An Integrated Framework of Positive-Unlabeled and Imbalanced Learning for Landslide Susceptibility Mapping

Zijin Fu, Hao Ma, Fawu Wang, Jie Dou, Bo Zhang, and Zhice Fang

***Abstract*—Machine learning is pivotal in data-driven landslide susceptibility mapping (LSM). However, the uncertainty of negative samples and the imbalance between positive and negative samples, which leads to misjudgments and overestimation, remain ongoing challenges. This study introduces a novel framework for LSM that integrates positive-unlabeled (PU) learning with imbalanced learning methods, making full and correct use of vast unlabeled samples. First, a prior model based on the spy algorithm is generated to obtain reliable negative (RN) samples, which is used to create imbalanced training and testing sets. Subsequently, four imbalanced learning models, namely synthetic minority oversampling technique-deep neural network (SMOTE-DNN), adaptive synthetic-DNN (ADASYN-DNN), balanced random forest (BRF), and EasyEnsemble (EE) are employed to process the imbalanced training and testing sets and generate the final prediction models. We have tested our LSM framework using a dataset of regional rainfall-induced landslides that occurred in Beijing, China. The positive impacts of RN samples are evaluated using baseline models and extensive saturation tests with various imbalance ratios are conducted. Imbalanced learning methods enhanced prediction for negative classes, with balance peaks observed in the saturation tests. BRF showed the best performance and stability across different imbalance ratios. This framework can improve the prediction accuracy for both positive and negative classes, which has the potential to reduce overestimation and misclassification and holds promise for significantly impacting future modeling strategies in LSM.**

***Index Terms*—Data-driven landslide susceptibility mapping (LSM), imbalanced learning, negative samples, positive-unlabeled (PU) learning.**

I. Introduction

LANDSLIDE presents a destructive geological hazard posing huge threats to both life and property, especially in mountainous regions. Landslides are particularly susceptible to being triggered by external events such as intense rainfall or earthquake in areas with fragile geological settings. The combination of surface erosion, softening of rock and soil masses, and increased pore water pressure resulting from heavy rainfall can easily lead to the occurrence of geo-hydrological disasters like landslides and mudflows [1]. In the context of climate change, frequent extreme weather events can easily trigger a large number of landslides in specific areas [2], [3]. For instance, during four consecutive days (from July 29 to August 1 in 2023) of heavy rainfall in the western mountainous areas of Beijing, China, over 15 000 landslides were triggered. These widely and randomly distributed landslides have brought huge challenges to regional disaster prevention and mitigation. Landslide susceptibility mapping (LSM) is used to predict the spatial probability of landslide occurring under certain conditions over a large-scale area. LSM is one of the tools for risk assessment of regional disasters, providing effective forecasts for disaster prevention and mitigation, and improving our understanding of the distribution pattern of such hazards.

The methods of LSM can be broadly divided into three categories: knowledge-driven, physics-based and data-driven methods [4]. While knowledge-driven methods tend to be subjective and the physics-based methods need multiple and accurate physical field parameters that can be laborious to acquire, the efficiency and accuracy of data-driven methods have made them increasingly popular in LSM. Advances in remote sensing techniques have provided a wealth of accessible multisource data, which also promotes the modeling process for data-driven LSM [5], [6], [7], [8]. Machine learning as a prevalent data-driven approach, has been widely applied and promoted in LSM, encompassing classical algorithms like logistic regression, support vector machine, decision tree (DT), random forest (RF), Bayes as well as deep learning methods like deep neural network (DNN), convolutional neural networks, residual network, transformer, etc. [9], [10], [11], [12]. However, within the supervised learning framework that most methods align with, the fundamental understanding of landslide data often gets overlooked. It is crucial to recognize that LSM based on supervised learning faces a positive-unlabeled (PU) and imbalanced problem.

The problem of PU in LSM is mainly due to the uncertainty of the unlabeled samples. Nonlandslide samples are not directly obtained but need to be generated in LSM. Many studies on LSM

Received 17 June 2024; revised 2 August 2024 and 23 August 2024; accepted 27 August 2024. Date of publication 30 August 2024; date of current version 13 September 2024. This work was supported by the National Natural Science Foundation of China under Grant 42230715. *(Corresponding author: Fawu Wang.)*

Zijin Fu, Hao Ma, and Bo Zhang are with the College of Civil Engineering, Tongji University, Shanghai 200092, China (e-mail: fuzijin@tongji.edu.cn; mah@tongji.edu.cn; bo_zhang@tongji.edu.cn).

Fawu Wang is with the Key Laboratory of Geotechnical and Underground Engineering of the Ministry of Education, Tongji University, Shanghai 200092, China (e-mail: wangfw@tongji.edu.cn).

Jie Dou is with the Badong National Observation and Research Station of Geohazards, China University of Geosciences, Wuhan 430074, China (e-mail: doujie@cug.edu.cn).

Zhice Fang is with the School of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China (e-mail: zhicefang@cug.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3452182

using supervised learning have employed random sampling as the most commonly used method based on the assumption that areas without landslides are likely to be stable [13]. Although this method is convenient and fast, areas without detected landslides may still have undetected landslides or conditions conducive to landslides, making them highly susceptible to landslides [14]. Thus, the unlabeled samples generated by random sampling actually contain a significant amount of uncertainty, which results in misjudgments. Researches show that negative samples screened by prior models can improve the prediction performance better than unlabeled samples [15], [16], [17], [18], [19]. Prior models like presence-only and PU learning methods can be used for this purpose. In addition, physics-based prior models can also be adapted to enhance the model's generalization capabilities and achieve a data-physics hybrid driver [20], [21], [22]. The quality of the negative samples generated from prior model trained by large amounts of unlabeled samples and landslide samples can also significantly affect the modeling of LSM. Some conclusions have been drawn from studies that extremely negative samples can lead to overestimation of susceptibility levels and critical negative samples can strengthen the decision boundary of models [23], [24]. In addition, innovative PU techniques have been applied in LSM, offering a promising direction for selecting reliable negative (RN) samples [25], [26].

The problem of imbalance within LSM is evident in the unequal or deeply imbalanced distribution of landslide and nonlandslide classes. This imbalance is similar to challenges encountered in various imbalanced classification tasks in industries like fault detection, fraud detection medical diagnoses, and so on [27], [28]. In LSM, the disparity between the majority class (nonlandslide) and the minority class (landslide) can be significant with ratios of 10:1, 100:1, or even larger in the reality [29]. It can be inferred from the most datasets of landslide inventory that landslides account for less than 5%, 1%, or much smaller of the study total area even in areas with strong earthquakes or heavy rainfall, although there may be incomplete interpretations due to invisibility [30], [31], [32]. Most standard machine learning methods usually treat all classes equally, assigning them the same misclassification cost and balanced proportion [27]. However, these methods are not optimal for imbalanced datasets due to inconsistency of sample distribution, the tendency to overlook misclassification of minority classes, and the neglect of valuable information in unlabeled samples. To address this, many algorithms have been optimized to adapt to imbalanced dataset by adjusting the sample weights for different classes. Studies have explored the effect of imbalanced datasets containing different numbers of negative samples on the LSM results, both sensitivity and specificity are significantly affected by the imbalance ratio [14], [33], [34], [35], [36]. In recent years, intelligent imbalanced learning has evolved into the following broad categories: data-level approach, ensemble approach, algorithm-level approaches, cost-sensitive learning, and hybrid algorithms [29], [37], [38]. Among these, the undersampling and oversampling methods of the data-level approach as well as ensemble learning methods are effective and worth investigating for imbalanced landslide susceptibility datasets. Some imbalanced learning study have already achieved superior results compared to common machine learning models in LSM [39], [40], [41].

The quality and quantity of negative samples used in the LSM profoundly influence the prediction. Recognizing LSM as a PU and imbalanced problem highlights the significance of optimizing PU learning and imbalanced learning methods. However, few studies have merged these approaches. This study proposes a novel framework combining PU learning with imbalanced learning for LSM. The main contributions of this study can be summarized as follows:

1) a susceptibility prior model based on the PU spy algorithm is generated to obtain a large number of RN samples;
2) four imbalanced learning models including synthetic minority oversampling technique-DNN (SMOTE-DNN), adaptive synthetic-DNN (ADASYN-DNN), balanced random forest (BRF), and EasyEnsemble (EE) from oversampling and ensemble learning are trained and tested on imbalanced training and testing sets;
3) extensive repeated sampling, detailed model comparison, and saturation test with different imbalanced ratios are conducted to compare the improvement brought by RN samples and the performance of different imbalanced models.

II. Study Area and Data

A. Study Area

From 29 July 2023 to 2 August 2023, an unusual and intense rainfall event occurred in Beijing, lasting for 83 h. The average cumulative rainfall across Beijing reached 331 mm, accounting for about 60% of the annual average rainfall (599.5 mm). The rainfall center is located in the western part of the city [see Fig. 1(a)], which is mountainous [see Fig. 1(b)]. Taking into account the rainfall distribution and geological conditions, the mountainous areas in western Beijing were selected as the study area, which covers all mountainous areas in Fangshan District and Mentougou District, as well as the southwest of Changping District. The study area covers an area of approximately 3250 km^2. Affected by this rainstorm, a large number of landslides were triggered in this area.

The western mountainous area is located at the northern end of the Taihang Mountains, and complex geological and geomorphic conditions make this area the highest prone area to geological hazards in Beijing. The NE compartmentalized fold structures shaped the basic tectonic framework. The strata are dominated by Cambrian and Ordovician carbonate rocks, as well as sandstone and mudstone of the Jurassic and Permian. In addition, igneous rocks and metamorphic rocks are also locally exposed. Affected by tectonic activities, many areas exist with massive fragmental rocks. The mountain ridges generally trend northeastward, consistent with the regional structural direction. The elevation of the mountainous areas generally ranges from 1000 to 1500 m. The highest altitude is over 2000 m. The main types of landforms include medium-sized mountains, low mountains, hills, and valleys. Erosion landforms are developed.

