*Article*

# Enhancing Seismic Landslide Susceptibility Analysis for Sustainable Disaster Risk Management through Machine Learning

Hailang He [1,2,3], Weiwei Wang [2,4], Zhengxing Wang [2], Shu Li [2,5] and Jianguo Chen [1,*]

1   School of Safety Science and Engineering, Tsinghua University, Beijing 100084, China
2   Hefei Institute for Public Safety Research, Tsinghua University, Hefei 230601, China;
    wang0weiwei@163.com (W.W.); wangzhengxing@tsinghua-hf.edu.cn (Z.W.); lishu@tsinghua-hf.edu.cn (S.L.)
3   Defense Engineering Institute, AMS, PLA, Beijing 100036, China
4   Primus Line (China) Ltd., Shanghai 201108, China
5   Anhui Key Laboratory of Bridge and Tunnel Engineering Testing, Hefei 230051, China
*   Correspondence: chenjianguo@tsinghua.edu.cn

**Abstract:** The accuracy of Seismic Landslide Susceptibility Maps (SLSMs) is imperative for the prevention of seismic landslide disasters. This study enhances the precision of SLSMs by integrating nine distinct machine learning methodologies with the GeoDetector version 0.0.4 tool to filter both numerical and physical factors contributing to landslide susceptibility. The dataset comprises 2317 landslide instances triggered by the 2013 Minxian Ms = 6.6 earthquake, from which redundant factors were pruned using the Recursive Feature Elimination technique. Subsequent evaluations of the optimized factors, both individually and in combination, were conducted through Frequency Ratio analysis and Factor Interaction assessment. The study juxtaposes the Area Under the Receiver Operating Characteristic Curve (AUC) and the accuracy of nine machine learning models before and after factor optimization. The findings indicate an increase in AUC from a maximum of 0.989 to 0.992 in the Random Forest model, and an 8.37% increase in AUC for the SVM model, signifying a notable enhancement in the stability across all models. The establishment of the SLSM notably elevated the frequency ratio in high-risk zones from 50.40 to 85.14, underscoring the efficacy of combining machine learning and detector optimization techniques in sustainable practices. This research proposes a universal framework aimed at eliminating redundancy and noise in SLSMs and hazard risk assessments, thereby facilitating sustainable geological disaster risk management.

**Keywords:** machine learning; GeoDetector; seismic landslide; sustainability

## 1. Introduction

Landslides represent a severe disaster globally, inflicting catastrophic socio-economic losses on communities residing in mountainous regions [1,2]. Moreover, the environmental issues associated with these phenomena are myriad [3]. In recent years, the frequency of landslides and slope instabilities has sharply increased, primarily due to the cumulative effects of rapidly changing climate conditions and heightened anthropogenic disturbances [4,5]. Landslide disasters are frequent in China, among which seismic landslides pose the most severe threats [6]. To more effectively address these disasters and promote sustainable development, the creation of Seismic Landslide Susceptibility Maps (SLSMs) is imperative [7]. SLSMs delineate the joint impact of multiple factors and are susceptible to variations, hence enhancing their accuracy is crucial for regional seismic hazard risk assessments. Susceptibility mapping is of paramount importance in understanding the likelihood of landslides occurring in any critical area [8]. Reliable and robust susceptibility maps will aid in informed planning and settlement in landslide-prone areas.

Earlier scholars devised numerous methods to develop landslide susceptibility models that can be categorized as physical-based, opinion-driven, and statistical models [9]. However, with the advent of faster and more economical computing tools, the use of machine learning techniques in the estimation of landslide hazard probability has become a feasible option [10]. Researchers found that the boundary between statistical and machine learning methods is quite narrow and forms a matter of discussion among geoscientists and geo-engineers [11]. Also, the machine learning models are devised simply based on a pattern in extracted datasets, while the statistical models have a certain set of mathematical rules to follow. Additionally, the optimization algorithms further empower the performance of ML models, and hence supersede statistical methods in terms of accuracy.

Seismic landslide, as a nonlinear problem, is challenging to be resolved accurately by conventional statistical methods [12], while the toolset of machine learning methods has powerful data processing capability and various types of models, which are widely used in seismic landslide research [13], which contains logistic regression model [14–17], random forest [18–20], support vector machine [21,22], artificial neural networks [23], Bayesian classification [24], and others. However, because machine learning models depend mainly upon databases, data training may have appeared as a disadvantage, such as overfitting or underfitting and not allowing to extract useful interpretation [25,26]. Therefore, the selection of crucial evaluation factors and reduction in redundant factors in the dataset can not only resolve the fitting problem of machine learning but also reduce the computational burden and improve the efficiency of the model. To enhance the model's accuracy and address potential overfitting issues in machine learning-based models, methodologies such as frequency ratio [27], deterministic factor [28], Pearson correlation coefficient [29], factor analysis [30], rough set [31], information gain [32], and recursive feature elimination [33] were employed.

This paper introduces a geographic detector (GeoDetector version 0.0.4) for the combined impact of screening factors. The geographic detector (GeoDetector) method considers the spatial and physical nature of geographic data concurrently and can further detect spatial disparity [34]. The Geographic Detector investigates and reveals the driving forces behind spatial characteristics by systematically analyzing quantitative types and incorporating the examination of physical phenomena and processes. In recent years, the GeoDetector method has been combined with various approaches as an environmental assessment model, such as integrating GeoDetector with methods like geographic weighted regression, for the analysis of the potential migratory habitability of villages in coal mining areas on the Loess Plateau [35,36]; GeoDetector combined with random forests, and Bayesian networks can effectively eliminate redundant factors and improve the accuracy of evaluation models [37]. However, whether the factors screened by GeoDetector can be applied to various machine learning models has not been researched explicitly in the literature so far.

In this study, we extend the research by integrating GeoDetector with multiple machine-learning models to investigate its applicability in the spatial discretization of seismic landslides. Utilizing the dataset from the 2013 Minxian earthquake-induced landslide region in Gansu Province, China, this paper explores the synergy between geographic detectors (GeoDetector) and nine machine learning models. We refine the evaluation of factors by employing a recursive methodology to discard those with lower importance scores, and examine the impact of these optimized factors on the SLSM. This involves analyzing the variability of the machine learning model's performance before and after the optimization process. By developing reliable and robust susceptibility maps, this research underpins informed decision-making for settlement and planning in mountainous areas. Such efforts not only contribute to the sustainable management of mountain resources but also aim to mitigate the socio-economic and environmental losses caused by landslide disasters, thus supporting sustainable development trajectories in mountainous regions.

## 2. Research Area and Data Collection

### 2.1. Research Area

The Minxian Ms = 6.6 earthquake occurred on 22 July 2013. It was caused by the uneven segmental activity of the Lintan-Tangchang fault zone in the region [38]. Its seismic faults are predominantly retrogradation and accompanied by a left-slip nature. The earthquake caused 95 deaths, 14 of which were due to seismic landslides, accounting for about 15% of the deaths. According to scholars [39], the distribution of landslide patches in the 2013 Minxian earthquake (Figure 1) demonstrates a significant variation in topography and lithology. The elevation range of the research area is between 2208 and 3339 m; the slope range varies from 0° to 62°. The underlying geology predominantly comprises conglomerate and sandstone of the Permian and Paleocene ages, which significantly influences the occurrence and distribution of landslides. The landslides are mainly small loess cliff avalanches, slides, and dips, triggering some deep coherent soil landslides. The distribution of landslides depicts that they are mainly distributed in the middle of the mountain as well as denser gullies, and no landslides occur in the low terrain. Thus, there is an apparent spatial differentiation is concluded. Therefore, the Minxian earthquake landslide is not only the result of multiple evaluation factors but also is under the influence of spatial characteristics, which meet the requirements of the research area that is investigated by multi-model combinations to screen influential factors.
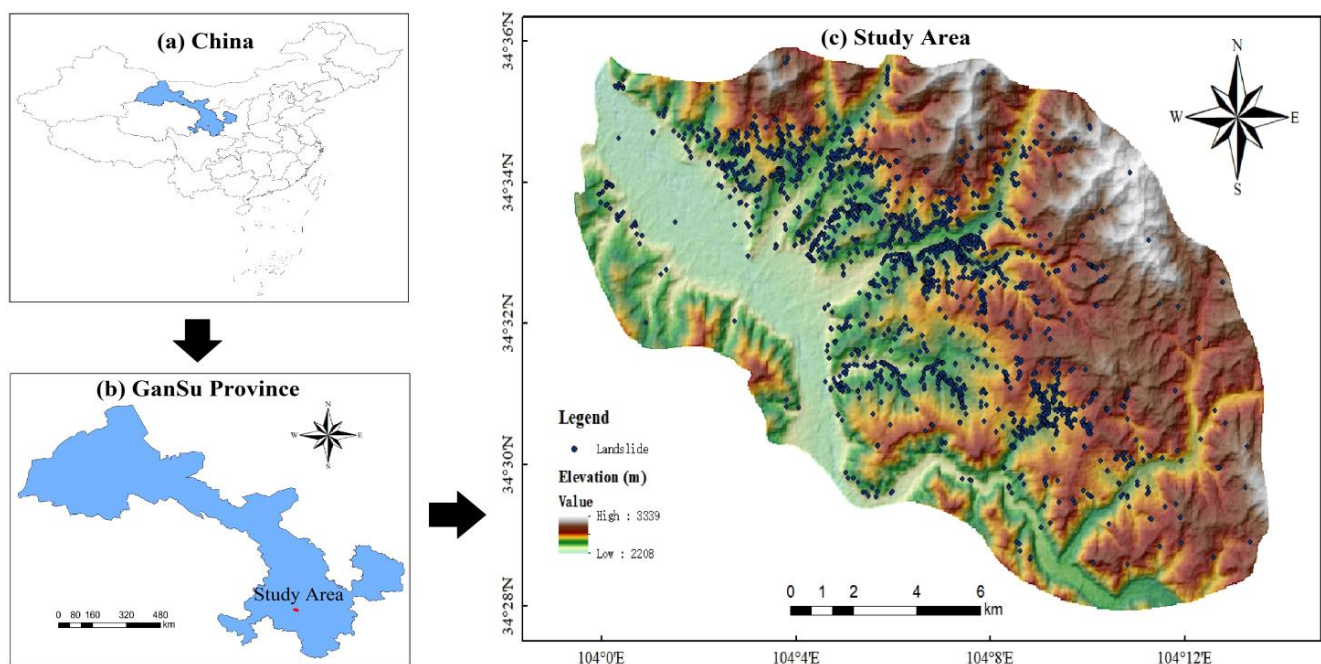


**Figure 1.** Location and landslide distribution of the research area: (**a**) Location of China; (**b**) Gansu Province Location; (**c**) Research Area.

### 2.2. Data

Investigating these unfortunate large-scale movements necessitates a comprehensive evaluation of land use, rainfall patterns, snowmelt, seismic activity, geology, geomorphology, and physical–mechanical properties, in addition to having a well-documented scientific inventory of landslides. The widespread adoption of Geographic Information Systems (GIS) effectively aids in identifying the geographical locations of previous slope instabilities and meticulously marking the Landslide Conditioning Factors (LCFs) [40]. In the literature, attributes such as slope angle, aspect, rainfall, lithology, land use, proximity to fault zones, slope height, temperature, elevation, and vegetation types are referred to as Landslide Conditioning Factors (LCFs), and are crucial for devising sustainable mountain planning and risk management [41,42]. Moreover, through the use of advanced tools and

technologies, LCFs will overlay susceptibility models, and their accuracy can be critically examined through statistical methods to ensure robustness and relevance.

