sustainability

MDPI

Article

Landslide Susceptibility Prediction: Improving the Quality of Landslide Samples by Isolation Forests

Qinghua Zhang [1,2], Zhu Liang [1,2,3,*], Wei Liu [1,2], Weiping Peng [1,2], Houzan Huang [1,2], Shouwen Zhang [1,2], Lingwei Chen [1,2], Kaihua Jiang [1,2] and Lixing Liu [1,2]

1 Guangzhou Urban Planning & Design Survey Research Institute, Guangzhou 510060, China
2 Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, Guangzhou 510060, China
3 School of Civil Engineering & Transportation, South China University of Technology, Guangzhou 510641, China
* Correspondence: liangzhu19@mails.jlu.edu.cn

Abstract: Landslide susceptibility prediction (LSP) is the first step to ease landslide disasters with the application of various machine learning methods. A complete landslide inventory, which is essential but difficult to obtain, should include high-quality landslide and non-landslide samples. The insufficient number of landslide samples and the low purity of non-landslide samples limit the performance of the machine learning models. In response, this study aims to explore the effectiveness of isolated forest (IF) to solve the problem of insufficient landslide samples. IF belongs to unsupervised learning, and only a small share of landslide samples in the study area were required for modeling, while the remaining samples were used for testing. Its performance was compared to another advanced integration model, adaptive boosting integrated with decision tree (Ada-DT), which belongs to two-class classifiers (TCC) and needs a sufficient number of samples. Huangpu District, Guangzhou City, Guangdong Province in China, was selected as the study area, and 13 predisposing factors were prepared for the modeling. Results showed that the IF proved its effectiveness with an AUC value of 0.875, although the Ada-DT model performed better (AUC = 0.921). IF outperformed the Ada-DT model in terms of recognizing landslides, and the sensitivity values of IF and the Ada-DT model were 90.00% and 86.67%, respectively, while the Ada-DT model performed better in terms of specificity. Two susceptibility maps obtained by the models were basically consistent with the field investigation, while the areas predicted by IF tended to be conservative as higher risk areas were presented, and the Ada-DT model was likely to be risky. It is suggested to select non-landslide samples from the very low susceptibility areas predicted by the IF model to form a more reliable sample set for Ada-DT modeling. The conclusion confirms the practicality and advancement of the idea of anomaly detection in LSP and improves the application potential of machine learning algorithms for geohazards.

Keywords: landslide susceptibility; isolation forest; landslide samples; machine learning

Citation: Zhang, Q.; Liang, Z.; Liu, W.; Peng, W.; Huang, H.; Zhang, S.; Chen, L.; Jiang, K.; Liu, L. Landslide Susceptibility Prediction: Improving the Quality of Landslide Samples by Isolation Forests. *Sustainability* **2022**, *14*, 16692. https://doi.org/10.3390/su142416692

Academic Editors: Shuihua Jiang, Zezhou Wang, Faming Huang and Jinsong Huang

Received: 4 November 2022
Accepted: 8 December 2022
Published: 13 December 2022

1. Introduction

The occurrence of landslides is frequent, sudden, and destructive. The annual economic losses and casualties caused by landslides are huge and hard to estimate [1]. Landslide susceptibility prediction (LSP) effectively reduces the loss by predicting landslide-prone areas, which is useful for disaster prevention [2].

Machine learning methods have been commonly applied to LSP with the development of computer technologies, and they can be divided into unsupervised and supervised learning [3–5]. Supervised learning methods, which need both labeled samples and influencing factors, usually perform well in terms of accuracy. Support vector machines (SVM) are one of the famous algorithms [6–8]. Unsupervised learning methods are easier to implement, as

labeled samples are unnecessary. Clustering analysis is a common unsupervised learning algorithm [9,10]. The label samples include landslide samples and non-landslide samples in LSP. A complete landslide inventory is key, but hard to obtain. Although remote sensing technology has been developed, high and steep landslides are difficult to be recognized. Furthermore, the quality of remote sensing images could be deteriorated by the weather and thus, the distribution of new landslides cannot be identified in time. The most common scenario is that the number of landslide samples is limited in a specific area, which hinders the application of supervised learning. On the other hand, the determination of non-landslide samples is also controversial, since they are unpredictable and usually selected based on some principles [11,12] that probably introduce noise into the data set. Therefore, it is necessary to explore a more effective method that can simultaneously solve the problems of insufficient landslide samples and the impurity of non-landslide samples for LSP.

Oversampling or the superior performance of some algorithms can solve the problem of data imbalance [13,14]. The integrated algorithms can also combat the noise in the data [15,16]. However, they are not easy to implement, especially for non-professional technical personnel. One-class classifier (OCC) needs only positive or negative labeled samples for modeling, which reduces the difficulty of sample labeling [17,18]. One-class SVM has been applied to LSP and proved its suitability [19]. The machine learning models obey a fundamental principle that landslides prefer to occur again under the same or similar environment that led to landslides in the past [20,21]. Consequently, landslide samples may be similar to each other but distinct from non-landslide samples. In the same way, non-landslide samples may be similar to each other but distinct from landslide samples. Therefore, landslide samples or non-landslide samples could be considered anomalous or isolated, which conforms to the concept of anomaly detection. IF is one of the most famous anomaly detection algorithms that has been widely applied to many areas but seldom to LSP [22,23]. More importantly, IF performs even better when the number of samples is limited or insufficient, which helps solve the problems of swamping and masking [24].

This study aimed to explore the effectiveness of IF to solve the problem of insufficient landslide samples and impurity of non-landslide samples. Its performance is compared to another advanced machine learning algorithm. Huangpu district, Guangzhou city, China, was selected as the study area, and 13 conditioning factors were prepared for the modeling. This work sought to provide a new perspective for machine learning in LSP and solve the problems caused by the low quality of samples (insufficient quantity or low purity) from the source.

2. Materials

2.1. Study Area

Huangpu district is located in the north of the Pearl River Delta, Guangzhou City, Guangdong Province in China. Its geographical coordinate ranges are E 113°23′29″~113°36′2″, N 23°01′57″~23°24′57″ (Figure 1). It has a population of more than 1.26 million with an area of about 484.17 km^2. The whole area can be divided into plains (accounting for about 45%) and low hills (accounting for about 55%). The terrain is roughly high in the north and low in the south, with an average elevation of 60 m (ranging from −26 to 433 m). It is a subtropical monsoon climate, with prevailing summer southwest monsoon and southeast monsoon. Rainfall is abundant in the study area, with an average annual rainfall of 1665.0 mm, a maximum monthly rainfall of 334 mm, and maximum daily rainfall of 542 mm. The lithology of the study area mainly consists of shale, limestone, dolomite, and granite. The geological structure of the southern area is complex with more faults and folds. The seismic intensity has a degree of VII on the modified Mercalli index and the ground motion peak acceleration is 0.10 g, which indicates a relatively stable area. The landslides in the area are shallow landslides that are mainly induced by rainfall during the annul flood period. The volume of the landslides in the study area ranged from 110 to 3000 m^3.

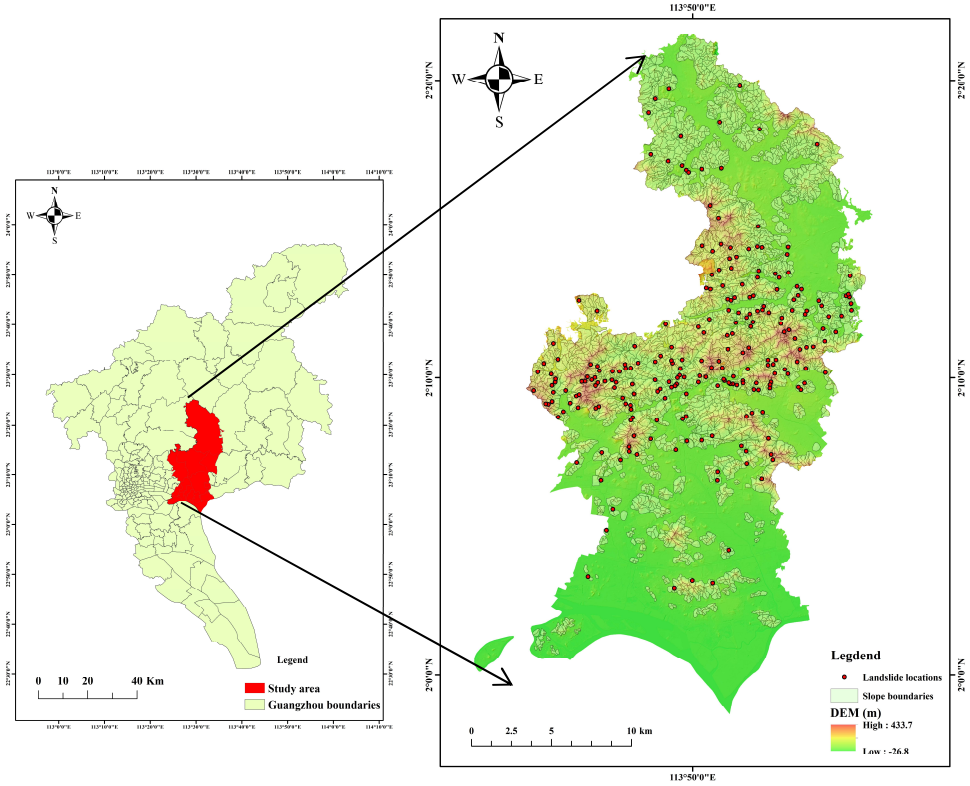

Figure 1. Location of the study area showing elevation and landslide samples.

