remote sensing

Article

Risk Factor Detection and Landslide Susceptibility Mapping Using Geo-Detector and Random Forest Models: The 2018 Hokkaido Eastern Iburi Earthquake

Yimo Liu [1,2], Wanchang Zhang [1,*], Zhijie Zhang [3], Qiang Xu [4] and Weile Li [4]

[1] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; liuym@aircas.ac.cn
[2] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Department of Geography, University of Connecticut, Storrs, CT 06269, USA; zhijie.zhang@uconn.edu
[4] State Key Laboratory of Geohazard Prevention and Geo-Environment Protection, Chengdu University of Technology, Chengdu 610059, China; xq@cdut.edu.cn (Q.X.); liweile08@mail.cudt.edu.cn (W.L.)
* Correspondence: zhangwc@radi.ac.cn; Tel.: +86-010-8217-8131

Abstract: Landslide susceptibility mapping is an effective approach for landslide risk prevention and assessments. The occurrence of slope instability is highly correlated with intrinsic variables that contribute to the occurrence of landslides, such as geology, geomorphology, climate, hydrology, etc. However, feature selection of those conditioning factors to constitute datasets with optimal predictive capability effectively and accurately is still an open question. The present study aims to examine further the integration of the selected landslide conditioning factors with Q-statistic in Geo-detector for determining stratification and selection of landslide conditioning factors in landslide risk analysis as to ultimately optimize landslide susceptibility model prediction. The location chosen for the study was Atsuma Town, which suffered from landslides following the Eastern Iburi Earthquake in 2018 in Hokkaido, Japan. A total of 13 conditioning factors were obtained from different sources belonging to six categories: geology, geomorphology, seismology, hydrology, land cover/use and human activity; these were selected to generate the datasets for landslide susceptibility mapping. The original datasets of landslide conditioning factors were analyzed with Q-statistic in Geo-detector to examine their explanatory powers regarding the occurrence of landslides. A Random Forest (RF) model was adopted for landslide susceptibility mapping. Subsequently, four subsets, including the Manually delineated landslide Points with 9 features Dataset (MPD9), the Randomly delineated landslide Points with 9 features Dataset (RPD9), the Manually delineated landslide Points with 13 features Dataset (MPD13), and the Randomly delineated landslide Points with 13 features Dataset (RPD13), were selected by an analysis of Q-statistic for training and validating the Geo-detector-RF- integrated model. Overall, using dataset MPD9, the Geo-detector-RF-integrated model yielded the highest prediction accuracy (89.90%), followed by using dataset MPD13 (89.53%), dataset RPD13 (88.63%) and dataset RPD9 (87.07%), which implied that optimized conditioning factors can effectively improve the prediction accuracy of landslide susceptibility mapping.

Keywords: Geo-detector; Random Forest; feature selection; landslide susceptibility mapping

Citation: Liu, Y.; Zhang, W.; Zhang, Z.; Xu, Q.; Li, W. Risk Factor Detection and Landslide Susceptibility Mapping Using Geo-Detector and Random Forest Models: The 2018 Hokkaido Eastern Iburi Earthquake. *Remote Sens.* **2021**, *13*, 1157. https://doi.org/10.3390/rs13061157

Academic Editor: Christian Bignami

Received: 20 February 2021
Accepted: 15 March 2021
Published: 18 March 2021

1. Introduction

Landslides are the most common geological disasters that damage property and infrastructure and result in loss of life. Landslide susceptibility mapping is an important tool to optimize land use planning and policy to reduce damage from landslides to public property, infrastructure and people's lives [1,2]. Landslide susceptibility mapping refers to a division of the land into zones of hazard classes ranked according to different landslide occurrence probabilities based on an estimated significance of conditioning factors to the causes of landslides [3–5]. Landslide susceptibility is determined by qualitative and quantitative analyses of the conditioning factors obtained in the former disaster area [3,5–7]. Maps of landslide susceptibility are usually prepared at regional scales at middle-to- high spatial resolutions, which favor regional studies allowing rapid assessment, and hence, larger areas can be covered in short duration [2].

A broad range of methods and techniques have been proposed for landslide susceptibility assessment over the last decades, and they can be grouped into heuristically based, physically based and statistically based methods [1,8]. The heuristically based approaches utilize contributing factors and their weights determined by export knowledge, which is partially subjective [5,8,9]. The physically based approaches generally provide accurate results by detailed data of geotechnical engineering at site-specific locations at a localized scale, but suffer from expensive cost and lack of data for large-scale areas [1,8–10].The statistically based approaches assume that conditions leading to slope failure in the past are likely to cause landslides in the future [1]. These methods train models using datasets generated from various conditioning factors and landslide inventory maps and produce a quantitative map of landslide occurrence probabilities [1,8]. In the literature, some common statistical models used in landslide susceptibility mapping include logistic regression [11–13], frequency ratio [14,15], support vector machines [1,13], decision trees [13,16,17], artificial neural networks [1,11], etc.

Landslide occurrence is highly correlated with intrinsic variables that contribute the most to initiating landslides, which can be described as geology, geomorphology, climate, hydrology, hydrogeology, land cover/use as well as human activities [4,18–20]. Significant conditioning factors are essential for high-accuracy landslide susceptibility mapping; however, redundant information may cause noise that restricts accuracy [1,6,12]. Many methods have been proposed in the literature to select factors with higher predictive capability. For examples, cross-tabulations [21], binary logistic regression [5], information gain ratio [6,22], Geo-detector [8,12] and convolutional neural networks [23] were used for the identification and selection of the causal parameters.

Weights assigned to causal factors differ as the characteristics of conditioning factors vary from region to region [6,8]. For the area of the Moldavian Plateau in Romania, the statistic interpretations revealed that the lithology, the geomorphology and the topography had stronger relations with the landslide compared to climate and land cover/use [4]. In both hilly areas and lower mountain regions in Romania, the prevailing landslide causal factors, whose weights were assessed using logistic regression, were slope, land cover/use and slope height above channel network [5]. For the earthquake-induced landslides in Sichuan Province, China, rock mass and slope were evaluated to have the greatest relative contributions to landslides [12]. For the shallow landslides in the hilly region in the Valnerina area in Italy, it was observed that the distance to faults and lithology had a significant impact on landslides [21]. For the semi-arid mountain environment in Granada, Spain, elevation, lithology, slope angle and aspect were proved as the most relevant landslide determining factors [24].

Despite the many studies on landslide susceptibility and hazard modeling, effect and accurate feature selection are still open questions [23]. Together with landslide susceptibility mapping, the geographical detection method has been introduced as a new and powerful tool for feature selection. The integration of Geo-detector and statistically based approaches have shown potential to the application of accuracy improvement of landslide susceptibility models [8,12]. Therefore, investigation of the combined methods and techniques in the different landslide-prone areas in the world is highly necessary to acquire adequate background to explore reasonable improvements for landslide susceptibility mapping.

The main purpose of this study is to improve landslide susceptibility mapping's performance by integrating Geo-detector and the Random Forest model. The first stage was to examine further the integration of the select landslide conditioning factors with Q-statistic in Geo-detector for determining stratification and selection of landslide conditioning factors. Then, the Random Forest model was adopted for the datasets optimized using Geo-detector as a classification method for building a landslide susceptibility model. Finally, the receiver operating characteristics (ROC) were used for the assessment, validation and comparison of the four different Geo-detector-RF models derived with four selected composites of conditioning factors. The location chosen for the study was the Atsuma Town, where more than 3000 landslides triggered by the 2018 Mw 6.6 Hokkaido Eastern Iburi Earthquake over an area of about 360 km^2 with extensive damage (https://www.gsi.go.jp/ accessed on 6 July 2020).

2. Materials and Methods

2.1. Study Area

On September 5, 2018, a magnitude Mw 6.6 earthquake, with its epicenter located at 42°41′10″ N, 141°55′44″ E and focal depth at 35.0 km, struck Eastern Iburi, Hokkaido, Japan (https://earthquake.usgs.gov/ accessed on 6 July 2020). According to the focal fault model, a high-angle slip striking in the north-south direction happened in a rectangle reverse fault at 14 km wide, 15.9 km long and 16.2 km upper edge deep and caused the severe earthquake (https://www.gsi.go.jp/ accessed on 6 July 2020). The earthquake and its derived disasters, such as landslides and mudslides, caused a large number of casualties and public facilities damaged.

