Potential of Ensemble Learning to Improve Tree-Based Classifiers for Landslide Susceptibility Mapping

Jiahui Song, Yi Wang, *Member, IEEE*, Zhice Fang, Ling Peng, and Haoyuan Hong

***Abstract*—Ensemble learning methods have been widely used due to their remarkable generalized performance, but their potential in landslide spatial prediction application is not fully studied. To take full advantage of ensemble learning techniques, the classification and regression tree classifier and four tree-based ensemble classifiers of random forest, extremely randomized tree, gradient boosting decision trees, and extreme gradient boosting decision trees are used in this study for landslide susceptibility assessment. Specifically, a stacking ensemble learning framework coupled with embedded feature selection is presented, consisting of multiple tree-based classifiers mentioned previously as base learners and logistic regression as a metalearner in a two-layer structure. In the study area of Yongxin, China, 364 historical landslide locations were first randomly partitioned into a ratio of 7/3 for training and testing the model. Then, a spatial database of 16 landslide causative factors was constructed for landslide prediction. Meanwhile, the relative importance of these factors were quantified by using the total number of feature splits and the average Gini index during the training process, and a novel embedded feature selection method was used in the base learner of the proposed framework to further improve the computational efficiency and predictive performance by allowing each base learner to obtain its own optimal subfeature space. Finally, different methods were assessed by using several evaluation criteria. Experimental results demonstrated that the proposed ensemble learning framework had the highest area under the curve value of 0.864, and it is more effective than the conventional tree-based classifiers and other ensemble learning methods.**

***Index Terms*—Embedded feature selection, ensemble learning, landslides susceptibility mapping, tree-based classifiers.**

I. Introduction

On a global scale, landslides are one of the most destructive geo-hazards, posing a serious threat to human life and causing a lot of economic losses [1]. China is one of the most active landslide-prone areas in the world. The increasing disasters and the demand for risk management make it urgent for professionals to assess and mitigate landslide risks [2]. In order to effectively formulate countermeasures to prevent landslide disasters from a macro perspective, it is necessary to identify potential landslide-prone areas [3]. In this regard, landslide susceptibility mapping (LSM), representing the spatial distribution of the probability of landslide occurrences in cartography, has been used as one of the most effective tools for landslide disaster management and mitigation [4], [5].

With the development of computer systems and geographic information system (GIS) tools, various machine-learning methods have been proposed for LSM in recent years, including artificial neural networks (ANN) [6], [7], logistic regression (LR) [8]–[10], decision trees [11]–[13], and support vector machines (SVM) [14]–[16]. Although these methods are not the same in principle, all of them are mainly based on the following assumptions [17]. First, landslide occurrence is controlled by some physical laws that can be analyzed and learned. Second, landslide causative factors are directly or indirectly related to landslide occurrence. Finally, future landslides are more likely to occur under the similar conditions that cause historical landslides.

Ensembles are well-established machine learning techniques that can obtain more accurate prediction results by integrating various base learners [18]. The state-of-the-art ensemble techniques can be generally divided into three groups: bagging, boosting, and stacking [19]. In comparison, ensemble learning methods are constructed with sequential or parallel base learners, and these learners can be homogeneous or heterogeneous in stacking ensembles, while they are homogeneous in bagging and boosting. In order to break through the limitations of a single machine learning algorithm, many ensemble learning methods have been applied to LSM in recent years [13], [20]–[26]. Most of them combine homogeneous ensemble frameworks of bagging or boosting with tree-based classifiers to further improve the performance of landslide susceptibility modeling. Unlike the two frameworks, stacking can combine multiple types of learning algorithms through combination algorithms to maximize the generalization accuracy [27], [28]. However, very little attention was paid to the application of the stacking ensemble to integrate multiple types of classifiers in LSM. Furthermore, the potential of ensemble learning methods in improving tree-based classifiers for LSM has been greatly limited. On the other hand,

Manuscript received May 16, 2020; revised July 7, 2020; accepted July 30, 2020. Date of publication August 4, 2020; date of current version August 26, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61271408 and Grant 41602362 and in part by the China Scholarship Council (201906860029). *(Corresponding authors: Yi Wang; Haoyuan Hong.)*

Jiahui Song, Yi Wang, and Zhice Fang are with the Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China (e-mail: 865610133@qq.com; cug.yi.wang@gmail.com; xmblb123@163.com).

Haoyuan Hong is with the Department of Geography and Regional Research, University of Vienna, 1010 Vienna, Austria (e-mail: hong_haoyuan@outlook.com).

Ling Peng is with the China Institute of Geo-Environment Monitoring, Beijing 100081, China (e-mail: pengl@cigem.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3014143

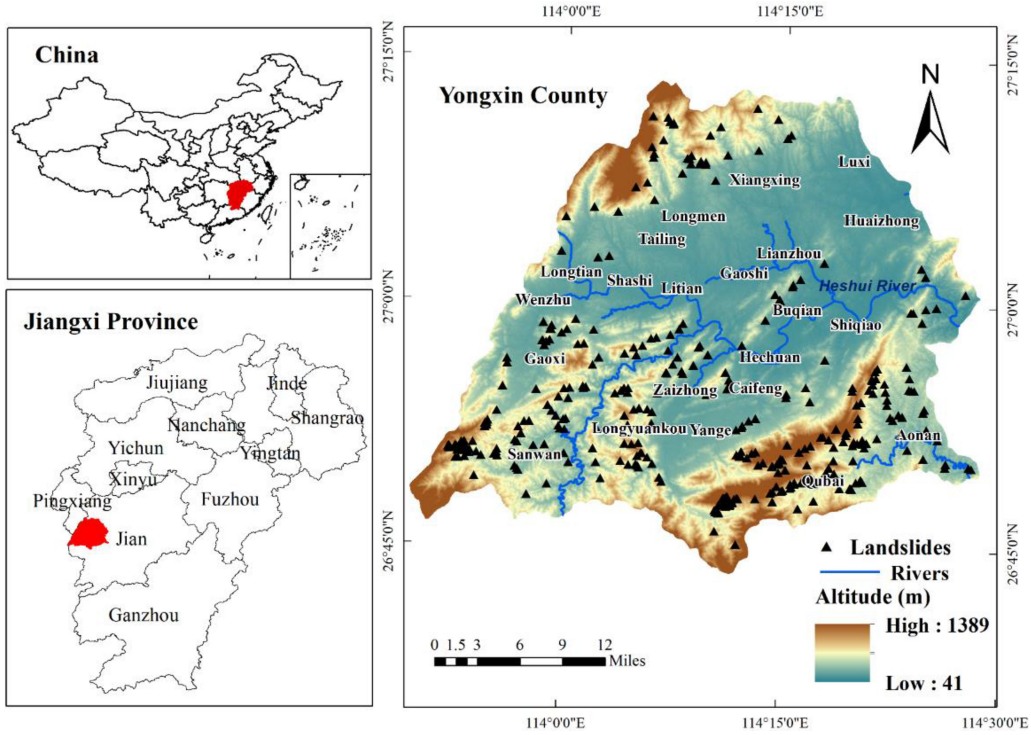


Fig. 1. Location of the study area.

feature selection is usually used to eliminate redundant landslide causative factors during the landslide susceptibility modeling process. Feature selection directly affects final landslide prediction results [29] and can be mainly classified into three categories: filter, wrapper, and embedded. Filter methods have been commonly used for landslide susceptibility assessment, including information gain ratio [30]–[32], correlation-based methods [33], [34], and relief-F [22], [35]. Although these methods are interpretable and easily calculated, they cannot effectively play a significant role in feature optimization [36]. In particular, the filter methods may delete features that have useful information to some learners before the base learner is trained during the heterogeneous ensemble learning process. However, the embedded methods incorporate feature selection as a part of the training process of the learner, which can conveniently obtain the optimal feature subset of the learner in the heterogeneous ensemble learning framework.

To solve the problems mentioned above, we present a two-layer stacking ensemble learning method (SELM) framework to explore the potential of ensemble methods for LSM. To the best of our knowledge, studies on stacking ensemble of different prediction models for LSM is very rare. In the proposed framework, five tree-based methods are used as base learners, including classification and regression tree (CART), two bagging methods of random forest (RF), extremely randomized tree (ERT), and two boosting methods of gradient boosting decision trees (GBDT) and extreme gradient boosting decision trees (XGBoost), and LR is used as a metalearner. Furthermore, a novel embedded feature selection (EFS) method is used in the base learner of the proposed framework to further improve computational efficiency and predictive performance by allowing each base learner to obtain its own optimal subfeature space. Compared with previous studies, this study is not a simple combination of single classifiers with different homogeneous integration methods, it is an exploration for the potential of heterogeneous (stacking) ensemble learning in LSM. The three main contributions of this study can be summarized as follows: First, the heterogeneous stacking strategy that is integrated into the EFS-SELM framework can maintain heterogeneity by combining different tree-based classifiers, which can solve the generalization problems to some extent and be more suitable for this field. Second, the discussion on the relationship between landslide inherent mechanism and geological structures can provide a certain explanation for final susceptibility results. Furthermore, the discussion on the impacts of susceptibility results on landslide disaster reduction and management can also provide reliable guides for researchers, engineers, and policy-makers. The effectiveness of the proposed stacking ensemble learning method with embedded feature selection (EFS-SELM) framework was systematically verified using landslide data from Yongxin County, China, and it was compared with the conventional tree-based classifiers mentioned previously.

II. Study Area and Data Preparation

A. Description

Yongxin County, located in the western part of Jiangxi Province, covers an area of about 2187 km^2 and has an elevation of 41–1398 m above sea level (see Fig. 1). The landforms in the study area are mainly mountainous and hilly, with high marginal terrain and low-lying central terrain, forming asymmetric basin landforms. In addition, its climate type is subtropical humid

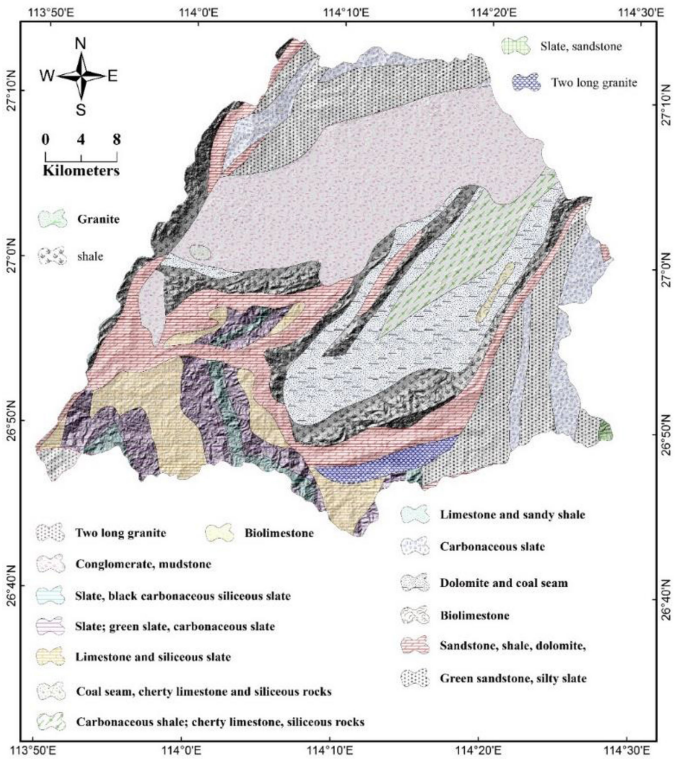


Fig. 2. Geological map of the study area.

monsoon climate with abundant rainfall, mild climate, and four distinct seasons. As the county is located in the mountainous and pluvial regions, it is one of the most serious landslide-prone areas, in China, especially during the rainy season. According to the local government report in Jiangxi Province, the lives of about 2174 people were affected by disastrous landslide events. As the number of catastrophic climatic events increases, landslide disasters are becoming more frequent in this area, posing a serious threat to human life and causing a lot of economic losses. Therefore, it is necessary to perform LSM in this area to prevent and mitigate the adverse effects of landslides. Furthermore, the geological environment and climatic conditions in this study area are typical for landslide-prone areas. As a result, the selection of this area will help the landslide spatial prediction models to obtain more comprehensive causative factors, which will facilitate the comparison of the various models and their application outside this study area where they were constructed.

Geologically, the study area is part of the fold system in southern China, and its tectonics is located in the eastern part of the Caledonides orogenic belt in China. It has abundant typical tectonic landforms and is characterized by a multiphase fault zone and a ductile shear zone from the northeast to Yanshanian. Except for the Sinian, Silurian, and Tertiary, the Cambrian to Quaternary strata are well distributed in the study area, with a total thickness of more than 20 000 m. The distribution of geological units in the study area is shown in Fig. 2. According to lithofacies and geological time, these units are divided into 17 groups, of which conglomerate, dolomite, sandstone, and limestone are the main outcrops.

