remote sensing

MDPI

Article

Feature Engineering of Geohazard Susceptibility Analysis Based on the Random Forest Algorithm: Taking Tianshui City, Gansu Province, as an Example

Xiao Ling [1,†], Yueqin Zhu [2,†], Dongping Ming [1,*], Yangyang Chen [3], Liang Zhang [1] and Tongyao Du [1]

[1] School of Information Engineering, China University of Geosciences (Beijing), 29 Xueyuan Road, Beijing 100083, China
[2] National Institute of Natural Hazards, Ministry of Emergency Management, Beijing 100085, China
[3] China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing 100083, China
* Correspondence: mingdp@cugb.edu.cn
† These authors contributed equally to this work.

Abstract: In this paper, Feature Engineering (FE) was applied to Landslide Susceptibility Mapping (LSM), while the most suitable conditioning feature dataset and analysis method were tested and analyzed. Tianshui city was taken as the study area, three types of geohazard (collapse, landslide, and unstable slopes) were used, while a total of twenty-three conditioning features were generated; two dimensionless methods (normalization and standardization) were tested afterward. Four Random-Forest-based (RF-based) feature selection methods using different indicators (Gini Impurity, GI; Out of Bag Accuracy, OOBA) were proposed and tested separately. The LSMs of four models were carried out under the guidance results of FE, namely Classification and Regression Tree (CART), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine for Classification (SVC). For feature enhancement, standardization had significant advantages over normalization. All RF-based methods were proven effective, lifting the AUC by 0.01~0.02. The RF model achieved the highest LSM accuracies, respectively, 0.949 (landslide), 0.957, and 0.949 (unstable slopes), improved by 0.008 (landslide), 0.005 (collapse), and 0.013 (unstable slopes). This proved that the FE helped to improve LSM and can help to decide the dominant conditioning factors for regional geohazards.

Keywords: landslide; feature engineering; landslide susceptibility mapping; random forest algorithm

Citation: Ling, X.; Zhu, Y.; Ming, D.; Chen, Y.; Zhang, L.; Du, T. Feature Engineering of Geohazard Susceptibility Analysis Based on the Random Forest Algorithm: Taking Tianshui City, Gansu Province, as an Example. *Remote Sens.* **2022**, *14*, 5658. https://doi.org/10.3390/rs14225658

Academic Editor: Andrea Ciampalini

Received: 1 September 2022
Accepted: 7 November 2022
Published: 9 November 2022

1. Introduction

Landslide is a natural phenomenon that includes mass down-slide movements of rocks, soil, or debris flows under gravity [1], and it has become the most common geohazard due to city expansion and climate change. Landslides can cause massive casualties and economic losses. In the last century, landslides caused over 16,000 casualties [2]. From 2004 to 2010, the Durham Fatal Landslide Database (DFLD) recorded 2620 non-seismic landslides and 32,322 related deaths [3]. China has long suffered from landslides and correlated hazards [4,5]. Hence, research on landslides and other geohazards is of great academic and social importance.

Landslide susceptibility represents the likelihood of landslide occurrence under a certain combination of geo-conditions [6]. In contrast, landslide susceptibility mapping (LSM) refers to its visualization through techniques such as geographical information science (GIS) and remote sensing (RS) [7]. LSM is the basis of geohazards prevention and mitigation and has always been a hot-spot research topic. In the quantitative LSM process, landslides are assumed to be the coupling results of multiple conditioning factors (also discussed as conditioning features when using ML models). The most frequently used methods are formula-based statistical methods and big-data-driven machine learning (ML) methods.

The statistical methods involve a series of calculations, weighting, assignment, and buffering that estimate the contribution of each factor to the hazard by the mathematical statistics and corresponding formulas, after which the relative rank of landslide susceptibility would be derived [8,9]. Such methods have broad applications in large-scale rough LSM with acceptable accuracy and interpretability. Statistical methods mainly include the binary discriminant statistics method [8], the multivariate statistical method [10], the Weighted-Certainty Factor method (WCF) [11], the Maximum Entropy method (MaxEnt) [12,13], and the Information Value model (IV) [14,15]. In recent years, with the rapid development of computer technology, various ML algorithms and deep learning (DL) algorithms have been adapted to LSM, making the process more automated and intelligent. In ML-based methods, samples and prior knowledge are used to pre-train the model, and the results are generated through complex non-linear models [16]. The most popular methods have been the Logistic Regression model (LR) [17], the Naïve Bayes model (NB) [18], the Decision Tree model (DT) and its boosting models [19], the Random Forest model (RF) [20,21], Rotation Forest model (ROF) [22,23], the Support Vector Machine model (SVM) [24,25], the Artificial Neural Network (ANN) [26], and the Convolutional Neural Network (CNN) [27–29]. Some researchers compared the differences between ML models [30–34], and some combined statistical criteria with it to further improve the model performance.

These LSM works focused chiefly on model tests and modification as researchers have attempted to make the best performance of different models. However, the commonly used conditioning factors are mainly selected by experience, and whether selection and combination would affect the model performance has not been fully discussed. The DT and its related models (such as CART, RF, ROF, and boosting DT) use criteria to calculate the input factors' weight in the model constructing process, thus estimating each factor's importance in LSM. Youssef et al. [33] compared the RF, boosted regression tree (BRT), classification and regression tree (CART), and general linear (GLM). The results showed that for the four models, aspect, altitude, and distance to faults have been of the highest importance. Hong et al. [35] joined the RF with three bivariate statistical models, namely Evidential Belief Function (EBF), Frequency Ratio (FR), and Multivariate Logistic Regression (MLR). The RF was used for feature importance calculation, and the results showed that distance to rivers, distance to roads, and slope gradient had been the dominant factors of landslides in the Lianhua area. Pham et al. [36] proposed a Random Subspace (RSS) and Classification And Regression Trees (CART) hybrid approach. They discovered that the rainfall, distance to road, and slope gradient had contributed the most to regional landslides. Chen et al. [37] used Gradient Boosting DT (GBDT), RF, and IV models to perform LSM and estimate each factor's importance in the LSM process in the Three Gorges Reservoir region. The results have shown that elevation was the dominant factor, followed by distance to rivers, vegetation, slope, etc. Cheng et al. [38] used RF to estimate 36 conditioning factors through the Gini index. The final LSM model used factors with a Gini value higher than 0.1 They found that the land cover, groundwater volume, and distance to rivers related the most to geohazards in this region. Zhou et al. [39] proposed two hybrid models by applying Geo-Detector and the Recursive Feature Elimination method (RFE) to RF to perform feature optimization and LSM. The results showed that after feature optimization, the AUC arose for both methods, proving the importance of feature optimization. These studies have shown that the DT-based models, especially RF, could perform feature selection, which is the most essential to the feature engineering (FE) process.

FE is a method that involves a series of data processions that transform raw data into training data that best suit the model [36]. The FE helps to select the optimal solution by the algorithm and the subset of features that best represent the dataset. This concept has been applied to many computing engineering fields yet has seldom been mentioned in the LSM. The FE helps to estimate the most suitable data pre-procession method and the best-matched conditioning feature dataset, thus making the ML-based LSM more convincing and explainable. Moreover, it helps to analyze the dominant conditioning factors of certain kinds of geohazards and compare the in between differences in between. In recent years,

some researchers have already used FE or feature selection to help analyze the LSM result, yet they mainly focused on feature importance ranking [21,32,33,35]. Sun et al. [40] used a simple feature importance ranking method, while Zhou et al. [35] used an iteration method; both authors retrained the ML model accordingly and concluded that the feature selection would improve the accuracy. However, whether the iteration method is superior has not been fully discussed.

The most suitable data pre-procession method and feature selection method type, and whether eliminating the unimportant features would improve the result, have not been fully discussed. In detail, there are three aims listed:

(1) Explore the most suitable data-preprocessing principles.
(2) Determine whether using the elected features to retrain the model would improve the accuracy.
(3) Compare the simple ranking feature selection idea and the iteration feature election idea.

Therefore, this paper presents a relatively comprehensive FE-guided LSM by setting experiments upon all the steps in the FE to explore the most suitable combination, namely feature extraction, feature enhancement, and feature selection.

2. Materials and Methods

Tianshui city was selected as the study area. The detailed workflow is shown in Figure 1. First, 23 conditioning features were generated, and 4 RF-based FE methods were proposed and tested. The LSM is performed with 4 ML models: the CART, RF, SVC (SVM for classification), and LR. The results were evaluated through the Receiver Operating Characteristic curve (ROC curve) and the Area Under Curve (AUC).

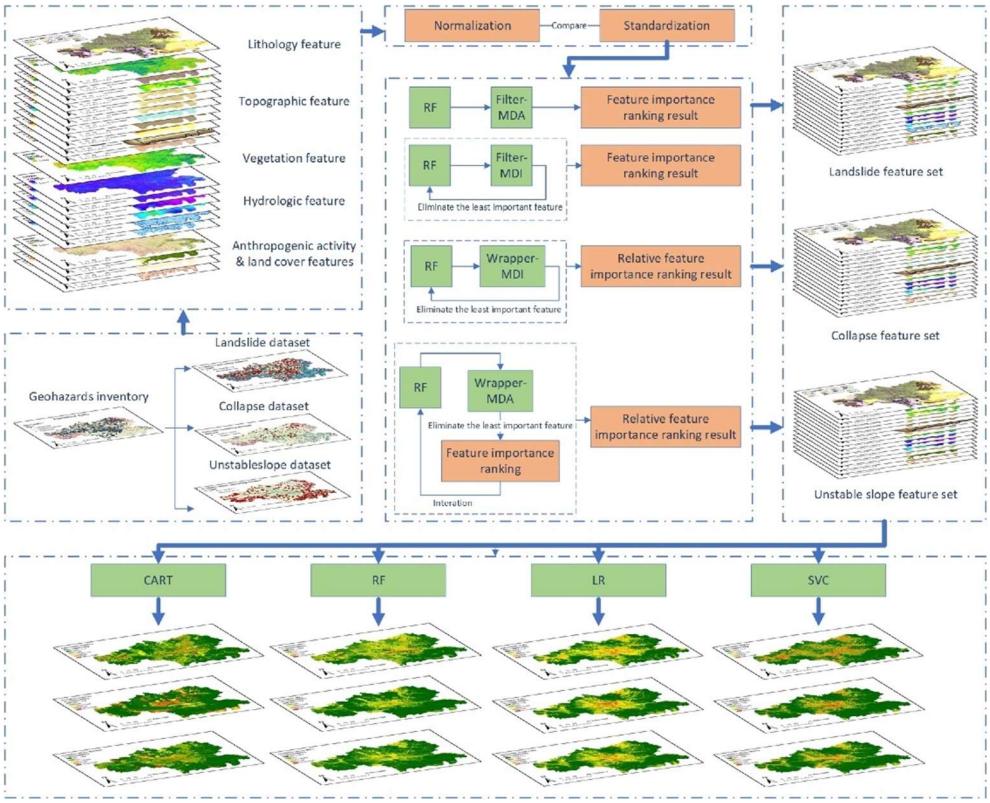


Figure 1. Workflow of this study.

2.1. Geological Conditions of the Study Area

Tianshui city, situated in the Gansu Province of western China, was selected as the study area. It is one of the prefecture-level cities within the Gansu province; both urbaniza-

tion and city expansion have been growing rapidly in recent years. Hence, the studies of local geohazards are important for economic development and urban safety.

The study area covered 16.4 km^2 in total, 104°35′1″~106°42′24″E (longitude), 34°4′57″~35°10′18″N (latitude). Located at the eastern end of the Qilian orogenic belt, the transition zone between the Guanzhong Plain and the Loess Plateau, the terrain is high in the west and relatively low in the east [41], ranging from 736~3118 m. Fractures are distributed widely, among which the most typical ones are the West Qinling North Rim Fault [42], the Tongwei Fault, and the Qingshui Fault. As a result, tectonic activities take place frequently. Tianshui city belongs to the semi-humid and semi-arid continental monsoon climate zone [41], and the precipitation is strongly affected by seasons and terrain, ranging from 500 to 600 mm. The annual temperature is from roughly 8 to 19 °C. The Wei River, the largest tributary of the Great Yellow River, traverses the entire study area from west to east. The basin has been covered with Quaternary deposits with loess covering 10 to 30 m at the top of the strata, which is unstable [42]. Lithology is gneiss (Proterozoic), rocks (Triassic sedimentary and metamorphic), and sandstone (Devonian and Cretaceous) [42].