The study area (Figure 1) is located in Atsuma Town, Iburi, Hokkaido, Japan, the central disaster district of the 2018 Hokkaido Eastern Iburi earthquake. It spans longitudes from 141°48′ E to 142°07′ E, latitudes from 42°35′ N to 42°52′ N, with an area of about 408 km^2. The landslides triggered by the earthquake are mostly shallow, several meters deep, and distributed with high density over an area of approximately 18 km^2 [25,26].

The topography of the study area is characterized by the remarkable transition zone from the southwest coastal plains on the Pacific Ocean to the northeast hilly land. The topographic inclination follows the NE–SW direction with elevation generally lower than 604 m above sea level (a.s.l.) and mean elevation about 137 m a.s.l. The terrain is relatively flat, with an average slope angle of about 9.93°. Areas with slope less than 15° account for more than 75% of the region, while areas greater than 25° cover less than 5%. Owning to its geographic location in coastal area of Hokkaido facing the Pacific Ocean to the south, the study area has a humid continental climate according to the Köppen Climate classification (https://www.jma.go.jp/ accessed on 6 July 2020). The recorded average annual precipitation of this area is about 1014.3 mm, with the highest value of 171.6mm in August and the lowest value of 31.2 mm in February (https://www.jma.go.jp/ accessed on 6 July 2020).

Tectonic forces and faults caused by relative movements under the interactions between the Pacific, North American, Eurasian, and Philippine plates make the geological structure and tectonic evolution of the study area rather complicated [26]. Most subsidences of the study area are marine and non-marine sediments, geologically aged from Holocene to Jurassic with a total thickness of 4–5 m and covered by surface soil inter-bedded with pumice and ash [26]. Such geological structure and stratigraphy make this area prone to geological disasters such as earthquakes and landslides.

2.2. Spatial Database

2.2.1. Landslide Inventory Map

The landslide inventory map plays an important role in hazard assessment [1,4,6,24]. The landslide inventory map was prepared at 1:10,000 scale through a combination of field surveys and photo interpretation based on stereoscopic and pseudo-stereoscopic aerial photography (Figure 2a,b). Different types of aerial photographs were analyzed visually to prepare the landslide inventory map by the Geographical Survey Institute, Japan (GSI, hereafter) (https://www.gsi.go.jp/ accessed on 6 July 2020). A total of about 1000 landslides covering an area of about 48 km^2 were extracted in the study area and plotted as polygon features on the digital inventory map for further analysis.

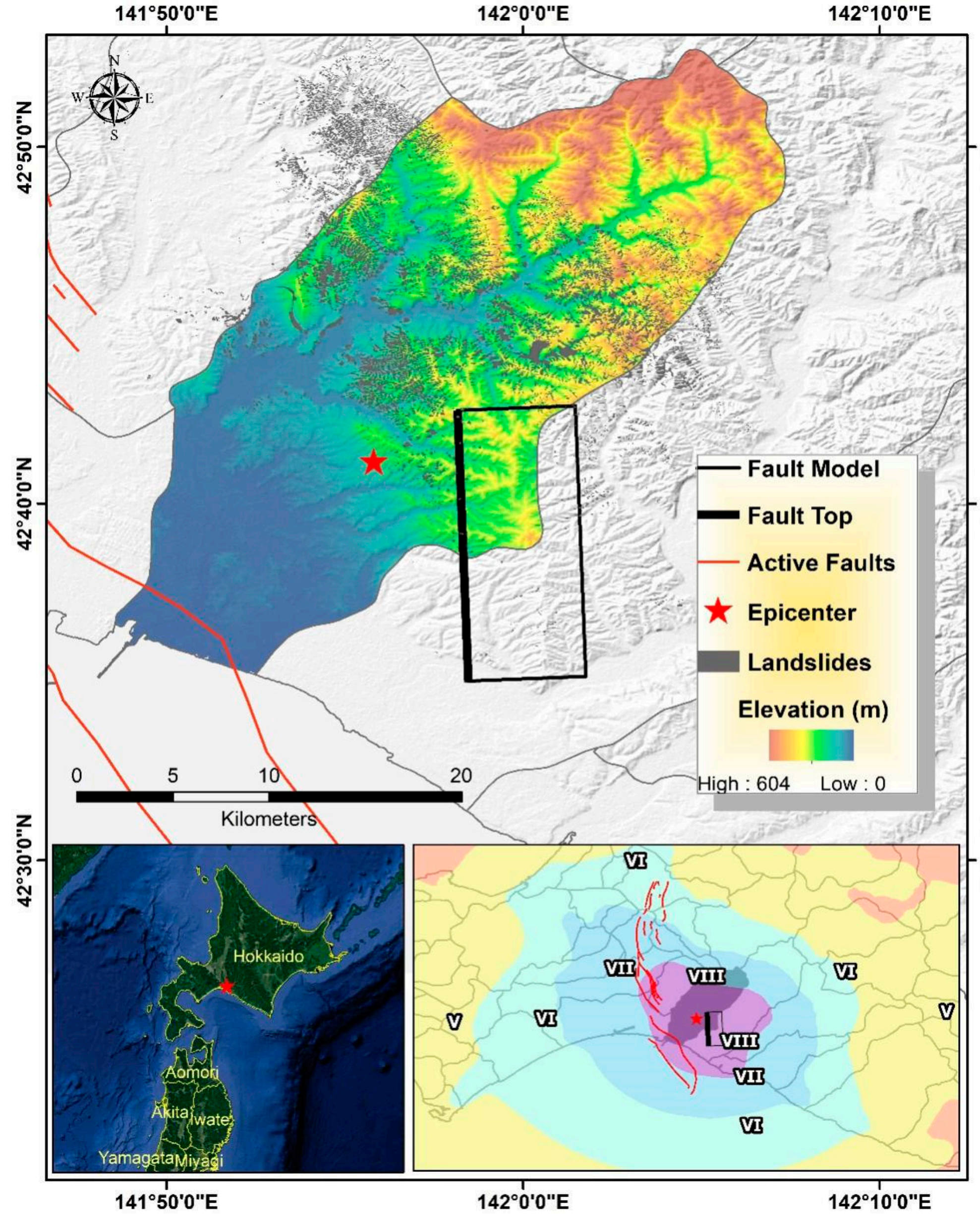

Figure 1. The digital maps showing the geographic location, topography, active faults, seismic magnitude and distribution of earthquake-triggered landslides over the study area, along with the location of the epicenter as well as the fault-model simulated results for the 2018 Eastern Iburi Earthquake at Hokkaido, Japan [2].

Only landslides with an area greater than 1000 m^2 were used for landslide susceptibility analysis [26]. Manual and random sampling approaches were utilized to generate landslide sample datasets. For manual sampling, given the high-resolution orthoimages and the existing landslide distribution vector from GSI Map Vector (https://maps.gsi.go.jp/vector/ accessed on 6 July 2020), the landslide sample points were extracted in the landslide source areas (Figure 2c), which were distinguished from debris deposits, while random sampling of landslide points (Figure 2d) and non-landslide points was derived with the aid of the software ArcGIS 10.6.