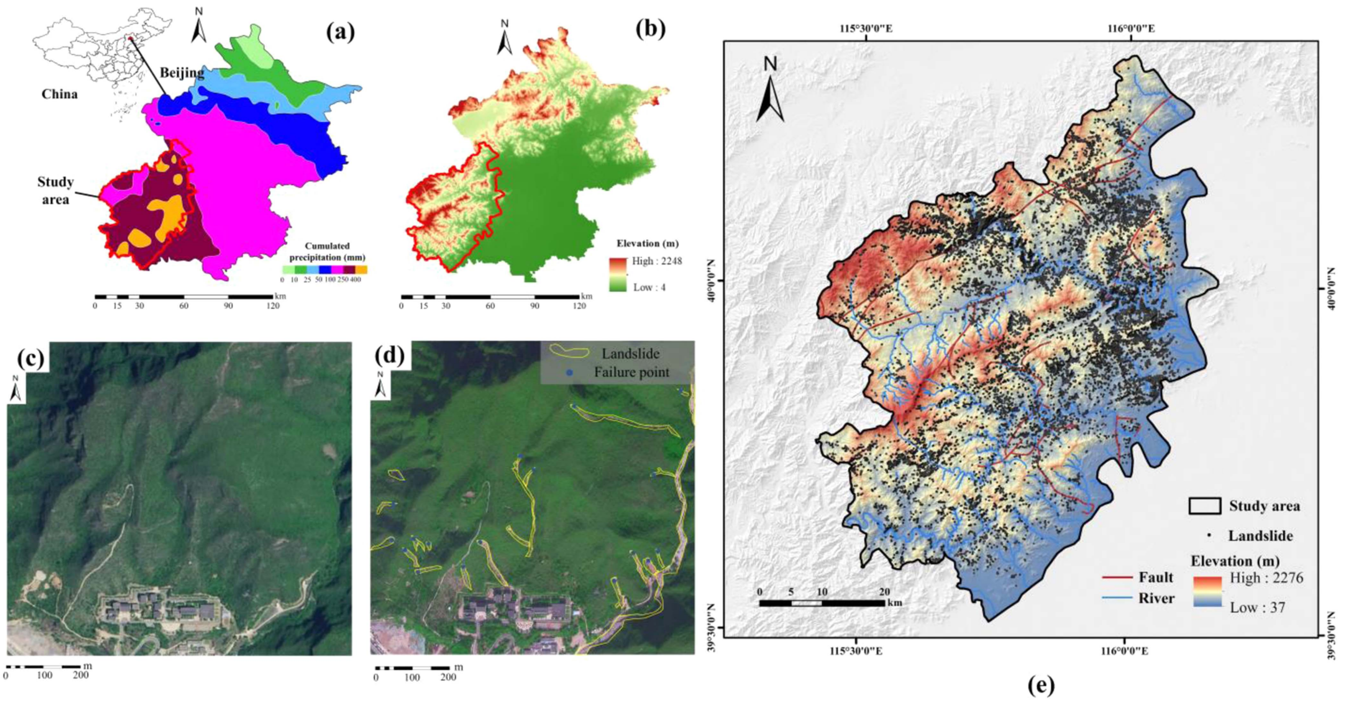

Fig. 1. (a) Cumulative rainfall distribution in Beijing from 8:00 PM on 29 July 2023, to 1:00 PM on 31 July 2023 (from Beijing meteorological bureau). (b) Elevation distribution of Beijing. (c) Satellite images one month prior to this rainstorm. (d) Satellite images one month after this rainstorm where noticeable rainfall-induced landslides can be identified. (e) Distribution of landslides in the study area.

B. Landslide Inventory

For establishing the landslide inventory, the high-resolution satellite images approximately one month before and one month after this rainstorm event were used to visually identify the landslides triggered these rainstorm events [see Fig. 1(c) and (d)]. The resolution of the images is between 0.8 and 2.0 m, and the data is from SIWEIearth platform. A total of 15 383 landslides were identified [see Fig. 1(e)], covering a combined area of 19.3 km^2. The area of an individual landslide ranges from about 40 m^2 to 2.1 × 10^5 m^2 in the landslide area. In total, 70% of the landslides caused by this rainfall event have an area of less than 1000 m^2. As most landslides triggered by this rainstorm are small at the initiation area and larger at the moving and deposit areas, the locations of the landslide main scarp are identified, which will serve as landslide training and testing samples for this study.

C. Landslide Conditioning Factors

Identifying landslide conditioning factors (LCFs) is a crucial step in LSM. LCFs can be divided into broad categories including topographic factors, geological factors, hydrological factors, and environmental factors in existing research works [42]. Due to extreme rainfall being the main cause of landslides in the study area in the short term, we added the accumulative rainfall during the landslide occurrence period as one of the LCFs, which is a common characteristic of event-induced LSM research works [43], [44]. Based on the availability of data and the characteristics of rainfall-induced landslides in the study area, we have preliminarily selected the following 16 LCFs: elevation, aspect, slope, plane curvature (PLC), profile curvature (PRC), terrain ruggedness index (TRI), topographic wetness index (TWI), topographic position index (TPI), stream power index (SPI), lithology, geological age (GA), distance to faults (DTF), distance to river (DTR), landcover, NDVI, and accumulative rainfall (AR). Almost all factors are commonly used in the LSM field, and all data types and sources are shown in Table I. The lithology is divided into the following seven categories: soil, andesite, volcanic clastic rock, gneiss, sandstone, carbonate, and granite. The GA includes Cambrian, Sinian, Ordovician, Carboniferous, Permian, Jurassic, Cretaceous, Quaternary, and Proterozoic. Landcover contains seven classes of bare ground, built area, crops, rangeland, trees, water, and flooded vegetation. All of the LCFs are processed on ArcGIS 10.8 and all layers are resampled with a resolution of 12.5 m.

III. Methodology

Fig. 2 depicts the detailed flow chart of this study. Initially, 16 LCFs are collected and examined by multicollinearity and importance analysis. Then, we apply the spy algorithm as a PU learning method to generate LSM prior models to obtain RN samples. Training and testing sets are constructed with a ratio of 7:3 using landslide and RN samples, and the testing set is fixed with an imbalanced ratio of 1:200. Next, four imbalanced learning models (SMOTE-DNN, ADASYN-DNN, BRF, and EE) are applied to process the imbalanced training and testing sets and generate the final prediction models. In this process, two benchmark models (DNN and RF) are employed to verify the positive impact of RN samples and saturation tests with different

TABLE I
DETAILED INFORMATION OF LCFs

LCFS	Values	Data type	Data source	Production time
Elevation (m)	[37, 2276]	Continuous	ALOS PLSAR	2009
Aspect (°)	[-1, 360]	Continuous	Derived from DEM	2009
Slope (°)	[0, 84.40]	Continuous	Derived from DEM	2009
PLC	[-57.50, 37.59]	Continuous	Derived from DEM	2009
PRC	[-61.07, 70.21]	Continuous	Derived from DEM	2009
TRI	[1.00, 10.24]	Continuous	Derived from DEM	2009
TWI	[0.13, 31.68]	Continuous	Derived from DEM	2009
TPI	[-119.41, 88.86]	Continuous	Derived from DEM	2009
SPI	[-9.33, 20.03]	Continuous	Derived from DEM	2009
Lithology	7 types	Discrete	1:1000000 geological map (https://geocloud.cgs.gov.cn/)	2002 to 2007
GA	9 types	Discrete		
DTF (m)	[0, 18398.70]	Continuous		
DTR (m)	[0, 12511]	Continuous	OpenStreetMap (https://www.openstreetmap.org/)	2023/7/17
Landcover	7 types	Discrete	Esri landcover (https://livingatlas.arcgis.com/landcoverexplorer/)	2022
NDVI	[-0.56, 1.00]	Continuous	Sentinel-2 L2A	2023/7/16
Accumulated rainfall (mm)	[241.69, 380.40]	Continuous	Interpolation by data from 17 meteorological stations within and around the research area (https://data.cma.cn/)	2023/7/29 to 2023/8/1

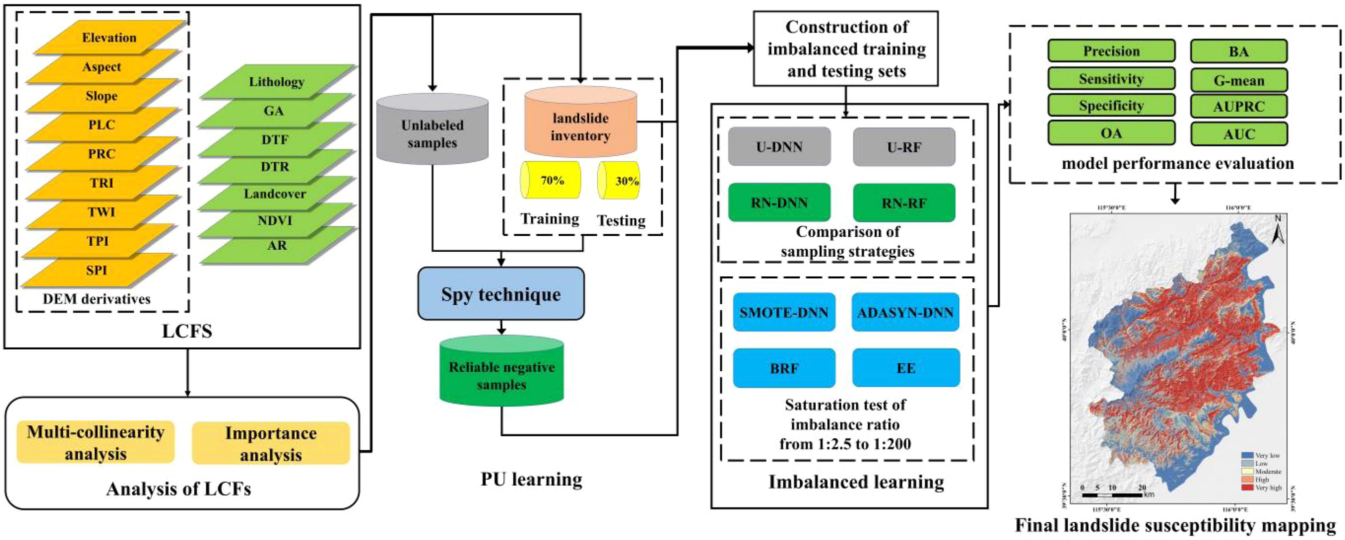

Fig. 2. Flow chart of this study.

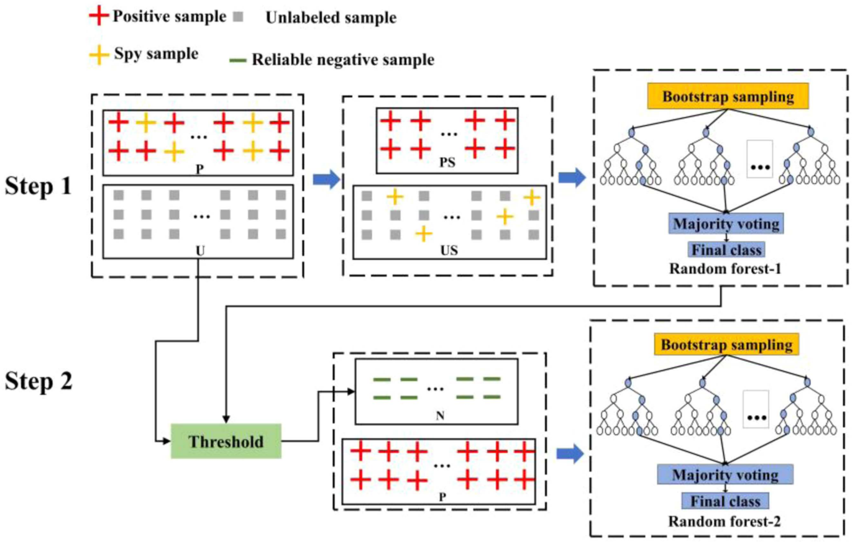

Fig. 3. Work flow of spy algorithm.

imbalance ratios are performed. Eight metrics are used to assess and compare the aforementioned models in detail, and the results of LSM are discussed.

