sustainability

MDPI

Article

Landslide Susceptibility Mapping in Guangdong Province, China, Using Random Forest Model and Considering Sample Type and Balance

Li Zhuo [1,2,3], Yupu Huang [1,2], Jing Zheng [4,*], Jingjing Cao [1,2] and Donghu Guo [1,5]

1 Guangdong Provincial Key Laboratory of Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510006, China; zhuoli@mail.sysu.edu.cn (L.Z.); huangyp66@mail2.sysu.edu.cn (Y.H.); caojj5@mail.sysu.edu.cn (J.C.); guodh5@mail2.sysu.edu.cn (D.G.)
2 Guangdong Provincial Engineering Research Center for Public Security and Disaster, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510006, China
3 Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519000, China
4 Guangdong Climate Center, Guangzhou 501641, China
5 Department of Earth Science and Engineering, Imperial College London, London SW7 2BX, UK
* Correspondence: zhengjing@gd121.cn

Abstract: Landslides pose a serious threat to human lives and property. Accurate landslide susceptibility mapping (LSM) is crucial for sustainable development. Machine learning has recently become an important means of LSM. However, the accuracy of machine learning models is limited by the heterogeneity of environmental factors and the imbalance of samples, especially for large-scale LSM. To address these problems, we created an improved random forest (RF)-based LSM model and applied it to Guangdong Province, China. First, the RF-based LSM model was constructed using rainfall-induced landslide samples and 13 environmental factors and by exploring the optimal positive-to-negative and training-to-test sample ratios. Second, the performance of the RF-based LSM model was evaluated and compared with three other machine learning models. The results indicate that: (1) the proposed RF-based model has the best performance with the highest area under curve (AUC) of 0.9145, based on optimal positive-to-negative and training-to-test sample ratios of 1:1 and 8:2, respectively; (2) the introduction of rainfall and global human modification (GHM) can increase the AUC from 0.8808 to 0.9145; and (3) rainfall and topography are two dominant factors in Guangdong landslides. These findings can facilitate landslide risk prevention and serve as a technical reference for large-scale accurate LSM.

Keywords: landslide susceptibility mapping (LSM); machine learning; random forest (RF); sample balance; rainfall-induced landslide; global human modification (GHM)

Citation: Zhuo, L.; Huang, Y.; Zheng, J.; Cao, J.; Guo, D. Landslide Susceptibility Mapping in Guangdong Province, China, Using Random Forest Model and Considering Sample Type and Balance. *Sustainability* **2023**, *15*, 9024. https://doi.org/10.3390/su15119024

Academic Editors: Luqi Wang, Lin Wang, Yankun Wang, Ting Xiao and Zhiyong Liu

Received: 13 May 2023
Revised: 31 May 2023
Accepted: 1 June 2023
Published: 2 June 2023

1. Introduction

As one of the main types of geological hazards, landslides refer to the sliding of rock and soil masses along a slope. They are generally influenced by both natural factors and human interference and are characterized by wide distribution and frequent occurrence [1]. Landslide disasters in China have caused considerable casualties and property damage due to their sudden and unpredictable nature [2–4]. During the period from 2002 to 2014, a total of 246,768 landslides occurred in China, which led to the injury or death of 17,296 people and direct economic losses of CNY 58.75 billion [5]. It is evident that landslide disasters have imposed constraints on sustainable development [6,7]. The International Consortium on Landslides (ICL) has advocated for harmonizing regulations of the Sendai Landslide Partnerships 2015–2025 and the 2030 Agenda Sustainable Development Goals, among others [7]. Understanding where landslides can occur is, therefore, of great importance for reducing casualties, addressing ecological and economic losses, and promoting sustainable development [8]. This understanding generally includes two main aspects: landslide

susceptibility assessment (LSA) and landslide susceptibility mapping (LSM). The LSA refers to obtaining the probability of landslide occurrence at a certain location under certain environmental conditions [9]. It is the basis of landslide hazard assessments and landslide risk assessments [1,3,10,11]. LSM is generated by the LSA. It can rapidly visualize and determine the distribution patterns of landslide susceptibility in a region using geographic information system (GIS) and remote sensing (RS) technologies. More importantly, an accurate LSM plays an essential role in identifying such disasters and developing principles for sustainable development [8,12,13]. However, there are many challenges in optimizing LSM due to the complex mechanism and many uncertainties involved in landslide disasters [14–17].

Recently, LSM studies have shifted from empirically driven qualitative analysis to knowledge-driven and data-driven quantitative analysis [18]. Among these quantitative studies, two types of models are often used, namely deterministic models and statistical models [5,19]. Deterministic models mainly adopt the traditional slope damage mechanics-based model and basic spatial data to estimate the probability of landslide occurrence [5]. These models rely on large amounts of data on geology and hydrology and, thus, are often used in small-scale studies [19]. The statistical model, a type of non-deterministic model, adopts statistical methods to determine the relationship between historical landslides and environmental factors and to estimate the probability of landslides on a regional scale [5,19]. Compared with deterministic models, statistical models do not require a large amount of data on the physical characteristics of landslides and are, therefore, more suitable for LSM on a large scale [18,19]. However, statistical models cannot be used to express the complex nonlinear relationships between landslide susceptibility and environmental factors [20,21]. Studies have found that machine learning models, such as random forest (RF) [22–27], support vector machine (SVM) [28,29], multilayer perceptron (MLP) [30,31], and logistic regression (LR) [21,32], have the ability to automatically learn and mine the complex relationships among high-dimensional complex data and can provide a promising pathway for LSM [17,18,33–35]. Machine learning models have shown better performance in LSM studies [17,18,36]. Although the best machine learning model for LSM is under debate [18], the RF model is currently relatively robust because of its excellent performance in handling large amounts of nonlinear data [17]. Recently, several deep learning models, including convolutional neural network, feedforward neural network, and recurrent neural network, have also been explored [11,37,38]. However, deep-learning-based landslide susceptibility models generally face the problem of data dependence. In particular, the computational complexity of deep-learning-based models may increase significantly when the sample size of landslides is insufficient [39]. Therefore, in contrast, machine learning models are more suitable for small-sample studies. However, the accuracy of machine learning models for LSM at a large scale is generally limited by the diversity of landslide types [40] and the heterogeneity of environmental factors affecting landslides [5,17,18].

To overcome the limitations faced by machine-learning-based LSM studies, scholars have made improvements in three aspects. One is to construct the machine learning model for LSM with multi-source data and improve the evaluation effect of machine learning by using a complete landslide sample dataset by increasing the sample number of all landslide types in the study area. Multi-source landslide data, including survey data [13,41], disaster reports [36,42], and remote sensing images [20,43], have been used to train machine learning models [14,16]. Meanwhile, semi-supervised learning has also been used to expand the landslide sample dataset [14]. However, this method cannot guarantee that each type of landslide will have the same number of samples, and the features of a landslide type with a small sample may not be fully learned [9,18]. The second is to enhance the accuracy of the machine learning model by introducing key environmental factors, such as elevation, slope, and lithology [16,22,44]. Recently, several studies have focused on choosing appropriate environmental factors by considering specific regional characteristics and landslide types [42,45]. In the existing LSM studies, the variables related to human activities mainly include distance to road and land use types [18]. Although

these variables may reflect one or a few aspects of human activity, this is not sufficient to reflect the comprehensive effect of human activities on landslides [46,47]. The third is to divide the study area into multiple regions using the clustering analysis method based on the distribution of landslides in the study area and the similarity of environmental factors and to use the category attribute of each region as one of the input variables of the machine learning model [16,41]. The introduction of more variables, however, may lead to a decrease in the generalization ability in the clustering analysis method [48]. In addition, it is difficult to quantitatively evaluate the clustering result, which may further increase the uncertainty of the LSM model [48].

Moreover, machine-learning-based LSM studies are often affected by evident sample imbalances. Yu et al. [5] indicated that the sample imbalance issue may affect the reliability of the LSM. Reichenbach et al. [18] also indicated that there are far more negative samples (i.e., non-landslide samples) than positive samples (i.e., landslide samples) in LSM studies, and the imbalance between positive and negative samples may affect the accuracy of the LSM. In particular, the selection of the training-to-test sample ratio is critical to the results of the machine-learning-based LSM models [18]. The training-to-test sample ratios of 7:3 [14,44] and 8:2 [49,50] are generally applied in existing LSM studies, while the performance of different ratios in constructing LSM models is seldom explored. In general, the selection of the positive-to-negative sample ratio and the training-to-test sample ratio affects the accuracy of LSM models. Most LSM studies have to rely on a small amount of sample data [14,51] and, hence, suffer from the problem of sample imbalance [18]; this is because it is difficult and expensive to obtain adequate landslide samples and their genesis types on a large scale in a tight timeframe [14,18]. Although some landslide inventories are available at national and global scales, they normally consist of major landslide events only. Therefore, making full use of the limited sample data and setting reasonable ratios of landslide samples are crucial for constructing LSM models.

In view of the above problems, this study proposes an improved RF-based model by considering sample type and balance for LSM in Guangdong Province. The proposed method aims to improve the existing LSM methods in terms of three aspects: (1) improving LSM accuracy by considering the comprehensive impact of multiple environmental factors, such as topography, geography, land cover type, rainfall, and human activities; (2) reducing the error caused by multiple types of landslide samples by using rainfall-induced landslide samples only as the model input; (3) mitigating the adverse effects of sample imbalance by optimizing the positive-to-negative sample ratio and training-to-test sample ratio of the LSM model based on quantitative analysis.

2. Materials and Methods

2.1. Study Area

The study area is Guangdong Province (20°13′ N~25°31′ N, 109°39′ E~117°19′ E) (Figure 1), located in Southern China, with a land area of 178,000 km^2. The area is made up of 31.4% mountains, 25.61% hills, 20.26% tablelands, and 22.73% plains and valleys [52]. The intrusive rocks dominated by granite account for one-third of the total area of Guangdong Province, while migmatite and gneiss also exist widely [45,52]. Guangdong belongs to a subtropical monsoon climate, with an average annual rainfall of 1758.8 mm and characterized by perennial mild and humid weather. Under the pressure of abundant rainfall and frequent landing typhoons, the probability of landslides in Guangdong Province is high [2,45,51]. Considering that Guangdong has a high population density and economic activity intensity, the multiple effects of geography, economy, and demographics in Guangdong create conditions susceptible to the occurrence of landslide disasters [4]. Due to the lack of data on the environmental factors of some islands (e.g., Dongsha Islands), this study mainly investigates the landslide susceptibility of the major land areas in Guangdong Province.