This research comprehensively evaluates landslide susceptibility by analyzing 2317 landslide events and 20 key evaluation factors grouped into five primary categories based on an extensive literature review and their documented impact on landslide triggers. Topographic Factors: This category includes elevation, slope, aspect, relief degree of the land surface (RDLS), and indices such as terrain ruggedness (TRI), wetness (TWI), stream power (SPI), and sediment transport (STI), along with various curvatures. These factors are integral for assessing the physical terrain's susceptibility to landslides [1]. Geological Factors: Factors like lithology and proximity to faults are included due to their significant roles in influencing subsurface stability and landslide vulnerability. Environmental Factors: This group comprises the normalized vegetation index (NDVI), distance from rivers, and land cover, which are essential for evaluating environmental conditions affecting slope stability. Human Engineering Activities: The density of human activity points of interest (POI) and proximity to roads reflect human modifications to the natural landscape, assessing their potential to exacerbate landslide risks. Seismic Factors: Seismic parameters such as peak ground acceleration (PGA) are considered for their impact on landslides, especially in earthquake-prone areas.

This set of evaluation factors summarizes all the evaluation indexes that are used, and the data of the topographic factors are mainly based on the numerical elevation model, which is based on the ASTER satellite 30 m data downloaded from the geospatial data cloud (https://www.gscloud.cn/) (accessed on 12 February 2024) [43]; the data of the geological factors are downloaded from China Geological Archives (https://www.ngac.cn/) (accessed on 2 February 2024) [44]; the NDVI is Landsat 8 satellite data downloaded from Geospatial Data Cloud (https://www.gscloud.cn) (accessed on 8 January 2024) [45]; the data called the distance from rivers is 1:200,000 scale downloaded from the Resource and Environmental Science Data Center of the Chinese Academy of Sciences (http://www.resdc.cn/) (accessed on 17 December 2023) [46]; the data of the land cover data are downloaded from GLOBELAND30 (http://www.globallandcover.com/) (accessed on 22 February 2024) [47], which downloads 30 m global land cover data; POI data from Google Earth 7.1.8, then ArcGIS 10.2 is utilized to generate POI kernel density.

## 3. Methodology

### 3.1. GeoDetector

GeoDetector is a statistical method designed to detect the spatial divergence of research targets and elucidate the factor drivers of their distributions [34,37]. It encompasses four primary functionalities: divergence and factor detection, interaction detection, risk area detection, and ecological detection. This method is employed dichotomously to examine the relationship between independent variables and dependent variables, specifically evaluating whether the spatial distributions of independent variables and landslides are correlated, which helps determine the significant influence of the independent variables on landslide occurrences.

$$q = 1 - SSW/SST.$$

$$SSW = \sum_{m}^{L} N_m \sigma_m^2. \tag{1}$$

$$SST = N\sigma^2.$$

Relation (1) details this relationship, where $q$ represents the explanatory power of the independent variable concerning the dependent variable, indicating the degree to which the independent variable explains variance in the dependent variable. The variable $q$ is defined on a scale from 0 to 1, where higher values signify greater explanatory importance. Furthermore, $SSW$. denotes the sum of variances within the strata of the evaluation factor, while $SST$ represents the total variance within the global factor evaluation. The indice $m = 1, \cdots, L$. refers to the stratification, classification, and partitioning of variable $Y$ (e.g.,

landslide susceptibility) or factor $X$ (e.g., rainfall, geological conditions), with $N_m$ and $N$ are the numbers of strata and the number of cells in the whole region, and $\sigma$ are the variance of $Y$ values in stratum m and the whole region.

It is essential to note that while GeoDetector effectively identifies spatial associations between factors and landslide occurrences, the reliability of predictions largely depends on the comprehensiveness and accuracy of the input data. The method is particularly effective in highlighting areas of high risk and potential ecological impacts, thereby providing valuable insights into the critical influences driving landslide susceptibility. However, the predictive reliability of this method might be contingent upon the spatial and temporal resolution of the data used in the analysis

### 3.2. Machine Learning Methods

This manuscript presents a comprehensive analysis of machine learning (ML) techniques applied to model earthquake-induced landslides, utilizing extensive datasets and multiple parameters to derive optimal solutions. The application of ML methods in landslide susceptibility modeling first gained significant attention in the early 2000s and achieved widespread popularity among scholars a decade later [48]. The study compares several ML approaches, including Logistic Regression, Artificial Neural Networks (ANNs), Support Vector Machines (SVM), Random Forest (RF), k-nearest Neighbors (KNN) [49], Naive Bayes (NB), and Decision Trees (DT) [50]. These methods have been scrupulously tested to establish robust models capable of predicting and analyzing the susceptibility of areas to landslides effectively.

(1) Logistic regression is a commonly used analytical model for dichotomous problems [17]. It divides the probability of occurrence by the probability of non-occurrence and then the logarithm of this ratio is computed. A linear relationship between the dependent and independent variables is constructed. It must be noted that while logistic regression facilitates a robust initial analysis for binary classification tasks, the prediction accuracy can be contingent on the specific characteristics and interdependencies of the variables involved. The mathematical expression for this relationship is denoted by

$$P = 1 \Big/ \left[ 1 + e^{-(\alpha + \beta_1 x_1 + \cdots + \beta_i x_i)} \right].$$
$$Ln(p/(1-p)) = \alpha + \beta_1 x_1 + \cdots + \beta_i x_i = \alpha + \beta x. \tag{2}$$

where $e$ is the probability of landslide occurrence, $P$ is in [0 1] and $i$ is the number of categories of evaluation factors.

(2) XGBoost is a machine learning algorithm implemented in the Gradient Boosting framework [51]. XGBoost uses Newton's method to calculate the extremes of the loss function by improving the gradient boosting algorithm and extending the loss function to the second order by Taylor expansion and adding the regularization term. The mathematical expression for this relationship is denoted by

$$L^{(t)} \cong \sum_{i=1}^{n} \left[ l\left(y_i, \hat{y}_i^{\;(t-1)}\right) + g_i f_i(x_i) + \tfrac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t).$$
$$\Omega(f_t) = \gamma^{T_t} + \tfrac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2. \tag{3}$$

where $L^{(t)}$ is the loss of a single sample of the truth iteration as a second-order continuously derivable convex function; $I$ denote the ith sample; $\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{\;(t-1)}\right)$ is the prediction error of round $t-1$ and is a constant; $y_i$ is the true value; $\hat{y}_i^{\;(t-1)}$ is the predicted value; $g_i$ and $h_i$ are the first-order and second-order derivatives; $f_i(x_i)$ is the weight assigned to the ith evaluation factor in round t; $f_t$ is the tree model generated in round $t$; $\Omega(f_t)$ is the regularization term; $^{T_t}$ is the round $t$ loop the number of leaf nodes in round t; $\omega_j$ is the weight of leaf node j; $\gamma$ and $\lambda$ are the hyperparameters set in advance.

(3) LightGBM like XGBoost is a machine learning algorithm implemented under the Gradient Boosting framework. Its original intention is to resolve the problem of memory storage when massive datasets are encountered. LightGBM uses unilateral gradient sampling to delete data with small gradients without affecting the prediction accuracy to improve the computing speed and save storage space. The mathematical expression is the same as that of XGBoost, except for the selection of the segmentation point based on the histogram. The value after feature discretization is only saved, and its memory ratio is only 1/8 of the latter.

(4) RandomForest, an ensemble model proposed by [42], is composed of many decision trees, in which each decision tree votes to select the mode as the final result. It uses the Bootstrap algorithm to extract features randomly based on the sample data set and train the evaluation factors, finally build a decision tree, and reduce its prediction error step by step. The mathematical expression for this relationship is denoted by

$$Y = E_\theta h(X, \theta).$$
$$GE = E_\theta E_{X,Y}[Y - h(X, \theta)]^2. \tag{4}$$

where $GE$ is the generalization error of each decision tree; $X$ and $Y$ are random variables drawn from the training set; $E_\theta$ is the expectation function; $h(X, \theta)$ is the decision tree prediction function; $E_{X,Y}$ is the joint expectation function of $X$ and $Y$.

(5) AdaBoost is an iterative model proposed by [52] to enhance the accuracy of the prediction, also known as reinforcement learning or boosting model. It performs different weak classifications on the same training set, afterward combines and iterates the weak classifications, and finally combines them into a robust classification. The mathematical expression for this relationship is denoted by

$$\varepsilon_t = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i.$$
$$D_t(i) = \frac{D_{t-1}(i)\beta_t^{-\varepsilon_i}}{z_t}.$$
$$W_t = \frac{1}{2} \ln\left(1 \big/ \beta_t\right). \tag{5}$$
$$\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}.$$
$$H(x) = \sum_{t=1}^{T} W_t h_t(x).$$

where $\varepsilon_i$ and $\varepsilon_t$ are the error and average error of ith weak classifier, respectively; $D_t(i)$ is the updated evaluation factor of the weight; $W_t$ is the weight of the weak classifier; $z_t$ is the normalization factor of $\sum_{i=1}^{N} D_t(x_t) = 1$; $H(x)$ is the final strong classifier; $h_t(x)$ is the weak classifier; $T$ is the iteration numbers.

(6) GaussianNB (GNB) and ComplementNB (CNB) are two forms of Naive Bayes. Naive Bayes is a supervised learning algorithm based on the Bayesian principle [53]. It assumes that the feature conditions are independent of each other. The joint probability distribution of input and output is learned through the selected training set, and the posterior probability is finally obtained. The mathematical expression for this relationship is denoted by

$$P(y|x_1, \cdots, x_n) = \frac{p(y)p(x_1, \cdots, x_n|y)}{p(x_1, \cdots, x_n)}. \tag{6}$$

where $P(y)$ is the prior probability, $P(x)$ is the evidence; $P(x|y)$ is the class conditional probability; $P(y|x)$ is the posterior probability.

There are four kinds of Naive Bayes in the sci-kit-learn: GaussianNB (GNB), MultinomialNB, BernoulliNB, and ComplementNB (CNB). The categories are defined based on the

distribution of prior probabilities, which are GaussianNB for the Gaussian distribution of prior probabilities, MultinomialNB for the polynomial distribution of prior probabilities, BernoulliNB for the Bernoulli distribution of prior probabilities, and ComplementNB for the modified version of the standard polynomial plain Bayesian classifier. In this study, both GaussianNB and ComplementNB are used.

(7) Multilayer Perceptron (MLP) is a mathematical model that simulates the principle of a biological neural network for distributed parallel information processing, which has the characteristics of massively parallel processing, high redundancy, and distributed storage. It achieves model convergence by changing the gradient so and forth, through the intersection of neurons at different levels and continuously learns to improve their weights. Stochastic gradient descent is performed on the variance of the output. The mathematical expression for this relationship is denoted by

$$y_i(t) = f\left(\sum_{i=1}^{n} \omega_{ji} x_i - \theta\right). \tag{7}$$

where $y_i$ is the output value of the neuron; $\theta$ is the threshold value of the neuron; $x_i$ is the input vector of the neuron; $\omega_{ji}$ is the connection weight between the layers and $f$ is the excitation function.

(8) Support vector machine (SVM) is a binary classifier to evaluate parameters utilizing the supervised learning scheme with sparsity and robustness. It classifies the evaluation parameters by the kernel function to achieve the overall classification and obtain the most suitable kernel function by iteration. The mathematical expression for this relationship is denoted by

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|}{\sigma^2}\right). \tag{8}$$

where $K$ is the kernel function known as the mapping function; $x$ is the center of the kernel function; $\sigma^2$ is the square of the width parameter of the function.

### 3.3. Recursive Feature Elimination

Recursive Feature Elimination is based on a greedy algorithm that utilizes a feature ranking technique proposed by [33]. It builds a model and eliminates features sequentially based on their contribution to model accuracy, which is iteratively determined. The stability of this method hinges on the initial model choice. To prevent overfitting, cross-validation is integrated during the elimination process, ensuring robustness and generalizability. Feature importance is evaluated through the model's internal metrics, such as information gain or coefficient magnitude, which influence the elimination sequence. This technique was applied using nine different machine learning models to generate classification parameters that meet basic requirements and achieve the intended testing purpose. Cross-model comparisons further ensure that the selected features maintain their predictive power across different model architectures, minimizing the risk of overfitting.