B. Data Preparation

1) Landslide Inventory Map: A landslide inventory is generally defined as a collection of historical landslide data that contains information on the area, type, activity, and physical properties [37]. It can provide important clues between landslide occurrences and causative factors to predict the area's future landslide possibility. In this study, a landslide inventory map was constructed by using historical landslide records, interpretations of satellite images, and field survey data provided by the local government. As shown in Fig. 1, a total 364 landslide locations were recorded, consisting both rotational slides (70%) and translational slides (30%). The largest and smallest landslides are 750 000 and 32 m^2, respectively. Among all the recorded landslides, 21.6%, 37.8%, and 40.6% were classified as large-scale ($>1000\,m^2$), medium-scale (400–1000 m^2), and small-scale ($<400\,m^2$), respectively.

2) Landslide Causative Factors: According to the previous research, the selection of the landslide causative factors should consider study area characteristics, scale of the analysis and the data availability [38], [39]. Base on this summarization, we use six geomorphic factors [altitude, slope, aspect, plan curvature, profile curvature, and sediment transport index (STI)], a tectonic factor (distance to fault), a geologic factor (lithology), a triggering factor (rainfall), three hydrological factors [stream power index (SPI), topographic wetness index (TWI) and distance to river], four land-related factors [land use, normalized difference vegetation index (NDVI), distance to road and soil]. The literature review demonstrated that most of the statistical methods used landslide predisposing factors related mainly to geomorphology [17]. These factors that can describe geomorphology are always obtained from digital elevation model (DEM) data, and some direct measures of geomorphology have been commonly used in LSM, including altitude, slope, aspect, plan curvature, profile curvature, and STI. Moreover, according to the Meteorological Bureau of Jiangxi Province, rainfall is the main landslide-triggering factor in the study area, rather than tectonic seismicity. Moreover, the study area is located in the hinterland of the Eurasian Plate, which has no active fault zone. Therefore, the crust is relatively stable. However, the rocks exposed near the faults generally have loose geotechnical structure, low shear strength and weather resistance. In these areas, the fracture pores are more developed, which is conducive to the infiltration of atmospheric precipitation and the migration of groundwater, providing dynamic conditions for the occurrence of landslides. Therefore, distance to faults was used as a tectonic-related predisposing factor in this study. For clarity, the detailed descriptions of these factors are listed in Table I, and all thematic maps of the factors were prepared using ArcGIS 10.2 and are shown in Fig. 3.

The DEM of the study area was generated from the ASTER GDEM,[1] Landsat 7 ETM+ satellite images of the study area were obtained from the Computer Network Information Center of Chinese Academy of Sciences,[2] the geological map was

[1][Online]. Available: http://gdem.ersdac.jspacesystems.or.jp
[2][Online]. Available: http://www.gscloud.cn

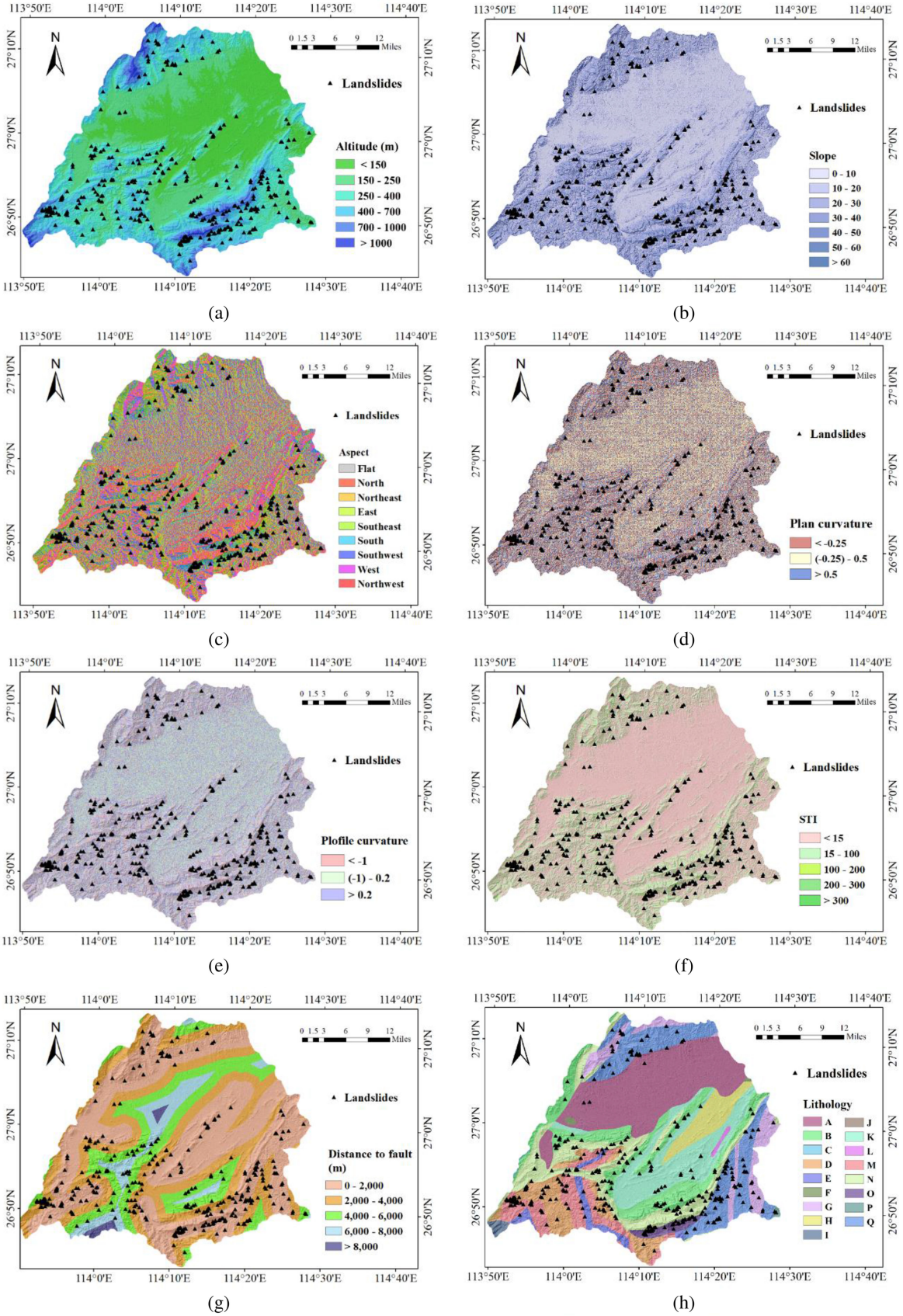


Fig. 3. Landslide causative factor maps. (a) Altitude, (b) slope, (c) aspect, (d) plan curvature, (e) profile curvature, (f) sediment transport index (STI), (g) distance to faults, (h) lithology, (i) rainfall, (j) stream power index (SPI), (k) topographic wetness index (TWI), (l) distance to river, (m) land use, (n) normalized difference vegetation index (NDVI), (o) distance to road, and (p) soil.

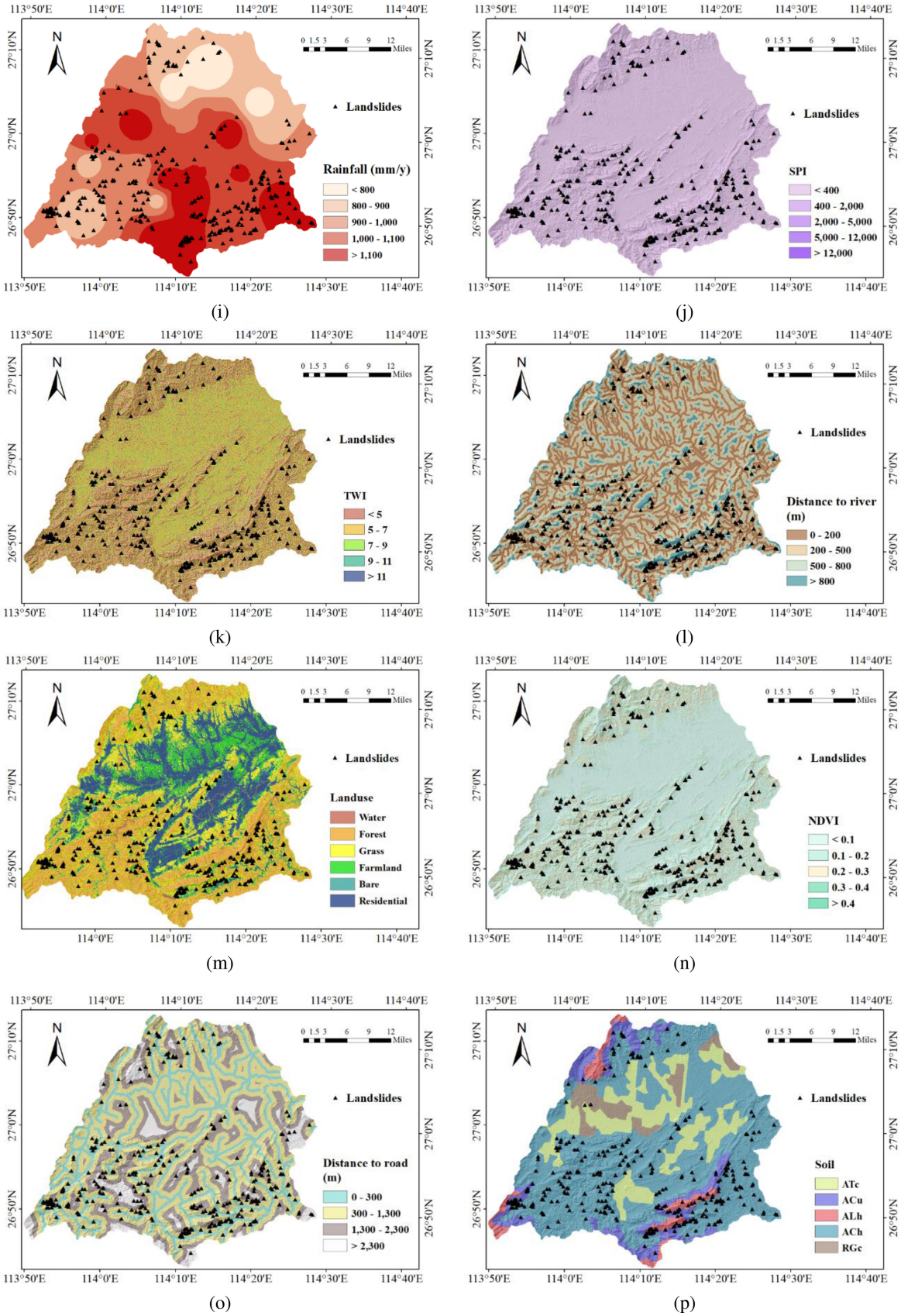


Fig. 3. Continued.

TABLE I
DESCRIPTIONS OF LANDSLIDE CAUSATIVE FACTORS

Data Type	Factors	Range	Source	Resolution/Scale
Geomorphic	**Altitude (m)**	(41,1389)	DEM	25m
	Slope (°)	(0, 67.60)	DEM	25m
	Aspect	Flat, North, West, South, Southeast, East, Northwest, Southwest, Northeast	DEM	25m
	Plan curvature	(-17.80, 18.00)	DEM	25m
	Profile curvature	(-15.42, 14.57)	DEM	25m
	Sediment Transport Index (STI)	(0, 1322)	DEM	25m
Tectonic	**Distance to fault (m)**	(0, 10217)	Geological map	1:200000
Geologic	**Lithology**	17 Groups	Geological map	1:200000
Triggering	**Rainfall**	(754, 1236)	GIS database	-
Hydrological	**Stream Power Index (SPI)**	(0, 46659)	DEM	25m
	Topographic Wetness Index (TWI)	(2.39, 32.78)	DEM	25m
	Distance to river (m)	(0, 1942)	GIS database	-
Land-related	**Land use**	water, residential area, forest land, bare land, farmland and grassland	Landsat ETM+ satellite images	30m
	Normalized Difference Vegetation Index (NDVI)	(-0.57, 0.43)	Landsat ETM+ satellite images	30m
	Distance to road (m)	(0, 5429)	GIS database	-
	Soil	RGc, ATc, ALh, ACu, Ach	Soil map	-

derived from the China Geology Survey,[3] and the soil map was prepared from the Institute of Soil Science, Chinese Academy of Sciences.[4] In addition, the GIS database of rainfall was derived from the average annual precipitation of 20 rain stations[5] distributed in the study area from 1960 to 2015, and precipitation values of the area are determined by the inverse distance weighted spatial interpolation method. The GIS databases[6] of distance to road and distance to river were obtained by buffer analysis using road and river vector map with buffer distance of four grades.