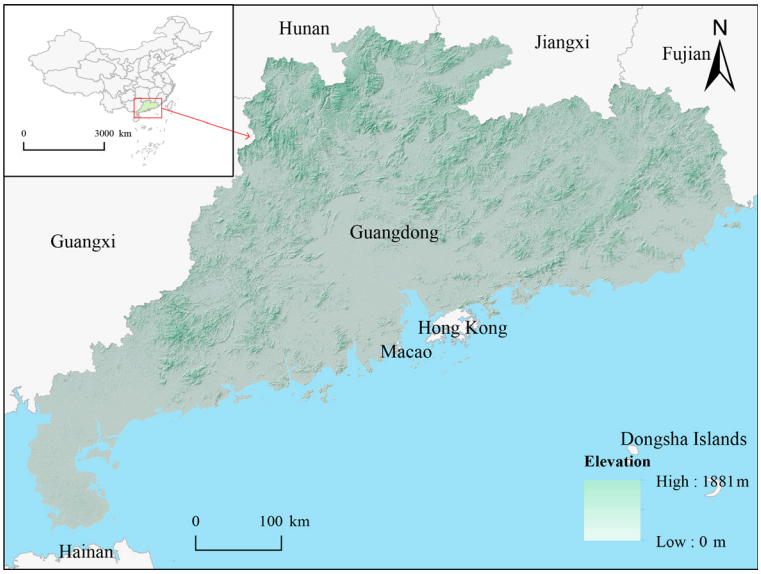


Figure 1. Location of the study area and the elevation image of Guangdong Province (using 50 m spatial resolution for better visualization).

2.2. Data

2.2.1. Landslide Inventory

Landslide inventory refers to a detailed register of the spatial distribution, geometry, and attributes of landslides [9]. In this study, two classes of landslide inventory were used: (1) Class I: data of 335 historical landslide points in Guangdong Province from 2015 to 2019, which were collected from geological disaster reports published by the Department of Natural Resources of Guangdong Province. These reports provide information on the occurrence time, detailed location, coordinates, primary causal factor (rainfall-induced or human activity–induced), and geological hazard type (rock fall, landslide, ground collapse, and debris flow), etc. Among these landslide events, 254 belong to the rainfall-induced type; the remainder are triggered by both rainfall and slope excavation and, thus, belong to the mixed type of rainfall-induced and human activity-induced landslides. Generally, the Class I landslide data are mainly rainfall-induced, with some mixed with human activity effects. (2) Class II: data of 586 landslide polygons in Longchuan County and Zijin County, Heyuan City in Guangdong Province. The Class II landslide data were obtained through a two-step procedure, which include visual interpretation of the GaoFen-2 images in 2021 and verification based on field investigation. This indicates that the Class II landslide data have a high accuracy and can be used as ground truth data in the region. The Class II landslide data used in this study are of mixed types, covering all landslide events detected in the study area; however, attribution categories are not specified. In this study, the Class I landslide data were used as the landslide samples for training and testing the RF-based model for LSM, while the Class II landslide data were employed to validate the effectiveness of the RF-based model for multiple types of landslides; there is no overlap between the two classes of landslide data.

2.2.2. Environmental Factors

Considering the regional characteristics of Guangdong Province and the main type of landslide samples (i.e., Class I landslide data), 13 environmental factors were selected, which can be categorized into topography, geology, land cover, meteorology, hydrology, and human activities. Detailed information and data sources of these factors are listed in Table 1. These environmental factors were resampled into 1 km spatial resolution grid cells using the WGS1984 coordinate system; the spatial distributions of these factors in Guangdong Province are shown in Figure 2. To ensure the consistency of the grid unit,

287 grids were labeled as landslide samples, with one or more historical landslide points falling into their grid area.

Table 1. The 5 categories of 13 environmental factors selected in this study and their information.

Category	Resolution	Factor	Data Type	Classes	Source
Topographic	30 m	Elevation (m)	Continuous	0~1701	ASTER GDEM V2 DEM (earthexplorer.usgs.gov, accessed on 15 November 2021)
		Slope (°)	Continuous	0~23.41	
		Aspect	Categorical	1: Flat, 2: North, 3: Northeast, 4: East, 5: Southeast, 6: South, 7: Southwest, 8: West, 9: Northwest	
		Profile curvature	Continuous	0~0.56	
		Plane curvature	Continuous	0~0.24	
Geological	1000 m	Lithology	Categorical	0: Unknown, 1: Extrusive rock, 2: Hypabyssal rock, 3: Plutonite rock, 4: Clastic rock, 5: Clay rock, 6: Clastic and Clay rock, 7: Biochemical sedimentary rock, 8: Metamorphic rock, 9: Melange rock, 10: Quaternary system	National Geological Archive (www.ngac.cn, accessed on 17 November 2021)
		Distance to fault (m)	Continuous	0~133,212	
Land cover	30 m	Normalized difference vegetation index (NDVI)	Continuous	0.04~0.85	Google Earth Engine (GEE) Landsat 8 Collection Tier 1 (earthengine.google.com, accessed on 16 November 2021)
		Land cover type	Categorical	10: Cropland, 20: Forest, 30: Grass, 40: Shrub, 50: Wetland, 60: Water, 80: Artificial, 90: Bareland	GlobeLand30 2020 (www.globallandcover.com, accessed on 2 December 2021)
Meteorological and hydrological	1000 m	Distance to river (m)	Continuous	0~13,347	OpenStreetMap (www.openstreetmap.org, accessed on 2 December 2021)
		Rainfall (mm)	Continuous	1403.6~2500.7	National Earth System Science Data Center (www.geodata.cn, accessed on 15 November 2021) [53]
Human activity	1000 m	Distance to road (m)	Continuous	0~13,642.6	OpenStreetMap (www.openstreetmap.org, accessed on 2 December 2021)
		Global human modification (GHM)	Continuous	0.08~0.98	GEE CSP GHM (earthengine.google.com, accessed on 1 December 2021) [54]

