symmetry

MDPI

Article

# Mapping Landslide Susceptibility Using Machine Learning Algorithms and GIS: A Case Study in Shexian County, Anhui Province, China

**Zitao Wang [1], Qimeng Liu [1,*] and Yu Liu [2]**

[1] School of Earth and Environment, Anhui University of Science and Technology, Huainan 232001, China; wzt3800@gmail.com

[2] State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines, Anhui University of Science and Technology, Huainan 232001, China; yliu@aust.edu.cn

* Correspondence: qmliu@aust.edu.cn; Tel.: +86-0554-663-3992

Received: 20 October 2020; Accepted: 23 November 2020; Published: 26 November 2020


**Abstract:** In this study, Logistics Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Machine (GBM), and Multilayer Perceptron (MLP) machine learning algorithms are combined with GIS techniques to map landslide susceptibility in Shexian County, China. By using satellite images and various topographic and geological maps, 16 landslide susceptibility factor maps of Shexian County were initially constructed. In total, 502 landslide and random safety points were then using the "Extract Multivalues To Points" tool in ArcGIS, parameters for the 16 factors were extracted and imported into models for the five algorithms, of which 70% of samples were used for training and 30% of samples were used for verification, which makes sense for date symmetry. The Shexian grid was converted into 260130 vector points and imported into the five models, and the natural breakpoint method was used to divide the grid into four levels: low, moderate, high, and very high. Finally, by using column results gained using Area Under Curve (AUC) analysis and a grid chart, susceptibility results for mapping landslide prediction in Shexian County was compared using the five methods. Results indicate that the ratio of landslide points of high or very high levels from LR, SVM, RF, GBM, and MLP was 1.52, 1.77, 1.95, 1.83, and 1.64, and the ratio of very high landslide points to grade area was 1.92, 2.20, 2.98, 2.62, and 2.14, respectively. The success rate of training samples for the five methods was 0.781, 0.824, 0.853, 0.828, and 0.811, and prediction accuracy was 0.772, 0.803, 0.821, 0.815, and 0.803, respectively; the order of accuracy of the five algorithms was RF > SVM > MLP > GBM > LR. Our results indicate that the five machine learning algorithms have good effect on landslide susceptibility evaluation in Shexian area, with Random Forest having the best effect.

**Keywords:** landslide susceptibility mapping; machine learning algorithms; Shexian country; logistics regression; support vector machine; random forest; gradient boosting machine; multilayer perceptron

## 1. Introduction

Landslides are a common geological disaster, resulting in economic losses of up to $100 billion globally and accounting for hundreds of deaths. Landslides have a serious effect on the lives and safety of populations, affecting the stable development of society [1]. Although China has a considerable land area, it is characterized by a generally flat topography in the north and the east, with higher land areas in the south and the west. Landslides in China are therefore predominantly concentrated in the south and western areas, especially near the Yangtze River Basin. Over the past 70 years, more than 20,000 people have died due to landslides in China; annual economic losses caused by landslides amount to

more than $50 million [2]. Shexian county is located in the core region of the mountainous area of the Southern Anhui Province, being the area most affected by landslides in Anhui Province [3]. Since 1970, non-seismic landslides in Shexian county have resulted in numerous deaths and injuries, with more than 1098 houses being destroyed; direct economic losses are calculated to be $ 1.549 million [4]. Previous investigations into landslides in Shexian county either mainly focused on simple geological hazard susceptibility zoning of Shexian and Huangshan City [4], or they analyzed the distribution law, formation mechanism, and influencing factors of landslides from geological or geotechnical points of view [5]. Quantitative analysis of landslide susceptibility in this area has been largely overlooked.

The main method of landslide hazard assessment combines the landslide susceptibility evaluation model with GIS data. This model includes qualitative, semi quantitative and quantitative evaluations. Qualitative evaluation is the main expert scoring method; the semi quantitative evaluation model includes analytic hierarchy processes [6,7] and Fuzzy Comprehensive Evaluation; and the quantitative evaluation model includes frequency ratio [8–10], the entropy index (IOE) model [11–13], Weight Of Evidence (WOE) [14–16], and frequently used machine learning algorithms [17,18], such as Logistic Regression [19,20], Support Vector Machine [21,22], Neural Network [23–25], and Decision Tree methods [26,27]. As each model has its own way of being applied, as well as different advantages and disadvantages, different models often lead to different evaluation results when used in the same region. In order to overcome these differences, researchers usually choose some methods for comparison [28–31]. However, as qualitative or semi quantitative evaluation methods are often affected by human factors, it is therefore more important to select the most appropriate quantitative evaluation model for a certain region.

Machine learning algorithm is a method that uses existing data to predict geological disasters. Although landslides are a main type of geological disasters, problems exist in analyzing the large amount of data collected on these phenomena. It is therefore very important to establish a machine learning model for landslide prediction, evaluation and interpretation using geological data [32]. At the same time, on the basis of establishing a certain model and using the model to predict outcomes, quantitative characteristics of a machine learning algorithm can reduce interference caused by human misjudgment, enabling it to be effectively applied to landslide susceptibility zoning [33]. At the same time, Python is used as the platform to build the model, enabling each grid point of a landslide to be analyzed and mapped individually, thereby having a good effect when processing large data sets. By taking Shexian County as a case study, we selected five representative quantitative evaluation machine learning supervision algorithms: Logistics Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Machine (GBM), and Multilayer Perception (MLP). Results obtained from these algorithms were evaluated and compared, enabling the most suitable landslide susceptibility evaluation model for Shexian and similar areas to be identified.

## 2. Study Area

Shexian County, located in the southernmost tip of Anhui Province, China, bordering Zhejiang Province, has an area of 2122 km$^2$ (118°15′00″–118°53′50″ E, 29°30′25″–30°07′00″ N). In 2020, the permanent resident population of Shexian County was 420000. Elevation in this county fluctuates greatly, ranging from 67 to 1777 m. In general, terrain in the east and the central area is relatively low, with areas in the south, west and north being higher. Shexian County belongs to the subtropical North Edge mountain thick monsoon humid climate; spring and autumn each account for two months and summer and winter each account for four months. Annual average temperature in this county is 16.3 °C, and annual precipitation is 1100–2000 mm (average annual precipitation is 1582.7 mm). The study area has a developed water system, comprising a dense river network. Almost all rivers in the study region converge into Xin'an River, the largest river in Shexian County. This river has a channel slope of 1.71% and an average annual discharge of 41.86 m$^3$/s. Shexian County also has a good transport network, having a total highway mileage of 1682.5 km [4]. In terms of geology, the study area is located in the South China stratigraphic area, and exposed strata include Sinian of Proterozoic,

Cambrian of Paleozoic, Ordovician, Jurassic and Cretaceous of Mesozoic, and Quaternary loose soil and deposition. In terms of structure, Shexian County has experienced many tectonic movements, with abundant folds and faults.

Landslide event refers to the phenomenon that downward movement of rock or soil when the gravity or other types of shear stress exceed the shear strength of the slope. Shexian County is frequently affected by landslides events, with a total of 502 landslide disaster points recorded in this area (Figure 1). In general, landslide disasters in the study area are characterized by a large number and a dense distribution.