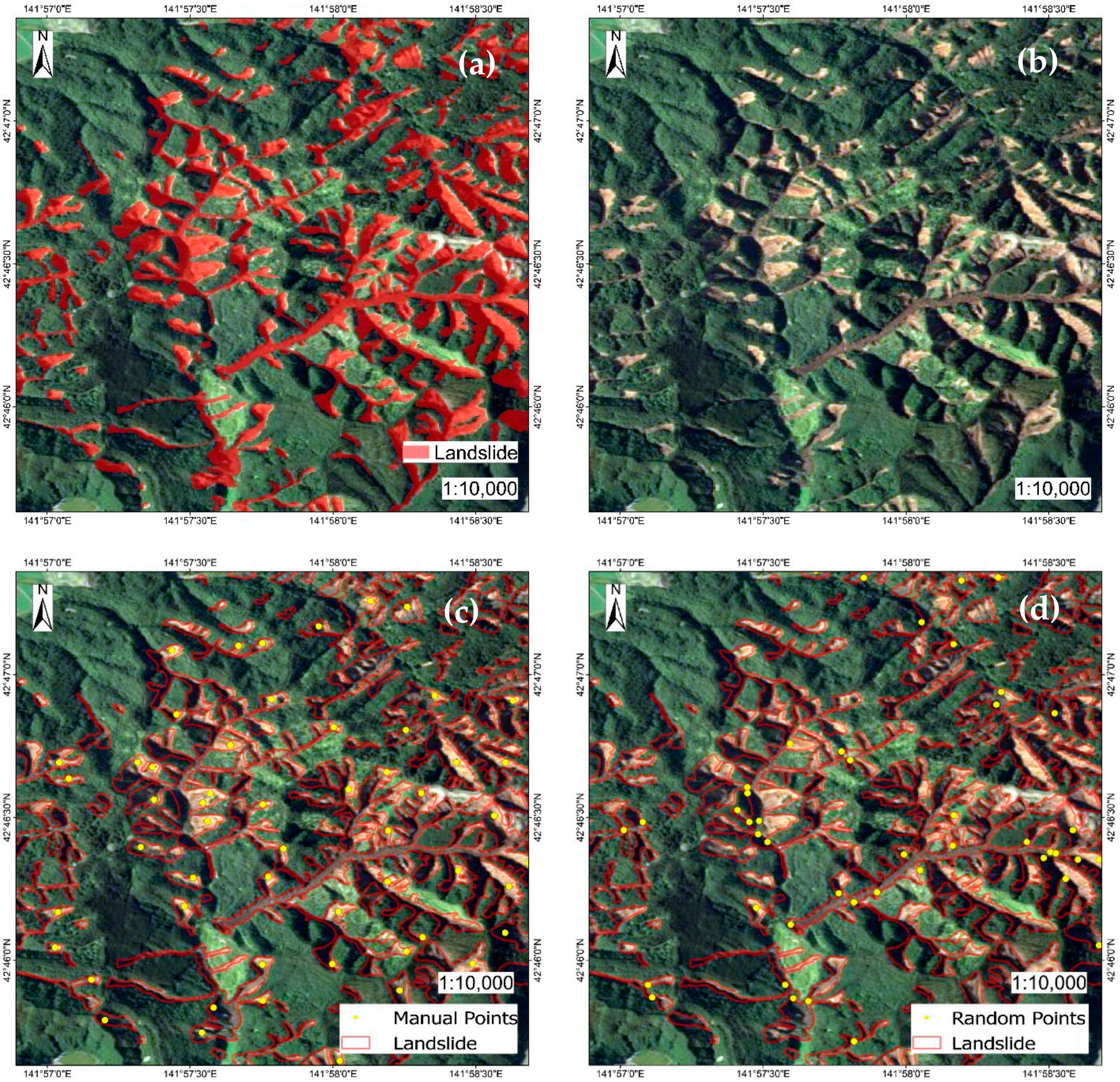

Figure 2. Part of (**a**) Landslide inventory map interpreted from (**b**) Aerial photograph by Geographical Survey Institute, Japan (GSI). (**c**) Distribution of manually sampled landslide points and (**d**) Distribution of randomly sampled landslide points.

1000 positive sample points were collected in the landslide area and divided into two subsets: 70% of the landslide points were used for model training, whereas the remaining 30% were used for the model validation, as shown in Figure 3. Meanwhile, 1000 negative sample points were randomly sampled from the landslide-free area and divided into training datasets and validation datasets at a ratio of 7:3 as well.

2.2.2. Landslide Conditioning Factors

Making a reference to the literature, i.e., Shao et al. (2019), Yi et al. (2019), etc., the appropriate initial set of conditioning factors was obtained using the data from the conventional field survey, observation station and remotely sensed information. The factors selected in this study could be divided into six categories: geology (lithology); geomorphology (elevation, slope, aspect, curvature); seismology (active fault, epicenter, peak ground velocity (PGV) and peak ground acceleration (PGA)); hydrology (topographic wetness index (TWI), river channels); land cover/use and human activity (roads).

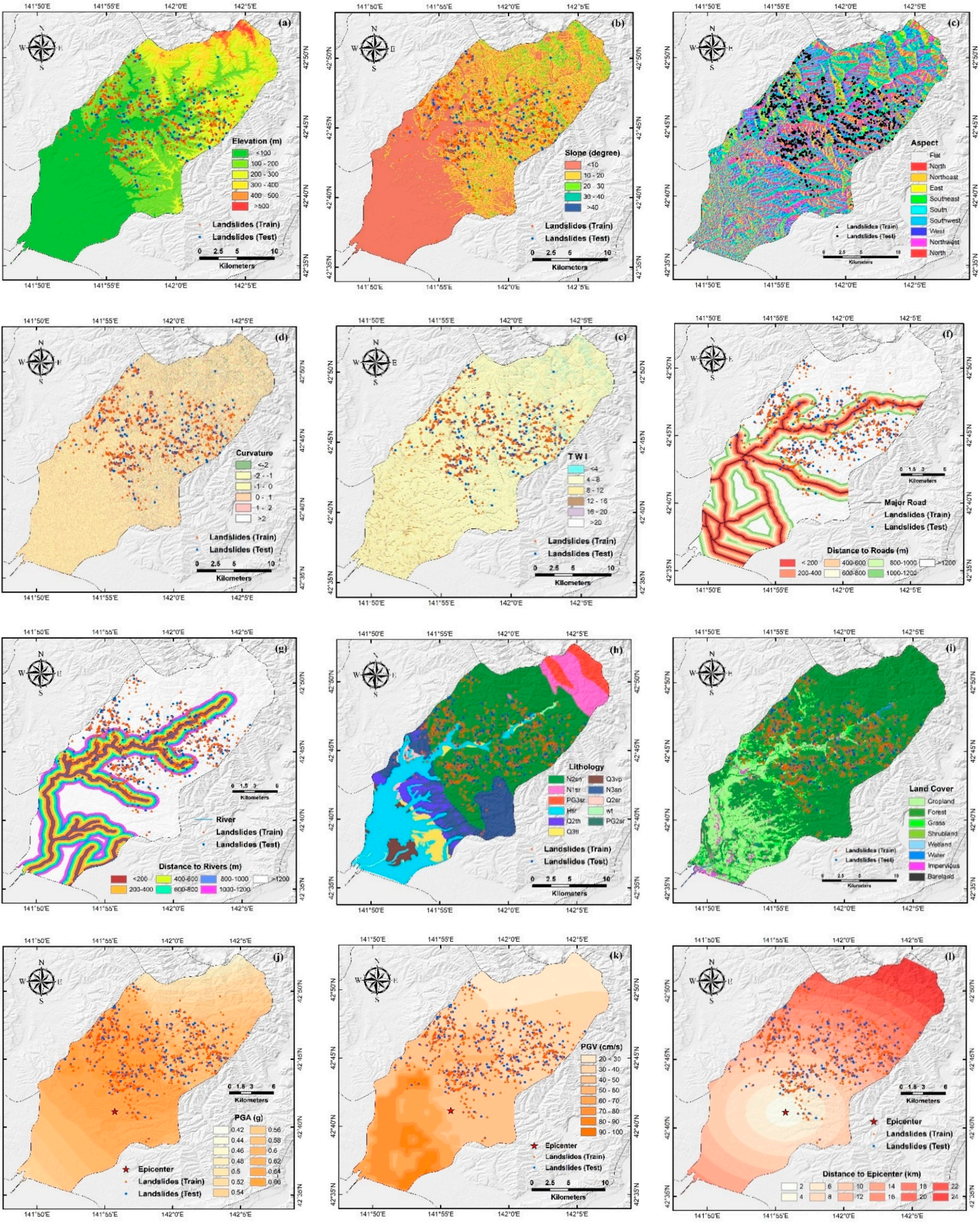

Figure 3. *Cont.*

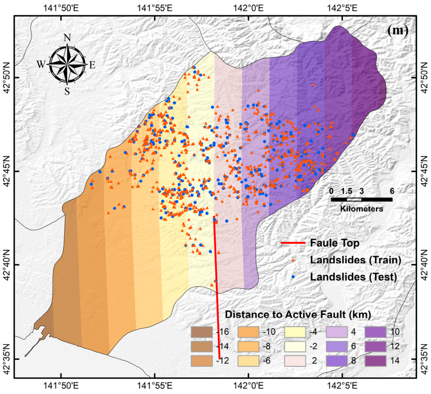

Figure 3. Landslide conditioning factor layers and their classes used for landslide susceptibility mapping in the study area. (**a**) Elevation, (**b**) Slope, (**c**) Aspect, (**d**) Curvature, (**e**) Topographic wetness index (TWI), generated from Shuttle Radar Topography Mission (SRTM) data, (**f**) Distance to roads, (**g**) Distance to rivers, digitized from GSI map, (**h**) Lithology, derived from GSI map, (**i**) Land cover, produced by Tsinghua University, (**j**) Peak ground acceleration (PGA), (**k**) Peak ground velocity (PGV), (**l**) Distance to epicenter, downloaded from the USGS and (**m**) Distance to active fault, collected from GSI. The landslide distribution used for training and testing was generated as described in Section 2.2.1.

