



Landslide Susceptibility Prediction Based on Positive Unlabeled Learning Coupled With Adaptive Sampling

Zhice Fang , Yi Wang , *Member, IEEE*, Ruiqing Niu, and Ling Peng

Abstract—Many studies consider landslide susceptibility prediction as a binary classification problem when using machine learning methods, which requires both landslide and nonlandslide samples for modeling. Nevertheless, there are only landslide and unlabeled areas in the real world, and directly considering unlabeled areas as nonlandslide areas may cause bias and incorrect label assignment. In this article, we present a positive unlabeled learning method coupled with adaptive sampling and random forest (AdaPU-RF) to predict landslide susceptibility in the Three Gorges Reservoir area, China. This method can make full use of the landslide and nonlandslide information contained in unlabeled areas. Experimental results show that the AdaPU-RF method achieves desirable prediction outcomes in terms of accuracy analysis, sensitivity analysis, and uncertainty analysis. Overall, the application of AdaPU-RF provides a new perspective for landslide susceptibility prediction, and can be recommended for other areas with similar geo-environmental conditions.

Index Terms—Adaptive sampling, landslide susceptibility prediction (LSP), positive unlabeled (PU) learning, sensitivity analysis, uncertainty analysis.

I. INTRODUCTION

LANDSLIDES are one of the most common and destructive geological disasters worldwide. According to the Emergency Events Database, 761 major landslide disasters occurred from 1900 to 2020, causing 67 058 deaths, approximately 14.6 million people affected and economic loss of about 10.9 billion dollars [1]. As a key step in landslide risk assessment, landslide susceptibility prediction (LSP) can predict where landslides are likely to occur and the likelihood of occurrence [2].

Manuscript received June 3, 2021; revised July 25, 2021, September 25, 2021, and October 25, 2021; accepted October 26, 2021. Date of publication November 8, 2021; date of current version November 24, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61271408 and Grant 41602362, in part by the State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences under Grant GBL12107, and in part by the Fundamental Research Funds for National Universities, China University of Geosciences (Wuhan). (*Corresponding author: Yi Wang.*)

Zhice Fang and Ruiqing Niu are with the Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China (e-mail: xmb123@163.com; rqnium@163.com).

Yi Wang is with the Institute of Geophysics and Geomatics and State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan 430074, China (e-mail: cug.yi.wang@gmail.com).

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

Digital Object Identifier 10.1109/JSTARS.2021.3125741

In the past few decades, different methods have been proposed for LSP, such as qualitative methods, physically based methods, and statistical methods. The qualitative methods are subjective and rely on the ability of experts to assess actual and potential landslides [3], [4]. The physically based methods perform mathematical modeling based on the controlling mechanisms of landslides, which require detailed geotechnical data [5]–[7]. The statistical methods are implemented by evaluating the relationship between landslide occurrence and related influencing factors, such as multicriteria decision analysis [8]–[10], bivariate analysis [11], [12], entropy-based methods [13], [14], and weight of evidence [15]–[17]. Recently, machine learning methods have flourished for LSP due to the improvement of computer science and the accessibility of high-quality data [18]. The most commonly used machine learning methods include logistic regression (LR) [19], [20], decision tree [21], support vector machine (SVM) [22]–[25], random forest (RF) [26], [27], artificial neural networks [28], [29], and deep learning methods [30]–[35].

When applying machine learning methods, most researchers consider LSP as a binary classification problem to distinguish whether an area will be landslide and predict the possibility of landslide occurrence. Consequently, positive (landslide) and negative (nonlandslide) samples are required for modeling. Positive samples are collected from known landslide locations identified by field investigation, historical records, and interpretation of remote sensing images. As for negative samples, most researchers consider the areas located outside landslide polygons as nonlandslide areas, and select negative samples from these areas [36]–[39]. However, the above-mentioned process has some disadvantages. First, the areas outside landslides are unlabeled areas, and we cannot know the true labels of samples. Therefore, it may cause bias when we directly regard unlabeled areas as nonlandslide areas and then use binary classification methods for prediction. Secondly, the unlabeled areas include nonlandslide areas and potential landslide areas. The nonlandslide sampling procedure mentioned above may select wrong samples, and cause incorrect and unreliable susceptibility prediction results. Therefore, it is necessary to find an appropriate method to solve the above problems.

Positive unlabeled (PU) learning can learn classifiers from positive and unlabeled samples [40]. The application scenarios of PU learning include many real-world classification problems,

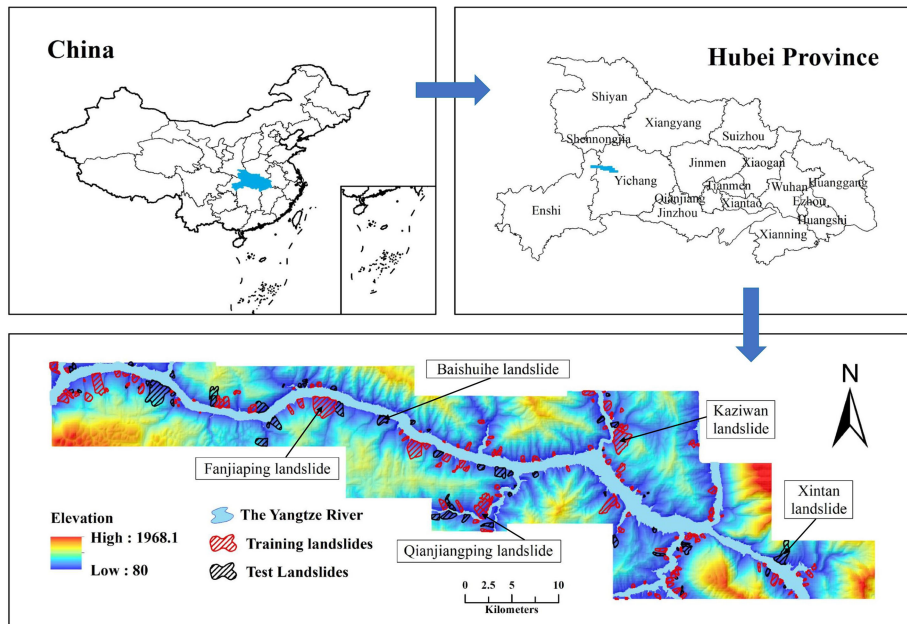


Fig. 1. Location of the study area.

such as medical diagnosis [41], fake reviews detection [42], and fault detection [43]. In fields related to natural disasters, we can only obtain disaster areas and unknown areas from the geographic space. Therefore, LSP is more precisely equivalent to a PU learning problem, rather than a binary classification problem.

Wu *et al.* [44] used a bagging-based PU learning method that integrates tree classifiers (PU-BaggingDT) to predict landslide susceptibility and achieved higher accuracy than traditional machine learning methods. The PU-BaggingDT method uses bootstrap sampling to reduce the instability caused by the non-landslide sampling process. However, it treats all unlabeled samples as nonlandslide data, so there is a problem of incorrect label assignment. Yang *et al.* [45] used a wrapper-based adaptive sampling PU learning method in synthetic and real datasets. The method can iteratively select reliable negative samples from unlabeled sets to train the model. However, it ignores the positive information hidden in the unlabeled set. In addition, although PU learning is theoretically more suitable for solving LSP problems, few studies have attempted to use this technology.

To fill the knowledge gap of LSP, this study aims to use a PU learning method coupled with adaptive sampling and random forest (AdaPU-RF) to predict landslide susceptibility in the Three Gorges Reservoir area, China. The main contributions of this study are summarized as follows. First, we explore the application potential of PU learning for predicting landslide susceptibility and introduce a PU learning strategy that considers the unlabeled areas as nonlandslide areas contaminated by hidden landslides. Compared to the PU-BaggingDT method proposed by Wu *et al.* [44], it can avoid the problem of incorrect label assignment. Second, the AdaPU-RF methods are proposed for predicting landslide susceptibility, which can make full use of

the landslide and nonlandslide information contained in unlabeled areas.