2.2. Landslide Inventories

Where geohazards have occurred tend to have more significant potential to breed new ones; hence, landslide inventories are required in LSM [1]. This study acquired the landslide inventory by the geological field survey of hazards [42]. It contained 968 landslides, 183 collapses, and 243 unstable slopes. All the spatial attributes of the records were verified manually. The accurate occurrence times were not logged due to the delayed field investigations and historical hazard records that took place decades ago. In Tianshui city, the most frequently occurring geohazards can be divided into three types: landslide, collapse, and unstable slope. The landslides often move along the horizontal direction, while collapses move along the tangential direction at a much higher speed. An unstable slope refers to a slope that is prone to sliding. It may not have slid or collapsed but obtains a high possibility of forming into a landslide or collapse.

Geological investigation results show that mudstone bedrock and overlying loess strata have been strongly affected by historical earthquakes in the study area, causing densely distributed geohazards [42]. Landslides aggregately distribute in river erosion zones and valleys. The overlying lithology grouping is mostly very soft, while the depositions are mainly loess, clayey, and debris, indicating that the landslides are typical mounded (earthy). Collapses mainly distribute along faults and are mostly caused by earthen slope slumps, metamorphic rock avalanches, and collapses.

The inventory contained 2027 records in total. Since the 0~10° slope covered area is considered gentle, the records were more likely to be deposits or misclassified loess; thus, this part of records was eliminated. Afterward, the remaining were 968 landslides, 183 collapses, and 243 unstable slopes. The detailed distributions are shown in Figure 2.

Table 1 shows the raw data resources of the conditioning feature extraction, while the detailed description is in Section 2.3.1. The geodetic reference system is unified to the WGS84 coordinate system using a Universal Transverse Mercator (UTM) projection 48 N band. Meanwhile, the spatial resolution is unified to 30 m. The raw data have a spatial resolution of 0.05 degrees for groundwater volume and precipitation. A 5-year mean value of each feature was calculated to reduce the impact of resampling errors. The platforms pre-processed the data obtained with Google Earth Engine, and codes completed the calculation.

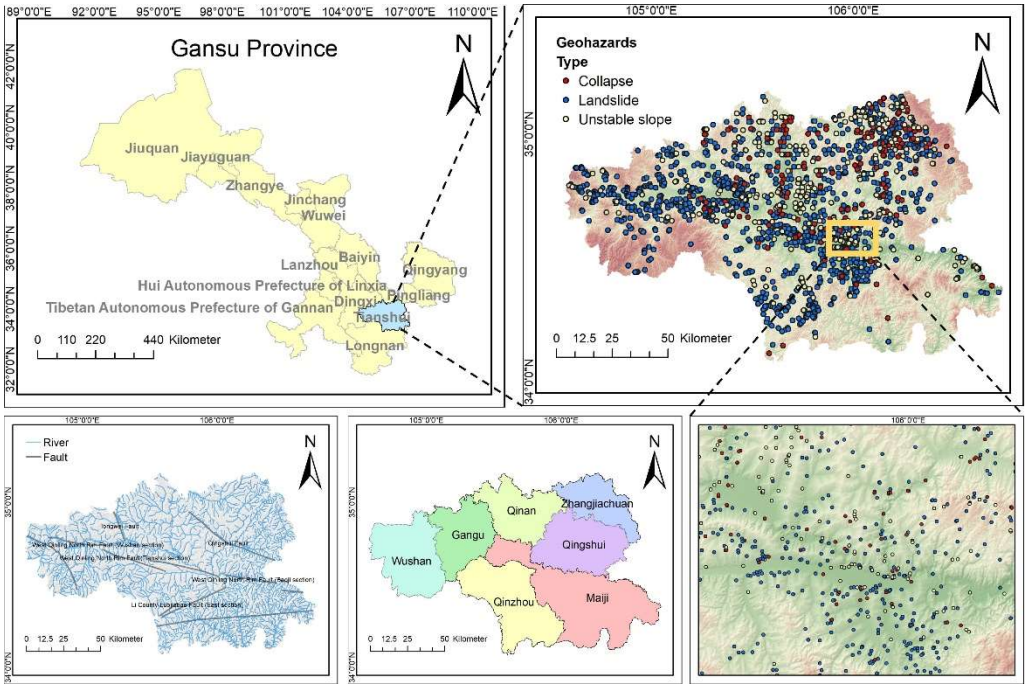


Figure 2. Location and geohazard distributions of Tianshui city.

Table 1. Raw data resources for feature extractions and geohazard inventories, the obtaining resources (platforms or websites) are listed as well.

Data	Raw Data Resource	Obtaining Resources
DEM	ASTER GEM (30 m)	Geospatial Data Cloud (http://www.gscloud.cn/, accessed on 28 October 2020)
Vegetation Building	Landsat 8-OLT images (2012~2017)	Google Earth Engine (https://explorer.earthengine.google.com/)
Precipitation	CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) data (2012~2017)	
Groundwater volume	GRACE (Gravity Recovery and Climate Experiment) data (2012~2017)	
Roads, rivers, faults, and boundary	Geographical vectors	Geological surveys
Geological map	Digital geological map (1:250,000, public ver.)	
Land cover	Global 30 m land cover classification products by the Chinese Academy of Sciences (2020)	http://data.casearth.cn/, accessed on 12 January 2021
Landslide inventories	Ground survey sheet	Geological field surveys

2.3. Feature Engineering

The FE process has 4 parts: feature extraction, feature enhancement, feature selection, and evaluation. As feature extraction and feature enhancement should be completed before model training, the two processes could be joined and described as feature pre-processing. The evaluation was merged into the LSM validation process and will not be discussed separately.

2.3.1. Feature Extraction

In quantitative analysis, LSM is a coupling result of multi-variables (conditioning features). However, the features used for LSM vary from region, study area scale, geohazard type, and evaluation model [39,43]. To generate a comprehensive analysis, the selected features should cover all formation aspects (topography, lithology, hydrogeology, vegetation, anthropogenic activity, and land cover). Finally, a total of 23 features were constructed. Except for lithology, aspect, distance to faults, distance to rivers, distance to

roads, and land cover, other features have not been classified and reassigned, the values of which are continuous. Table 2 is an aggregation of the classified features' properties.

Table 2. The properties of features that have been classified.

Feature Name	Types	Classification Standard
Lithology	15	1. 1-I; 2. 1-II; 3. 1-III; 4. 2-I; 5. 2-II; 6. 2-III; 7. 3-I; 8. 3-II; 9. 3-III; 10. 4-I; 11. 4-II; 12. 4-III; 13. 5-I; 14. 5-II; 15. 5-III.
Aspect	9	1. 0~22.5 and 337.5~360°; 2. 22.5~67.5°; 3. 67.5~112.5°; 4. 112.5~157.5°; 5. 157.5~202.5°; 6. 202.5~ 247.5°.
Distance to faults	11	1. <2 km; 2. 2~4; 3. 4~6; 4. 6~8; 5. 8~10; 6. 10~12; 7. 12~14; 8. 14~16; 9. 16~18; 10. 18~20; 11. >20 km.
Distance to rivers	11	1. <0.2; 2. 0.2~0.4; 3. 0.4~0.6; 4. 0.6~0.8; 5. 0.8~1; 6. 1~1.2; 7. 1.2~1.4; 8. 1.4~1.6; 9. 1.6~1.8; 10. 1.8~2; 11. >2 km.
Distance to roads	11	1. <0.2; 2. 0.2~0.4; 3. 0.4~0.6; 4. 0.6~0.8; 5. 0.8~1; 6. 1~1.2; 7. 1.2~1.4; 8. 1.4~1.6; 9. 1.6~1.8 km.
Landcover	8	1. farmland; 2. forest land; 3. grassland; 4. shrubs; 5. wetlands.

Based on the previous work of authors [44], we have discovered that the factor classification principles have no specific impact on model training or accuracies, only the distinctness in the LSM maps. This paper adopted the equal-interval classification method by considering the geographical significance of distance decay. Each class in the reassigned results of distances to faults, rivers, and roads represents a natural range. Within each interval, the geological influence can be equally seen. The total distances to faults, roads, and rivers were determined after a comprehensive analysis of the study area scale and its natural impacts. As the influence declined with the distance, the contribution to LSM out of the maximum distance could be neglected. Therefore, they were all reassigned to "11".

1. Lithology feature:

Different lithology types would lead to differences in slope stability and chemical properties. The study area is located at the eastern end of the Qilian orogenic belt. The lithology mainly consists of Precambrian metamorphic rocks with a small amount of magmatic rock [42]. Since the lithology feature in the study was mixed and complex, this paper grouped it referring to hardness and reassigned values of 1~15. The approximate lithology types included under each grouping are shown in Table 3.

2. Topographic features:

This paper selected elevation, slope, aspect, curvature, plan curvature, profile curvature, Topographical Roughness Index (TRI), Topographical Wetness Index (TWI), distance to faults, fault density, and cumulative solar radiation as topographic features.

Features such as elevation, slope gradient, aspect, and cumulative solar radiation contribute to landslides indirectly, as they influence the catchment area and vegetation coverage, thus affecting landslide susceptibility. Areas with a higher slope gradient tend to be more unstable, as the gravitational potential energy gradually transforms into kinetic energy along with the increasing slope gradient [33]. Curvature (the second-order derivative of slope gradient), profile curvature (the curvature in the maximum slope direction), plan curvature (the curvature perpendicular to the maximum slope direction) [19] were all calculated. The Topographic Wetness Index (TWI) and the Topographical Roughness Index (TRI) are micro-geomorphic indices. TWI is a quantitative simulation of soil's dry and wet conditions in a watershed, while TRI describes regional topographic changes [45].

Table 3. Detailed lithology types in the study area. The lithology types are divided into 5 primary categories labeled 1 to 5. Category 1 is marked as the Very Hard (VH) rock group while category 5 is the Very Soft (VS) rock group, each group is further divided with 3 sub-categories labeled I–III with the same sorting standard.

Primary Category	Secondary Category	Representative Lithology Types	Strata
1	I	Diorite and granite	Hercynian, Caledonian, Himalayan, Upper Paleozoic, Mesozoic
	II	Acidic volcanic rocks, quartzite, dacite, phyllite	Caledonian, Sinian, Lower Paleozoic
	III	Quartz sandstone, pebbled sandstone, siltstone	Devonian, Permian
2	I	Schist, gneiss, mixed rock with volcanic rock	Caledonian, Pre-Sinian, Upper Paleozoic
	II	Pegmatite, syenite, volcanic metamorphic rock, purple and purple-red rhyolite porphyry	Hercynian, Caledonian, Himalayan, Sinian
	III	Limestone, gray-green slate	Permian, Upper Paleozoic
3	I	Argillaceous purple-red siltstone, mudstone, sandy shale, gray-green SLATE shale, pebbly sandstone	Devonian, Cenozoic, Upper Paleozoic
	II	Conglomerate, glutenite, siltstone, sandy mudstone	Cretaceous, Tertiary, Triassic, Permian
	III	Biotite calcarenite schist, bimica schist, hornblende schist, chlorite Muscovite schist	Lower Paleozoic
4	I	Melaleite, metamorphosed siltstone, metamorphosed fine sandstone	Sinian, Devonian, Permian, Carboniferous
	II	Shale, siltstone, sandstone, sandy limestone, shell limestone, oolitic limestone	Cretaceous, Permian, Triassic, Upper Paleozoic
	III	Conglomerate, sandy conglomerate, clay rock with calcareous nodules, purplish-red sandy mudstone with sandstone	Tertiary, Triassic
5	I	Red, purplish-red clay with gray matter nodules, red sandstone, conglomerate, conglomerate	Tertiary
	II	Alluvial secondary loess, silty loess, gravel	Quaternary
	III	Riverbed alluvial gravel, sand, silt, boulders, sub-sandy soil, secondary alluvial loess and loam	Quaternary, modern

TWI can be calculated according to Equation (1),

$$TWI = \ln\left(\frac{A_s}{tan\beta}\right) \quad (1)$$

where A_s refers to the catchment area that can be calculated by flow accumulation, and β refers to slope gradient.

TRI can be calculated according to Equation (2),

$$TRI = H_{max} - H_{min} \quad (2)$$

where H refers to the altitude of the calculation unit, and H_{max} is the highest in the region, while H_{min} is the lowest.

These above features were generated using DEM with continuous values, except for the aspect feature, which was divided into 9 directions and was further reassigned accordingly to 1~9 [27] at 45° intervals. Table 2 shows the detailed sorting results.