A. LCFs Selection

1) Multicollinearity Analysis: Multicollinearity analysis is an essential process for selecting LCFs for LSM. It is generally believed that a certain conditioning factor does not have multicollinearity with other factors when the variance inflation factor (VIF) is smaller than 5. When VIF is between 5 and 10, this factor has weak multicollinearity with other factors. Factors with VIF values above 10 are considered to have moderate or higher multicollinearity [45]. In LSM, it is recommended to remove the conditioning factors of VIF above 10 to avoid redundant indicators leading to distorted predictions [46].

2) Importance Analysis: Gini importance or mean decrease in impurity can be used to evaluate the importance of every feature using impurity-based method [47]. The Gini coefficient in RFs is used to evaluate the importance of LCFs and to assist in the selection of LCFs.

B. Spy Algorithm

Two-step approach (TSA) is another common way to solve PU problems like LSM and consists of the following two main steps: 1) identify relatively RN samples (RRN); 2) construct a final classifier based on RRN samples [25], [48]. Two classifiers need to be constructed in the TSA, the first one uses positive and unlabeled samples and the second one uses positive and RRN samples identified from the first classifier. Building upon the TSA, this study adopts the spy technique and RF classifier to solve the PU problem in LSM.

The spy technique is an auxiliary technique in the two-step framework that is widely used because it outperforms other methods and is computationally fast [49], [50]. Fig. 3 shows the process of TSA with the spy technique. In step 1, 15% of the landslide samples are randomly selected as spy samples (S) to be added to the unlabeled sample set (U) to form a new sample set (US). Then train the first classifier with the positive sample set excluding spy samples (PS) and the US sample set and make predictions for all the samples. Determine the threshold based on the probability value of S, by which RRN samples are filtered in the set of unlabeled samples. In step 2, another classifier is trained based on the P and RRN as the final model for LSM. RF classifier is chosen as the two classifiers described above, which is widely used in LSM for its excellent predictive performance and efficiency.

C. Data-Level Imbalanced Learning

1) Synthetic Minority Oversampling Technique-DNN: One of the best-known techniques in oversampling methods is SMOTE [51]. SMOTE employs linear interpolation to generate new samples within the minority class. Random oversampling randomly copies existing minority class samples, which may easily cause the model to overfit due to too many repeated samples. SMOTE can mitigate the issue of over-fitting caused by random oversampling [52]. The process of SMOTE is shown in Fig. 4. For each minority class sample, find its K nearest neighbors in the minority class, then randomly select one of the K nearest neighbors, the synthetic sample is generated randomly on the connection between the selected neighbor and the current minority class sample. Repeat the above-mentioned operations until the number of minority class samples and majority class samples are balanced. The pseudo samples it creates are not the same with the original samples but retain the similar relationship with them. In order to pay more attention to samples near the decision boundary, some models such as SMOTE-borderline, SMOTEBoost, and ADASYN are developed based on SMOTE [53], [54].

2) Adaptive Synthetic: ADASYN is another technique similar to SMOTE that generates synthetic samples from the minority

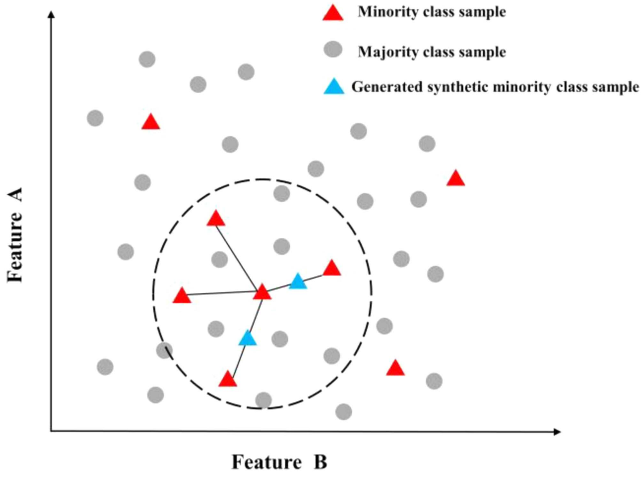

Fig. 4. Schematic diagram of SMOTE.

class to balance the dataset from data-level [55]. Similar to borderline-SMOTE and SMOTE-boost, ADASYN pays more attention to minority samples that are hard to learn by determining the number of synthetic samples to each minority sample based on density distribution [53]. ADASYN first calculates the weight of all the minority class samples through the proportion of majority class samples around the sample. Then, each minority class sample is assigned the number of synthetic samples needed to be generated based on its weight. This method generates more synthetic samples for the minority with blurred boundaries to the majority, which helps to train the decision boundaries of the model.

D. Ensemble Imbalanced Learning

1) Balanced Random Forest: The BRF is proposed to improve the performance of normal RF when facing an extremely imbalanced dataset [56]. RF is a supervised learning algorithm based on bagging, while bootstrap sampling is conducted to obtain input samples of each DT and random features of the sample are selected for splitting. Due to strong randomness, it has high robustness and is not easy to overfit. However, when facing extremely imbalanced data, most of the DT may receive few or even none minority class data, which greatly reduces its predictive performance on minority classes. In order to solve the category bias of the training samples on each tree, BRF introduces the sampling strategies to improve RF. Each tree of BRF receives a balanced subset generated by undersampling or oversampling.

2) EasyEnsemble: EE is one of the effective undersampling techniques based on ensemble learning, which demonstrates high accuracy while solving imbalanced problems [27], [57]. This method operates by randomly sampling multiple subsets from the majority class and then training base-classifiers through a combination of these subsets with minority class data. Each subset matches the minority class in number, ensuring that every base-classifier is trained on a balanced dataset. These base-classifiers use the AdaBoost algorithm, and the AdaBoost algorithm itself is also an ensemble learning method. AdaBoost is an iterative algorithm that adjusts the sample weight according to whether the sample can be correctly classified after each base classifier is trained. The weights of each base-classifiers are adjusted according to their predictive performance and AdaBoost is constructed of these weighted classifiers [58]. As the representative algorithm of Boosting, AdaBoost has extremely strong predictive performance and robustness. This structure suggests that EE is an ensemble of ensembles, which is a bagging ensemble based on the boosting ensembles.

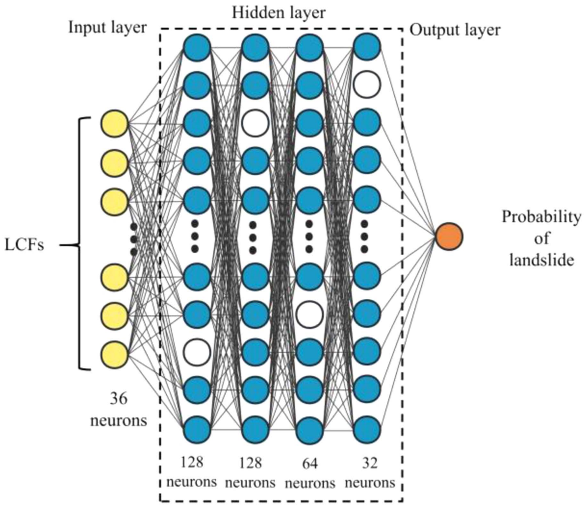

Fig. 5. Structure of DNN.

E. Benchmark Prediction Models

RF and DNN are employed as benchmark models for the comparation with the framework of PU and imbalanced learning. These two models are frequently used in the field of LSM due to their nonlinear processing ability and high features [10], [25]. RF is also deployed as base learners for spy algorithm and as the benchmark for the 1:1 trained balanced model in the imbalanced learning section, considering its good fitting effect and fast training speed on small datasets. When oversampling in data-level imbalanced methods, millions of samples will be generated, which requires a huge amount of time for training an RF model. Hence, we choose to use DNN models to train the extensive number of samples generated by data-level method. Neural networks contain massive computational parameters, making them suitable for handling complex relationships between large-scale datasets. In addition, with the support of hardware devices such as GPUs, DNN has a very high training efficiency in dealing with millions of data. After using the data-level method, the DNN model is used to process the generated sample sets on different imbalanced ratios. Fig. 5 shows the DNN structure used in the study, where the input layer is the results of the partial discretization of LCFs.

F. Model Evaluation Metrics

Confusion matrix is the most commonly used way to measure the accuracy of classification results. This matrix shows the correct classification and misclassification of positive and negative samples by the prediction and provides very useful information to measure the predictive performance of the model. The confusion matrix of the two-classification problem contains four parameters: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Based on the confusion matrix indicators, six metrics including overall accuracy (OA), precision, sensitivity, specificity, Balanced Accuracy (BA) and G-mean are selected for model assessment. OA reflects the OA of the positive and the negative predictions, while precision highlights the success rate of predicted positive samples from a forecasting perspective. Sensitivity and specificity measure the rate of successful prediction of real positive and negative samples from an objective perspective. BA and G-mean are metrics for evaluating models on imbalanced datasets, which reflect the balance between positive and negative predictions [39]. We also specify area under ROC curve (AUC) and area under the precision-recall curve (AUPRC) as the comprehensive predictive ability metrics. The AUPRC is a more useful metric for evaluating a model's ability to correctly classify positive samples when processing imbalanced datasets than AUROC, although it is rarely used in LSM evaluation [59]. Due to the differences in the use of precision and false positive rate (FPR) within the calculation of AUPRC and AUC, in imbalanced datasets where TN is very large and TP is very small, AUPRC is much more sensitive to changes in FP. The relevant formulas of these metrics are as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{1}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$\mathrm{sensitivity} = \mathrm{recall} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right) \tag{5}$$

$$G - \mathrm{mean} = \sqrt{\mathrm{sensitivity} * \mathrm{specificity}} \tag{6}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{7}$$

G. Experimental Setting

As advocated by the framework, due to the uncertainty that exist in unlabeled samples, unlabeled samples cannot be directly used as negative samples in the testing set. The testing set needs to consist of deterministic samples, but we cannot guarantee that any unlabeled sample must be a nonlandslide. Therefore, we propose a PU and imbalanced LSM testing method as the following steps show:

1) divide the landslide samples into training and testing sets;
2) to match the landslide samples in the training and testing set, unlabeled samples in the training and testing set need to be extracted from unlabeled samples without repetition independently and randomly;
3) train two prior model using landslide samples from the training and testing set, as well as matching unlabeled samples;
4) extract points from very low and low susceptibility (LS) areas of training and testing prior models as training and testing negative samples, and combine them with matched landslides to form imbalanced training and testing sets, respectively.