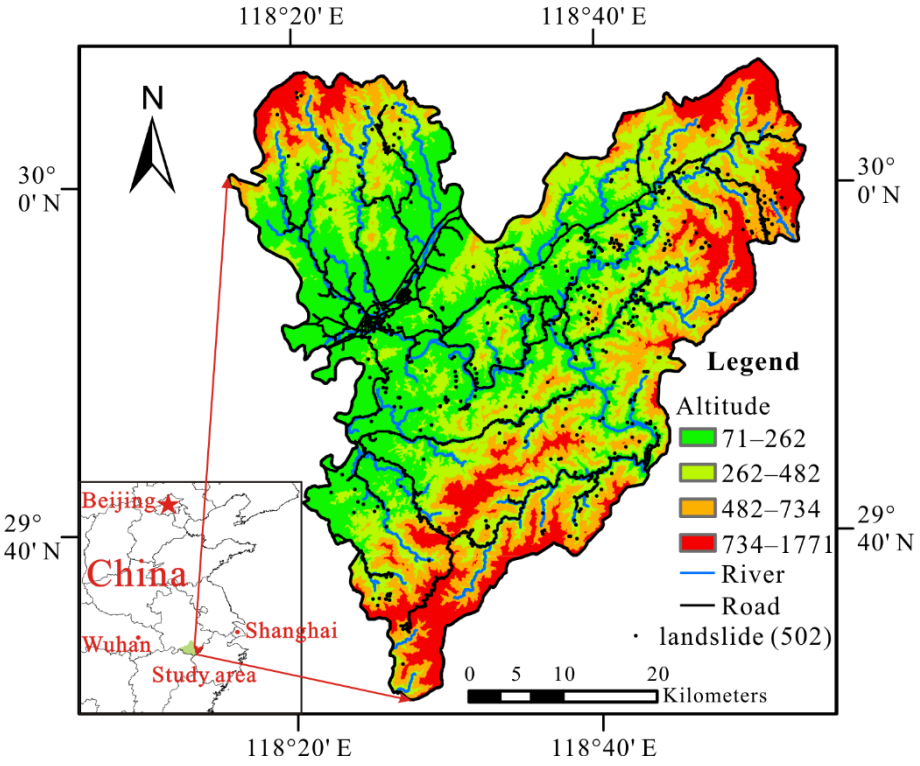


**Figure 1.** The study area characteristics and the location of landslide events.

## 3. Data

In order to map and predict landslide susceptibility in Shexian County, 502 landslide points were collected in this study. It is important to note that the range of thematic data types used for susceptibility assessment has not changed significantly over time [18]. With reference to previous studies, and according to the geological environment of the study area and the development characteristics of landslides, a total of 16 condition factors were selected and divided into four categories according to type: topography and geology, hydrology, and others [34]. Among the different condition factors, topography mainly incorporated downloaded DEM geospatial data from the official cloud website (http://www.gscloud.cn/search). This category included aspect, slope, plane curvature, profile curvature, topographic relief, surface roughness, and landform. Geology was mainly derived from vectorization of the regional geological map, including faults and lithology. Hydrology was derived from vectorization of the topographic map, grid calculation of GIS, and extraction of the rainfall distribution map, which were divided into river and net flow intensity indices. The other category included road and vegetation coverage. All of the conditional categories were comprised into grid images with a pixel size of 30 m × 30 m.

### *3.1. Topographical and Geological Factors*

Among the terrain factors, slope aspect, slope degree, plane curvature, profile curvature, topographic relief, and surface roughness were extracted from DEM data in GIS. Although slope aspect does not directly affect landslide stability, different slope aspects are affected by different levels of solar radiation and weathering, resulting in different slope characteristics, thus indirectly affecting stability [35,36]. According to different directions, the slope aspect of Shexian County can be divided into flat (0), north (>337.5 or ≤22.5), northeast (22.5< and ≤67.5), east (67.5< and ≤112.5), southeast (112.5< and ≤157.5), south (157.5< and ≤202.5), southwest (202.5< and ≤247.5), west (247.5< and ≤292.5), northwest (292.5< and ≤337.5), west (247.5< and ≤292.5), northwest (292.5< and ≤337.5), west (247.5< and ≤292.5), northwest (292.5< and ≤ 337.5 < and ≤337.5), and northwest (292.5< and ≤337.9) (Figure 2a).

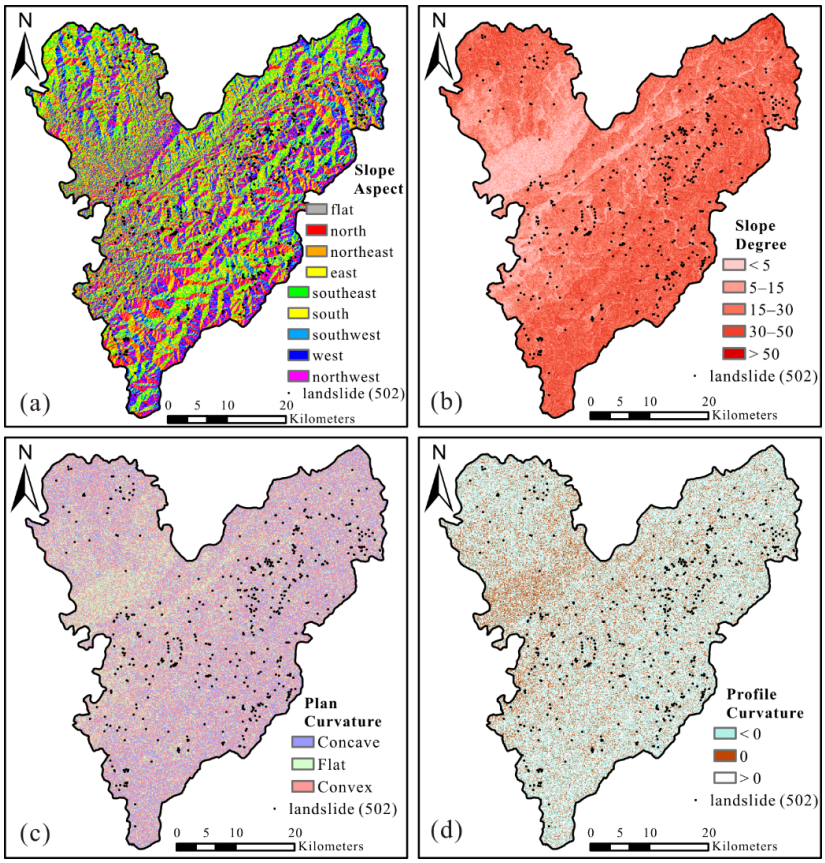


**Figure 2.** Topographic and Geology parameter maps of the study area and landslide location points: (**a**) Slope aspect; (**b**) Slope degree; (**c**) Plan curvature; (**d**) Profile curvature.

Slope degree is the angle between the slope section and the horizontal plane. This not only determines the spatial distribution characteristics of a landslide, it also controls the geotechnical distribution of the slope. This factor has an important effect on slope stability [37]. Slope variation in Shexian County was found to notably differ, being divided into five categories: <5, 5–15, 15–30, 30–50, and >50° (Figure 2b).

Terrain curvature, a quantitative measure of the degree of change of each point on a slope [29], is decomposed into horizontal and vertical directions, termed plan curvature and profile curvature, respectively [4,34,38]. Plane curvature extracts aspect from DEM data before extracting slope from aspect. This can be divided into three categories: concave (<0), flat (0), and convex (>0) (Figure 2c). Profile curvature extracts slope twice from DEM data, being divided into <0, 0, and >0 (Figure 2d).

Topographical relief refers to the difference between the highest and the lowest altitudes in a certain area. Landslides developed on slopes with different elevation differences are often different [39,40]. The relief degree of Shexian County can be divided into five grades: <15, 15–30, 30–45, 45–60, and >60 (Figure 3a).

Surface roughness of Shexian reflects the degree of surface erosion [41], being the ratio of the surface area of the surface unit to the projected area on the horizontal plane. By dividing surface roughness of Shexian County into <1.05, 1.05–1.15, 1.15–1.30, and >1.30, landslides can be seen to be mainly distributed in areas with relatively low surface roughness (Figure 3b).

Different geomorphic phenomena in Shexian County have had different effects on the occurrence of landslides in this region [34,42]. Landforms in the study area include plains, hills and mountains. Mountainous areas and hills account for 95% of the total study area, and plains account for 5%. Geomorphology of the study area was divided into eight grades: I-1 (plain), I-2 (shallow hilly plain), II-1 (medium hill), II-2 (high hill), III-1 (low undulating low mountain), III-2 (high undulating low mountain), IV-1 (low undulating mountain), and IV-2 (high undulating mountain) [4] (Figure 3c).

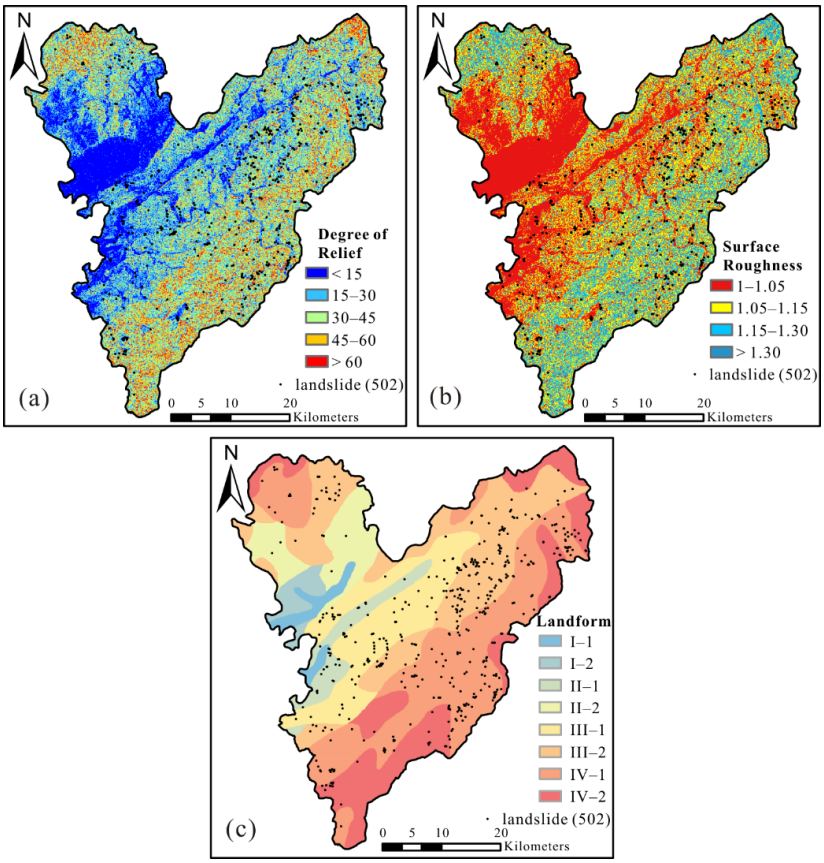


**Figure 3.** Topographic and Geology parameter maps of the study area and landslide location points: (**a**) Degree of relief; (**b**) Surface roughness; (**c**) Landform.