Distance to faults and fault density factor reflect the influence of tectonic fractures, which reduce the strength of the rock mass, causing geological activities. These two features were generated by the faults vector. For distance to fault, a total of 10 ring buffers were generated at an interval of 2 km; for the fault density, the analysis radius was set to 5 km.

3. Vegetation feature:

In this paper, the Normalized Differences Vegetation Index ($NDVI$) is used as a vegetation feature [32]. Landsat 8-OLT images are applied to calculate a 5-year mean NDVI value (2012~2017), the calculation is shown in Equation (3):

$$NDVI = \frac{Band_{NIR} - Band_{Red}}{Band_{NIR} + Band_{Red}} \quad (3)$$

for Landsat 8-OLT images, the $Band_{Red}$ is the 4th band, while the $Band_{NIR}$ is the 5th band.

4. Hydrologic features:

The erosive force directly affects the slope foot and the river incision, while precipitation and groundwater jointly affect the infiltration, thereby affecting the stability of the slope. Precipitation and groundwater can reduce the shear strength and change the lithology composition through chemical interaction [46]. In this paper, the selected hydrologic features were precipitation [47], groundwater volume, Normalized Difference Water Index (NDWI), Normalized Difference Water Index (MNDWI), distance to rivers [32], and river density.

Five-year mean (2012~2017) precipitation and groundwater volume features were generated. Both $NDWI$ and $MNDWI$ [48] can be representative of waterbody distribution. To test whether either feature would contribute more to landslides, both features at the 5-year mean (2012~2017) value were generated. The calculations are shown in Equations (4) and (5)

$$NDWI = \frac{Band_{Green} - Band_{NIR}}{Band_{Green} + Band_{NIR}} \quad (4)$$

$$MNDWI = \frac{Band_{Green} - Band_{SWIR1}}{Band_{Green} + Band_{SWIR1}} \quad (5)$$

For Landsat 8-OLT images, the $Band_{Green}$ is the 3rd band, the $Band_{NIR}$ is the 5th band, while the $Band_{SWIR1}$ is the 6th band.

Rivers have an important influence on landslides, especially seismic-induced geological hazard development [49]. In this study, distance to rivers and river density were generated to estimate the influence upon landslide of rivers. Considering that evapotranspiration would cause a decrease in the influence of rivers, for distance to rivers, a total of 10 ring buffers were generated at an interval of 200 m; for the river density, the analysis radius was set to 0.5 km.

5. Anthropogenic activity and land cover features:

Anthropogenic activities (unreasonable artificial slope cutting, soil and water damage due to road construction, mining deposits, water storage, and drainage purposes [1,22]) lead to a decrease in slope stability. This study used land cover, distance to roads, road density, and Normalized Difference Building Index (NDBI) [22,40,46].

The land cover feature used in this paper is the global 30 m land cover classification products that were released in 2020, provided by the Institute of Air and Space Information Innovation, Chinese Academy of Sciences (http://data.casearth.cn/, accessed on 12 January 2021). For distance to roads, a total of 10 ring buffers were generated at an interval of 200 m; for the road density, the analysis radius was set to 0.1 km.

The $NDBI$ is applied as a quantitative estimator of buildings in the study area. This paper calculated the mean $NDBI$ of 2012~2017, while the calculation formula is shown in Equation (6).

$$NDBI = \frac{Band_{SWIR2} - Band_{NIR}}{Band_{SWIR2} + Band_{NIR}} \quad (6)$$

For Landsat 8-OLT images, the $Band_{NIR}$ is the 5th band, while the $Band_{SWIR2}$ is the 7th band.

2.3.2. Feature Enhancement

The commonly used feature enhancement methodology is dimensionless, i.e., data with different ranges or distributions converted into a uniform format. Linear methods are commonly used, such as valorization, centralization, normalization (Min–Max scaling), standardization (Z-score scaling), weighting, log function conversion [4], and inverse tangent function conversion. For ML, normalization and standardization are the most used methods. In this study, both methods were tested.

1. Normalization (Min–Max scaling):

The normalization [39,40] (Min–Max scaling) is the extreme difference scaling. The process is to scale the data between [0, 1], with the formula shown as Equation (7). For each column, x_{min} is the minimum value in the dataset, while x_{max} is the maximum.

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{7}$$

2. Standardization:

Data standardization refers to scaling the data distribution to a normal distribution with 0 as the mean and 1 as the standard deviation (i.e., the standard normal). The formula is shown in Equation (8). For each column, μ is the mean value of the data, while σ is the standard deviation.

$$x^* = \frac{x - \mu}{\sigma} \tag{8}$$

2.3.3. Feature Selection

Generally, the feature selection methods can be sorted into filter, wrapper, and embedded methods. The filter methods use statistical metric calculation results (such as variance, correlation coefficient, chi-square index, maximum information, etc.) by setting a threshold to eliminate features with lower importance and more noise [50–53]. While the wrapper methods are also known as recursive elimination feature selection methods, in each round, features that do not meet the evaluation threshold are eliminated, and the remaining features are used as input for the next round of training [54,55]. With the development of ML models, the embedded methods have been popular. Methods are usually algorithmically built in, such as the feature importance ranking for DT-based models and penalty term ranking for LR models.

This study selected the RF, the integrated algorithm of the DT, as the basis of the feature selection method. By combining 2 indictors, a total of 4 feature selection methods were built.

1. Random Forest algorithm:

The RF is a typical bagging ensemble ML algorithm [56]. An RF is constructed using several decision trees, and each tree obtains its classification result using individual classification. The modeling operates in parallel while the output of the whole forest is obtained by voting on all judgment results. Figure 3 shows the workflow of the RF-based classification algorithm using random bootstrap sample selection.

The basic process of RF-based classification is as follows:

(1) Construct the original training sample dataset for DTs, the number of cases is N while the number of input variables is M.
(2) Generate sub-training datasets by sampling with the replacement bootstrap method for n times, meaning that the generated RF has n trees in total.
(3) To select the features for each non-leaf node (internal node), the model first randomly selects a certain number of features from all features and uses them as split features and then selects the best-performing one for node splits.
(4) The classifier output is determined by a majority vote of by each tree in the RF.

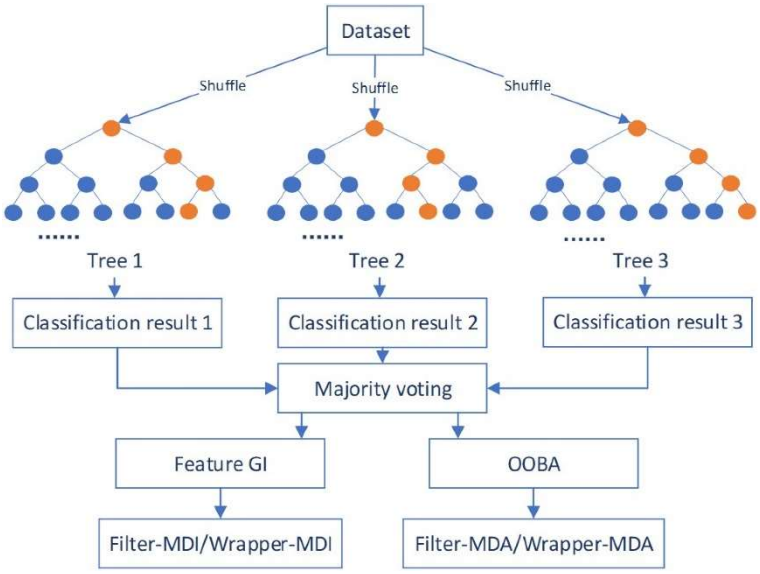


Figure 3. RF-based feature selection frame; for the FE, the classification results were not used.

Gini Impurity (GI) and Entropy are both commonly used in the RF algorithm. After repeated training processes, the model scores and accuracies using GI were always higher than using Entropy; therefore, this paper adopted GI as the indicator. The calculation is shown in Equation (9), where $I_{Gini}(f)$ represent the GI of each feature f, m is the total number of samples, and f_i is the probability each sample's occurrence.

$$I_{Gini}(f) = 1 - \sum_{i=1}^{m} f_i^2 \tag{9}$$

GI should be at a maximum when a node is equally divided among all classes, which means that the split uses the least helpful information. The split is kept until the terminal nodes have a few cases or are all pure. Therefore, the RF forms according to Equation (10), and the multiple independent DT classifiers can be written as $\{h(X, \theta_k), n = 1, 2, \ldots N\}$.

$$H(x) = arg_z^{max} \sum_{i=1}^{N} h_i(x) \tag{10}$$

where N represents the number of DTs, and $h_i(x)$ is the classifier result of DT_i.

2. Feature selection methods:

While the RF is applied as the base feature selection model, this paper constructed the filter-embedded method and the wrapper-embedded method accordingly. RF uses 2 indicators to measure feature importance: GI and Out of Bag Accuracy (OOBA). The method using GI is sorted as the Mean Decrease Impurity method (MDI), while the method using OOBA is the Mean Decrease Accuracy (MDA) method. By combining the MDI and MDA with filter and wrapper thoughts, a total of 4 methods were proposed: filter-MDI, filter-MDA, wrapper-MDI, and wrapper-MDA.

In the RF training process, after n sampling rounds, a subset of n samples equal in size to the original training set is obtained, and the possibility p of each sample being selected is calculated after Equation (11). Since the sampling process is put back, while n is large enough, p would converge to $1 - \left(\frac{1}{e}\right)$. Therefore, approximately 37% of the training data would never be involved in the modeling; these data are called the Out of Bag Data (OOB), and OOBA can also be used as an evaluating criterion for model accuracy [57].

$$p = 1 - \left(1 - \frac{1}{n}\right)^n \tag{11}$$

In both the MDI and MDA processes, an RF model is first built, and the criterion is calculated accordingly. For MDI, the features are ranked referring to GI, while for MDA, the features are ranked referring to the change of OOBA. By replacing each feature with noise data and calculating the variance, the change of OOBA is measured. More significant variance indicates the higher feature importance.

For the filter-MDI, the feature selection is performed by setting a threshold of feature remanence. It took only one round of training with the lowest time complexity. The threshold is set at 17, which refers to 70% of total features, equally in both filter-MDI and filter-MDA. The wrapper-MDI recursively performed the feature selection. In each round, the RF model is re-trained, the feature of the lowest GI is eliminated, and the remaining features forms into new subsets for model training. To determine how many features remained that would make out the highest LSM accuracy, we drew an AUC-changing curve that ranked the relative feature importance by the order in which features are eliminated. The same is the wrapper-MDA, while the metrics switched from GI to OOBA. The wrapper-MDA had the highest time complexity for its two-tier iteration.

2.4. Landslide Susceptibility Modeling

To explore the FE's effect on different ML models except for RF, another 3 ML models were used, namely LR, SVC (SVM for classification), and CART. Python 3.7 was used for modeling; all ML models (including RF) were built according to the scikit-learn module. These models are all the most-used models in binary classification. The CART is compared to the integrated method (the RF), while the LR and SVM are selected to compare whether the tree-based model or functional model performs better in LSM.

2.4.1. Logistic Regression Model

The LR is a generalized linear regression analysis model commonly used for dichotomous classification. This method has the advantages of simplicity, parallelizability, and strong interpretability [31,58,59]. This paper chose the "L2" penalty to avoid overfit. "C" (the hyperparameter controlling regularization degree) was set as 0.04. The "lib linear" solver was set accordingly. As the LSM was a binary classification process, the "multi_class" was set as "ovr".

2.4.2. Classification and Regression Tree Model

The CART model is one typical model using DT as the base estimator. It is a non-parametric supervised ML method that generates tree-formed decision rules to complete the tasks. It consists of a Root Node, a series of Internal Nodes, and Leaf Nodes; each Internal Node represents an attribute judgment, each branch represents a judgment result output, and each Leaf Node represents a classification result [32,36]. In the CART modeling procession, the hyperparameters were set as 15 (max_depth), "gini" (criterion), 30 (min_samples_leaf), and 15 (min_sample_split).

2.4.3. Support Vector Machine (for Classification)

The SVC model is a supervised generalized classifier for binary classification. It is a nonlinear learning algorithm developed from pairwise theory using certain kernel functions to calculate a margin hyperplane that can maximize the heterogeneity between samples [24,59]. In the SVC modeling procession, "rbf", representing the Gauss kernel function, was selected. The Gamma hyperparameter was set as 0.07.