On 21 May 2020, Huangpu District experienced a once-in-a-century torrential rain in which the maximum rainfall reached 288.5 mm in 3 h and 378.6 mm in 24 h. The extremely heavy rainfall triggered multiple landslides. A shallow landslide, the relative elevation difference of which reached 36 m, occurred in Mingquan villa. The volume of the landslide reached 3000 m^3, which is one of the largest landslides in the area, and ultimately caused 2 deaths and destroyed a home (Figure 2). Therefore, it is of great significance to make an accurate landslide susceptibility map for the study area.

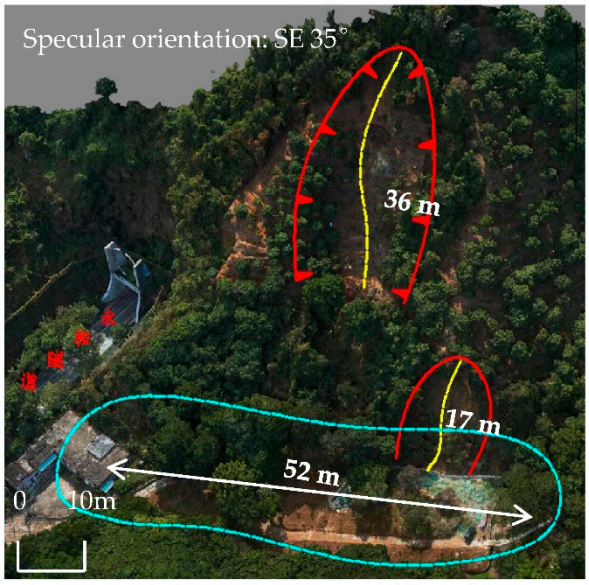

Figure 2. Shallow landslides in Mingquan Villa (taken on 9 July 2021).

2.2. Data Preparation

The modeling of landslide susceptibility mainly involves two data sets as landslide inventory and conditioning factors. A complete and accurate landslide inventory is the key to modeling, especially for machine learning arithmetic [25]. Landslide events induced by rainfall and earthquakes are considered in the study. A total of 239 landslide locations were compiled in the area based on the data from historical reports (1978~2020) and remote sensing interpretation. Landslide locations were further confirmed by field investigations during 2019~2022 (Figures 3 and 4). Landslide locations were extracted as single red points, as shown in Figure 1.

Figure 3. Shallow landslides that occurred on Jinkang Forest farm, Jiulong Town: (**a**) on 6 July 2021; (**b**) on 11 June 2018.

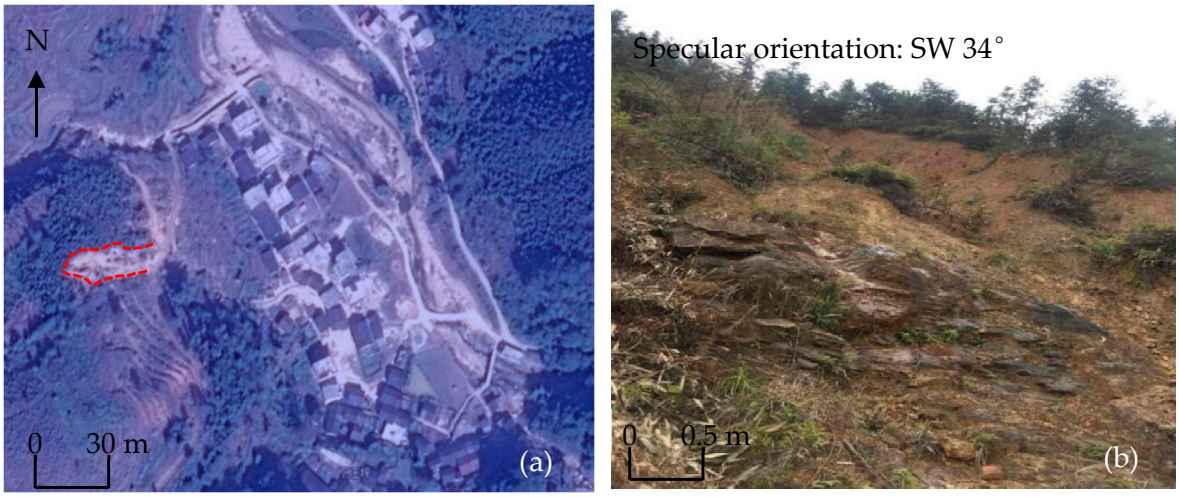

Figure 4. Remote sensing image interpretation: (**a**) shallow landslide on Lianhe Street (occurred on 23 June 2019); (**b**) field investigation (17 May 2019).

2.2.1. Data Sources

The occurrence of a landslide is mainly affected by environmental factors and triggering factors [26,27]. Environmental factors mainly refer to topographical and geological factors. Considering the previous experience and the availability of data sources, 13 conditioning factors were selected and shown in Table 1.

Table 1. Landslide conditioning factors in this study.

Category	Conditioning Factors	Type	Data Source	Values
Topographical	Elevation (m)	Continuous	DEM	−26.8~433.7
	Curvature	Continuous	DEM	−11.7~11.3
	Slope angle (°)	Continuous	DEM	0~81.2
	MED (m)	Continuous	DEM	3.8~370.1
	Slope aspect	Categorical	DEM	Flat; East; Northeast; North; Southeast; South; Southwest; West; Northwest
	TWI	Continuous	DEM	1.8~25.6
Geological	Distance to faults (km)	Continuous	Geological map	0~1838.5
	Distance to streams (km)	Continuous	GESI	0~3940.4
	Lithology	Categorical	Geological map	Metasandstone; Gneiss; Glutenite; Siltite; Granite; Calcareous mudstone; Diorite
Triggering factors	24-Maximum Rainfall (mm)	Continuous	GZB	66.6~215.6
	72-Maximum Rainfall (mm)	Continuous	GZB	162.2~380.6
	Monthly Maximum Rainfall (mm)	Continuous	GZB	250.6~743.0
	Distance to roads (km)	Continuous	GESI	0~1838.5

2.2.2. Landslide-Related Influencing Factors

Topographical factors were derived from a digital elevation model (DEM) with a resolution of 5 × 5 m (obtained from Guangzhou Urban Planning Survey and Design Institute, GZPI) and included elevation, curvature, slope angle, maximum elevation difference (MED), slope aspect and topographic wetness index (TWI). Elevation affects the occurrence of landslides, as it is related to both rainfall and runoff and has been referenced many times [28,29]. The curvature reflects the undulation of the ground, which affects surface water runoff [30,31]. Slope gradient reflects potential energy and pore pressure, which help to form a free surface. Slope angle and MED were considered the essential factors [32,33]. The slope aspect affects the solar radiation, which further affects the vegetation and rock weathering degree and ultimately results in differences in the stability of a slope [34]. TWI reflects both the basic terrain and soil moisture content [35].

Geological factors such as fault information and lithology were acquired from the geological map at a 1:50,000 scale. Distance to faults, which was calculated by the Euclidean distance tool in ArcGIS, was selected. Landslides tend to be distributed along rivers and roads, and the information on rivers and roads was obtained from Google Earth satellite imagery (GESI). Similarly, distances to rivers and roads were also selected.