### 3.4. Confusion Matrix

Nine machine learning models are employed to evaluate the susceptibility of seismic landslide areas. Given that the assessment of seismic landslides is framed as a binary classification problem, with outcomes represented as 0 (non-landslide) or 1 (landslide), a confusion matrix is introduced as a critical evaluation index [54]. This matrix facilitates a detailed analysis of the differential performance of each model, using various evaluation parameters to accurately assess the susceptibility of the final model.

The confusion matrix organizes the predicted and actual situations along the horizontal and vertical axes, respectively, generating results in the form of four distinct outcomes:

True Positive (*TP*): Both the predicted and actual situations are landslides;

True Negative (*TN*): Both the predicted and actual situations are non-landslides;

False Positive (*FP*): The model predicts a landslide where there is actually a non-landslide;

False Negative (*FN*): The model predicts a non-landslide where there is actually a landslide.

These classifications serve as the basis for calculating key metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and $F_1$-score, along with the ROC curve and AUC value. The ROC curve is plotted with sensitivity on the vertical axis against "1-specificity" on the horizontal axis. By altering the threshold value for prediction classification, different points on this curve are obtained and transformed into a continuous curve. The area under the ROC curve (AUC), a comprehensive metric, quantifies the overall ability of the model to discriminate between classes and is commonly employed as a principal model evaluation index. These detailed computations are depicted in Table 1.

**Table 1.** The explanations of the parameters in the confusion matrix.

| Statistical Criteria | Meaning | Formula |
|---|---|---|
| Accuracy | the proportion of true results (both true positives and true negatives) in the total cases examined. | $ACC = \dfrac{TP + TN}{TP + TN + FP + FN}.$ |
| Sensitivity | the proportion of actual positives that were correctly identified. | $TPR = \dfrac{TP}{TP + PN}.$ |
| Specificity | the proportion of actual positives that were correctly identified. | $TNR = \dfrac{TN}{TN + FP}.$ |
| Positive predictive value | the proportion of positive identifications that were correct. | $PRE = \dfrac{TP}{TP + FP}.$ |
| Negative predictive value | the proportion of negative identifications that were correct. | $NPE = \dfrac{TN}{TN + FN}.$ |
| $F_1$-value | the harmonic mean of precision and sensitivity. | $F_1 = \dfrac{2PRE \times TPR}{PRE + TPR}.$ |

## 4. Modeling

The model can be divided into five stages in total:

(1) Aggregating the seismic landslide evaluation parameters in the study area and establishing a spatial database.
(2) Designing three kinds of samples for the training set, validation set, and test set of the machine learning model.
(3) Analyzing the importance ranking of evaluation factors obtained from GeoDetector and nine machine learning methods.
(4) Using recursive feature elimination means screening the evaluation factors; exploring the relationship between evaluation factors and seismic factors;
(5) Exploring the relationship between evaluation factors and landslides.

### 4.1. Spatial Database

As a phenomenon caused by multiple factors, including natural and induced conditions, seismic landslides are investigated, and 20 well-known evaluation indexes are selected as research parameters in this paper. The chosen evaluation factors have been demonstrated to be sensitive to various triggering conditions, such as seismic activities and environmental changes, in previous research. For the evaluation factors that cannot be directly measured, their values are attained according to the acquired means presented in Table 2 and then are resampled from the exact 30 m resolution. Afterward, a classification scheme is established for each constant factor for reclassification to establish the spatial database of the evaluation factors, and each evaluation factor is shown in Figure 2. The classification scheme is stopped when the distributions of the continuous factors are used, and then the generated values are rounded to obtain scores presented in Table 3.

This methodology ensures that our spatial database accurately captures and reflects the influence of triggering conditions on landslide susceptibility.

**Table 2.** The calculation of the evaluation factors.

| Parameters | Meaning | Acquisition Means |
|---|---|---|
| Elevation | Surface to sea level height. | ASTER satellite 30 m data for projection analysis. |
| Slope | The degree of steepness of the surface unit. | ArcGIS:3D Analyst tools-Raster Surface-Slope. |
| Aspect | The direction of the projection of the normal to the slope on the horizontal plane. | ArcGIS:3D Analyst tools-Raster Surface.-Aspect. |
| RDLS | The height difference between the highest and lowest point in the defined area. | ArcGIS: Spatial Analyst Tools-Neighborhood-point Statistics. |
| Slope length | The maximum horizontal projection length of the trajectory between the fixed point upstream and the beginning of the flow. | $L = \dfrac{DEM}{\sin\left(\frac{slope*\pi}{180}\right)}.$ |
| TRI | The ratio between their projected areas in the specified area. | $TRI = \dfrac{1}{\cos\left(\frac{slope*\pi}{180}\right)}.$ |
| TWI | Physical indicators of the influence of regional topography on runoff direction and accumulation. | $TWI = \ln\left(\dfrac{SCA}{\tan\left(\frac{slope*\pi}{180}\right)}\right).$ <br> Note: SCA is the sink flow per unit area. |
| SPI | Quantitative description of the erosive capacity of surface water. | $SPI = SCA \cdot \tan\left(\dfrac{slope*\pi}{180}\right).$ |
| STI | Quantitative description of surface water sand transport capacity. | $STI = \left(\dfrac{SCA}{22.13}\right)^{0.6} \cdot \left(\dfrac{\sin\left(\frac{slope*\pi}{180}\right)}{0.0896}\right)^{1.3}.$ |
| Curvature | The slope of the surface slope | ArcGIS:3D Analyst tools-Raster Surface-Curvature |
| Profile Curvature | Second-order derivative along the steepest descending slope | ArcGIS:3D Analyst tools-Raster Surface-Curvature |
| Plan Curvature | Second-order derivative perpendicular to the downward gradient | ArcGIS:3D Analyst tools-Raster Surface-Curvature |

**Table 3.** The classification of the evaluation factors.

| Evaluation Factors | Class | Classification Standards |
|---|---|---|
| Elevation (m) | 8 | 1. <2350; 2. 2350–2469; 3. 2469–2572; 4. 2572–2672; 5. 2672–2883; 6. 2883–3028; 7. 3028–3339; 8. >3339. |
| Lithology | 5 | 1. Devonian (D); 2. Permian (P); 3. Paleocene (E); 4. Neocene (N); 5. Quaternary (Q). |
| Distance from roads | 7 | 1. <500; 2. 500–1000; 3. 1000–1500; 4. 1500–2000; 5. 2000–2500; 6. 2500–3000; 7. >3000. |
| Aspect | 9 | 1. Flat; 2. North; 3. Northwest; 4. East; 5. Southwest; 6. South; 7. Southwest; 8. West; 9. Northwest |
| RDLS | 8 | 1. 0–42; 2. 42–72; 3. 72–95; 4. 95–117; 5. 117–140; 6. 140–167; 7. 167–202; 8. 202–314. |
| Slope length | 6 | 1. <4683; 2. 4683–5003; 3. 5003–6941; 4. 6941–18,648; 5. 18,648–89,393; 6. >89,393. |
| TRI | 6 | 1. <1.05; 2. 1.05–1.1; 3. 1.1–1.17; 4. 1.17–1.27; 5. 1.27–1.4; 6. >1.4. |
| TWI | 7 | 1. <2.7; 2. 2.7–5.1; 3. 5.1–6.9; 4. 6.9–9.4; 5. 9.4–11.4; 6. 11.4–13.3; 7. >13.3. |
| SPI | 6 | 1. <6520; 2. 6520–13,041; 3. 13,041–19,562; 4. 19,562–39,124; 5. 39,124–136,935; 6. >136,935. |
| STI | 7 | 1. <100; 2. 100–200; 3. 200–300; 4. 300–400; 5. 400–500; 6. 500–600; 7. >600. |
| PGA | 7 | 1. <0.3; 2. 0.3–0.35; 3. 0.35–0.4; 4. 0.4–0.45; 5. 0.45–0.5; 6. 0.5–0.55; 7. >0.55. |
| Curvature | 8 | 1. <−3; 2. −3 to −2; 3. −2 to −1; 4. −1 to 0; 5. 0–1; 6. 1–2; 7. 2–3; 8. >3. |
| Profile Curvature | 8 | 1. <−2; 2. −2 to −1; 3. −1 to −0.5; 4. −0.5 to 0; 5. 0–0.5; 6. 0.5–1; 7. 1–2; 8. >2. |
| Plan Curvature | 8 | 1. <−2; 2. −2 to −1; 3. −1 to −0.5; 4. −0.5 to 0; 5. 0–0.5; 6. 0.5–1; 7. 1–2; 8. >2. |
| Distance from faults | 8 | 1. <500; 2. 500–1000; 3. 1000–1500; 4. 1500–2000; 5. 2000–2500; 6. 2500–3000; 7. 3000–3500; 8. >3000. |
| NDVI | 8 | 1. <0.25; 2. 0.25–0.3; 3. 0.3–0.35; 4. 0.35–0.4; 5. 0.4–0.45; 6. 0.45–0.5; 7. 0.5–0.55; 8. >0.55. |
| Distance from rivers | 7 | 1. <1000; 2. 1000–2000; 3. 2000–3000; 4. 3000–4000; 5. 4000–5000; 6. 5000–6000; 7. >6000. |
| Land cover | 5 | 1. Arable land; 2. Forest; 3. Grassland; 4. Land for waters and water conservancy facilities; 5. Artificial surface. |
| Slope (°) | 9 | 1. <5; 2. 5–10; 3. 10–15; 4. 15–20; 5. 20–25; 6. 25–30; 7. 30–35; 8. 35–40; 9. >40. |
| POI kernel density | 7 | 1. <0.2; 2. 0.2–0.4; 3. 0.4–0.6; 4. 0.6–0.8; 5. 0.8–1; 6. 1–2; 7. >2. |

### 4.2. Sample Design

Seismic landslides are categorized dichotomously, meaning they take either 0 or 1, where 1 represents the occurrences, and 0 denotes the non-occurrences. Since there exist 2317 seismic landslides in the research area, an equal number of 2317 are randomly selected as non-seismic landslide points in the non-seismic landslide area of the research region to maintain a 1:1 ratio.

