 321 2019). However, the vegetation coverage of the research area is low, having a weak ability to resist landslides. The
 322 primary type of land-use in this area is wetland filled with groundwater, which is one of the significant external factors
 323 inducing landslide. Groundwater will sharply increase the weight of the rock and soil and reduce the anti-sliding
 324 resistance, which leads to the increase of sliding force and slope instability, resulting in landslides. Hence, LSM can
 325 be applied to land-use planning and in the prioritizing the management of countermeasures to mitigate potential losses
 326 by landslides and also helps the government formulate relevant scientific policies according to different susceptibility
 327 levels as a means of mitigating landslides. Moreover, a LSM could also be used to raise public awareness of landslides
 328 and then reduce related activities in hazardous areas.

329 5. Validation of landslide susceptibility maps

330 The ROC curves of the six models are shown in Figure 7. The AUC values of the six models are 83.5%, 83.9%,
 331 89%, 88.8%, 84.4%and 91.3%, respectively, indicating that all the six models are suitable for LSM in this region.
 332 Based on the ROC curves results, the weighted methods are generally better than the unweighted methods (except the
 333 LightGBM/WLightGBM), and the WRF model with the highest AUC value (AUC = 91.3%) is probably considered
 334 to be the most appropriate model. The ROC curve cannot evaluate the models' performance perfectly because it
 335 cannot directly reflect the overall cost expectation of the models in case of unequal costs. Furthermore, the model's
 336 ability to predict landslides should be emphasized rather than non-landslides in the landslide susceptibility evaluation.
 337 Therefore, we selected more appropriate evaluation indicators to compare the pros and cons of the models. Table 5
 338 shows the Balanced accuracy, G-mean, Recall, Accuracy, and AUC of the six models.

The Recall values of the six models are 0.000, 0.774, 0.321, 0.842, 0.150 and 0.821, respectively. The Recall value of the LR model is 0, meaning that it cannot predict landslides. The weighted models (WLR, WLighGBM, WRF) are better than the unweighted models (LR, LightGBM, RF) in terms of Recall, suggesting that the weighted models have a more powerful ability to predict landslides. The six models have distinctive Accuracy values, with the figures of 0.950, 0.736, 0.952, 0.793, 0.950 and 0.772, respectively. The weighted models (WLR, WLighGBM, WRF) are worse than the unweighted models (LR, LightGBM, RF) in terms of Accuracy values, denoting that the unweighted models have the stronger ability to predict non-landslides. The G-mean values and Balanced accuracy values of the six models are 0.000, 0.774, 0.321, 0.842, 0.150, 0.821 and 0.500, 0.776, 0.550, 0.844, 0.511, 0.823, respectively. The G-mean and Balanced accuracy values imply that the weighted models are better than the unweighted models in LSM when a class-imbalanced problem is viewed as a cost-sensitive issue. In line with the AUC results, the Balanced accuracy and G-mean scores indicate that the WRF model has achieved much better performance than the other weighted models. Landslide events not only reduce the financial losses but also cost human lives. A landslide susceptibility map is an essential tool for developing preventive measures in landslide-prone areas. Therefore, many scholars are committed to improving LSM models' performance. Recently, machine learning models and ensemble machine learning models had good performance in LSM. However, few studies have focused on the class-imbalanced problem, which will lead to poor performance in LSM whether the machine learning or ensemble machine learning models are utilized. Thus, we carried out the application of the class-weighted algorithm combined with traditional machine learning (LR) and ensemble machine learning models (LightGBM and RF) to the LSM based on a case study of the Wanzhou section of the Three Gorges Reservoir, China, in the present study. The results proved that the weighted methods (WLR, WLighGBM, WRF) are better than unweighted methods (LR, LightGBM, RF), shown as higher AUC, G-mean, and Balanced Accuracy values generally. Moreover, the WRF model has much better performance than WLR and WLighGBM models. Although the unweighted models have higher Accuracy value, they are incapable of evaluating landslide susceptibility because their accuracy rates come from the prediction of the negative class (non-landslides) rather than the positive class (landslides). A vital advantage of the weighted models is that the class-weighted algorithm turned the susceptibility evaluation problem into a cost-sensitive issue by setting unequal weights for different classes, which improves the performance of LSM, manifesting in higher Recall values. On the other hand, the weighted models (WLR/WLighGBM/WRF) tend to divide more high and very high susceptibility areas than the unweighted models (LR/LightGBM/RF) (Fig 5, 6). Landslide susceptibility map is the basis of landslide risk evaluation. Suppose the high susceptibility area is incorrectly classified as a low susceptibility zone, which may lead to a false judgment on the risk of landslides and then result in considerable threats to the safety of human life and property. Furthermore, the weighted models pay more attention to landslide samples' classification accuracy, which is the actual concern in the landslide susceptibility evaluation. Although every study area has its own unique landslide contributing

371 factors and geological conditions, the weighted models proposed in this paper will provide significant clues for the
372 landslide susceptibility evaluation concerning the imbalanced landslide samples. Regardless, the weighted models
373 still have several disadvantages. For instance, the cost matrix should be processed before classification using weighted
374 models, which is affected by the processing method and is time-consuming. Moreover, a high-resolution DEM for the
375 study area is not freely available, resulting in the poor performance of weighted models. If high-resolution DEM were
376 utilized for extracting landslide-related parameters, these weighted models could achieve better results.