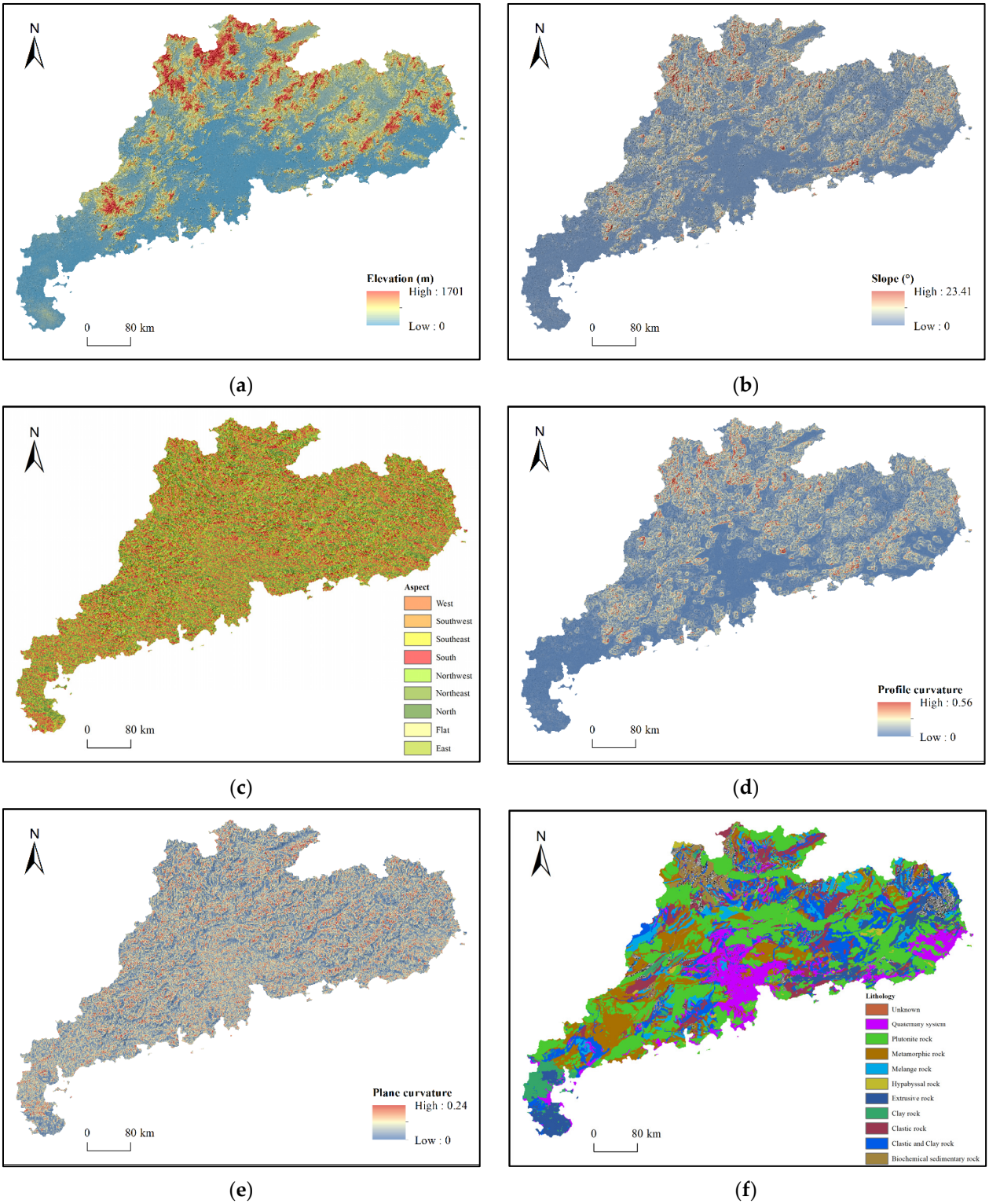


Figure 2. *Cont.*

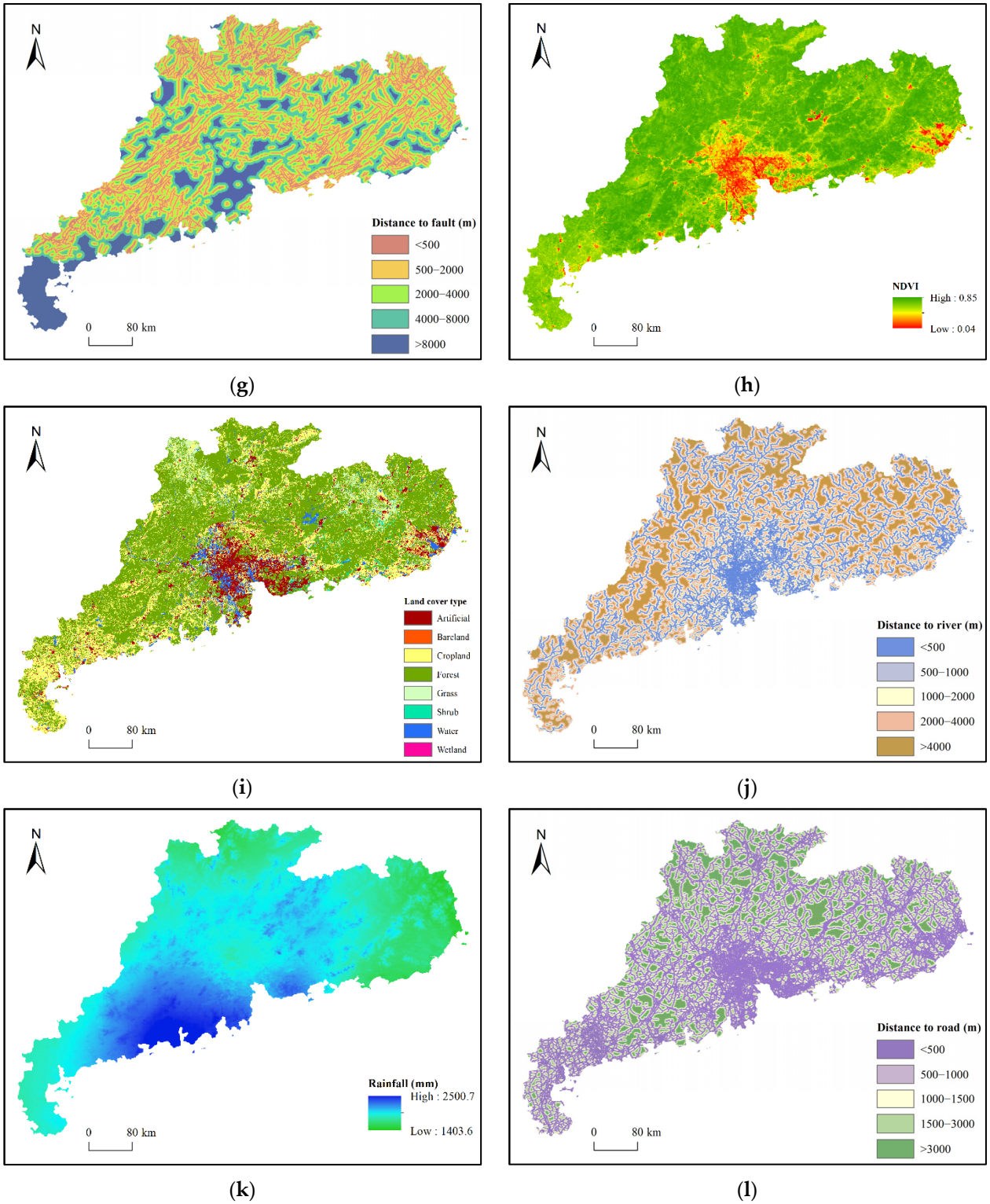
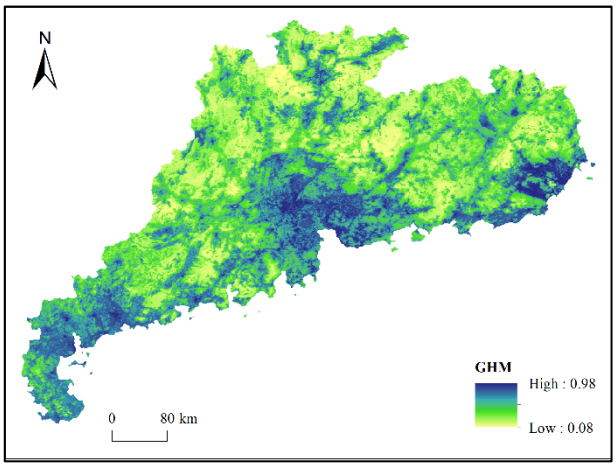


Figure 2. *Cont.*



(**m**)

Figure 2. The spatial distributions of 13 environmental factors in Guangdong Province: (**a**) elevation; (**b**) slope; (**c**) aspect; (**d**) profile curvature; (**e**) plane curvature; (**f**) lithology; (**g**) distance to fault; (**h**) NDVI; (**i**) land cover type; (**j**) distance to river; (**k**) rainfall; (**l**) distance to road; (**m**) GHM.

1. Topographic data
   The elevation, slope, aspect, profile, and plane curvature were extracted from the 30 m spatial resolution ASTER global digital elevation model (GDEM) V2 version data after resampling (data year 2011). The ASTER GDEM V2 version was adopted in this study because it has been proven to be more accurate than the V1 version and it has more practical use and allows for greater verification than the V3 version [55]. Topographic data are the most commonly used factor in LSM studies [18]. Landslides can occur when elevation and slope exhibit certain conditions [18,48]; a higher curvature means that the slope has a stronger capacity for water accumulation and is more prone to landslides [34,48]; meteorological processes regulate sunlight, hydrological elements, and wind direction, which affect slope stability [48].
2. Geological data
   Landslides are related to geological conditions [18]. They primarily occur in areas with a weathered soil layer on the bedrock surface [51] as well as with tectonic activity [18]. Thus, the geological data of the study area mainly include lithology and distance to fault, which were derived from the geological vector map of Guangdong Province provided by the National Geological Archive (NGA). The lithology factor was reclassified into the quaternary system, plutonic rock, metamorphic rock, melange (or mélange) rock, hypabyssal rock, extrusive rock, clay rock, clastic rock, and biochemical sedimentary rock, in accordance with the original data description and the study of Liu et al. [41]. The distance to fault factor was calculated using the Euclidean distance between each grid and the nearest fault structure.
3. Land cover data
   The normalized difference vegetation index (NDVI) and land cover type were used to represent land cover in the study area, as they can affect slope stability by altering the density of vegetation, soil moisture content, land evapotranspiration, and root strength [18,35]. The NDVI was generated via the Landsat8 Collection 1 Tier 1 8-Day from 2015 to 2019 on the Google Earth Engine (GEE) platform; the maximum values during the study period were adopted as the NDVI factor. The land cover type was obtained from the land cover product GlobeLand30 for 2020. Since Guangdong Province is located in a low-latitude area and has a tropical and subtropical climate, eight land cover types, including cultivated land, forest, grassland, shrubland, wetland, water bodies, artificial surfaces, and bare land, were considered in the study area (excluding tundra and permanent snow and ice).
4. Meteorological and hydrological data

Rainfall is an important trigger of landslides [56]. Because the rainfall in Guangdong Province is significant and the samples input to the model are all rainfall-induced landslides, it is necessary to consider the rainfall factor. The annual average rainfall represents the average rainfall condition over the long term in the region [24], which has been validated in previous machine learning LSM studies [16,41]. The average annual rainfall of the study area from 2015 to 2019 is used as the rainfall factor; it was calculated by processing the primitive data of the monthly rainfall dataset provided by the National Earth System Science Data Center [53]. The river can carry away the rock and soil mass at the slope toe, resulting in the slope toe near the river easily forming an aerial surface and promoting the occurrence of landslides [24]. Thus, the factor of distance to river was added; it was obtained by calculating each grid's Euclidean distance to the nearest river. The river network data were obtained from OpenStreetMap.

5. Human activity data
   The road construction near slopes can considerably reduce the stability of slopes [24] and even directly lead to landslide occurrence [47]. Therefore, the factor of distance to road indicates a trigger of human activity; it was obtained by calculating the Euclidean distance from each grid to the nearest road based on road data downloaded from the OpenStreetMap. In particular, to fully measure the complicated impact of human activities on landslides [47], the Global Human Modification (GHM) data with a 1 km spatial resolution on the GEE platform were also incorporated (data year 2016) [54]. The GHM is a continuous index ranging from 0 (no human impact) to 1 (high impact), where a greater value indicates more intense human modification of terrestrial lands. It provides an insight into the impact of human activities by analyzing various types of data, including human settlements, agricultural activities, transportation, mining and energy production activities, and electricity infrastructure. As an integrated variable, the GHM considers the impact of different types of human modification.

2.3. Methods

This study constructed an RF-based LSM model to analyze the distribution pattern of landslide susceptibility in Guangdong Province and its dominant environmental factors. First, the normalization and correlation test for the preprocessing of the input data of LSM was performed. Second, the RF-based LSM model with the optimal ratio of positive-to-negative samples and training-to-test samples was constructed. Third, the performance of the proposed RF-based LSM model was evaluated and compared with the SVM, MLP, and LR-based models using the area under curve (AUC) method. All steps were performed through the scikit-learn open-source library in Python and the ESRI ArcGIS software.

2.3.1. Normalization and Correlation Test of Model Inputs

To avoid the inconsistent influence in the data dimension of different input factors for the LSM model, min–max normalization pre-processing was performed for the 13 environmental factors using the following formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x is the original value of the independent variable (i.e., environmental factor), x_{max} and x_{min} are the maximum and minimum values of the independent variable, respectively, and x' is the normalized independent variable.

Considering the correlation effect between 13 environmental factors in the machine learning model [33], the variance inflation factor (*VIF*) was calculated to detect multicollinearity among these factors (*VIF* values above 5 indicate the presence of multicollinear-

ity) [34]. An appropriate combination of environmental factors could then be determined. The formula for *VIF* is as follows:

$$VIF_i = \frac{1}{1 - {R_i}^2} \tag{2}$$

where R_i is the correlation coefficient of regression analysis of x_i for the remaining independent variables.

2.3.2. Construction of LSM Model Considering Sample Type and Balance

This study employed the RF algorithm to construct the LSM model based on Class I landslide data and environmental factors. To minimize the model fitting error caused by the heterogeneity of landslide samples, the rainfall-induced landslides derived from the Class I landslide data were used as positive samples. In addition, the negative samples were selected from non-landslide regions and defined as non-road and non-river regions more than 1 km away from the positive samples [44].

The selection of the positive-to-negative sample ratio (i.e., landslide to non-landslide sample ratio) and the training-to-test sample ratio is essential for machine-learning-based LSM models [5,18]. The grid cells where the landslides occur were selected as the landslide samples, while the grid cells in the area outside the occurrence of landslides were used as the non-landslide samples [18]. Thus, the number of available landslide samples was much lower than the number of the non-landslide samples. To obtain the optimal positive-to-negative sample ratio of the LSM model, a positive-to-negative sample ratio of 1:1 was used to represent the case of balanced positive and negative samples, and ratios of 1:2 and 1:3 represented the case of unbalanced positive and negative samples. Considering that the area under curve (AUC) is insensitive when there is a quantitative difference between positive and negative samples [57], the *Sensitivity* from the confusion matrix was employed to evaluate the landslide identification precision. As shown in Figure 3, TP, FN, TN, and FP are four metrics of the confusion matrix. TP is the true positive, i.e., the number of landslide samples correctly predicted by the model; FN is the false negative, i.e., the number of samples where the model misclassifies landslides as non-landslides; TN is the true negative, i.e., the number of non-landslide samples correctly predicted by the model; and FP is the false positive, i.e., the number of samples where the model misclassifies non-landslides as landslides. The indexes of *Sensitivity* and $Specificity$ can be calculated using the following formulas:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

where *Sensitivity* denotes a true-positive rate, i.e., the rate of landslides correctly predicted number by the model to the number of landslide samples; $Specificity$ denotes the true-negative rate, i.e., the rate of non-landslides correctly predicted number by the model to the number of non-landslide samples. In this study, the positive-to-negative sample ratio with the greatest *Sensitivity* was selected as the optimal ratio.