The selection of the mapping unit is an important prerequisite for LSM. The grid-based method is by far the most popular for landslide susceptibility modeling using raster data [17]. In order to effectively compute landslide susceptibility of each grid unit, in addition to unifying all landslide causative factors into a raster form with respect to the DEM spatial resolution (25 m), the data of each factor must be reclassified according to its essential structure. Specifically, the continuous-valued factors of altitude, slope, plan curvature, profile curvature, SPI, STI, TWI, NDVI, rainfall, distance to fault, distance to river, and distance to road were reclassified into several discrete subcategories. For land use, the study area was classified into six classes: water, forest, grassland, bare land, farm land, and residential area, with an overall accuracy of 92.4% using maximum likelihood. The factor of slope aspect was classified into eight directions and flat (no aspect), and the soil was divided into RGc, ATc, ALh, ACu, and Ach. According to the geological map, the lithology of the study area was divided into 17 units, and the descriptions of each units are listed in Table II [15]. Therefore, all the causative factors were reclassified so that each grid cell corresponds to a new class of values of all 16 factors, and these new classes were used as input data in modeling. The normalized classes of all the causative factors and the corresponding frequency ratios (FRs) based on landslide densities are listed in Table III.

[3][Online]. Available: http://www.cgs.gov.cn
[4][Online]. Available: http://www.issas.ac.cn
[5][Online]. Available: http://www.weather.org.cn
[6][Online]. Available: http://www.geodata.cn/

III. METHODOLOGY

As shown in Fig. 4, the proposed method in this study mainly consists of four steps. First, we constructed a spatial database containing the landslide inventory map and causative factors of the study area, and historical landslide locations are divided into two groups for training and verification. Second, we used the spearman's rank correlation coefficient to quantify the correlation between landslide causative factors, and the importance of each factor is quantified when training the tree-based models. Third, CART, four tree-based ensemble methods, and SELM and ESF-SELM methods are used to assess the susceptibility of landslides. Finally, we evaluated and compared the performance of these models mentioned previously by receiver operating characteristic (ROC) curve and five metrics.

A. Tree-Based Single Model and Ensemble Models

1) CART: CART is a classic machine learning method [40]. Unlike the C4.5 decision tree, CART is essentially a binary

TABLE II
DESCRIPTIONS OF LITHOLOGY IN THE STUDY AREA

Group name	Unit name	Lithology
A	Lianhe group, Tangbian group, Hekou group	Conglomerate, mudstone
B	Zhangzong group, Zhongpeng group, Yunshan group	Shale
C	Changlong group, Oujia chong group	Limestone and sandy shale
D	Huamian gong group, Shi kou group	Limestone and siliceous slate
E	Duier shi group, Jueshan gou group	Slate, black carbonaceous siliceous slate
F	Longtang group, Qi baoshan group, Chang xing group	Coal seam, cherty limestone and siliceous rocks
G	Ba cun group, Liu jiaohe group	Carbonaceous slate
H	Gu feng group, Qixia group, Xiao Jiangbian group	Carbonaceous shale; cherty limestone, siliceous rocks
I	Huang xie group, Hai hui group, Xi hua group	Two long granite
J	Chang lejie group, Gu poshan group, San jiangkou group	Granite
K	Zishan group, Yang jiayuan group	Dolomite and coal seam
L	Hu tian group	Biolimestone
M	Dui ershi group, Shi kou group	Slate; green slate, carbonaceous slate
N	Xia shan group, Yi jiawang group, Qi ziqiao group	Sandstone, shale, dolomite
O	Fu fang group, Taihe group, Tang hu group	Two long granite
P	Ba cun group, Shui shi group	Slate, sandstone
Q	Ba cun group, Gao tang group	Green sandstone, silty slate

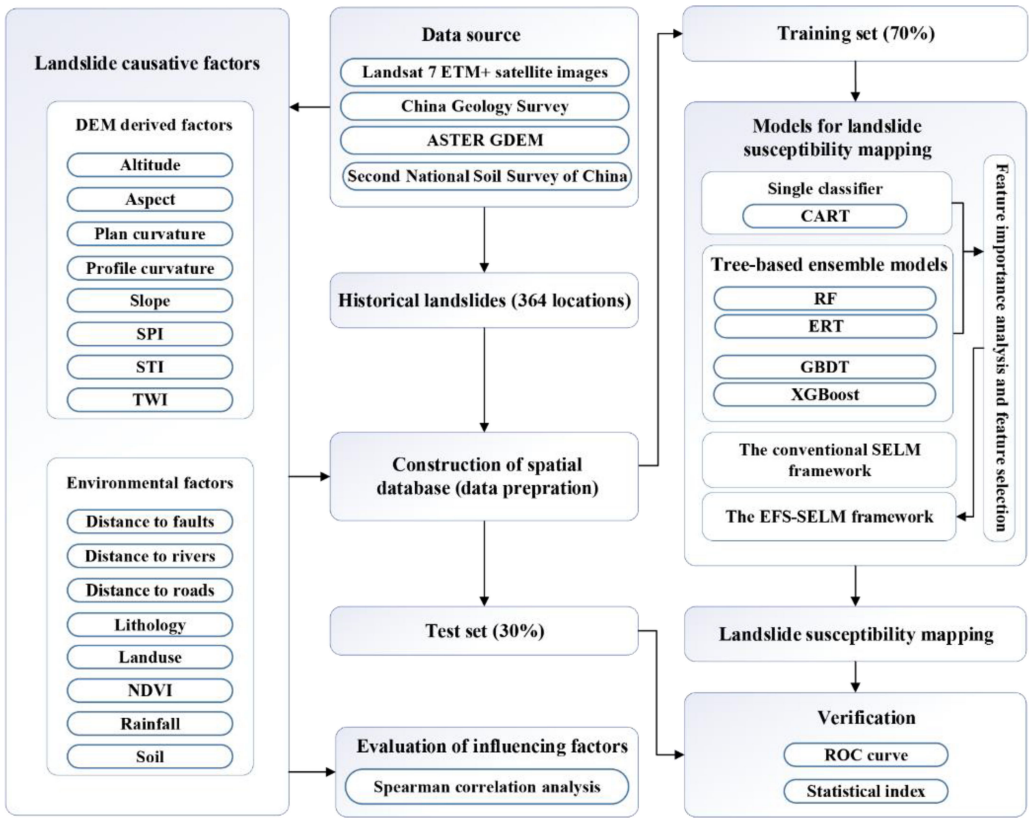


Fig. 4. Flowchart of this study.

TABLE III
RECLASSIFICATION INFORMATION OF LANDSLIDE CAUSATIVE FACTOR

Causative factors	Class	Normalized class	Number of pixels in domain	Percent of domain (%)	Number of landslides	Percent of landslide (%)	FR
Altitude (m)	<150	1	876142	24.84	4	1.10	0.04
	150-250	2	947989	26.88	55	15.11	0.56
	250-400	3	822620	23.33	127	34.89	1.50
	400-700	4	663718	18.82	146	40.11	2.13
	700-1000	5	180954	5.13	30	8.24	1.61
	>1000	6	35309	1.00	2	0.55	0.55
Aspect	Flat	1	26229	0.74	13	3.57	4.80
	North	2	460854	13.07	16	4.40	0.34
	Northeast	3	409633	11.62	53	14.56	1.25
	East	4	472022	13.38	90	24.73	1.85
	Southeast	5	458681	13.01	113	31.04	2.39
	South	6	466466	13.23	48	13.19	1.00
	Southwest	7	381226	10.81	23	6.32	0.58
	West	8	422817	11.99	8	2.20	0.18
	Northwest	9	428804	12.16	0	0.00	0.00
Distance to fault (m)	0-2000	1	1605946	45.54	191	52.47	1.15
	2000-4000	2	1030730	29.23	110	30.22	1.03
	4000-6000	3	626737	17.77	55	15.11	0.85
	6000-8000	4	234027	6.64	8	2.20	0.33
	>8000	5	29292	0.83	0	0.00	0.00
Land use	Water	1	42077	1.19	0	0.00	0.00
	Forest	2	1231417	34.92	103	28.30	0.81
	Grass	3	926977	26.28	213	58.52	2.23
	Farmland	4	468788	13.29	12	3.30	0.25
	Bare	5	150316	4.26	30	8.24	1.93
	Residential	6	707157	20.05	6	1.65	0.08
Lithology	Group A	1	784838	22.25	3	0.82	0.04
	Group B	2	429108	12.17	62	17.03	1.40
	Group C	3	2311	0.07	0	0.00	0.00
	Group D	4	233843	6.63	39	10.71	1.62
	Group E	5	77394	2.19	6	1.65	0.75
	Group F	6	6728	0.19	0	0.00	0.00
	Group G	7	189655	5.38	18	4.95	0.92
	Group H	8	145633	4.13	0	0.00	0.00
	Group I	9	17333	0.49	0	0.00	0.00
	Group J	10	5	0.00	0	0.00	0.00
	Group K	11	423127	12.00	14	3.85	0.32
	Group L	12	9764	0.28	0	0.00	0.00
	Group M	13	234472	6.65	46	12.64	1.90
	Group N	14	446643	12.66	67	18.41	1.45
	Group O	15	48645	1.38	36	9.89	7.17
	Group P	16	4670	0.13	2	0.55	4.15
	Group Q	17	472563	13.40	71	19.51	1.46
NDVI	<0.1	1	1821056	51.64	60	16.48	0.32
	0.1-0.2	2	860462	24.40	74	20.33	0.83
	0.2-0.3	3	763478	21.65	195	53.57	2.47
	0.3-0.4	4	81671	2.32	35	9.62	4.15
	>0.4	5	65	0.00	0	0.00	0.00
Plan curvature	<-0.25	1	1242908	35.24	164	45.05	1.28
	-0.75	2	1369415	38.83	98	26.92	0.69
	>0.5	3	914409	25.93	102	28.02	1.08
Profile curvature	<-1	1	626562	17.77	95	26.10	1.47
	(-1)-0.2	2	1387942	39.35	91	25.00	0.64
	>0.2	3	1512228	42.88	178	48.90	1.14
Rainfall	<800	1	27289	0.77	0	0.00	0.00
	800-900	2	309590	8.78	8	2.20	0.25
	900-1000	3	1015248	28.79	113	31.04	1.08
	1000-1100	4	1057419	29.98	83	22.80	0.76
	>1100	5	1117186	31.68	160	43.96	1.39
Distance to river (m)	<200	1	1198190	33.97	88	24.18	0.71
	200-500	2	1321706	37.48	154	42.31	1.13
	500-800	3	735755	20.86	94	25.82	1.24
	>800	4	271081	7.69	28	7.69	1.00

TABLE III
CONTINUED

Distance to road (m)	<300	1	737514	20.91	45	12.36	0.59
	300-1300	2	1736532	49.24	213	58.52	1.19
	1300-2300	3	787621	22.33	84	23.08	1.03
	>2300	4	265065	7.52	22	6.04	0.80
Slope (°)	0-10	1	1466412	41.58	64	17.58	0.42
	10-20	2	1036916	29.40	111	30.49	1.04
	20-30	3	620572	17.60	95	26.10	1.48
	30-40	4	305484	8.66	66	18.13	2.09
	40-50	5	87469	2.48	24	6.59	2.66
	50-60	6	9679	0.27	4	1.10	4.00
	>60	7	200	0.01	0	0.00	0.00
Soil	ATc	1	584915	16.59	1	0.27	0.02
	ACu	2	346372	9.82	77	21.15	2.15
	ALh	3	136843	3.88	22	6.04	1.56
	ACh	4	2275477	64.52	263	72.25	1.12
	RGc	5	183125	5.19	1	0.27	0.05
SPI	<400	1	3411208	96.72	334	91.76	0.95
	400-2000	2	99870	2.83	22	6.04	2.13
	2000-5000	3	12610	0.36	8	2.20	6.15
	12000	4	2675	0.08	0	0.00	0.00
	>12000	5	369	0.01	0	0.00	0.00
STI	<15	1	2701877	76.61	198	54.40	0.71
	15-100	2	781291	22.15	155	42.58	1.92
	100-200	3	34803	0.99	8	2.20	2.23
	200-300	4	6041	0.17	3	0.82	4.81
	>300	5	2720	0.08	0	0.00	0.00
TWI	<5	1	1129914	32.04	159	43.68	1.36
	5-7	2	1678597	47.60	141	38.74	0.81
	7-9	3	546447	15.49	40	10.99	0.71
	9-11	4	150377	4.26	19	5.22	1.22
	>11	5	21397	0.61	5	1.37	2.26