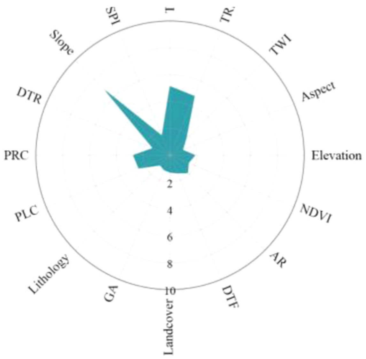

Fig. 6. VIF value distribution of the LCFs.

Separating landslide samples and unlabeled samples from the beginning in this testing method can effectively prevent information leakage between the training and testing sets. This proposed testing method can provide an objective and accurate evaluation of our proposed models.

In this research, the ratio of the training set to the testing set is 7:3. The unlabeled samples extracted by the prior model are three times the number of corresponding landslide samples, which is set by experience. In order to unify the measurement standards, an imbalanced testing set with a positive and negative sample ratio of 200:1 is fixed to evaluate all of the models, which is determined by the above-mentioned PU and imbalanced testing methods. In addition, different imbalanced ratios between positive and negative samples in the training are set from 1:2.5 to 1:200 for conducting saturation tests of the amounts of negative samples on our imbalanced models. In the training process, the hyperparameters of machine models are determined by grid search within the range we specify.

IV. Results and Analysis

A. Analysis of LCFs

The VIF values of all LCFS are shown in Fig. 6. The values of all the LCFs are smaller than 10, it can be inferred that no factor exhibits high multicollinearity. The VIF values of Slope and TPI are relatively large, with values of 6.80 and 5.13, respectively, showing slight multicollinearity. The VIF values of all other influencing factors are smaller than 5, indicating a low level of collinearity between them. Fig. 7 shows the importance of all

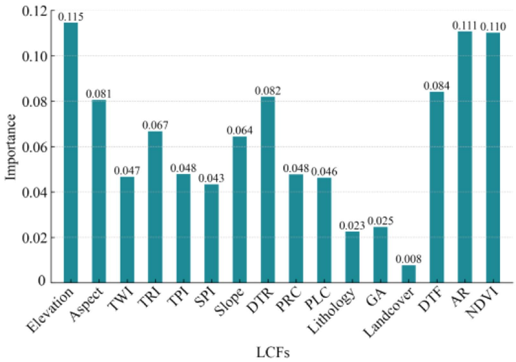

Fig. 7. Results of importance value of the LCFs.

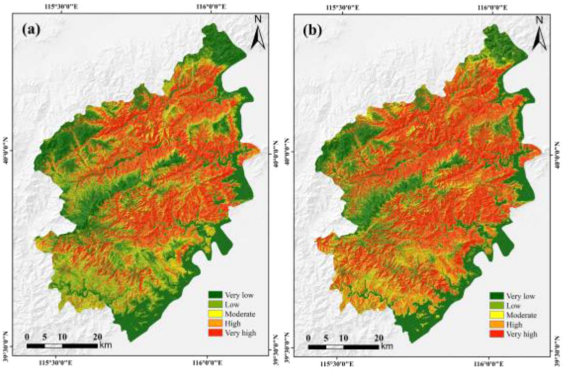

Fig. 8. Training and testing prior models generated by spy-algorithm. (a) Prior model generated by training set. (b) Prior mode generated by testing set. RN samples are selected from the very low and LS area of above models.

LCFs. The importance of Elevation, AR, and NDVI are relatively high, with values of 0.115, 0.111, and 0.110, respectively. Various terrain factors, fault and river have also certain importance. The importance of lithology and GA are relatively low, and landcover gets the lowest importance of 0.008. It can be seen that the terrain, vegetation, and hydrologic factors play a crucial role in landslide occurrence. Although the importance of landcover is small, it can still assist in distinguishing nonlandslide samples (such as built area and water), which means all LCFs have promoting effects on model prediction. Considering the absence of severe collinearity and redundant variables in LCFs, as well as the strong adaptability of the machine learning methods chosen next, we choose to retain all LCFs in the following modeling process.

B. *Improvement of the RN Samples on the Performance of LSM*

Fig. 8 shows the results of prior models generated by the spy algorithm using training and testing sets. After constructing the imbalanced testing set, the comparative tests are carried out to discover the improvement on the performance of LSM brought by RN samples using RF and DNN. In order to reduce stochasticity, 10 repeated random sampling of unlabeled samples and RN samples with the same quantity as landslide samples are conducted.

Table II presents the average test results obtained by training models using DNN with unlabeled samples (U-DNN), DNN with RN samples (RN-DNN), RF with unlabeled samples (U-RF), and RF with RN samples (RN-RF). U-DNN and U-RF are baseline models that do not use either PU methods or imbalanced learning methods. From unlabeled samples to RN samples, the metrics are improved mainly in sensitivity and AUPRC. From U-RF to RN-RF, AUPRC increased from 0.5810 to 0.6881, and sensitivity increased from 0.7833 to 0.9265. From U-DNN to RN-DNN, AUPRC increased from 0.1685 to 0.3528, and sensitivity increased from 0.8064 to 0.9072. There is also a small improvement in metrics such as BA, AUC, and G-mean in both DNN and RF. However, the use of RN samples resulted in a decrease in precision, specificity, and OA. In addition, whether on unlabeled samples or RN samples, the overall performance of RF is generally better than that of DNN.

From the difference in metrics in Table II, it can be inferred that the improvement of using RN samples is mainly reflected in the improvement of the predictive ability of positive samples (seen from the large improvement in sensitivity and AUPRC). OA and specificity have decreased, which is the side effect of the decrease in negative sample predictive ability while the positive sample prediction performance has increased. At the same time, because the negative samples in the testing set are the majority class, this side effect will cause OA to drop significantly. However, using RN samples can definitely improve the overall prediction performance of the model, because the comprehensive predictive performance metrics of the positive reflection model, BA, AUC, and G-mean, have been improved to a certain extent. In particular, BA and G-mean reflect the improvement of the model's trade-off between positive and negative predictive capabilities.

C. *Comparison of Different Imbalanced Learning Models*

A total of 12 different imbalance ratios from 1:200 to 1:2.5 are set to compare the four imbalanced learning models including SMOTE-DNN, ADASYN-DNN, BRF, and EE. In this section, we selected the largest ratio (1:200) and the smallest imbalance ratio (1:2.5) to show the performance difference between the four imbalance models. Figs. 9 and 10 display the box plots of test results of models trained on 10 sets of RN samples. Table III shows the mean values of the testing metrics of the benchmark models (RN-DNN and RN-RF) and imbalanced learning models under an imbalance ratio of 1:2.5 and 1:200.

In the ratio of 1:200, four imbalanced learning methods have consistently improved in precision, specificity, OA, BA, and G-mean than benchmark models as shown in Fig. 9. SMOTE-DNN achieved the highest precision and specificity (0.1093 and 0.9645), followed by ADASYN-DNN (0.0932 and 0.9572), BRF (0.0659 and 0.9361), and EE (0.0539 and 0.9204). The increase in precision and specificity compared to the benchmark models indicates an enhanced ability to predict negative classes and reduction of FP samples in imbalanced learning models. Since negative classes are the majority in the testing set, the OA follows a similar pattern to specificity. Regarding model balance, BRF achieved the highest BA and G-mean (0.9185 and 0.9184),

TABLE II
Average Test Results of DNN and RF Trained With Unlabeled Samples and RN Samples in 10 Times Sampling

Models	Precision	Sensitivity	Specificity	OA	BA	G-mean	AUPRC	AUC
U-DNN	0.0530	0.8064	0.9264	0.9258	0.8664	0.8642	0.1685	0.9470
U-RF	0.1673	0.7833	0.9805	0.9795	0.8819	0.8764	0.5810	0.9690
N-DNN	0.0398	0.9072	0.8882	0.8883	0.8977	0.8976	0.3528	0.9619
N-RF	0.0399	0.9265	0.8886	0.8888	0.9076	0.9074	0.6881	0.9720

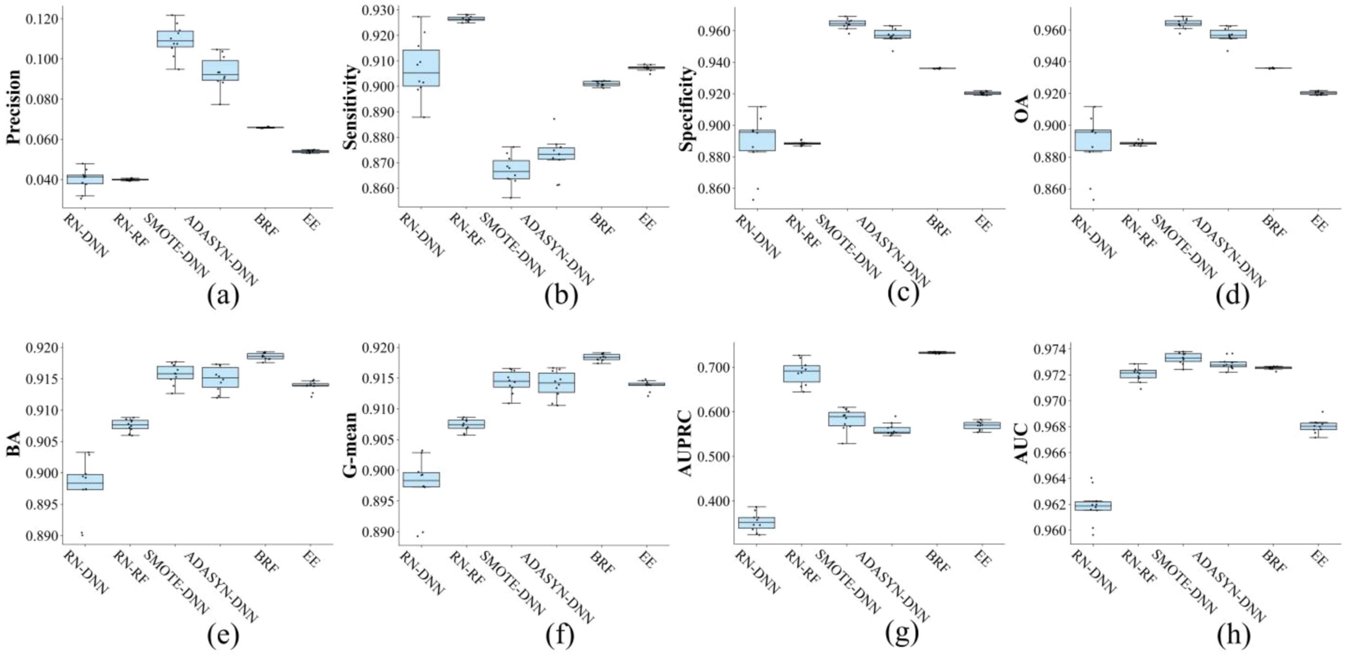

Fig. 9. Metrics results of benchmark models and imbalanced learning models with imbalance ratio of 1:200. (a) Precision. (b) Sensitivity. (c) Specificity. (d) OA. (e) BA. (f) G-mean. (g) AUPRC. (h) AUC.