377 6. Conclusions

378 In the present study, the class-weighted algorithm combined with traditional machine learning (logistic regres-
379 sion) and ensemble machine learning models (LightGBM and random forest) was utilized to improve the accuracy
380 of the LSM models disturbed by the imbalanced landslide samples based on a case study of Wanzhou section of
381 the Three Gorges Reservoir, China. The result demonstrated that the weighted models (weighted logistic regression,
382 weighted LightGBM, weighted random forest) performed better than unweighted models (logistic regression, Light-
383 GBM, weighted random forest), achieving higher AUC, G-mean, and Balanced accuracy values, with the weighted
384 random forest model has a much better performance. The class-weighted algorithm turned the susceptibility evaluation
385 problem into a cost-sensitive issue by setting unequal weights for different classes, which improves the accuracy of
386 the landslide susceptibility evaluation. The weighted models (especially weighted random forest) are probably to be
387 applied to solve the class-imbalanced problem of the landslide susceptibility evaluation in other areas for regarding the
388 harm resulted from landslides.

389 7. Acknowledgments

390 The authors would like to acknowledge Prof. Chong Xu for helpful discussions. This research was funded by
391 Project Digital frequency spectrum analysis and mineralization precise prediction for continental su-pergene U-Re
392 (No. 41872243), East China University of Technology Doctoral Research Startup Fund (No. DHBK2019218), Jiangxi
393 Provincial Nuclear and Geoscience Data Science and System Engineering Technology Research Center (No.JETRCNGDSS20200
394 National Natural Science Foundation of China (No. 41807297), and Jiangxi Engineering Laboratory on Radioactive
395 Geoscience and Big Data Technology (No. JELRGBDT202004).

396 **Code availability section**

397 ArcGIS 10.8 and QGIS 3.16 were used to extract landslide factors, visualize landslide factors and export result
 398 maps.

399 The source codes are available for downloading at the link: <https://github.com/songyingxu/LspModelsForCageo>

400 **References**

- 401 Abuzied, S., Ibrahim, S., Kaiser, M., Saleem, T., 2016. Geospatial susceptibility mapping of earthquake-induced landslides in nuweiba area, gulf
 402 of aqaba, egypt. *Journal of Mountain Science* 13, 1286–1303.
- 403 Ada, M., San, B.T., 2017. Comparison of machine-learning techniques for landslide susceptibility mapping using two-level random sampling (2lrs)
 404 in alakir catchment area, antalya, turkey. *Natural Hazards* 90, 237–263. doi:10.1007/s11069-017-3043-8.
- 405 Aditian, A., Kubota, T., Shinohara, Y., 2018. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and
 406 artificial neural network in a tertiary region of ambon, indonesia. *Geomorphology* 318, 101–111. doi:10.1016/j.geomorph.2018.06.006.
- 407 Akgun, A., 2012. A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio
 408 methods: a case study at zmir, turkey. *Landslides* 9, 93–106.
- 409 Aksoy, B., Ercanoglu, M., 2012. Landslide identification and classification by object-based image analysis and fuzzy logic: An example from the
 410 azdavay region (kastamonu, turkey). *Computers & Geosciences* 38, 87–98. doi:10.1016/j.cageo.2011.05.010.
- 411 Aktas, H., San, B.T., 2019. Landslide susceptibility mapping using an automatic sampling algorithm based on two level random sampling. *Computers
 412 & Geosciences* 133, 104329. doi:10.1016/j.cageo.2019.104329.
- 413 Althuwaynee, O.F., Pradhan, B., Park, H.J., Lee, J.H., 2014. A novel ensemble decision tree-based CHi-squared automatic interaction detec-
 414 tion (CHAID) and multivariate logistic regression models in landslide susceptibility mapping. *Landslides* 11, 1063–1078. doi:10.1007/
 415 s10346-014-0466-0.
- 416 Anbalagan, R., Kumar, R., Lakshmanan, K., Parida, S., Neethu, S., 2015. Landslide hazard zonation mapping using frequency ratio and fuzzy logic
 417 approach, a case study of lachung valley, sikkim. *Geoenvironmental Disasters* 2. doi:10.1186/s40677-014-0009-y.
- 418 Ayalew, L., Yamagishi, H., 2005. The application of gis-based logistic regression for landslide susceptibility mapping in the kakuda-yahiko moun-
 419 tains, central japan. *Geomorphology* 65, 15–31.
- 420 Bai, S.B., Wang, J., Lü, G.N., Zhou, P.G., Hou, S.S., Xu, S.N., 2010. GIS-based logistic regression for landslide susceptibility mapping of the
 421 zhongxian segment in the three gorges area, china. *Geomorphology* 115, 23–31. doi:10.1016/j.geomorph.2009.09.025.
- 422 Barredo, J., Benavides, A., Hervás, J., van Westen, C.J., 2000. Comparing heuristic landslide hazard assessment techniques using GIS in the
 423 tirajana basin, gran canaria island, spain. *International Journal of Applied Earth Observation and Geoinformation* 2, 9–23. doi:10.1016/
 424 s0303-2434(00)85022-9.
- 425 Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32. URL: <https://doi.org/10.1023/A:1010933404324>, doi:10.1023/A:
 426 1010933404324.
- 427 Bui, D.T., Tsangaratos, P., Nguyen, V.T., Liem, N.V., Trinh, P.T., 2020. Comparing the prediction performance of a deep learning neural network
 428 model with conventional machine learning models in landslide susceptibility assessment. *CATENA* 188, 104426. doi:10.1016/j.catena.
 429 2019.104426.
- 430 Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z., Ma, J., 2017. A comparative study of logistic model tree, random forest,
 431 and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* 151, 147–160.