Machine learning involves distinct datasets for training and testing to ensure robust model evaluation. Specifically, the dataset was divided into training and test sets, with the test set constituting 15% of the total sample (*N* = 695 cases). This partition was performed randomly to mitigate selection bias. To target high prediction accuracy and enhance the reliability of our models, cross-validation was implemented. In our methodology, the training set underwent 10-fold cross-validation, where it was randomly split into 10 equal subsets. During each cross-validation fold, nine subsets were used for training the model, and the remaining subset served as the validation set. This process not only maximizes learning outcomes by effectively using diverse training data but also prevents overfitting by ensuring no single subset influences the model excessively. The difference between the validation and test set results was carefully monitored, with a criterion for success being that the absolute difference did not exceed 10% of the test set outcomes. This rigorous approach minimizes random errors due to sample selection variability across different machine learning models and facilitates a comparative analysis of the accuracy between test and training results, thereby confirming the models' generalization ability on unseen data.
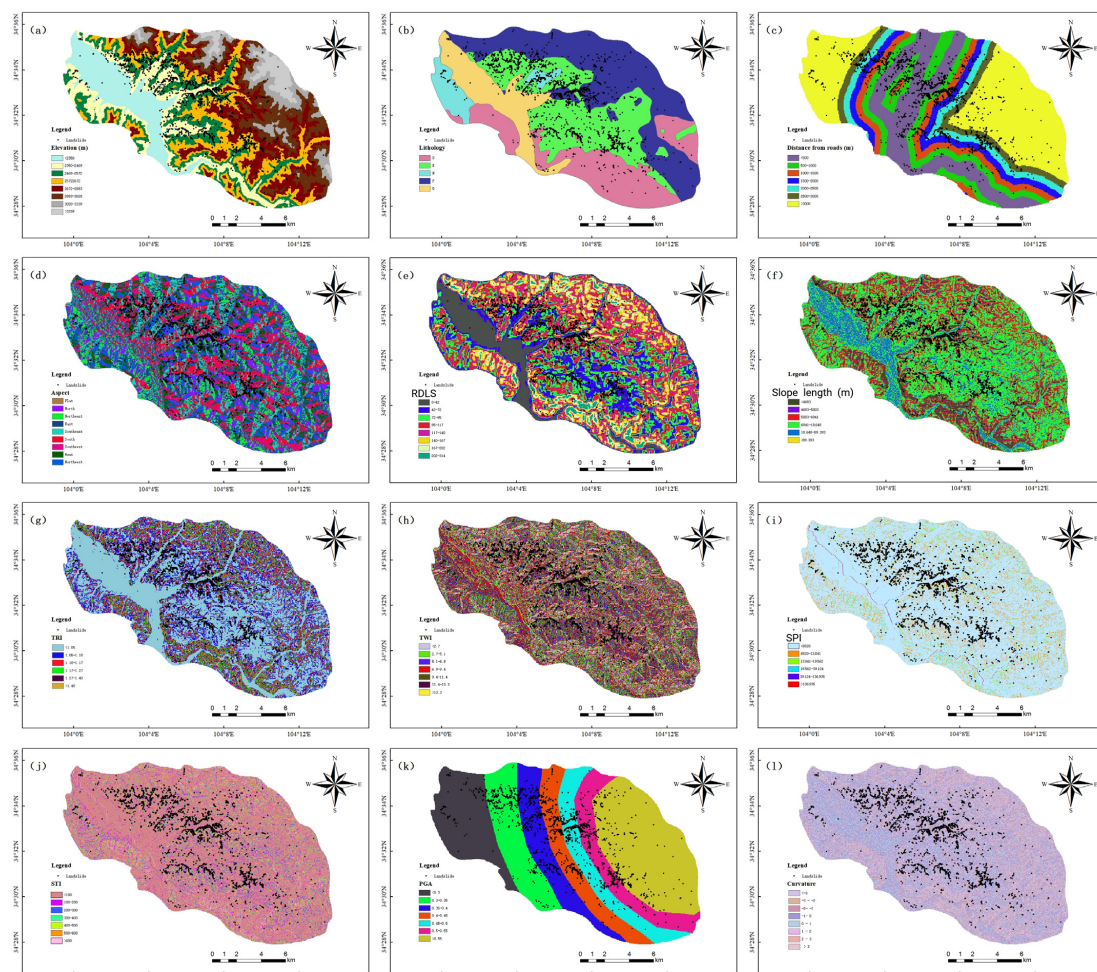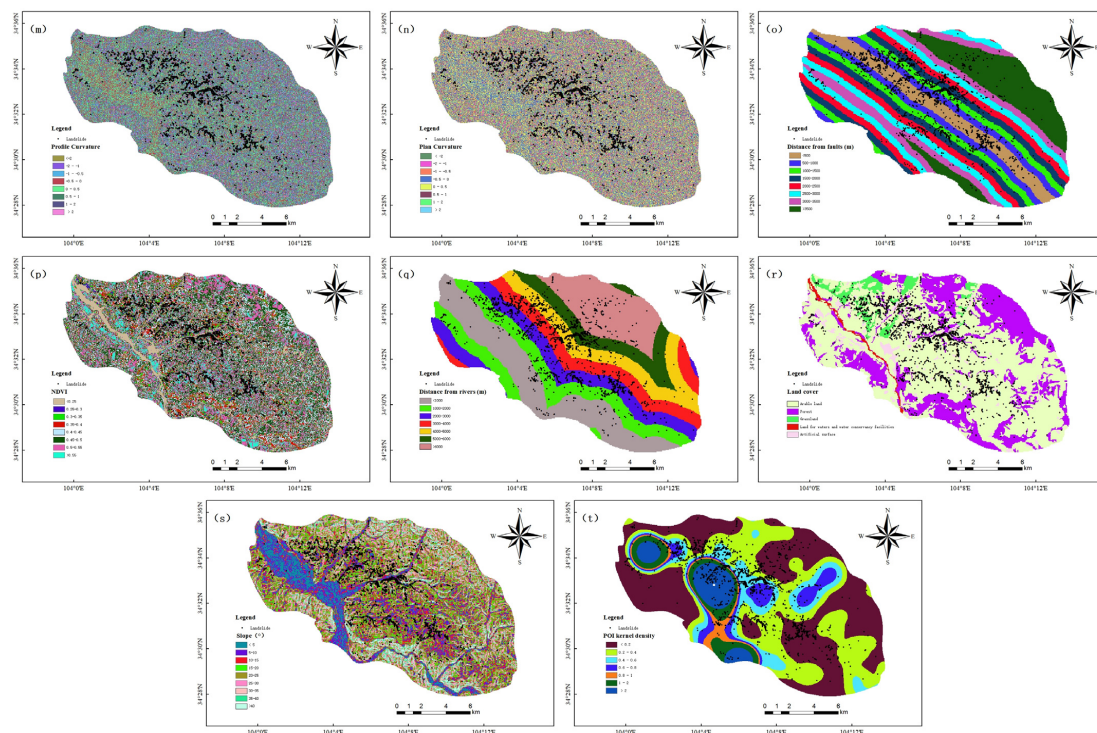


**Figure 2.** *Cont.*

**Figure 2.** Classification diagram of seismic landslide evaluation factors: (**a**) Elevation; (**b**) Lithology; (**c**) Distance from roads; (**d**) Aspect; (**e**) RDLS; (**f**) Slope length; (**g**) TRI; (**h**) TWI; (**i**) SPI; (**j**) STI; (**k**) PGA; (**l**) Curvature; (**m**) Profile curvature; (**n**) Plan curvature; (**o**) Distance from faults; (**p**) NDVI; (**q**) Distance from rivers; (**r**) Land cover; (**s**) Slope; (**t**) POI kernel density.

*4.3. Ranking of Factor Importance*

4.3.1. Evaluation Factors of GeoDetector

The classification numbers of the 20 evaluation factors were put into the GeoDetector table, and the interpretive q-values of different evaluation factors were calculated by the divergence and factor detector, which gave the following results (Figure 3). Lithology (0.236), Elevation (0.199), Distance from faults (0.131), Distance from rivers (0.102), Distance from roads (0.088), NDVI (0.086), PGA (0.080), RDLS (0.065), TWI (0.058), STI (0.044), POI kernel density (0.041), Slope (0.031), Aspect (0.025), Land cover (0.022), TRI (0.011), SPI (0.010), Slope length (0.010), Profile curvature (0.007), Curvature (0.006), and Plane curvature (0.004). The interpretative q-values obtained from the divergence and factor detectors in the geographic probe are relative scores. So, they need to be assessed statistically to decide whether they are significant or not. All 20 evaluation indicators are tested and found to be statistically significant since all *p*-values are 0.000 < 0.05, indicating that each evaluation factor is classified reasonably. Indicators such as lithology and elevation have a strong ability to explain seismic landslides are found when the q-values are under consideration. In addition, indicators such as profile curvature, curvature, and plane curvature have a limited ability to explain seismic landslides are derived. The analysis presented above is also consistent with the characteristics of the research area, where seismic landslides occur in relatively hilly and mountainous regions. Thus, the complex topography leads to more confusing values of curvature characteristics in the research area, and the spatial distribution of curvature is poorly consistent with the seismic landslides. However, the spatial differentiation of lithology and elevation is correlated with seismic landslides and plays an important role in seismic landslides.
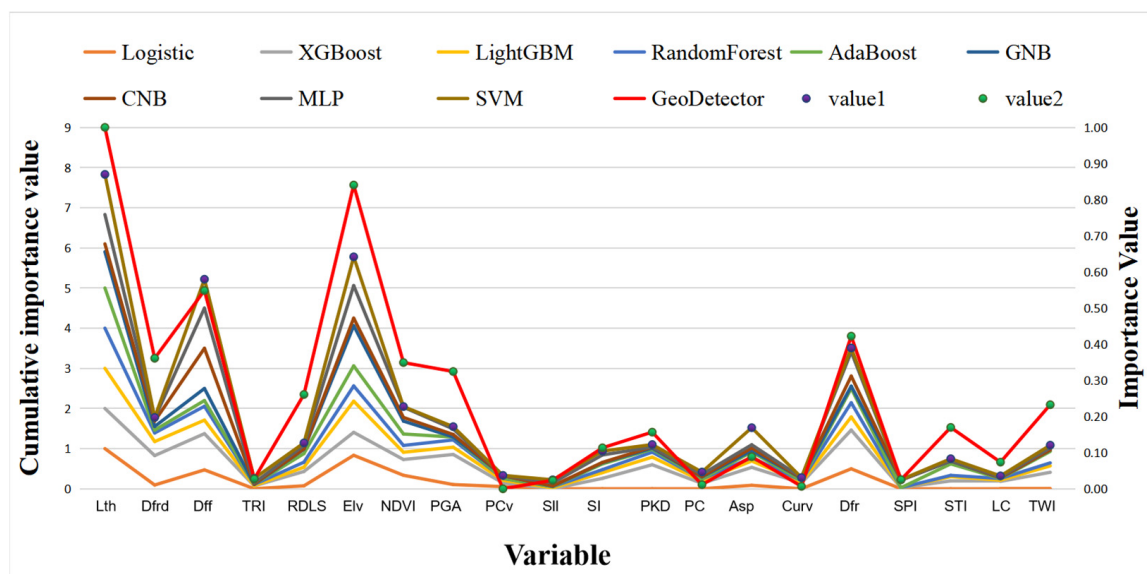
**Figure 3.** The stacking line graph of the important evaluation factors: Lth (Lithology), Dfrd (Distance from roads), Dff (Distance from faults), TRI (TRI), RDLS (RDLS), Elv (Elevation), NDVI (NDVI), PGA (PGA), PCv (Plan Curvature), Sll (Slope length), Sl (Slope), PKD (POI Kernel density), PC (Profile Curvature), Asp (Aspect), Curv (Curvature), Dfr (Distance from rivers), SPI (SPI), STI (STI), LC (Land cover), TWI (TWI).

### 4.3.2. Evaluation Factors for Machine Learning Modeling

The importance analysis of the evaluation factors was performed for nine machine learning methods. The importance of different evaluation factors was obtained through multiple iterations and by setting parameters corresponding to their models and quantitative characteristics. To standardize the importance scores across these diverse models, a normalization procedure was applied whereby each factor's importance score obtained from a model was divided by the total sum of importance scores from that model. This ensured that the normalized scores for each model were summed to one, allowing for equitable comparison and aggregation of scores across models.

The normalized importance scores were then summed across all nine models to derive a cumulative importance measure for each factor. This cumulative scoring approach ensures that our analysis reflects a balanced consideration of each factor's significance as determined by a variety of predictive modeling techniques. The results are easily summarized to create a stacked line graph (Figure 3), displaying the aggregated importance across models. When combined with the important evaluation factors identified by geographic detectors, a consistent ranking among them is observed. The following combined importance rankings were derived from the aggregated data: Lithology (21.36%), Elevation (16.00%), Distance from faults (13.95%), Distance from rivers (9.50%), NDVI (5.80%), Distance from roads (5.17%), PGA (4.53%), Aspect (3.89%), RDLS (3.39%), TWI (3.19%), POI kernel density (3.04%), Slope (2.54%), STI (2.22%), Profile curvature (1.04%), Land cover (0.96%), Plan curvature (0.81%), TRI (0.96%), Curvature (0.69%), SPI (0.61%), Slope length (0.60%).

The most appropriate parameters for each machine learning modeling are found to be as follows:

(1) Logistic regression: iteration number, regularization type, convergence metric, and regularization factor are set to 100, L1, 0.0001, and 0.01084366, respectively.

(2) XGBoost with binary optimization objective function: learning rate, maximum tree depth, minimum bifurcation weight, and L2 regularization coefficient are assigned to 0.3, 8, 6, and 0.5, respectively;

(3) LightGBM with gbdt type algorithm: learning rate, maximum tree depth, and the maximum number of leaves, and maximum number of leaves are set to 0.001, 20, 50, 5, respectively;

(4) RandomForest with metric Gini: minimum divergence purity gain, number of trees are assigned to 0.0, and 100, respectively;

(5) AdaBoost: n_learning rate and n_estimators number of single models are set to 0.3, and 50, respectively;

(6) GNB: var_smoothing $1 \times 10^{-7}$;

(7) CNB: alpha (additive (Laplace/Lidstone) smoothing): 0;

(8) MLP with nonlinear logistic function: hidden layer width and iteration number are set to (30,30), and 20, respectively;

(9) SVM: C (coefficient of regularization): 1.0, kernel (kernel type): rbf, tol (convergence metric): 0.001.