Rainfall, as the major triggering factor in the area, should be considered. Landslides are frequently accompanied by heavy or continuous rainfall, which has been verified by many researchers and landslide events [36–38]; 24-maximum rainfall, 72-maximum rainfall and

monthly maximum rainfall were selected. The thematic maps were generated by the spatial interpolation tool in ArcGIS with the information from 63 precipitation stations (2000~2022) in the study area. Rainfall data were collected from the Guangzhou Meteorological Bureau (GZB). Earthquakes are another triggering factor, but their seismic intensity was the same in the study area. Therefore, rainfall was selected as the natural triggering factor, and distance to roads as the human triggering factor.

2.2.3. Mapping Units

The selection of a mapping unit is another essential factor for LSP, as it may affect the accuracy and precision of the results. Slope units are more reasonable and reliable compared to other terrain units, since landslides occur naturally on a slope [39–41]. Accordingly, the study area was divided into 3171 slope units based on the hydrological analysis tool in the ArcGIS platform (Figure 1). The mean value of each influencing factor was taken as an attribute of each slope unit. The thematic maps are shown in Figure 5. Nevertheless, the slope units, which were limited by drainage and water dividing lines, needed boundary revision artificially based on satellite imagery.

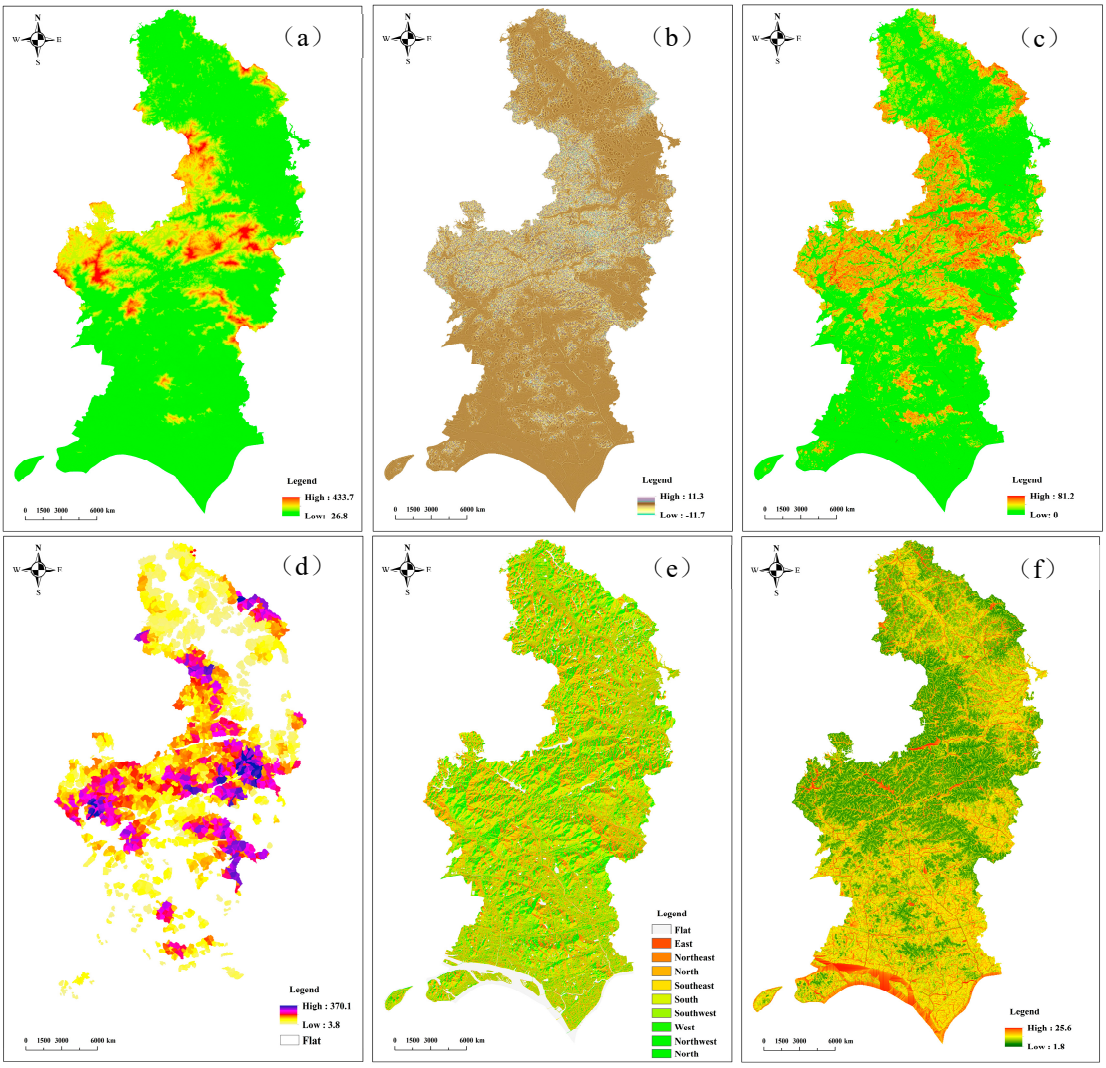

Figure 5. *Cont.*

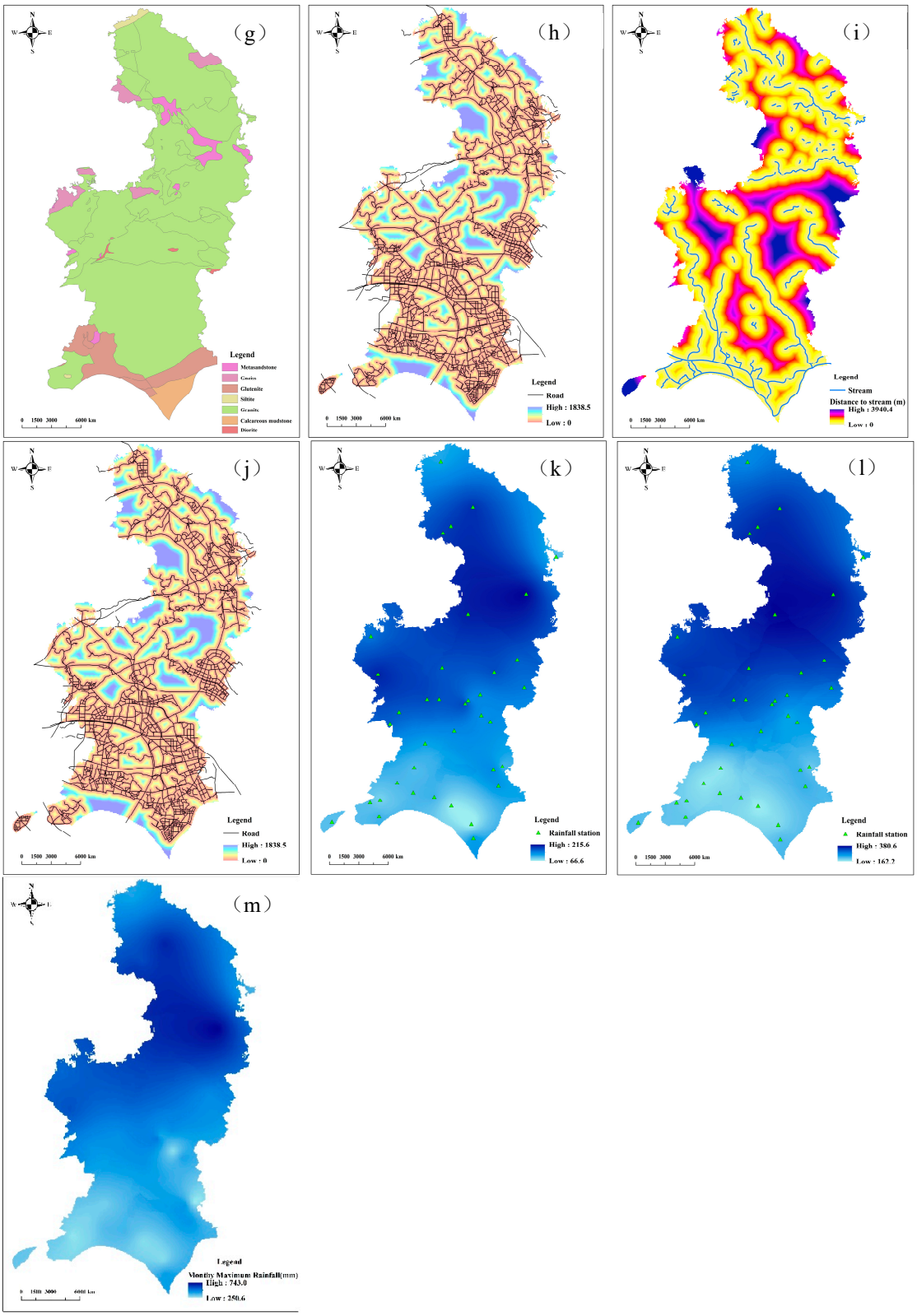

Figure 5. Study area thematic maps: (**a**) Elevation; (**b**)Curvature; (**c**) Slope; (**d**) MED; (**e**) Slope Aspect; (**f**) TWI; (**g**) Lithology; (**h**) DTF; (**i**) DTS; (**j**) DTR; (**k**) 24-Maximum Rainfall; (**l**) 72-Maximum Rainfall; (**m**) Monthly Maximum Rainfall.