- 432 Dağ, S., Akgün, A., Kaya, A., Alemdağ, S., Bostancı, H.T., 2020. Medium scale earthflow susceptibility modelling by remote sensing and geo-
433 graphical information systems based multivariate statistics approach: an example from northeastern turkey. Environmental Earth Sciences 79.
434 doi:10.1007/s12665-020-09217-7.
- 435 Dahal, R.K., Hasegawa, S., Nonomura, A., Yamanaka, M., Masuda, T., Nishino, K., 2007. GIS-based weights-of-evidence modelling of
436 rainfall-induced landslides in small catchments for landslide susceptibility mapping. Environmental Geology 54, 311–324. doi:10.1007/
437 s00254-007-0818-3.
- 438 Eeckhaut, M.V.D., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., Vandekerckhove, L., 2006. Prediction of landslide susceptibility using
439 rare events logistic regression: A case-study in the flemish Ardennes (belgium). Geomorphology 76, 392–410. doi:10.1016/j.geomorph.
440 2005.12.003.
- 441 Fang, Z., Wang, Y., Peng, L., Hong, H., 2020. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility
442 mapping. International Journal of Geographical Information Science 35, 321–347. doi:10.1080/13658816.2020.1808897.
- 443 Finlay, P.J., Fell, R., Maguire, P.K., 1997. The relationship between the probability of landslide occurrence and rainfall. Canadian Geotechnical
444 Journal 34, 811–824. doi:10.1139/t97-047.
- 445 Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38, 367–378.
- 446 Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S., Tiede, D., Aryal, J., 2019. Evaluation of different machine learning methods and
447 deep-learning convolutional neural networks for landslide detection. Remote Sensing 11, 196. doi:10.3390/rs11020196.
- 448 Gui, L., Yin, K., Glade, T., 2016. Landslide displacement analysis based on fractal theory, in wanzhou district, three gorges reservoir, china.
449 Geomatics, Natural Hazards and Risk 7, 1707–1725. doi:10.1080/19475705.2015.1137241.
- 450 Guzzetti, F., Carrara, A., Cardinali, M., Reichenbach, P., 1999. Landslide hazard evaluation: a review of current techniques and their application in
451 a multi-scale study, central italy. Geomorphology 31, 181–216.
- 452 Hong, H., Liu, J., Zhu, A.X., 2020. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes
453 with the bagging ensemble. Science of The Total Environment 718, 137231. doi:10.1016/j.scitotenv.2020.137231.
- 454 Jr, R.G.P., Si, K., 2014. The total operating characteristic to measure diagnostic ability for multiple thresholds. International Journal of Geographical
455 Information Science 28, 570–583.
- 456 Kim, J.C., Lee, S., Jung, H.S., Lee, S., 2017. Landslide susceptibility mapping using random forest and boosted tree models in pyeong-chang, korea.
457 Geocarto International 33, 1000–1015. doi:10.1080/10106049.2017.1323964.
- 458 Kocaman, S., Tavus, B., Nefeslioglu, H.A., Karakas, G., Gokceoglu, C., 2020. Evaluation of floods and landslides triggered by a meteorological
459 catastrophe (ordu, turkey, august 2018) using optical and radar data. Geofluids 2020, 1–18. doi:10.1155/2020/8830661.
- 460 Napoli, M.D., Carotenuto, F., Cevasco, A., Confuorto, P., Martire, D.D., Firpo, M., Pepe, G., Raso, E., Calcaterra, D., 2020. Machine
461 learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. Landslides 17, 1897–1914. doi:10.1007/
462 s10346-020-01392-9.
- 463 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas,
464 J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine
465 Learning Research 12, 2825–2830.
- 466 Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth-
467 Science Reviews 180, 60–91. doi:10.1016/j.earscirev.2018.03.001.
- 468 Roy, J., Saha, D.S., 2019. GIS-based gully erosion susceptibility evaluation using frequency ratio, cosine amplitude and logistic regression ensembled
469 with fuzzy logic in hinglo river basin, india. Remote Sensing Applications: Society and Environment 15, 100247. doi:10.1016/j.rsase.