### *4.4. Factor Screening*

Machine learning is prone to overfitting or underfitting due to its characteristics; therefore, selecting influential factors from a large amount of underlying data can reduce the workload of models and provide better prediction accuracy for models. The recursive feature elimination method is used as an efficient and accurate screening tool to evaluate factors. Hence, this paper adopts two ways of screening to evaluate factors based on their principles and compares them. In the first phase, the nine machine learning approaches are ranked according to their evaluated factor importance, and the parameters with the lowest importance scores are eliminated. Therefore, their AUC, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 value of self-evaluation scores related to factor recursive feature elimination maps are obtained. In the second stage rankings based on the nine machine learning methods employing the importance scores of the comprehensive evaluation, factors are researched by eliminating the parameters with low importance scores. Similarly, the recursive feature elimination map of comprehensive evaluation factors was obtained (Figure 4).

From the recursive feature elimination plot representing the AUC values, it is known that if the factors are removed without any necessary change in AUC values, it means that they are non-essential parameters and if the factors are removed, their AUC values show a significant decrease, it means that they are essential parameters and can be utilized as valid factors. From the AUC graph of MOI, the AUC value starts to decrease when the number of removed factors reaches 15. In other words, the five most essential evaluation factors can reach information provided by the full 20 evaluation factors. The rest of the confusion matrix values are also within a particular control range. However, the evaluation factors with the highest self-evaluation importance scores are not precisely the same as for different machine learning methods, and valid evaluation factors with universal applicability are not usable. From the graph of AUC values of CI, the AUC values start to decrease when the number of deleted factors reaches 10. The 10 evaluation factors with the highest importance scores can ensure that each machine learning method reaches a higher value. By analyzing and comparing the remaining model indicators of the two groups of recursive feature elimination, the trends and values remained the same in the first 10 decreasing factors without significant variability being found. On the other hand, the last 10 decreasing factors attained by the comprehensively evaluated factors observe decreases in the values. It further verifies that the influence of the first 10 positions of the important comprehensive evaluation factor on the model indicators can be utilized as validity factors.

Figure 5 illustrates the differences in AUC values by subtracting the composite factor (CI) AUC from the self-assessment factor (MOI) AUC. Specifically, Figure 5a demonstrates that in the training set, AUC values with the decreasing self-assessment factor are generally higher than those achieved by the composite factor. In contrast, Figure 5b shows that for the test set, AUC values with the decreasing self-assessment factor are generally lower than those from the composite factor.
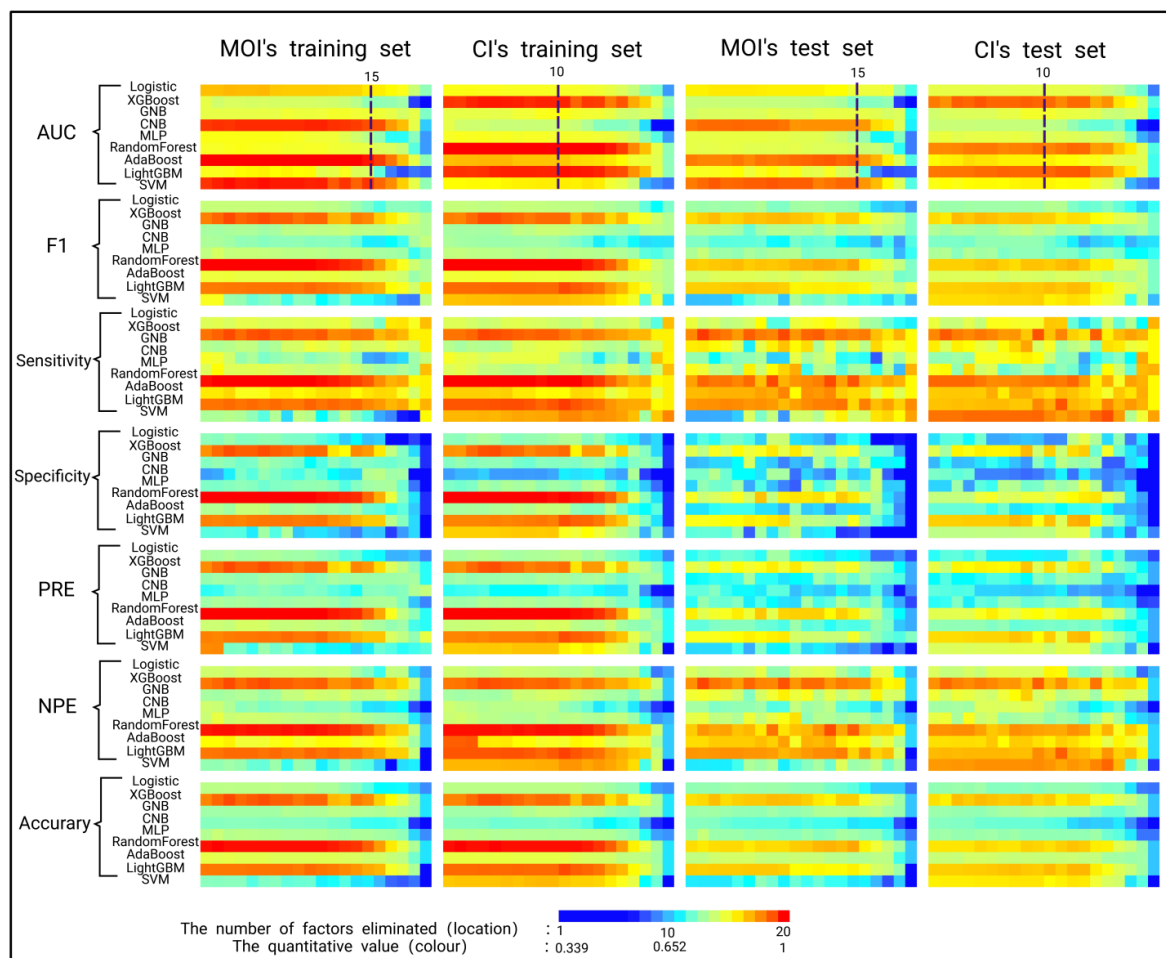
**Figure 4.** The recursive feature elimination map. MOI: The elimination of the factors according to the importance degree of the model; CI: The elimination of the factor according to the comprehensive importance degree.
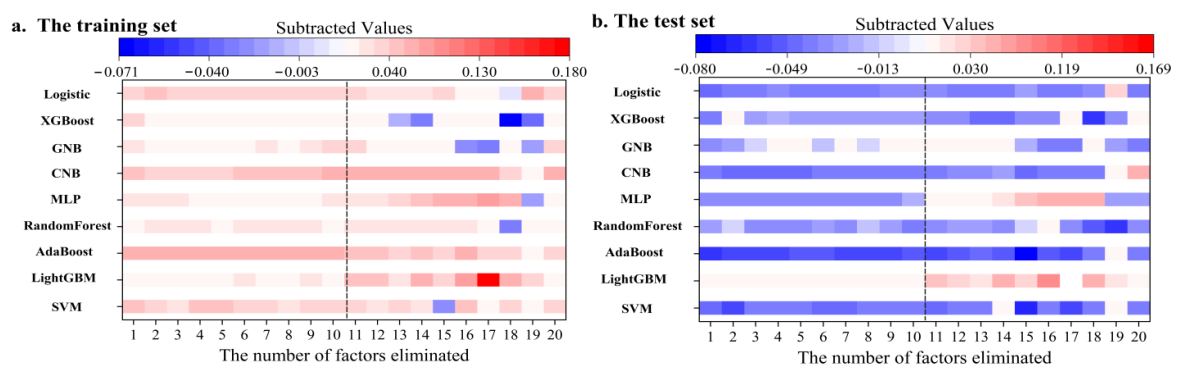


**Figure 5.** The obtained difference in AUC by subtracting CI from MOI.

These findings indicate that the AUC values derived from the decreased composite factor exhibit less variability between the training and test sets and are more consistently accurate than those derived from the decreased self-assessment factor. The analysis further reveals that the differences between the two methods remain unchanged until the elimination of 10 features, after which the AUC values begin to show significant fluctuations, highlighting an increase in variability and a divergence in the effectiveness of the two methods. This integrated analysis supports the utility of the top 10 influential factors identified by the integrated evaluation in supporting the nine machine learning models used in this

study, without compromising their original predictive accuracy. The detailed examination of these dynamics in Figure 5 enriches our understanding of the critical evaluation factors essential for robust machine learning applications.

Through the above analysis, the 10 influential evaluation factors obtained by the decreasing integrated factors can be utilized in the effective evaluation systems of nine machine learning models. Thus, the factors after screening and optimization are found to be lithology, elevation, distance from faults, distance from rivers, NDVI, distance from roads, PGA, aspect, RDLS, and TWI.

*4.5. Factor Analysis*

4.5.1. Single Factor Analysis

This section analyzes the relationship between seismic landslides and various evaluation factors across defined classification intervals, with key visualizations presented in Figure 6a–d. Lithological statistics reveal that the dominant lithologies, Permian and Paleoproterozoic, account for 32.89% and 27.91% of the region's geological composition, respectively. The Paleoproterozoic lithology exhibits the highest frequency of seismic landslides, representing 59.3% of the region's total (Figure 6a). Elevational data show a concentration of landslides between 2350 and 2572 m, accounting for 63.62% of occurrences, with a decline in landslide frequency at higher elevations (Figure 6b). Proximity to fault lines demonstrates a negative correlation with landslide occurrences, stabilizing beyond a distance of 2.5 km from fault lines (Figure 6c). The relationship between hydrological distance, elevation, and landslide occurrences indicates maximum activity 2–3 km from water systems, with a decreasing trend noted both nearer to and farther from these sources (Figure 6d). The frequency ratio of seismic landslides, calculated as the ratio of the observed to the expected number of landslides based on the proportional area, is detailed in Figure 6. The highest ratios are found in Paleoproterozoic regions, while other lithological intervals display frequency ratios below one.
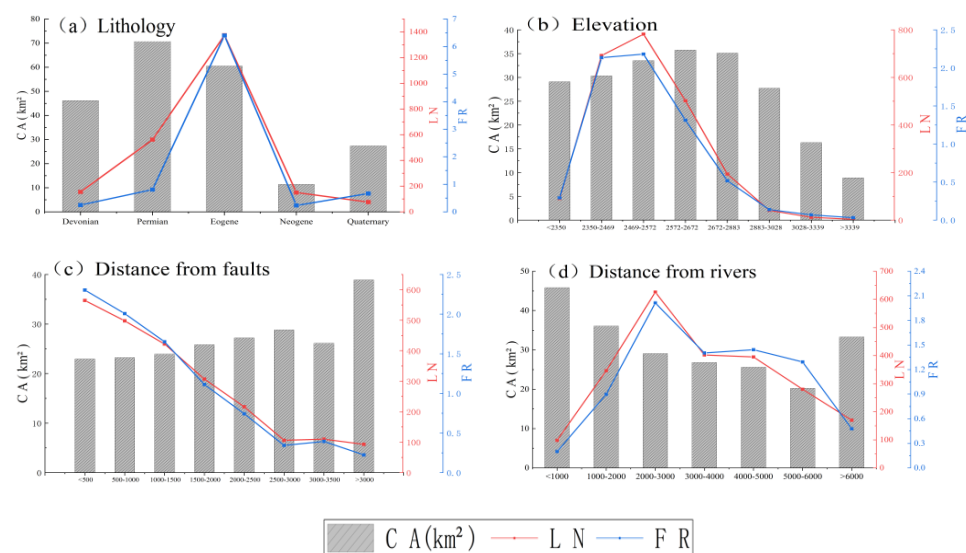


**Figure 6.** Relationships between each factor and LD and FR values: CA is the area of each category (km$^2$, shaded column); LN is the landslide number (km$^2$, red polylines); FR is the frequency ratio (blue polylines).