Referring to several LSM studies, the training-to-test sample ratio of the sample set was set to 7:3 [14,44] and 8:2 [49,50]. AUC values of the LSM model with different ratios were calculated, and average AUC values were obtained by repeating the previous process ten times. The AUC denotes the area enclosed by the receiver operating characteristic (ROC) curve and the coordinate axis. It has been widely used to measure the accuracy of models [34]. The range of the AUC value is from 0.5 to 1, where a higher value indicates a stronger discriminative capability of the model [34]. Generally, values of AUC above 0.7 indicate high accuracy, while values above 0.9 indicate very high accuracy [58]. The ROC curve is generated with *Sensitivity* as the horizontal axis and $Specificity$ as the vertical axis.

In this study, the training-to-test sample ratio with the highest average AUC value was selected as the optimal ratio. The optimal combination of hyperparameters was determined using the cross-validation and grid search methods.

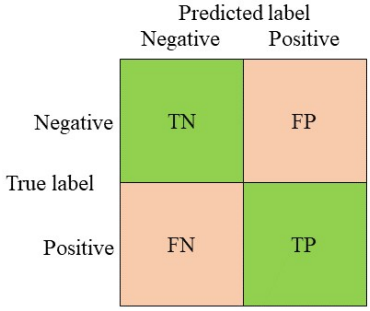


Figure 3. Confusion matrix diagram.

2.3.3. Machine-Learning-Based LSM Model Comparison

To verify the reliability of the proposed RF-based LSM model, the performance of this model was compared with three other machine learning models, i.e., SVM, MLP, and LR, using the optimal positive-to-negative sample ratio and training-to-test sample ratio. Specifically, all samples were first divided into ten sample sets, each of which had the optimal ratio of positive-to-negative samples, i.e., 287 positive samples and 287n (n = 1, 2, or 3) negative samples. For each sample set, the optimal ratio of training-to-test samples was adopted, i.e., 70% or 80% of the samples were used as the training set, while the remaining 30% or 20% were used as the test set. Then, these ten sample sets were used to train and test four machine learning models, and their performances were evaluated based on the testing set and the ROC. Finally, AUC values corresponding to the sample sets of each machine learning model were calculated, and the model with the largest average AUC value indicated the highest accuracy of the LSM.

2.3.4. Model Application and Multidimensional Analysis

The model with the largest AUC value was employed to map the landslide susceptibility of Guangdong Province. We used the natural breaks method to identify four levels, i.e., low susceptibility, moderate susceptibility, high susceptibility, and very high susceptibility. The distribution pattern of landslide susceptibility in the study area was further analyzed at the provincial scale and the city scale.

To evaluate the effect of the model on landslide samples, the percentage of the number of Class I landslides at each landslide susceptibility level to the total number of Class I landslides (Pn_i) was calculated; it is defined as follows:

$$Pn_i = \frac{n_i}{n} \times 100\% \quad (5)$$

where n_i donates the number of Class I landslides at each susceptibility level i (four levels, including low, moderate, high, and very high), and n donates the overall number of Class I landslides.

The susceptibility levels of landslides for each city in Guangdong Province were evaluated by calculating the percentage of the area at different susceptibility levels in each city of Guangdong Province (Pa_{ij}). The formula of Pa_{ij} is as follows:

$$Pa_{ij} = \frac{A_{ij}}{A_j} \times 100\% \quad (6)$$

where A_{ij} donates the area of each susceptibility level i (four levels, including low, moderate, high, and very high) of city j, and A_j donates the administrative area of city j.

Furthermore, the landslide susceptibility distribution pattern was the result of various environmental factors. Considering that the RF model has also proven to be an effective method for quantifying the importance of different variables, the RF model was used to analyze the impact of 13 environmental factors on the landslides in Guangdong Province.

3. Results

3.1. The Correlation of Environmental Factors

Table 2 shows the multicollinearity test results, i.e., the VIF values of 13 environmental factors. As shown in Table 2, the VIF values of the environmental factors are all less than 5, which indicates there is no multicollinearity among these factors [34]. Therefore, this study employed all 13 environmental factors for LSM.

Table 2. The *VIF* values of 13 environmental factors.

Factor	*VIF* Value	Factor	*VIF* Value	Factor	*VIF* Value
Elevation	2.4	NDVI	2.69	Land cover type	1.53
Slope	2.21	Distance to river	1.26	Rainfall	1.04
Aspect	1	Distance to road	1.52	GHM	2.69
Profile curvature	1.53	Distance to fault	1.09		
Plane curvature	1.04	Lithology	1.14		

3.2. Performance Evaluation of Sample Ratios

3.2.1. Analysis of Positive-to-Negative Sample Ratio on Sensitivity

Figure 4 shows the confusion matrices of the RF-based LSM model with different ratios of positive-to-negative samples (i.e., landslide samples and non-landslide samples) using the testing sets. It can be found that the *Sensitivity* value (multiplied by 100% to show the percentage) is the highest (79.31%) when the positive-to-negative sample ratio is 1:1. For ratios of 1:2 and 1:3, the *Sensitivity* values drop to 44.83% and 25.86%, respectively. We can, therefore, conclude that the ratio of positive-to-negative samples has a significant effect on the sensitivity index, i.e., the precision of landslide identification. In other words, the model's ability to identify landslides would be greatly limited when there is a significant quantitative difference between positive and negative samples. Therefore, 1:1 is selected as the optimal positive-to-negative sample ratio.

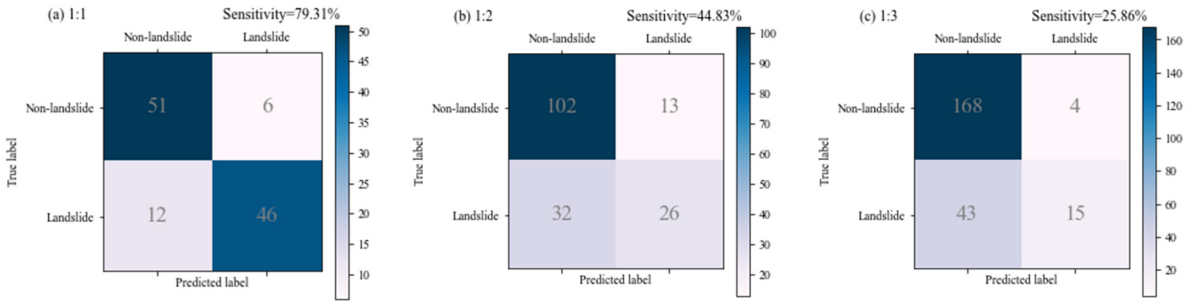


Figure 4. The confusion matrices and sensitivity values of the RF model with three positive-to-negative sample ratios: (**a**) ratio of 1:1, (**b**) ratio of 1:2, (**c**) ratio of 1:3.

3.2.2. Analysis of Training-to-Test Sample Ratio of AUC

Figure 5 shows that the ROC curves and AUC values are based on the same sample set with a positive-to-negative sample ratio of 1:1. The RF model was constructed using training-to-test sample ratios of 8:2 and 7:3, and the model construction process was repeated 10 times. The results show that the AUC values range from 0.7609 to 0.8852, with an average of 0.831, when the training-to-test sample ratio is set as 8:2. Meanwhile, the AUC values range from 0.7468 to 0.8569, with an average value of 0.7973, when the training-to-test sample ratio is set as 7:3. The values of both the highest AUC and the

average AUC with a training-to-test sample ratio of 8:2 are higher than those of the 7:3 ratio. Thus, in this study, the optimal training-to-test sample ratio for the RF model was set as 8:2.

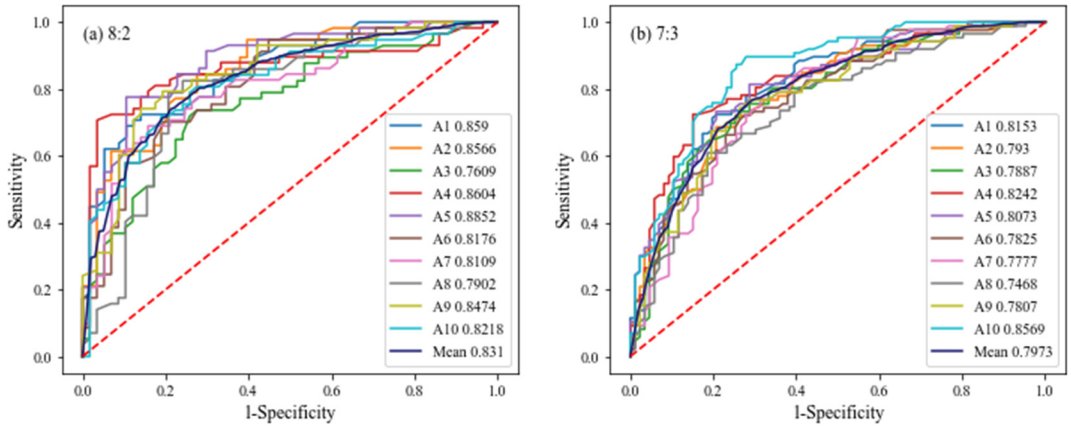


Figure 5. The ROC curves and AUC values of the RF model with two training-to-test sample ratios (A*i* represents the AUC value of the *i*-th sample set; and the red dash line represents the reference line, above which the ROC curves are valid): (**a**) ratio of 8:2; (**b**) ratio of 7:3.

3.3. Comparison of Machine Learning Models

The proposed RF-based LSM model was compared with three other machine learning models, namely SVM, MLP, and LR, which were all trained and tested using 10 different sample sets with the optimal positive-to-negative sample ratio (1:1) and training-to-test sample ratio (8:2). In addition, the optimal combination of hyperparameters was found and is shown in Table 3. Figure 6 illustrates the ROC and AUC values of these four models. In general, the AUC values of SVM, MLP, and LR models are between 0.6322 and 0.7901, with average AUC values around 0.7. Among these machine learning models, the RF model has the best performance on landslide susceptibility mapping, with the highest average AUC value of 0.8116 and the highest AUC value of 0.9145 (Figure 6a). These findings indicate that the RF model is superior to the other three machine learning models. This is likely because the RF model can prevent overfitting of the training set by creating multiple decision trees and handling missing values and outliers [17].