2.5. Validation

In binary classification processes, samples can be divided into positive or negative samples [6,60]. The results are sorted into 4: positive samples that are predicted to be positive (true positive, TP); negative samples that re-predicted to be positive (false positive, FP); negative samples that are predicted to be negative (true negative, TN); positive samples that are predicted to be negative (false negative, FN). The ROC curve is drawn with the

True Positive Rate (TPR) on the Y-axis and the False Positive Rate (FPR) on the X-axis. The closer the curve is to the upper left corner, the more accurate the model is. TPR and FPR can be calculated according to Equations (12) and (13).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{P} = \frac{FP}{FP + TN} \tag{13}$$

where P is the number of positive samples in the original dataset,FN can be calculated by $P - TP$, while TN can be calculated by $P - FN$.

The AUC refers to the area under the ROC curve, with a range of [0.5, 1]. This paper evaluated the FE by comparing the AUC of LSM before and after the FE process. The ROC curve was plotted using python, while the AUC was also calculated.

3. Results

In this section, the detailed experimental procedure and the corresponding results are presented in the order of sampling construction, FE and LSM. As there are three types of geohazards in the study area, the experiments and results were constructed separately.

3.1. Sampling Dataset Preparations and Feature Extraction

As the quantitative LSM can be regarded as a binary classification process, landslide inventories were used to construct the sampling dataset, which contained positive samples (landslide points) and negative samples (non-landslide points). In this study, three geohazard sampling datasets were built separately. Considering the scale of Tianshui city, the landslide inventories are relatively small. Thus, it is necessary to enhance the sampling dataset size to avoid overfitting. The differences between positive and negative sample numbers should not be too large, while the total should not be too small. For negative samples, the spatial distribution should be homogeneous. After a comprehensive analysis, the ratio of positive and negative samples was set to 1:3 (landslide sampling dataset), 1:4 (collapse sampling dataset), and 1:4 (unstable slope sampling set). The sample number within each dataset is shown in Table 4.

Table 4. Sample number within each sampling dataset.

Geohazards	Positive Samples	Negative Samples	Total
Landslide	968	2958	3926
Collapse	183	732	915
Unstable slope	243	972	1215

While Figure 4 shows the points' distribution of each dataset, to ensure the sample purity and thus improve the separability, the negative sample points were separated from the positive sample points in spatial location by at least 2 km.

Figure 5 shows the graphics for all feature layers after the pre-procession.

3.2. Results and Comparison of Feature Enhancement Methods

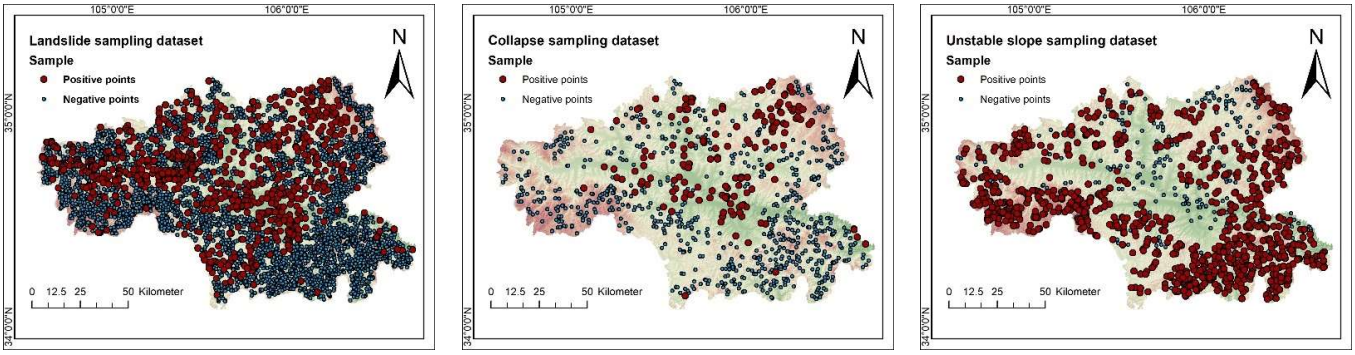
This paper compared the model score and AUC under different ML methods to decide whether the Min–Max scaling or the Z-score scaling is superior. The results are shown in Table 5.

Comparing the data in Table 5, it is evident that in most cases, the standardization-processed results are better than normalization-processed ones, the maximum improvement in model score was 12.54, while the maximum improvement in AUC is 0.095. Moreover, the improvement was more obvious for models with higher data requirements, such as SVC and LR. Thus, the following feature selection and LSM results were all based on data that used the standardization method.

3.3. Results and Comparison of Feature Selection Methods

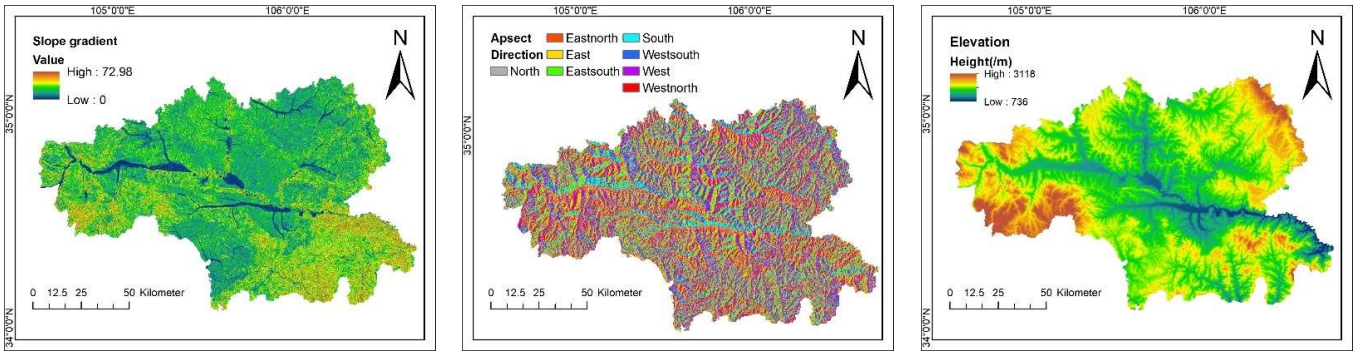
Four feature selection methods were separately performed, referring to Section 2.3.3. The multicollinearity test was first performed using the SPSS platform (ver. 2021). The multi-correlation between features is also an indicator of FE's effectiveness, for features with a high correlation may cause model instability. We compared the numbers of features with multi-correlation before and after the FE to add extra validation to the experiment. In multicollinearity analysis, the Variance Inflation Factor (VIF) has been a standard evaluation index. The pre-multicollinearity analysis suggested that high correlations existed among the condition factor dataset, with a high VIF up to over 2000; multi-correlations are specifically high among NDVI, NDWI, and MNDWI; as well as curvature, plane curvature, and profile curvature features.

The ratio of the training dataset to the testing dataset was set to 7:3. To control the variables, the other hyperparameters of RF were set to the same value, namely 600 (n_estimators), "gini" (criterion), 80 (random_state), and 20 (max_depth). Since the sample set in this study was small, the results converged under this parameter combination.



(**a**) (**b**) (**c**)

Figure 4. Graphics of sampling datasets, positive points represent the geohazard points: (**a**) Landslide; (**b**) Collapse; (**c**) Unstable slope.



(**a**) (**b**) (**c**)