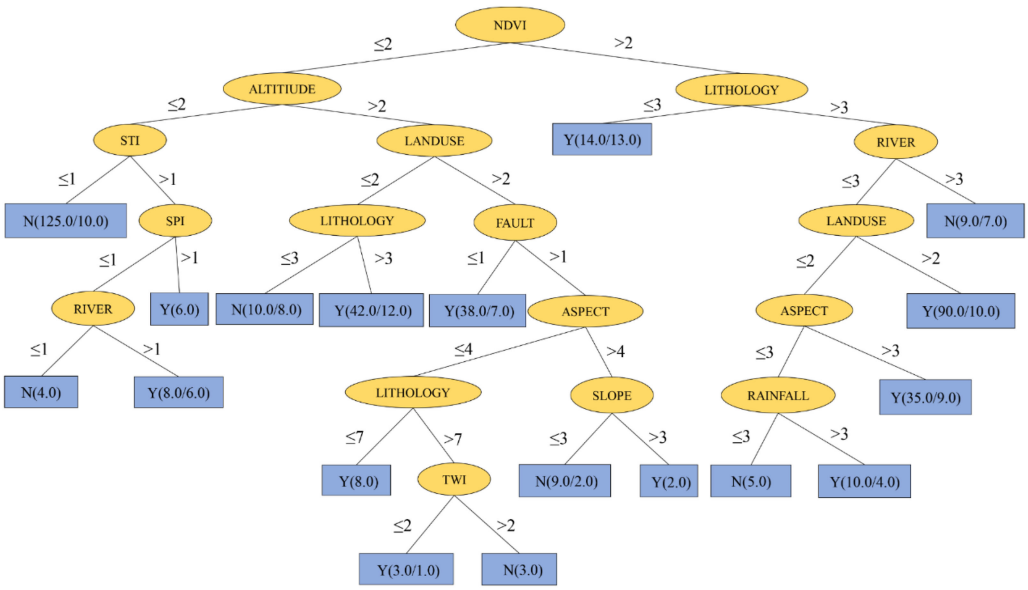


Fig. 5. Flowchart of training CART for landslide prediction.

partition of the recursive feature space, and the partitioning process can be graphically represented [41].

In this study, CART is used for landslide susceptibility assessment. The CART classifier for landslide prediction is shown in Fig. 5. The landslide causative factors are the bifurcation points of CART, and the leaf nodes Y and N represent landslide and nonlandslide, respectively. Although CART is an effective method and can be easily visualized or even extract classification rules, its prediction ability and generalization ability can be further improved through ensemble frameworks.

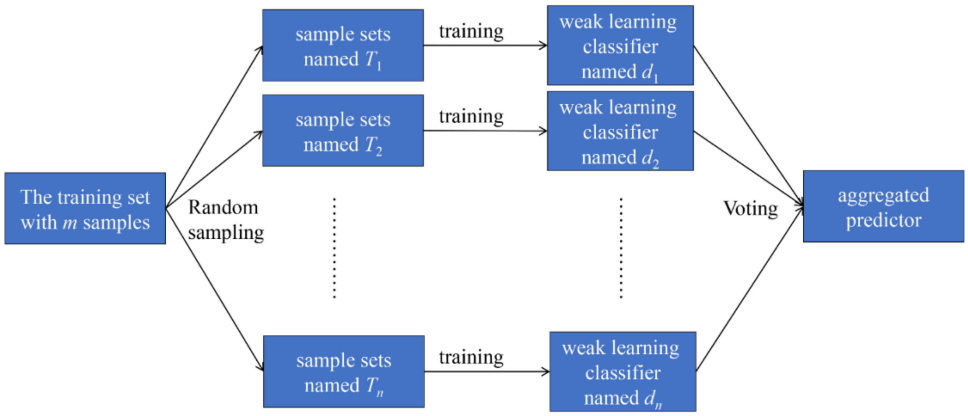


Fig. 6. General overview of the bagging process.

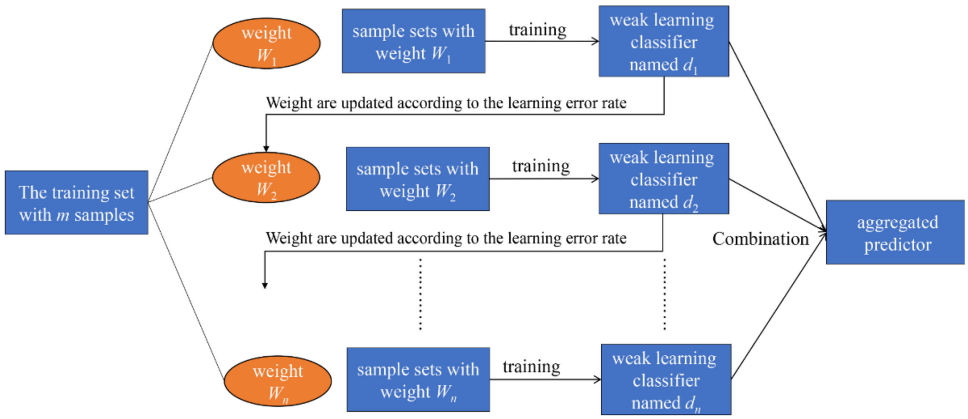


Fig. 7. General overview of the boosting process.

2) Tree-Based Ensemble Learning Methods: The bootstrap aggregating method of bagging [42] is one of the earliest ensemble method. The general process of bagging is shown in Fig. 6. First, the subtraining sets are obtained from random subsampling in the training set. Then, these subtraining sets are used to train weak learners. Finally, an aggregated predictor is obtained by voting on these weak learners.

RF is a widely used algorithm that is derived from the idea of the bagging ensemble. It is a collection of CART, so each of its trees depends on the value of independently sampled random vector [43]. The optimal split of each node can be obtained by searching a random subset of candidate attributes.

ERT shares multiple features with RF and further play the randomness in tree splitting. ERT is also a representative algorithm combining bagging and CART. However, the main difference between ERT and RF is that instead of choosing the best cut-point in each node based on local samples, the ERT algorithm randomly selects the best splitting point in a node [44].

Boosting is to sequentially produce weak learners in an iterative manner in sequence. It does not randomly select training samples like the Bagging ensemble, but focuses on samples that do not have accurate predictions. A general overview of the boosting process is shown in Fig. 7.

GBDT is a novel and representative boosting-based algorithm. This algorithm uses CART as the base learner, and provides a competitive and highly robust tool for regression and classification [45]. For the binary classification problem, GBDT uses a negative gradient similar to the log-likelihood loss function of LR to fit the approximate value of the loss.

XGBoost is a scalable end-to-end tree boosting system, which is widely used in different machine learning tasks. Compared with GBDT, its computational speed and accuracy have been significantly improved. Its main innovations can be summarized as follows [46]. First, its loss function is optimized. Second, the candidate split value is efficiently generated by parallel approximate histogram algorithm. Finally, it presents an effective cache-aware block structure for out-of-core tree learning and a novel sparsity-aware algorithm for parallel tree learning.

B. Evaluation Methods for Causative Factor Analysis

In this study, the relative importance of landslide causative factors and the correlation between the factors are quantified using the measures of feature importance measure (FIM) and spearman's rank correlation coefficient, respectively, and then the analyzed results are used as a measure for the base learner to perform embedded feature selection in the first layer of the EFS-SELM framework.

1) FIM: The importance of a feature implies how much it contributes to the accuracy of the output during the prediction process [47]. The relative importance of landslide causative factors may vary because of different prediction methods. In this study, CART uses the Gini index (GI) to calculate the importance of features in the training process, which represents the probability that randomly selected samples are misclassified in a subset. The smaller the GI, the higher the purity of the dataset. Let $X = \{X_1,\ X_2,\ \ldots, X_J\}$ be the features, *D* a training set, *K* the number of categories, and p_k the probability that a sample is classified into the *k*th class, the GI is given by the following:

$$Gini\,(D) = \sum_{i=1}^{K} p_k \cdot (1 - p_k) = 1 - \sum_{i=1}^{K} p_k^2 \tag{1}$$

Then, the FIM of X_j at the *m*th node of CART is calculated as follows:

$$FIM_{jm} = GI_m - GI_l - GI_r \tag{2}$$

where GI_m represents the Gini index of the *m*th node, and GI_l and GI_r indicate the Gini indexes of two new nodes after the branch, respectively. Assuming that nodes split by feature X_j in CART are in the set *M*, then the FIM of this feature in the CART is calculated as follows:

$$FIM_j^{(CART)} = \sum_{m \in M} FIM_{jm} \tag{3}$$

Assuming that there are *n* trees in RF, the FIM of the feature X_j in RF is calculated as follows:

$$FIM_j^{(RF)} = \sum_{i=1}^{n} \sum_{m \in M} FIM_{jm} \tag{4}$$

The FIM calculation of ERT, GBDT, and RF are basically the same. As the ensemble model has a certain randomness in selecting feature splitting, the feature importance of each tree is usually averaged. When quantifying the importance of a feature, XGBoost uses the number of times that the feature is used as a partition node in all trees as an evaluation indicator to measure its importance. Thus, the more times a feature is selected to be split, the greater the importance inside the tree. Finally, each feature's FIM of each model is normalized and converted into a number from 0 to 1. The closer it is to 1, the greater the role of this factor as a feature in the prediction of the corresponding model.

2) Spearman's Rank Correlation Coefficient: In order to estimate the correlation between landslide causative factors, the spearman correlation analysis method is used in this study. Spearman's rank correlation coefficient is a nonparametric statistical correlation measure of rank correlation to evaluate how well the relationship is between elements in two different sets. Given a pair of feature vectors *X* and *Y* of length *N*, the spearman's rank correlation coefficient can be calculated as follows:

$$Rs = 1 - \frac{6 \cdot \sum_{i=1}^{N} |R\,(X_i) - R\,(Y_i)|^2}{N \cdot (N^2 - 1)} \tag{5}$$

where $R(X_i)$ and $R(Y_i)$ represent the rank of elements X_i and Y_i in the feature vectors *X* and *Y*, respectively. It can be clearly seen that spearman's rank correlation coefficient ranges from –1 to 1, representing the total negative linear correlation and positive linear correlation, respectively. The higher the absolute value of the coefficient, the more related the two factors are. In practice, if the absolute value of the correlation coefficient between the two features is greater than 0.7, indicating the correlation is too strong [48], one of these two features should be excluded.

C. Proposed Framework

Over the last two decades, Bagging and Boosting have been the most representative homogeneous ensemble techniques, while stacking has become a commonly used technique for heterogeneous ensemble approach since Wolpert first presented the related study in 1992 [49]. In fact, stacking is a general two-level framework, where the metalearner (second layer) is trained by the prediction values produced by the first-layer base learners to make the final prediction. The first layer of the SELM framework is like a highly complex nonlinear feature converter. After this conversion, the samples have new representations. Therefore, the second layer does not require complex classifiers, and generalized linear models such as LR are a suitable choice. In this study, we present a novel strategy to predict landslide susceptibility using the SELM framework, which combines the advantages of various tree-based algorithms (CART, RF, ERT, GBDT, and XGBoost) with a meta-learner (LR) to maximize the generalization accuracy.

The training process of the constructed SELM framework is illustrated in Fig. 8. Given an original training dataset $T = \{(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots, (x_n,\ y_n),\ \ldots, (x_N,\ y_N)\}$, where x_n and y_n represent the feature vector and the target value of the *n*th instance. Then, the original training set is randomly divided into *K* folds $T_1,\ T_2,\ \ldots, T_K$, subsequently, T_k and $T^{(-k)} = T - T_k$ are sequentially selected as the validation and training sets, respectively, of the *k*th fold in the *K*-fold cross validation. Each base learner is trained by $T^{(-k)}$ and then predicts each instance in T_k. Let $P_i(x)$ represent the prediction of the *i*th base learner on an instance with a feature vector x in T_k, and then we will get the predictions of the *i*th base learner on all instances in the original training dataset as follows:

$$P_{in} = P_i\,(x_n)\,, \quad n = 1,\ 2,\ \ldots,\ N \tag{6}$$

where P_{in} can be a class predicted by the base learner (landslide or nonlandslide), or can be a probability value of the class (landslide susceptibility). After the cross-validation process of each base learner is completed, the set of predicted values of each instance output by the base learners and its corresponding target value are combined to construct the training set $T_{meta} = \{(P_{1n},\ \ldots, P_{in},\ \ldots, P_{5n},\ y_n), n = 1,\ 2,\ \ldots, N\}$ of the metalearner (LR) in the second layer. Finally, the metalearner is trained by T_{meta}.