3. Methods

3.1. Flow Chart of Landslide Susceptibility Prediction

The aim of the study was to explore the potential possibilities of IF for LSP, and its performance was compared to another advanced machine learning method. The flow chart of methods applied in the study is illustrated in Figure 6. The methodology mainly included five steps. First, data included conditioning factors, mapping units and landslide inventory were prepared for modeling. Then, the samples were divided into two parts for model training and validation. Subsequently, isolation forest and adaptive boosting were applied to the modeling of landslide susceptibility prediction using Python 3.7 with the ensemble library of scikit-learn. After that, the performance of the two models was compared and discussed based on the parameters of accuracy and specificity. Finally, the potential possibilities of IF for LSP were determined.

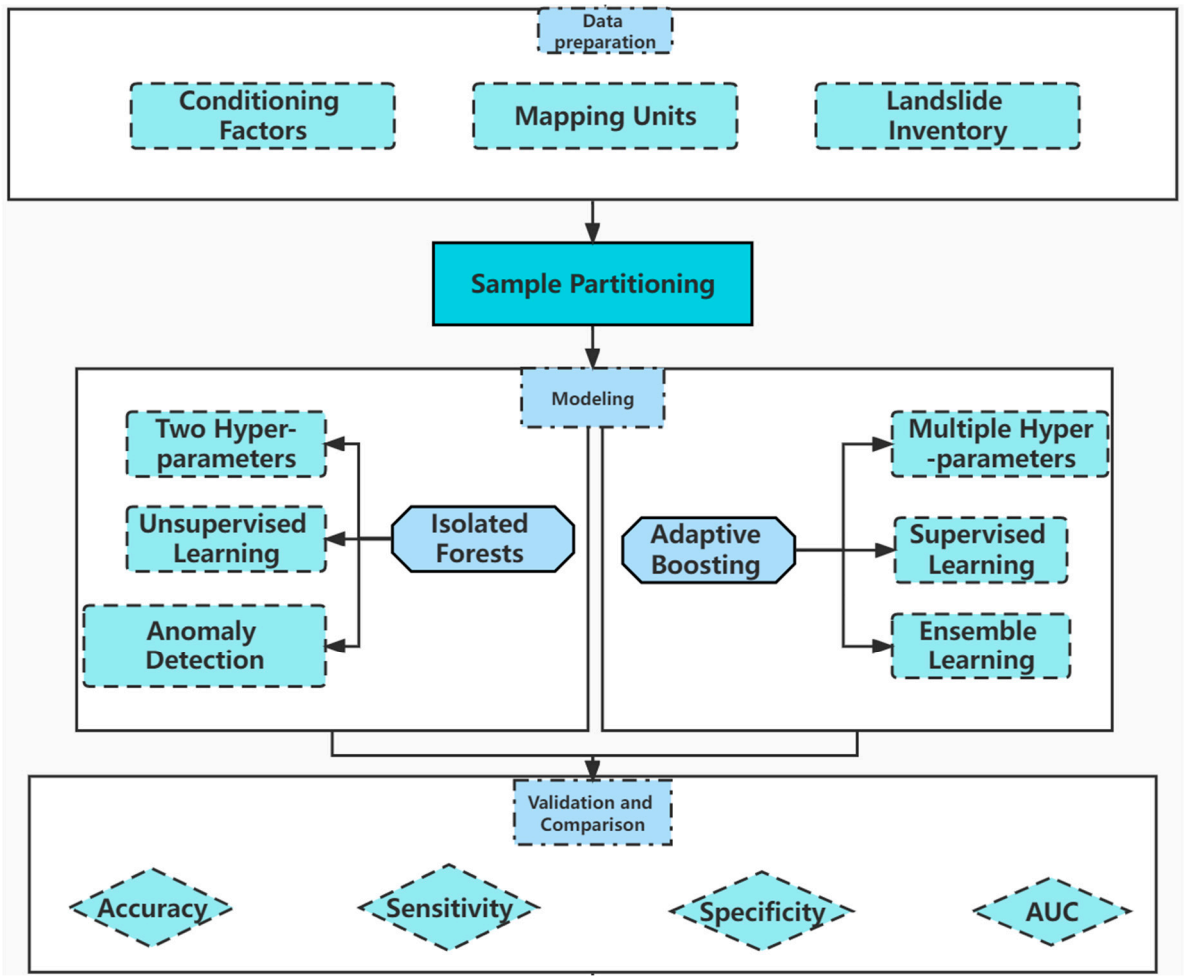

Figure 6. Flow chart of the methods applied in the study.

3.2. Sampling Strategy

The sampling strategy involved the problem of sample selection and allocation. LSP can be treated as a binary regression problem that needs both positive (landside) and negative (non-landside) samples. Thus, an equal number of non-landslide samples (which were selected randomly from the "safe area" where landslide density was low) and validated landslides were applied to the modeling of Ada-DT. A small fraction of anomalous samples (landslide samples) were required for the modeling of IF. The ratio of anomalous samples, which is also called contamination, was set to be 0.1 (25 of 239 validated landslide samples were randomly selected).

On the other hand, the samples should be divided into two parts, one for model training, and the other for validation. Leaving-one-out is the most popular measure for the allocation of samples, which divides the data into independent training and validation sets [42]. K-fold cross-validation (K = 10 in this study) was applied to evaluate the performance of the models. It divided the samples into 10 independent groups, with one group for validation and the other nine for training, repeating the process 10 times until all samples were involved [43].

3.3. Isolation Forests

IF was first introduced by Liu in 2008 and further improved in 2011. It belongs to an unsupervised learning method and needs outliers that are more likely to be separated [44]. Its main idea is based on the isolation and ensemble concept, which is usually applied to anomaly detection [45]. Considering that outliers are different from the normal samples and occupy a small ratio of a total dataset, they are more susceptible to being separated, which is referred to as isolation. A series of random binary isolated trees are constructed for data of different dimensions describing the same object. These IFs are either leaf nodes with no children or internal nodes with only two children. Normally, the anomalies have more chances to be separated closer to the root of an IF, while the normal samples will finish the process at the deeper end of an IF.

Path length and anomaly score are two indicators for testing anomaly samples, which can be defined as follows:

$$C(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{1}$$

where $H(i)$ is the harmonic number, and the value can be estimated as $\ln(i) + 0.5772156649$; $C(n)$ is the average value of the path length when the number of samples n is given, used to normalize the path length $h(x)$ of sample x.

$$s(x, n) = 2^{-\frac{E(h(x))}{C(n)}} \tag{2}$$

where $E(h(x))$ represents the expected path length of sample x in a batch of IF; $s(x, n)$ is the normalization of the path corresponding to this value. The path length is inversely proportional to the anomaly score. The smaller the depth, the higher the anomaly score, that is, the greater the probability that the sample belongs to an abnormal sample. Finally, the discrimination basis of anomaly detection is as follows:

$$\begin{cases} s(x, n) \to 1, abnormal \\ s(x, n) \to 0, normal \\ s(x, n) \to 0.5, uncertain \end{cases} \tag{3}$$

Two hyper-parameters, sub-sampling size λ and the number of trees n, need to be determined in advance. Normally, λ and n should be set to be defaults, which are 256 and 100, respectively [44]. Python 3.7 was responsible for the modeling with the use of the ensemble library of scikit-learn. The observed landslide samples were involved in the training and validation of the model.