A digital elevation model (DEM) of the study area was derived from Shuttle Radar Topography Mission (SRTM) with 1 arc-second (about 30 m) spatial resolution and was divided into 6 classes (Figure 3a). Based on DEM data, another four topographical conditioning factors were generated with the same resolution using the tools provided by software ArcGIS 10.6: slope angles were divided into 5 classes (Figure 3b), aspects were divided into 9 directions (Figure 3c), surface curvatures were divided into 6 classes (Figure 3d) and TWIs were divided into 6 categories (Figure 3e).

The surface curvature is a topographic index which represents the physical characteristics of the river basin [27,28]. Given an 3 × 3 altitude submatrix as shown in Figure 4, the surface curvature of its central point is calculated as the second derivative of its corresponding surface (https://desktop.arcgis.com/ accessed on 26 July 2020) [28]:

$$\begin{array}{c} D = \left[\frac{(Z_4+Z_6)}{2} - Z_5\right]/L^2 \\ E = \left[\frac{(Z_2+Z_8)}{2} - Z_5\right]/L^2 \\ Curvature = -2(D+E) \times 100 \end{array}, \tag{1}$$

where Z_i are the corresponding altitudes as shown in Figure 4, and 100 is the constant coefficient, as curvature is generally small [28].

The TWI is a topographic index which represents a theoretical estimation of the accumulation of flow at any point and is calculated as follows [29]:

$$TWI = Ln(a/\tan\beta), \tag{2}$$

where, in the terms of a raster DEM, a stands for upslope contributing area per unit contour, and β stands for local slope angle. a is calculated as follows:

$$a = \begin{cases} S{\cdot}A/L, & flow\ direction = Z_2, Z_4, Z_6, Z_8 \\ S{\cdot}A/\sqrt{2}L, & flow\ direction = Z_1, Z_3, Z_7, Z_9 \end{cases}, \tag{3}$$

where, in the terms of a raster DEM, S stands for area of a cell, A represents flow accumulation of the cell, L stands for size of a cell and Z_i are the corresponding locations as shown in Figure 4.

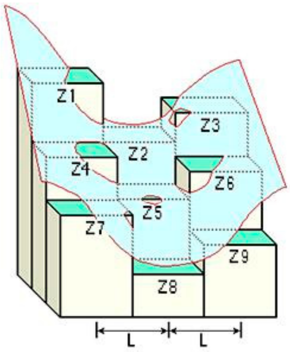

Figure 4. 3 × 3 altitude submatrix (https://desktop.arcgis.com/ accessed on 26 July 2020). Z_1 to Z_9 are the nine submatrix altitudes. Z_5 is the altitude of the central point. L is the distance between matrix points in the row and column directions and must be in the same units as Z [28].

Undercutting and erosion functions of rivers and roads on the natural topography may affect landslide susceptibility; therefore, distance to roads and rivers has often been taken into account in landslide susceptibility analysis [26]. In this study, the roads and rivers were digitized from the standard GSI Map Vector (https://maps.gsi.go.jp/ accessed on 6 July 2020), and both of them were classified into 7 classes with a 200m buffer interval and then converted to raster format with the same resolution as the DEM, respectively (Figure 3f,g).

Lithology is directly related to landslide susceptibility, so it is usually taken as conditioning factor in landslide analysis [5,12,14,26,30]. The lithology data of the study area was collected from the Seamless Digital Geological Map of Japan at 1:200,000 scale produced by the Geological Survey of Japan [31]. In total, about 11 geological formation units were classified according to age and lithology (Figure 3h), including Hsr (Late Pleistocene to Holocene marine and non-marine sediments), N1sr (Early Miocene to Middle Miocene marine and non-marine sediments), N2sn (Middle to Late Miocene non-marine sediments), N3sn (Late Miocene to Pliocene non-marine sediments), PG2sr (Middle Eocene marine and non-marine sediments), PG3sr (Late Eocene to Early Oligocene marine and non-marine sediments), Q2sr (Middle Pleistocene marine and non-marine sediments), Q2th (Middle Pleistocene higher terrace), Q3tl (Late Pleistocene lower terrace), Q3vp (Late Pleistocene non-alkaline pyroclastic flow volcanic rocks) and wt (water), specifically. Most landslides slid down over the air-fall pumice and ash layers [22]. Over 90% of the landslides occurred in central Atsuma, where the main lithology was N2sn. Worth mentioning at this time is that N2sn is the dominant lithology in the region, accounting for around 48%, followed by Hsr (20%) and N3sn (12%). In addition, landslide occurrence ratios within each geological formation units were calculated for a deeper understanding of the relation between lithology and landslides. The lithology with the largest landslide occurrence ratio is N2sn (18%), followed by Q2sr (8%), N3sn (3%), wt (2%), Hsr (1%), Q3tl (1%), Q2th and the others are no more than 1%.

Land cover affects surface water and soil conditions and thus influences landslide susceptibility. In this study, the land cover map (Figure 3i) was downloaded from the 10-m resolution global land cover datasets from 2017, produced by Tsinghua University [32].

Amplitude and position of ground deformation are strongly related to seismic landslides, and earthquake-triggered landslides are usually found in the vicinity of active faults [14]. Based on the availability of seismic data, four conditioning factors were constructed: PGA, PGV, distance to active fault and distance to epicenter. The PGA data ranging from 0.42 g (gravitational acceleration, approximately 9.8 m/s^2) to 0.66 g in the study area was classified into 13 classes with 2 g intervals (Figure 3j), while the PGV data ranging from 24 cm/s to 92 cm/s in the study area was divided into 8 categories with 10 cm/s intervals (Figure 3k). Both of them were downloaded from the USGS (https://earthquake.usgs.gov/ accessed on 6 July 2020) and were converted to raster format with the same resolution as the DEM, respectively. The distance was generated from the ring buffers in 2 km intervals with the epicenter as the origin (Figure 3l). The active fault data was derived from the fault model of the 2018 Hokkaido Eastern Iburi Earthquake constructed by GSI and was made up of buffers with 2 km intervals in the study area. The distance to the active fault was divided into 15 categories with a sigh and distance combined (Figure 3m), where the positive value represented the location to the west of the fault and the negative value represented the location to the east of the fault, respectively.

2.3. Methodology

As a binary classification approach, the quality and quantity of independent conditioning factors adopted in a model for landslide susceptibility mapping greatly affect the accuracy of the model. To assess the optimization effects of Geo-detector on landslide conditioning factors, two main steps were conducted in this study. First, preparation of training and validation datasets to select and determine weights of each conditioning factor by using Geo-detector; in this process, 4 datasets were generated according to different optimized combinations of conditioning factors, respectively. Next, the Random Forest Model for landslide susceptibility mapping was trained with 4 different optimized combinations of conditioning factors and validated for final mapping. The receiver operating characteristics (ROC), a statistical evaluation measure, was used to assess the predictive capability of the model trained by the 4 different datasets.