Geological faults control the weak structural plane of a slope. Faults act to cut rock and soil mass of a slope into a discontinuous whole, forming a potential landslide point. At the same time, faults can destroy rock and soil structure, providing a channel for rainfall and other factors [43]. Based on the distance from faults, the study area can be divided into five grades: <400, 400–800, 800–1200, 1200–2000, and >2000 m (Figure 4a). The closer the study area is to the fault, the greater is the impact of the fault.

Lithology is the basis of landslide development, and different lithology compositions affect the type and scale of a landslide [40,44]. Through vectorization of the bedrock geological map, a stratigraphic lithologic map of Shexian County was created and divided into six types (Table 1, Figure 4b).

**Table 1.** Lithology classification.

| Group | Stratum and Lithology |
|---|---|
| 1 | Mesoproterozoic. Black slate with light metamorphic lithic arkose and siltstone. |
| 2 | Lower Sinian. This is composed of grayish green and purplish red argillaceous conglomerate, gravelly sandstone and gravelly sandstone with a small amount of mudstone. |
| 3 | Middle Sinian. Gray black, light gray, thin-thick layer siliceous rock and siliceous shale. |
| 4 | Cambrian. Argillaceous limestone and dolomitic limestone are mainly mixed with a thin marl, carbonaceous and calcareous mudstone. |
| 5 | Quaternary. Pebble sandy fine gravel layer, silty sand clay layer, gravel, sandy soil layer and gravelly loam. |
| 6 | Granite |

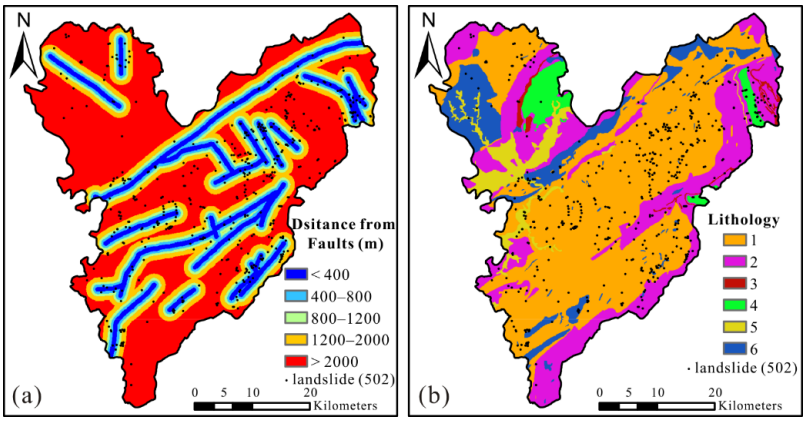


**Figure 4.** Topographic and Geology parameter maps of the study area and landslide location points: (**a**) Distance from faults; (**b**) Lithology.

### *3.2. Hydrological Factors*

Water is an important factor affecting the development of landslides, playing a key role in the evaluation of landslide susceptibility. Hydrological factors mainly derive from vectorization of the topographic map and grid calculation by GIS. Rainfall derives from the precision monitoring points established by Shexian meteorological stations in 28 towns and villages, as well as the interpolation of observation data of other meteorological stations nearby (Tunxi District, Huangshan District) [4].

The spatial distance from rivers or water systems expresses the influence of water level change on landslide formation [45]. Slopes can be easily eroded by a river water system, forming a steep free surface. These changes can result in a change in slope stress, resulting in a landslide [46]. By vectorizing rivers in a topographic map, buffer zones were established using GIS. Based on the distance from faults, the study area can be divided into five grades: <400, 400–800, 800–1200, 1200–2000, and >2000 m (Figure 5a).

The catchment area (*As*) was calculated using GIS hydrological tools such as "Flow Direction" and "Flow Accumulation" from DEM data in the study area. The Stream Power Index (SPI) was calculated using the "Raster Calculator" [47–49] of:

$$SPI = As \times \tan\beta \tag{1}$$

where, $\beta$ is the slope. In this equation, slope was converted to radians as:

$$radians = \frac{degrees \times \pi}{180} \tag{2}$$

SPI of Shexian County can be divided into three categories: <2, 2–4, and >4 (Figure 5b).

Topographic Wetness Index (TWI) is a quantitative description of soil moisture in a watershed. An understanding of the influence terrain change has on a soil can be gained using the following equation [50,51]:

$$TWI = \ln \frac{As}{\tan \beta} \quad (3)$$

Results for TWI enabled Shexian County to be divided into three categories: <6, 6–8 and >8 (Figure 5c).

Flow length is the projection length, referring to the maximum ground distance from a ground point along the flow direction to the start or end point on the horizontal plane, directly affects the speed of surface runoff and causes different erosivity to surface soil. [52]. Flow length of Xin'an River and its tributaries, extracted from DEM data by GIS, can be divided into five categories: <500, 500–1000, 1000–1500, 1500–2500, and >2500 m (Figure 5d).

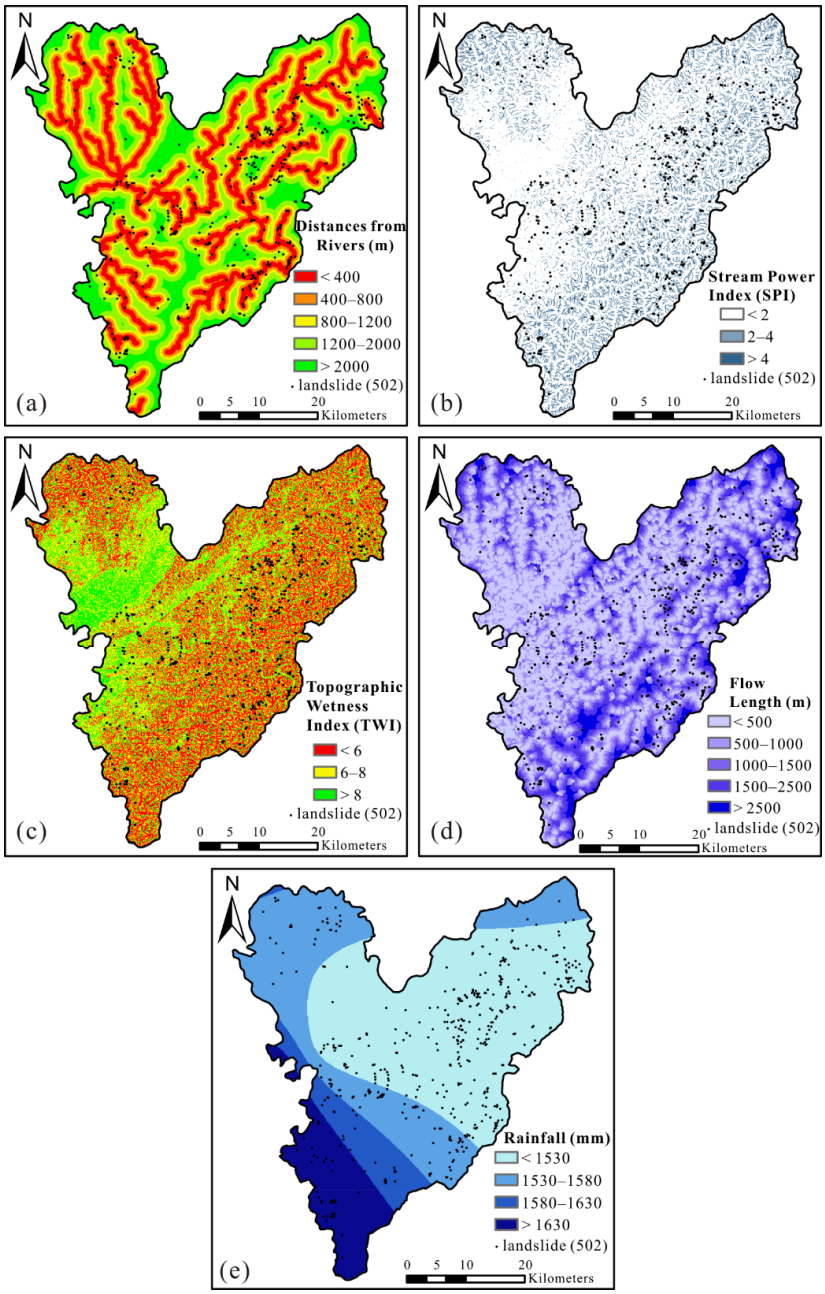


**Figure 5.** Hydrology parameter maps of the study area and landslide location points: (**a**) Distance from Rivers; (**b**) Stream Power Index; (**c**) Topographic Wetness Index; (**d**) Flow Length; (**e**) Rainfall.