3.4. Adaptive Boosting (Adaboost)

Boosting is an important integrated learning technique that enhances the weak classifier into a strong classifier with high accuracy. The Adaboost algorithm was first introduced by Freund and Schapire [46] and has been proven to be an effective and practical boosting algorithm. It selects the weak classifier with the smallest weight coefficient from the trained weak classifiers by adjusting the weight of samples and classifiers and combines them into a final strong classifier. AdaBoost belongs to an iterative algorithm where the misclassified

samples are given more weight and accordingly reduce both bias and variance [47]. An ensemble model that the decision tree (DT) involves as the weak classifier, called Ada-DT, was constructed for LSP [48].

The Ada-DT model is applied to solve the problem of binary-class classification, which needs both positive and negative samples. The observed landslides were set to be the positive samples with the label of "1", while non-landslide locations were set to be the negative samples with "0". It was modeled in Python 3.7 with the use of the AdaBoost class library of scikit-learn. The number of iterations was set to 100, and the other involved parameters were left at their default values.

Finally, two landslide susceptibility zoning maps were constructed using IF and Ada-DT models, and the study area was classified into five categories of landslide susceptibility levels as very low (0~0.2), low (0.2~0.4), moderate (0.4~0.6), high (0.6~0.8) and very high (0.8~1.0) based on the equal spacing principle.

3.5. Susceptibility Model Evaluation

It was necessary to compare the performance of the two models using related indices such as accuracy, sensitivity, specificity, and receiver operating characteristic curve (ROC). These indices have been widely applied to evaluate binary classification models and adopted by many LSP studies [49,50]. Accuracy measures the proportion of all samples that are correctly divided, and sensitivity reflects the positives that are correctly identified, while specificity is for the negatives. The area under the ROC curve (AUC) ranges from 0.5 to 1, and the higher the value, the greater the performance. Detailed information can be found in another study [51]. Related indices can be calculated based on the following equations:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Sensitivity &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{FP+TN} \end{aligned} \tag{4}$$

where True Positives (*TP*) and True Negatives (*TN*) are the number of units that are predicted correctly; False Positives (*FP*) and False Negatives (*FN*) are the number of units that are predicted incorrectly.

The performance of unsupervised learning is inherently difficult to evaluate by the above indicators because it lacks the use of labeled samples. The sample set consisting of 239 landslides and 239 non-landslides was applied for the modeling of Ada-DT by 10-fold cross-validation, while 25 observed landslide samples and 190 non-landslides (about 90% of 239 non-landslides) were applied for the training of IF, and the remaining samples were for validation. Thus, the values of TP, TN, FP, and FN could be calculated.

4. Results

4.1. Landslide Susceptibility Maps

It is generally believed that the observed landslide samples should be located in the very high or high susceptibility zones as much as possible, while non-landslide samples should appear in a safe area whose susceptibility is low or very low. A total of 478 samples consisting of 239 landslides and 239 non-landslides were applied to model Ada-DT, and the landslide susceptibility map was obtained based on the probability value, which is also called the landslide susceptibility index (LSI). Similarly, the second landslide susceptibility map was obtained based on the anomaly score. The distribution of different susceptible levels is shown in Figure 7. As for the IF model, the high level occupied the biggest proportion, at 36.01% of the study area, while the low level occupied the smallest proportion, at 8.26%. Very high and very low levels accounted for 22.96% and 22.90%, and moderate levels accounted for 9.87%. The landslide samples (black points) were basically located in red areas whose susceptibility was very high, while the non-landslide samples (blue points)

were mostly in the green or deep green areas whose susceptibility was low or very low. Obviously, some non-landslide samples were misclassified into red or orange areas.

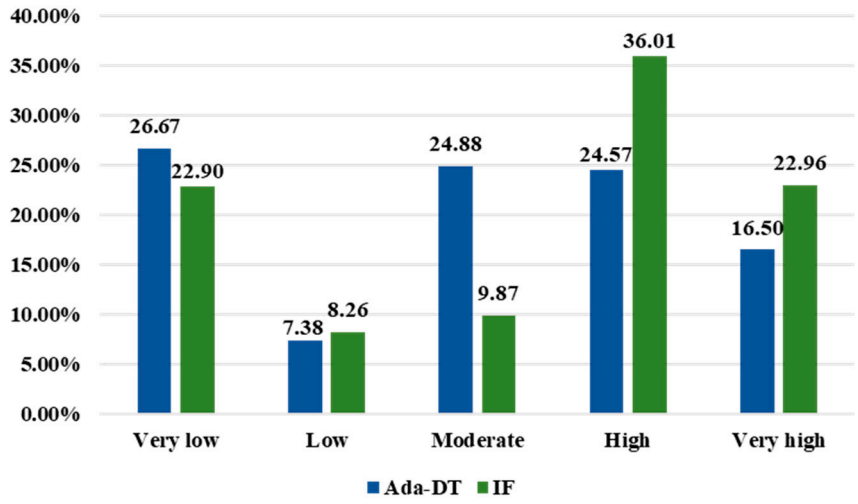

Figure 7. The percentages of the different susceptibility classes in the two models.

Compared with the IF model, the Ada-DT model classified a larger ratio of moderate and very low levels, as 24.88% and 26.67% of the study area, respectively. The low, high and very high levels accounted for 7.83%, 24.57%, and 16.50%, respectively, which were all smaller than those of the IF model. More landslide locations were predicted in the yellow areas, while non-landslides were in green areas (Figure 8). Basically, no non-landslide samples could be found in the red area.

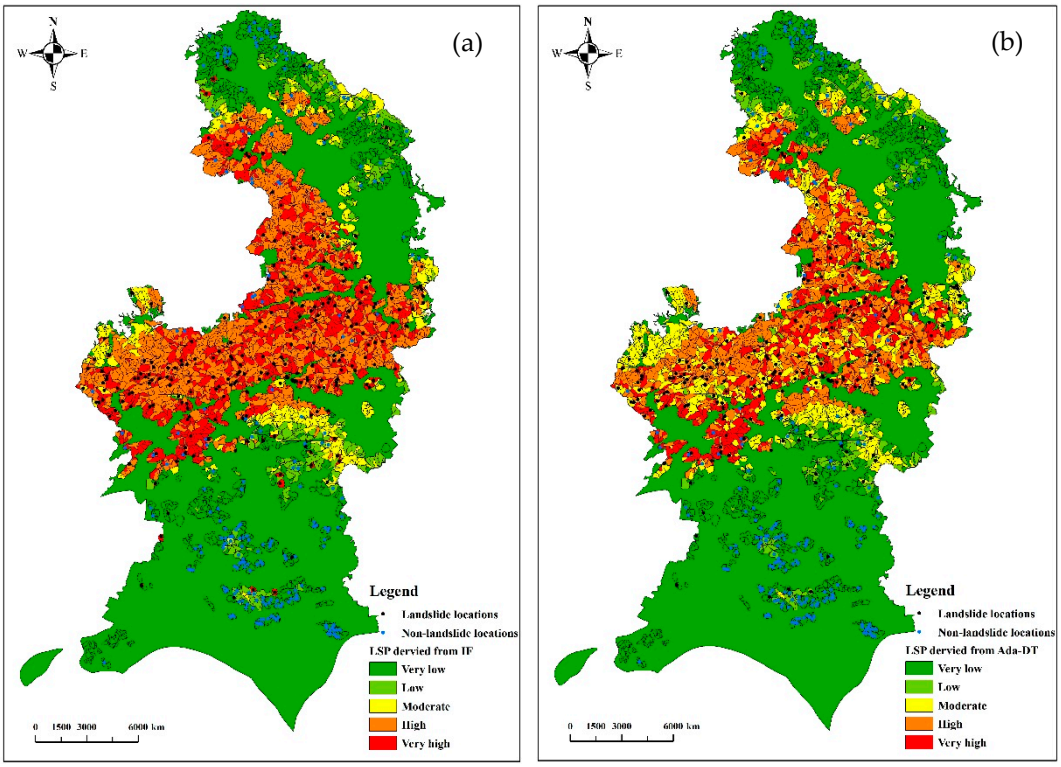

Figure 8. Landslide susceptibility map: (**a**) IF model; (**b**) Ada-DT model.

It is obvious that the high and very high susceptibility areas predicted by IF were larger. Risk areas belonging to high or very high susceptibility levels were similar, mainly

distributed in the central region, including Xinlong town, Changling street, and Linhe street, while the low-risk areas with low or very low susceptibility included Yunpo street and Jiufo street, which are located in the northern and southern areas. These results are basically consistent with the field investigation and historical records.