The rest of this article is organized as follows. Section II introduces the study area. Section III presents the data used in experiments and explains the proposed AdaPU-RF method. Section IV analyzes the landslide susceptibility results from accuracy, sensitivity, and uncertainty. In Section V, we discuss the susceptibility prediction results of the proposed method. Finally, Section VI concludes the article.

II. STUDY AREA

The study area is located in the Zigui-Badong section of the Three Gorges Reservoir area, China. The area is 446.32 km² and its altitude ranges from 80 to 1968.1 m (see Fig. 1). The mainstream of the Yangtze River runs from west to east, with a flow path about 80 km. Moreover, the secondary tributaries of the Yangtze River, such as Yandu River, Qinggan River, Xiangxi River, and Jiuwanxi River, are staggered to form a dendritic hydrological network with a river network density of 1.2 kilometers per square kilometer. The farmland accounts for 42.73% of the entire study area and the residential accounts for only 5.16%.

The stratum of the study area is generally intact from the Sinian to Quaternary, extending from east to west. Specifically, the lithological categories of limestone, dolomite, and silicalite are distributed in the Miaohé-Xiangxi section with the stratum from the Sinian to the Lower Triassic, and the lithological categories of sandstone, shale, mudstone, and marlstone, which are prone to landslide occurrence, are distributed in the Xiangxi-Badong section with the stratum from the Middle Triassic to Jurassic.

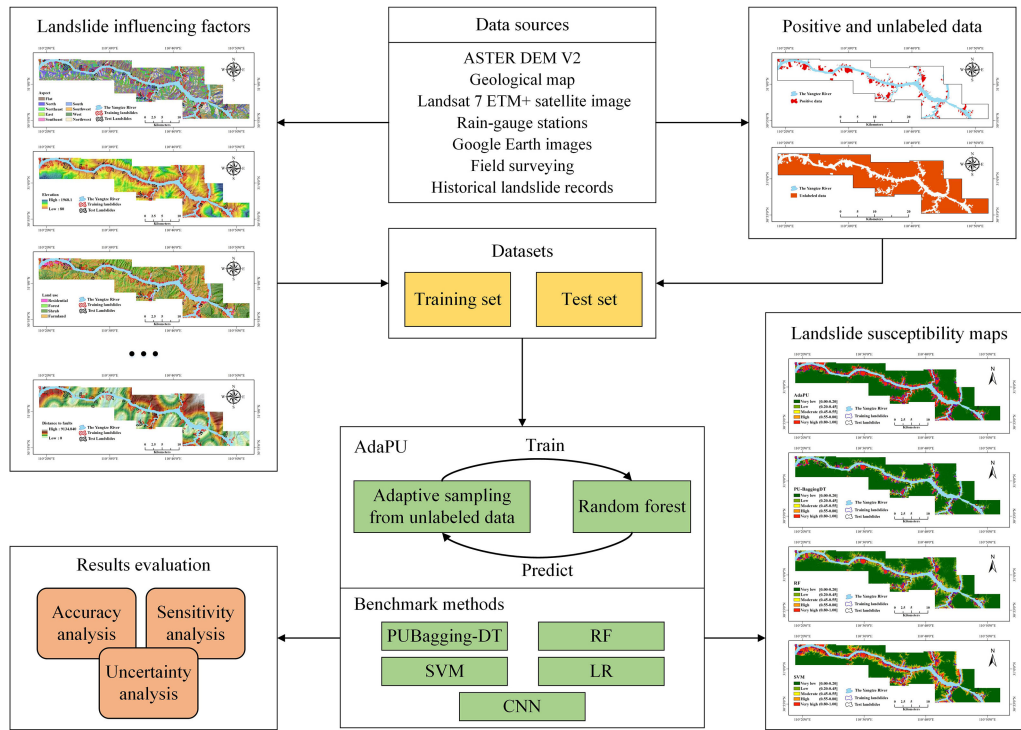


Fig. 2. Flowchart of the study.

The construction of the Three Gorges Reservoir greatly influences the natural environment of the study area. The slope changes in the dredging process of large-scale excavation and abandoned slag severely destroy the morphological structure and stress balance of the original natural slope, causing geological disasters. During the impoundment period, the periodic fluctuation of water level adversely influences the rock and soil mass near the bank slope. Reservoir storage has significantly increased the number of geological disasters in this area, especially landslide disasters [46]. Therefore, it is necessary to obtain accurate landslide susceptibility maps for disaster management and land resource planning in the Three Gorges Reservoir area.

III. MATERIAL AND METHODOLOGY

The flowchart of our study is illustrated in Fig. 2. First, we extracted landslide influencing factors, positive data and unlabeled data from multiple data sources. Then, we built landslide susceptibility models using the training set. Next, the proposed AdaPU-RF method was compared with five benchmark methods in terms of accuracy, sensitivity, and uncertainty. Finally, we predicted landslide susceptibility of each grid cell and output landslide susceptibility maps.

A. Landslide Inventory Map

An accurate landslide inventory map is particularly important for LSP. In this study, 196 landslide polygons were identified through field surveys, historical landslide records, and visual interpretation of Google Earth images. Fig. 1 shows the spatial distribution of landslides. Among these landslides, the smallest

Kuihua street landslide has an area of 2068.8 m², and the largest Fanjiaping landslide is about 1.51 km². The 196 landslide polygons were randomly divided into two parts: 70% (137 polygons) for training and the remaining 30% (59) for testing.

B. Landslide Influencing Factors

The selection of landslide influencing factors is important for constructing landslide susceptibility models [2]. Many scholars have conducted LSP in the Three Gorges Reservoir area [47]–[51]. In this study, 13 landslide influencing factors were selected for modeling based on previous publications and the characteristics of the study area, including ten continuous factors and three discrete factors. The continuous factors are elevation, plan curvature, profile curvature, slope, terrain position index (TPI), topographic wetness index (TWI), distance to faults, normalized difference vegetation index (NDVI), rainfall, and distance to rivers. The discrete factors include aspect, land use, and stratum. Fig. 3 shows the thematic maps of landslide susceptibility factors.

The elevation determines the distribution of the free surface and directly influences the movement of the landslide [52]. The plan curvature reflects the surface runoff, and the profile curvature reflects the slope shape that affects the risk of landslides [51], [53]. Slope is a key factor because landslides only occur in sloped terrain. In addition, the slope can directly or indirectly reflect the surface runoff, vegetation characteristics, and the stress distribution on the slope [37]. TPI is defined as the difference between the target central grid cell and its surrounding cells, and can measure the topographic slope position [49]. The TWI determines the dry and humid conditions of soil

