GIS-Based Landslide Susceptibility Mapping using Logistic Regression, Instability Index, and Support Vector Machine: Case Study of the Jingshan River, Taiwan

Hsun-Chuan Chan
Department of Soil and Water Conservation, National Chung Hsing University,Taichung, Taiwan (ROC)

Yu-Chin Chen
Department of Soil and Water Conservation, National Chung Hsing University, Taichung, Taiwan (ROC)

Jung-Tai Lee
Department of Forestry and Natural Resources, National Chiayi University, Chiayi City, Taiwan (R.O.C.),
jtlee@mail.ncyu.edu.tw

Yu-Ting Wen
Department of Soil and Water Conservation, National Chung Hsing University, Taichung, Taiwan (ROC)

RESEARCH ARTICLE

GIS-Based Landslide Susceptibility Mapping Using Logistic Regression, Instability Index, and Support Vector Machine: Case Study of the Jingshan River, Taiwan

Hsun-Chuan Chan [a], Yu-Chin Chen [a], Jung-Tai Lee [b],*, Yu-Ting Wen [a]

[a] Department of Soil and Water Conservation, National Chung Hsing University, 250 Kuo Kuang Rd., Taichung 402, Taiwan, ROC
[b] Department of Forestry and Natural Resources, National Chiayi University, No. 300 Syuefu Rd., Chiayi City 60004, Taiwan, ROC

Abstract

Many techniques have been developed to produce landslide susceptibility maps. However, mapping the spatial distribution of landslide susceptibility involves critical steps. This study evaluated susceptibility to the occurrence of shallow landslides in upstream areas of the Jingshan River, Taiwan, where heavy rainfall has resulted in increased landslide occurrence and reservoir sedimentation. The landslide susceptibility case study was conducted using logistic regression, an instability index method, and support vector machine (SVM). A selection procedure was first developed to identify the factors influencing landslide occurrence. Historical landslide data were used to assess the corresponding parameters of each model. The receiver operating characteristic curve was then used to evaluate the accuracy of model results. The results indicated that the instability index method underestimated landslide susceptibilities in areas near the river. The instability index method was highly affected by the classifications of model factors. Of the models, the SVR model was the most accurate regarding landslide susceptibility in the study area. Employing this framework, governments may implement suitable, cost-effective improvements to structural measures and land-use planning.

Keywords: Logistic regression, Instability index, Support vector machine, landslide susceptibility

1. Introduction

Taiwan is situated on a plate boundary, exhibits a diverse and complex terrain and a fragile geological structure. It also features complex and changeable weather involving concentrated and intense rainfall. Therefore, landslides frequently occur in its high-sensitivity mountain zones. Consequently, the slope disaster occurrence rate has increased annually in areas with steep slopes. To safeguard the lives and property of residents from typhoons and heavy rain, disaster prevention measures should be combined with landslide susceptibility analysis to enhance their comprehensiveness and subsequent slope treatment work.

Landslide susceptibility analysis typically involves identifying a set of topographic, geologic, and hydrologic factors that explain landslide occurrences, constructing a landslide susceptibility model, calculating susceptibility values, and illustrating a landslide susceptibility map. Generally, the methods employed to investigate landslide susceptibility in previous studies can be divided into qualitative and quantitative [7]. Qualitative methods do not involve objective evaluation, thus hindering comparisons of analyses results by different

scholars on the same region. The recent maturation of geographic information system and remote sensing technology has indirectly facilitated the development of landslide susceptibility analysis and mapping methods. Researchers have used different quantitative methods for landslide susceptibility classification, for example, discriminant analysis, fixed-value analysis, logistic regression, instability index, artificial neural network, and support vector machine (SVM).

Of the aforementioned quantitative methods, Discriminant analysis can only employ continuous factors (e.g., slopes and elevations) as the independent variables and cannot incorporate classified factors (e.g., aspects, lithological characters). Fixed-value analysis is based on the theory of mechanics and features rigorous analytical processes. However, the predictive capability of this method on large areas is limited because of the difficulty in acquiring geotechnical parameters. The logistic regression and instability index methods statistically analyze the effects of various factors on landslides, and a set of equations are formulated accordingly. Logistic regression involves establishing a regression model through the use of a set of factors that discriminate landslide and nonlandslide data and thereby calculate landslide susceptibility. The instability index method involves exploring the frequency of landslide occurrence in a region according to the classes of various factors, calculating the landslide ratio of each region, identifying the weight of each factor, and estimating the landslide susceptibility in each analyzed region through linear superposition. The results of these two methods are simple to

incorporate and compare. An artificial neural network generates results with more ideal fit than do general linear equations. However, the trial-and-error process in learning in such a network is time-consuming, and the calculation process of the network involves a black-box analysis, rendering analysis of the relationships between various factors and landslide occurrence difficult. Moreover, over-fitting may occur in the network. SVM is another type of learning machine that follows the principle of structural risk minimization, thus, preventing over-fitting during its calculation process.