4.2. Analysis of Main Influencing Factors

Landslides were mainly distributed in the central region of the study area, while few appeared in the north and south. Analyzing the major conditioning factors helps understand the reason for the difference, which is also significant for landslide prevention. The Gini index, which measures the correlation degree between variables and results, was applied to evaluate the relative importance of the conditioning factor, and the results were normalized [52]. The 24-maximum rainfall, slope, monthly maximum rainfall, and MED were regarded as the key factors, the weight values of which were relatively large, at 0.37, 0.14, 0.13, and 0.13, respectively (Figure 9 and Table 2). The weight values of lithology, DTR, DTF, and TWI were small, indicating they had little influence on a landslide.

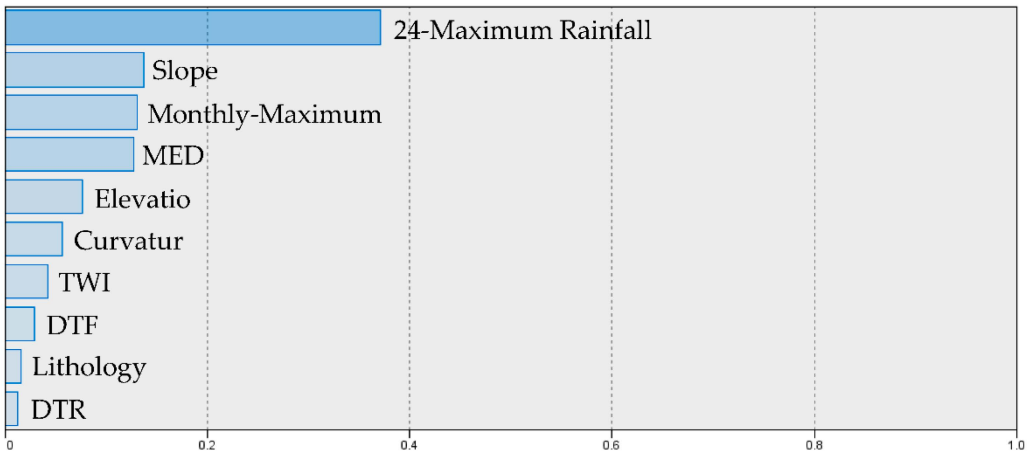

Figure 9. Parametric importance graphics based on Gini index.

Table 2. Conditioning factors assigned by the Ada-DT.

Method	24-Maximum Rainfall	Slope	Monthly Maximum Rainfall	MED	Elevation	Curvature	TWI	DTF	Lithology	DTR
Gini	0.37	0.14	0.13	0.13	0.08	0.06	0.04	0.03	0.02	0.01

Compared with the northern and southern regions, the central region is steeper and rainfall is concentrated, which promotes the occurrence of landslides. More importantly, population is concentrated and slope excavation is common in the central area due to the construction of tenements and roads. Thus, the slope toes have become steeper and slopes are prone to be unstable. Accordingly, landslides occurred more frequently in the central region, which is consistent with the findings of our field investigation.

5. Discussion

5.1. Model Validation and Comparison

The statistical indices provided a more detailed and comprehensive comparison between the models, and the results are shown in Table 3. The results confirmed that the Ada-DT model performed better in terms of accuracy, specificity, and AUC, the values of which were 87.50%, 91.67%, and 0.906, respectively, followed by the IF model (accuracy = 85.83%, specificity = 81.67%, and AUC = 0.875). As for sensitivity, the IF model performed better (sensitivity = 90.00%) than the Ada-DT model (sensitivity = 83.33%). The performance of the two models declined compared to the training data, except for the sensitivity value of

the IF model. Overall, the performance of the two models was satisfactory, as the values of the AUC were all above 0.87 (Figure 10).

Table 3. Model performance using related indices.

	Parameter			
Method	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**AUC**
IF	85.83	90	81.67	0.875
Ada-DT	87.50	83.33	91.67	0.910

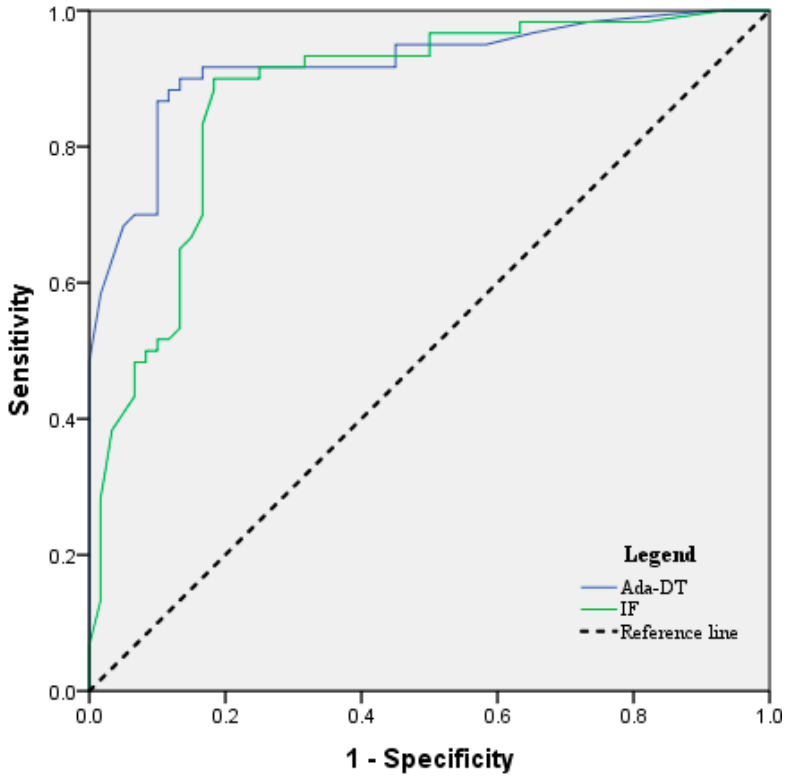

Figure 10. The AUC values were calculated using the IF and Ada-DT models.

The results predicted by the IF were biased since only a small fraction of the positive samples were involved in the training, and there were not enough samples to compete with negative samples. The areas were more likely to be predicted as very high susceptibility, and the sensitivity value was higher compared to TCC, while the specificity value was smaller compared to that of the Ada-DT model. For TCC, both positive and negative samples are necessary, but the negative samples will inevitably (more or less) appear in areas prone to landslides, as they are selected randomly. In other words, the results predicted by the IF model were too conservative, while the TCC tended to be risky.

5.2. The Applicability and Advancement of Isolation Forests

New machine learning methods such as gradient boost decision tree (GBDT) and artificial neural networks usually involve several or more hyper-parameters that need to be tuned to optimize the models [53,54]. Although some approaches such as the genetic algorithm method and grid search method have been applied to tune the parameters, they are difficult to implement and sometimes even pointless, whereas IF is easier to implement, as only two parameters need to be tuned. The values of the two parameters have been determined by Tony [44] (2012).

On the other hand, small sub-samples are preferred for IF because they help solve the problems of swamping and masking, which have been emphasized in anomaly de-

tection [55]. The efficiency of IF remains high under normal sample participation. The prediction effect will increase with the participation of abnormal samples and normal samples, but not significantly. The ratio of abnormal to normal samples, also known as contamination, is set automatically to 0.1 or even lower. In other words, the application of IF requires a low number of abnormal samples. Most machine learning methods expect more data to train the models, which is not easy to satisfy in LSP, especially for an area where landslide inventory is incomplete or the landslide samples are limited [56,57]. On the other hand, hypothetical negative samples are not necessary, as positive samples are more credible. Therefore, IF outperforms some machine learning algorithms, as (1) IF is easier to implement; and (2) IF has lower requirements for the quality of samples, which will further enhance the practicability of the machine learning algorithm in geological hazard evaluation. However, the performance of IF was slightly worse than that of Ada-DT.