2.3.1. Geo-Detector

The quality and the quantity of the conditioning factor data layers used are essential for the reliability of landslide susceptibility mapping and the evaluation and optimization of conditioning factors regarding their predictive capabilities are vital to improving the accuracy of a model for landslide susceptibility mapping [2].

Geo-detector is a tool to detect and utilize spatial stratified heterogeneity (SSH), a basic characteristic of geographical phenomena being applied in many fields of natural and social sciences for assessing the optimization effects of the combination of conditioning factors [33]. An assumption beyond the Geo-detector is that the greater the influence of an independent variable on a dependent variable, the more similar their spatial distribution will be. The Q-statistic in Geo-detector that is used to measure spatial distribution similarity is calculated according to Wang et al. (2012) as follows:

$$q = 1 - \frac{\sum_1^l N_i \sigma_i^2}{N\sigma^2}, \tag{4}$$

where i = 1, 2, . . . , L, represent the stratum i of variables and N_i and N represent the unit numbers of stratum i and the whole region of interest, respectively. σ_i^2 and σ^2 represent the variance of stratum i and the whole area, respectively, and the second term in the right side of the equation above denotes the ratio of within sum of squares to the total sum of squares, $q \in [0, 1]$. According to different bases of stratification, the Q-statistic contains three meanings: SSH detection, factor detection and interactive factor detection. In SSH detection, stratification is based on the variable itself, and the more obvious the SSH is, the larger the Q-statistic will be; in factor detection, stratification of the dependent variable Y is based on the independent variable X, so a larger Q-statistic represents stronger explanatory power of the independent variable X on the dependent variable Y; while in interactive factor detection, the basis of stratification is the overlay of two independent variable X1 and X2 and the Q-statistic obtained can be compared with the Q-statistic of two single variables to evaluate the interaction, which can be divided into five types: nonlinear weakening, single-factor nonlinear weakening, double-factor enhancement, independent and nonlinear enhancement. For detailed explanations of Geo-detector, please refer to Wang et al. (2017).

For samples in training datasets, landslide samples were assigned to a value of "1" while non-landslide samples were assigned to a value of "0"; they were taken as the dependent variable Y used in Geo-detector. Values for the 13 initial landslide conditioning factors that were extracted from the conditioning factor maps generated previously were taken as the independent variables Xs for Geo-detector. The Q-statistic of each independent variable X was calculated according to Equation (4) to perform feature selection. The larger the Q-statistic was, the stronger the explanatory power of the conditioning factor on landslide occurrence, and vice versa.

2.3.2. Dataset Generation Based on Geo-Detector

As mentioned above, initial conditioning factor datasets were discretized in the first step. Next, Geo-detector was used to detect the contributions of all potential influencing factors to the spatial stratified heterogeneity of the landslides. Landslide sample points with truth values and conditioning factor classification values as attributes were input into Geo-detector. Next, according to the distribution of output Q-statistics and the geographical correlation between the determinants and landslide distribution, variables that ranked higher in the Q-statistics were selected as the independent variables in the Random Forest model.

2.3.3. Random Forest Model

A decision tree is a hierarchical model that divides feature space into sub-spaces as homogeneously as possible by recursive partitioning in order to achieve sample classification [34,35]. The implementation of an individual decision tree is simple and easy to explain, but it is prone to over-fitting, which is manifested as good performance on the training set but a poor generalization of the data on the validation set. A Random Forest is a kind of ensemble algorithm that combines tree predictors to vote for the most popular class to achieve significant reduction of generalization error and improvement of classification accuracy [36,37].

The base tree classifier used in this study was the Classification and Regression Tree (CART, hereafter), using the Gini Index as a measure of impurity to determine the features and thresholds for splitting nodes, and pruning was performed by the validation data independently from training data to suppress over-fitting [34].

For binary classification, where categories are represented by 0, 1, respectively, the Gini Index of nodes, where there is a total of samples with features, can be calculated as follows [35]:

$$G(m) = 2 \cdot \frac{\sum_{i \in N_m} y_i = 0}{N_m} \cdot \frac{\sum_{i \in N_m} y_i = 1}{N_m}, \quad (5)$$

When node m is split using feature x_j and threshold t, the impurity of the partition is evaluated as follows [35]:

$$I(j,t) = G(m(x_j <= t)) + G(m(x_j > t)), \quad (6)$$

Lower impurity represents the stronger partition effects. The parameters j and t, which minimize the impurity, are selected to split the node m.

The Random Forest reduces variance and improves the robustness of the model by injecting randomness and combining the average value of the predictors. Randomness is mainly realized by two parameters, the number of selected features and samples. Each tree is trained by a certain number of samples randomly selected from the whole dataset. When splitting nodes, the impurities of the partition are calculated only using a certain number of features rather than all of them [35,36]. After all the trees are generated, the scores of all trees are combined to obtain an average value [35].

2.3.4. The Receiver Operating Characteristic Curve

For validation and comparison of landslide susceptibility models with different datasets of combined conditioning factors as inputs, the receiver operating characteristic (ROC) curve was adopted on training datasets for examining training accuracy as well as on validation datasets for prediction accuracy.

To investigate the prediction performance of the models, the false positive rate (FPR) and true positive rate (TPR) were taken as the horizontal and vertical axis, respectively, to obtain the ROC and area under the ROC (AUC) [35,38]. For a binary classification, samples were assigned as positive and negative, representing landslide and non-landslide, respectively. False positive (FP) stands for the samples wherein prediction class is landslide but the ground truth is not, while true positive (TP) stands for the samples wherein both prediction class and ground truth are landslide. Hence, FPR denotes the ratio of the number of FP over the number of all real landslides; meanwhile, TPR denotes the ratio of the number over the number of all the non-landslides in the ground truth. When plotting the ROC curve, the model prediction scores of all samples are ranked in descending order. The evaluation method considers that the sample should be assigned to the positive class if its score is greater than c, where the threshold c denotes the likelihood of occurrence of a landslide and varies from [0, 1]. A fixed threshold c corresponds to a certain set of coordinates composed of (FPR, TPR). In particular, a threshold c, equaling 0, corresponds to a coordinate (1, 1), while a threshold c, equaling 1, corresponds to a coordinate (0, 0). With the group of sets of coordinates, the ROC is given, as well as the AUC [35,38]. The AUC ranges from 0 to 1, where 1 represents that all pixels are correctly classified and 0 indicates a poor predictive capacity of the model [39].

3. Results

3.1. Geo-Detector and Dataset Generation

In this study, the explanatory powers of conditioning factors on landslide distribution were identified by Q-statistics using Geo-detector. Two training datasets of landslide and non-landslide points were generated by random generation and manual sampling, respectively, and were taken as the inputs to Geo-detector, while Q-statistics of every conditioning factor (Figure 5) were the output. The results demonstrated that the distribution of relative contribution of all the potential landslide conditioning factors calculated from the Manual Points with 13 Features Dataset (MPD13, hereafter) and Random Points with 13 Features Dataset (RPD13, hereafter) were basically in the similar variation trend with the same significant data gap, in which seven conditioning factors with larger Q-statistics, including elevation, slope, lithology, distance to fault, distance to epicenter, PGA and PGV, and the remaining six factors with lower explanatory power, including aspect, curvature, TWI, distance to road, distance to river and land cover can be recognized. Lithology had the highest relative contribution to the landslide distribution for the two examined datasets, with a Q-statistic value of 0.318 in MPD and a Q-statistic value of 0.277 in RPD, respectively, followed by PGV, epicenter, and fault, etc., while river and road contributed the least.

Considering the dimensions of feature space and their explanatory powers, the first nine variables with satisfied critical values ($q > 0.04$ both in MPD and RPD) regarding to their relative contributions were selected as the optimized feature space. The values of the selected variables were extracted and combined with landslide tags to generate two new datasets: Manual Points with 9 Features Dataset (MPD9, hereafter) and Random Points with 9 Features Dataset (RPD9, hereafter). Together with the two original datasets composed of all 13 conditioning factors and landslide labels, a total of four datasets, including the training dataset and the validation dataset, were used in the following Random Forest model for landslide susceptibility mapping.