The testing process of the proposed SELM framework is shown in Fig. 9. Let L_{ik} represent the *i*th base learner that is trained by $T^{(-k)}$, given a new test instance, L_{ik}produces a prediction P_{ik} for this instance. The mean value of $P_{ik}(k = 1, 2,\ \ldots, K)$is calculated as the prediction of the *i*th

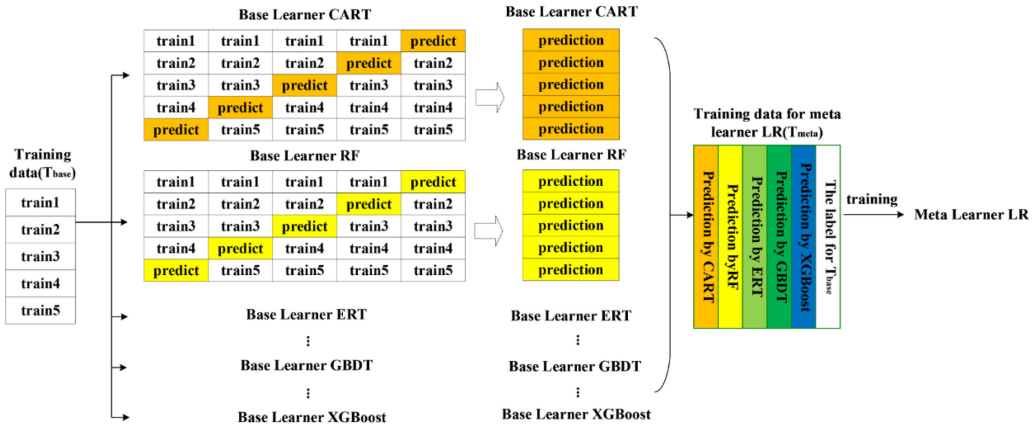


Fig. 8. Training process of the proposed SELM framework.

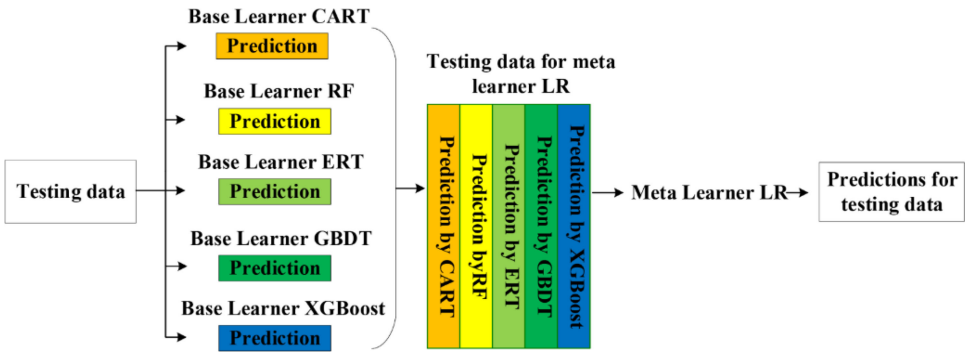


Fig. 9. Test process of the proposed SELM framework.

base learner for this instance. In this way, the base learners produce a vector of predictions, which is used as the input of the trained metalearner (LR) to make the final prediction.

Furthermore, we improve the performance of the traditional SELM framework by embedded feature selection (see Fig. 10), and explore ways to build a more efficient SELM framework. Specifically, in the first layer of SELM, the top j (j is less than the number of features and will be used as a parameter in EFS-SELM to determine its optimal value) important features of each base learner are selected according to the FIM calculated in the training process, and then correlation tests are carried out on these features. If the spearman's rank correlation coefficient between the two features is greater than 0.7, the features with lower FIM in the corresponding learners will be deleted. In this way, each base learner obtains the feature space that it is "good at," and finally, it is combined with metalearner that synthesizes them. The process of EFS-SELM using embedded feature selection to obtain multiple feature subspaces of different base learners not only reduces the dimension of the dataset, but also further mines the advantages of the SELM framework in knowledge discovery and feature extraction.

D. Evaluation Metrics

In landslide susceptibility analysis, assessing the validity of the model used is absolutely an essential component, because it has no scientific significance without verification [50], [51]. The ROC plots the true positive rate on the y-axis and the false positive rate on the x-axis, which helps to indicate the quality of the probabilistic prediction system [52], and has been used for this study. In addition, some commonly used statistical measures such as area under the ROC (AUC), accuracy, recall, precision, and F-measure are also used to assess the predictive capability of landslide models. These statistical measures are calculated by the respective following formulas:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$F - \text{measure} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

where TP (true positive) and TN (true negative) denote the number of correctly classified landslide and nonlandslide samples, whereas FP (false positive) and FN (false negative) mean the number of incorrectly classified landslide and nonlandslide samples, respectively. To apply the evaluation measures mentioned previously to the study area, we use training and test datasets to reflect the fitting ability and predictive ability, respectively.

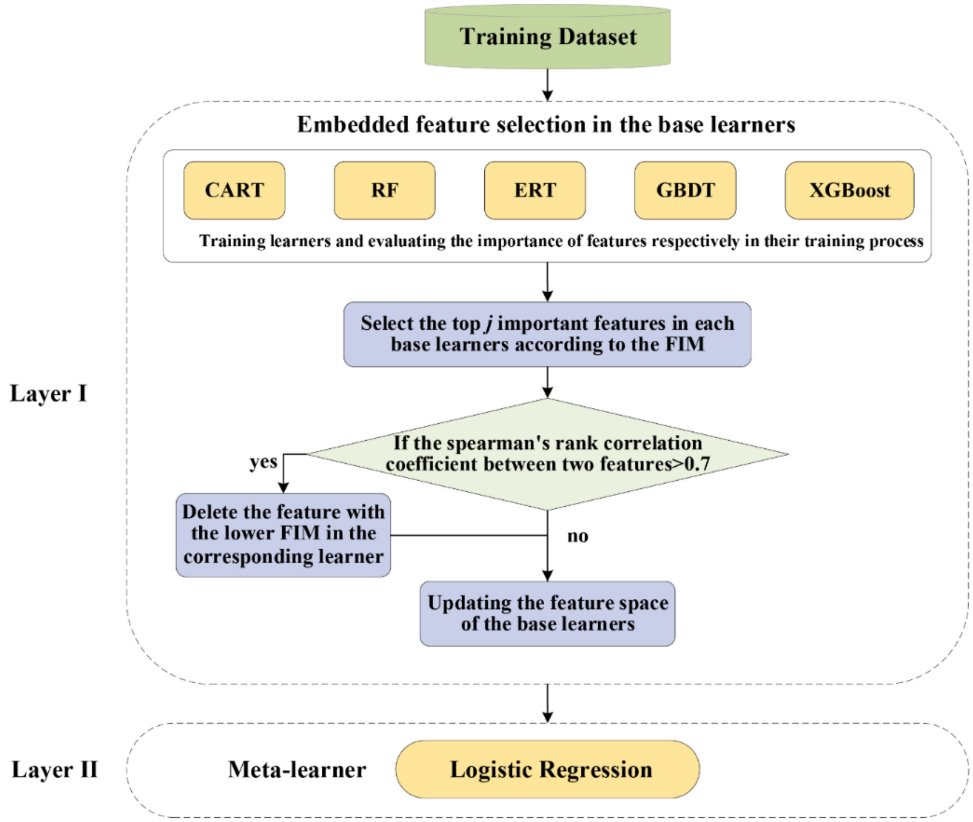


Fig. 10. Flowchart of the proposed EFS-SELM framework.

It should be noted that the performance on the test set better reflects the predictive accuracy and generalization capability of the model, since it is not used for training the model. For the measures of accuracy, precision, recall, and *F*-measure metrics, the higher the value, the better the model.

IV. Results and Analysis

A. Landslide Causative Factors Analysis

1) Relative Importance of Causative Factors: The average FIMs of causative factors in different tree-based models are shown in Fig. 11. First, all models showed that the factors of altitude, land use, lithology, NDVI, and STI contribute significantly to landslide modeling, while the factors of plan, profile, distance to road, and TWI are relatively low. It should be noted that the FIM values of plan, profile, distance to road, and soil in the CART are zero because the CART was pruned. Moreover, the importance of STI is very different between the models, specifically, it has the highest FIM of 0.2153 in XGBoost, while it has a relatively low FIM in the other models.

2) Correlation Analysis Between Causative Factors: A visualized heat map of the spearman's rank correlation coefficient between the causative factors is shown in Fig. 12, where blue represents positive correlation, while red indicates negative correlation. It can be seen that the altitude is negatively correlated with land use with a correlation coefficient of –0.51, while it is positively correlated with NDVI and slope with correlation coefficients of 0.45 and 0.52, respectively, indicating that the higher the altitude, the less land use and development. Meanwhile, the denser the vegetation, the steeper the slope. The slope is positively correlated with STI as the correlation coefficient between them is 0.53, and negatively correlated with TWI as the correlation coefficient between them is –0.48, which indicates that the steeper the slope, the higher the STI and the lower the TWI. It can be observed that the correlation between these factors is very low, because all the correlation coefficients are less than the critical value of 0.7, so no factor was eliminated in this study.

B. Training Models and Constructing Landslide Susceptibility Maps

To apply these models for LSM of the study area, the past landslide events were randomly divided according to a common sampling strategy [16], [53], [54], i.e., 70% of landslides (255) for training and the remaining 30% of landslides (109) for testing. In addition, to maintain class balance, the same number of nonlandslide sites (255 and 109) was randomly selected from the landslide-free areas to construct the training and test sets.

In this section, the training dataset was input into the methods, which were implemented in Python under the scikit-learn[7] framework. To automatically obtain an optimal combination of

[7][Online]. Available: https://scikit-learn.org/stable/

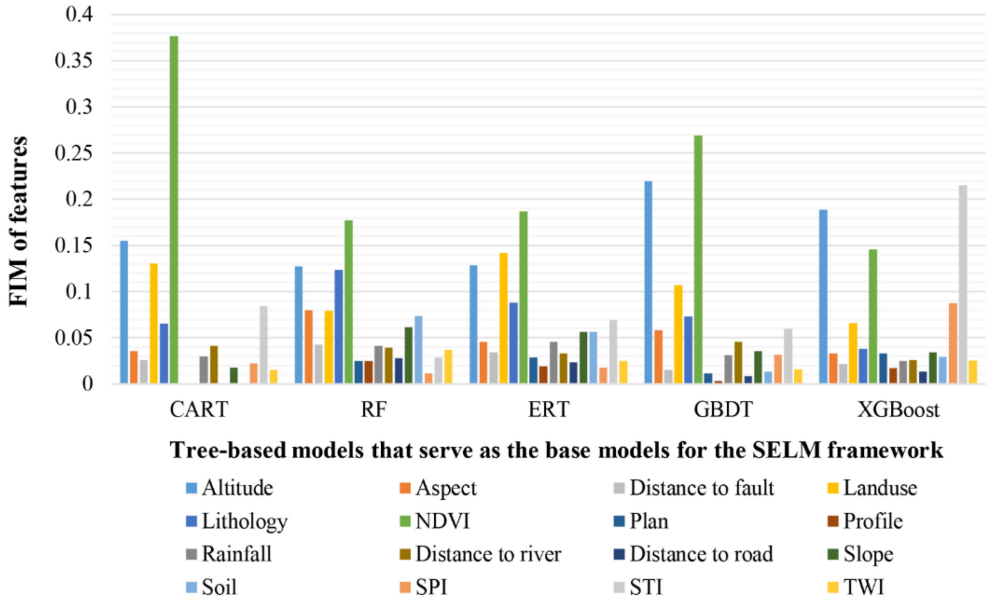


Fig. 11. Feature importance measures (FIMs) of landslide causative factors in different models.