TABLE III
Average Test Results of DNN and RF Trained With Unlabeled Samples and RN Samples in 10 Times Sampling

Ratio	Model	Precision	Sensitivity	Specificity	OA	BA	G-mean	AUPRC	AUC
1	RN-DNN	0.0398	0.9072	0.8882	0.8883	0.8977	0.8976	0.3528	0.9619
	RN-RF	0.0399	0.9265	0.8886	0.8888	0.9076	0.9074	0.6881	0.9720
2.5	SMOTE-DNN	0.0552	0.9032	0.9215	0.9214	0.9124	0.9123	0.4478	0.9692
	ADASYN-DNN	0.0401	0.9215	0.8865	0.8867	0.9040	0.9037	0.4170	0.9682
	BRF	0.0646	0.9002	**0.9348**	0.9346	**0.9175**	**0.9173**	**0.7157**	0.9720
	EE	0.0529	0.9061	0.9189	0.9188	0.9125	0.9125	0.5485	0.9667
200	SMOTE-DNN	0.1093	0.8669	**0.9645**	0.9640	0.9157	0.9144	0.5820	0.9733
	ADASYN-DNN	0.0932	0.8728	0.9572	0.9568	0.9150	0.9140	0.5606	0.9729
	BRF	0.0659	0.9010	0.9361	0.9359	**0.9185**	**0.9184**	**0.7326**	0.9725
	EE	0.0539	0.9072	0.9204	0.9204	0.9138	0.9138	0.5690	0.9680

The marked value is the highest in the group.

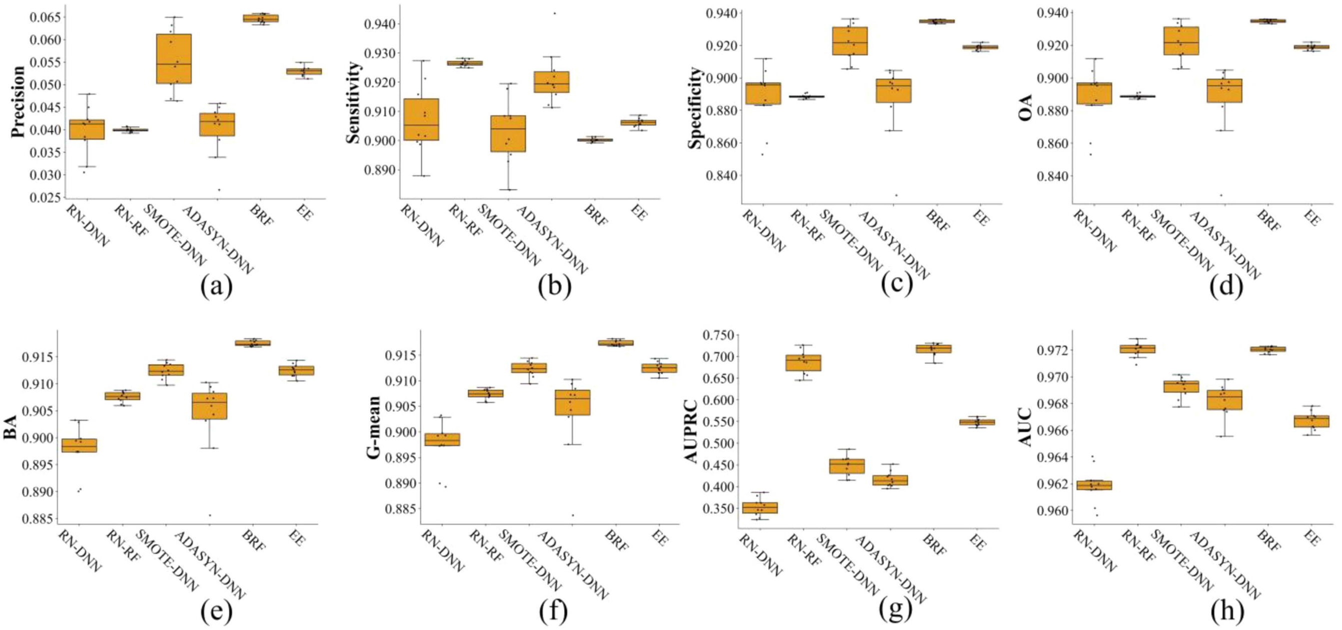

Fig. 10. Metrics results of benchmark models and imbalanced learning models with imbalance ratio of 1:2.5. (a) Precision. (b) Sensitivity. (c) Specificity. (d) OA. (e) BA. (f) G-mean. (g) AUPRC. (h) AUC.

followed by SMOTE-DNN (0.9157 and 0.9144), ADASYN-DNN (0.9150 and 0.9140), and EE (0.9138 and 0.9138). In terms of AUPRC, the DNN-based methods showed significant improvement through oversampling. Among the ensemble learning methods, BRF achieved the highest AUPRC (0.7326), whereas EE showed a decline (0.5690). Due to the extremely imbalanced nature of the testing set, the AUC values for the four models did not show significant changes. In addition, sensitivity and specificity show a trend of mutual restriction. Compared with the benchmark models, SMOTE-DNN and ADASYN_DNN decrease the most in sensitivity, while BRF and EE decrease less.

In the ratio of 1:2.5, it is observed that most of the metrics including precision, specificity, OA, BA, and G-mean show an improvement over the benchmark models, but the extent of the improvement is limited compared to the scenario in ratio of 1:200, and some models even exhibit a decline. BRF achieves the highest precision and specificity (0.0646 and 0.9348), followed by SMOTE-DNN (0.0552 and 0.9215), EE (0.0529 and 0.9189), and ADASYN-DNN (0.0401 and 0.8865). In terms of model balance, both BA and G-mean are higher for SMOTE-DNN and ADASYN-DNN compared to the benchmark RN-DNN, while BRF and EE are higher than benchmark RN-RF. BRF has the highest balance (0.9175 and 0.9173). In the oversampling methods and imbalanced learning methods, AUPRC shows an improvement compared to their respective benchmarks, except for the EE model. BRF still exhibits the best performance in AUPRC (0.7157).

The above-mentioned analysis highlights that imbalanced learning significantly enhances the model's predictive capability for negative samples, while also improves precision and maintains a good level of sensitivity in positive instances, thereby enhancing model balance. Simultaneously, there is a noticeable reduction in FPRs (referred in AUPRC). BRF achieves the best balance in both the maximum and minimum ratios. Based on oversampling, DNN models can exhibit stronger predictive power for negative classes compared to ensemble learning when given sufficient samples. However, oversampling-based DNN models are highly susceptible to the influence of imbalanced ratios. In addition, SMOTE outperforms the ADASYN method across most of the metrics.

D. Exploring the Effects of Different Imbalanced Ratio on Imbalanced Learning Model

In order to investigate the impact of imbalanced ratios on imbalanced learning models, we set up 12 groups of imbalanced ratios: 1:2.5, 1:5, 1:7.5, 1:10, 1:15, 1:20, 1:25, 1:50, 1:100, 1:150, and 1:200. This process is referred to as saturation test, aimed at exploring at what point the model performance reaches a bottleneck as the number of negative samples increases. The higher density of ratio settings in the earlier stages is aimed at capturing the significant variations in model performance during this phase. For each ratio, the models are trained on 10 sets of data, which are repeatedly sampled and tested on the fixed 1:200 imbalanced testing set.

The average evaluation results of precision, sensitivity, specificity, G-mean, AUPRC, and AUC of imbalanced learning models trained with different imbalanced ratios and benchmark models on imbalanced testing set are shown in Fig. 11. The results of OA and BA are omitted because they are very close to specificity and G-mean respectively and have similar meanings. In oversampling methods, as the imbalance ratio increases, the

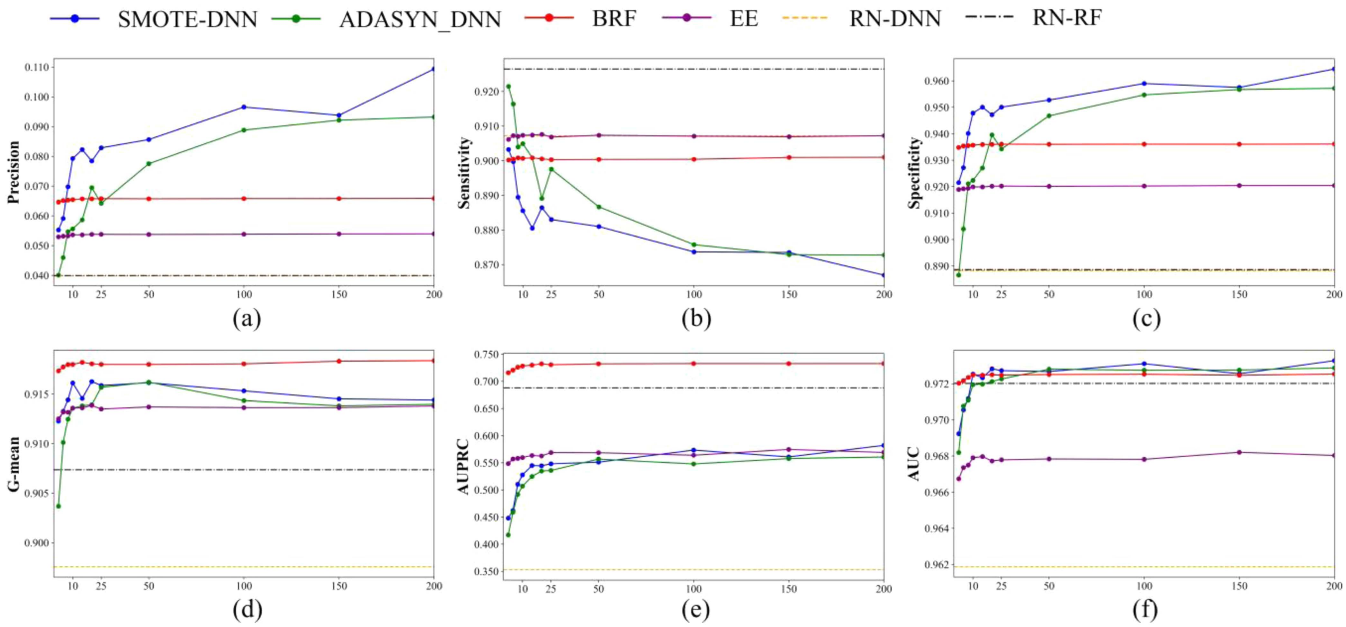

Fig. 11. Mean values of metrics results of benchmark models and imbalanced learning models during saturation test of different imbalance ratios. (a) Precision. (b) Sensitivity. (c) Specificity. (d) G-mean. (e) AUPRC. (f) AUC.