3.2. Model Accuracy Assessment and Comparison

Random Forest models were built using the four abovementioned datasets generated from different sampling methods and feature combinations based on the results of Geo-detector. To investigate the prediction performance of the models, FPR and TPR were taken as the horizontal and vertical axes, respectively, to obtain the ROC and AUC as shown in Figure 6. Performances of the Random Forest models trained from those four datasets were satisfactory in general, with all AUC values above 85%. More detailed observation suggested that the model trained by MPD9 outperformed the others with the highest AUC value of 89.90%, about 0.4% higher than the model trained by MPD13, which implied that a smaller number of conditioning factors was sufficient to produce reasonable results even with higher accuracy. Generally, model performance was ranked the best from the model trained by MPD followed by the model trained by RPD13 (88.63%) and the model trained by RPD9 (87.07%), respectively, which suggested that the precision of the model trained by manually selected samples was indeed improved compared with the one trained by randomly selected samples.

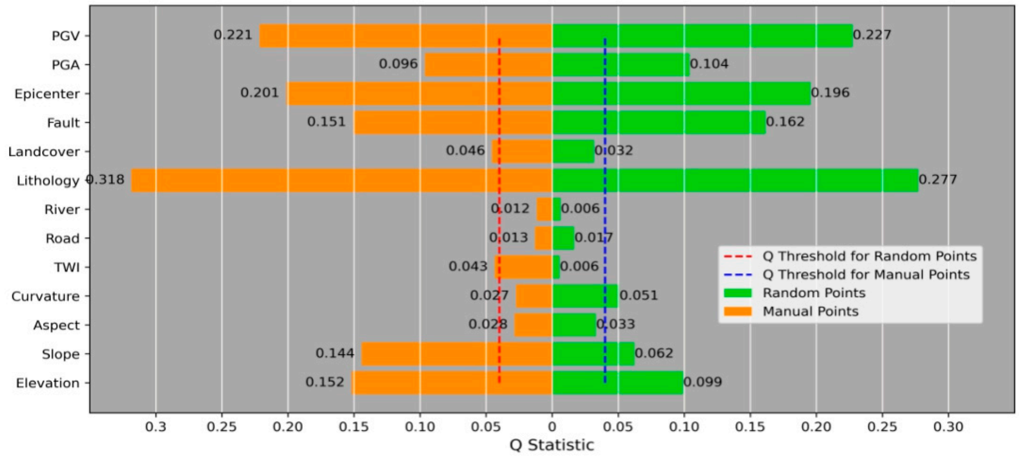

Figure 5. Q-statistics of landslide conditioning factors evaluated by Geo-detector.

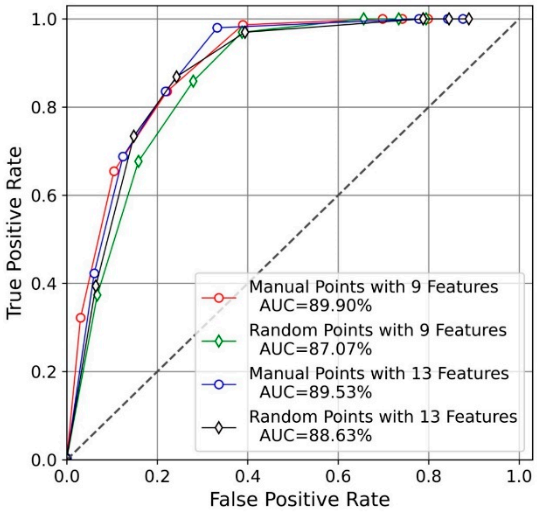

Figure 6. The receiver operator characteristics (ROC) curves for evaluating performances of Random Forest Models trained by Manual Points with 9 Features Dataset (MPD9), Random Points with 9 Features Dataset (RPD9), Manual Points with 13 Features Dataset(MPD13), Random Points with 13 Features Dataset (RPD13) datasets in landslide susceptibility mapping. AUC is the area under the ROC.

3.3. *Landslide Susceptibility Mapping*

The landslide susceptibility indexes of all the pixels in the study area were calculated by using the trained classification models. In order to clearly visualize the predicted landslide susceptibility distribution, the indexes of landslide susceptibility over the study area were classified into five categories as presented in Figure 7: very high (80–100%), high (60–80%), moderate (40–60%), low (20–40%) and very low (0–20%) by using the equal interval method, and the relationship between the percentage of landslide truth value and the percentage of grading area was exhibited as a curve [1,14,40].

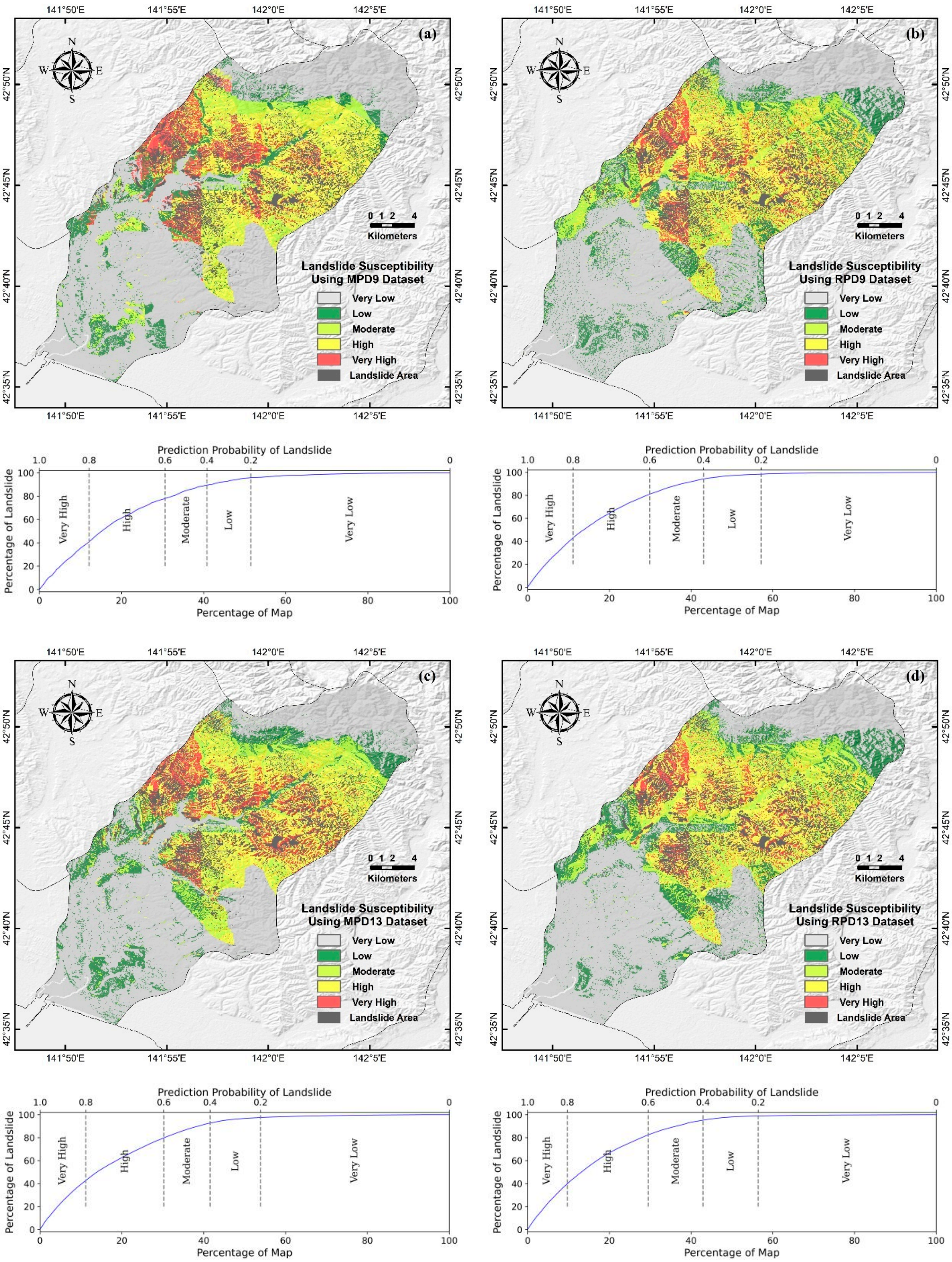

Figure 7. Landslide susceptibility map generated by the models trained by (**a**) MPD9, (**b**) RPD9, (**c**) MPD13, and (**d**) RPD13, respectively.