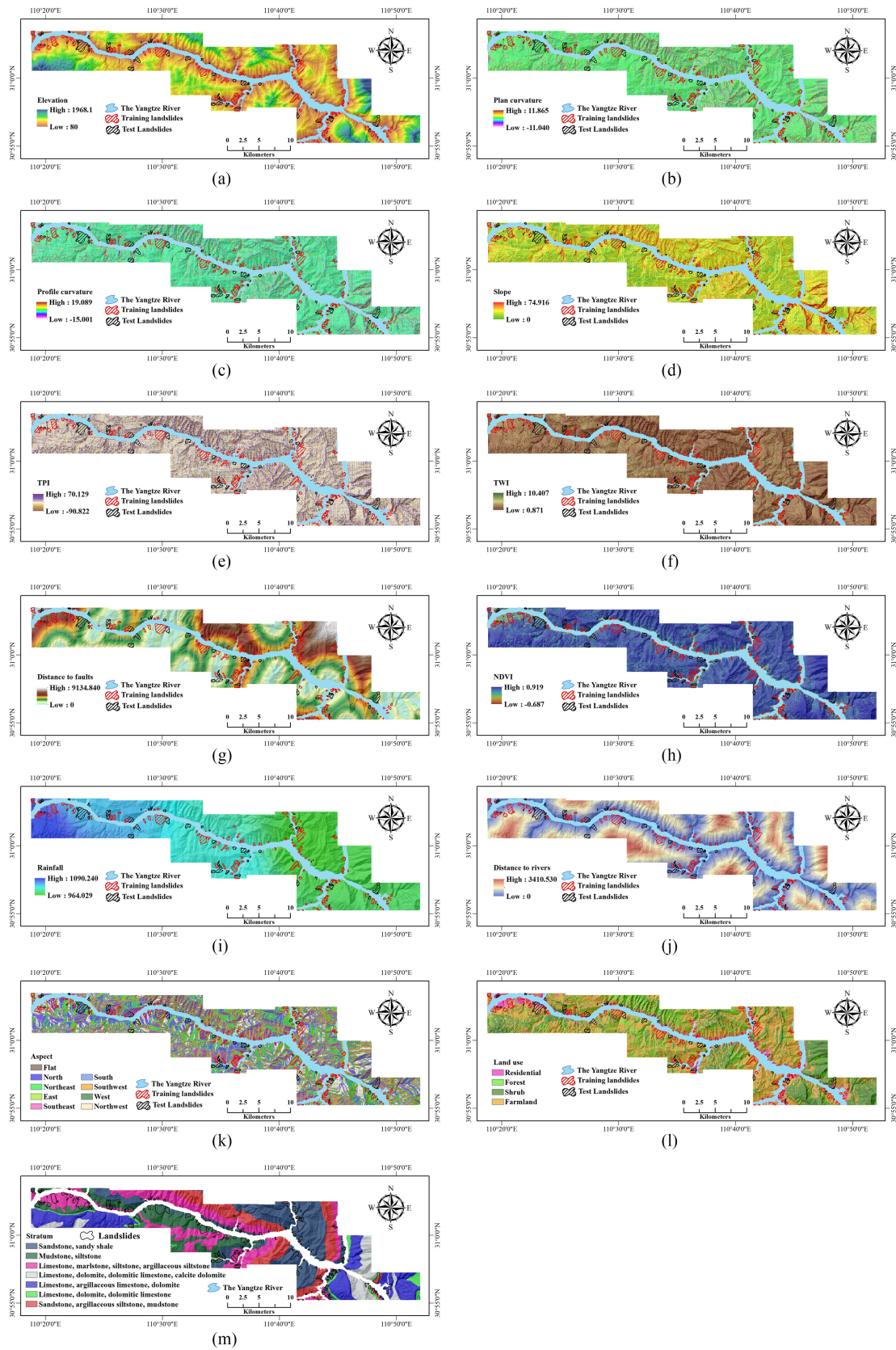


Fig. 3. Thematic maps of landslide influencing factors. (a) Elevation. (b) Plan curvature. (c) Profile curvature. (d) Slope. (e) TPI. (f) TWI. (g) Distance to faults. (h) NDVI. (i) Rainfall. (j) Distance to rivers. (k) Aspect. (l) Land use. (m) Stratum.

moisture, which strongly influences the occurrence of landslides [54]. The distance to faults and stratum are selected because geological conditions play an important role in the formation of landslides [51], [55]. NDVI reflects the growth conditions of green vegetation in an area [56]. Vegetation can retain water and reduce rain erosion on slopes. Rainfall is one of the main external forces that cause slope deformation and destruction [57]–[59]. The Three Gorges Reservoir has a great impact on landslides in this area, and rivers affect the stability of the slope by weakening the resistance of the front edge and increasing the open surface. The aspect determines the intensity of solar radiation and then affects slope evaporation and erosion [19]. Land use reflects the intensity of human activities, especially the construction of the Three Gorges Dam greatly affects the stability of landslides in this study area [46], [60]

The elevation, plan curvature, profile curvature, slope, TPI, TWI, and river networks were calculated from the ASTER GDEM V2 data. The fault network and stratum were extracted from the geological map provided by Hubei Geological Bureau.¹ The Landsat 7 ETM+ satellite images were used to generate the NDVI and land use factors. The average annual rainfall factor was constructed using the precipitation data during 2003–2010 at seven rainfall stations.

C. Adaptive Sampling

The adaptive sampling technique selects samples based on the information learned from previous surveys [61]. In this sense, the sampling design is adaptive and the sampling manner may change based on previous observations. In our study, the adaptive sampling used in the proposed method is an extension of the bootstrap sampling technique [45], [62]. Different from the bootstrap sampling that randomly selects samples with replacement, adaptive sampling selects a sample with replacement according to the probability associated with each sample in D (which is the original sample set), and repeats n times. Finally, we get n samples for further modeling.

D. Modeling Process of AdaPU-RF

Considering a series of influencing factors, landslide susceptibility refers to the possibility of landslide occurrence in a given area [2]. We aim to predict the landslide susceptibility of each grid cell in the study area. It is assumed that the input data of landslide susceptibility models are organized in raster format. Each grid cell in the study area represents a sample containing factors and landslide information. The data extracted from the actual landslide polygon are positive samples. Grid cells located outside the landslide polygon are considered as unlabeled samples. In fact, the unlabeled areas contain both landslide and nonlandslide samples, but the true label of these samples is unknown. Moreover, nonlandslide areas are much larger than landslide areas in the real world. Therefore, unlabeled data can be regarded as negative (nonlandslide) data contaminated by hidden positive (landslide) samples. In this study, we treat LSP as a PU learning problem to predict landslide susceptibility based on landslide and unlabeled datasets.

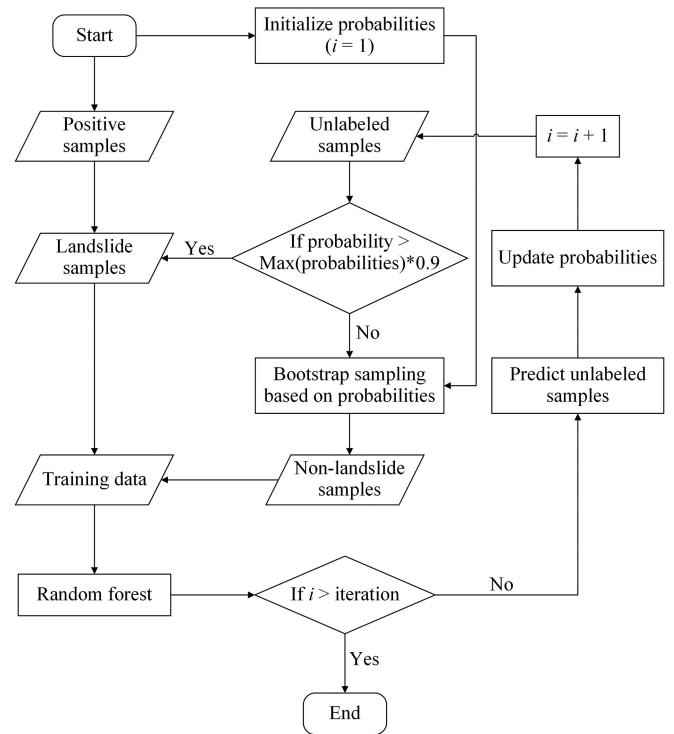


Fig. 4. Detailed procedure of the AdaPU-RF method.