Table 3. The optimal combination of main hyperparameters involved in RF.

Hyperparameter	Explanation	Value
n_estimators	The number of decision trees.	200
criterion	The sample segmentation criteria.	gini
max_depth	The maximum depth of decision trees.	None
min_samples_split	the minimum number of samples required to split an internal node.	2
min_samples_leaf	the minimum number of samples required to be at a leaf node.	1

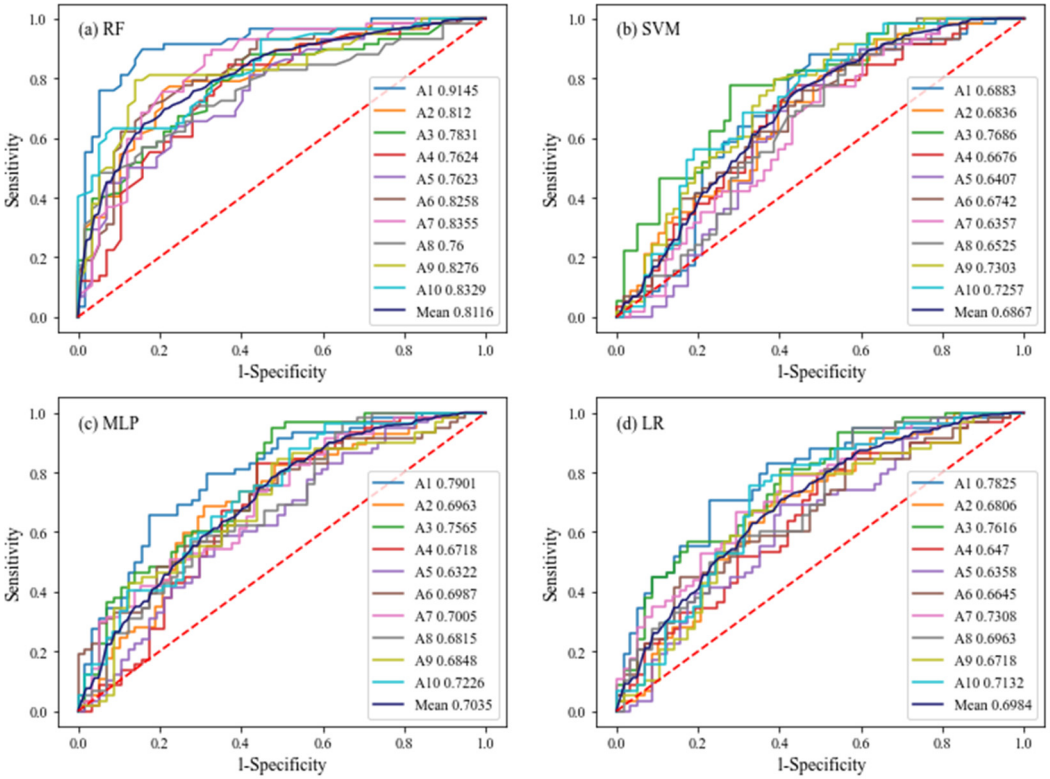


Figure 6. The ROC curves and AUC values of four machine learning models (the red dash line represents the reference line, above which the ROC curves are valid), (**a**) random forest (RF), (**b**) support vector machine (SVM), (**c**) multilayer perceptron (MLP), and (**d**) logistic regression (LR), with 10 different sample sets (A*i* represents the AUC value of the *i*-th sample set; mean represents the average AUC value of the 10 sample sets).

3.4. Analysis of Landslide Susceptibility Distribution Pattern and Factor Importance

3.4.1. Distribution Pattern Analysis of Landslide Susceptibility at the Provincial Scale

Following the analysis outlined above, the optimal positive-to-negative sample ratio of 1:1 and training-to-test sample ratio of 8:2 were determined. We also found that the RF model outperformed the SVM, MLP, and LR models. Therefore, the RF-based LSM model with the highest AUC value (A1 in Figure 6a) was employed in Guangdong Province to obtain the landslide susceptibility index for each grid. The landslide susceptibility raster layer of the study area was then classified into four levels using the natural breaks method, with the four levels being low susceptibility (0–0.28), moderate susceptibility (0.29–0.43), high susceptibility (0.44–0.6), and very high susceptibility (0.61–0.98). Figure 7 shows the distribution of the four landslide susceptibility levels in Guangdong Province. As shown in Figure 7, regions of high landslide susceptibility are mainly located in the north and east part of Guangdong, while regions of low landslide susceptibility are mainly located in the south and west part of Guangdong Province. It can be found that mountainous regions tend to have higher landslide susceptibility, while regions with gentle terrain, such as the Pearl River Delta, the Hanjiang Delta, and the coastal areas, tend to have lower landslide susceptibility. This also indicates that the landslides of Guangdong are closely related to topography.

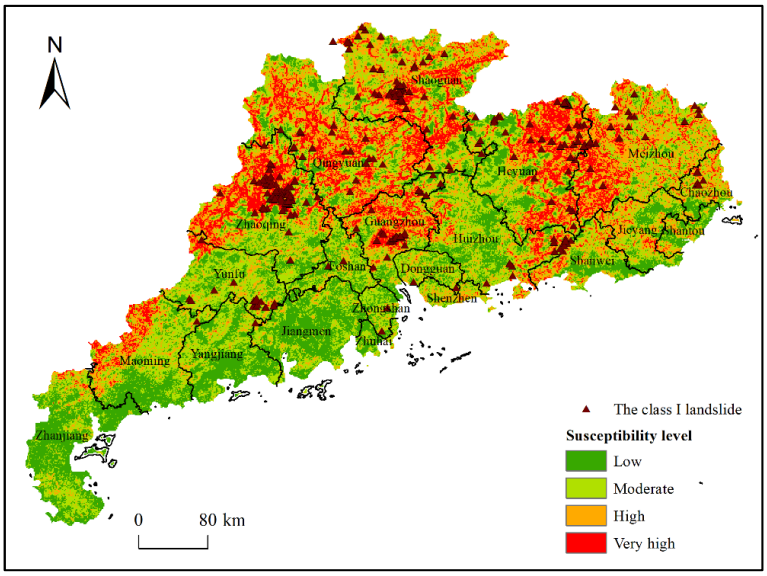


Figure 7. The landslide susceptibility map of Guangdong Province and the recorded 335 historical landslide points provided by the Class I landslide data (note: since the elevation data of the sea around some offshore land is missing, its slope, aspect, and curvature could not be calculated, so the boundary of this figure may be slightly different from the boundary of Figure 1).

Table 4 shows the percentage of the Class I landslide number in each landslide susceptibility level region to its overall number in the study area (Pn_i). The results showed that 98.5% of historical landslide points in the study area have high and very high landslide susceptibility. Moreover, as shown in Figure 7, the areas surrounding the landslide points are mainly categorized as high landslide susceptibility regions. This indicates that the landslide susceptibility levels of Guangdong Province have excellent spatial consistency with the Class I landslide data.

Table 4. The Pn_i values of four landslide susceptibility levels in Guangdong Province.

Susceptibility Level	Low	Moderate	High	Very High
Pn_i (%) *	0.6	0.9	6.56	91.94

* Pn_i is calculated using Equation (5).

3.4.2. Distribution Pattern Analysis of Landslide Susceptibility at the City Scale

To further investigate the hierarchical structure of the landslide susceptibility level at the city scale, the Pa_{ij} values for all the 21 cities in Guangdong Province were calculated and ranked in descending order. Figure 8 shows that the landslide susceptibility level varies from city to city. The landslide susceptibility of these cities can be classified into four categories for multi-level construction works for disaster reduction. The first category includes cities with large Pa_{ij} values at the very high susceptibility level, among which Heyuan has the largest Pa_{ij} value of 29.37%; the cities of Qingyuan, Zhaoqing, Shaoguan, Guangzhou, and Meizhou have Pa_{ij} values over 20% at the very high susceptibility level. These cities should, thus, further strengthen landslide prevention measures. The second category includes cities with large Pa_{ij} values at the high susceptibility level, such as Chaozhou, Shenzhen, Jieyang, Shanwei, and Shantou. Although their Pa_{ij} values at the very high landslide susceptibility level were not prominent, their Pa_{ij} values at the high landslide susceptibility level were over 25%, indicating that these cities have potential risks of landslides. The third category includes Jiangmen, Zhuhai, Yangjiang, and Zhanjiang, which are all coastal cities. Their Pa_{ij} values at the very high susceptibility level were less than 2%, while the Pa_{ij} values at the low susceptibility level were more than 50%. This result shows that the probability of landslides in these cities is low. The fourth category

consisted of the cities with relatively large Pa_{ij} values at the moderate susceptibility level, including Zhongshan, Dongguan, Foshan, Yunfu, Huizhou, and Maoming; these areas require regular inspection of potential landslide hazards.

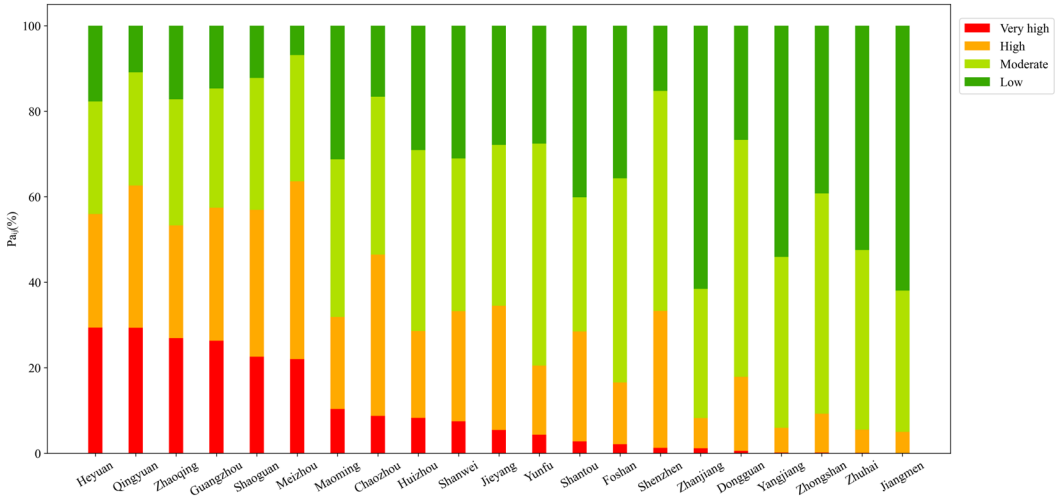


Figure 8. Pa_{ij} values in each city of Guangdong Province (Pa_{ij} is calculated using Equation (6)).