- 470 2019.100247.
- 471 Saha, S., Arabameri, A., Saha, A., Blaschke, T., Ngo, P.T.T., Nhu, V.H., Band, S.S., 2021. Prediction of landslide susceptibility in rudraprayag, india
472 using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. *Science of The Total Environment*
473 764, 142928. doi:10.1016/j.scitotenv.2020.142928.
- 474 San, B.T., 2014. An evaluation of SVM using polygon-based random sampling in landslide susceptibility mapping: The candir catchment area
475 (western antalya, turkey). *International Journal of Applied Earth Observation and Geoinformation* 26, 399–412. doi:10.1016/j.jag.2013.
476 09.010.
- 477 Sevgen, Kocaman, Nefeslioglu, Gokceoglu, 2019. A novel performance assessment approach using photogrammetric techniques for landslide
478 susceptibility mapping with logistic regression, ANN and random forest. *Sensors* 19, 3940. doi:10.3390/s19183940.
- 479 Shahabi, H., Hashim, M., Ahmad, B.B., 2015. Remote sensing and gis-based landslide susceptibility mapping using frequency ratio, logistic
480 regression, and fuzzy logic methods at the central zab basin, iran. *Environmental Earth Sciences* 73, 1–22.
- 481 Sittadewi, E.H., Tejakusuma, I.G., 2019. THE ROLE OF ROOT PLANT ARCHITECTURE IN LANDSLIDE AND EROSION DISASTER
482 MITIGATION. *Jurnal Sains dan Teknologi Mitigasi Bencana* 14, 54–61. doi:10.29122/jstmb.v14i1.3552.
- 483 Song, Y., Niu, R., Xu, S., Ye, R., Peng, L., Guo, T., Li, S., Chen, T., 2018. Landslide susceptibility mapping based on weighted gradient boosting
484 decision tree in wanzhou section of the three gorges reservoir area (china). *ISPRS International Journal of Geo-Information* 8, 4. doi:10.3390/
485 ijgi8010004.
- 486 Wang, M., ping Qiao, J., 2013. Reservoir-landslide hazard assessment based on GIS: A case study in wanzhou section of the three gorges reservoir.
487 *Journal of Mountain Science* 10, 1085–1096. doi:10.1007/s11629-013-2498-7.
- 488 Wilde, M., Günther, A., Reichenbach, P., Malet, J.P., Hervás, J., 2018. Pan-european landslide susceptibility mapping: ELSUS version 2. *Journal*
489 *of Maps* 14, 97–104. doi:10.1080/17445647.2018.1432511.
- 490 Yalcin, A., 2008. Gis-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in ardesen (turkey): Com-
491 parisons of results and confirmations. *Catena* 72, 1–12.
- 492 Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., Al-Katheeri, M.M., 2015. Landslide susceptibility mapping using random forest, boosted
493 regression tree, classification and regression tree, and general linear models and comparison of their performance at wadi tayyah basin, asir
494 region, saudi arabia. *Landslides* 13, 839–856. doi:10.1007/s10346-015-0614-1.
- 495 Yu, X., Wang, Y., Niu, R., Hu, Y., 2016. A combination of geographically weighted regression, particle swarm optimization and support vector
496 machine for landslide susceptibility mapping: A case study at wanzhou in the three gorges area, china. *Int J Environ Res Public Health* 13, 487.
- 497 Zhu, A.X., Wang, R., Qiao, J., Qin, C.Z., Chen, Y., Liu, J., Du, F., Lin, Y., Zhu, T., 2014. An expert knowledge-based approach to landslide
498 susceptibility mapping using gis and fuzzy logic. *Geomorphology* 214, 128–138.

499 List of Figures

500 1	Location of the study area. (a) Elevation of the study area. (b) Three Gorges Reservoir area. (c)	18
501	Wanzhou District, the image is a Landsat 8 image with true color (R:band 4; G: band 3; B:band 2).	
502 2	The geological and tectonic sketch of the study area.	19
503 3	Landslide factors used in the study.(a) Elevation.(b) Slope.(c) Aspect.(d) Curvature.(e) Distance to the	
504	river (m).(f) NDVI.(g) NDWI.(h) Rainfall. (i) Seismic intensity. (j) Landuse. (k) TRI. (l) Lithology.	21
505 4	Overall workflow of this study.	22
506 5	Misclassification weights of the WLighGBM models.	23
507 6	LSM results of 6 models. (a) LSM using LR model. (b) LSM using WLR model. (c) LSM using	
508	LightGBM model. (d) LSM using WLighGBM model. (e) LSM using RF model. (f) LSM using	
509	WRF model.	24
510 7	Distribution ratio of different landslide susceptibility classes for 6 models.	25
511 8	The ROC curve of 6 LSM models.	26

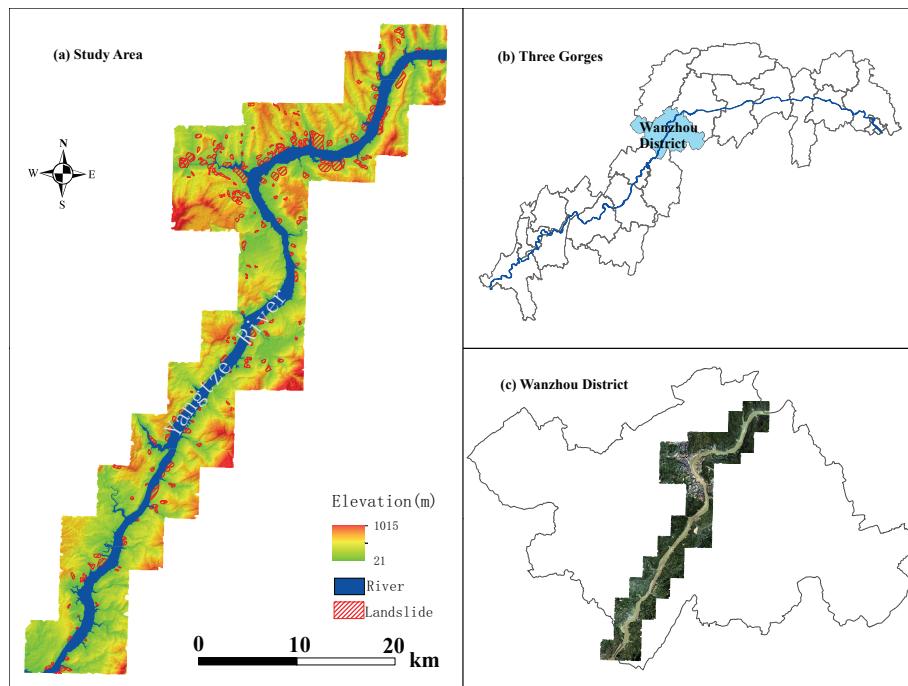


Figure 1: Location of the study area. (a) Elevation of the study area. (b) Three Gorges Reservoir area. (c) Wanzhou District, the image is a Landsat 8 image with true color (R:band 4; G: band 3; B:band 2).