Landslides caused by rainfall events refer to the natural phenomenon whereby rainfall infiltration leads changes in pore pressure and rock and soil mass strength, resulting in rocks and soil to slide downward along a certain weak surface under the action of gravity. Rainfall magnitude constitutes the risk degree of a landslide, with heavy rainfall having a direct effect on landslide occurrence [53]. Rainfall threshold can be used to determine the landslide data set quantitatively, that is to say, landslides may occur when a certain rainfall threshold is reached. Thus we can discard the unavailable or unreliable landslide records [53]. Based on the global landslide disaster database compiled by Froude and Petrey, 3285 out of 5318 non-seismic landslides in the world were related to rainfall from 2004 to 2016 [54–56]. The landslides in Shexian county are also related to rainfall events, and 81.4% of the total landslides occurred in rainy season (from May to July) [4,5]. Rainfall causes an increase in groundwater level, a decrease in effective stress between rock and soil particles, an increase in pore water pressure, and a decrease in the shear strength of a rock and soil mass. As these changes can result in the occurrence of a landslide, rainfall is therefore regarded as the main factor of a landslide occurrence [57]. According to the distribution of atmospheric rainfall in Shexian County, rainfall can be divided into four grades: <1530, 1530–1580, 1580–1630, and >1630 mm (Figure 5e).

### *3.3. Other Factors*

Among other factors, roads were derived from the vectorization of geographical location maps, and normalized vegetation index (NDVI) was calculated based on the extraction of remote sensing images.

Roads have been categorized as the main man-made factor causing landslides [13]. As Shexian County is located in a mountainous area, roads are therefore predominantly constructed along mountains. Engineering activities used in road construction destroy the integrity of the mountain along the road, resulting in an increase in landslide susceptibility. According to the distance from the road, buffer zones of <400, 400–1200, 1200–2000, and >2000 m were set (Figure 6a).

Vegetation acts as a soil anchor via its root system, thereby improving the shear resistance of a soil. At the same time, transpiration of plants can reduce soil moisture to a certain extent [21,47]. NDVI of Shexian was extracted from remote sensing images and divided into five grades: <0.2, 0.2–0.3, 0.3–0.35, 0.35–0.4 and >0.4 (Figure 6b). Our results indicate that NDVI values were higher in areas further away from towns and smaller in areas with concentrated populations [2,4,9].

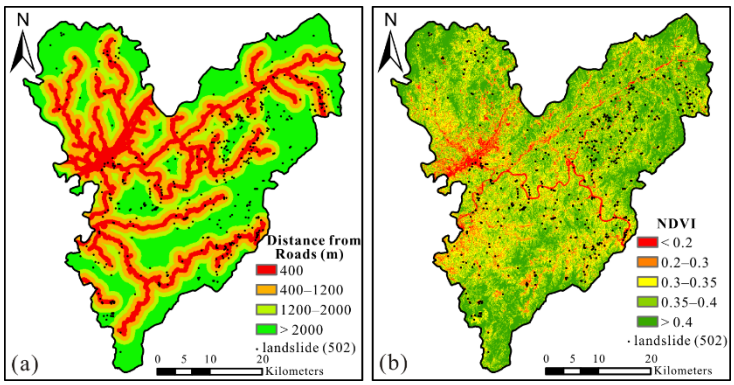


**Figure 6.** Other parameter maps for the study area and landslide location points: (**a**) Distance from Roads; (**b**) normalized vegetation index (NDVI).

## 4. Methodology

Firstly, on the basis of the 16 landslide susceptibility factors, information values for the factors for 502 landslide points and safety points were extracted using the "Extract Multi Value To Point" GIS tool. Of these, 351 landslide points and safety points (70%) were used for training, and 151 landslide points and safety points (30%) were used for verification [31,40,58]. All data (1004 groups) were imported into

Scikit-Learn (Sklearn) repository of Python for training and verification. Data were analyzed using algorithms of LR, SVM, RF, GBM, and MLP. After the model was established, the "Raster To Point" GIS tool was used to convert the raster image of the study area into 260,130 vector points; the information values of the 16 factors for the 260,130 points were extracted using the "Extract Multi Value To Point" tool. Overall, 260,130 groups of data were imported into the established models to calculate the score of landslide susceptibility. Finally, by using the "point to raster " GIS tool, five Landslide Susceptibility Maps (LSMs) for the five methods were obtained. The process of drawing LSMs using a machine learning algorithm is shown in Figure 7.

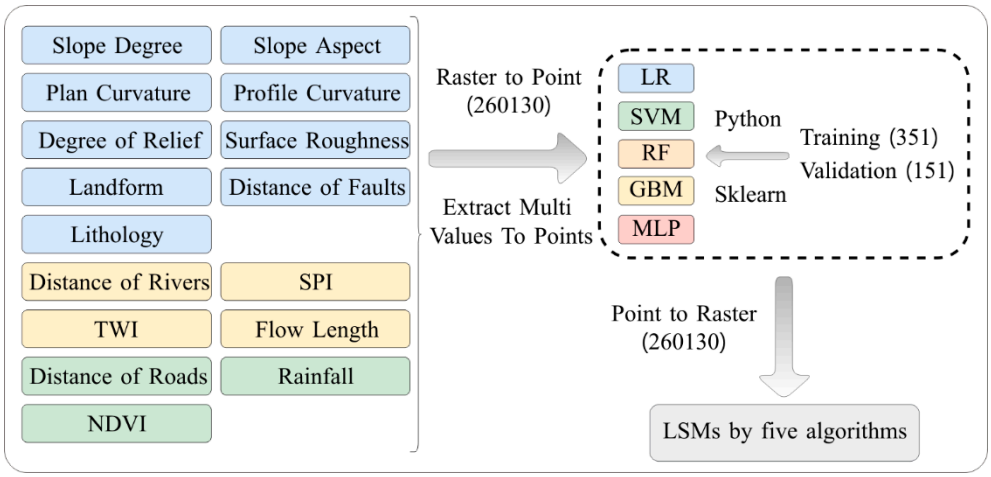


**Figure 7.** A flow chart depicting the process used to create Landslide Susceptibility Maps (LSMs) using machine learning algorithms.

### *4.1. Logistic Regression (LR)*

LR is a generalized linear model [59,60] having many similarities with linear regression. As LR assumes that dependent variable $y$ obeys a Bernoulli distribution (in contrast, linear regression assumes that dependent variable $y$ obeys Gaussian distribution [61]), LR is therefore supported by linear regression theory. However, LR can be used to deal with 0/1 classification problems after introducing the logic function (i.e., sigmoid function) [43].

### *4.2. Support Vector Machine (SVM)*

In order to enhance the generalization ability of the model, an additional optimization objective was added to the linear discriminant method. This method was termed SVM [20,36].

SVM is a linear classifier with a maximum possible safe interval, in which the support vectors are the points on both sides of the safe interval (Figure 8). SVM can be regarded as solving two problems: on the one hand, it finds an appropriate way to measure the correlation between input vectors, i.e., the kernel function $K(x, y)$; on the other hand, it constructs a linear structure by combining the output of training samples with new test samples. The output of training samples is measured by similarity [21,61]. The more similar the input samples, the greater the contribution to the output. As per the original nearest neighbor classifier, it can be approximately expressed as:

$$\sum_{i=1}^{\ell} y_i \lambda_i^* K(x, x_i) \tag{4}$$

where, $l$ is the number of training samples; $y_i$ is the output of training samples; and $x_i$ and $x$ are the new test samples to be classified. When kernel is used to calculate the point product of data points mapped

by function $\varphi(x)$, it is not necessary to calculate the mapping function. The equation can therefore be expressed as:

$$K(x, x_i) = \varphi(x) \cdot \varphi(x_i) \tag{5}$$

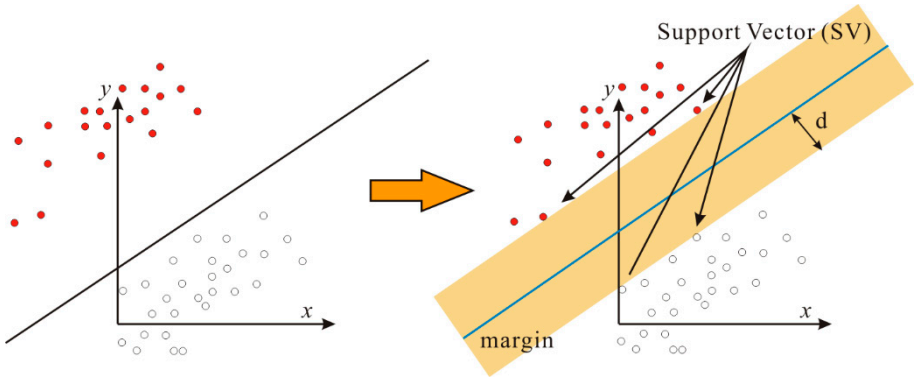


**Figure 8.** Support Vector Machine.

### *4.3. Random Forest (RF)*

The tree-based learning algorithm enables the prediction model to be accurate, stable and easily explained. Different from the linear model, tree-based algorithms can also effectively map the nonlinear relationship. Common tree-based models include decision trees, random forests, and promoted trees [27,28,62,63].