This section provides a concise, quantitative evaluation of the factors influencing the susceptibility of seismic landslides, supported by targeted visualizations and statistical analysis. When combined with the practical factors obtained from the previous screening, further observation shows that the frequency ratios of their influential factors hardly change with the size of the graded area. The frequency ratios even show an inverse correlation with the graded area. In contrast, the frequency ratios of the factors rejected by the GeoDetector

and machine learning show the same trend with the graded area. It further confirms that the eliminated evaluation factors, which do not have strong characteristics in terms of seismic landslide susceptibility, and their influence on the number of seismic landslides, are mainly determined by the area of the graded area.

### 4.5.2. Analysis of Factor Interactions

The interactions of multiple factors in seismic landslides are called facilitation, exclusion, and independence among their factors. In this study, a two-factor interaction analysis was conducted using a geographic detector (GeoDetector) to interact with the independent variables x1 and x2 to obtain their importance q(x1∩x2) values, which were then compared with the values of Min(q(x1), q(x2)), Max(q(x1), q(x2)), and q(x1) + q(x2), and the comparison results can be divided into the following five categories:

(1)   Non-linear weakening relationship: q(x1∩x2) < Min(q(x1), q(x2));
(2)   single-factor nonlinear weakening: Min(q(x1), q(x2)) < q(x1∩x2) < Max(q(x1), q(x2));
(3)   two-factor enhancement: q(x1∩x2) > Max(q(x1), q(x2)).
(4)   independent q(x1∩x2) = q(x1) + q(x2);
(5)   nonlinear enhancement: q(x1∩x2) > q(x1) +q(x2).

When all are applied to this paper, a graph of factor interaction evaluations is obtained (Figure 7). The horizontal coordinates of the graph denote all ranked scores according to their comprehensive evaluation importance scores, and their interactions are obtained to show both two-factor enhancement and nonlinear enhancement. For example, the interaction between elevation (q = 0.199) and topographic relief (q = 0.065) with q = 0.254 has a greater importance score than both, and the interaction relationship shows a two-factor enhancement since the effect of RDLS (elevation) on the occurrence of seismic landslides is different from RDLS (elevation) on the non-occurrence of seismic landslide. Hence, elevation and terrain relief can both promote each other, resulting in a larger effect on the occurrence of seismic landslides. In this paper, the filtered effective factors are boxed in Figure 7. The number of bifactor enhancements in their interaction accounts for 66.7% of the total number of bifactor enhancements found, which also further confirms the effective factors meeting the bifactor enhancement characteristics more than the non-effective factors and are more closely interconnected. Therefore, the influence degree of excluding effective factors would be greater than that of the non-effective factors.
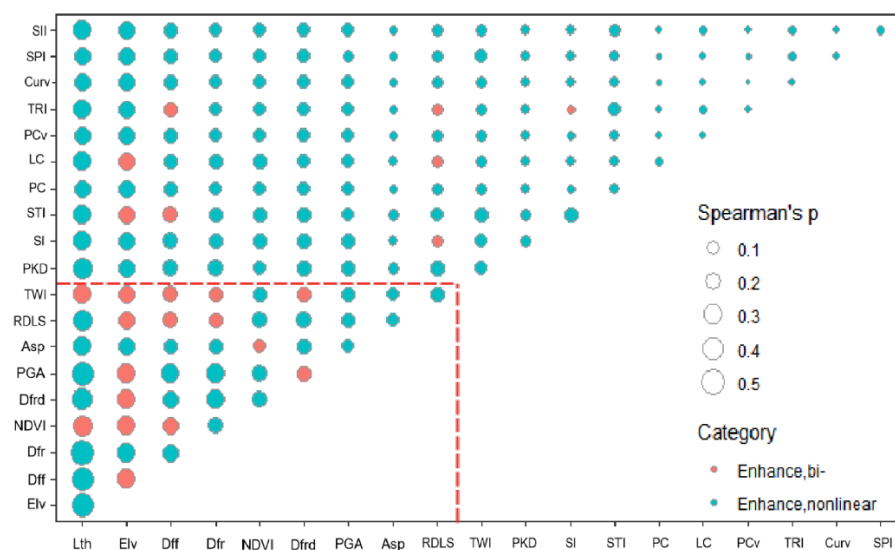


**Figure 7.** The evaluation of the factor interactions. (The red dashed line indicates the interaction among the ten factors selected after screening).

## 5. Results

### 5.1. Model Validation

The 20 evaluation factors were analyzed in the model comparisons with the screened 10 valid evaluation factors to obtain the variability of the two cases and produce the multi-model evaluation system diagram of the factors (Figures 8 and 9). From Figures 8a,b and 9a,b, improved AUC values for the model test set were observed, effective screening of evaluation factors was performed, and the AUC values of the five models such as GNB, MLP, Random-Forest, SVM, and XGboost were improved with the following 3.61%, 1.19%, 0.30%, 8.37%, and 1.12%, respectively, and the remaining four models had no change in their AUC values. Therefore, the AUC value of the highest model (random forest) in this test set also is improved from 0.989 to 0.992.



**Figure 8.** The map of the multi-model evaluation system with 20 factors: (**a**) ROC curve of the training set with 20 factors; (**b**) ROC curve of the validation set with 20 factors; (**c**) DCA curve for the validation set of 20 factors.
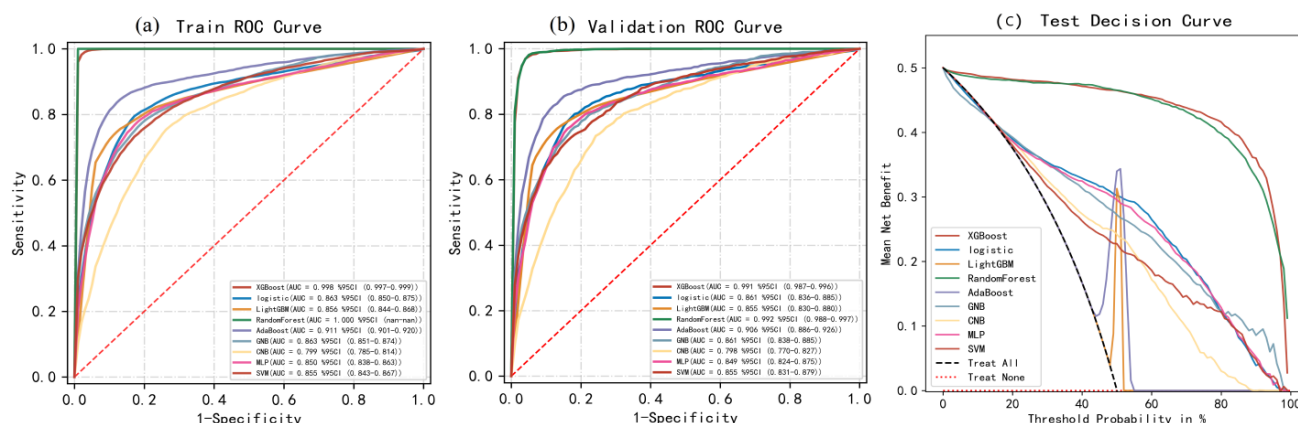


**Figure 9.** The map of the multi-model evaluation system with 10 factors: (**a**) ROC curve of the training set with 10 factors; (**b**) ROC curve of the validation set with 10 factors; (**c**) DCA curve for the validation set of 10 factors.

From Figures 8c and 9c, the decision curve analysis shows a larger area between the curves for the revised model set-up, reflecting greater model stability and reliability in decision-making contexts. After careful analysis of the distribution of each model, there is no change in the relationship between the derived models, which further shows no effect of the factor screening on the calculation characteristics of the model. So, implementation was found to be practical and feasible.

*5.2. Optimized Landslide Susceptibility Mapping*

In this study, nine machine learning algorithms along with the GeoDetector were employed to identify critical evaluation factors for landslides triggered by the Minxian earthquake, resulting in the selection of 10 experimental evaluation factors. Using the Random Forest model, a seismic landslide susceptibility map was generated and classified into five susceptibility levels (deficient, low, medium, high, very high) using the Jenks natural breaks method in ArcGIS [55].

The susceptibility maps generated with 20 and 10 evaluation factors (shown in Figures 10a and 10b, respectively) exhibited a high degree of visual consistency. Both the kernel and point density maps of actual landslide locations (Figure 10c,d) indicated that the highest susceptibility and actual landslide densities were concentrated in the northwestern and central parts of the study area. Notably, the map with 10 evaluation factors (Figure 10b) displayed a more focused distribution in the middle and high susceptibility regions compared to the more dispersed trend in Figure 10a. This suggests that the map with 10 factors more accurately reflects the actual landslide distribution.
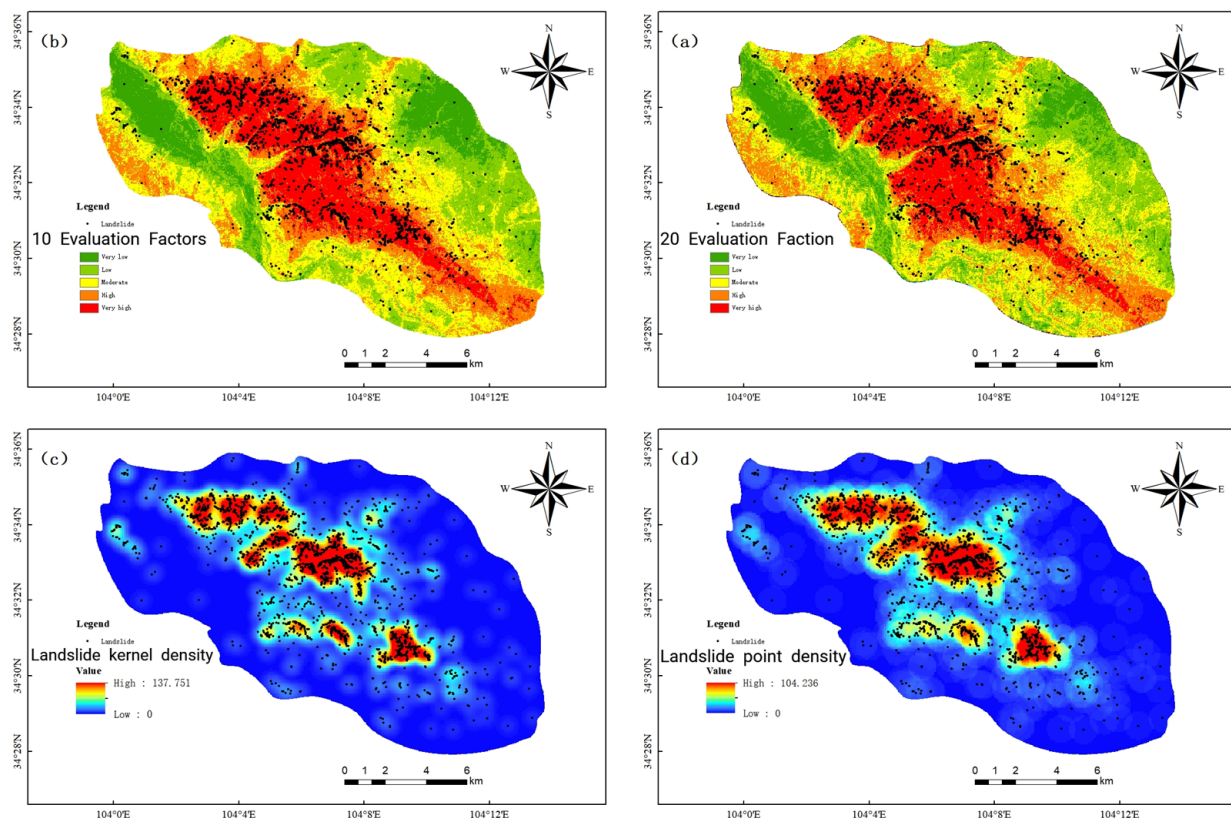


**Figure 10.** The sensitivity map of the seismic landslide: (**a**) 20 Evaluation Faction; (**b**) 10 Evaluation Faction; (**c**) Landslide kernel density; (**d**) Landslide point density.