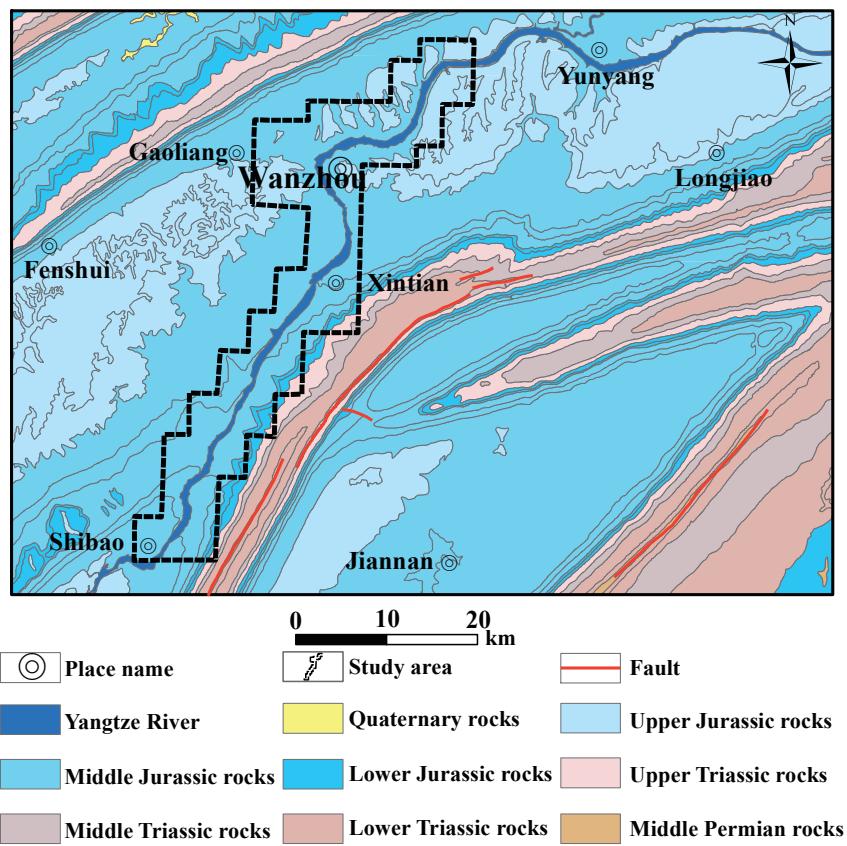
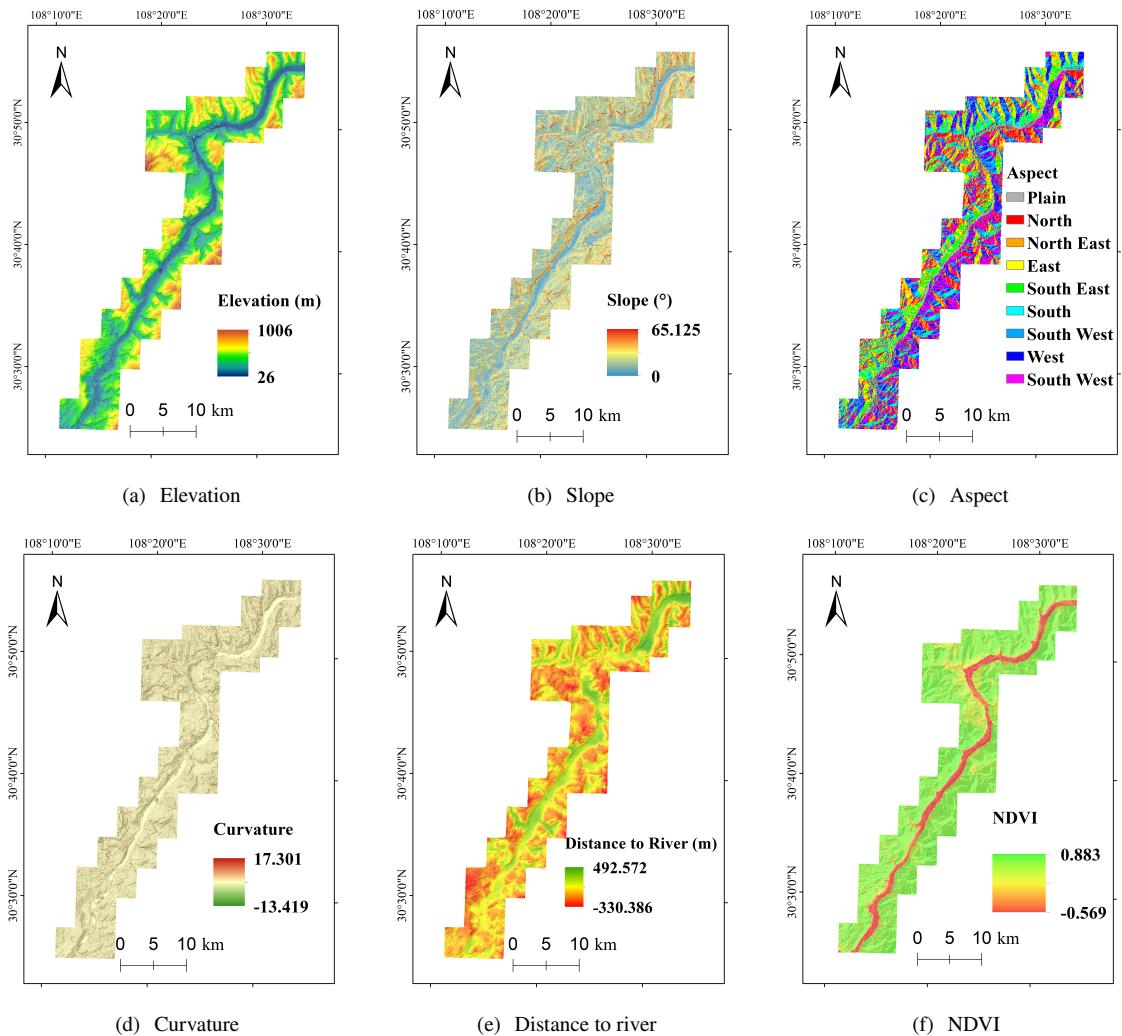


Figure 2: The geological and tectonic sketch of the study area.