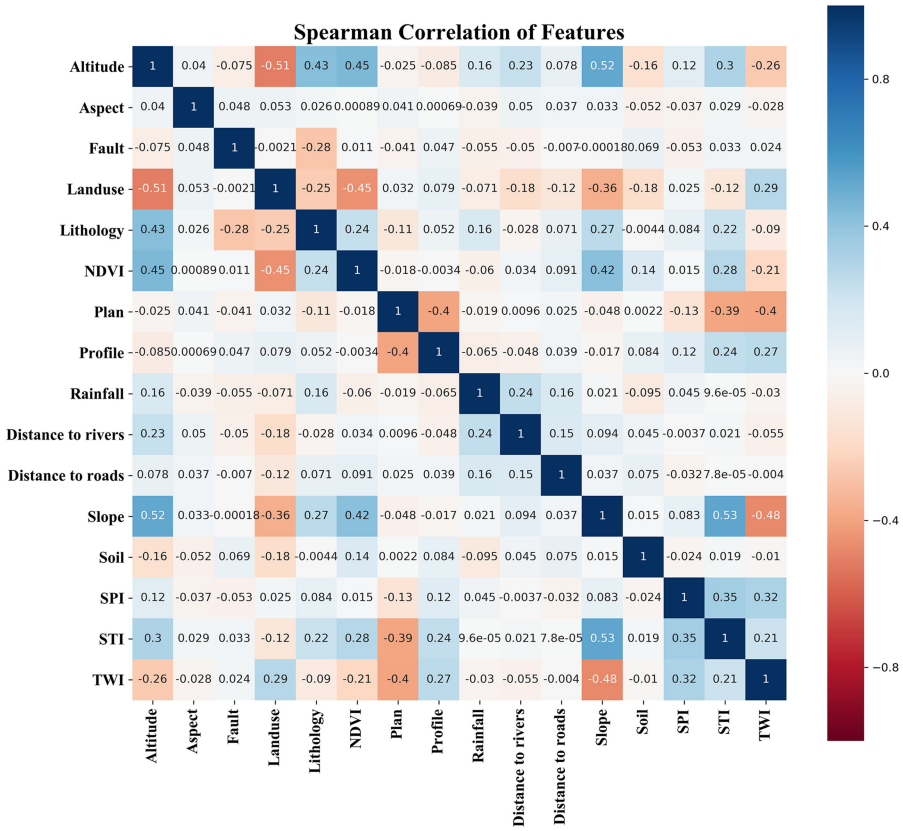


Fig. 12. Spearman's rank correlation coefficient between landslide causative factors.

parameters of these methods, the grid search was used in the scikit-learn framework package to traverse the given method parameter combination and determine the best parameter combination through cross validation. The optimal parameters of all the methods obtained through the above tuning process are listed in Table IV.

After training the landslide prediction models, landslide susceptibility maps were constructed in the form of probability grids in the ArcGIS environment. Each grid cell in the map was assigned a landslide susceptibility index that indicates the probability of landslide output by the methods. For better visualization, the indices were reclassified into five levels: very low,

TABLE IV
PARAMETERS OF DIFFERENT METHODS USED IN THIS STUDY

Methods	Learning Algorithms	Parameters
Single classifier	CART	minimum number of samples required to split an internal node, 2; maximum leaf nodes of tree, 18; maximum depth of the tree, 8;
Bagging	RF	number of trees, 25; maximum depth of the tree, 8; minimum number of samples required to split an internal node, 2;
	ERT	number of trees, 25; maximum depth of the tree, 8; minimum number of samples required to split an internal node, 3;
Boosting	GBDT	number of boosting stages to perform, 50; learning rate, 0.1, maximum depth of the tree, 3; minimum number of samples required to split an internal node, 5;
	XGBoost	number of trees to fit,100; boosting learning rate, 0.1; initial prediction score of all instances, 0.5; maximum depth of the tree, 3;
Stacking	SELM	Base learners: CART, RF, ERT, GBDT, XGBoost (their parameters are the same as above) Meta-learner: Logistic Regression (tolerance for stopping criteria, 0.01; random state, 5; maximum number of iterations taken for the solvers to converge, 8;)
	EFS-SELM	Base learners: EFS-CART, EFS-RF, EFS-ERT, EFS-GBDT, EFS-XGBoost (their parameters are the same as above) Meta-learner: Logistic Regression (the parameters are the same as the Logistic Regression in the SELM framework) *j* mentioned in Section III-*C*: 14

low, moderate, high, and very high, using the commonly used natural breaks method. Landslide susceptibility maps obtained by different methods are illustrated in Fig. 13. Then, to understand the overall pattern of landslide distribution and different classes of landslide susceptible areas, the landslide density distribution was obtained in Fig. 14, which shows the distribution of each class in susceptibility maps and the percentage of landslides in different susceptible classes.

It can be observed that the spatial distribution of the landslide susceptibility maps produced by different prediction methods share some similar rules. For example, the regions with relatively high susceptibility are mainly distributed in the north and south of the study area, and the central region is classified to the relatively low susceptibility class. Furthermore, the historical landslide occurrences are mostly located in very high and the high susceptible areas. Among them, the very high susceptibility class of the map produced by CART occupied 30% of the study area, while the low and high susceptibility classes only account for 0.76% and 5.47% in area. As a single tree classifier, the spatial distribution of the landslide susceptibility map produced by CART is not desirable, but the regional distribution of landslide susceptibility produced by the other ensemble-based methods was more in line with the spatial distribution of landslide.

C. Model Assessment and Comparison

In this study, the performance of the models was assessed using both training and testing sets. The results of five evaluation statistical metrics are listed in Table V. Using the training set, RF had the best fitting ability, followed by ERT, SELM, EFS-SELM, XGBoost, GBDT, and CART. However, the results of the methods using the test set is very different, and EFS-SELM achieved the best performance using the test set, followed by SELM, XGBoost, GBDT, RF, ERT, and CART. As the performance on the test set can better demonstrate the predictive and generalization accuracy of the model than the using the training set, it can be observed that the EFS-SELM method obtained the highest accuracy in landslide prediction, and RF may be overfitted using the training set.

The ROC curve and the calculated AUC used for the overall predictive capability estimation of all the methods are shown in Fig. 15. It can be seen that all the ensemble-based methods obtained satisfactory predictive performance with an AUC above 0.8, and EFS-SELM had the highest AUC value of 0.864, followed by SELM (0.860), XGBoost (0.856), GBDT (0.851), RF (0.841), ERT (0.835), and CART (0.778).

In order to further validate the effectiveness of the proposed EFS-SELM framework, which demonstrated the best performance in the previous experiments, we compared this framework with some traditional machine learning algorithms such as SVM and ANN. The optimal parameters of SVM and ANN were obtained using the grid search that has been mentioned in Section IV-B. Moreover, Table VI lists the results of five evaluation statistical metrics of the proposed framework, SVM, and ANN. It turns out that the proposed EFS-SELM framework achieved highest AUC value (0.864) on test sets, higher than SVM (0.853) and ANN (0.843).

V. DISCUSSION

A. Prediction Performance of Different Methods

Landslide is a very complicated process. So far, in order to accurately evaluate and predict landslide susceptibility, scholars have been trying to explore new methods [55], [56]. Because tree-based models are easy to visualize and are suitable for a small amount of sample data, scholars often apply them to

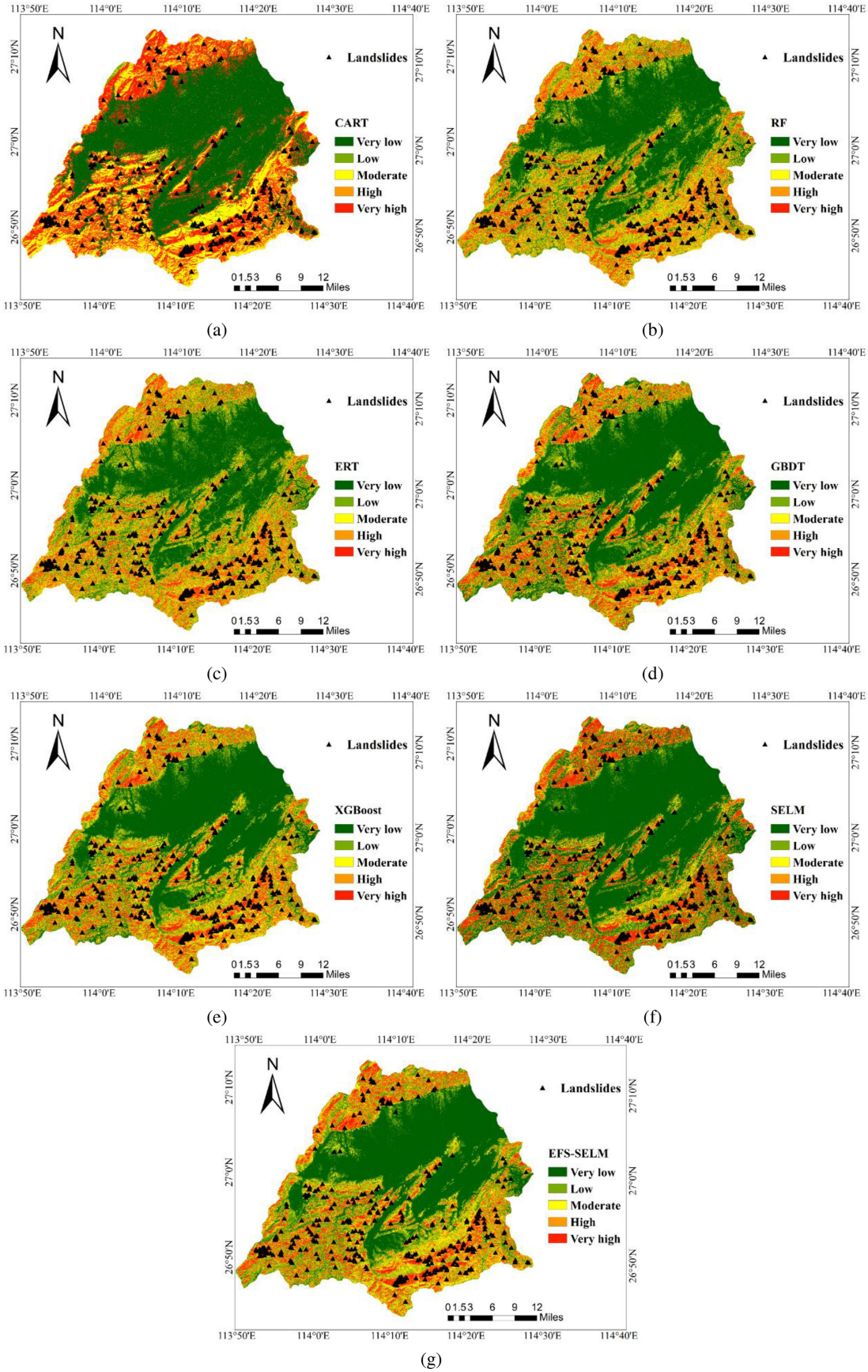


Fig. 13. Landslide susceptibility maps by different methods. (a) CART. (b) RF. (c) ERT. (d) GBDT. (e) XGBoost. (f) SELM. (g) EFS-SELM.

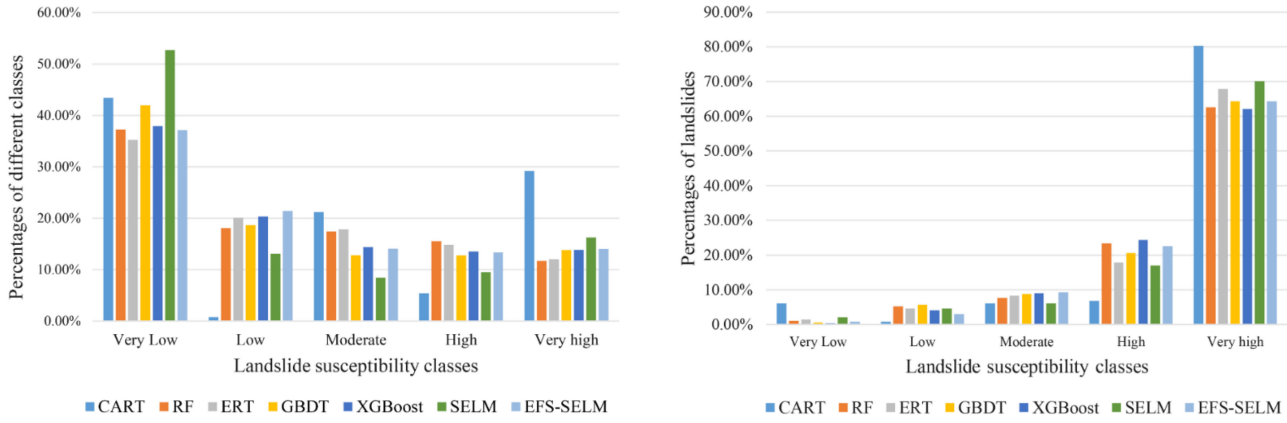


Fig. 14. Percentages of different classes and landslides. (a) Landslide susceptibility classes. (b) Landslides in the corresponding susceptible classes.