3.4.3. Analysis of Environmental Factor Importance

The selection of environmental factors may also influence the model's accuracy. Thus, the 13 environmental factors of landslide susceptibility in the study area were analyzed and ranked according to their importance. As shown in Figure 9, rainfall is the most important factor contributing to landslides in the model, which is consistent with the Class I landslide data from disaster reports, indicating that most landslides are rainfall-induced. The elevation, profile curvature, slope, and plane curvature rank from second to fifth place, which indicates that topography factors (except for the aspect), also have great influence on landslides in Guangdong Province. The GHM has higher importance than the distance to roads, which means it can serve as a better index for representing human activities.

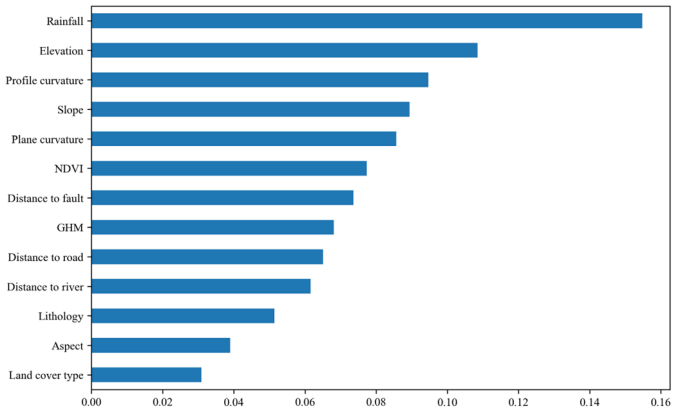


Figure 9. Ranking of importance of 13 environmental factors.

Additionally, the NDVI and the distance to fault ranked sixth and seventh, respectively, whereas the importance of the distance to river, lithology, slope direction, and land cover type is relatively low, representing the relatively low contribution of these environmental factors to landslide susceptibility in Guangdong Province.

Given that the landslides are generally related to geological conditions, the landslide susceptibility model was constructed by adopting two geological factors commonly used in existing LSM studies, i.e., lithology and distance to fault [18]. The landslides of Guangdong Province have proven to be predominantly rainfall-induced; the rising groundwater levels

have led to an increase in the soil pore water pressure, reduction in the geotechnical strength, and an increase in slope stress state [52,59]. Therefore, lithology has a relatively lower impact on landslides; its influence may be considered theoretical and indirect.

4. Discussion

4.1. Impact of Rainfall and GHM on Landslide Susceptibility Mapping

In this study, GHM was introduced to LSM for the first time, and the importance of rainfall as an environmental factor for landslides in Guangdong Province was verified. As shown in Table 5, the AUC values of the optimal model (RF) with and without the rainfall and GHM factors were calculated to investigate the effectiveness of these two environmental factors. When the GHM factor was removed from the RF model, the AUC value decreased from 0.9145 to 0.8991; when the rainfall factor was removed, the AUC value decreased from 0.9145 to 0.8917; when both factors were removed, the AUC value dramatically decreased from 0.9145 to 0.8808. The results indicated that the introduction of rainfall and GHM can significantly improve the mapping accuracy of landslide susceptibility.

Table 5. Effectiveness of rainfall and GHM on AUC values of the optimal model.

Model Category	Optimal Model	Without GHM	Without Rainfall	Without Rainfall and GHM
AUC value	0.9145	0.8991	0.8917	0.8808

The model accuracy and factor weights vary at different spatial resolutions in the same region [41]. Considering the limitations on the spatial resolution, the importance and impact of some factors with high heterogeneity (such as GHM) may also be underestimated. Therefore, the increase in the resolution can help us to analyze the impact of factors more accurately, as long as the computational power allows.

Other rainfall characteristics, such as intensity and duration, have also proven helpful for LSM studies [18]; these factors are more applicable to the modeling of landslide inventories after rainfall or typhoon events [51,60] as well as for landslide prediction under multiple scenarios [59,61].

4.2. Analysis of Landslide Environmental Factors in Guangdong Province

To reveal the nonlinear relationship between landslide susceptibility and each environmental factor, partial dependence plots were drawn. As shown in Figure 10, among all the factors, the most dramatic increase in the probability of landslide occurrence is observed when the rainfall range is 1650–1900 mm. The probability of landslide occurrence is also more prominent at 50–250 m of elevation and for the land cover type of artificial surfaces (code is 80). Moreover, the profile curvature, plane curvature, distance to road, and distance to fault show a linear relationship with the probability of landslide occurrence within a certain range. It is worth mentioning that the GHM increases rapidly around the interval 0.25–0.35 and then decreases slowly. The reasons may be twofold. On the one hand, initial human activities are destructive to the land (e.g., agricultural reclamation and large-scale deforestation [47]); however, as human activities increase, greater attention is paid to the construction of infrastructure to prevent landslide occurrence (e.g., conversion of farmland to forestry [47] and the installation of well-maintained drainage systems [46]). On the other hand, the most intensive human activities in Guangdong Province occur in the Pearl River Delta, which mostly consists of plains and, thus, has a very low landslide probability. Although the linear relationship between distance to road and the probability of landslide occurrence is more obvious, the potential advantage of the GHM is that it quantifies the intensity of human modification. In other words, GHM may be able to show to what extent human modification would promote landslide occurrence and, thus, help identify regions under threat.

However, it should be noted that the partial dependence plots are incapable of revealing the combined effect of high-dimensional data [62]. The partial dependence plots in

Figure 10 should not be interpreted as a simple linear relationship between the generation of landslides and individual environmental factors. Occurrences of landslides are, in most cases, the result of the combined effects of multiple environmental factors, with intricate and complex mechanisms [13,17]. For example, Yangjiang City and Jiangmen City are located in an area with the most abundant rainfall in Guangdong and boast intense human activity. However, the terrain in this area consists of mostly plains and low hills. The occurrence of landslides is, therefore, quite low in this region. This explains why there is a sharp decline when the annual rainfall amount surpasses 2100 mm in the partial dependence plot of rainfall.

To summarize, the partial dependence plots can help elucidate the relationship between landslide susceptibility and different environmental factors to some extent; however, the complexity of the interactions among these factors should be taken into account.

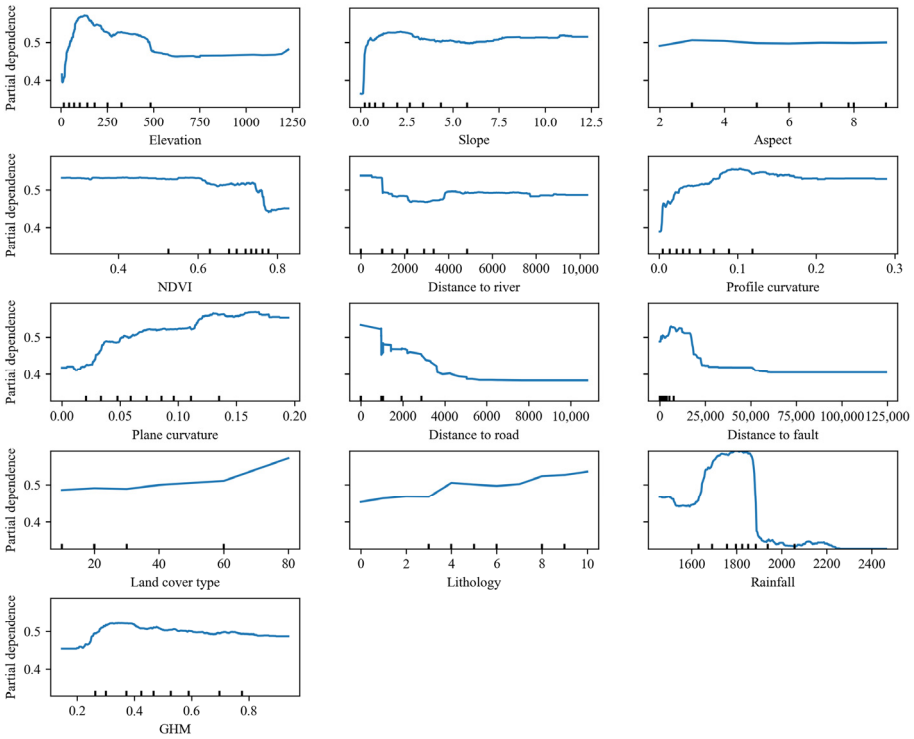


Figure 10. Partial dependence plots of 13 environmental factors. The horizontal axis represents the values of 13 environmental factors; the vertical axis is the partial dependence, representing the probability of landslide occurrence estimated by the model.

Interestingly, in the LSM studies that have adopted the RF algorithm, almost all of them have used a larger number of factors (more than 10) as inputs [22–27]. The reason may be that the RF model allows for the input of both categorical and continuous factors [17]; thus, it does not require much preprocessing and avoids the loss of information. This may also be one of the reasons why the RF model accuracy was better than the other models in the study. However, it is noteworthy that different classification criteria for categorical factors may still affect the result [18], especially for the factors that lack uniform classification criteria, such as lithology.

4.3. Model Evaluation for Multiple Types of Landslides

To further verify the reliability of the LSM model, multiple types of landslides were involved in the validation process. Longchuan County and Zijin County in Heyuan City were selected as the study cities because Heyuan City had the highest susceptibility value among all the 21 cities (Pa_{ij} = 29.37%). The landslide susceptibility map of these two counties was overlaid with Class II landslide data. Figure 11 shows that the distribution

patterns of landslides in Longchuan County and Zijin County are positively associated with the landslide susceptibility level. As shown in Table 6, the percentage of the area at different susceptibility levels in each county (such as Pa_{ij}) was applied in these two counties. The percentages of landslide areas at high and very high susceptibility levels are over 88% for both Longchuan County and Zijin County. Regions with high and very high susceptibility levels in Longchuan County and Zijin County are spatially consistent with the actual landslide locations. This further proves the reliability and generalizability of the model for application in Guangdong Province. Nevertheless, this paper only used remote sensing interpretation data in local areas for preliminary validation, because it is difficult to obtain a comprehensive sample of landslides in the whole province. Additional validation should be performed when more landslide samples are available.

The Class II landslide data were obtained via remote sensing interpretation and field verification. The Class II landslide data cover almost all landslide events that can be detected in the study area but do not contain their attribution categories. Thus, we considered the Class II landslide data as the mixed type and applied them to validate the LSM model. To further improve the model effect, detailed information on landslide type is needed.