Both classification and regression trees belong to the branch of decision tree. The category predicted by a classification tree is the most common category of observation value of training samples in a certain region, namely the mode response of training observation values. In order to achieve the purpose of classification, the system typically predicts a group of categories and their probability of occurrence. Generally, recursive binary segmentation is used to generate a classification tree. However, as Residual Sum of Squares (RSS) cannot be used as a binary segmentation standard in a classification tree, it is therefore necessary to define the impure quantity (QM) of a leaf node to replace RSS, that is, a method to measure the homogeneity of target variables in the subset region R1, R2,..., RJ [61]. In node m, we can express the frequency of the category of $R_M$ in a region by nm sample observations, and the frequency of the $k$th class training in the m region can be expressed as:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \tag{6}$$

RF decorrelates all trees by random disturbance. The core idea of RF is that it is the same as bagging tree, thus its variance is reduced. In addition, a large number of predictors can be considered for RF, not only because this method reduces bias, but also because local feature predictors play an important role in tree structure. RF can use a large number of predictors, even more than the number of samples observed. The most significant advantage of RF is that it can obtain more information to reduce deviation of the fitting value and estimation segmentation [64,65]. As RF computes enough decision tree models, each predictor has at least a few chances to become a predictor for defining segmentation. In most cases, the leading predictor and the feature predictor have the opportunity to define the segmentation of the dataset.

### *4.4. Gradient Boosting Machine (GBM)*

GBM, an ensemble learning method, combines multiple decision trees to build a more powerful model which can be used for classification or regression [28,61,66]. Different from RF, GBM constructs trees in a continuous way, and each tree tries to correct the error of the previous tree. The idea behind

GBM is to combine multiple weak learners to improve their performance. The main parameters of the GBM tree model include the number of trees and the learning rate.

### *4.5. Multilayer Perceptron (MLP)*

MLP consists of three layers: an input layer, a hidden layer and an output layer (Figure 9). The different layers in MLP are fully connected [25,67,68]. In the classification task, the softmax function is used as the activation function in the output layer of the perceptron to ensure that the output is a probability value, and its sum is equal to 1. The softmax function receives a fractional vector of random real values and converts it into a plurality of vector values between 0 and 1, and the sum of which is 1 [69].

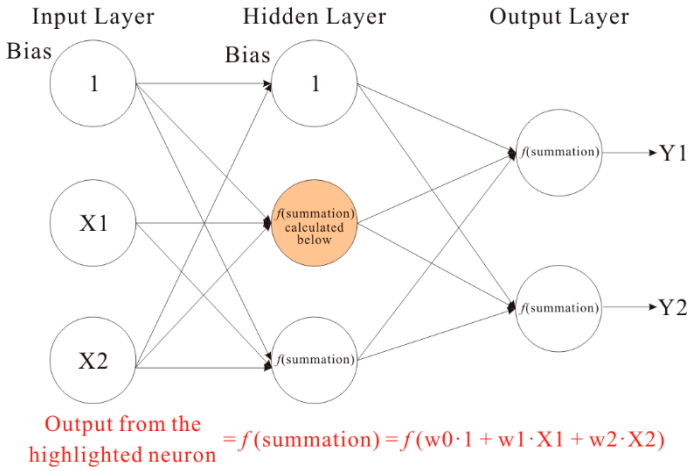


**Figure 9.** Multilayer Perceptron.

## 5. Results and Discussion

Results gained using the five algorithms were imported into GIS, and the Janks natural breakpoint method [70] was used to divide landslide susceptibility into four grades [29]: low, moderate, high, and very high. Using this method, landslide susceptibility zoning maps were generated for each algorithm (Figure 10). Results from this analysis recorded certain similarities. For example, landslide susceptibility was high in the middle and northeastern areas, and low in the south and northwestern areas.

In order to quantify and compare differences between zoning results gained using the five methods, a grid distribution histogram of different grades of landslides was drawn (Figure 11). Here, the proportion of landslide prone areas in the total area (gray bars), and the proportion of known landslide points in each grade area (red bars) are shown. Results from this analysis indicate that higher red columns in high and very high areas coupled with lower gray columns in different landslide grades indicate a higher level of model success and a better fit. When the proportion of landslide points in high or very high areas were divided by the proportion of the grade area, the five algorithms had results of 1.52 (LR), 1.77 (SVM), 1.95 (RF), 1.83 (GBM), and 1.64 (MLP); when the proportion of very high landslide points were directly divided by the proportion of the grade area, results were 1.92, 2.20, 2.98, 2.62, and 2.14, respectively. Based on these results, the five models can therefore be ranked in the order of: RF > SVM > MLP > GBM > LR.

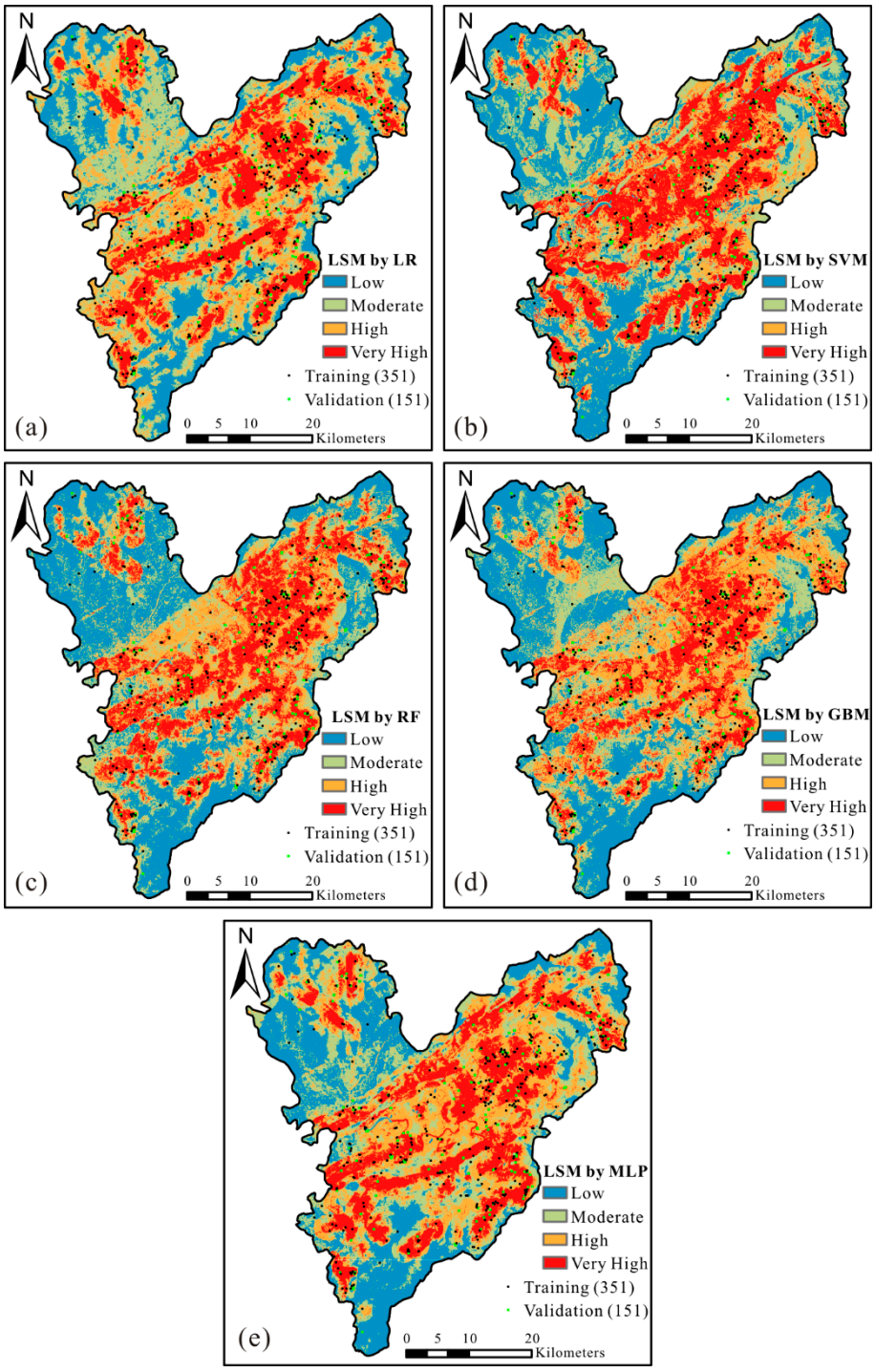


**Figure 10.** Landslide susceptibility maps based on (**a**) Logistics Regression; (**b**) Support Vector Machine; (**c**) Random Forest; (**d**) Gradient Boosting Machine; (**e**) Multilayer Perceptron.