Details of the distribution of landslides in each relative grading area are listed in Table 1. It was observed that the landslide susceptibility maps generated by the models trained from the four datasets had consistent trends with slight differences. The category that accounted for the largest percentage of the study area was "very low", nearly half, followed by category "high", with approximately 20 percent. The remaining three categories, "very high", "moderate" and "low", were basically at the same level; however, for MPD9, the proportion of "very high" was the least, while for RPD13, it was slightly greater than the other two. In addition, it was obvious that the higher the landslide susceptibility, the higher the landslide distribution density tended to be. The classification of landslide susceptibility obtained from the four datasets suggested that in the category "very high", there were 30–40 km^2 of landslide per 100 km^2 on average, while in the "very low" category, there were almost no landslides. The distribution density of landslides in the middle three levels decreased by approximately half with the decrease of landslide susceptibility. At the same time, the proportion of landslides in the region showed the same trend, which decreased with the decrease of the susceptibility. Except for MPD13, category "high" was about 5 percent higher than category "very high". In general, for the four datasets, landslides in areas of the landslide susceptibility levels of "high" and "very high" accounted for nearly 80 percent of the total landslide areas, while level "moderate" accounted for half or more of the remaining 20% and level "low" was slightly higher than level "very low".

Table 1. Landslide distribution of landslide susceptibility derived from the models trained by using MPD9, RPD9, MPD13 and RPD13, respectively.

A	B	C	D	E	F	G	H	I
Datasets	**Susceptibility**	**Pixels**	**Landslide Pixels**	**Density**	**Map Pixels**	**Percentage of Map**	**Map Landslide Pixels**	**Percentage of Landslide**
(A)	**(B)**	**(C)**	**(D)**	**(D/C)**	**(F)**	**(C/F)**	**(H)**	**(100D/H)**
MPD9	Very High	54066	17339	0.32	574471	9.41	52190	33.22
	High	132482	24693	0.19	574471	23.06	52190	47.31
	Moderate	55040	5476	0.10	574471	9.58	52190	10.49
	Low	52676	2599	0.05	574471	9.17	52190	4.98
	Very Low	280207	2083	0.01	574471	48.78	52190	3.99
RPD9	Very High	53832	19429	0.36	574471	9.37	52190	37.23
	High	127100	24195	0.19	574471	22.12	52190	46.36
	Moderate	74226	6114	0.08	574471	12.92	52190	11.71
	Low	75861	1787	0.02	574471	13.21	52190	3.42
	Very Low	243452	665	0.00	574471	42.38	52190	1.27
MPD13	Very High	56965	20252	0.36	574471	9.92	52190	38.80
	High	121146	22143	0.18	574471	21.09	52190	42.43
	Moderate	65411	6331	0.10	574471	11.39	52190	12.13
	Low	66750	2272	0.03	574471	11.62	52190	4.35
	Very Low	264199	1192	0.00	574471	45.99	52190	2.28
RPD13	Very High	47632	18392	0.39	574471	8.29	52190	35.24
	High	131947	25964	0.20	574471	22.97	52190	49.75
	Moderate	73086	5671	0.08	574471	12.72	52190	10.87
	Low	71690	1608	0.02	574471	12.48	52190	3.08
	Very Low	250116	555	0.00	574471	43.54	52190	1.06

4. Discussion

A new, integrated model for landslide susceptibility analysis based on Geo-detector and the Random Forest method has been proposed in this study. The new model was carried out in the main disaster area of eastern Iburi in Hokkaido in Japan where there were widespread and extremely severe earthquake-triggered landslides. The prediction accuracy of landslide susceptibility mapping showed improvement by taking full advantage of spatial structure information and removing redundant information.

In this study, two sampling methods of landslide points were used for considering that the landslide distribution map downloaded from GSI did not distinguish between landslide source and debris deposits, while the location of selected samples may have had a certain influence on the prediction accuracy of the landslide susceptibility mapping model [26]. The results demonstrated that the accuracy of landslide prediction was improved by limiting the landslide samples of the landslide source to 1–2 percent compared to randomly sampling points in the case of the same dimension of feature spaces.

The accuracy of the prediction depends not only on the method chosen but also on the quantity and quality of conditioning factors. Previous studies have shown that a small number of conditioning factors are sufficient to produce landslide susceptibility maps with a reasonable quality [1,6,12]. After abandoning the last four factors with smaller contributions, as evaluated by Geo-detector in manually selected datasets, results suggested that the AUC of the classifier had a particular rising. This finding revealed that a less complex landslide susceptibility model of more reasonable quality could be expected for a small but sufficient number of conditioning factors integration. In the case of random sampling, the prediction accuracy of the model decreases, with the dimension of the feature space decreasing from 13 to 9, showing an opposite trend compared with previous results. These results demonstrated that the quality and quantity of conditioning factors determine the performance of landslide susceptibility modeling together. Removing redundant information should be carried out while maintaining the dimension of feature space at a reasonable level based on the analysis of landslide types and the characteristics of the study area.

However, there are still some limitations to the proposed method that need to be improved. Firstly, a detailed and accurate landslide inventory map is the first and most important stage of landslide susceptibility analysis, which is defined as using the conditions of slope failure in the past and present to assess the likelihood of causing landslides in the future [1,4,14]. In this study, the extracted landslide distribution area comprised landslide scarps and debris deposits in which there were shallow landslides of planar types as well as some deep-seated landslides of dip-slipping types [25]. We did not accurately distinguish the landslide types before due to the limitations of historical images. Further work should be carried out on the type purification of more detailed landslide inventories. Secondly, landslide susceptibility mapping employs topological, environmental, geological and hydrological parameters; an increase in the number of conditioning factors could result in improved accuracy, whereas the noise in the independent variables may well reduce predictive quality [1,6,12]. In the proposed method, different accuracy variation trends were obtained in the two datasets with the reduction of the dimensions of feature spaces. The effect of the dynamic combination of quantity and quality of conditioning factors on the predictive ability of the model is not explored further here. Regardless, if more variation in dimensions of feature space were included, the results would have been more convincing.

5. Conclusions

Landslide susceptibility mapping is considered to be an important step in landslide risk assessment, carrying great significance for reducing damage to life and properties caused by disasters. The process of landslide susceptibility analysis consists of three stages. At the first stage, landslide inventory maps and conditioning factors datasets should be collected and preprocessed into common formats, in which procedure features should be evaluated and cleaned. Modeling should include model training and model evaluation based on the detailed and accurate datasets generated in the first phase. Next, the resulting model should be applied to predict the likelihood of occurrence of landslides in the area of interest and make further analyses or decisions based on the results. In this study, with multi-geoscience information data as the inputs and Geo-detector as the feature selection method, the Random Forest model was used for landslide susceptibility analysis and model evaluation was carried out through ROC. The results illustrated the improvements and reliability after purification and putting limitations on the location of landslide sampling points and demonstrated a reasonable predictive quality when removing factors with little predictive ability. Geo-detector-RF model has shown a reliable assessment and predictive ability in landslide susceptibility analysis, and the integration of Geo-detector and some other machine learning techniques should be taken into account for advancements in future studies.

Author Contributions: Conceptualization, Y.L., W.Z., Z.Z., Q.X. and W.L.; methodology, Y.L., W.Z., Z.Z., Q.X. and W.L.; validation, Y.L.; analysis, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L., W.Z. and Z.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was jointly financed by the National Key R & D Program of China [Grant No. 2016YFA0602302] and the Key R & D and Transformation Program of Qinghai Province [Grant No. 2020-SF-C37].

Data Availability Statement: The landslide inventory map data presented in this study are openly and freely available at https://www.gsi.go.jp/ accessed on 6 July 2020; DEM data presented in this study are openly and freely available at https://earthexplorer.usgs.gov/ accessed on 6 July 2020; Roads and rivers data presented in this study are obtained at https://www.gsi.go.jp/ accessed on 6 July 2020; Lithology data presented in this study are obtained at https://www.gsi.go.jp/ accessed on 6 July 2020; Land cover data presented in this study are obtained at http://data.ess.tsinghua.edu.cn/ accessed on 6 July 2020; Earthquake related data presented in this study are obtained at https://earthquake.usgs.gov/ accessed on 6 July 2020.