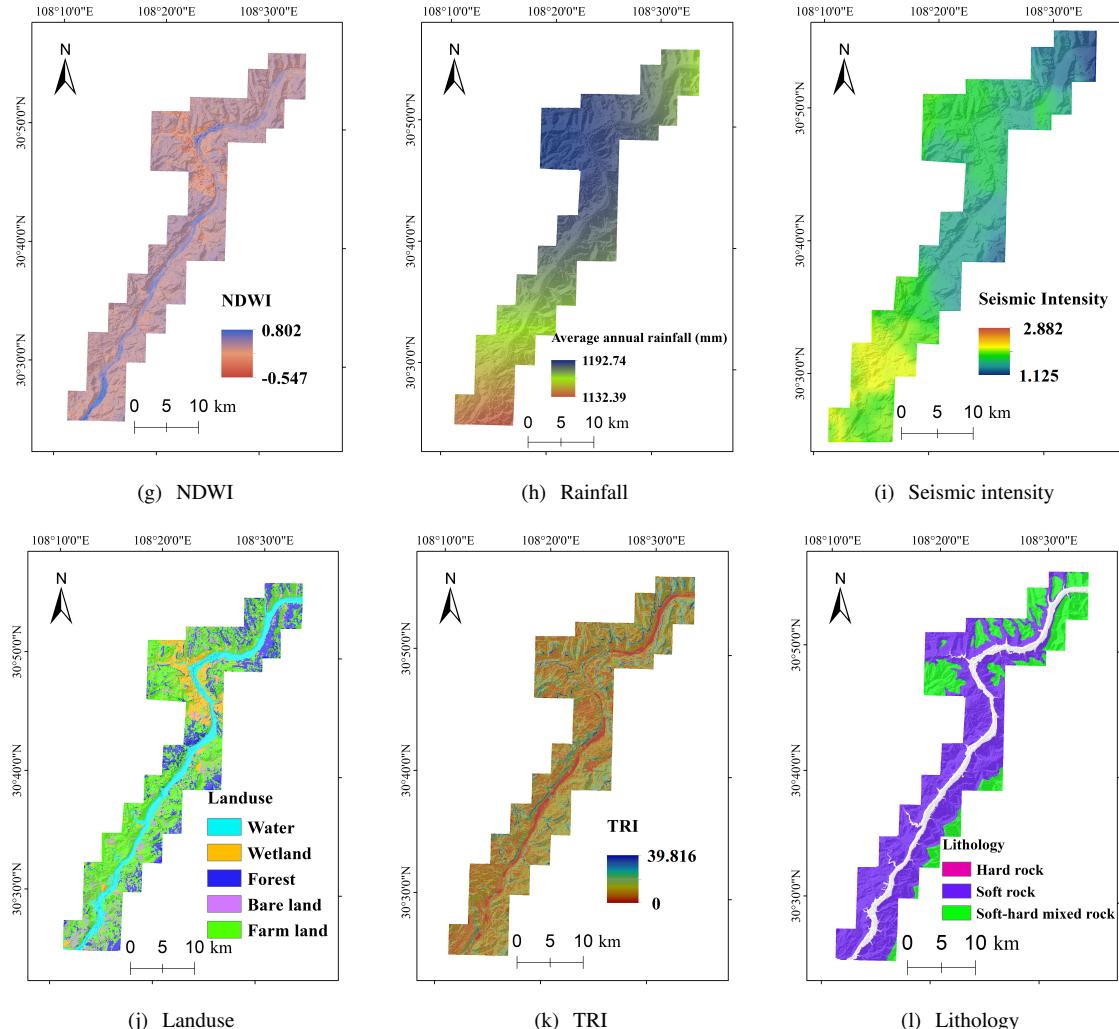


Figure 3: Landslide factors used in the study. (a) Elevation. (b) Slope. (c) Aspect. (d) Curvature. (e) Distance to the river (m). (f) NDVI. (g) NDWI. (h) Rainfall. (i) Seismic intensity. (j) Landuse. (k) TRI. (l) Lithology.