In AdaPU-RF, we simplify PU learning into a traditional binary classification problem, which repeatedly selects reliable positive samples and negative samples from unlabeled data to train the landslide susceptibility model. Let P (label $y = 1$) and U denote positive samples and unlabeled samples, respectively. In the first iteration, a negative sample N is selected with equal probability from U and combined with P as follows:

$$[D^1, y] = [P, y = 1] \cup [N^1, y = 0] \quad (1)$$

where $N^1 \subset U$ and the superscript 1 denotes the first iteration index; D^1 are the sample data used for PU learning in the first iteration. The learned model is expressed as follows:

$$p(y|x) = h(x; [D^1, y]). \quad (2)$$

The learned model is used to predict the probability of unlabeled samples belonging to the positive class. For the next iteration, a sample with a higher probability of $p(y = 1|U)$ in U is selected as the newly added positive data L . A negative sample N is selected from U with replacement according to the probability of $1 - p(y = 1|U)$. Thus, the training data for PU modeling are updated as follows:

$$[D^i, y] = [P, y = 1] \cup [L^i, y = 1] \cup [N^i, y = 0] \quad (3)$$

where i represents the current iteration of sampling. During the modeling process, the training data are updated according to (3). Once the iteration is completed, the RF method is trained based on the optimized training set, and then used to predict landslide susceptibility. Generally, the AdaPU-RF method can make full use of unlabeled information and repeatedly select high-quality

¹[Online]. Available: <http://dzj.hubei.gov.cn>

TABLE I
PARAMETERS SETTING OF DIFFERENT METHODS

Methods	Parameters	Search space	Final setting
AdaPU-RF	Iterations	[1, 2, 3, ..., 15]	13
	Number of trees	[10, 20, 30, ..., 100] and [100, 150, 200, ..., 500]	10
PU-BaggingDT		Wu, et al. [44]	
RF	Number of trees	[10, 20, 30, ..., 100] and [100, 150, 200, ..., 500]	10
SVM	C	[0.1, 1, 10, 100, 1000]	1000
	gamma	[10, 1, 0.1, 0.001, 0.0001]	0.001
LR	Penalty	[L1, L2]	L2
	C	[0.001, 0.01, 0.1, 1, 10, 100]	0.1
CNN		Wang, et al. [32]	

samples for modeling. Fig. 4 shows the detailed process of the proposed AdaPU-RF method.

IV. RESULTS

A. Model Construction

In this study, we applied AdaPU-RF and five benchmark methods of PU-BaggingDT, RF, SVM, LR, and convolutional neural network (CNN) to predict landslide susceptibility. PU-BaggingDT is a PU learning method that was recently proposed to predict landslide susceptibility [44]. RF, SVM, and LR are popular and robust machine learning methods, which have been widely used for LSP [26], [47]. CNN is a typical and powerful deep learning method.

Initially, we converted all landslide influencing factors into a raster format with a spatial resolution of 30 m. A normalization process was used to eliminate bias that existed in continuous factors. The one-hot encoding procedure was applied to quantify these discrete influencing factors. For the AdaPU-RF and PU-BaggingDT methods, we used 70% of landslides and all unlabeled areas for modeling, and used the remaining 30% of landslides to test model performance. For RF, SVM, LR, and CNN, we randomly selected 70% of the landslides for training and the remaining 30% for testing. Meanwhile, the same number of nonlandslide grid cells was randomly sampled from unlabeled areas to construct training and test sets [39], [59], [63]. Parameter optimization is an important step, which has a great influence on the final susceptibility results. In our experiments, we used a three-fold cross-validation procedure to find the optimal parameters of different models. In this study, we referred to the recent work to set the parameters of the PU-BaggingDT and CNN method [32], [44]. Table I presents the parameters setting of different methods.

B. Landslide Susceptibility Maps

The constructed models were used to predict landslide susceptibility of each grid cell in the study area. Fig. 5 shows landslide susceptibility maps obtained by the different methods. All landslide susceptibility values were reclassified into five classes based on the natural break algorithm, namely very low, low, moderate, high, and very high classes. All susceptibility maps have similar spatial distribution characteristics, i.e., high and very high susceptible regions are mainly located near the

Yangtze River. The moderate susceptibility class has the smallest area among all susceptibility classes. Compared with the benchmark methods, AdaPU-RF can obtain larger very low and very high susceptibility areas, but has smaller low, moderate, and high susceptibility areas. Fig. 6 presents the frequency analysis of landslides on different susceptibility maps. The very high susceptibility areas obtained by AdaPU-RF achieved the highest landslide percentage, followed by RF, PU-BaggingDT, CNN, SVM, and LR. This indicates that the AdaPU-RF method achieves more accurate landslide susceptibility map. The PU-BaggingDT had the highest landslide percentage in very low susceptibility areas. This may be because it occurred incorrect label assignment during the training process.

C. Model Accuracy Analysis

We plotted the success rate and prediction rate curves, and calculated the area under the curve (AUC) to evaluate the model accuracy [64]. First, we sorted all landslide susceptibility values in descending order. Then, we divided equally these values into 100 classes [56]. The x-axis is the cumulative percentage of the study area, and the y-axis is the cumulative percentage of landslide areas. The success rate curve and the prediction rate curve are plotted based on training landslides and testing landslides, respectively. The larger the AUC value, the better the fit and prediction accuracy. Fig. 7 shows the success rate and prediction rate curves of different methods. For the success rate curves, PU-BaggingDT achieved the highest AUC value (0.999), followed by RF (0.988), AdaPU-RF (0.979), and SVM (0.908). For the prediction rate curves, AdaPU-RF had the highest AUC value of 0.906, which is 0.042–0.07 higher than the three benchmark methods. We can observe that PU-BaggingDT and RF were better than AdaPU-RF with success rate curves. However, they have lower accuracy than AdaPU-RF in terms of prediction rate curves. This means PU-baggingDT and RF are overfitted during training process, and AdaPU-RF keeps a good balance between fit ability and prediction ability.

D. Sensitivity Analysis

Generally, statistical methods are very sensitive to the input data. If the input data vary within a reasonable range, the results of a reliable and robust model will not change significantly [55]. In this study, we evaluated the sensitivity of the model from two aspects: one is the impact of different training sample

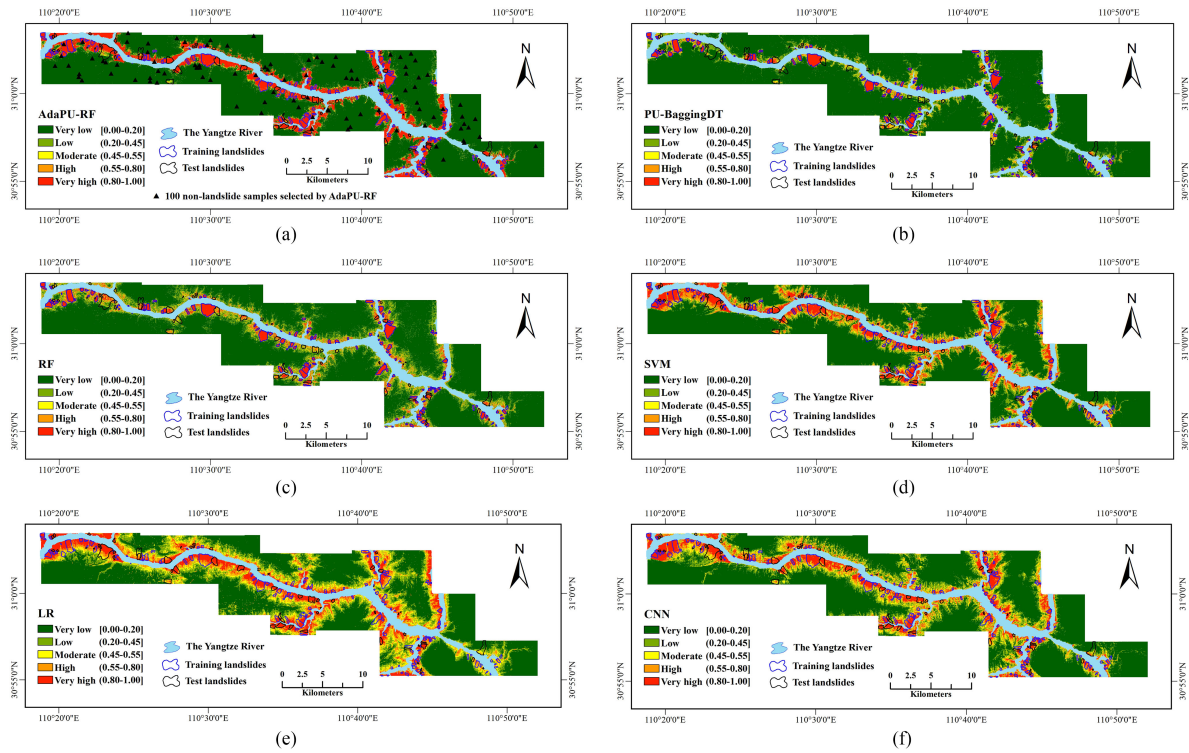


Fig. 5. Landslide susceptibility maps of different methods. (a) AdaPU-RF. (b) PU-BaggingDT. (c) RF. (d) SVM. (e) LR. (f) CNN.