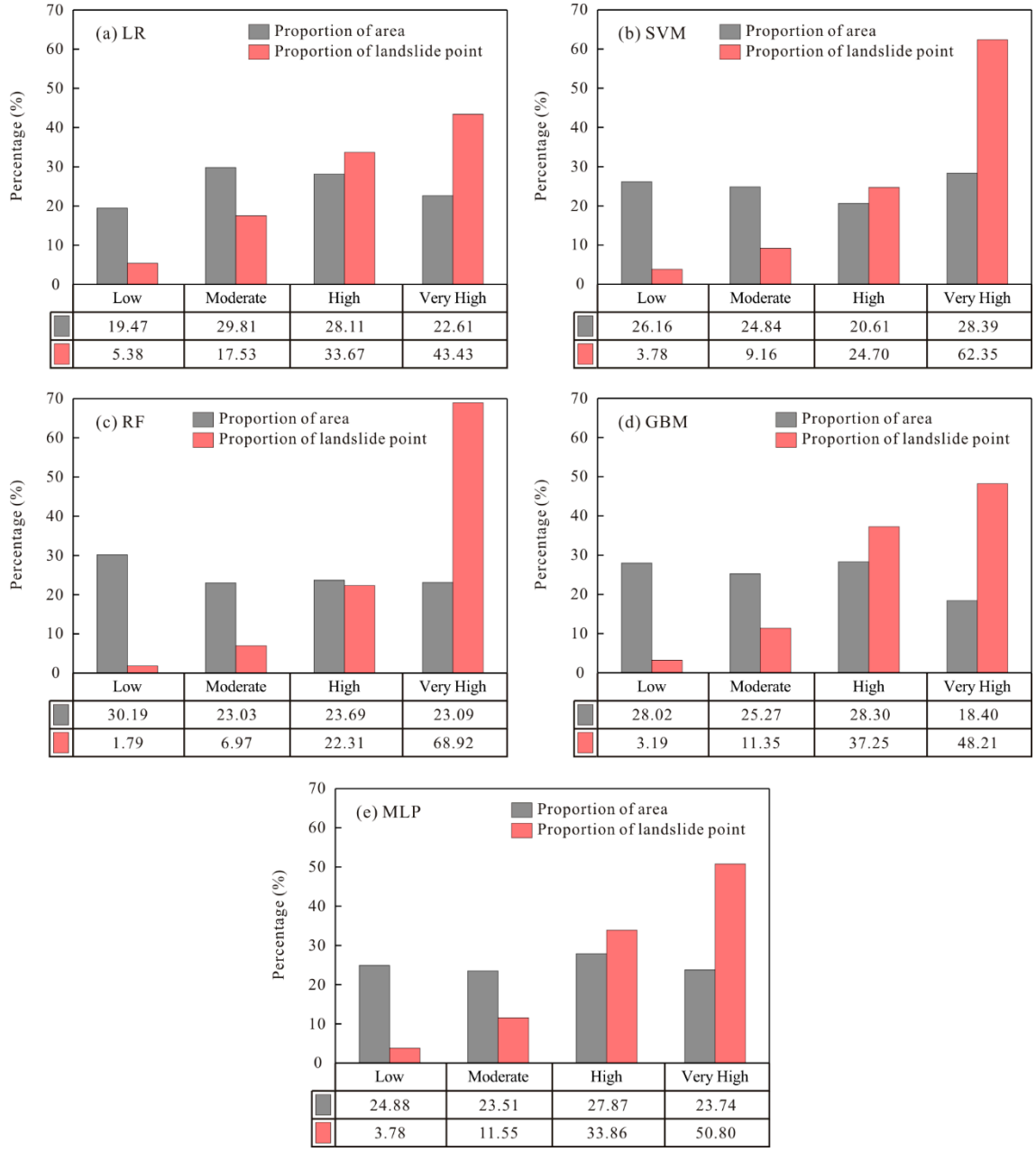


**Figure 11.** Raster distributions of landslides of different classes.

The accuracy of the landslide susceptibility test can be shown using a Receiver Operate Curve (ROC) [18]. Here, the *x*-axis has a false positive rate, i.e., 1-specificity, indicating the probability of non-disaster points being mis-predicted; the *y*-axis is a true positive rate, that is, susceptibility, indicating the probability of correct prediction of disaster points [20,40,71]. Area Under Curve (AUC) was used to analyze the accuracy of the prediction results. AUC represents the area under the curve enclosed by the coordinate axis. The closer the AUC value is to 1, the more accurate the prediction results of the model are [29,31,72]. Python was used in this analysis to derive the success rate curve of training samples (Figure 12) and prediction rate curves of test samples (Figure 13). Results from these analyses indicate that the success rates of LR, SVM, RF, GBM, and MLP models under training samples were 0.781, 0.824, 0.853, 0.828, and 0.811, respectively (Figure 12), and the prediction rates of test samples were 0.772, 0.803, 0.821, 0.815, and 0.803, respectively (Figure 13). These results indicate that the probability of the RF model was relatively high. Combined with the success rate and prediction

rate, we can see that the accuracy of the five algorithms was in the order of: RF > GBM > SVM > MLP > LR. Accuracy ranking was generally consistent with results gained using the histogram of raster distribution (Figure 11). In the five machine learning algorithms involved in the landslide susceptibility mapping of Shexian County, RF had the best effect, followed by SVM and GBM; MLP and LR had the lowest level of accuracy. Since RF is an integrated model, its generalization ability, anti-interference ability, and fitting ability are all stronger than models using a single factor. Because the landslide factors of Shexian county are derived from 16 factors, the relationship between the factors is complex. In response to this practical problem, the RF integrated model achieved good results.

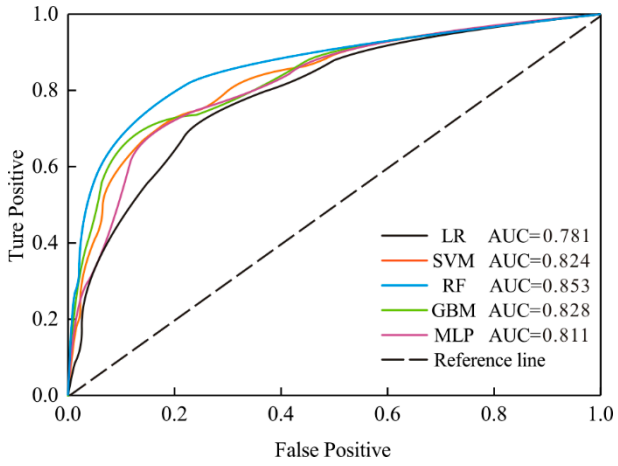


**Figure 12.** Success rate curve.

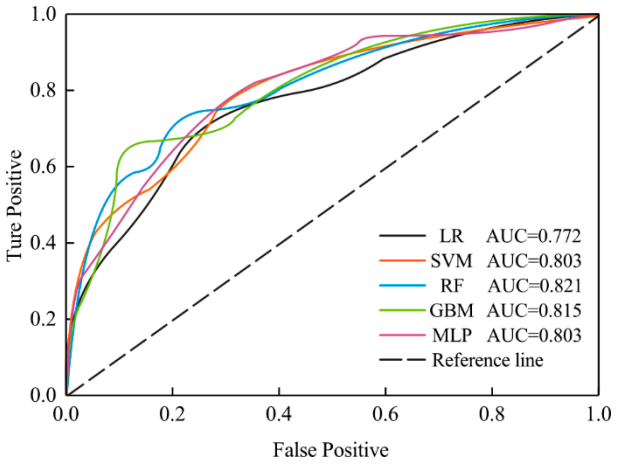


**Figure 13.** Prediction rate curve.

## 6. Conclusions

In this study, GIS and Sklearn were used in conjunction with five machine learning algorithms (LR, SVM, RF, GBM, and MLP) to analyze slope aspect, slope, plane curvature, profile curvature, terrain emergence degree, surface roughness, landform, fault, lithology, river buffer zone, net flow intensity index, and land surface roughness. On the basis of 16 condition factors, including shape humidity index, flow length, rainfall, road buffer and NDVI, 502 landslide and safety points were analyzed using the five algorithms. Overall, 70% of the landslide points were used as training data and 30% were used for verification. After establishing models for the five algorithms, 260 and 130 grids for the study region were converted into points. These points were then imported into the models for calculation. Finally, landslide susceptibility maps based on the five algorithms were created. According to the degree of susceptibility, landslides were divided into four grades: low, moderate, high, and very high.