TABLE V
STATISTICAL MEASURES OF DIFFERENT METHODS USING TRAINING AND TEST SETS

Dataset	Learning methods	Performance				
		Accuracy	AUC	Recall	Precision	F-measure
Training set	CART	0.822	0.892	0.839	0.811	0.825
	RF	**0.932**	**0.986**	**0.964**	**0.907**	**0.934**
	ERT	0.924	0.983	0.953	0.892	0.927
	GBDT	0.853	0.935	0.894	0.826	0.859
	XGBoost	0.863	0.944	0.904	0.837	0.869
	SELM	0.894	0.962	0.925	0.871	0.897
	EFS-SELM	0.876	0.960	0.922	0.845	0.882
Test set	CART	0.752	0.778	0.826	0.720	0.769
	RF	0.782	0.841	0.863	0.743	0.799
	ERT	0.762	0.835	0.874	0.715	0.786
	GBDT	0.789	0.851	0.862	0.752	0.803
	XGBoost	0.793	0.856	0.844	0.760	0.803
	SELM	0.807	0.860	0.890	0.764	0.822
	EFS-SELM	**0.816**	**0.864**	**0.899**	**0.772**	**0.831**

TABLE VI
STATISTICAL MEASURES OF THE PROPOSED FRAMEWORK AND TRADITIONAL MACHINE LEARNING ALGORITHMS USING TEST SETS

Learning methods	Performance				
	Accuracy	AUC	Recall	Precision	F-measure
EFS_SELM	0.816	0.864	0.899	0.772	0.831
SVM	0.775	0.853	0.853	0.738	0.791
ANN	0.779	0.841	0.817	0.761	0.788

LSM and use the tree visualization to understand the rules of landslide prediction [11]–[13]. However, the generalization and overfitting problems of tree-based methods make landslide prediction more complicated, which may cause uncertainty in the LSM process. In addition, the high complexity of landslide forecasting and the uncertainty of various sources in the modeling process limit the prediction method and reduce the generalization performance of the method. To solve these problems, this study applies ensemble learning methods to improve the generalization accuracy of tree-based methods in landslide susceptibility assessment.

This article presents the SELM framework for LSM, where five tree-based machine learning methods, namely CART, RF, ERT, GBDT, and XGBoost, are combined through a metalearner (LR). Subsequently, this article compares the three commonly used ensemble ideas of bagging, boosting, and stacking for landslide modeling. The experimental results show that ensemble learning methods can improve the prediction performance of landslide modeling. Specifically, the prediction performance based on the representative boosting algorithms (GBDT and XGBoost) is superior to the representative bagging algorithms (RF and ERT). This finding is the same as the previous research [57]. Moreover, the performance of these ensemble methods is better than a single tree (CART).

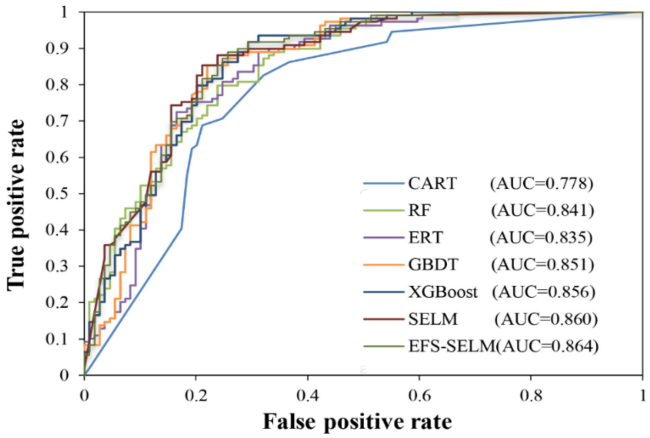


Fig. 15. ROC curves by different methods using the test set.

In addition, among these machine learning ensembles, the stacking algorithm provided the greatest improvement. This can be explained by the fact that in the SELM framework, if the base learner mistakenly learns a specific region in the feature space and leads to a misclassification, the metalearner at the second-layer may classify correctly based on other learners [58]. On the other hand, in order to further explore the potential of the SELM framework, an embedded feature selection process is added to the EFS-SELM framework. This feature selection method can not only reduce the complexity of the model and the dimension of the dataset, but also further explore the advantages of the SELM framework in knowledge discovery and feature extraction. Compared with the SELM framework, the EFS-SELM framework does achieve higher prediction and generalization accuracy.

However, a recent study evaluated and compared the predictive capabilities of SVM hybrid ensemble algorithms (bagging, boosting, stacking) for landslide susceptibility modeling, and the SVM-stacking model was found it has the lowest performance [23]. In this previous relevant research, the use of the stacking method reduced the accuracy of the single classifier SVM from 0.813 to 0.741, probably because the advantages of stacking ensemble method have not been fully explored. Notably, using the stacking ensemble method, we achieved higher prediction performance.

B. Importance of Landslide Causative Factors

The importance of evaluating causative factors is of practical significance for geological experts to analyze the relationship between landslides and environmental variables to prevent them. Machine learning algorithms are increasingly helping decision makers gain new insights [59]. The experimental results show that five causative factors, altitude, NDVI, land use, SPI, and lithology, have an important role in modeling landslide in the study area. In mountainous areas, altitude affects vegetation distribution, land use, and slope, which can also be observed from the correlation analysis mentioned in Section IV-*A*, and altitude further affects the stress value of the slope body, thereby affecting the occurrence of landslides. Vegetation can play an active role in the stability of shallow soil on the slope through the roots, and NDVI can reflect the distribution of vegetation, so NDVI has a strong correlation with the occurrence of shallow landslides [60]. The importance of land use factors illustrates the huge impact of human activities on landslides, which reminds people to carry out more detailed geological survey before land development and utilization in this area. Finally, from the perspective of engineering geology, lithology is the fundamental factor that determines the anti-slip force of a slope, and unfavorable geological tectonic background and characteristics of the rock-soil body are the prerequisite for landslide. Due to the bifurcated nature of tree nodes, some features can only play an important role if other specific feature spaces have been well divided [61]. Therefore, the bifurcation process of trees is similar to the process of judging how various lithologies are prone to landslide under the influence of environmental factors. For example, the slope of sandy soil cannot exceed its internal friction angle, but hard and intact rocks (such as granite and silicalite) can form very steep high slopes without losing their stability. In addition, due to the difference in porosity between different lithological rock and soil mineral particles, the pore water pressure of some susceptible rock-soil bodies will change drastically during rainfall, and the shear strength will decrease, causing slope instability.

C. Reliability From Geological Perspective

Generating the sample dataset containing both positive and negative samples is a primary step before landslide susceptibility modeling. Positive samples are prepared from historical landslides. There are several sampling methods for LSM, including seed cells [62], single pixels [63]–[65], and all pixels [66]. However, by using some sampling methods, mixed types of landslides might affect the modeling outcomes. Therefore, in our study, sampling the centroids of landslide polygons as positive data can significantly mitigate the negative effects of mixed landslide types of rotational and translational landslides.

On the other hand, in this study, the LSM obtained by EFS-SELM has good flexibility and practicability, and it has been determined that one-third of the study areas shows high and very high susceptibility to landslides, which are mainly concentrated in faults, the soft rock mass distribution areas of Devonian Jurassic and Carboniferous, as well as the areas with high altitude and relatively lack of vegetation. From a geological point of view, the main exposed lithologies in highly susceptible areas are sandstone, shale, dolomite, and carbonaceous slate, which have loose geotechnical structure, low shear strength, and weather resistance, and their properties are easy to change under the action of water and prone to landslides. In addition, the magmatic activities in the study area are frequent and have experienced long-term multicycle tectonic movement, resulting in multistage magmatic activities from Nanhua to Cretaceous and forming a wide distribution of granite and a small amount of basic intrusive rocks. In the distribution area of granite, the thickness of clastic rock is relatively large, and the division of clastic rocks is relatively good. The weathering products are mostly granular quartz and clay that are loose in structure, and

they have good water permeability and moisture content, which may cause surface water to penetrate into the soil and fill it with water, and significantly reduce the shear strength of rock-soil that is likely to cause landslides. Furthermore, the folding structure in the study area is relatively developed, showing an "S"-shaped turn. It is an arc-shaped compact-isoclinic fold, sloping westward and axially northeastward, and most of it is damaged by late fractures, so its shape is incomplete. The fold strata are composed of Cambrian, Ordovician and part of Devonian, and the two flanks of the fold strata are prone to landslides.

Overall, the above geological analysis of high landslide susceptibility is consistent with the prediction of the EFS-SELM framework in high-risk areas. Before carrying out land planning and construction in these areas, more professional and detailed geological surveys and engineering prevention are required. Specifically, engineering geology, soil properties, and geotechnical techniques should be considered. In addition, for the "very high" risk area, we recommend engineering measures to enhance slope stability and real-time monitoring of environmental factors (such as rainfall) to help increase the agility and efficiency of disaster prevention measures.

D. Impacts on Disaster Reduction and Management

Effective and accurate space forecasting is highly conducive to the systematic construction of landslide-resistant sustainable human settlements, thereby reducing the probability and vulnerability of poor mountain inhabitants to extreme weather-induced landslides [67], which is one of the sustainable development goals of the 2030 Agenda [68].

The tree-based models are computationally lightweight and easier to extract rules than other black-box models [40], which can help decision makers gain new insights into understanding the landslide disasters. This study explored the potential of ensemble learning methods in improving tree-based classifiers for LSM and proposed an EFS-SELM framework. Such a strategy and framework can free researchers from focusing on the improvement of single statistical algorithm, as it not only is able to combine the advantages of many different algorithms, but also has the potential to be visualized to provide researchers with a machine learning prediction perspective. It turns out that the proposed framework can effectively carry out knowledge discovery, and accurately recognize landslide-prone areas with the highest AUC value of 0.864.

Although the prediction of the EFS-SELM framework should not be considered deterministic, its ability to quickly discover a large number of data patterns reduces the time required to identify susceptible areas and provides a quick and valuable starting point that needs to be supplemented and evaluated by experts. Moreover, the application of the tree-based models and ensemble methods in constructing the macroscopic landslide susceptibility map is very effective, and it is of great significance for people to conduct targeted landslide prevention and control. In addition, it can also be used as a basic tool for land management and planning for future construction projects in such areas. Finally, it is worth emphasizing that the ensemble idea of combining different conventional classifiers provides great potential and possibility for the assessment and mitigation of geo-hazards, which can be effectively used for landslide spatial prediction modeling.

Since research works have shown that improving the technical and scientific capacity to identify, understand, and predict potentially hazardous landslides does not automatically translate into effective practices for landslide risk reduction [69]. Therefore, in order to reduce disaster risk, there is a need to address existing challenges and prepare for future ones by focusing on 1) making full use of advanced remote sensing technology to monitor, assess, and understand disaster risk and share such information; 2) strengthening cooperation and coordination among relevant institutions in disaster risk governance, as well as the full and meaningful participation of relevant stakeholders at appropriate levels [70].

VI. Conclusion

In this study, a novel EFS-SELM framework is proposed to explore the potential of ensemble learning to improve the performance of tree-based classifiers for landslide susceptibility assessment. Base on the experimental results, we can draw some conclusions as follows.

1) For homogeneous ensemble learning, the boosting methods, XGBoost and GBDT, obtained higher prediction accuracies than those of the bagging methods, RF and ERT. Moreover, the prediction accuracies by them are significantly higher than that of the single-tree classier, CART.
2) For heterogeneous ensemble learning, stacking ensemble and embedded feature selection used by the EFS-SELM framework can significantly improve prediction and generalization accuracies of tree-based classifiers for LSM. Therefore, the proposed framework can effectively analyze the relationship between various landslide causative factors, carry out effective feature extraction and knowledge discovery, and accurately recognize landslide-prone areas. The landslide susceptibility maps of EFS-SELM comprehensively consider the influence of all the causative factors, and have a guiding significance as a whole.
3) Finally, it is worth emphasizing that the ensemble idea of combining different conventional classifiers provides great potential and possibility for the assessment and mitigation of geo-hazards, which can be effectively used for landslide modeling.

Acknowledgment

The authors would like to thank the handling editors and two anonymous reviewers for their valuable comments and suggestions, which significantly improved the quality of this article.

References

[1] D. Petley, "Global patterns of loss of life from landslides," *Geology*, vol. 40, no. 10, pp. 927–930, 2012.
[2] S. Lacasse and F. Nadim, *Landslide Risk Assessment and Mitigation Strategy*. Berlin, Germany: Springer-Verlag, 2009, pp. 31–61.
[3] K. T. Chang, A. Merghadi, A. P. Yunus, B. T. Pham, and J. Dou, "Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques," *Sci. Rep.*, vol. 9, no. 1, Aug. 2019, Art. no. 12296.

[4] V. Moosavi and Y. Niazi, "Development of hybrid wavelet packet-statistical models (WP-SM) for landslide susceptibility mapping," *Landslides*, vol. 13, no. 1, pp. 97–114, 2016.

[5] A. Pradhan and Y. Kim, "Evaluation of a combined spatial multi-criteria evaluation model and deterministic model for landslide susceptibility mapping," *Catena*, vol. 140, pp. 125–139, 2016.

[6] A. A. Shahri, J. Spross, F. Johansson, and S. Larsson, "Landslide susceptibility hazard map in southwest Sweden using artificial neural network," *Catena*, vol. 183, 2019, Art. no. 104225.

[7] L. V. Lucchese, G. G. de Oliveira, and O. C. Pedrollo, "Attribute selection using correlations and principal components for artificial neural networks employment for landslide susceptibility assessment," *Environ. Monitoring Assessment*, vol. 192, no. 2, 2020, Art. no. 129.