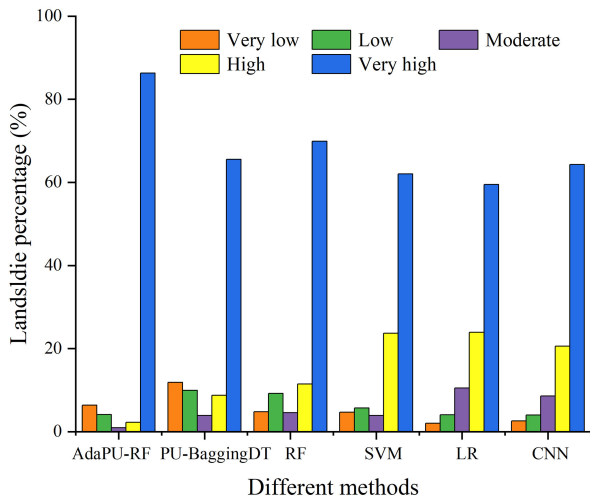


Fig. 6. Frequency analysis of landslides on susceptibility maps.

sizes on model performance, and the other is the impact of the random sampling process that generates training and test sets on model performance. We randomly selected samples with different proportions as the training set, and used the corresponding remaining samples as the test set. Meanwhile, we repeated the random sampling process ten times. Fig. 8 shows the model sensitivity of different methods. When the percentage of training set is between 10% and 70%, the AdaPU-RF method obtained the highest prediction accuracy than the benchmark methods. Meanwhile, only two results obtained by AdaPU-RF showed

higher accuracy than the other methods when the training samples account for 90% of the total samples. This phenomenon still shows that AdaPU-RF maintains good predictive ability in the face of different numbers of training samples.

In addition, all the methods that use 10% and 90% training samples had greater AUC variations than methods that use 30%, 50%, and 70% training samples. This is because the test set only contains 10% of the total number of samples, which can hardly reflect the true prediction performance of the model. Also, we can find that AdaPU-RF was less sensitive to the randomness of the training/test splitting process than the other methods when the percentage of training data is 10%, indicating its stable predictive ability with a small number of training samples.

E. Uncertainty Analysis

To analyze the uncertainty of landslide susceptibility mapping models, we selected the models [the same as Fig. 8(d)] that are trained with 70% of the training data. Therefore, we obtained ten landslide susceptibility estimates for each model. Fig. 9 shows the uncertainty analysis of different methods. These figures plot the average susceptibility estimate on the x -axis against two standard deviations (2σ) on the y -axis [55]. The 2σ values of the four models were low for grid cells classified as very high (probability > 0.8) and very low susceptibility areas (probability ≤ 0.2), indicating that the uncertainty in these areas is small. The grid cells classified as moderate susceptibility areas had higher 2σ values than other grid cells, which indicates that model cannot predict these grid cells as landslides or nonlandslides stably.

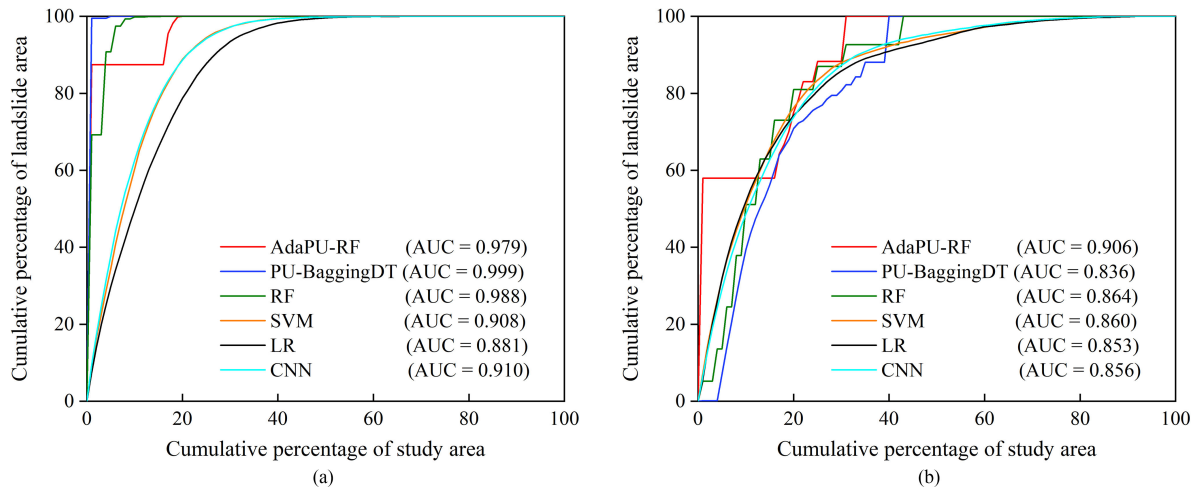


Fig. 7. (a) Success rate and (b) prediction rate curves of four methods.

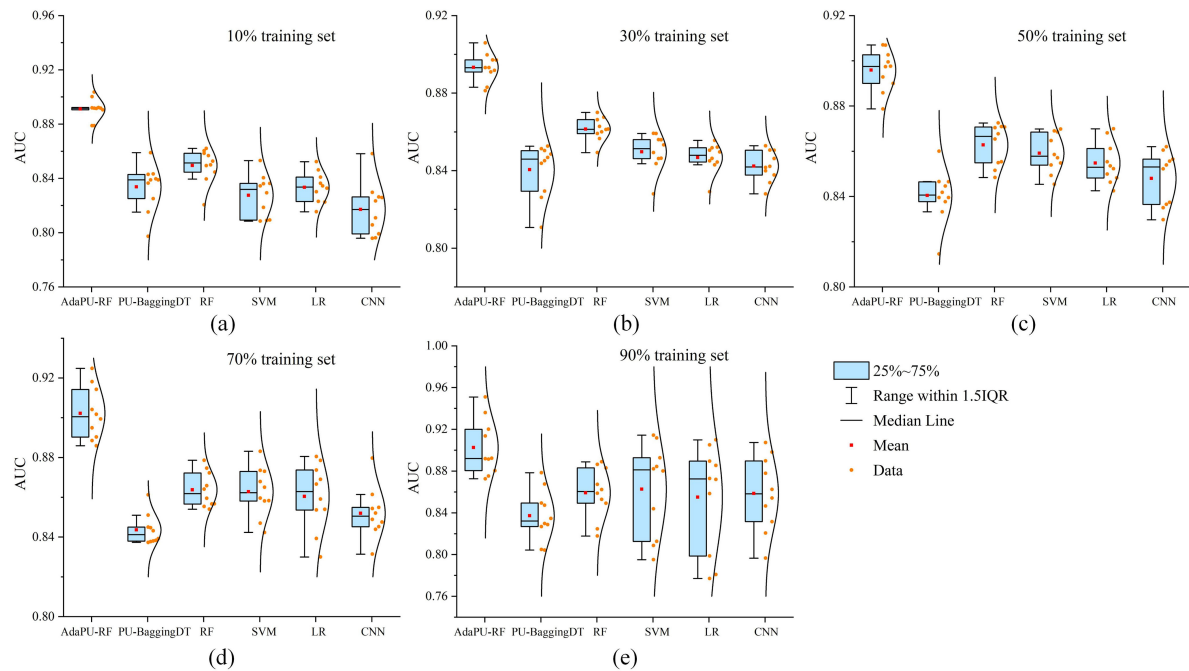


Fig. 8. Sensitivity analysis of four different methods. (a)–(e) AUC values of models with different percentages of the training data. Each modeling process is repeated ten times, and the ten yellow dots are the AUC values. The red dot is the mean value of the ten AUC values. The median line denotes the median value of the ten AUC values. The blue rectangle contains AUC values between the first quartile and the third quartile.