At the same time, the distribution histogram of grid and landslide points in different grades, as well as ROC and AUC, were used to compare the effect of these algorithms using landslide susceptibility maps. Results indicated that the five models were relatively successful in predicting landslide susceptibility occurrence. The ratio of high or very high landslide points to grade area defined by LR, SVM, RF, GBM, and MLP was 1.52, 1.77, 1.95, 1.83, and 1.64, and the ratio of very high landslide points to grade area was 1.92, 2.20, 2.98, 2.62, and 2.14, respectively. The success rates of training samples were 0.781, 0.824, 0.853, 0.828, and 0.811, and the prediction rates of test samples were 0.772, 0.803, 0.821, 0.815, and 0.803, respectively. The success rate and prediction rate of the other five algorithms were greater than 0.8, apart from LR which was slightly lower than 0.8. By ordering the five algorithms from good to bad (RF > SVM > MLP > GBM > LR), our results indicated that RF had the best landslide susceptibility evaluation. By combining machine learning algorithms with GIS to map landslide susceptibility and evaluate susceptibility, results from this investigation provide a greater level of information for relevant staff. The method presented in this study is not only suitable for Shexian County, it can also be expanded to include other mountainous areas in the Southern Anhui Province.

**Author Contributions:** Methodology, Q.L.; Software, Z.W.; Supervision, Q.L.; Validation, Y.L.; Visualization, Z.W. and Y.L.; Writing—original draft preparation, Z.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Natural Science Foundation of Anhui Province under grant number [1908085ME145].

**Acknowledgments:** Thanks to the anonymous reviewers for their valuable feedback on the manuscript.

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

## References

1. Lee, S.; Oh, H.J. Ensemble-Based Landslide Susceptibility Maps in Jinbu Area, Korea. In *Terrigenous Mass Movements*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 193–220.
2. Wang, Q.; Li, W.; Wu, Y.; Pei, Y.; Xing, M.; Yang, D. A comparative study on the landslide susceptibility mapping using evidential belief function and weights of evidence models. *J. Earth Syst. Sci.* **2016**, *125*, 645–662. [CrossRef]
3. He, H.; Hu, D.; Sun, Q.; Zhu, L.; Liu, Y. A landslide susceptibility assessment method based on GIS technology and an AHP-weighted information content method: A case study of southern Anhui, China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 266. [CrossRef]
4. Liao, Y. Study on Division of Geological Disasters Susceptibility and Meteorological Forecasting and Warning of She County Anhui Province. Ph.D. Thesis, Chengdu University of Technology, Chengdu, China, 2015.
5. Pan, G. Study on Landslide Distribution, Failure Mechanism and Monitoring in Shexian County of Southern Anhui Province. Ph.D. Thesis, Hefei University of Technology, Hefei, China, 2015.
6. Banerjee, P.; Ghose, M.K.; Pradhan, R. Analytic hierarchy process and information value method-based landslide susceptibility mapping and vehicle vulnerability assessment along a highway in Sikkim Himalaya. *Arab. J. Geosci.* **2018**, *11*, 139. [CrossRef]
7. El Jazouli, A.; Barakat, A.; Khellouk, R. GIS-multicriteria evaluation using AHP for landslide susceptibility mapping in Oum Er Rbia high basin (Morocco). *Geoenviron. Disasters* **2019**, *6*, 3. [CrossRef]
8. Hepdeniz, K. Using the analytic hierarchy process and frequency ratio methods for landslide susceptibility mapping in Isparta-Antalya highway (D-685), Turkey. *Arab. J. Geosci.* **2020**, *13*, 1–16. [CrossRef]
9. Liu, H.; Li, X.; Meng, T.; Liu, Y. Susceptibility mapping of damming landslide based on slope unit using frequency ratio model. *Arab. J. Geosci.* **2020**, *13*, 1–19. [CrossRef]
10. Senanayake, S.; Pradhan, B.; Huete, A.; Brennan, J. Assessing Soil Erosion Hazards Using Land-Use Change and Landslide Frequency Ratio Method: A Case Study of Sabaragamuwa Province, Sri Lanka. *Remote Sens.* **2020**, *12*, 1483. [CrossRef]
11. Mondal, S.; Mandal, S. Landslide susceptibility mapping of Darjeeling Himalaya, India using index of entropy (IOE) model. *Appl. Geomat.* **2019**, *11*, 129–146. [CrossRef]
12. Shirani, K.; Pasandi, M.; Arabameri, A. Landslide susceptibility assessment by dempster–shafer and index of entropy models, Sarkhoun basin, southwestern Iran. *Nat. Hazards* **2018**, *93*, 1379–1418. [CrossRef]

13. Wang, Q.; Li, W.; Yan, S.; Wu, Y.; Pei, Y. GIS based frequency ratio and index of entropy models to landslide susceptibility mapping (Daguan, China). *Environ. Earth Sci.* **2016**, *75*. [CrossRef]
14. Gadtaula, A.; Dhakal, S. Landslide susceptibility mapping using Weight of Evidence Method in Haku, Rasuwa District, Nepal. *J. Nepal Geol. Soc.* **2019**, *58*, 163–171. [CrossRef]
15. Kumar, R.; Anbalagan, R. Landslide susceptibility mapping of the Tehri reservoir rim area using the weights of evidence method. *J. Earth Syst. Sci.* **2019**, *128*, 153. [CrossRef]
16. Sifa, S.F.; Mahmud, T.; Tarin, M.A.; Haque, D.M.E. Event-based landslide susceptibility mapping using weights of evidence (WoE) and modified frequency ratio (MFR) model: A case study of Rangamati district in Bangladesh. *Geol. Ecol. Landsc.* **2019**, 1–14. [CrossRef]
17. Chen, X.; Chen, W. GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. *CATENA* **2021**, *196*, 104833. [CrossRef]
18. Reichenbach, P.; Rossi, M.; Malamud, B.D.; Mihir, M.; Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **2018**, *180*, 60–91. [CrossRef]
19. Regmi, N.R.; Giardino, J.R.; McDonald, E.V.; Vitek, J.D. A comparison of logistic regression-based models of susceptibility to landslides in western Colorado, USA. *Landslides* **2014**, *11*, 247–262. [CrossRef]
20. Shan, Y.; Chen, S.; Zhong, Q. Rapid prediction of landslide dam stability using the logistic regression method. *Landslides* **2020**, *17*, 2931–2956. [CrossRef]
21. Pandey, V.K.; Pourghasemi, H.R.; Sharma, M.C. Landslide susceptibility mapping using maximum entropy and support vector machine models along the Highway Corridor, Garhwal Himalaya. *Geocarto Int.* **2020**, *35*, 168–187. [CrossRef]
22. Pourghasemi, H.R.; Jirandeh, A.G.; Pradhan, B.; Xu, C.; Gokceoglu, C. Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. *J. Earth Syst. Sci.* **2013**, *122*, 349–369. [CrossRef]
23. Harmouzi, H.; Nefeslioglu, H.A.; Rouai, M.; Sezer, E.A.; Dekayir, A.; Gokceoglu, C. Landslide susceptibility mapping of the Mediterranean coastal zone of Morocco between Oued Laou and El Jebha using artificial neural networks (ANN). *Arab. J. Geosci.* **2019**, *12*, 696. [CrossRef]
24. Sameen, M.I.; Pradhan, B.; Lee, S. Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment. *Catena* **2020**, *186*, 104249. [CrossRef]
25. Shahri, A.A.; Spross, J.; Johansson, F.; Larsson, S. Landslide susceptibility hazard map in southwest Sweden using artificial neural network. *Catena* **2019**, *183*, 104225. [CrossRef]
26. Niu, F.; Chen, L. Forecasting of Landslide Stability Based on Gradient Boosting Decision Tree Model. *Int. Core J. Eng.* **2019**, *5*, 42–48.
27. Wu, Y.; Ke, Y.; Chen, Z.; Liang, S.; Zhao, H.; Hong, H. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *Catena* **2020**, *187*, 104396. [CrossRef]
28. Chen, W.; Fan, L.; Li, C.; Pham, B.T. Spatial prediction of landslides using hybrid integration of artificial intelligence algorithms with frequency ratio and index of entropy in nanzheng county, china. *Appl. Sci.* **2020**, *10*, 29. [CrossRef]
29. Jaafari, A.; Najafi, A.; Pourghasemi, H.; Rezaeian, J.; Sattarian, A. GIS-based frequency ratio and index of entropy models for landslide susceptibility assessment in the Caspian forest, northern Iran. *Int. J. Environ. Sci. Technol.* **2014**, *11*, 909–926. [CrossRef]
30. Li, R.; Wang, N. Landslide susceptibility mapping for the Muchuan county (China): A comparison between bivariate statistical models (woe, ebf, and ioe) and their ensembles with logistic regression. *Symmetry* **2019**, *11*, 762. [CrossRef]
31. Wang, Q.; Li, W.; Chen, W.; Bai, H. GIS-based assessment of landslide susceptibility using certainty factor and index of entropy models for the Qianyang County of Baoji city, China. *J. Earth Syst. Sci.* **2015**, *124*, 1399–1415. [CrossRef]
32. Dikshit, A.; Pradhan, B.; Alamri, A.M. Pathways and challenges of the application of artificial intelligence to geohazards modelling. *Gondwana Res.* **2020**. [CrossRef]
33. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D.T.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
34. Yu, X. Study on the Landslide Susceptibility Evaluation Method Based on Multi-Source Data and Multi-Scale Analysis. Ph.D. Thesis, China University, Wuhan, China, 2016.