Further quantitative analysis was performed on these maps (Figure 11). Figure 11a showed that the areas covered by each susceptibility category were similar across both maps, indicating good generalizability [7,18–20]. Figure 11b,c indicated that the very high susceptibility areas contained 67.6% and 67.77% of the total landslides for the maps with 20 and 10 evaluation factors, respectively, with frequency ratios of 3.47 and 3.52. This confirms that the concentration of landslides remains high with fewer factors.
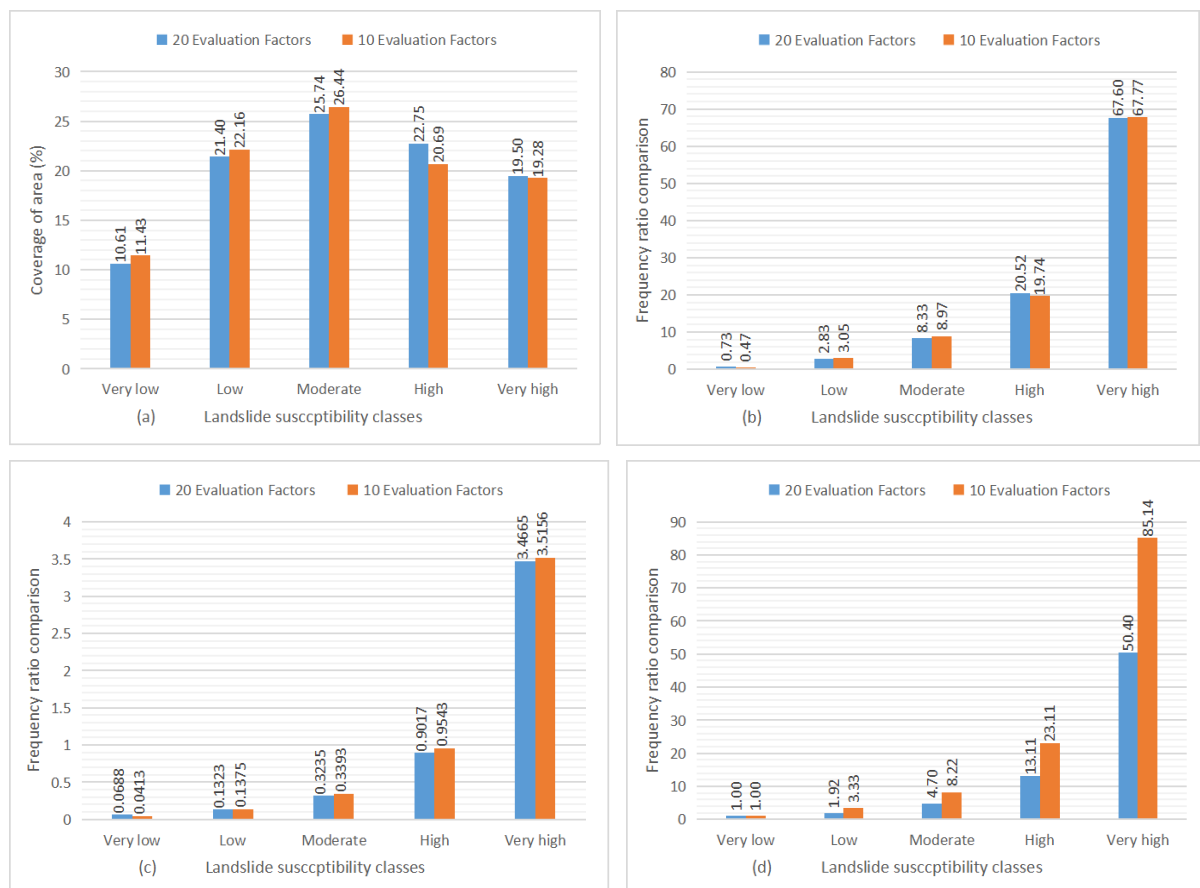
**Figure 11.** The susceptibility map of the quantitative analysis of seismic landslide: (**a**) the graded area of the susceptibility map; (**b**) the landslide data on each graded area of the susceptibility map; (**c**) the landslide frequency ratio of each graded area of the susceptibility map; (**d**) the ratio of the landslide frequency ratio of the susceptibility levels to the deficient level.

The frequency ratio multiplier for the very high susceptibility category was found to be significantly higher (85.14) with the optimized 10 factors compared to 50.40 with 20 factors, an improvement of 68.93%. This analysis confirms that optimizing evaluation factors using machine learning and GeoDetector significantly enhances the accuracy and efficiency of seismic landslide susceptibility mapping.

## 6. Discussion

### 6.1. The Optimization of the Evaluation Factors of the Seismic Landslide

Optimizing evaluation factors in seismic landslide susceptibility mapping presents a complex, nonlinear challenge often compounded by redundancy and missing factors. Enhancing this optimization significantly improves the speed and accuracy of susceptibility map production. In this study, we employ a novel combination of data and spatial features through nine machine-learning models, alongside geographic detectors, pioneering this approach based on the foundational work cited in references [6–8]. Although various optimization methods ranging from simple to complex models have been explored by scholars, often yielding improved results, our approach addresses the lack of spatial interpretation inherent in many machine learning models, thereby enhancing the practical application of these optimized factors.

We analyzed the core area of the Minxian earthquake landslide using recursive feature elimination in a conventional machine learning model to narrow down 20 factors to the most influential ones, as shown in Figure 4. Different models identified different influential factors, leading to varied optimization outcomes as noted in previous studies [25]. To

overcome this, we normalized, accumulated, and integrated the importance of each factor from all models with the results from the geographic detectors (Figure 3), achieving a universally applicable set of evaluation factors. This approach allowed us to filter down to 10 influential factors effective across all models, enhancing prediction accuracy as evidenced by improved AUC and DCA plots (Figures 8 and 9). This method not only confirmed the individual model findings but also demonstrated superior stability and predictive power across the regional dataset, aligning with the successes reported in [18,28].

### 6.2. Impact of Effective Evaluation Factors on Seismic Landslides

The evaluation factors for seismic landslides are extensive but can be broadly categorized into seismic, topographic, geological, environmental, and artificial activity parameters. Through a comprehensive analysis using machine learning models and geographic detectors, we identified ten practical factors: lithology, elevation, distance from faults, rivers, roads, NDVI, PGA, aspect, RDLS, and TWI. Notably, geological and environmental factors proved highly effective.

Our analysis revealed significant relationships between these factors and the occurrence of seismic landslides, detailed in Figure 6. Lithology and elevation emerged as the most influential factors, with landslides predominantly occurring in the Paleoproterozoic era and at elevations between 2350 and 2572 m—terrain characterized by hilly, gully-filled, and valley regions. This finding correlates with the bedrock composition in the area—conglomerate and sandstone from the lower Paleozoic strata (Eb), known for their susceptibility to seismic activities as documented in [56]. Hydrological patterns in the area also influence landslide distribution, predominantly towards the southwest and west, comprising 38.37% of all landslides. Human activities further modulate landslide occurrence, particularly through POI nucleation density and proximity to roads [57,58], validating the relevance of these factors in lower elevation areas.

Our study also analyzed the interaction of multiple evaluation factors on seismic landslides. The double-dual evaluation interaction of 20 factors was explored, revealing that influential factors often have stronger inter-factor interactions, accounting for 66.7% of all enhancements observed in Figure 7. This suggests that effective factors not only operate individually but also synergistically to influence landslide dynamics. Moreover, we computed correlation graphs between the frequency ratio of each factor and the graded area, finding that effective factors typically show negative correlations with the graded area, whereas non-effective factors display high consistency. This specificity underlines the distinct roles that different categories of the same factor may play—either promoting or inhibiting landslide occurrence. However, these observations have not been universally verified and require further validation across diverse regions with multiple seismic landslide events to confirm their reliability.

This analysis confirms the complex and interdependent nature of evaluation factors in seismic landslide susceptibility, highlighting the necessity of integrating diverse data sets for accurate predictions.

### 6.3. Data Processing and Sampling in Landslide Prediction

Despite achieving robust results through the optimization of evaluation factors via nine machine learning models and geographic detectors, this study has its uncertainties and limitations. The pre-processing resolution varied as data were resampled from $30 \times 30$ m across the dataset, a common practice in prior studies [59], yet potentially distorting peripheral values of seismic landslide data. We addressed these distortions by eliminating 13 outliers identified through field comparison.

Furthermore, the ratio of landslide points to non-landslide points was maintained at 1:1, a ratio found reliable in earlier research [60]. However, the reliability of non-landslide points, summarized from historical data where landslides never occurred [61], remains unverified against other studies. This paper's reliability of non-landslide data differs from that in existing literature, where landslide susceptibility maps are often created through

self-organized mapping networks and low probability regions are used for selecting non-landslide points [62].

## 7. Conclusions

In this study, nine machine learning models were integrated with the GeoDetector to evaluate the screening factors in the Minxian earthquake-induced landslide area. Ten reliable and effective evaluation factors were identified, eliminating superfluous ones, thereby enhancing the predictive capability of machine learning models in the region and ultimately producing a Seismic Landslide Susceptibility Map (SLSM). Moreover, several other significant conclusions were summarized as follows:

(i) The combination of machine learning models with the GeoDetector addressed the lack of spatial features in machine learning models and improved the interpretability of model outcomes. This offers a reference solution for subsequent machine learning model studies on regional issues.

(ii) Through recursive screening of integrated evaluation factors across nine machine learning models, the effective factors identified were proven to be applicable across these models. This comprehensive evaluation approach overcame the limitation where effective factors initially screened in one way were only usable in that particular model, laying the foundation for establishing a regional master database of effective evaluation factors and offering new directions for formulating earthquake disaster prevention and mitigation plans.

This paper employs multiple machine learning models in conjunction with the GeoDetector for Seismic Landslide Susceptibility Mapping (SLSM) and provides a universal framework for filtering the influencing factors of landslides triggered by earthquakes. Furthermore, the effective factors derived in this study can be applied to SLSMs in regions or countries with mountainous terrains similar to the study area.

The limitations of the research are expressed as follows: (1) the reclassification process could lead to a difference between the actual situation of the seismic landslide points and observed samples, which could distort the actual data on peripheral values, which leads to the removal of some observations, (2) the reliability related to the non-landside points of this paper is not compared to the studies conducted by other scholars.