TABLE II
AVERAGE VALUES OF 2σ IN VERY LOW AND VERY HIGH SUSCEPTIBILITY AREAS

Methods	Mean of 2σ	
	≤ 0.2 (Very low)	> 0.8 (Very high)
AdaPU-RF	0.046	0.148
PU-BaggingDT	0.039	0.382
RF	0.087	0.151
SVM	0.031	0.080
LR	0.032	0.060
CNN	0.050	0.121

Table II lists the average of 2σ in the very low and very high susceptibility areas. SVM and LR achieved the lowest average value of 2σ in the very low and very high susceptibility areas, respectively. RF had the highest average value of 2σ in the very low and very high susceptibility areas, indicating that RF has the largest uncertainty. In general, the uncertainty of AdaPU-RF is within an acceptable range.

V. DISCUSSION

Since the mid-2000s, a large number of studies have applied machine learning methods to predict landslide susceptibility

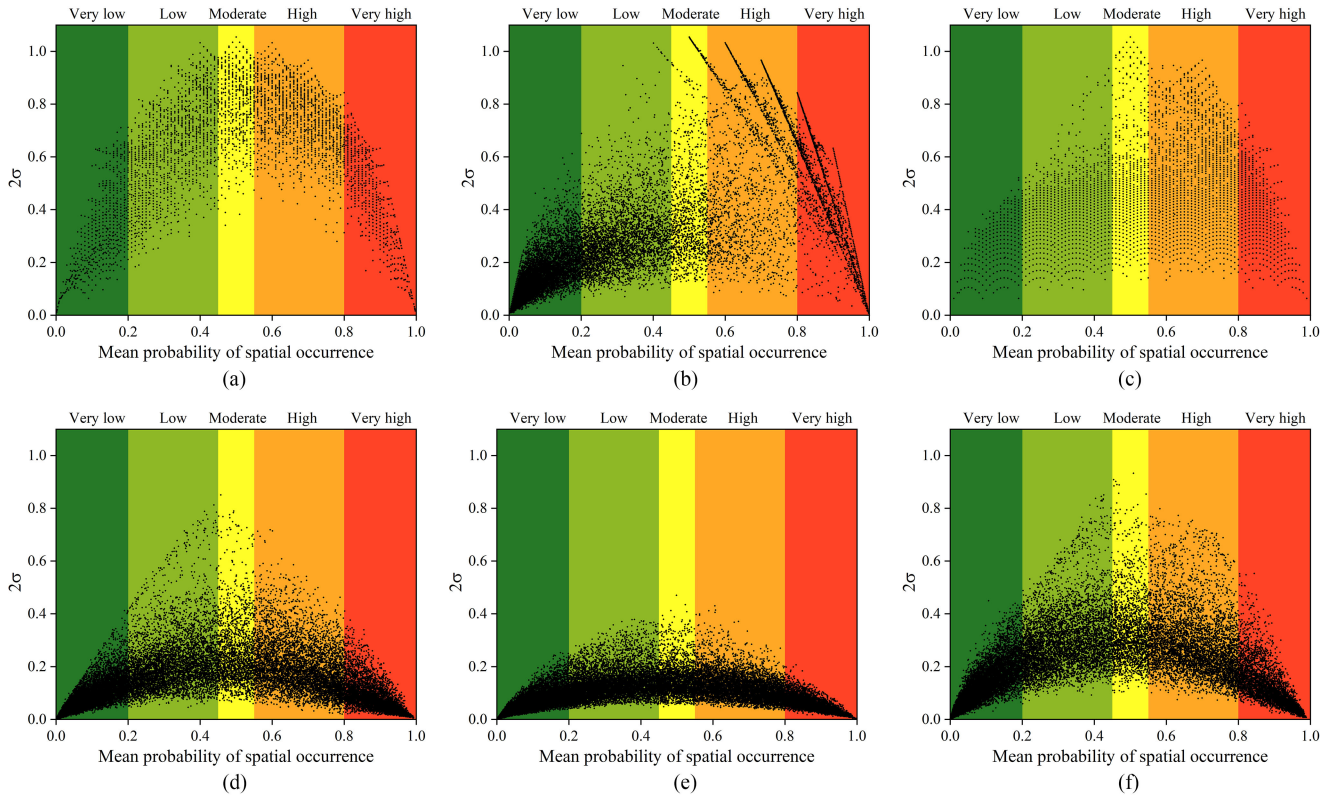


Fig. 9. Uncertainty analysis of different methods. x -axis denotes the mean of ten susceptibility estimates. y -axis denotes two standard deviations of the susceptibility estimates. (a) AdaPU-RF. (b) PU-BaggingDT. (c) RF. (d) SVM. (e) LR. (f) CNN.

[18]. These machine learning methods suffer the deficiency that requires both landslide samples and nonlandslide samples for modeling [36]. In the real-world, landslide information can be collected from existed landslide events. However, it is difficult for field investigators to identify real nonlandslide areas. This is because, on the one hand, the unknown area is much larger than the landslide area, so investigating all the unknown areas will spend a lot of resources. On the other hand, it is difficult to judge whether a landslide will occur in a certain area in the future. In this study, we treat LSP as a PU learning problem to predict landslide susceptibility based on landslide and unlabeled data.

The RF, SVM, LR, and CNN are commonly used machine learning method in landslide susceptibility analysis. They regard LSP as a binary classification problem and randomly select nonlandslide samples from the areas outside the landslide polygon. This selection procedure may select wrong samples, and cause bias and uncertainty. The PU-BaggingDT method uses bootstrap sampling to reduce the instability caused by the nonlandslide sampling process and obtain stable prediction results. However, it treats all unlabeled samples as nonlandslide data, so there is a problem of incorrect label assignment. The AdaPU-RF method trains a model based on landslide and unlabeled information. By iteratively selecting reliable landslide and nonlandslide samples, it can make full use of unlabeled information to update the training set. It can avoid the problem of incorrect label assignment that occurs in traditional nonlandslide sampling procedures and improve prediction performance.

A comprehensive assessment of landslide susceptibility models is important and necessary [65]. Guzzetti *et al.* [55] proposed a series of criteria to rank the quality of landslide susceptibility evaluation with the quality scale from 0 to 7. In our experiment, we used the highest quality assessment scale (level 7) to validate the AdaPU-RF method and compared it with the benchmark methods. Regarding the degree of model fitting, we found that the AdaPU-RF method cannot obtain the highest AUC value (see Fig. 7), which is lower than PU-baggingDT and RF. For the degree of model prediction, the AdaPU-RF and PU-BaggingDT methods obtained the highest and lowest AUC values, respectively. This indicates that PU-baggingDT has overfitted during training, whereas AdaPU-RF is well trained. The reason for this phenomenon is because PU-baggingDT is an ensemble method that combines multiple decision tree models, which will exacerbate the problem of label assignment errors.

Model sensitivity analysis can describe the robustness of the model. In this study, we tested the changes in the modeling results when two factors change: one is the training sample size, and the other is the random splitting process of the training and test sets. The former factor is important because a good model should maintain a high accuracy when the training sample size changes [66]. The latter will change the landslide distribution of the training set and cause uncertainty. Results show that the AdaPU-RF method is more robust and less sensitive to the randomness of the training/test splitting process, compared to the other methods. In addition, when the training samples account for 90% of the total number of samples, the AUC values

of all the methods will change greatly as the random splitting process is repeated [see Fig. 8(e)]. This is because the test set only contains 10% of the total number of samples, which can hardly reflect the true prediction performance of the model.

Many studies focus on the quantitative assessment of model performance on training and test sets [18], [21], [53], [59], which is useful, but does not fully evaluate the accuracy and reliability of the susceptibility prediction model. We investigated the uncertainty of the susceptibility estimate corresponding to each grid cell in the study area. 2σ was used as a quantitative measure of landslide susceptibility (see Fig. 9). We find that all susceptibility models are stable in predicting very low and very high susceptible areas. When estimating the susceptibility of grid cells in the moderate susceptible area, the model will lose the ability to distinguish whether the grid cell is stable. The finding is consistent with previous studies [47], [55], [67]. From the statistical results of 2σ (see Table II), the AdaPU-RF method is more stable than PU-baggingDT, RF, and CNN, but the stability is lower than SVM and LR. This may be because the repeated selection of landslide and nonlandslide samples during the training process will increase the uncertainty of the AdaPU-RF method. However, considering the improvement of prediction accuracy and calculation efficiency, it is more recommended to use the AdaPU-RF method to predict landslide susceptibility. Compared with the traditional application of machine learning methods, PU learning provides a new application perspective for researchers to carry out the LSP task.