precision, specificity, G-mean, and AUPRC of the oversampling models increase significantly, and the sensitivity decreases significantly. This change is most obvious, especially between the ratio of 1:2.5 and 1:25. When the ratio exceeds 1:25, the G-mean shows a slight decline and AUPRC tends to be stable. This trend is also slightly reflected in the ensemble learning imbalanced models (BRF and EE). Metrics of ensemble learning tend to be stable after the imbalance ratio reaches 1:10. Ensemble learning especially the BRF is insensitive to fluctuations in the imbalance ratio, and performs at similar levels at lower and larger ratios. The performance of BRF and EE can stabilize after the imbalance ratio reaches a threshold. After reaching the threshold, although G-mean and AUPRC tend to be stable, the changes in precision, sensitivity, and specificity are still obvious in oversampling models.

In the aspect of models' balance (G-mean), there are imbalanced peaks existing in the imbalanced models. Comparing the above four imbalanced models, BRF can always maintain a better balance than the benchmark models in different ratios, followed by SMOTE-DNN, ADASYN-DNN, and EE. In terms of AUC, although the differences among the four models are small, the other models can outperform RN-DNN after reaching a certain ratio threshold except for EE.

E. Comparison of LSM Results

Fig. 12 shows the LSM results of the balanced models using unlabeled samples (U-DNN and U-RF), the balanced models using RN samples (RN-DNN and RN-RF), and four imbalanced learning models with an imbalance ratio of 1:200. In order to provide a more quantitative distribution pattern of susceptibility levels, Fig. 13 shows the distribution percentages of very low susceptibility (VLS), LS, moderate susceptibility (MS), high susceptibility (HS), and very HS (VHS) for each model. When using U-DNN and U-RF, the VHS area is significantly smaller, and VHS of RF is smaller than DNN's. When reliable samples are used, the VHS area of RN-DNN and RN-RF increases significantly, and the VLS and LS areas also decrease slightly. This can be also seen from the changes in Sensitivity and specificity in Table II. The use of RN samples in the training process significantly increases the model's predictive ability for positive samples. When the imbalanced learning model is compared with the balanced benchmark models RN-DNN and RN-RF models, the most intuitive change is reflected in the significant increase in the area ratio of VLS and LS. The area ratio of VHS is slightly reduced in imbalanced learning models, too. The oversampling models ADASYN-DNN and SMOTE-DNN can provide a higher proportion of VLS and LS than the ensemble learning BRF and EE. This is the positive benefit brought by the improvement of specificity and precision through oversampling models. Introducing a large number of RN samples can better identify VLS and LS regions, reduce the false prediction of false positive samples while maintaining a good level of reliable prediction for positive classes.

The high scores of AUC indicate that the models can correctly classify most of the samples. The slight differences in metrics come from the ability to correctly classify the ambiguous samples. To highlight the comparison between models, focus areas where the models have significant disagreements are manually selected. On one hand, these areas can highlight the models' ability to handle ambiguous samples. On the other hand, these areas are small and abnormal due to divergence, which cannot reflect the overall prediction performance of the region. Fig. 14 shows the comparison of LSM results of three selected subfocus areas. It can be seen that U-DNN and U-RF easily misjudge VHS

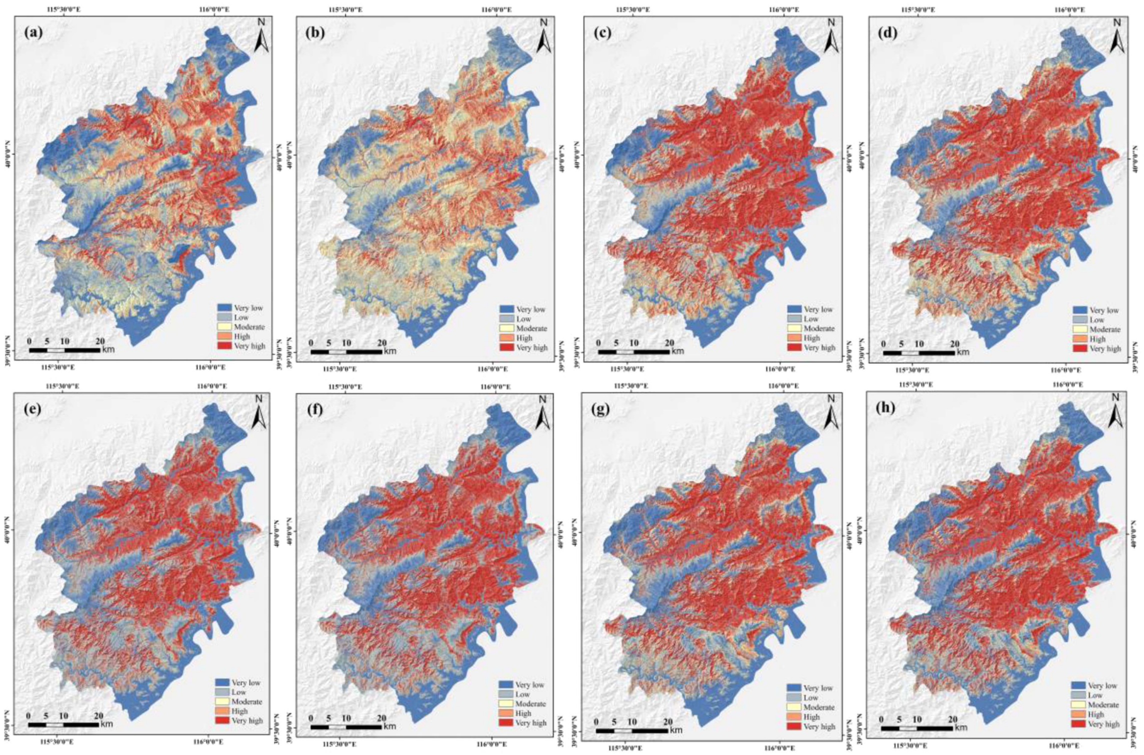

Fig. 12. LSM results of all models. (a) U-DNN. (b) U-RF. (c) RN-DNN. (d) RN-RF. (e) SMOTE-DNN. (f) ADASYN-DNN. (g) BRF. (h) EE. The imbalance ratios of (e)–(h) are 1:200.

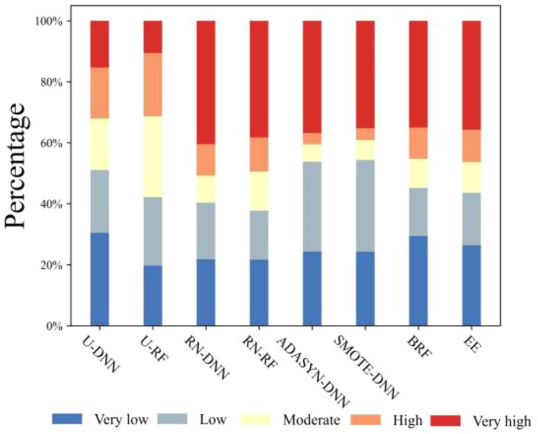

Fig. 13. Distribution percentages of VHS, LS, MS, HS, and VHS of each model.

as VLS, LS, or MS. After using RN samples, this misjudgment is greatly corrected. After further using the imbalanced learning methods, most of the correctly predicted VLS areas are maintained, and the prediction of stable areas, namely VLS and LS, is enhanced. In the oversampling method, as the proportion of VLS and LS areas increases significantly, it is possible to generate a false prediction of FN samples, while ensemble learning can maintain a correct judgement.

In summary, the combination of the PU learning and imbalanced learning models can improve the prediction accuracy of VHS, VLS, and LS. The ambiguous MS and HS areas in the middle level are reduced, which has a clearer indicative significance for disaster prevention and mitigation.

V. Discussions

In this study, a novel framework combining PU learning and imbalanced learning is proposed for accurate LSM work and tested in a regional rainfall-induced landslides dataset in Beijing, China. This framework demonstrates its advanced nature in terms of principles and results, especially in its ability to uncover underlying laws and make accurate predictions in complex scenarios. Through the logic of this framework, two common modeling problems are identified in most LSM works: uncertainty in negative samples and the imbalanced nature of positive and negative classes. To solve these problems, we utilize the spy PU learning model to obtain RN samples and incorporate four imbalanced learning models. The ultimate goal is to improve the model's predictive capabilities for both positive and negative samples, resulting in an improved balance of the model. In terms of model selection rationale, we have chosen four representative and easily implementable models considering data-level and ensemble learning as two mainstream directions for imbalanced learning. As this framework is an initial attempt, we also encourage the exploration and comparison of various other PU or imbalanced learning models.