5.3. The Validation of Unsupervised Learning and One Class-Classifier

The validation process for unsupervised learning and OCC are awkward, as samples without labels or only positive or negative samples are involved for modeling. Accordingly, the performance of unsupervised learning and OCC is difficult to comprehensively compare with TCC such as random forest [58]. Some researchers have compared the performance of OCC with TCC based on the values of AUC, spatial variation of the susceptibility value, and landslide distribution density [59]. However, the prediction for a low susceptibility area is also non-negligible, and validation is necessary. The modeling of IF requires a majority of normal samples and a small number of abnormal samples. This study prepared non-landslide samples from the "safe area" randomly, where landslides are seldom or not distributed. Five-fold cross-validation was adopted to evaluate the performance of Ada-DT, and the values of TN and FN were obtained. The non-landslide samples were also distributed on the primal landslide susceptibility map predicted by IF. We can roughly compare the accuracy of the two models in predicting low-susceptibility areas, although uncertainty is obvious.

5.4. Improving the Quality of Samples

Neither conservative nor risky results are conducive to practical application. Therefore, it is reasonable to improve the performance of LSP by combining the advantages of IF and TCC. In this study, the performance of IF was more guaranteed, and the primal landslide susceptibility map deserves further application. It is suggested that the non-landslide samples be selected from the very low susceptibility areas displayed on the primal map [60–62]. The samples filtered in this way are almost impeccable, which will further improve the performance of TCC.

6. Conclusions

This study applied isolation forest, which performs well in anomaly detection to ease the influence from the low quality of samples on LSP, and its performance was compared to another ensemble learning method, AdaBoost-DT.

IF was proved to be effective for LSP, as its performance was satisfactory when landslide samples were incomplete or insufficient. It improves the application possibilities for machine learning algorithms in geohazards. IF has no dependence on the selection of non-landslide samples, so noise is eliminated directly from the data. The results predicted by IF tended to be conservative compared to Ada-DT. Consequently, for a more robust sampling approach, it is suggested that the non-landslide samples should be selected from the very low susceptibility areas predicted by IF.

Landslide susceptibility maps predicted by IF and Ada-DT were both reasonable in predicting the risk areas, and Xinlong town, Changling street and Linhe street in the center of the study area should be focused on. Steep terrain, sufficient rainfall, and complex geological conditions promote the development of landslides in the south of the study area.

The research on the prediction of landslide susceptibility remains to be elucidated in more detail. The conclusions are of great significance for the research on artificial intelligence in geohazards. However, the methods adopted also have limitations:

1. The performance of IF could be compared to more modeling methods;
2. Data preprocessing could be performed before modeling;
3. Different values of contamination for IF could be discussed.

Author Contributions: Q.Z., writing—original draft, methodology and software; Z.L., W.L., W.P. and L.C., review and validation; K.J., L.L., H.H. and Q.Z., reviewing and editing. S.Z. is responsiable for the review and validation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Key-Area Research and Development Program of Guangdong Province (Grant No. 2020B0101130009), Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning (No. 2020B121202019), The Science and Technology Foundation of Guangzhou Urban Planning & Design Survey Research Institute (Grant No. RDI2210204140, RDI2210204146, RDI2220204031, RDI2220204037, RDI2220204125), and Postdoctoral Research Project of Guangzhou (20220402).

Data Availability Statement: The related machine learning code applied in the study is available at https://github.com/Liangzhu-mz, accessed on 9 December 2020.

Acknowledgments: The authors would like to thank the Editors and anonymous reviewers for their valuable comments, which improved this paper.

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

References

1. Yi, Y.; Zhang, Z.; Zhang, W.; Xu, Q.; Deng, C.; Li, Q. GIS-based earthquake-triggered-landslide susceptibility mapping with an integrated weighted index model in Jiuzhaigou region of Sichuan Province, China. *Nat. Hazards Earth Syst. Sci.* **2019**, *19*, 1973–1988. [CrossRef]
2. Kavoura, K.; Sabatakakis, N. Investigating landslide susceptibility procedures in Greece. *Landslides* **2020**, *17*, 127–145. [CrossRef]
3. Bravo-López, E.; Del Castillo, T.F.; Sellers, C.; Delgado-García, J. Landslide susceptibility mapping of landslides with artificial neural networks: Multi-approach analysis of backpropagation algorithm applying the neuralnet package in Cuenca, Ecuador. *Remote. Sens.* **2022**, *14*, 3495. [CrossRef]
4. Merghadi, A.; Abderrahmane, B.; Bui, D.T. Landslide susceptibility assessment at Mila Basin (Algeria): A comparative assessment of prediction capability of advanced machine learning methods. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 268. [CrossRef]
5. Huang, F.; Zhang, J.; Zhou, C.; Wang, Y.; Huang, J.; Zhu, L. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. *Landslides* **2020**, *17*, 217–229. [CrossRef]
6. Miao, F.; Wu, Y.; Xie, Y.; Li, Y. Prediction of landslide displacement with step-like behavior based on multialgorithm optimization and a support vector regression model. *Landslides* **2018**, *15*, 475–488. [CrossRef]
7. Huang, F.; Cao, Z.; Guo, J.; Jiang, S.H.; Li, S.; Guo, Z. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. *Catena* **2020**, *191*, 104580. [CrossRef]
8. Peethambaran, B.; Anbalagan, R.; Kanungo, D.P.; Goswami, A.; Shihabudheen, K.V. A comparative evaluation of supervised machine learning algorithms for township level landslide susceptibility zonation in parts of Indian Himalayas. *Catena* **2020**, *195*, 104751. [CrossRef]
9. Liang, Z.; Wang, C.; Han, S.; Ullah Jan Khan, K.; Liu, Y. Classification and susceptibility assessment of debris flow based on a semi-quantitative method combination of the fuzzy C-means algorithm, factor analysis and efficacy coefficient. *Nat. Hazards Earth Syst. Sci.* **2020**, *20*, 1287–1304. [CrossRef]
10. Tang, R.X.; Kulatilake, P.H.; Yan, E.; Cai, J.S. Evaluating landslide susceptibility based on cluster analysis, probabilistic methods, and artificial neural networks. *Bull. Eng. Geol. Environ.* **2020**, *79*, 2235–2254. [CrossRef]
11. Kornejady, A.; Ownegh, M.; Bahremand, A. Landslide susceptibility assessment using maximum entropy model with two different data sampling methods. *Catena* **2017**, *152*, 144–162. [CrossRef]
12. Jiang, S.H.; Huang, J.; Huang, F.; Yang, J.; Yao, C.; Zhou, C.B. Modelling of spatial variability of soil undrained shear strength by conditional random fields for slope reliability analysis. *Appl. Math. Model.* **2018**, *63*, 374–389. [CrossRef]
13. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2019**, *513*, 429–441. [CrossRef]
14. Xie, Y.; Qiu, M.; Zhang, H.; Peng, L.; Chen, Z. Gaussian distribution based oversampling for imbalanced data classification. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 667–679. [CrossRef]
15. Chang, L.; Zhang, R.; Wang, C. Evaluation and prediction of landslide susceptibility in Yichang section of Yangtze River Basin based on integrated deep learning algorithm. *Remote Sens.* **2022**, *14*, 2717. [CrossRef]