VI. CONCLUSION

This study develops a new AdaPU-RF method to predict landslide susceptibility in the Three Gorges Reservoir area, China. The AdaPU-RF method combines PU learning with an adaptive sampling strategy to make full use of landslide and unlabeled information, and avoid the problem of incorrect label assignment. The main conclusions are summarized as follows. First, the AdaPU-RF method can obtain accurate landslide susceptibility results. In terms of the AUC value of the prediction rate curve, the proposed method is 0.042–0.07 higher than the benchmark methods. Second, the AdaPU-RF method was not sensitive to the randomness of training/test splitting process, compared with the other methods. In addition, the proposed method can retain higher prediction accuracy than the benchmark methods when using different percentages of the training set. Third, the uncertainty of the susceptibility estimation obtained by the AdaPU-RF method is within a reasonable and satisfactory range. Generally, the AdaPU-RF method is enlightening and more recommended for predicting landslide susceptibility. The PU learning provides a new application perspective for the problem of LSP. Meanwhile, we expect that the proposed method can promote other researchers to further explore the application potential of PU learning.

ACKNOWLEDGMENT

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

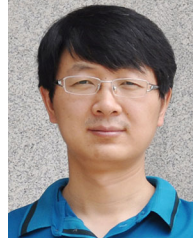
The authors are grateful to the Headquarters of Prevention and Control of Geo-Hazards in Area of Three Gorges Reservoir for providing data and material.

REFERENCES

- [1] D. Guha-Sapir, R. Below, and P. Hoyois, "EM-DAT: International disaster database," Université catholique de Louvain, Brussels, Belgium, 2020. Accessed: March 3, 2020. [Online]. Available: <http://www.emdat.be>.
- [2] F. Guzzetti, A. Carrara, M. Cardinali, and P. Reichenbach, "Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study, Central Italy," *Geomorphology*, vol. 31, no. 1–4, pp. 181–216, 1999.
- [3] E. A. C. Abella and C. J. Van Westen, "Qualitative landslide susceptibility assessment by multicriteria analysis: A case study from San Antonio del Sur, Guantánamo, Cuba," *Geomorphology*, vol. 94, no. 3/4, pp. 453–466, 2008.
- [4] D. Bălăteanu *et al.*, "National-scale landslide susceptibility map of Romania in a European methodological framework," *Geomorphology*, vol. 371, 2020, Art. no. 107432.
- [5] V. Medina, M. Hürlimann, Z. Guo, A. Lloret, and J. Vaunat, "Fast physically-based model for rainfall-induced landslide susceptibility assessment at regional scale," *Catena*, vol. 201, 2021, Art. no. 105213.
- [6] S. Wang, K. Zhang, L. P. van Beek, X. Tian, and T. A. Bogaard, "Physically-based landslide prediction over a large region: Scaling low-resolution hydrological model results for high-resolution slope stability assessment," *Environ. Model. Softw.*, vol. 124, 2020, Art. no. 104607.
- [7] H. Zhang, G. Zhang, and Q. Jia, "Integration of analytical hierarchy process and landslide susceptibility index based landslide susceptibility assessment of the Pearl River Delta Area, China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 11, pp. 4239–4251, Nov. 2019.
- [8] S. A. Ali *et al.*, "GIS-based landslide susceptibility modeling: A comparison between fuzzy multi-criteria and machine learning algorithms," *Geosci. Frontiers*, vol. 12, no. 2, pp. 857–876, 2021.
- [9] V. Ghiasi, S. A. R. Ghasemi, and M. Yousefi, "Landslide susceptibility mapping through continuous fuzzification and geometric average multi-criteria decision-making approaches," *Natural Hazards*, vol. 107, no. 1, pp. 795–808, 2021.
- [10] A. Bera, B. P. Mukhopadhyay, and D. Das, "Landslide hazard zonation mapping using multi-criteria analysis with the help of GIS techniques: A case study from Eastern Himalayas, Namchi, South Sikkim," *Natural Hazards*, vol. 96, no. 2, pp. 935–959, 2019.
- [11] Q. B. Pham *et al.*, "A comparison among fuzzy multi-criteria decision making, bivariate, multivariate and machine learning models in landslide susceptibility mapping," *Geomatics, Natural Hazards Risk*, vol. 12, no. 1, pp. 1741–1777, 2021.
- [12] A. Arabameri, B. Pradhan, K. Rezaei, M. Sohrabi, and Z. Kalantari, "GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms," *J. Mountain Sci.*, vol. 16, no. 3, pp. 595–618, 2019.
- [13] T.-Y. Zhang, L. Han, H. Zhang, Y.-H. Zhao, X.-A. Li, and L. Zhao, "GIS-based landslide susceptibility mapping using hybrid integration approaches of fractal dimension with index of entropy and support vector machine," *J. Mountain Sci.*, vol. 16, no. 6, pp. 1275–1288, 2019.
- [14] V. K. Pandey, H. R. Pourghasemi, and M. C. Sharma, "Landslide susceptibility mapping using maximum entropy and support vector machine models along the Highway Corridor, Garhwal Himalaya," *Geocarto Int.*, vol. 35, no. 2, pp. 168–187, 2020.
- [15] Y. Tang *et al.*, "Integrating principal component analysis with statistically-based models for analysis of causal factors and landslide susceptibility mapping: A comparative study from the Loess plateau area in Shanxi (China)," *J. Clean Prod.*, vol. 277, 2020, Art. no. 124159.
- [16] J.-H. Lee, M. I. Sameen, B. Pradhan, and H.-J. Park, "Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods," *Geomorphology*, vol. 303, pp. 284–298, 2018.
- [17] P. Goyes-Peñafiel and A. Hernandez-Rojas, "Landslide susceptibility index based on the integration of logistic regression and weights of evidence: A case study in Popayan, Colombia," *Eng. Geol.*, vol. 280, 2021, Art. no. 105958.
- [18] A. Merghadi *et al.*, "Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance," *Earth-Sci. Rev.*, vol. 2020, 2020, Art. no. 103225.