A progressive testing procedure is used to justify the development of this framework. It can be found in the comparation of U-DNN, U-RF, RN-DNN, and RN-RF that the inclusion

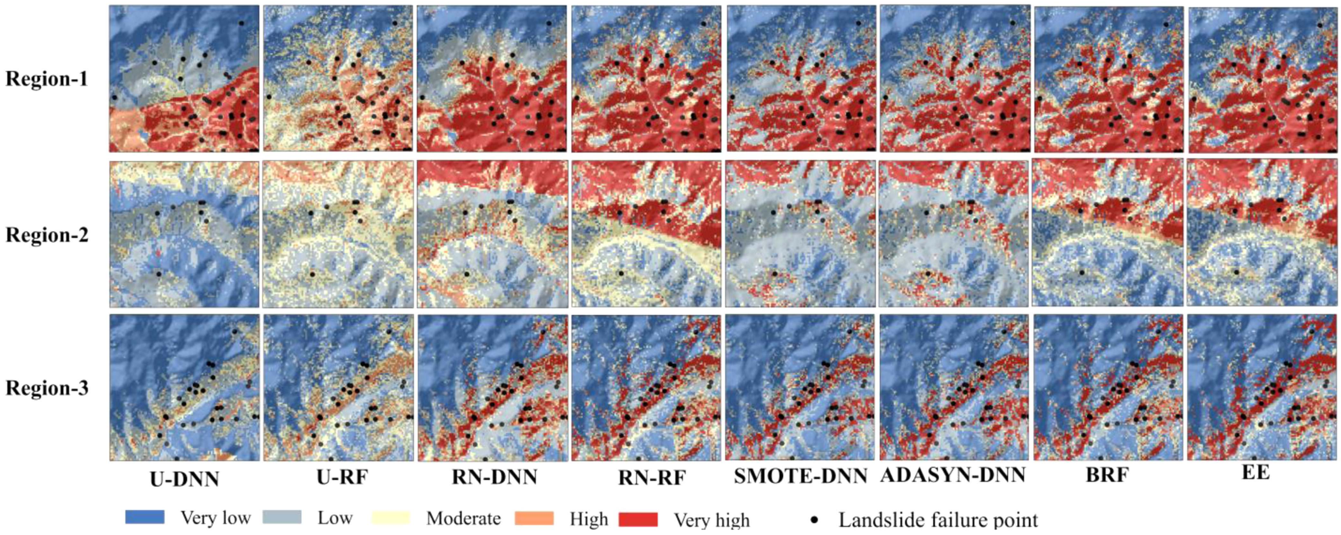

Fig. 14. Comparison of LSM results of some selected subfocus regions with divergence generated by different models. In these regions: landslide failure sites in U-DNN and U-RF are not easily identified as VHS, while in RN-DNN and RN-RF models, landslide failure sites are more easily corrected to VHS; the support of the imbalanced learning methods can maintain the correct identification of VHS while enhancing the identification of VLS and LS around landslides.

of RN samples can enhance the predictive ability of positive classes. In the comparation of four imbalanced learning models and the balanced benchmark models of RN-DNN and RN-RF, the predictive ability for negative samples is enhanced while maintaining the good level of predicting the positive samples by imbalanced learning. This framework enhances both positive and negative predictions, ultimately resulting in an improved balance of the model (G-mean and BA). When comparing imbalanced models and conducting saturation tests on imbalanced ratios, it is observed that the balance of four models has increased significantly comparing the benchmark models, with BRF being the most stable. It is found that balance peaks exist in the varying imbalanced ratios, while the balance of ensemble learning models BRF and EE maintain a stable state, the balance of the oversampling methods SMOTE and DNN decreases after this threshold. This phenomenon has been mentioned in some studies that the oversampling method can generate a large number of false landslide samples [39]. Once the false landslide samples overwhelm the true landslide samples, it can lead to the model deviating from predicting the positive class. This phenomenon is closely related to the decrease in sensitivity in Fig. 11. Taking this into account, using imbalanced models advocates the utilization of imbalanced ratios up to the balance peaks in LSM work. This can provide the model with sufficient negative sample information and have a positive impact on the prediction. For ensemble imbalanced learning models, the ratio can exceed this peak due to their stability. However, for oversampling models, it may not be appropriate to increase the ratio after reaching its balance peak. Nonetheless, employing BRF is both convenient and stable.

In the field of LSM research, there is a lack of attention given to model balance. Some LSM models that lose balance can also show extremely high accuracy, but this may be due to blindly overestimating the level of susceptibility [23]. Such predicted results do not align with reality, while accurate assessment of VLS and LS is neglected. The main reasons for this phenomenon are the lack of imbalanced modeling ideas and unified imbalanced model evaluation criteria. On the one hand, only using negative samples equivalent to the number of landslides makes it difficult to reflect the true situation of negative class predictive ability in both the training and testing process. In a specific study area, there are a large number of unlabeled samples that are not entirely without value. In fact, learning these unlabeled samples helps the model comprehend the overall situation in the area and enables it to make more reasonable assessments. On the other hand, AUC is the most commonly used metric in LSM evaluation, but this metric is quite insensitive to the balance of models in this study and other imbalanced learning field [59]. Some studies have used imbalanced testing set and some useful imbalanced evaluation metrics like G-mean, sensitivity, and specificity for evaluating the LSM model's performance [33], [39]. The use of imbalanced testing set and suitable evaluation metrics urgently requires to become a standard procedure in LSM modeling. We simply used BA and G-mean to evaluate the balance of the models and also monitored other indicators like AUPRC. Innovative and convenient metrics are needed in future LSM works.

There are still some limitations and points to be studied in this study. We only use DNN due to the huge amount of computation brought by oversampling. The performance of oversampling combined with other algorithms in combination with algorithm optimization and hardware improvement remains to be tested. Besides, whether more advanced PU methods and imbalanced learning models can improve the performance of this framework remains to be compared and tested in the future. In addition, in this study, framework of PU learning and imbalanced learning are divided into two parts, a one-step efficient model that combines the advantages of both is yet to be developed. The selection of imbalanced ratios requires the adaptation of intelligent methods. More importantly, convolutional networks are becoming more accurate and commonly used in LSM, and the applicability of this framework to convolutional modeling remains to be tested. This study demonstrates that the framework has strong classification capabilities for binary classification problems, making it suitable for addressing the complex patterns of landslides induced by heavy rainfall. In theory, the framework can be applicable to landslide datasets from different regions with different causes. In the future, further testing on similar landslide datasets or other types of landslide datasets (like earthquake-induced landslides) will be necessary to validate its effectiveness. Our next steps will focus on the generalized application of this model across landslide datasets from different regions and with various causes. Ultimately, we aim to develop a transfer learning model based on this framework to predict real-world landslides in advance. Similar PU and imbalanced binary classification problems similar to regional landslides (not limited to geological disasters) can also consider adopting a TSA based on this LSM framework: 1) use PU methods to construct a prior model to screen out RN samples from the unlabeled sample set; 2) train imbalanced learning models to obtain reliable predictors.

VI. Conclusion

In this article, we propose a novel framework combining PU learning and imbalanced learning, aiming to solve the problems of uncertainty in negative samples and imbalance between positive and negative classes in the LSM modeling process. We used the spy algorithm to generate the prior model and obtain RN samples. We then deployed four imbalanced learning models from the oversampling method and ensemble learning method including SMOTE-DNN, ADASYN-DNN, BRF, and EE to process the imbalanced training and testing sets. This framework significantly improves the predictive performance of both positive and negative classes in LSM. We have tested our LSM framework on a dataset of regional rainfall-induced landslides in Beijing, China, and conducted detailed comparisons using benchmark models (DNN and RF) and saturation tests of imbalanced ratios on the imbalanced training and testing sets. The main findings of the study are as follows.

1) By using the RN samples generated by a prior model based on the spy algorithm instead of randomly selecting unlabeled samples, the overall predictive performance and balance of RN-DNN and RN-RF have been enhanced, especially the predictive ability of positive samples.
2) Based on the use of imbalanced learning methods, the ability for predicting negative classes is significantly enhanced. Moreover, the ability for predicting positive classes can be maintained at a good level. As the imbalance ratio increases, imbalanced models reach a point of saturation on balance (balance peaks) after reaching a certain threshold. Beyond these thresholds, the balance and other metrics of ensemble imbalanced learning models (BRF and EE) tend to be stable. However, the balance of the oversampling models (SMOTE-DNN and ADASYN-DNN) slightly decreases and the predictive ability of the positive class also decreases after exceeding these thresholds. Overall, BRF is the best performing and stable model in this study.
3) By combining the above PU technique and imbalanced learning, the predictive capabilities of both positive and negative classes can be improved as well as the overall balance of the models can be improved simultaneously. This model helps to address the issue of overestimation or misjudgment of susceptibility, resulting in a more accurate and balanced susceptibility assessment.

References

[1] F. Wang, X. Peng, G. Zhu, K. Nam, Y. Chen, and K. Yan, “The Hongchi landslide triggered by heavy rainfall from super typhoon In-Fa on 25 July 2021 in Hangzhou City, Zhejiang Province, China,” *Bull. Eng. Geol. Environ.*, vol. 81, no. 10, Sep. 2022, Art. no. 411.

[2] S. L. Gariano and F. Guzzetti, “Landslides in a changing climate,” *Earth-Sci. Rev.*, vol. 162, pp. 227–252, Nov. 2016.

[3] F. Wang et al., “Impact of a clustered rainfall-induced geo-hydrological disaster on densely populated gully villages in Fuyang District, Hangzhou City, Zhejiang Province, China on 22 July 2023,” *Landslides*, vol. 21, no. 5, pp. 1149–1154, May 2024.

[4] J. Corominas et al., “Recommendations for the quantitative analysis of landslide risk,” *Bull. Eng. Geol. Environ.*, vol. 73, no. 2, pp. 209–263, May 2014.

[5] G. J. Hearn and A. B. Hart, “Landslide susceptibility mapping: A practitioner’s view,” *Bull. Eng. Geol. Environ.*, vol. 78, no. 8, pp. 5811–5826, Dec. 2019.

[6] W. Jiang et al., “Deep learning for landslide detection and segmentation in high-resolution optical images along the Sichuan-Tibet transportation corridor,” *Remote Sens.*, vol. 14, no. 21, Jan. 2022, Art. no. 5490.

[7] J. Wandong, X. I. Jiangbo, L. I. Zhenhong, D. Mingtao, Y. Ligong, and X. I. E. Dashuai, “Landslide detection and segmentation using Mask R-CNN with simulated hard samples,” *Geomatics Inf. Sci. Wuhan Univ.*, vol. 48, no. 12, pp. 1931–1942, Dec. 2023.

[8] S. Gao et al., “Optimal and multi-view strategic hybrid deep learning for old landslide detection in the loess plateau, Northwest China,” *Remote Sens.*, vol. 16, no. 8, Jan. 2024, Art. no. 1362.

[9] A. Merghadi et al., “Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance,” *Earth-Sci. Rev.*, vol. 207, Aug. 2020, Art. no. 103225.

[10] H. Wang, L. Zhang, H. Luo, J. He, and R. W. M. Cheung, “AI-powered landslide susceptibility assessment in Hong Kong,” *Eng. Geol.*, vol. 288, Jul. 2021, Art. no. 106103.

[11] A. Kaushal and V. K. Sehgal, “Landslide susceptibility detection using ResNet,” in *Proc. 3rd Asian Conf. Innov. Technol.*, Aug. 2023, pp. 1–5.

[12] T. Chen, Q. Wang, Z. Zhao, G. Liu, J. Dou, and A. Plaza, “LCFSTE: Landslide conditioning factors and swin transformer ensemble for landslide susceptibility assessment,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6444–6454, 2024.

[13] A. Shirzadi et al., “Uncertainties of prediction accuracy in shallow landslide modeling: Sample size and raster resolution,” *CATENA*, vol. 178, pp. 172–188, Jul. 2019.