16. Mao, Y.; Mwakapesa, D.S.; Wang, G.; Nanehkaran, Y.A.; Zhang, M. Landslide susceptibility modelling based on AHC-OLID clustering algorithm. *Adv. Space Res.* **2021**, *68*, 301–316. [CrossRef]
17. Alam, S.; Sonbhadra, S.K.; Agarwal, S.; Nagabhushan, P. One-class support vector classifiers: A survey. *Knowl.-Based Syst.* **2020**, *196*, 105754. [CrossRef]
18. Zhu, L.; Wang, G.; Huang, F.; Li, Y.; Chen, W.; Hong, H. Landslide susceptibility prediction using sparse feature extraction and machine learning models based on gis and remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3001505. [CrossRef]
19. Chen, S.; Miao, Z.; Wu, L.; He, Y. Application of an Incomplete Landslide Inventory and One Class Classifier to Earthquake-Induced Landslide Susceptibility Mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 1649–1660. [CrossRef]
20. Wang, H.; Zhang, L.; Yin, K.; Luo, H.; Li, J. Landslide identification using machine learning. *Geosci. Front.* **2021**, *12*, 351–364. [CrossRef]
21. Varnes, D.J. Landslide hazard zonation: A review of principles and practice, Commission on Landslides of the IAEG. *UNESCO Nat. Hazards* **1984**, *3*, 61.
22. Du, J.; Han, G.; Lin, C.; Martinez-Garcia, M. ITrust: An anomaly-resilient trust model based on isolation forest for underwater acoustic sensor networks. *IEEE Trans. Mob. Comput.* **2020**, *21*, 1684–1696. [CrossRef]
23. Zou, Z.; Xie, Y.; Huang, K.; Xu, G.; Feng, D.; Long, D. A docker container anomaly monitoring system based on optimized isolation forest. *IEEE Trans. Cloud Comput.* **2019**, *10*, 134–145. [CrossRef]
24. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the ICDM '08: 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE Computer Society: Manhattan, NY, USA, 2008; pp. 413–422.
25. 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]
26. 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]
27. Rossi, M.; Guzzetti, F.; Salvati, P.; Donnini, M.; Napolitano, E.; Bianchi, C. A predictive model of societal landslide risk in Italy. *Earth-Science Rev.* **2019**, *196*, 102849. [CrossRef]
28. Conforti, M.; Pascale, S.; Robustelli, G.; Sdao, F. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). *Catena* **2014**, *113*, 236–250. [CrossRef]
29. Huang, F.; Pan, L.; Fan, X.; Jiang, S.H.; Huang, J.; Zhou, C. The uncertainty of landslide susceptibility prediction modeling: Suitability of linear conditioning factors. *Bull. Eng. Geol. Environ.* **2022**, *81*, 182. [CrossRef]
30. Camilo, D.C.; Lombardo, L.; Mai, P.M.; Dou, J.; Huser, R. Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized Generalized Linear Model. *Environ. Model. Softw.* **2017**, *97*, 145–156. [CrossRef]
31. Miao, F.; Wu, Y.; Török, Á.; Li, L.; Xue, Y. Centrifugal model test on a riverine landslide in the Three Gorges Reservoir induced by rainfall and water level fluctuation. *Geosci. Front.* **2022**, *13*, 101378. [CrossRef]
32. Liang, Z.; Wang, C.-M.; Zhang, Z.-M.; Khan, K.-U. A comparison of statistical and machine learning methods for debris flow susceptibility mapping. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 1887–1907. [CrossRef]
33. Miao, F.; Zhao, F.; Wu, Y.; Li, L.; Xue, Y.; Meng, J. A novel seepage device and ring-shear test on slip zone soils of landslide in the Three Gorges Reservoir area. *Eng. Geol.* **2022**, *307*, 106779. [CrossRef]
34. Liang, Z.; Wang, C.; Duan, Z.; Liu, H.; Liu, X.; Ullah Jan Khan, K. A hybrid model consisting of supervised and unsupervised learning for landslide susceptibility mapping. *Remote Sens.* **2021**, *13*, 1464. [CrossRef]
35. Zhou, C.; Yin, K.; Cao, Y.; Ahmed, B.; Li, Y.; Catani, F.; Pourghasemi, H.R. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China. *Comput. Geosci.* **2018**, *112*, 23–37. [CrossRef]
36. Cho, S.E. Prediction of shallow landslide by surficial stability analysis considering rainfall infiltration. *Eng. Geol.* **2017**, *231*, 126–138. [CrossRef]
37. Pradhan, A.M.S.; Lee, S.-R.; Kim, Y.-T. A shallow slide prediction model combining rainfall threshold warnings and shallow slide susceptibility in Busan, Korea. *Landslides* **2019**, *16*, 647–659. [CrossRef]
38. Xing, X.; Wu, C.; Li, J.; Li, X.; Zhang, L.; He, R. Susceptibility assessment for rainfall-induced landslides using a revised logistic regression method. *Nat. Hazards* **2021**, *106*, 97–117. [CrossRef]
39. Ba, Q.; Chen, Y.; Deng, S.; Yang, J.; Li, H. A comparison of slope units and grid cells as mapping units for landslide susceptibility assessment. *Earth Sci. Inform.* **2018**, *11*, 373–388. [CrossRef]
40. Huang, F.; Tao, S.; Chang, Z.; Huang, J.; Fan, X.; Jiang, S.-H.; Li, W. Efficient and automatic extraction of slope units based on multi-scale segmentation method for landslide assessments. *Landslides* **2021**, *18*, 3715–3731. [CrossRef]
41. Chang, Z.; Catani, F.; Huang, F.; Liu, G.; Meena, S.R.; Huang, J.; Zhou, C. Landslide susceptibility prediction using slope unit-based machine learning models considering the heterogeneity of conditioning factors. *J. Rock Mech. Geotech. Eng.* **2022**. [CrossRef]
42. Stock, M.; Pahikkala, T.; Airola, A.; Waegeman, W.; De Baets, B. Algebraic shortcuts for leave-one-out cross-validation in supervised network inference. *Briefings Bioinform.* **2020**, *21*, 262–271. [CrossRef] [PubMed]
43. Wong, T.-T.; Yeh, P.-Y. Reliable Accuracy Estimates from *k*-Fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 1586–1594. [CrossRef]

44. Karczmarek, P.; Kiersztyn, A.; Pedrycz, W.; Al, E. K-Means-based isolation forest. *Knowl.-Based Syst.* **2020**, *195*, 105659. [CrossRef]
45. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data (TKDD)* **2012**, *6*, 1–39. [CrossRef]
46. Freund, Y.; Schapire, R.E. A desicion-theoretic generalization of on-line learning and an application to boosting. *Eur. Conf. Comput. Learn. Theory* **1995**, *55*, 23–37.
47. Shahraki, A.; Abbasi, M.; Haugen, Ø. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103770. [CrossRef]
48. Bui, D.T.; Ho, T.-C.; Pradhan, B.; Pham, B.-T.; Nhu, V.-H.; Revhaug, I. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, bagging, and MultiBoost ensemble frameworks. *Environ. Earth Sci.* **2016**, *75*, 1101.
49. Dou, J.; Yunus, A.P.; Bui, D.T.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.W.; Han, Z.; Pham, B.T. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides* **2020**, *17*, 641–658. [CrossRef]
50. Liang, Z.; Wang, C.; Khan, K.U.J. Application and comparison of different ensemble learning machines combining with a novel sampling strategy for shallow landslide susceptibility mapping. *Stoch. Environ. Res. Risk Assess.* **2020**, *35*, 1243–1256. [CrossRef]
51. Parmigiani, G. Receiver operating characteristic curves with an indeterminacy zone. *Pattern Recognit. Lett.* **2020**, *136*, 94–100. [CrossRef]
52. Kausar, N.; Majid, A. Random forest-based scheme using feature and decision levels information for multi-focus image fusion. *Pattern Anal. Appl.* **2016**, *19*, 221–236. [CrossRef]
53. 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]
54. Bragagnolo, L.; da Silva, R.V.; Grzybowski, J.M.V. Artificial neural network ensembles applied to the mapping of landslide susceptibility. *Catena* **2020**, *184*, 104240. [CrossRef]
55. Iwata, T.; Toyoda, M.; Tora, S.; Ueda, N. Anomaly detection with inexact labels. *Mach. Learn.* **2020**, *109*, 1617–1633. [CrossRef]
56. Zhu, A.-X.; Miao, Y.; Yang, L.; Bai, S.; Liu, J.; Hong, H. Comparison of the presence-only method and presence-absence method in landslide susceptibility mapping. *Catena* **2018**, *171*, 222–233. [CrossRef]
57. Zhu, Q.; Chen, L.; Hu, H.; Pirasteh, S.; Li, H.; Xie, X. Unsupervised feature learning to improve transferability of landslide susceptibility representations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 3917–3930. [CrossRef]
58. Yuan, Y.-H.; Li, J.; Li, Y.; Gou, J.; Qiang, J. Learning unsupervised and supervised representations via general covariance. *IEEE Signal Process. Lett.* **2020**, *28*, 145–149. [CrossRef]
59. 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]
60. Tehrani, F.S.; Santinelli, G.; Herrera, M.H. Multi-Regional landslide detection using combined unsupervised and supervised machine learning. *Geomat. Nat. Hazards Risk* **2021**, *12*, 1015–1038. [CrossRef]
61. Zhiyong, L.; Liu, T.; Wang, R.Y.; Benediktsson, J.A.; Saha, S. Automatic Landslide Inventory Mapping Approach Based on Change Detection Technique With Very-High-Resolution Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 6000805. [CrossRef]
62. Tang, Q.; Chen, Y.; Jia, R.; Guo, W.; Chen, W.; Li, X.; Gao, H.; Zhou, Y. Effect of clay type and content on the mechanical properties of clayey silt hydrate sediments. *J. Pet. Sci. Eng.* **2022**, *220*, 111203. [CrossRef]