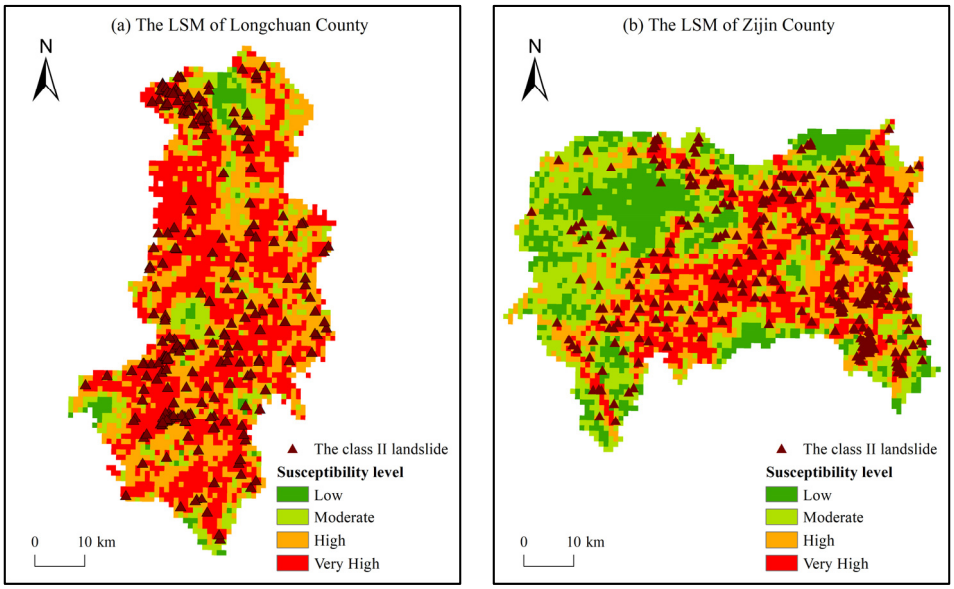


Figure 11. The Class II landslide data and the landslide susceptibility map of two counties in Heyuan City: (**a**) Longchuan County, (**b**) Zijin County.

Table 6. Percentage of landslide area at each susceptibility level to the overall landslide area.

Susceptibility Level	Longchuan County	Zijin County
Low	0%	1.11%
Moderate	2.33%	10.25%
High	25.79%	37.31%
Very high	71.88%	51.33%

4.4. Importance of Sample Type and Balance

This study employed the RF model for mapping the landslide susceptibility of Guangdong Province. Compared with previous large-scale landslide susceptibility studies [16,41,44], the RF-based LSM model can provide satisfactory mapping accuracy when multiple landslide sample types and sample imbalance issues are considered. It is, therefore, necessary to select the landslide type with similar genesis and relevant environmental factors as input variables and to optimize the positive-to-negative sample ratio and training-to-test sample ratio. The sample balance issue has been reported in some LSM studies [5,18]; however, the landslide

sample ratio selection is rarely explored. In this study, the sample ratios were analyzed quantitatively using confusion matrices and AUC values. The RF-based LSM model with optimal sample ratios has better generalization ability and operability, as it does not require a large amount of landslide data or the introduction of many variables [16]. Moreover, compared with other machine learning models, the improved RF-based model is more suitable for LSM studies on a large scale, especially when the sample size is small and there are multiple sample types.

5. Conclusions

To address the issues of environmental heterogeneity and sample imbalance in large-scale landslide susceptibility mapping, this study proposes an improved RF-based LSM model considering the sample type and balance and introducing GHM to the LSM model for the first time. The results indicate the importance of the impact of sample type and balance as well as appropriate environmental factors on the accuracy of LSM.

For the LSM models, the optimal positive-to-negative sample ratios of 1:1 and training-to-test sample ratio of 8:2 were obtained with confusion matrices and AUC values. Compared with the SVM, MLP, and LR models, the RF model, with the highest AUC value of 0.9145, had excellent accuracy for LSM in Guangdong Province. In the study area, rainfall and topography are the two primary influencing factors of landslides, with higher importance rankings. Moreover, the newly introduced GHM can be used for the LSM model to improve the AUC value. In Guangdong Province, the regions of high landslide susceptibility are mainly located in the northeast, while the regions of low landslide susceptibility are mainly located in the southwest. Heyuan, Qingyuan, Zhaoqing, Guangzhou, Shaoguan, and Meizhou in Guangdong Province show higher landslide susceptibility. These findings can provide an effective technical reference for LSM studies and can contribute to landslide prevention and disaster reduction, ensuring sustainable development of the regional economy.

Author Contributions: J.Z. and Y.H. collected and processed the data, Y.H. and D.G. proposed the model and analyzed the results, L.Z. and Y.H. drafted the manuscript, J.Z. and J.C. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2022B1515130001), the National Natural Science Foundation of China (No. 41971372) and the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (No. 311021004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We express our gratitude to Zhenghai Wang and his team at the School of Earth Sciences and Engineering, Sun Yat-sen University, for providing the remote sensing interpretation of landslide data, and the data support from National Earth System Science Data Center, National Science & Technology Infrastructure of China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guzzetti, F.; Reichenbach, P.; Cardinali, M.; Galli, M.; Ardizzone, F. Probabilistic Landslide Hazard Assessment at the Basin Scale. *Geomorphology* **2005**, *72*, 272–299. [CrossRef]
2. Lin, Q.; Wang, Y. Spatial and Temporal Analysis of a Fatal Landslide Inventory in China from 1950 to 2016. *Landslides* **2018**, *15*, 2357–2372. [CrossRef]
3. Yang, J.; Cheng, C.; Song, C.; Shen, S.; Ning, L. Visual Analysis of the Evolution and Focus in Landslide Research Field. *J. Mt. Sci.* **2019**, *16*, 991–1004. [CrossRef]
4. Zhang, F.; Peng, J.; Huang, X.; Lan, H. Hazard Assessment and Mitigation of Non-Seismically Fatal Landslides in China. *Nat. Hazards* **2021**, *106*, 785–804. [CrossRef]

5. Yu, X. Study on the Landslide Susceptibility Evaluation Method Based on Multi-Source Data and Multi-Scale Analysis. Ph.D. Thesis, China University of Geosciences, Wuhan, China, 2016.
6. Shi, P. The Natural Disasters, Constructions works for Disaster Reduction and Sustainable Development of China. *J. Nat. Resour.* **1995**, *3*, 267–275. [CrossRef]
7. Alcantara-Ayala, I.; Sassa, K. Contribution of the International Consortium on Landslides to the Implementation of the Sendai Framework for Disaster Risk Reduction; Engraining to the Science and Technology Roadmap. *Landslides* **2021**, *18*, 21–29. [CrossRef]
8. Wang, D.; Hao, M.; Chen, S.; Meng, Z.; Jiang, D.; Ding, F. Assessment of Landslide Susceptibility and Risk Factors in China. *Nat. Hazards* **2021**, *108*, 3045–3059. [CrossRef]
9. Guzzetti, F.; Reichenbach, P.; Ardizzone, F.; Cardinali, M.; Galli, M. Estimating the Quality of Landslide Susceptibility Models. *Geomorphology* **2006**, *81*, 166–184. [CrossRef]
10. Van Westen, C.J.; Castellanos, E.; Kuriakose, S.L. Spatial Data for Landslide Susceptibility, Hazard, and Vulnerability Assessment: An Overview. *Eng. Geol.* **2008**, *102*, 112–131. [CrossRef]
11. Bui, D.T.; Tuan, T.A.; Klempe, H.; Pradhan, B.; Revhaug, I. Spatial Prediction Models for Shallow Landslide Hazards: A Comparative Assessment of the Efficacy of Support Vector Machines, Artificial Neural Networks, Kernel Logistic Regression, and Logistic Model Tree. *Landslides* **2015**, *13*, 361–378. [CrossRef]
12. Ganguly, M.; Aynyas, R.; Nandan, A.; Mondal, P. Hazardous Area Map: An Approach of Sustainable Urban Planning and Industrial Development—A Review. *Nat. Hazards* **2018**, *91*, 1385–1405. [CrossRef]
13. Zhou, J.; Tan, S.; Li, J.; Xu, J.; Wang, C.; Ye, H. Landslide Susceptibility Assessment Using the Analytic Hierarchy Process (AHP): A Case Study of a Construction Site for Photovoltaic Power Generation in Yunxian County, Southwest China. *Sustainability* **2023**, *15*, 5281. [CrossRef]
14. Huang, F.; Cao, Z.; Jiang, S.; Zhou, C.; Huang, J.; Guo, Z. Landslide Susceptibility Prediction Based on a Semi-Supervised Multiple-Layer Perceptron Model. *Landslides* **2020**, *17*, 2919–2930. [CrossRef]
15. Chang, Z.; Du, Z.; Zhang, F.; Huang, F.; Chen, J.; Li, W.; Guo, Z. Landslide Susceptibility Prediction Based on Remote Sensing Images and GIS: Comparisons of Supervised and Unsupervised Machine Learning Models. *Remote Sens.* **2020**, *12*, 502. [CrossRef]
16. Wang, Y.; Feng, L.; Li, S.; Ren, F.; Du, Q. A Hybrid Model Considering Spatial Heterogeneity for Landslide Susceptibility Mapping in Zhejiang Province, China. *Catena* **2020**, *188*, 104425. [CrossRef]
17. Ma, Y.; Li, H.; Wang, L.; Zhang, W.; Zhu, Z.; Yang, H.; Wang, L.; Yuan, X. Machine Learning Algorithms and Techniques for Landslide Susceptibility Investigation: A Literature Review. *J. Civ. Environ. Eng.* **2022**, *44*, 53–67. [CrossRef]
18. Reichenbach, P.; Rossi, M.; Malamud, B.D.; Mihir, M.; Guzzetti, F. A Review of Statistically-Based Landslide Susceptibility Models. *Earth-Sci. Rev.* **2018**, *180*, 60–91. [CrossRef]
19. Wang, Z.; Li, R.; Wang, X. Review of Researches on Regional Landslide Susceptibility Mapping Model. *J. Yangtze River Sci. Res. Inst.* **2012**, *29*, 78–85. [CrossRef]
20. Pradhan, B.; Lee, S. 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.* **2010**, *25*, 747–759. [CrossRef]
21. Wang, L.; Guo, M.; Sawada, K.; Lin, J.; Zhang, J. A Comparative Study of Landslide Susceptibility Maps Using Logistic Regression, Frequency Ratio, Decision Tree, Weights of Evidence and Artificial Neural Network. *Geosci. J.* **2016**, *20*, 117–136. [CrossRef]
22. Kim, J.C.; Lee, S.; Jung, H.S.; Lee, S. Landslide Susceptibility Mapping Using Random Forest and Boosted Tree Models in Pyeong-Chang, Korea. *Geocarto Int.* **2018**, *33*, 1000–1015. [CrossRef]
23. Arabameri, A.; Pradhan, B.; Rezaei, K.; Lee, C.W. Assessment of Landslide Susceptibility Using Statistical- and Artificial Intelligence-Based FR–RF Integrated Model and Multiresolution DEMs. *Remote Sens.* **2019**, *11*, 999. [CrossRef]
24. Sun, D.; Wen, H.; Wang, D.; Xu, J. A Random Forest Model of Landslide Susceptibility Mapping Based on Hyperparameter Optimization Using Bayes Algorithm. *Geomorphology* **2020**, *362*, 107201. [CrossRef]
25. Rong, G.; Alu, S.; Li, K.; Su, Y.; Zhang, J.; Zhang, Y.; Li, T. Rainfall Induced Landslide Susceptibility Mapping Based on Bayesian Optimized Random Forest and Gradient Boosting Decision Tree Models—A Case Study of Shuicheng County, China. *Water* **2020**, *12*, 3066. [CrossRef]
26. Wu, X.; Song, Y.; Chen, W.; Kang, G.; Qu, R.; Wang, Z.; Wang, J.; Lv, P.; Chen, H. Analysis of Geological Hazard Susceptibility of Landslides in Muli County Based on Random Forest Algorithm. *Sustainability* **2023**, *15*, 4328. [CrossRef]
27. Akinci, H.; Kilicoglu, C.; Dogan, S. Random Forest-Based Landslide Susceptibility Mapping in Coastal Regions of Artvin, Turkey. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 553. [CrossRef]
28. Huang, F.; Yao, C.; Liu, W.; Li, Y.; Liu, X. Landslide Susceptibility Assessment in the Nantian Area of China: A Comparison of Frequency Ratio Model and Support Vector Machine. *Geomat. Nat. Hazards Risk* **2018**, *9*, 919–938. [CrossRef]
29. Zhao, S.; Zhao, Z. A Comparative Study of Landslide Susceptibility Mapping Using SVM and PSO-SVM Models Based on Grid and Slope Units. *Math. Probl. Eng.* **2021**, *2021*, 8854606. [CrossRef]
30. Zare, M.; Pourghasemi, H.R.; Vafakhah, M.; Pradhan, B. Landslide Susceptibility Mapping at Vaz Watershed (Iran) Using an Artificial Neural Network Model: A Comparison between Multilayer Perceptron (MLP) and Radial Basic Function (RBF) Algorithms. *Arab. J. Geosci.* **2013**, *6*, 2873–2888. [CrossRef]