- [19] L. Lombardo and P. M. Mai, "Presenting logistic regression-based landslide susceptibility results," *Eng. Geol.*, vol. 244, pp. 14–24, 2018.
- [20] V.-H. Nhu *et al.*, "Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, Naïve Bayes Tree, artificial neural network, and support vector machine algorithms," *Int. J. Environ. Res. Public Health*, vol. 17, no. 8, 2020, Art. no. 2749.
- [21] J. Song, Y. Wang, Z. Fang, L. Peng, and H. Hong, "Potential of ensemble learning to improve tree-based classifiers for landslide susceptibility mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4642–4662, Aug. 2020.
- [22] S. Chen, Z. Miao, L. Wu, and Y. He, "Application of an incomplete landslide inventory and one class classifier to earthquake-induced landslide susceptibility mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1649–1660, Apr. 2020.
- [23] 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.
- [24] 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, 2020.
- [25] R. Q. Niu, X. L. Wu, D. K. Yao, L. Peng, L. Ai, and J. H. Peng, "Susceptibility assessment of landslides triggered by the Lushan Earthquake, April 20, 2013, China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 9, pp. 3979–3992, Sep. 2014.
- [26] D. Sun, J. Xu, H. Wen, and D. Wang, "Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest," *Eng. Geol.*, vol. 281, 2021, Art. no. 105972.
- [27] B. F. Tanyu, A. Abbaspour, Y. Alimohammadlou, and G. Tecuci, "Landslide susceptibility analyses using Random forest, C4.5, and C5.0 with balanced and unbalanced datasets," *Catena*, vol. 203, 2021, Art. no. 105355.
- [28] L. Bragagnolo, R. da Silva, and J. Grzybowski, "Artificial neural network ensembles applied to the mapping of landslide susceptibility," *Catena*, vol. 184, 2020, Art. no. 104240.
- [29] M. Di Napoli *et al.*, "Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability," *Landslides*, vol. 17, no. 8, pp. 1897–1914, 2020.
- [30] 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.
- [31] M. I. Sameen, B. Pradhan, and S. Lee, "Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment," *Catena*, vol. 186, 2020, Art. no. 104249.
- [32] 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.
- [33] Y. Chen, D. Ming, X. Ling, X. Lv, and C. Zhou, "Landslide susceptibility mapping using feature fusion-based CPCNN-ML in Lantau Island, Hong Kong," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3625–3639, Mar. 2021.
- [34] H. Cai, T. Chen, R. Niu, and A. Plaza, "Landslide detection using densely connected convolutional networks and environmental conditions," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5235–5247, May 2021.
- [35] X. Gao, T. Chen, R. Niu, and A. Plaza, "Recognition and mapping of landslide using a fully convolutional DenseNet and influencing factors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7881–7894, Aug. 2021.
- [36] X. Yao, L. G. Tham, and F. C. Dai, "Landslide susceptibility mapping based on support vector machine: A case study on natural slopes of Hong Kong, China," *Geomorphology*, vol. 101, no. 4, pp. 572–582, 2008.
- [37] Á. M. Felicísimo, A. Cuartero, J. Remondo, and E. Quirós, "Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: A comparative study," *Landslides*, vol. 10, no. 2, pp. 175–189, 2013.
- [38] A. Aditian, T. Kubota, and Y. Shinohara, "Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia," *Geomorphology*, vol. 318, no. 2018, pp. 101–111, 2018.
- [39] Z. Fang, Y. Wang, L. Peng, and H. Hong, "A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping," *Int. J. Geographical Inf. Sci.*, vol. 35, pp. 321–347, 2020.
- [40] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 213–220.
- [41] M. A. Zuluaga, D. Hush, E. J. D. Leyton, M. H. Hoyos, and M. Orkisz, "Learning from only positive and unlabeled data to detect lesions in vascular CT images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2011, pp. 9–16.
- [42] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 899–904.
- [43] J. Zhang, Z. Wang, J. Meng, Y.-P. Tan, and J. Yuan, "Boosting positive and unlabeled learning for anomaly detection with multi-features," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1332–1344, May 2018.
- [44] B. Wu, W. Qiu, J. Jia, and N. Liu, "Landslide susceptibility modeling using bagging-based positive-unlabeled learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 766–770, May 2020.
- [45] P. Yang, W. Liu, and J. Y. H. Yang, "Positive unlabeled learning via wrapper-based adaptive sampling," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3273–3279.
- [46] H. Tang, J. Wasowski, and C. H. Juang, "Geohazards in the Three Gorges Reservoir area, China—Lessons learned from decades of research," *Eng. Geol.*, vol. 261, 2019, Art. no. 105267.
- [47] L. Peng, R. Niu, B. Huang, X. Wu, Y. Zhao, and R. Ye, "Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China," *Geomorphology*, vol. 204, pp. 287–301, 2014.
- [48] K. Xu *et al.*, "Landslide susceptibility evaluation based on BPNN and GIS: A case of Guojiaba in the Three Gorges Reservoir area," *Int. J. Geographical Inf. Sci.*, vol. 29, no. 7, pp. 1111–1124, 2015.
- [49] X. Yu, Y. Wang, R. Niu, and Y. Hu, "A combination of geographically weighted regression, particle swarm optimization and support vector machine for landslide susceptibility mapping: A case study at Wanzhou in the Three Gorges area, China," *Int. J. Environ. Res. Public Health*, vol. 13, no. 5, 2016, Art. no. 487.
- [50] 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.
- [51] C. Zhou *et al.*, "Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China," *Comput. Geosci.*, vol. 112, pp. 23–37, 2018.
- [52] A. M. Youssef, H. R. Pourghasemi, Z. S. Pourtaghi, and M. M. Al-Katheeri, "Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir region, Saudi Arabia," *Landslides*, vol. 13, no. 5, pp. 839–856, 2016.
- [53] 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.
- [54] Q. Zhu, L. Chen, H. Hu, S. Pirasteh, H. Li, and X. Xie, "Unsupervised feature learning to improve transferability of landslide susceptibility representations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3917–3930, Jul. 2020.
- [55] F. Guzzetti, P. Reichenbach, F. Ardizzone, M. Cardinali, and M. Galli, "Estimating the quality of landslide susceptibility models," *Geomorphology*, vol. 81, no. 1/2, pp. 166–184, 2006.
- [56] B. Pradhan and S. Lee, "Landslide susceptibility assessment and factor effect analysis: Backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling," *Environ. Model. Softw.*, vol. 25, no. 6, pp. 747–759, 2010.
- [57] A. Erenér and H. S. B. Düzgün, "Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of more and Romsdal (Norway)," *Landslides*, vol. 7, no. 1, pp. 55–68, 2010.
- [58] A. Erenér, A. Mutlu, and H. S. Düzgün, "A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM)," *Eng. Geol.*, vol. 203, pp. 45–55, 2016.
- [59] M. Juliev, M. Mergili, I. Mondal, B. Nurtaev, A. Pulatov, and J. Hübl, "Comparative analysis of statistical methods for landslide susceptibility mapping in the Bostanlik district, Uzbekistan," *Sci. Total Environ.*, vol. 653, pp. 801–814, 2019.

- [60] D. Zhu, T. Chen, Z. Wang, and R. Niu, "Detecting ecological spatial-temporal changes by remote sensing ecological index with local adaptability," *J. Environ. Manage.*, vol. 299, 2021, Art. no. 113655.
- [61] P. J. Lavrakas, *Encyclopedia of Survey Research Methods*. London, U.K.: Sage Publ., 2008.
- [62] P. Yang, J. T. Ormerod, W. Liu, C. Ma, A. Y. Zomaya, and J. Y. Yang, "Adasampling for positive-unlabeled and label noise learning with bioinformatics applications," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1932–1943, May 2019.
- [63] J. Samia *et al.*, "Implementing landslide path dependency in landslide susceptibility modelling," *Landslides*, vol. 15, no. 11, pp. 2129–2144, 2018.
- [64] C.-J. F. Chung and A. G. Fabbri, "Validation of spatial prediction models for landslide hazard mapping," *Natural Hazards*, vol. 30, no. 3, pp. 451–472, 2003.
- [65] P. Reichenbach, M. Rossi, B. Malamud, M. Mihir, and F. Guzzetti, "A review of statistically-based landslide susceptibility models," *Earth-Sci. Rev.*, vol. 180, pp. 60–91, 2018.
- [66] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts, "Mapping land-cover modifications over large areas: A comparison of machine learning algorithms," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 2272–2283, 2008.
- [67] M. Rossi, F. Guzzetti, P. Reichenbach, A. C. Mondini, and S. Peruccacci, "Optimal landslide susceptibility zonation based on multiple forecasts," *Geomorphology*, vol. 114, no. 3, pp. 129–142, 2010.



Ruiqing Niu received the Ph.D. degree in earth exploration and information technology from China University of Geosciences, Wuhan, China, in 2005.

He is currently a Professor with the Institute of Geophysics and Geomatics, China University of Geosciences. His research interests include remote sensing, geographic information system, and engineering geology.



Ling Peng received the Ph.D. degree in earth exploration and information technology from 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.



Zhice Fang received the B.E. degree in geoinformatics, in 2017, from China University of Geosciences, Wuhan, China, where he is currently working toward the Ph.D. degree in earth exploration and information technology.

His research interests include natural disaster susceptibility mapping and remote sensing applications.



Yi Wang (Member, IEEE) received the B.S. degree in printing engineering and 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 and Chinese Association of Automation. In 2019, he was named CUG Outstanding Young Talent.