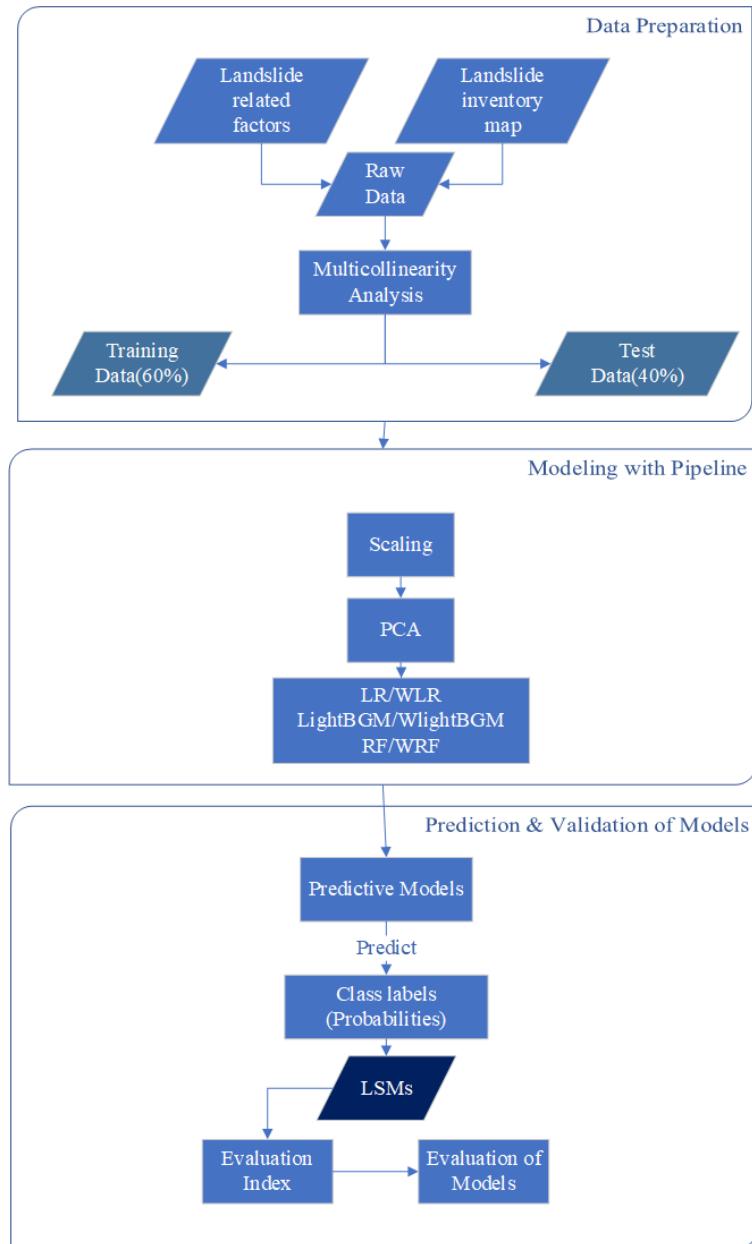


Figure 4: Overall workflow of this study.

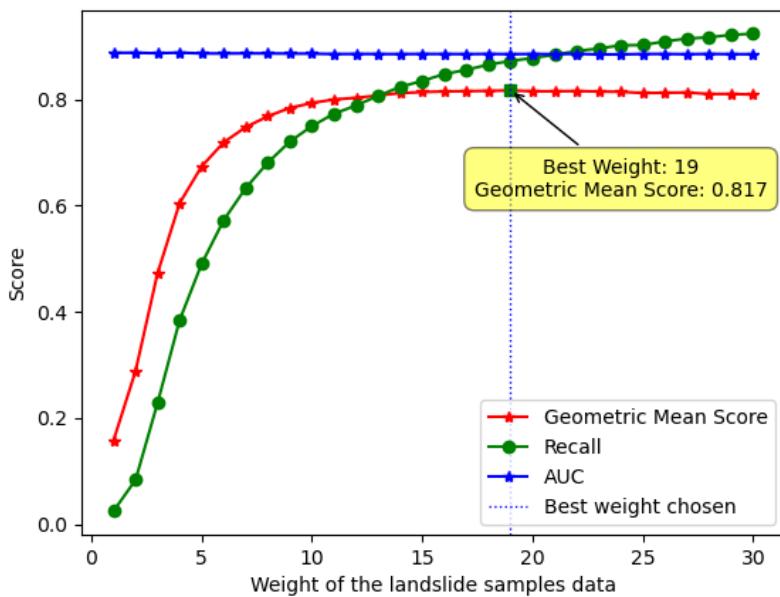


Figure 5: Misclassification weights of the WLightGBM models.