Acknowledgments: The authors would like to thank Gongpeng's group, GSI and USGS for providing the research data; the graduate students in Wanchang Zhang's group for open discussions in weekly seminars; and all the scholars for their reference researches.

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

References

1. Tien Bui, D.; Tuan, T.A.; Klempe, H.; Pradhan, B.; Revhaug, I. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* **2015**, *13*, 361–378. [CrossRef]
2. Bălteanu, D.; Micu, M.; Jurchescu, M.; Malet, J.-P.; Sima, M.; Kucsicsa, G.; Dumitrică, C.; Petrea, D.; Mărgărint, M.C.; Bilaşco, Ş.; et al. National-scale landslide susceptibility map of Romania in a European methodological framework. *Geomorphology* **2020**, 371. [CrossRef]
3. Guzzetti, F.; Carrara, A.; Cardinali, M.; Reichenbach, P. Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology* **1999**, *31*, 216. [CrossRef]
4. Mărgărint, M.C.; Niculiţă, M. Landslide Type and Pattern in Moldavian Plateau, NE Romania. In *Landform Dynamics and Evolution in Romania*; Springer Geography; Springer: Cham, Switzerland, 2017; pp. 271–304. [CrossRef]
5. Mărgărint, M.C.; Grozavu, A.; Patriche, C.V. Assessing the spatial variability of weights of landslide causal factors in different regions from Romania using logistic regression. *Nat. Hazards Earth Syst. Sci. Discuss.* **2013**, *1*, 1774. [CrossRef]
6. Jebur, M.N.; Pradhan, B.; Tehrany, M.S. Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. *Remote Sens. Environ.* **2014**, *152*, 150–165. [CrossRef]
7. 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]
8. Luo, W.; Liu, C.-C. Innovative landslide susceptibility mapping supported by geomorphon and geographical detector methods. *Landslides* **2017**, *15*, 465–474. [CrossRef]
9. van Westen, C.J.; Castellanos, E.; Kuriakose, S.L. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Eng. Geol.* **2008**, *102*, 112–131. [CrossRef]
10. Canli, E.; Thiebes, B.; Petschko, H.; Glade, T. Comparing physically-based and statistical landslide susceptibility model outputs—A case study from Lower Austria. In Proceedings of the EGU General Assembly 2015, Vienna, Austria, 12–17 April 2015.
11. Pradhan, B.; Lee, S. Landslide susceptibility assessment and factor effect analysis: Backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environ. Model. Softw.* **2010**, *25*, 747–759. [CrossRef]
12. Yang, J.; Song, C.; Yang, Y.; Xu, C.; Guo, F.; Xie, L. New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway Basin, Sichuan Province, China. *Geomorphology* **2019**, *324*, 62–71. [CrossRef]
13. Kavzoglu, T.; Sahin, E.K.; Colkesen, I. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* **2013**, *11*, 425–439. [CrossRef]
14. Yi, Y.; Zhang, Z.; Zhang, W.; Xu, Q.; Deng, C.; Li, Q. GIS-based earthquake-triggered-landslide susceptibility mapping with an integrated weighted index model in Jiuzhaigou region of Sichuan Province, China. *Nat. Hazards Earth Syst. Sci.* **2019**, *19*, 1973–1988. [CrossRef]

15. Lee, S.; Pradhan, B. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* **2006**, *4*, 33–41. [CrossRef]
16. Nefeslioglu, H.A.; Sezer, E.; Gokceoglu, C.; Bozkir, A.S.; Duman, T.Y. Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey. *Math. Probl. Eng.* **2010**, *2010*, 1–15. [CrossRef]
17. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* **2013**, *51*, 350–365. [CrossRef]
18. Capecchi, V.; Perna, M.; Crisci, A. Statistical modelling of rainfall-induced shallow landsliding using static predictors and numerical weather predictions: Preliminary results. *Nat. Hazards Earth Syst. Sci.* **2015**, *15*, 75–95. [CrossRef]
19. Fell, R.; Corominas, J.; Bonnard, C.; Leroi, E.; Cascini, L. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Eng. Geol.* **2008**, *102*, 83–84. [CrossRef]
20. Guzzetti, F.; Ardizzone, F.; Cardinali, M.; Galli, M.; Reichenbach, P.; Rossi, M. Distribution of landslides in the Upper Tiber River basin, central Italy. *Geomorphology* **2008**, *96*, 105–122. [CrossRef]
21. Donati, L.; Turrini, M.C. An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: Application to an area of the Apennines (Valnerina; Perugia, Italy). *Eng. Geol.* **2002**, *63*, 289. [CrossRef]
22. Martínez-Álvarez, F.; Reyes, J.; Morales-Esteban, A.; Rubio-Escudero, C. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. *Knowl.-Based Syst.* **2013**, *50*, 198–210. [CrossRef]
23. Fang, Z.; Wang, Y.; Peng, L.; Hong, H. Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. *Comput. Geosci.* **2020**, 139. [CrossRef]
24. Jiménez-Perálvarez, J.D.; Irigaray, C.; El Hamdouni, R.; Chacón, J. Landslide-susceptibility mapping in a semi-arid mountain environment: An example from the southern slopes of Sierra Nevada (Granada, Spain). *Bull. Eng. Geol. Environ.* **2010**, *70*, 265–277. [CrossRef]
25. Yamagishi, H.; Yamazaki, F. Landslides by the 2018 Hokkaido Iburi-Tobu Earthquake on September 6. *Landslides* **2018**, *15*, 2521–2524. [CrossRef]
26. Shao, X.; Ma, S.; Xu, C.; Zhang, P.; Wen, B.; Tian, Y.; Zhou, Q.; Cui, Y. Planet Image-Based Inventorying and Machine Learning-Based Susceptibility Mapping for the Landslides Triggered by the 2018 Mw6.6 Tomakomai, Japan Earthquake. *Remote Sens.* **2019**, *11*, 978. [CrossRef]
27. Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital terrain modelin—A review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **1991**, *7*, 18. [CrossRef]
28. Zevenbergen, L.W.; Thorne, C.R. Quantitative Analysis of Land Surface Topography. *Earth Surf. Process. Landf.* **1987**, *12*, 56. [CrossRef]
29. Quinn, P.F.; Beven, K.J.; Lamb, R. The in(a/tan/β) index: How to calculate it and how to use it within the topmodel framework. *Hydrol. Process.* **2010**, *9*, 22. [CrossRef]
30. Falaschi, F.; Giacomelli, F.; Federici, P.R.; Puccinelli, A.; D'Amato Avanzi, G.; Pochini, A.; Ribolini, A. Logistic regression versus artificial neural networks: Landslide susceptibility evaluation in a sample area of the Serchio River valley, Italy. *Nat. Hazards* **2009**, *50*, 551–569. [CrossRef]
31. Seamless Digital Geological Map of Japan. Available online: https://gbank.gsj.jp/seamless/index_en.html?p=download (accessed on 1 February 2021).
32. Gong, P.; Liu, H.; Zhang, M.; Li, C.; Wang, J.; Huang, H.; Clinton, N.; Ji, L.; Li, W.; Bai, Y.; et al. Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. [CrossRef]
33. Wang Jinfeng, X.C. Geo-detector: Principle and prospective. *Acta Geogr. Sin.* **2017**, *72*, 116–134. [CrossRef]
34. Breiman, L.F.; Jerome, H.; Olshen, R.A.; Charles, J.C. *Classification and Regression Trees*; Wadsworth International Group: Wadsworth, OH, USA, 1984.
35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2830.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Leo, B. Arcing Classifiers. *Ann. Stat.* **1999**, *26*, 801–849.
38. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
39. Walter, S.D. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat. Med.* **2002**, *21*, 1237–1256. [CrossRef] [PubMed]
40. Pradhan, B.; Lee, S. Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. *Landslides* **2009**, *7*, 13–30. [CrossRef]