35. 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.
36. Feizizadeh, B.; Roodposhti, M.S.; Blaschke, T.; Aryal, J. Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. *Arab. J. Geosci.* **2017**, *10*, 1–13. [CrossRef]
37. Van Den Eeckhaut, M.; Vanwalleghem, T.; Poesen, J.; Govers, G.; Verstraeten, G.; Vandekerckhove, L. Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). *Geomorphology* **2006**, *76*, 392–410. [CrossRef]
38. Ohlmacher, C.G. Plan curvature and landslide probability in regions dominated by earth flows and earth slides. *Eng. Geol.* **2007**, *91*, 117–134. [CrossRef]
39. Wang, Z.; Hu, Z.; Liu, H.; Gong, H.; Zhao, W.; Yu, M.; Zhang, M. Application of the relief degree of land surface in landslide disasters susceptibility assessment in China. *Geoinformatics* **2010**, 1–5. [CrossRef]
40. Zhang, J.; Yin, K.; Wang, J.; Liu, L.; Huang, F. Evaluation of landslide susceptibility for Wanzhou district of Three Gorges Reservoir. *Chin. J. Rock Mech. Eng.* **2016**, *35*, 284–296. (In Chinese)
41. Cristinicu, I. Frequency ratio and GIS-based evaluation of landslide susceptibility applied to cultural heritage assessment. *J. Cult. Herit.* **2017**, *28*, 172–176. [CrossRef]
42. Chen, C.-W.; Wei, L.-W.; Lin, G.-W.; Iida, T.; Yamada, R. Evaluating the susceptibility of landslide landforms in Japan using slope stability analysis: A case study of the 2016 Kumamoto earthquake. *Landslides* **2017**, *14*, 1793–1801. [CrossRef]
43. Erener, A.; Mutlu, A.; Duzgun, S. A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM). *Eng. Geol.* **2016**, *203*, 45–55. [CrossRef]
44. Dai, C.F.; Lee, F.C.; Li, J.; Xu, W.Z. Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. *Environ. Earth Sci.* **2001**, *40*, 381–391.
45. Park, S.; Choi, C.; Kim, B.; Kim, J. Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area, Korea. *Environ. Earth Sci.* **2012**, *68*, 1443–1464. [CrossRef]
46. Yalcin, A. GIS-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in Ardesen (Turkey): Comparisons of results and confirmations. *Catena* **2008**, *72*, 1–12. [CrossRef]
47. Gayen, A.; Saha, S.; Pourghasemi, H.R. Soil erosion assessment using RUSLE model and its validation by FR probability model. *Geocarto Int.* **2020**, *35*, 1750–1768. [CrossRef]
48. Moore, D.I.; Grayson, B.R.; Ladson, R.A. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **1991**, *5*, 3–30. [CrossRef]
49. Mind'je, R.; Li, L.; Nsengiyumva, J.B.; Mupenzi, C.; Nyesheja, E.M.; Kayumba, P.M.; Gasirabo, A.; Hakorimana, E. Landslide susceptibility and influencing factors analysis in Rwanda. *Environ. Dev. Sustain.* **2019**, *22*, 7985–8012. [CrossRef]
50. Tran, Q.C.; Minh, D.D.; Jaafari, A.; Al-Ansari, N.; Minh, D.D.; Van, D.T.; Nguyen, D.A.; Tran, T.H.; Ho, L.S.; Nguyen, D.H.; et al. Novel Ensemble Landslide Predictive Models Based on the Hyperpipes Algorithm: A Case Study in the Nam Dam Commune, Vietnam. *Appl. Sci.* **2020**, *10*, 3710. [CrossRef]
51. Yilmaz, I.; Keskin, I. GIS based statistical and physical approaches to landslide susceptibility mapping (Sebinkarahisar, Turkey). *Bull. Eng. Geol. Environ.* **2009**, *68*, 459–471. [CrossRef]
52. Dong, J.-J.; Tung, Y.-H.; Chen, C.-C.; Liao, J.-J.; Pan, Y.-W. Discriminant analysis of the geomorphic characteristics and stability of landslide dams. *Geomorphology* **2009**, *110*, 162–171. [CrossRef]
53. Gariano, S.L.; Sarkar, R.; Dikshit, A.; Dorji, K.; Brunetti, M.T.; Peruccacci, S.; Melillo, M. Automatic calculation of rainfall thresholds for landslide occurrence in Chukha Dzongkhag, Bhutan. *Bull. Eng. Geol. Environ.* **2019**, *78*, 4325–4332. [CrossRef]
54. Froude, M.J.; Petley, D.N. Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2161–2181. [CrossRef]
55. Dikshit, A.; Sarkar, R.; Pradhan, B.; Segoni, S.; Alamri, A.M. Rainfall Induced Landslide Studies in Indian Himalayan Region: A Critical Review. *Appl. Sci.* **2020**, *10*, 2466. [CrossRef]
56. Dikshit, A.; Sarkar, R.; Pradhan, B.; Acharya, S.; Alamri, A.M. Spatial Landslide Risk Assessment at Phuentsholing, Bhutan. *Geosciences* **2020**, *10*, 131. [CrossRef]

57. Y Bui, T.D.; Bui, T.D.; Nguyen, P.Q.; Hoang, N.-D.; Klempe, H. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. *Landslides* **2017**, *14*, 1–17.
58. Xianyu, Y.; Yi, W.; Ruiqing, N.; Youjian, H. A Combination of Geographically Weighted Regression, Particle Swarm Optimization and Support Vector Machine for Landslide Susceptibility Mapping: A Case Study at Wanzhou in the Three Gorges Area, China. *Int. J. Environ. Res. Public Health* **2016**, *13*, 487. [CrossRef]
59. Das, G.; Lepcha, K. Application of logistic regression (LR) and frequency ratio (FR) models for landslide susceptibility mapping in Relli Khola river basin of Darjeeling Himalaya, India. *SN Appl. Sci.* **2019**, *1*, 1453. [CrossRef]
60. Hosmer, W.D.; Stanley, L. Applied logistic regression. *Contemp. Sociol.* **2000**. [CrossRef]
61. Battiti, R.; Brunato, M. *Machine Learning Plus Intelligent Optimization*; Lionsolver Inc.: Los Angeles, CA, USA, 2013.
62. Lee, S.; Lee, M.-J.; Lee, S. Spatial prediction of urban landslide susceptibility based on topographic factors using boosted trees. *Environ. Earth Sci.* **2018**, *77*, 656. [CrossRef]
63. Wang, Y.; Sun, D.; Wen, H.; Zhang, H.; Zhang, F. Comparison of Random Forest Model and Frequency Ratio Model for Landslide Susceptibility Mapping (LSM) in Yunyang County (Chongqing, China). *Int. J. Environ. Res. Public Health* **2020**, *17*, 4206. [CrossRef]
64. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
65. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
66. Sahin, K.E. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2020**, *2*, 1–17. [CrossRef]
67. Pham, B.T.; Tien Bui, D.; Pourghasemi, H.R.; Indra, P.; Dholakia, M.B. Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: A comparison study of prediction capability of nave bayes, multilayer perceptron neural networks, and functional trees methods. *Theor. Appl. Climatol.* **2015**, *122*, 1–19. [CrossRef]
68. Zare, M.; Pourghasemi, H.R.; Vafakhah, M.; Pradhan, B. Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: A comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. *Arab. J. Geosci.* **2013**, *6*, 2873–2888. [CrossRef]
69. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
70. Chen, J.; Yang, S.; Li, H.; Zhang, B.; Lv, J. Research on geographical environment unit division based on the method of natural breaks (Jenks). *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *3*, 47–50. [CrossRef]
71. Hanley, A.J.; McNeil, J.B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
72. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2005**, *27*, 861–874. [CrossRef]