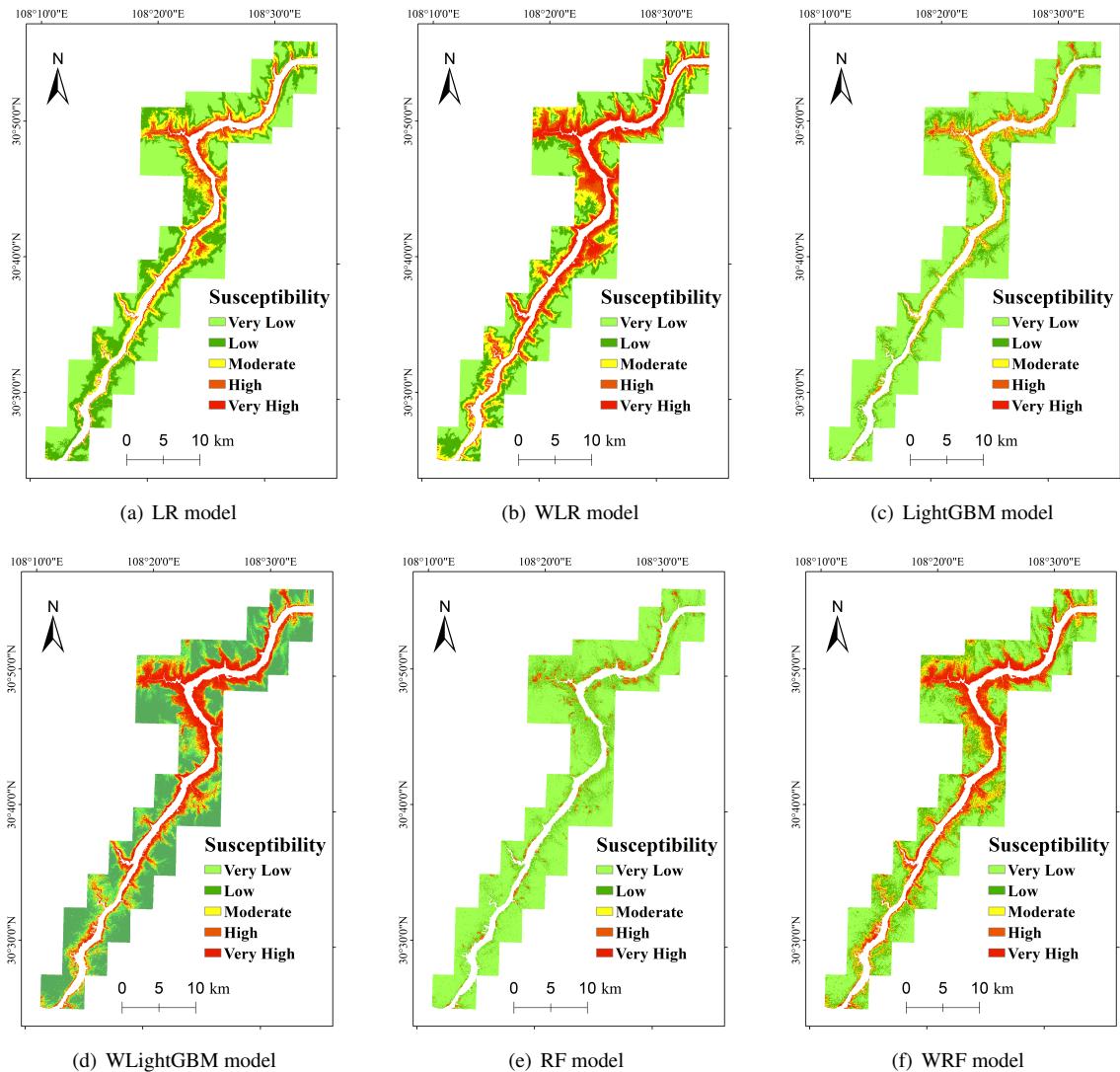


Figure 6: LSM results of 6 models. (a) LSM using LR model. (b) LSM using WLR model. (c) LSM using LightGBM model. (d) LSM using WLighGBM model. (e) LSM using RF model. (f) LSM using WRF model.

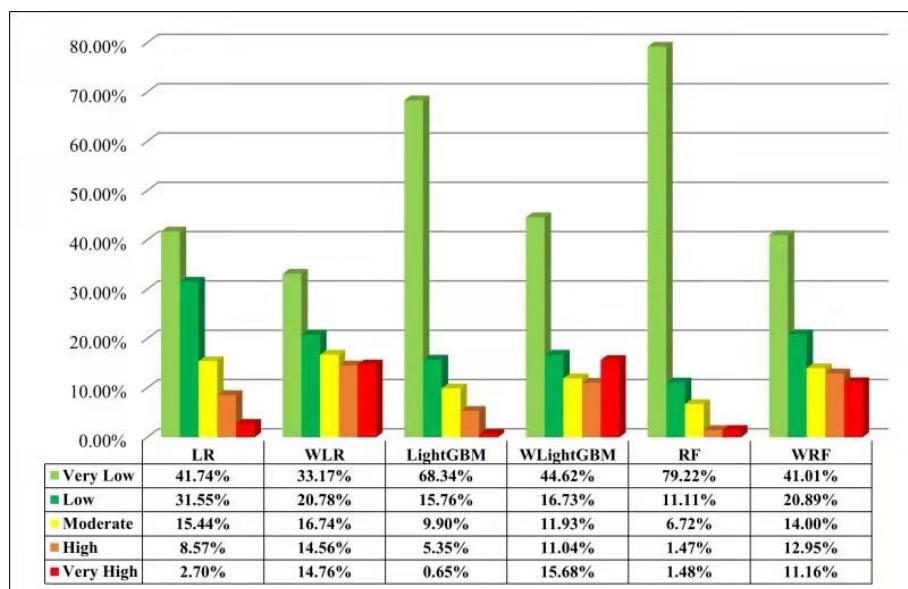


Figure 7: Distribution ratio of different landslide susceptibility classes for 6 models.

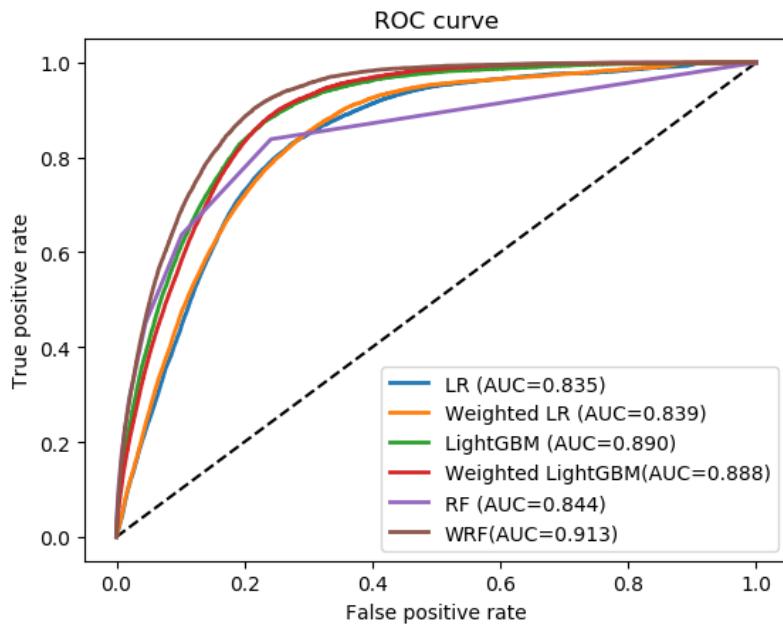


Figure 8: The ROC curve of 6 LSM models.