Figure 5. *Cont.*

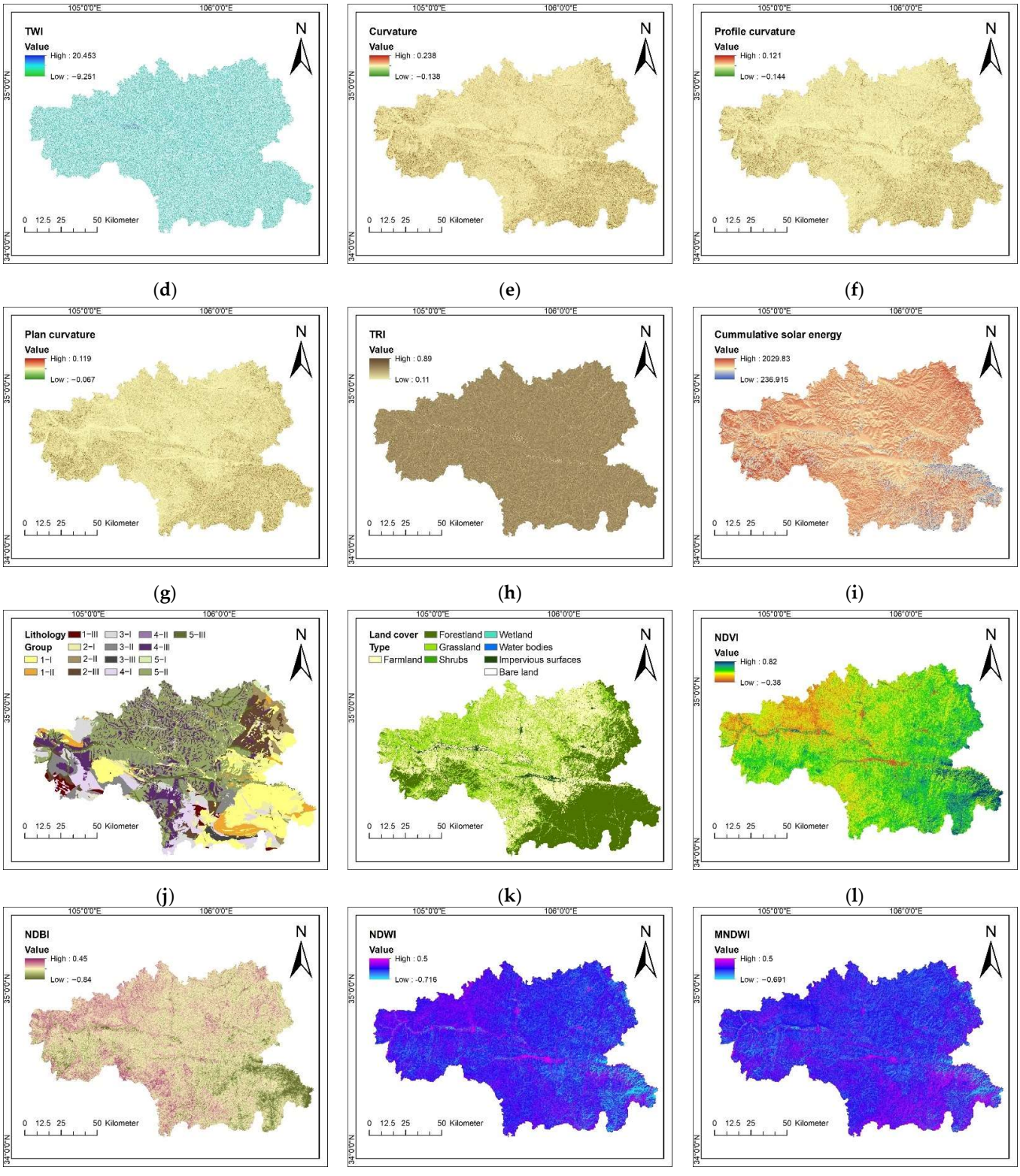


Figure 5. *Cont.*

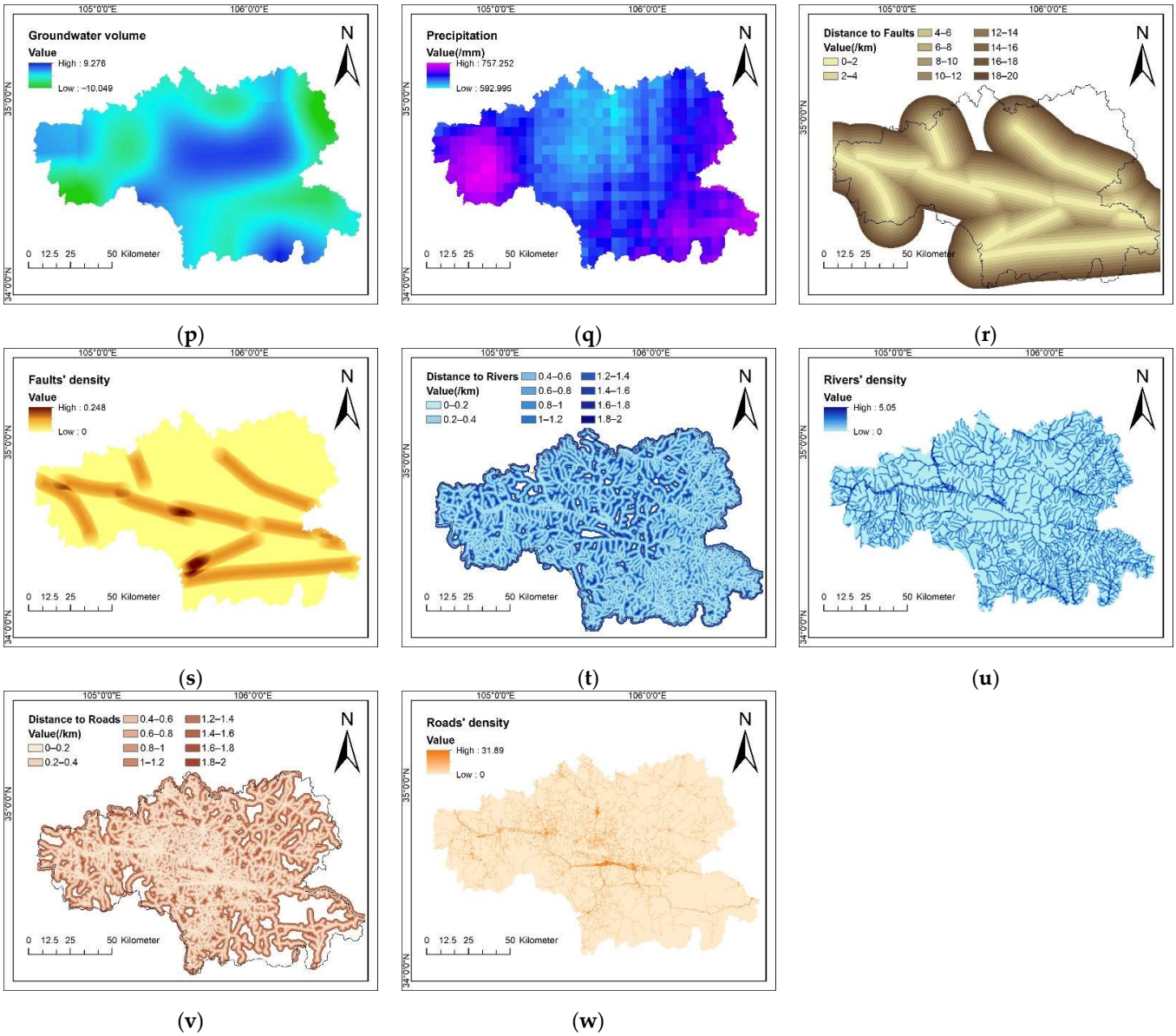


Figure 5. Graphics of the conditioning factors: (**a**) Slope; (**b**) Aspect; (**c**) Elevation; (**d**) TWI; (**e**) Curvature; (**f**) Profile curvature; (**g**) Plan curvature; (**h**) TRI; (**i**) Cumulative solar radiation; (**j**) Lithology; (**k**) Land cover; (**l**) NDVI; (**m**) NDBI; (**n**) NDWI; (**o**) MNDWI; (**p**) Ground water volume; (**q**) Precipitation; (**r**) Distance to faults; (**s**) Fault density; (**t**) Distance to rivers; (**u**) River density; (**v**) Distance to roads; (**w**) Road density.

For filter methods, the elimination threshold should first be settled. Commonly used methods are setting fixed thresholds, while for RF-based FE methods, the threshold could also be settled by drawing a learning curve of hyperparameters (max_features). However, the RF model had certain randomness that led to max_feature values varied from datasets, causing low stability in threshold selection based on the learning curve; thus, this paper adopted a fixed threshold instead. In most of the LSM processions, the number of factors ranges from 2 to 22, with an average of 9 [43]. This paper calculated the GI differences according to the feature ranking result as Equation (14) to decide the most suitable threshold that would bring the most minor information loss and the most representative of the feature dataset.

$$D_{GI} = GI_k - GI_{k-1} (k = 1, 2, 3, \ldots, n) \quad (14)$$

where GI_k is the related GI of feature k, and Figure 6 shows the result.

Table 5. Results of model scores and AUCs using normalization and standardization as feature enhancement method; standardization made for a better performance in model scores in all cases except for CART.

Geohazards	ML Model	Normalization		Standardization		Range	
		Model Score	AUC	Model Score	AUC	Model Score	AUC
Landslide	CART	79.17	0.827	81.24	0.845	2.07	0.018
	RF	85.97	0.927	87.52	0.932	1.55	0.005
	LR	76.08	0.794	78.31	0.826	2.24	0.032
	SVC	74.1	0.818	83.13	0.878	9.04	0.06
Collapse	CART	86.91	0.858	84.36	0.917	−2.55	0.059
	RF	89.82	0.946	91.27	0.97	1.45	0.025
	LR	78.91	0.816	84.36	0.911	5.45	0.095
	SVC	78.91	0.868	87.27	0.929	8.36	0.061
Unstable slope	CART	81.48	0.805	84.05	0.848	2.56	0.042
	RF	85.75	0.919	92.31	0.938	6.55	0.019
	LR	76.35	0.845	84.33	0.881	7.98	0.036
	SVC	75.5	0.862	88.03	0.886	12.54	0.023

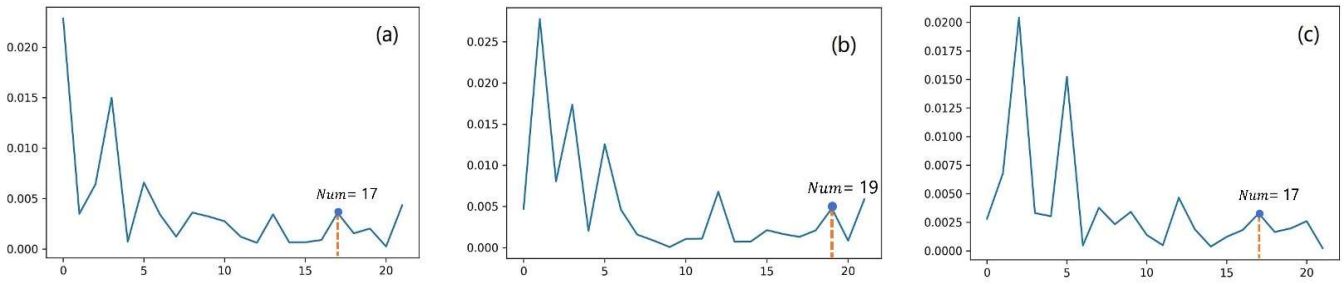


Figure 6. Difference of feature importance by ranking order. X-axis refers to the feature numbers, while the Y-axis is the D_{GI}: (**a**) Landslide; (**b**) Collapse; (**c**) Unstable slope.

After repeated experiments, the last peak always appeared in the range of 16~20 (feature numbers), indicating that when the remaining features are appropriately 70% (i.e., 17 features) remaining, the GI is relatively low and will not vary dramatically. Therefore, 70% (i.e., 17 features) was finally selected as the threshold.

3.3.1. Results of Filter-MDI

In the filter-MDI procession, the Gini impurity value for each feature was output through the interface feature_importance. Figure 7 shows the ranking results.

The changes in model score and AUC are shown in Table 6. Improvements can be observed after the feature selection, where the most significant improvement appears in the collapse dataset, up to 1.45 of model score and 0.018 of AUC.

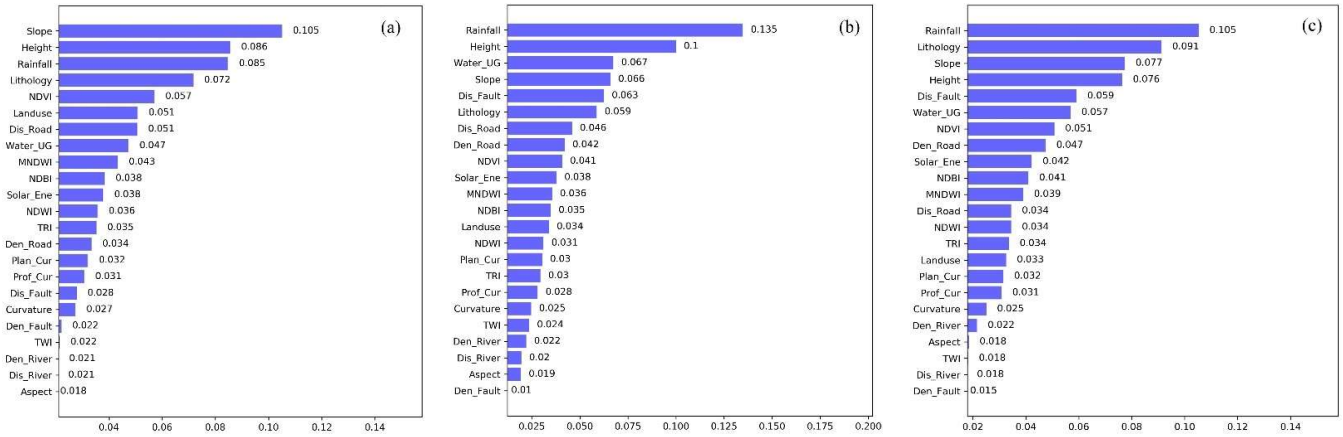


Figure 7. Feature importance ranking results using filter-MDI. The values are the average impurity of each feature: (**a**) Landslide; (**b**) Collapse; (**c**) Unstable slope.

Table 6. Changes in model score and AUC after using filter-MDI feature selection method.

Geohazards	Before Feature Selection		Before Feature Selection		Range	
	Model Score	AUC	Model Score	AUC	Model Score	AUC
Landslide	87.09	0.929	87.87	0.942	0.77	0.013
Collapse	88.36	0.951	89.82	0.969	1.45	0.018
Unstable slope	91.17	0.939	91.74	0.944	0.57	0.006

3.3.2. Results of Filter-MDA

For the filter-MDA procession, as the OOBA change is a relative value, its ranking indicated the relative feature importance as well. The results are shown in Figure 8. Ranking differed from the results processed by filter-MDA, such as NDBI, slope direction, lithology, site type, geologic lithology, distance to rivers, distance to faults, and NDVI. However, the features with the highest importance are still slope, elevation, and precipitation.

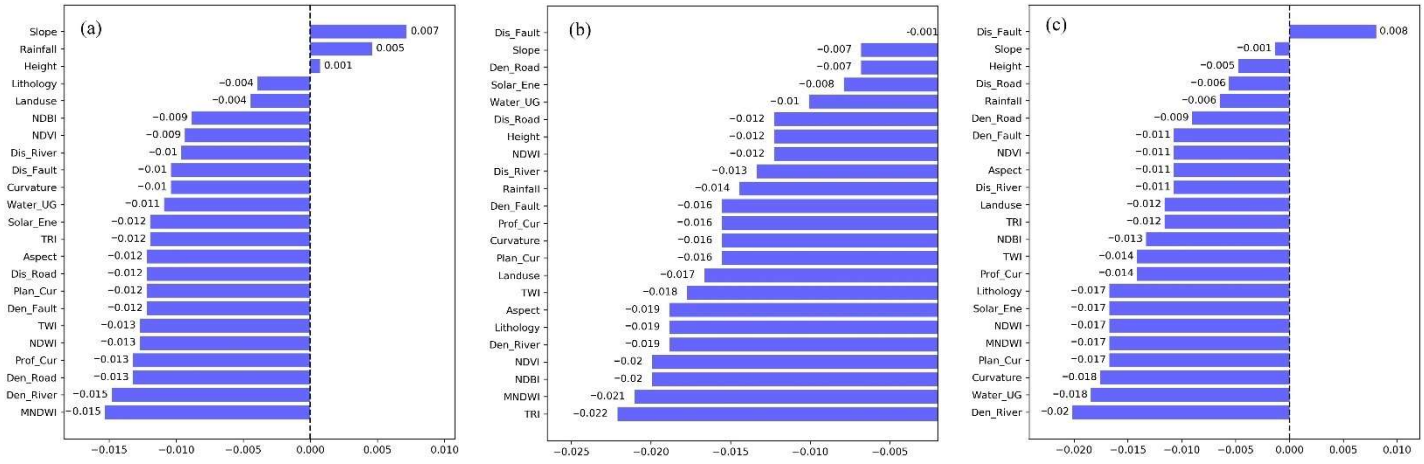


Figure 8. Feature importance ranking results using the filter-MDA feature selection method: (**a**) Landslide; (**b**) Collapse; (**c**) Unstable slope.

The changes in model score and AUC are shown in Table 7. The most significant improvement of model score appeared in the collapse dataset, up to 1.17, while the AUC improvements were even.

Table 7. Changes in model score and AUC after using filter-MDA feature selection method.