# References

1. Reichenbach, A.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.

2. Kainthola, A.; Sharma, V.; Pandey, V.H.; Jayal, T.; Singh, M.; Srivastav, A.; Singh, P.K.; Champati, R.P.K.; Singh, T.N. Hill slope stability examination along lower tons valley, garhwal himalayas, India. *Geomat. Nat. Hazards Risk* **2021**, *12*, 900–921. [CrossRef]

3. Tiwari, V.N.; Pandey, V.H.R.; Kainthola, A.; Singh, P.K.; Singh, K.H.; Singh, T.N. Assessment of karmi landslide zone, bageshwar, uttarakhand, India. *J. Geol. Soc. India* **2020**, *96*, 385–393. [CrossRef]

4. Mahanta, B.; Singh, H.O.; Singh, P.K.; Kainthola, A.; Singh, T.N. Stability analysis of potential failure zones along NH-305, India. *Nat. Hazards* **2016**, *83*, 1341–1357. [CrossRef]

5. Ansari, T.; Kainthola, A.; Singh, K.H.; Singh, T.N.; Sazid, M. Geotechnical and microstructural characteristics of phyllite derived soil; implications for slope stability, Lesser Himalaya, Uttarakhand, India. *Catena* **2021**, *196*, 104906. [CrossRef]

6. Zhong, R.; He, D.; Hu, J.M.; Duan, X.; Huang, J.C.; Cheng, X.P. Distribution and susceptibility assessment of collapses and landslides in the riparian zone of the xiaowan reservoir. *Chin. Geogr. Sci.* **2019**, *29*, 70–85. [CrossRef]

7. Xiao, T.; Segoni, S.; Chen, L.; Yin, K.; Casagli, N. A step beyond landslide susceptibility maps: A simple method to investigate and explain the different outcomes obtained by different approaches. *Landslides* **2019**, *17*, 627–640. [CrossRef]

8. Guzzetti, F.; Reichenbach, P.; Ardizzone, F.; Cardinali, M.; Galli, M. Estimating the quality of landslide susceptibility models. *Geomorphology* **2006**, *81*, 166–184. [CrossRef]

9. Pham, B.T.; Prakash, I.; Dou, J.; Singh, S.K.; Trinh, P.T.; Tran, H.T.; Le, T.M.; Phong, T.V.; Khoi, D.K.; Shirzadi, A.; et al. A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. *Geocarto Int.* **2019**, *35*, 1267–1292. [CrossRef]

10. Chang, K.T.; Merghadi, A.; Yunus, A.P.; Pham, B.T.; Dou, J. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. *Sci. Rep.* **2019**, *9*, 12296. [CrossRef]

11. Bzdok, D.; Altman, N.; Krzywinski, M. Machine learning: Supervised methods. *Nat. Methods* **2018**, *15*, 233–234. [CrossRef]

12. Van Westen, C.J.; van Asch, T.W.; Soeters, R.; Westen, C.J. Landslide hazard and risk zonation—Why is it still so difficult? *Bull. Eng. Geol. Environ.* **2006**, *65*, 167–184. [CrossRef]

13. Abraham, M.T.; Satyam, N.; Lokesh, R.; Pradhan, B.; Alamri, A. Factors affecting landslide susceptibility mapping: Assessing the influence of different machine learning approaches, sampling strategies, and data splitting. *Landslides* **2021**, *10*, 989. [CrossRef]

14. Pourghasemi, H.R.; Gayen, A.; Park, S.; Lee, C.W.; Lee, S. Assessment of landslide-prone areas and their zonation using logistic regression, logitboost, and naïvebayes machine-learning algorithms. *Sustainability* **2018**, *10*, 3697. [CrossRef]

15. Lee, S.; Ryu, J.-H.; Kim, I.-S. Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: Case study of Youngin, Korea. *Landslides* **2007**, *4*, 327–338. [CrossRef]

16. Colkesen, I.; Sahin, E.K.; Kavzoglu, T. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines, and logistic regression. *J. Afr. Earth Sci.* **2016**, *118*, 53–64. [CrossRef]

17. Tang, R.X.; Yan, E.C.; Wen, T.; Yin, X.M.; Tang, W. Comparison of logistic regression, information value, and comprehensive evaluating model for landslide susceptibility mapping. *Sustainability* **2021**, *13*, 3803. [CrossRef]

18. 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]

19. Sun, D.; Shi, S.; Wen, H.; Xu, J.; Zhou, X.; Wu, J. A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping. *Geomorphology* **2021**, *379*, 107623. [CrossRef]

20. Sevgen, E.; Kocaman, S.; Nefeslioglu, H.A.; Gokceoglu, C. A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression, ANN, and random forest. *Sensors* **2019**, *19*, 3940. [CrossRef]

21. Hong, H.; Liu, J.; Zhu, A.X.; Shahabi, H.; Pham, B.T.; Chen, W.; Pradhan, B.; Bui, D.T. A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China). *Environ. Earth Sci.* **2017**, *76*, 652–670. [CrossRef]

22. Li, J.Y.; Wang, W.D.; Han, Z. A variable weight combination model for prediction on landslide displacement using AR model, LSTM model, and SVM model: A case study of the Xinming landslide in China. *Environ. Earth Sci.* **2021**, *80*, 386. [CrossRef]

23. Kawabata, D.; Bandibas, J. Landslide susceptibility mapping using geological data, a DEM from ASTER images, and an Artificial Neural Network (ANN). *Geomorphology* **2009**, *113*, 97–109. [CrossRef]

24. 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]

25. Jones, S.; Kasthurba, A.K.; Bhagyanathan, A.; Binoy, B.V. Landslide susceptibility investigation for Idukki district of Kerala using regression analysis and machine learning. *Arab. J. Geosci.* **2021**, *14*, 838–854. [CrossRef]

26. Sun, X.; Chen, J.; Han, X.; Bao, Y.; Zhan, J.; Peng, W. Application of a GIS-based slope unit method for landslide susceptibility mapping along the rapidly uplifting section of the upper Jinsha River, South-Western China. *Bull. Eng. Geol. Environ.* **2020**, *79*, 533–549. [CrossRef]

27. Oh, H.-J.; Lee, S. Cross-application used to validate landslide susceptibility maps using a probabilistic model from Korea. *Environ. Earth Sci.* **2011**, *64*, 395–409. [CrossRef]

28. Preti, F.; Letterio, T. Shallow landslide susceptibility assessment in a data-poor region of Guatemala (Comitancillo municipality). *J. Agric. Eng.* **2015**, *46*, 85–94. [CrossRef]

29. Djukem, W.D.L.; Braun, A.; Wouatong, A.S.L.; Guedjeo, C.; Dohmen, K.; Wotchoko, P.; Fernandez-Steeger, T.M.; Havenith, H.-B. Effect of soil geomechanical properties and geo-environmental factors on landslide predisposition at Mount Oku, Cameroon. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6795. [CrossRef] [PubMed]

30. Sarkar, S.; Samanta, M. Stability analysis and remedial measures of a landslip at Keifang, Mizoram—A case study. *J. Geol. Soc. India* **2017**, *89*, 697–704. [CrossRef]

31. Dung, N.V.; Hieu, N.; Phong, T.V.; Amiri, M.; Costache, R.; Al-Ansari, N.; Prakash, I.; Van Le, H.; Nguyen, H.B.T.; Pham, B.T. Exploring novel hybrid soft computing models for landslide susceptibility mapping in Son La hydropower reservoir basin. *Geomat. Nat. Hazards Risk* **2021**, *12*, 1688–1714. [CrossRef]

32. Malek, Ž.; Zumpano, V.; Hussin, H. Forest management and future changes to ecosystem services in the Romanian Carpathians. *Environ. Dev. Sustain.* **2018**, *20*, 1275–1291. [CrossRef]

33. Jennifer, J.J. Feature elimination and comparison of machine learning algorithms in landslide susceptibility mapping. *Environ. Earth Sci.* **2022**, *81*, 489. [CrossRef]

34. Zhang, F.H.; Zhu, Y.Y.; Zhao, X.Z.; Zhang, Y.; Shi, L.H.; Liu, X.D. Research on the spatial distribution characteristics and identification of landslide potential hazard points supported by geographic factors. *Geomat. Inf. Sci. Wuhan Univ.* **2020**, *45*, 1233–1244.

35. Zhang, J.; Yuan, X. COVID-19 risk assessment: Contributing to maintaining urban public health security and achieving sustainable urban development. *Sustainability* **2021**, *13*, 4208. [CrossRef]

36. Tang, J.; Lichun, S. Geodetector-Based Livability Analysis of Potential Resettlement Locations for Villages in Coal Mining Areas on the Loess Plateau of China. *Sustainability* **2022**, *14*, 8365. [CrossRef]

37. Rong, G.; Li, K.; Han, L.; Alu, S.; Zhang, J.; Zhang, Y. Hazard mapping of the rainfall–landslides disaster chain based on GeoDetector and bayesian network models in shuicheng county, China. *Water* **2020**, *12*, 2572. [CrossRef]

38. Zheng, W.J.; Min, W.; He, W.G.; Ren, Z.K.; Liu, X.W.; Wang, A.G.; Xu, C.; Li, F. Distribution of the related disaster and the causative tectonic of the Minxian Zhangxian MS6.6 earthquakes on July 22, 2013, Gansu, China. *Seismol. Geol.* **2013**, *35*, 604–615.

39. Xu, C.; Xu, X.W.; Zhang, W.J. Compiling an inventory of landslides triggered by the Minxian-Zhangxian earthquake of July 22, 2013, and their spatial distribution analysis. *J. Eng. Geol.* **2013**, *21*, 736–749.

40. Pham, B.T.; Pradhan, B.; Bui, D.T.; Prakash, I.; Dholakia, M.B. 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]

41. Kavzoglu, T.; Sahin, E.K.; Colkesen, I. Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. *Eng. Geol.* **2015**, *192*, 101–112. [CrossRef]

42. Chen, W.; Zhang, S.; Li, R.; Shahabi, H. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Sci. Total Environ.* **2018**, *644*, 1006–1018. [CrossRef] [PubMed]

43. ASTER Global Digital Elevation Model. Available online: https://www.gscloud.cn (accessed on 12 February 2024).

44. China Geological Archives. Available online: http://www.ngac.cn (accessed on 2 February 2024).

45. Landsat 8 Satellite Data. Available online: https://www.gscloud.cn (accessed on 8 January 2024).

46. Resource and Environmental Science Data Center of the Chinese Academy of Sciences. Available online: http://www.resdc.cn (accessed on 17 December 2023).

47. GLOBELAND30. Available online: http://www.globallandcover.com (accessed on 22 February 2024).

48. 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]

49. Guo, Z.Z.; Shi, Y.; Huang, F.M.; Fan, X.M.; Huang, J.S. Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. *Geosci. Front.* **2021**, *12*, 101249. [CrossRef]

50. Sinarta, I.N.; Basoka, I.A. Safety factor analysis of landslides hazard as a result of rain condition infiltration on Buyan-Beratan Ancient Mountain. *J. Phys. Conf. Ser.* **2019**, *1402*, 22002. [CrossRef]

51. Shi, N.; Li, Y.; Wen, L.; Zhang, Y. Rapid prediction of landslide dam stability considering the missing data using XGBoost algorithm. *Landslides* **2022**, *19*, 2951–2963. [CrossRef]

52. 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]

53. Pham, B.T.; Prakash, I.; Jaafari, A.; Bui, D.T. Spatial prediction of rainfall-induced landslides using aggregating one-dependence estimators classifier. *J. Indian Soc. Remote Sens.* **2018**, *46*, 1457–1470. [CrossRef]

54. Lin, L.; Lin, Q.; Wang, Y. Landslide susceptibility mapping on a global scale using the method of logistic regression. *Nat. Hazards Earth Syst. Sci.* **2017**, *17*, 1411–1424. [CrossRef]

55. Aditian, A.; Kubota, T.; Shinohara, Y. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. *Geomorphology* **2018**, *318*, 101–111. [CrossRef]

56. Lan, H.; Peng, J.; Zhu, Y.; Li, L.; Pan, B.; Huang, Q.; Zhang, Q. Research on geological and surficial processes and major disaster effects in the Yellow River Basin. *Sci. China Earth Sci.* **2022**, *65*, 234–256. [CrossRef]

57. Devkota, K.C.; Regmi, A.D.; Pourghasemi, H.R.; Yoshida, K.; Pradhan, B.; Ryu, I.C.; Dhital, M.R.; Althuwaynee, O.F. Landslide susceptibility mapping using certainty factor, index of entropy, and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya. *Nat. Hazards* **2013**, *65*, 135–165. [CrossRef]
58. Schlögl, M.; Richter, G.; Avian, M.; Thaler, T.; Heiss, G.; Lenz, G.; Fuchs, S. On the nexus between landslide susceptibility and transport infrastructure—An agent-based approach. *Nat. Hazards Earth Syst. Sci.* **2019**, *19*, 201–219. [CrossRef]
59. Ju, Z.Y.; Wang, Z.H. Bayesian classification algorithm based on selective patterns. *J. Comput. Res. Dev.* **2020**, *57*, 1605–1616.
60. Sun, D.L.; Wen, H.J.; Wang, D.Z.; Xu, J.H. A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. *Geomorphology* **2020**, *362*, 107201. [CrossRef]
61. 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]
62. Huang, F.; Yin, K.; Huang, J.; Gui, L.; Wang, P. Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. *Eng. Geol.* **2017**, *223*, 11–22. [CrossRef]