31. Li, D.; Huang, F.; Yan, L.; Cao, Z.; Chen, J.; Ye, Z. Landslide Susceptibility Prediction Using Particle-Swarm-Optimized Multilayer Perceptron: Comparisons with Multilayer-Perceptron-Only, BP Neural Network, and Information Value Models. *Appl. Sci.* **2019**, *9*, 3664. [CrossRef]
32. Zhu, L.; Huang, J. GIS-Based Logistic Regression Method for Landslide Susceptibility Mapping in Regional Scale. *J. Zhejiang Univ. Sci. A* **2006**, *7*, 2007–2017. [CrossRef]
33. Drobnič, F.; Kos, A.; Pustišek, M. On the Interpretability of Machine Learning Models and Experimental Feature Selection in Case of Multicollinear Data. *Electronics* **2020**, *9*, 761. [CrossRef]
34. Achour, Y.; Pourghasemi, H.R. How Do Machine Learning Techniques Help in Increasing Accuracy of Landslide Susceptibility Maps? *Geosci. Front.* **2020**, *11*, 871–883. [CrossRef]
35. Fang, H.; Shao, Y.; Xie, C.; Tian, B.; Shen, C.; Zhu, Y.; Guo, Y.; Yang, Y.; Chen, G.; Zhang, M. A New Approach to Spatial Landslide Susceptibility Prediction in Karst Mining Areas Based on Explainable Artificial Intelligence. *Sustainability* **2023**, *15*, 3094. [CrossRef]
36. Wu, C.-Y.; Lin, S.-Y. Performance Assessment of Event-Based Ensemble Landslide Susceptibility Models in Shihmen Watershed, Taiwan. *Water* **2022**, *14*, 717. [CrossRef]
37. Zhang, W.; Li, H.; Li, Y.; Liu, H.; Chen, Y.; Ding, X. Application of Deep Learning Algorithms in Geotechnical Engineering: A Short Critical Review. *Artif. Intell. Rev.* **2021**, *54*, 5633–5673. [CrossRef]
38. Wei, X.; Zhang, L.; Luo, J.; Liu, D. A Hybrid Framework Integrating Physical Model and Convolutional Neural Network for Regional Landslide Susceptibility Mapping. *Nat. Hazards* **2021**, *109*, 471–497. [CrossRef]
39. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. *IEEE Int. Conf. Neural Netw.* **2018**, *11141*, 270–279. [CrossRef]
40. Cruden, D.M.; Varnes, D.J. Landslide Types and Processes. *Spec. Rep.-Natl. Res. Counc. Transp. Res. Board* **1996**, *247*, 36–75. [CrossRef]
41. Liu, L.; Li, S.; Li, X.; Jiang, Y.; Wei, W.; Wang, Z.; Bai, Y. An Integrated Approach for Landslide Susceptibility Mapping by Considering Spatial Correlation and Fractal Distribution of Clustered Landslide Data. *Landslides* **2019**, *16*, 715–728. [CrossRef]
42. Li, J.; Wang, W.; Li, Y.; Han, Z.; Chen, G. Spatiotemporal Landslide Susceptibility Mapping Incorporating the Effects of Heavy Rainfall: A Case Study of the Heavy Rainfall in August 2021 in Kitakyushu, Fukuoka, Japan. *Water* **2021**, *13*, 3312. [CrossRef]
43. Nourani, V.; Pradhan, B.; Ghaffari, H.; Sharifi, S.S. Landslide Susceptibility Mapping at Zonouz Plain, Iran Using Genetic Programming and Comparison with Frequency Ratio, Logistic Regression, and Artificial Neural Network Models. *Nat. Hazards* **2014**, *71*, 523–547. [CrossRef]
44. Zhao, J.; Zhang, Q.; Wang, D.; Wu, W.; Yuan, R. Machine Learning-Based Evaluation of Susceptibility to Geological Hazards in the Hengduan Mountains Region, China. *Int. J. Disaster Risk Sci.* **2022**, *13*, 305–316. [CrossRef]
45. Feng, W.; Bai, H.; Lan, B.; Wu, Y.; Wu, Z.; Yan, L.; Ma, X. Spatial–Temporal Distribution and Failure Mechanism of Group-Occurring Landslides in Mibei Village, Longchuan County, Guangdong, China. *Landslides* **2022**, *19*, 1957–1970. [CrossRef]
46. Van Den Eeckhaut, M.; Poesen, J.; Vandekerckhove, L.; Van Gils, M.; Van Rompaey, A. Human-Environment Interactions in Residential Areas Susceptible to Landsliding: The Flemish Ardennes Case Study. *Area* **2010**, *42*, 339–358. [CrossRef]
47. Li, Y.; Wang, X.; Mao, H. Influence of Human Activity on Landslide Susceptibility Development in the Three Gorges Area. *Nat. Hazards* **2020**, *104*, 2115–2151. [CrossRef]
48. Mwakapesa, D.S.; Mao, Y.; Lan, X.; Nanehkaran, Y.A. Landslide Susceptibility Mapping Using DIvisive ANAlysis (DIANA) and RObust Clustering Using LinKs (ROCK) Algorithms, and Comparison of Their Performance. *Sustainability* **2023**, *15*, 4218. [CrossRef]
49. Rabby, Y.W.; Li, Y. Landslide Susceptibility Mapping Using Integrated Methods: A Case Study in the Chittagong Hilly Areas, Bangladesh. *Geosciences* **2020**, *10*, 483. [CrossRef]
50. Wubalem, A.; Meten, M. Landslide Susceptibility Mapping Using Information Value and Logistic Regression Models in Goncha Siso Eneses Area, Northwestern Ethiopia. *SN Appl. Sci.* **2020**, *2*, 807. [CrossRef]
51. Ma, S.; Shao, X.; Xu, C. Characterizing the Distribution Pattern and a Physically Based Susceptibility Assessment of Shallow Landslides Triggered by the 2019 Heavy Rainfall Event in Longchuan County, Guangdong Province, China. *Remote Sens.* **2022**, *14*, 4257. [CrossRef]
52. Lin, B.; Yang, S.; Zhu, B.; Wu, H. Geological Structure and Basic Geotechnical Characteristics in Guangdong Province. *Chin. J. Rock Mech. Eng.* **2006**, *25*, 3337–3346. [CrossRef]
53. Peng, S.; Ding, Y.; Liu, W.; Li, Z. 1 Km Monthly Temperature and Precipitation Dataset for China from 1901 to 2017. *Earth Syst. Sci. Data* **2019**, *11*, 1931–1946. [CrossRef]
54. Kennedy, C.M.; Oakleaf, J.R.; Theobald, D.M.; Baruch Mordo, S.; Kiesecker, J. Managing the Middle: A Shift in Conservation Priorities Based on the Global Human Modification Gradient. *Glob. Change Biol.* **2019**, *25*, 811–826. [CrossRef]
55. Tang, X.; Li, S.; Li, T.; Gao, Y.; Zhang, S.; Chen, Q.; Zhang, X. Review on global digital elevation products. *Natl. Remote Sens. Bull.* **2021**, *25*, 167–181. [CrossRef]
56. Gnyawali, K.; Dahal, K.; Talchabhadel, R.; Nirandjan, S. Framework for Rainfall-Triggered Landslide-Prone Critical Infrastructure Zonation. *Sci. Total Environ.* **2023**, *872*, 162242. [CrossRef]
57. Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning–ICML '06, Pittsburgh, PA, USA, 25–29 June 2006; Volume 148, pp. 233–240. [CrossRef]

58. Swets, J. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285–1293. [CrossRef]
59. Mai, J.; Xian, Y.; Liu, G. Predicting potential rainfall-triggered landslides sites in Guangdong Province (China) using MaxEnt model under climate changes scenarios. *J. Geo Inf. Sci.* **2021**, *23*, 2042–2054. [CrossRef]
60. Yang, S.R. Assessment of Rainfall-Induced Landslide Susceptibility Using GIS-Based Slope Unit Approach. *J. Perform. Constr. Facil.* **2017**, *31*, 8. [CrossRef]
61. Yang, S.R. Probability of Road Interruption Due to Landslides under Different Rainfall-Return Periods Using Remote Sensing Techniques. *J. Perform. Constr. Facil.* **2016**, *30*, C4015002. [CrossRef]
62. Shi, H.; Yang, N.; Yang, X.; Tang, H. Clarifying Relationship between $PM_{2.5}$ Concentrations and Spatiotemporal Predictors Using Multi-Way Partial Dependence Plots. *Remote Sens.* **2023**, *15*, 358. [CrossRef]