Geohazards	Before Feature Selection		Before Feature Selection		Range	
	Model Score	AUC	Model Score	AUC	Model Score	AUC
Landslide	86.64	0.936	87.69	0.948	1.05	0.012
Collapse	89.38	0.955	90.55	0.965	1.17	0.011
Unstable slope	89.01	0.935	89.17	0.947	0.16	0.012

3.3.3. Results of Wrapper-MDI

In the wrapper-MDI procession, the feature importance is ranked in the order of iterated elimination. By continuously eliminating features, the model was rebuilt, and AUC also changed. The number of remaining features corresponding with the highest AUC determined the feature selection threshold. Figure 9 shows the detailed AUC changing curve during the procession.

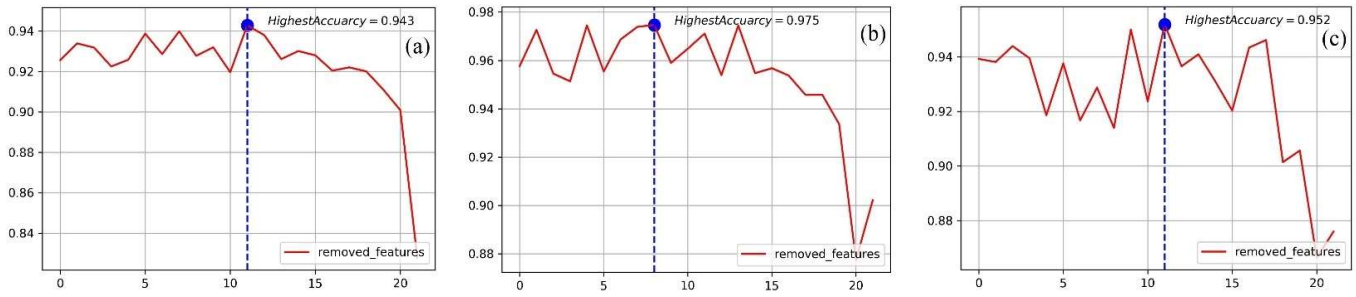


Figure 9. AUC changing curves of wrapper-MDI feature selection method. These results reflect the different numbers of remaining features in the dataset while the model obtained the highest AUCs. (**a**). Landslide; (**b**) Collapse; (**c**) Unstable slope.

To be noted, the optimal number of features is not a fixed value but varies with the dataset division. In the wrapper-MDI feature selection procession, the impurity is only a rejection indicator and would not be shown in the final ranking. The feature importance results are shown in Table 8.

Table 8. Relative ranking results of feature importance using the wrapper-MDI feature selection method.

Geohazards	Landslide	Collapse	Unstable Slope
Ranking	Precipitation	Precipitation	Precipitation
	Slope	Slope	Elevation
	Elevation	Elevation	Slope
	NDVI	Groundwater volume	Groundwater volume
	MNDWI	Distance to faults	Lithology
	Ground water volume	Lithology	NDVI
	Distance to roads	Cumulative solar radiation	Distance to faults
	Lithology	Distance to roads	Cumulative solar radiation
	Cumulative solar radiation	NDVI	MNDWI
	Land cover	MNDWI	Road density
	NDBI	NDBI	NDBI
	Plan curvature	Plan curvature	NDWI
	NDWI	Profile curvature	Profile curvature
	Profile curvature	Road density	TRI
	TRI	TRI	Plan curvature
	Distance to faults	NDWI	Land cover
	Road density	Land cover	Distance to roads
	Curvature	Curvature	River density
	River density	River density	Curvature
	TWI	TWI	Aspect
	Fault density	Aspect	Distance to rivers
	Distance to rivers	Distance to rivers	TWI
	Aspect	Fault density	Fault density

The changes in AUC are shown in Table 9. To be noted, due to frequent iterations, the results represent the improvement after feature selection; the highest AUC may not correspond to the best model performance.

Table 9. Changes in AUC after using the filter-MDA feature selection method. As the models were rebuilt in every iteration, the model scores are not logged.

Geohazards	Before Feature Selection	After Feature Selection	Range
Landslide	0.926	0.943	0.017
Collapse	0.958	0.975	0.016
Unstable slope	0.939	0.952	0.013

3.3.4. Results of Wrapper-MDA

Figure 10 shows the detailed AUC changing curve during the wrapper-MDA procession.

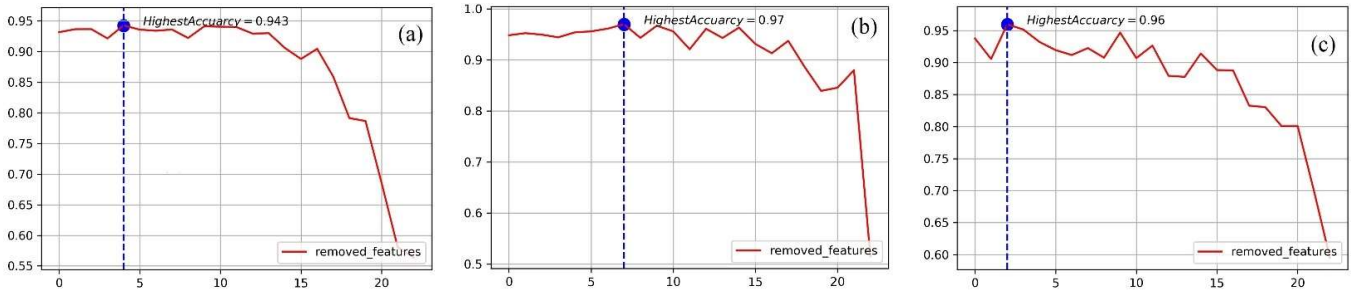


Figure 10. AUC changing curves of the wrapper-MDA feature selection method: (**a**) Landslide; (**b**) Collapse; (**c**) Unstable slope.

The ranking results are shown in Table 10.

Table 10. Relative ranking results of feature importance using the wrapper-MDA feature selection method.

Geohazards	Landslide	Collapse	Unstable Slope
Ranking	NDBI	TWI	NDWI
	River density	Rainfall	Lithology
	NDVI	Aspect	Distance to roads
	NDWI	NDWI	NDVI
	Distance to faults	Slope	TRI
	Rainfall	Distance to roads	Distance to faults
	Slope	Curvature	Rainfall
	Plan curvature	Elevation	NDBI
	Ground water volume	Distance to faults	Plan curvature
	Elevation	Lithology	Cumulative solar radiation
	Land cover	NDBI	Ground water volume
	Aspect	Profile curvature	Elevation
	Road density	Cumulative solar radiation	Land cover
	Distance to roads	Fault density	Slope
	Distance to rivers	NDVI	Profile curvature
	Curvature	Distance to rivers	Aspect
	Profile curvature	Land cover	Fault density
	TWI	MNDWI	Road density
	Lithology	TRI	Distance to rivers
	MNDWI	Ground water volume	TWI
	TRI	River density	River density
	Cumulative solar radiation	Road density	MNDWI
	Fault density	Plan curvature	Curvature

Table 11 shows the changes in AUC. Compared with the filter-indicator methods, the two wrapper-indicator methods achieved higher AUCs, and the wrapper-MDA performed better than the wrapper-MDI.

Table 11. Changes in AUC after using the filter-MDA feature selection method. The highest improvement appeared in the Collapse and Unstable slope datasets, up to 0.022.

Geohazards	Before Feature Selection	After Feature Selection	Range
Landslide	0.932	0.943	0.011
Collapse	0.948	0.970	0.022
Unstable slope	0.938	0.960	0.022

3.4. Landslide Susceptibility Mapping

By comparing the repeated experimental process and the results, this study finally chose to retain 70% of the total number of features (i.e., 16). Seven features were eliminated from each dataset. While deciding which feature to be eliminated, ones that ranked in the bottom 40% under all four feature selection methods were first chosen. If the total did not account for seven, then features ranked in the bottom 40% under three feature selection methods were chosen. Table 12 shows the eliminated features for each dataset.

Table 12. Eliminated features in each dataset for final LSM, this table is in the order of feature importance. Features at the top are ones that were eliminated first.

Geohazards	Landslide	Collapse	Unstable Slope
Eliminated features	Fault density	River density	River density
	TWI	Aspect	Curvature
	Road density	Land cover	TWI
	River density	Distance to rivers	Fault density
	Curvature	Fault density	Distance to rivers
	MNDWI	Curvature	Aspect
	Aspect	MNDWI	MNDWI

In Section 3.3, this paper discussed the discovery of high multi-correlation, while the VIF decreased significantly after the FE. The VIFs among the remaining features ranged from 1 to 7. When the VIF value is less than 10, the multi-correlation among features is low and acceptable, indicating that the FE has reduced data redundancy.

The remaining features were used to model the prediction. The LSMs were completed by outputting the probability of "predicted landslide hazard". Four ML models (CART, RF, LR, and SVC) were prepared and used. This paper sliced the prediction dataset in the order of county administrative boundaries. The results were divided into five classes: very low, low, medium, high, and very high, with an interval of 0.2. The ranking represented the susceptibility of geohazards. Figure 11 shows the LSM results.

To validate the results, this paper adopted the ROC curves and AUC. Figure 12 shows the ROC curves.

In addition, we compared the AUC changes before and after the FE. The results are shown in Table 13. For models with high data quality requirements (such as LR and SVC), the AUC improvement after the FE can be more than 0.24.

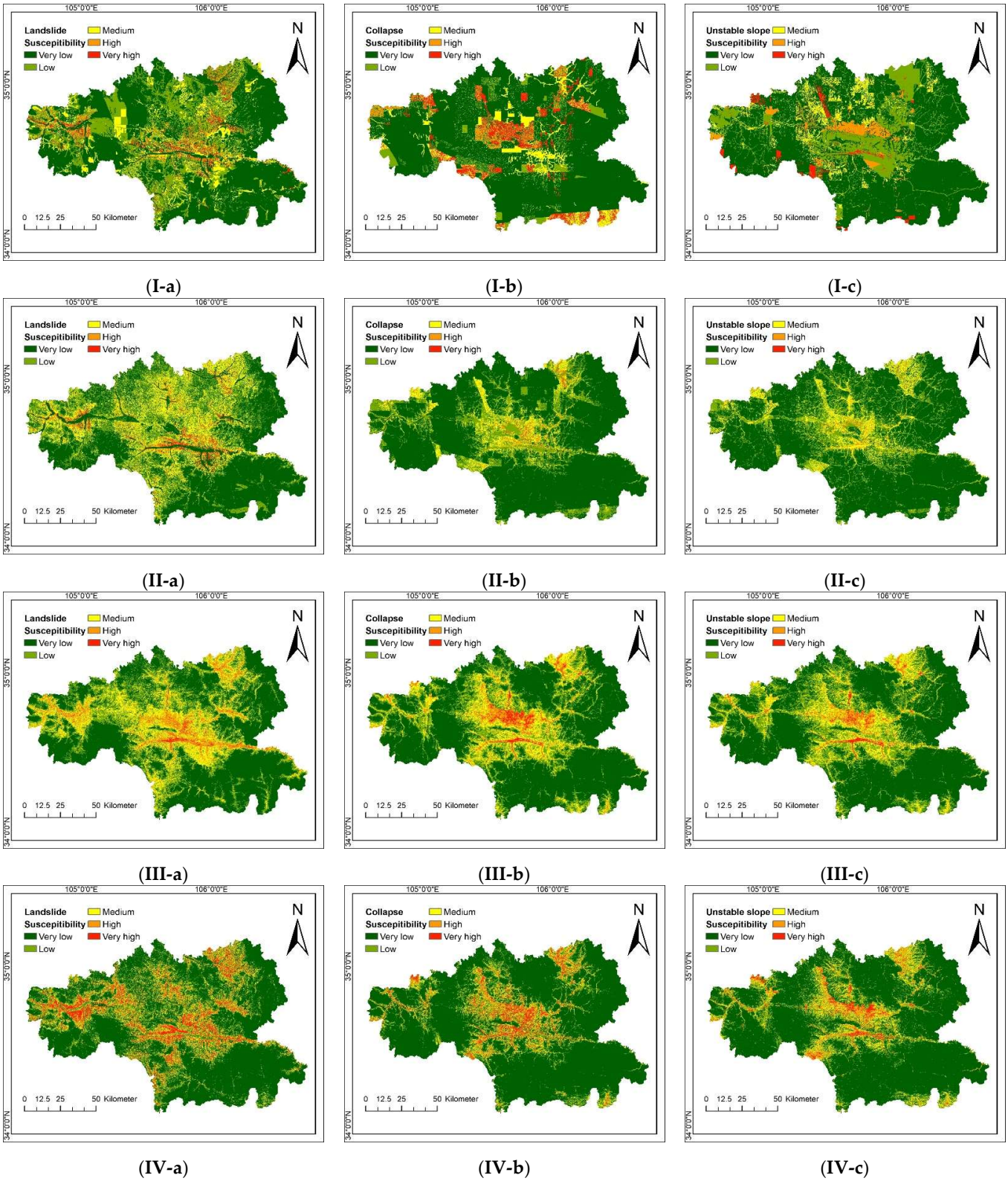


Figure 11. LSM results in Tianshui city. **Line I**: results of CART; (**I-a**): Landslide; (**I-b**): Collapse; (**I-c**): Unstable slope; **Line II**: results of RF; (**II-a**): Landslide; (**II-b**): Collapse; (**II-c**): Unstable slope; **Line III**: results of LR; (**III-a**): Landslide; (**III-b**): Collapse; (**III-c**): Unstable slope; **Line IV**: results of SVC; (**IV-a**): Landslide; (**IV-b**): Collapse; (**IV-c:**) Unstable slope.