[14] F. Huang et al., “Modelling landslide susceptibility prediction: A review and construction of semi-supervised imbalanced theory,” *Earth-Sci. Rev.*, vol. 250, Jan. 2024, Art. no. 104700.

[15] F. Huang, K. Yin, J. Huang, L. Gui, and P. Wang, “Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine,” *Eng. Geol.*, vol. 223, pp. 11–22, Jun. 2017.

[16] Y. Li, X. Deng, P. Ji, Y. Yang, W. Jiang, and Z. Zhao, “Evaluation of landslide susceptibility based on CF-SVM in Nujiang prefecture,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 21, Jan. 2022, Art. no. 14248.

[17] C. Ye, R. Tang, R. Wei, Z. Guo, and H. Zhang, “Generating accurate negative samples for landslide susceptibility mapping: A combined self-organizing-map and one-class SVM method,” *Front. Earth Sci.*, vol. 10, 2023, Art. no. 1054027.

[18] Y. W. Rabby, Y. Li, and H. Hilafu, “An objective absence data sampling method for landslide susceptibility mapping,” *Sci. Rep.*, vol. 13, no. 1, Jan. 2023, Art. no. 1740.

[19] R. Zhang et al., "Interferometric synthetic aperture radar (InSAR)-based absence sampling for machine-learning-based landslide susceptibility mapping: The three gorges reservoir area, China," *Remote Sens.*, vol. 16, no. 13, Jan. 2024, Art. no. 2394.

[20] C. Xi et al., "Effectiveness of newmark-based sampling strategy for coseismic landslide susceptibility mapping using deep learning, support vector machine, and logistic regression," *Bull. Eng. Geol. Environ.*, vol. 81, no. 5, Apr. 2022, Art. no. 174.

[21] X. Wei et al., "Comparison of hybrid data-driven and physical models for landslide susceptibility mapping at regional scales," *Acta Geotechnica*, vol. 18, pp. 4453–4476, Mar. 2023.

[22] S. Liu et al., "A physics-informed data-driven model for landslide susceptibility assessment in the three gorges reservoir area," *Geosci. Front.*, vol. 14, no. 5, Sep. 2023, Art. no. 101621.

[23] Z. Fu, F. Wang, J. Dou, K. Nam, and H. Ma, "Enhanced absence sampling technique for data-driven landslide susceptibility mapping: A case study in Songyang County, China," *Remote Sens.*, vol. 15, no. 13, Jan. 2023, Art. no. 3345.

[24] A.-X. Zhu et al., "A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods," *Catena*, vol. 183, 2019, Art. no. 104188.

[25] J. Yao, S. Qin, S. Qiao, X. Liu, L. Zhang, and J. Chen, "Application of a two-step sampling strategy based on deep neural network for landslide susceptibility mapping," *Bull. Eng. Geol. Environ.*, vol. 81, no. 4, Mar. 2022, Art. no. 148.

[26] Z. Fang, Y. Wang, R. Niu, and L. Peng, "Landslide susceptibility prediction based on positive unlabeled learning coupled with adaptive sampling," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11581–11592, 2021.

[27] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[28] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 79:1–79:36, Aug. 2019.

[29] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: A review," *Int. J. Comput. Bus. Res.*, vol. 5, no. 4, pp. 1–29, 2014.

[30] W. Li, R. Huang, C. Tang, Q. Xu, and C. van Westen, "Co-seismic landslide inventory and susceptibility mapping in the 2008 Wenchuan earthquake disaster area, China," *J. Mountain Sci.*, vol. 10, no. 3, pp. 339–354, Jun. 2013.

[31] F. Wang et al., "Coseismic landslides triggered by the 2018 Hokkaido, Japan (Mw 6.6), earthquake: Spatial distribution, controlling factors, and possible failure mechanism," *Landslides*, vol. 16, no. 8, pp. 1551–1566, Aug. 2019.

[32] C. Xie, Y. Huang, L. Li, T. Li, and C. Xu, "Detailed inventory and spatial distribution analysis of rainfall-induced landslides in Jiexi County, Guangdong Province, China in august 2018," *Sustainability*, vol. 15, no. 18, Jan. 2023, Art. no. 13930.

[33] Q. Liu, A. Tang, and D. Huang, "Exploring the uncertainty of landslide susceptibility assessment caused by the number of non–landslides," *CATENA*, vol. 227, Jun. 2023, Art. no. 107109.

[34] K. Nam and F. Wang, "The performance of using an autoencoder for prediction and susceptibility assessment of landslides: A case study on landslides triggered by the 2018 Hokkaido Eastern Iburi earthquake in Japan," *Geoenviron Disasters*, vol. 6, no. 1, Dec. 2019, Art. no. 19.

[35] Y. Song et al., "Landslide susceptibility mapping based on weighted gradient boosting decision tree in Wanzhou section of the three gorges reservoir area (China)," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 1, Jan. 2019, Art. no. 4.

[36] L. Tang, X. Yu, W. Jiang, and J. Zhou, "Comparative study on landslide susceptibility mapping based on unbalanced sample ratio," *Sci. Rep.*, vol. 13, no. 1, Apr. 2023, Art. no. 5823.

[37] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 04, pp. 687–719, 2009.

[38] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 4, pp. 42–47, 2012.

[39] S. K. Gupta and D. P. Shukla, "Handling data imbalance in machine learning based landslide susceptibility mapping: A case study of Mandakini River Basin, North-Western Himalayas," *Landslides*, vol. 20, no. 5, pp. 933–949, May 2023.

[40] Y. Song et al., "Evaluating landslide susceptibility using sampling methodology and multiple machine learning models," *ISPRS Int. J. Geo-Inf.*, vol. 12, no. 5, May 2023, Art. no. 197.

[41] X. Wu and J. Wang, "Application of bagging, boosting and stacking ensemble and EasyEnsemble methods for landslide susceptibility mapping in the three Gorges Reservoir Area of China," *Int. J. Environ. Res. Public Health*, vol. 20, no. 6, Jan. 2023, Art. no. 4977.

[42] 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, May 2018.

[43] F. C. Dai and C. F. Lee, "A spatiotemporal probabilistic modelling of storm-induced shallow landsliding using aerial photographs and logistic regression," *Earth Surf. Processes Landforms*, vol. 28, no. 5, pp. 527–545, 2003.

[44] C.-T. Lee, C.-C. Huang, J.-F. Lee, K.-L. Pan, M.-L. Lin, and J.-J. Dong, "Statistical approach to storm event-induced landslides susceptibility," *Natural Hazards Earth Syst. Sci.*, vol. 8, no. 4, pp. 941–960, Aug. 2008.

[45] Z. Zhao, T. Chen, J. Dou, G. Liu, and A. Plaza, "Landslide susceptibility mapping considering landslide local-global features based on CNN and transformer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7475–7489, 2024.

[46] T. Chen et al., "BisDeNet: A new lightweight deep learning-based framework for efficient landslide detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3648–3663, 2024.

[47] F. S. Tehrani, G. Santinelli, and M. Herrera Herrera, "Multi-regional landslide detection using combined unsupervised and supervised machine learning," *Geomatics, Natural Hazards Risk*, vol. 12, no. 1, pp. 1015–1038, Jan. 2021.

[48] A. Kaboutari, J. Bagherzadeh, and F. Kheradmand, "An evaluation of two-step techniques for positive-unlabeled learning in text classification," *Int. J. Comput. Appl. Technol. Res.*, vol. 3, no. 9, pp. 592–594, Sep. 2014.

[49] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 387–394.

[50] K. Yu, Y. Liu, L. Qing, B. Wang, and Y. Cheng, "Positive and unlabeled learning for user behavior analysis based on mobile internet traffic data," *IEEE Access*, vol. 6, pp. 37568–37580, 2018.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[52] C.-R. Wang and X.-H. Shao, "An improving majority weighted minority oversampling technique for imbalanced classification problem," *IEEE Access*, vol. 9, pp. 5069–5082, 2021.

[53] I. Dey and V. Pratap, "A comparative study of SMOTE, borderline-SMOTE, and ADASYN oversampling techniques using different classifiers," in *Proc. 3rd Int. Conf. Smart Data Intell.*, 2023, pp. 294–302.

[54] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases: PKDD 2003*, vol. 2838, N. Lavrač, D. Gamberger, L. Todorovski, and H. Blockeel, Eds. Berlin, Germany: Springer, 2003, pp. 107–119.

[55] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, 2008, pp. 1322–1328.

[56] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *Univ. California, Berkeley*, vol. 110, no. 1–12, 2004, Art. no. 24.

[57] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[58] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Japanese Soc. Artif. Intell.*, vol. 14, no. 771–780, 1999, Art. no. 1612.

[59] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big data," *J. Big Data*, vol. 10, no. 1, Apr. 2023, Art. no. 42.

Zijin Fu received B.E. degree in geological engineering from the China University of Geosciences, Wuhan, China, in 2022. He is currently working toward the Ph.D. degree in geological resource and geological engineering with Tongji University, Shanghai, China.

His research interests include AI application in landslide science, landslide susceptibility mapping, and remote sensing.

Hao Ma received the M.S. degree majoring in civil engineering from the Chongqing University, Chongqing, China, in 2019. He is currently working toward the Ph.D. degree in geological engineering with the Tongji University, Shanghai, China.

His current main research interests include landslide mapping, rock slope engineering, and deep-seated slope gravitational deformations based on remote sensing technology.

Fawu Wang received the Ph.D. degree in geophysics from the Kyoto University, Kyoto, Japan, in 1999.

He is Chairholder of UNESCO Chair on Geoenvironmental Disaster Reduction, Professor on geodisaster reduction with School of Civil Engineering, Tongji University, Shanghai, China, and Emeritus Professor with Shimane University, Matsue, Japan. His primary research interests include to clarify the common mechanisms of landslides initiated by different triggers such as earthquake, rainfall, water level variation, etc., and to find a way to predict the occurrence and motion of landslides, for the purpose of landslide disaster mitigation.

Jie Dou received the Ph.D. degree in natural environmental studies from the University of Tokyo, Tokyo, Japan, in 2015.

He is a Professor with China University of Geosciences, Wuhan, China. He is the Director of the Big Data and Intelligent Disaster Prevention Center. His research interests include AI-based Big Data for geological disaster prevention, numerical simulation, and the mechanisms of geological disaster.

Bo Zhang received the B.E. degree in geological engineering, in 2022, from Tongji University, Shanghai, China, where he is currently working toward the Ph.D. degree in geological resource and geological engineering.

His research interests focus on studying the mechanisms of geo-disasters through field investigation, laboratory experiment, and remote sensing.

Zhice Fang received the M.S. degree in geoinfomatics, in 2020, from the 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.