[8] Y. Zhao, R. Wang, Y. Jiang, H. Liu, and Z. Wei, "GIS-based logistic regression for rainfall-induced landslide susceptibility mapping under different grid sizes in Yueqing, Southeastern China," *Eng. Geol.*, vol. 259, 2019, Art. no. 105147.

[9] Q. Wang, Y. Wang, R. Niu, and L. Peng, "Integration of information theory, K-means cluster analysis and the logistic regression model for landslide susceptibility mapping in the Three Gorges Area, China," *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 938.

[10] R. P. Riegel *et al.*, "Assessment of susceptibility to landslides through geographic information systems and the logistic regression model," *Natural Hazards*, vol. 103, pp. 497–511, 2020.

[11] S. J. Park, C. W. Lee, S. Lee, and M. J. Lee, "Landslide susceptibility mapping and comparison using decision tree models: A case study of Jumunjin Area, Korea," *Remote Sens.*, vol. 10, no. 10, Oct. 2018, Art. no. 1545.

[12] P. Tsangaratos and I. Ilia, "Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece," *Landslides*, vol. 13, no. 2, pp. 305–320, 2016.

[13] B. T. Pham, D. T. Bui, and I. Prakash, "Landslide susceptibility modelling using different advanced decision trees methods," *Civil Eng. Environ. Syst.*, vol. 35, nos. 1-4, pp. 139–157, Oct. 2018.

[14] Y. Huang and L. Zhao, "Review on landslide susceptibility mapping using support vector machines," *Catena*, vol. 165, pp. 520–529, 2018.

[15] Y. Wang, H. Duan, and H. Hong, "A comparative study of composite kernels for landslide susceptibility mapping: A case study in Yongxin County, China," *Catena*, vol. 183, 2019, Art. no. 104217.

[16] Z. L. Chang *et al.*, "Landslide susceptibility prediction based on remote sensing images and GIS: Comparisons of supervised and unsupervised machine learning models," *Remote Sens.*, vol. 12, no. 3, Feb. 2020, Art. no. 502.

[17] P. Reichenbach, M. Rossi, B. D. Malamud, M. Mihir, and F. Guzzetti, "A review of statistically-based landslide susceptibility models," *Earth-Sci. Rev.*, vol. 180, pp. 60–91, 2018.

[18] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, vol. 1857, J. Kittler and F. Roli, Eds. Berlin, Germany: Springer-Verlag, 2000, pp. 1–15.

[19] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," in *Proc. Asian Conf. Intell. Inf. Database Syst.*, 2010, pp. 340–350.

[20] P. Kadavi, C.-W. Lee, and S. Lee, "Application of ensemble-based machine learning models to landslide susceptibility mapping," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1252.

[21] B. T. Pham, D. Tien Bui, and I. Prakash, "Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: A comparative study," *Geotech. Geological Eng.*, vol. 35, no. 6, pp. 2597–2611, 2017.

[22] H. Hong *et al.*, "Landslide susceptibility mapping using J48 decision tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)," *Catena*, vol. 163, pp. 399–413, 2018.

[23] J. Dou *et al.*, "Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan," *Landslides*, vol. 17, no. 3, pp. 641–658, Mar. 2020.

[24] E. Kutlug Sahin and I. Colkesen, "Performance analysis of advanced decision tree-based ensemble learning algorithms for landslide susceptibility mapping," *Geocarto Int.*, pp. 1–23, 2019.

[25] H. Hong, J. Liu, and A. X. Zhu, "Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble," *Sci. Total Environ.*, vol. 718, 2020, Art. no. 137231.

[26] B. T. Pham *et al.*, "Ensemble modeling of landslide susceptibility using random subspace learner and different decision tree classifiers," *Geocarto Int.*, pp. 1–23, 2020.

[27] Y. Zhang, M. Li, S. Han, Q. Ren, and J. Shi, "Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms," *Sensors (Basel)*, vol. 19, no. 18, Sep. 2019, Art. no. 3914.

[28] W. Sun, "River ice breakup timing prediction through stacking multi-type model trees," *Sci. Total Environ.*, vol. 644, pp. 1190–1200, Dec. 2018.

[29] Z. Fang, Y. Wang, L. Peng, and H. Hong, "Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping," *Comput. Geosci.*, vol. 139, 2020, Art. no. 104470.

[30] Y. Wang, Z. Fang, and H. Hong, "Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China," *Sci. Total Environ.*, vol. 666, pp. 975–993, 2019.

[31] Y. Wang *et al.*, "A hybrid GIS multi-criteria decision-making method for flood susceptibility mapping at Shangyou, China," *Remote Sens.*, vol. 11, no. 1, 2019, Art. no. 62.

[32] D. T. Bui, T.-C. Ho, B. Pradhan, B.-T. Pham, V.-H. Nhu, and I. Revhaug, "GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks," *Environmental Earth Sci.*, vol. 75, no. 14, 2016, Art. no. 1101.

[33] B. T. Pham, B. Pradhan, D. T. Bui, I. Prakash, and M. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environmental Model. Softw.*, vol. 84, pp. 240–250, 2016.

[34] U. Đurić, M. Marjanović, Z. Radić, and B. Abolmasov, "Machine learning based landslide assessment of the Belgrade metropolitan area: Pixel resolution effects and a cross-scaling concept," *Eng. Geol.*, vol. 256, pp. 23–38, 2019.

[35] D. Kumar, M. Thakur, C. S. Dubey, and D. P. Shukla, "Landslide susceptibility mapping & prediction using support vector machine for Mandakini River Basin, Garhwal Himalaya, India," *Geomorphology*, vol. 295, pp. 115–125, 2017.

[36] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[37] A. Rosi *et al.*, "The new landslide inventory of Tuscany (Italy) updated with PS-InSAR: Geomorphological features and landslide distribution," *Landslides*, vol. 15, no. 1, pp. 5–19, 2017.

[38] V. H. Nhu *et al.*, "Shallow landslide susceptibility mapping by random forest base classifier and its ensembles in a semi-arid region of Iran," *Forests*, vol. 11, no. 4, Apr. 2020, Art. no. 421.

[39] J. Dou *et al.*, "Evaluating GIS-based multiple statistical models and data mining for earthquake and rainfall-induced landslide susceptibility using the LiDAR DEM," *Remote Sensing*, vol. 11, no. 6, p. 638, 2019.

[40] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[41] D. a. Glenn and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000.

[42] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[44] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[45] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[46] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[47] J. H. Friedman and J. J. Meulman, "Multiple additive regression trees with application in epidemiology," *Statist. Med.*, vol. 22, no. 9, pp. 1365–1381, 2003.

[48] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018.

[49] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[50] A. M. S. Pradhan and Y.-T. Kim, "Spatial data analysis and application of evidential belief functions to shallow landslide susceptibility mapping at Mt. Umyeon, Seoul, Korea," *Bull. Eng. Geol. Environ.*, vol. 76, no. 4, pp. 1263–1279, 2016.

[51] Y. Wang, Z. Fang, M. Wang, L. Peng, and H. Hong, "Comparative study of landslide susceptibility mapping with different recurrent neural networks," *Comput. Geosci.*, vol. 138, 2020, Art. no. 104445.
[52] A. Akgun, "A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir, Turkey," *Landslides*, vol. 9, no. 1, pp. 93–106, 2012.
[53] M. I. Sameen, B. Pradhan, D. T. Bui, and A. M. Alamri, "Systematic sample subdividing strategy for training landslide susceptibility models," *CATENA*, vol. 187, 2020, Art. no. 104538.
[54] Y. Achour and H. R. Pourghasemi, "How do machine learning techniques help in increasing accuracy of landslide susceptibility maps?" *Geosci. Frontiers*, vol. 11, no. 3, pp. 871–883, 2020.
[55] W. Chen *et al.*, "Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques," *Geomorphology*, vol. 297, pp. 69–85, 2017.
[56] L. Zhu *et al.*, "Landslide susceptibility prediction modeling based on remote sensing and a novel deep learning algorithm of a cascade-parallel recurrent neural network," *Sensors*, vol. 20, no. 6, Mar. 2020, Art. no. 1576.
[57] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, and H. Hong, "Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping," *CATENA*, vol. 187, 2020, Art. no. 104396.
[58] W. Sun and Z. Li, "Hourly PM2.5 concentration forecasting based on feature extraction and stacking-driven ensemble model for the winter of the Beijing-Tianjin-Hebei area," *Atmospheric Pollution Res.*, vol. 11, no. 6, pp. 110–121, 2020.
[59] R. C. Deo and M. Şahin, "Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia," *Atmospheric Res.*, vol. 153, pp. 512–525, 2015.
[60] B. Yu, F. Chen, and C. Xu, "Landslide detection based on contour-based deep learning framework in case of national scale of Nepal in 2015," *Comput. Geosci.*, vol. 135, 2020, Art. no. 104388.
[61] S. A. Naghibi *et al.*, "Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential," *Environmental Monitoring Assessment*, vol. 191, no. 4, Mar. 2019, Art. no. 248.
[62] M. L. Süzen and V. Doyuran, "Data driven bivariate landslide susceptibility assessment using geographical information systems: A method and application to Asarsuyu catchment, Turkey," *Eng. Geol.*, vol. 71, no. 3, pp. 303–321, 2004.
[63] P. M. Atkinson and R. Massari, "Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy," *Comput. Geosci.*, vol. 24, no. 4, pp. 373–385, 1998.
[64] M. Van Den Eeckhaut, T. Vanwalleghem, J. Poesen, G. Govers, G. Verstraeten, and L. Vandekerckhove, "Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium)," *Geomorphology*, vol. 76, no. 3, pp. 392–410, 2006.
[65] D. Piacentini *et al.*, "Statistical analysis for assessing shallow-landslide susceptibility in South Tyrol (south-eastern Alps, Italy)," *Geomorphology*, vol. 151, pp. 196–206, 2012.
[66] L. Ayalew and H. Yamagishi, "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan," *Geomorphology*, vol. 65, no. 1, pp. 15–31, 2005.
[67] A. Depicker *et al.*, "The added value of a regional landslide susceptibility assessment: The western branch of the East African Rift," *Geomorphology*, vol. 353, 2020, Art. no. 106886.
[68] A. Aitsi-Selmi *et al.*, "Reflections on a science and technology agenda for 21st century disaster risk reduction," *Int. J. Disaster Risk Sci.*, vol. 7, no. 1, pp. 1–29, 2016.
[69] A. Tozier de la Poterie and M.-A. Baudoin, "From Yokohama to Sendai: Approaches to participation in international disaster risk reduction frameworks," *Int. J. Disaster Risk Sci.*, vol. 6, no. 2, pp. 128–139, 2015.
[70] J. Weichselgartner and P. Pigeon, "The role of knowledge in disaster risk reduction," *Int. J. Disaster Risk Sci.*, vol. 6, no. 2, pp. 107–116, 2015.

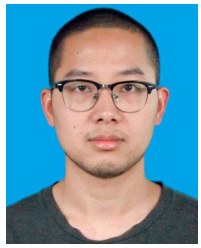
Jiahui Song received the B.E. degree in geographic information science from Yangtze University, Jingzhou, China, in 2018. She is currently working toward the M.S. degree in the China University of Geosciences, Wuhan, China.

Her current research interests include landslide susceptibility mapping as well as remote sensing applications.

Yi Wang (Member, IEEE) received the B.S. degree in printing engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He is currently a Professor with the Institute of Geophysics and Geomatics, China University of Geosciences (CUG), Wuhan. He is the Head of the Department of Geoinformatics. His research interests include remote sensing technology and application, geoinformation data mining, and environmental impact assessment.

Dr. Wang is a member of Geological Society of China (GSC) and Chinese Association of Automation (CAA). In 2019, he was named CUG Outstanding Young Talent.



Zhice Fang received the B.E. degree in geoinfomatics, in 2017, from the China University of Geosciences, Wuhan, China, where he is currently working toward the Ph.D. degree.

His current research interests include natural disaster susceptibility mapping as well as remote sensing applications.

Ling Peng received the Ph.D. degree in earth exploration and information technology from the China University of Geosciences, Wuhan, China, in 2013.

Since 2013, he has been with the China Institute of Geo-Environment Monitoring, Beijing, China, where he is currently a Senior Engineer. His research interests include remote sensing applications for geohazard prevention and geo-environment protection.

Haoyuan Hong received the B.S. degree in geographic information system and the M.S. degree in climate system and global change from Nanjing University of Information Science and Technology, Nanjing, China, in 2007 and 2011, respectively.

Since 2020, he has been with the Department of Geography and Regional Research, University of Vienna, Vienna, where he is currently a Lecturer of Geomorphology. His research interests include machine learning, GIS, and natural hazard assessment.