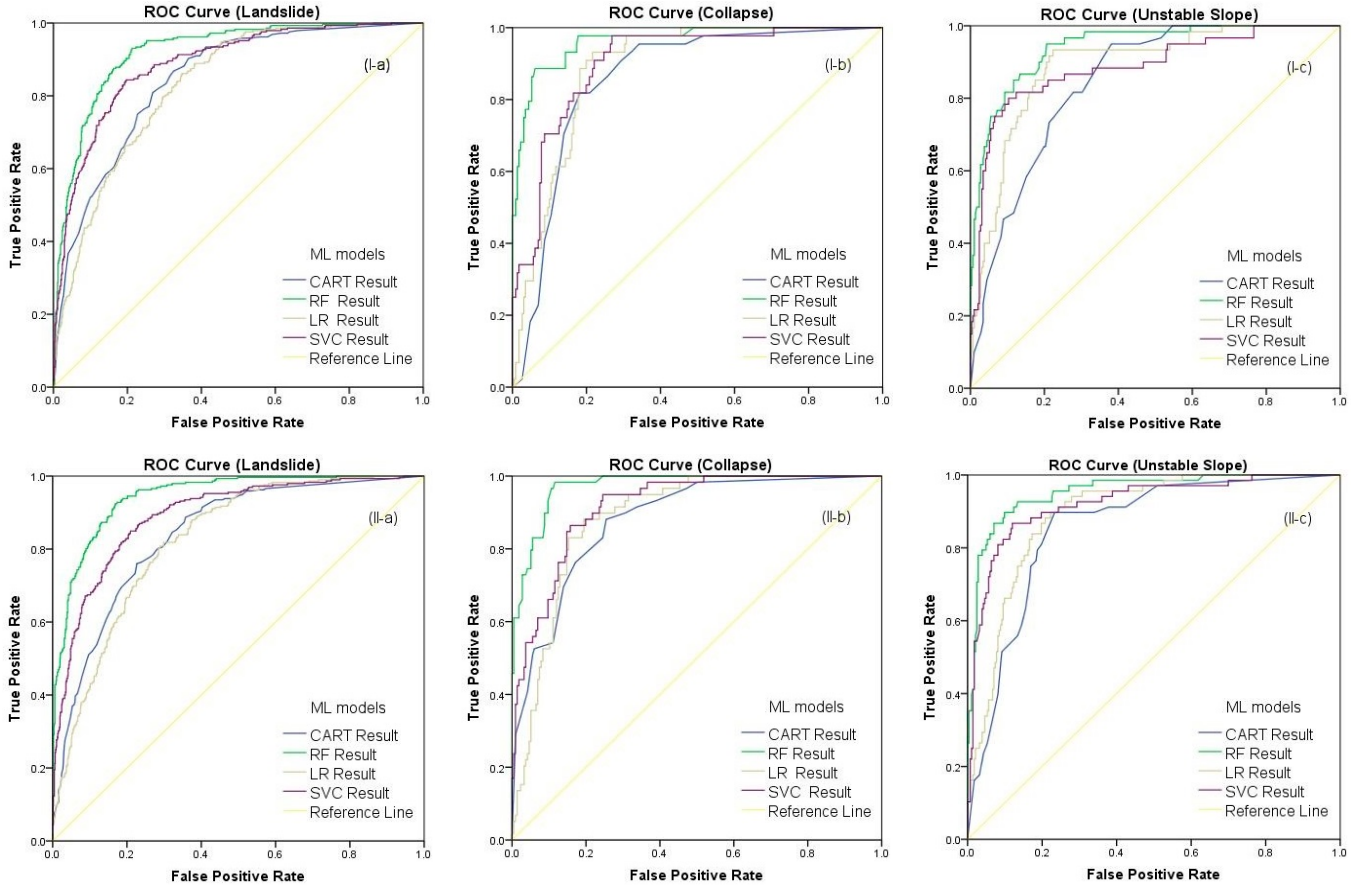


Figure 12. ROC curves of each model before and after the FE. **Line I**: results before the FE; (**I-a**): Landslide; (**I-b**): Collapse; (**I-c**): Unstable slope; **Line II**: results after the FE; (**II-a**): Landslide; (**II-b**): Collapse; (**II-c**): Unstable slope.

Table 13. Changes in AUC of different methods before and after FE procession.

Geohazards		CART	RF	LR	SVC
landslide	Before FE	0.844	0.933	0.836	0.877
	After FE	0.854	0.941	0.85	0.896
	Range	0.009	0.008	0.014	0.019
Collapse	Before FE	0.874	0.952	0.903	0.905
	After FE	0.878	0.957	0.898	0.913
	Range	0.004	0.005	−0.0056	0.008
Unstable Slope	Before FE	0.854	0.936	0.881	0.901
	After FE	0.878	0.949	0.901	0.912
	Range	0.024	0.013	0.019	0.011

4. Discussion

In this section, discussions that reflect the experiments are proposed. This paper summarizes the results and sorts the discourse into geohazards' FE results, FE improvement for different ML methods, and regional LSM.

1. Effectiveness of the FE:

As discussed in Section 3.4, the most suitable feature enhancement method is standardization; the improvements are more evident for LR or SVC. Four RF-based methods were proposed and tested for feature selection. As the RF modeling required parameter adjust-

ment and the dataset needed to be shuffled and re-divided, the accuracy is not fixed. Both wrapper-indicator methods appeared to be more unstable than the filter-indicator methods.

The filter-MDI feature selection method is the most stable, as it requires one-time modeling, with the simplest structure and the highest time efficiency. While the wrapper-MDA achieved the highest AUC improvement, the process is time-consuming and unstable. In most cases, the AUC changing curve oscillated hard, affecting the results and leading to apparent differences between the results obtained by other methods. This indicated that the iteration algorithm might not be the most suitable for LSM. Although it achieved the highest AUC improvement, the simple ranking filter methods are already enough to reach the goal of FE.

Still, by comparing the accuracy change of LSM models before and after the FE, it can be concluded that in most cases, the FE could bring a promised improvement upon LSM, especially for the LR and the SVC. Repeated experiments have been carried out to ensure that the FE would improve the results. For landslide, unstable slope, and most cases of collapse, the AUC after the FE was always higher than the original, although the increases sometimes seemed to be very tiny. For the CART and the RF, the accuracy improvements are not much. The reason may be that the FE methods proposed in this paper are all RF-based.

2. High susceptibility area distribution analysis and the correlated domain conditioning features for different geohazards:

Referring to Section 3.4 and the previous work of [42], the geohazards in Tianshui city mainly consist of loess-cutting and loess-plating forms. The high susceptibility areas are mainly distributed on the riverbanks, where the soil structure is easily cut and eroded by flowing water and rainfall. Since Tianshui is in the arid region of northwest China, the rivers are mostly surrounded by residential areas, and human activities can also damage the stability of the slopes.

The highest and lowest important conditioning features among the three hazards had certain similarities. In the order of ranking results, the dominant features are shown as follows:

(1) Slope gradient, elevation, precipitation, NDVI, lithology, land cover, and groundwater volume (landslide);
(2) Precipitation, elevation, slope gradient, distance to roads, distance to faults, groundwater volume, and lithology (collapse);
(3) Precipitation, NDVI, lithology, slope gradient, elevation, distance to roads, distance to faults, and road density (unstable slope).

By comparing the results for different hazards, the features most closely associated with regional geohazards are slope gradient, elevation, and precipitation. Moreover, details can be observed for different geohazards. Landslides and unstable slopes are more closely associated with vegetation cover (i.e., the NDVI). Meanwhile, the collapse and unstable slopes are related more to roads, reflecting that these two types of geohazards are highly influenced by human activities such as artificial slope cutting during road works. Distance to faults is another domain reason for collapse and unstable slopes, yet it is not equally important for landslides. Groundwater volume has been an important conditioning feature for landslides and collapses but is not prominent for unstable slopes. Meanwhile, comparing the eliminated features in Table 12, we can conclude that this aspect has a weak connection between all types of geohazards. This is because the overall low vegetation cover of the study area, collapse, and unstable slopes has little relation with rivers. However, they are still affected by rainfall and groundwater.

These discussions show that the principles concluded by ML algorithms are somehow consistent with the laws of geography and geology, thus proving the efficiency of the FE. However, the results are subject to the study area due to the limited landslide inventory and study area. The geohazard type in Tianshui city is simple (loess geohazard mainly), while the triggering condition is mostly rainfall-flush of gravity. Multiple study areas

should be involved for further studies to test the FE and to reveal the hazard formation modes thoroughly.

3. Accuracy and validation of LSM:

From Table 13, the ML model with the best analysis results and highest AUC is always the RF model. Considering the cartographic mapping results, the CART model performed the worst with too many block patches and a clear slice boundary of the dataset. However, this phenomenon appeared in both DT-based models due to the large spatial resolution scale that caused homogeneous value distribution of some conditioning factors (e.g., precipitation, groundwater volume, etc.).

The overall test dataset is sliced into the administrative boundary, as we aimed to give clear advice for each county. All four ML models had short optimization in the boundaries. The LR and SVC had more continuous mapping results; therefore, the administrative boundaries are relatively insignificant. Regarding the zoning effect, the analysis results of the RF model had the highest differentiation between the medium, high, and very high susceptibility areas. The very high susceptibility area is the smallest, making it the best differed in the susceptibility results in riverbanks and other low-flat areas. Above all, the RF model outperformed the other three ML models, but the mapping results appeared to be dispersed; thus, the result is slightly worse than the one obtained by the SVC.

4. The correlating mitigation advice:

For landslides, the high and very high susceptibility areas are located in Qinzhou County, north Maiji County, west Zhangjiachuan County, and east Wushan County, mostly in farming and wetland areas. Along the Shaoxing River, Weihe River, Baimao River, and Qingshui River, with a slope gradient ranging mostly below 20°, the corresponding elevation is below 2 km. Lithology types are mostly sandy loess and clay. For collapses, higher susceptibility to landslides is distributed in the central part of the study area, mainly on both sides of the West Qinling-Beiyuan Faults and the Lixian-Luojiabao Faults. The high and very high susceptibility areas of landslides and collapses overlapped, indicating that the two types of geohazards have similar characteristics. Since unstable slopes can be regarded as developing landslides or collapses, the high susceptibility area of unstable slopes is small and located within the other two.

The high susceptibility areas of three geohazards in Tianshui city are generally distributed along the low and gentle terrain of rivers and plates that are used mostly for farming and construction, indicating that the soil and water conservation capacity has been damaged. In addition, all three geohazards show a strong tendency to be distributed along both sides of the road, indicating that the stability decreases in rock–soil bodies due to anthropogenic activities have been severe.

5. Conclusions

By summarizing Section 4, the conclusions are as follows:

The FE has been proven effective and can help improve the accuracy of LSM. The wrapper-indicator methods performed better than the filter-indicator methods in accuracy yet appeared to be more unstable and time-consuming. However, the improvement differences between feature selection methods are not significant enough to ignore the time cost and the overfitting of iteration methods; the simplest filter-MDI method can already meet the demand for efficient and accurate LSM. This paper recommends using the simple method to balance the modeling complexity, accuracy, and robustness. What matters the most is whether the FE is involved in the LSM, not which method is applied for feature selection.

By analyzing the feature importance ranking results, this paper lists the features of the most significant influence:

(1) Slope gradient, elevation, precipitation, NDVI, lithology, land cover, and groundwater volume (landslide);

(2) Precipitation, elevation, slope gradient, distance to roads, distance to faults, groundwater volume, and lithology (collapse);
(3) Precipitation, NDVI, lithology, slope gradient, elevation, distance to roads, distance to faults, and road density (unstable slope).

This paper jointly analyzed the high-susceptibility regions of each geohazard with the dominant conditioning features and discovered that the geohazards in Tianshui city are strongly influenced by hydrology and anthropogenic activities. It is recommended to replan the farming areas and increase the protection of wetlands to enhance the soil water conservation capacity, as well as carry out geotechnical reinforcement along the road to avoid casualties and economic losses due to disasters.

Author Contributions: Conceptualization, X.L., Y.Z. and D.M.; data supplement, Y.Z., methodology, X.L.; experiments and validation, X.L., T.D. and Y.C.; funding acquisition, Y.Z.; writing, X.L. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been jointly supported by the National Natural Science Foundation of China (41872253), the National Key R & D Program of China (2022YFB3900017), the China Geological Survey [DD2021364], and the Fundamental Research Funds for the Central Universities.

Data Availability Statement: Parts of related data can be found at Geospatial Data Cloud (http://www.gscloud.cn/, accessed on 28 October 2020), Google Earth Engine (https://explorer.earthengine.google.com/), and http://data.casearth.cn/, accessed on 12 January 2021.

Acknowledgments: The authors would like to thank Tao Wang's team and the Chinese Academy of Geological Sciences Institute of Geomechanics for data support.

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

References

1. Highland, L.; Bobrowsky, P.T. *The Landslide Handbook: A Guide to Understanding Landslides*; US Geological Survey Reston: Reston, VA, USA, 2008.
2. Nadim, F.; Kjekstad, O.; Peduzzi, P.; Herold, C.; Jaedicke, C. Global landslide and avalanche hotspots. *Landslides* **2006**, *3*, 159–173. [CrossRef]
3. Petley, D. Global patterns of loss of life from landslides. *Geology* **2012**, *40*, 927–930. [CrossRef]
4. Xu, C.; Xu, X.; Shen, L.; Yao, Q.; Tan, X.; Kang, W.; Ma, S.; Wu, X.; Cai, J.; Gao, M.J. Optimized volume models of earthquake-triggered landslides. *Sci. Rep.* **2016**, *6*, 29797. [CrossRef] [PubMed]
5. Qing, Y.; Ming, D.; Wen, Q.; Weng, Q.; Xu, L.; Chen, Y.; Zhang, Y.; Zeng, B. Operational earthquake-induced building damage assessment using CNN-based direct remote sensing change detection on superpixel level. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102899. [CrossRef]
6. Brabb, E.E. Innovative approaches to landslide hazard and risk mapping. In Proceedings of the International Landslide Symposium Proceedings, Toronto, ON, Canada, 23–31 August 1985; pp. 17–22.
7. Chacón, J.; Irigaray, C.; Fernandez, T.; El Hamdouni, R. Engineering geology maps: Landslides and geographical information systems. *Bull. Eng. Geol. Environ.* **2006**, *65*, 341–411. [CrossRef]
8. Neuland, H. A prediction model of landslips. *Catena* **1976**, *3*, 215–230. [CrossRef]
9. Shahabi, H.; Ahmad, B.; Khezri, S. Evaluation and comparison of bivariate and multivariate statistical methods for landslide susceptibility mapping (case study: Zab basin). *Arab. J. Geosci.* **2013**, *6*, 3885–3907. [CrossRef]
10. He, Y.; Beighley, R. GIS-based regional landslide susceptibility mapping: A case study in southern California. *Earth Surf. Processes Landf. J. Br. Geomorphol. Res. Group* **2008**, *33*, 380–393. [CrossRef]
11. Van Westen, C. Statistical landslide hazard analysis. *Ilwis* **1997**, *2*, 73–84.
12. Chen, W.; Pourghasemi, H.R.; Kornejady, A.; Zhang, N. Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. *Geoderma* **2017**, *305*, 314–327. [CrossRef]
13. Liu, Y.; Zhao, L.; Bao, A.; Li, J.; Yan, X. Chinese High Resolution Satellite Data and GIS-Based Assessment of Landslide Susceptibility along Highway G30 in Guozigou Valley Using Logistic Regression and MaxEnt Model. *Remote Sens.* **2022**, *14*, 3620. [CrossRef]
14. Van Westen, C.J. Application of Geographic Information Systems to Landslide Hazard Zonation. Ph.D. Thesis, Delft University of Technology, Delft, The Netherlands, 1993.
15. El Abidine, R.Z.; Abdelmansour, N. Landslide susceptibility mapping using information value and frequency ratio for the Arzew sector (North-Western of Algeria). *Bull. Miner. Res. Explor.* **2019**, *160*, 197–211. [CrossRef]
16. Jordan, M.I.; Mitchell, T. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]

17. Das, I.; Sahoo, S.; van Westen, C.; Stein, A.; Hack, R. Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India). *Geomorphology* **2010**, *114*, 627–637. [CrossRef]
18. Mao, Y.-m.; Zhang, M.-s.; Wang, G.-l.; Sun, P.-P. Landslide hazards mapping using uncertain Naïve Bayesian classification method. *J. Cent. South Univ.* **2015**, *22*, 3512–3520. [CrossRef]
19. 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.
20. Catani, F.; Lagomarsino, D.; Segoni, S.; Tofani, V. Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 2815–2831. [CrossRef]
21. Hong, H.; Miao, Y.; Liu, J.; Zhu, A.-X. Exploring the effects of the design and quantity of absence data on the performance of random forest-based landslide susceptibility mapping. *Catena* **2019**, *176*, 45–64. [CrossRef]
22. Hong, H.; Liu, J.; Bui, D.T.; Pradhan, B.; Acharya, T.D.; Pham, B.T.; Zhu, A.-X.; Chen, W.; Ahmad, B.B. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *Catena* **2018**, *163*, 399–413. [CrossRef]
23. Zhao, L.; Wu, X.; Niu, R.; Wang, Y.; Zhang, K.J. Using the rotation and random forest models of ensemble learning to predict landslide susceptibility. *Nat. Hazards Risk* **2020**, *11*, 1542–1564. [CrossRef]
24. Huang, Y.; Zhao, L. Review on landslide susceptibility mapping using support vector machines. *Catena* **2018**, *165*, 520–529. [CrossRef]
25. Marjanović, M.; Kovačević, M.; Bajat, B.; Voženílek, V. Landslide susceptibility assessment using SVM machine learning algorithm. *Eng. Geol.* **2011**, *123*, 225–234. [CrossRef]
26. Moayedi, H.; Mehrabi, M.; Mosallanezhad, M.; Rashid, A.S.A.; Pradhan, B. Modification of landslide susceptibility mapping using optimized PSO-ANN technique. *Eng. Comput.* **2019**, *35*, 967–984. [CrossRef]
27. Chen, Y.; Ming, D.; Ling, X.; Lv, X.; Zhou, C. Landslide Susceptibility Mapping Using Feature Fusion-Based CPCNN-ML in Lantau Island, Hong Kong. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3625–3639. [CrossRef]
28. Wang, Y.; Fang, Z.; Hong, H. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. *Sci. Total Environ.* **2019**, *666*, 975–993. [CrossRef]
29. Zhang, Y.; Chen, Y.; Ming, D.; Zhu, Y.; Ling, X.; Zhang, X.; Lian, X. Landslide Hazard Analysis Based on SBAS-InSAR and MCE-CNN Model: A case study of Kongtong, Pingliang. *Geocarto Int.* **2022**, 1–20, *just-accepted*. [CrossRef]
30. Kavzoglu, T.; Sahin, E.K.; Colkesen, I. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* **2014**, *11*, 425–439. [CrossRef]
31. Tsangaratos, P.; Ilia, I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena* **2016**, *145*, 164–179. [CrossRef]
32. Chen, W.; Xie, X.; Wang, J.; Pradhan, B.; Hong, H.; Bui, D.T.; Duan, Z.; Ma, J. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* **2017**, *151*, 147–160. [CrossRef]
33. Youssef, A.M.; Pourghasemi, H.R.; Pourtaghi, Z.S.; Al-Katheeri, M.M. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides* **2016**, *13*, 839–856. [CrossRef]
34. Luo, X.; Lin, F.; Zhu, S.; Yu, M.; Zhang, Z.; Meng, L.; Peng, J.J. Mine landslide susceptibility assessment using IVM, ANN and SVM models considering the contribution of affecting factors. *PLoS ONE* **2019**, *14*, e0215134. [CrossRef] [PubMed]
35. Hong, H.; Pourghasemi, H.R.; Pourtaghi, Z.S. Landslide susceptibility assessment in Lianhua County (China): A comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology* **2016**, *259*, 105–118. [CrossRef]
36. Pham, B.T.; Prakash, I.; Bui, D.T. Spatial prediction of landslides using a hybrid machine learning approach based on random subspace and classification and regression trees. *Geomorphology* **2018**, *303*, 256–270. [CrossRef]
37. Chen, T.; Zhu, L.; Niu, R.-Q.; Trinder, C.J.; Peng, L.; Lei, T. Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. *J. Mt. Sci.* **2020**, *17*, 670–685. [CrossRef]
38. Cheng, Y.-S.; Yu, T.-T.; Son, N.-T. Random Forests for Landslide Prediction in Tsengwen River Watershed, Central Taiwan. *Remote Sens.* **2021**, *13*, 199. [CrossRef]
39. Zhou, X.; Wen, H.; Zhang, Y.; Xu, J.; Zhang, W. Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. *Geosci. Front.* **2021**, *12*, 101211. [CrossRef]
40. 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]
41. Zhou, X.; Chen, F.; Wu, X.; Qian, R.; Liu, X.; Wang, S. Variation Characteristics of Stable Isotopes in Precipitation and Response to Regional Climate Conditions during Pre-monsoon, Monsoon and Post-monsoon Periods in the Tianshui Area. *Water* **2020**, *12*, 2391. [CrossRef]
42. Zhang, Z.-l.; Wang, T.; Wu, S.-R. Distribution and features of landslides in the Tianshui Basin, Northwest China. *J. Mt. Sci.* **2020**, *17*, 686–708. [CrossRef]

43. 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]
44. Ling, X.; Liu, J.; Wang, T.; Zhu, Y.; Yuan, L.; Chen, Y. Application of information value model based on symmetrical factors classification method in landslide hazard assessment. *Remote Sens. Nat. Resour.* **2021**, *33*, 172–181.
45. Weiss, A. Topographic position and landforms analysis. In Proceedings of the Poster Presentation, ESRI User Conference, San Diego, CA, USA, 9–13 July 2001.
46. Pham, B.T.; Pradhan, B.; Bui, D.T.; Prakash, I.; Dholakia, M. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environ. Model. Softw.* **2016**, *84*, 240–250. [CrossRef]
47. Lee, S. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. *Int. J. Remote Sens.* **2005**, *26*, 1477–1491. [CrossRef]
48. McFeeters, S. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
49. Su, L.-J.; Hu, K.-H.; Zhang, W.-F.; Wang, J.; Lei, Y.; Zhang, C.-L.; Cui, P.; Pasuto, A.; Zheng, Q.-H. Characteristics and triggering mechanism of Xinmo landslide on 24 June 2017 in Sichuan, China. *J. Mt. Sci.* **2017**, *14*, 1689–1700. [CrossRef]
50. Liu, H.; Setiono, R. A probabilistic approach to feature selection-a filter solution. In Proceedings of the ICML, Bari, Italy, 3–6 July 1996; pp. 319–327.
51. Sánchez-Marono, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter methods for feature selection–a comparative study. In *International Conference on Intelligent Data Engineering and Automated Learning*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 178–187.
52. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 856–863.
53. Zhukov, A.V.; Sidorov, D.N.; Foley, A.M. Random forest based approach for concept drift handling. In Proceedings of the International Conference on Analysis of Images, Social Networks and Texts, Yekaterinburg, Russia, 7–9 April 2016; pp. 69–77.
54. Hall, M.A.; Smith, L.A. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In Proceedings of the FLAIRS Conference, Orlando, FL, USA, 1–5 March 1999; pp. 235–239.
55. Maldonado, S.; Weber, R. A wrapper method for feature selection using support vector machines. *Inf. Sci.* **2009**, *179*, 2208–2217. [CrossRef]
56. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
57. Stumpf, A.; Kerle, N. Object-oriented mapping of landslides using Random Forests. *Remote Sens. Environ.* **2011**, *115*, 2564–2577. [CrossRef]
58. Yesilnacar, E.; Topal, T. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Eng. Geol.* **2005**, *79*, 251–266. [CrossRef]
59. Yilmaz, I. Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: Conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environ. Earth Sci.* **2010**, *61*, 821–836. [CrossRef]
60. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]