In Taiwan, statistical analyses of landslide susceptibility have yielded considerable results. SVM features a high fitting capability during its calculation process; however, it has limited applications to landslide susceptibility analysis. In this study, two statistical analyses (i.e., logistic regression and instability index methods) and the SVR method were used to produce a landslide susceptibility map of the Jingshan River, where shallow landslides are particularly recurrent. The performance of statistical analyses was compared with the performance of the SVR method. The accuracy of the maps was evaluated, and the differences between the maps were examined. The aim was to determine which of the three landslide susceptibility models was the most suited to the study area.

2. Study Area

The upstream areas of the Jingshan River, a tributary of Da'an River, were selected as the study area. The downstream area of the Jingshan River features
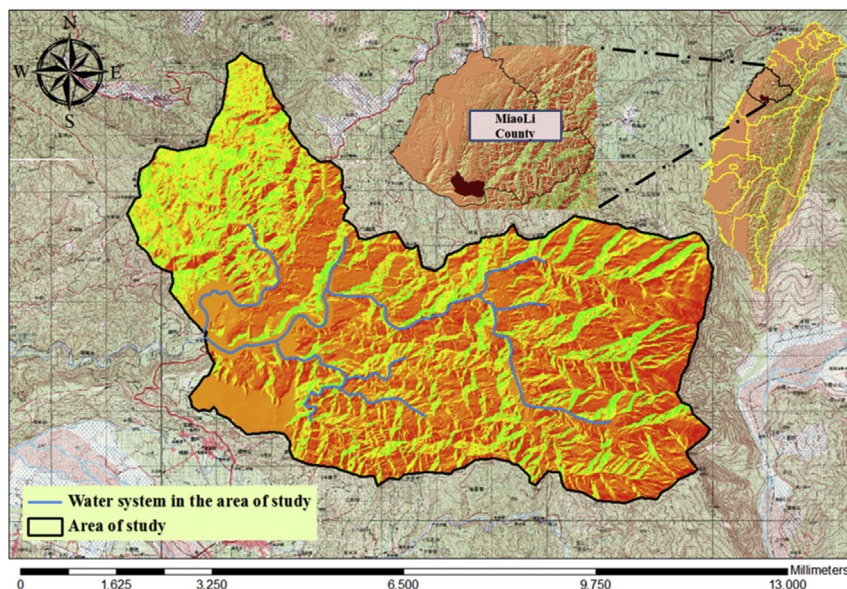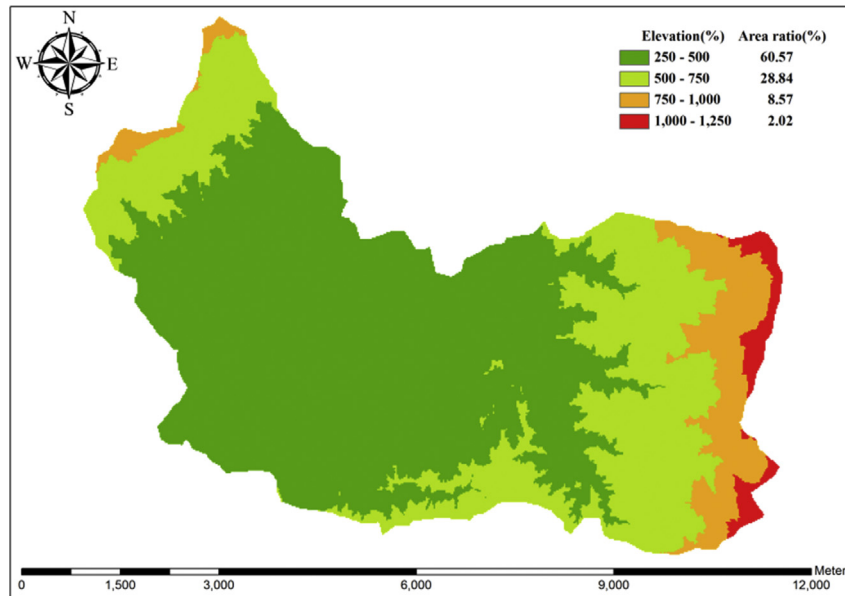

Fig. 1. Area of study.

Fig. 2. Spatial distribution of elevations in the area of study.

the Liyutan Dam, a reservoir that provides tourism, irrigation, flood prevention, and electricity; the reservoir is the primary water source for central Taiwan. Recently, typhoon-induced rainfall has caused shallow landslides in upstream areas of the Jingshan River. Sediments have flowed into the reservoir and shortened its lifespan, threatening the stability of water quality in the region.

The Jingshan River is located in the Miaoli county, central Taiwan. It features a catchment area of approximately 4,662 hectares (Fig. 1). This region is

located in a subtropical climate zone. Northeastern seasonal wind is prevalent from October of each year to April of the following year, and rainfall is limited during this period because of obstruction by the Central Mountain Range. Conversely, southwestern seasonal wind, plum rain, and typhoons are prevalent from April to September of each year, leading to abundant rainfall. Fig. 2 illustrates the distribution of the elevations of the catchment area, which range from 294 m to 1,224 m; 60.57% of the area features elevations of 250 to 500 m. The slopes
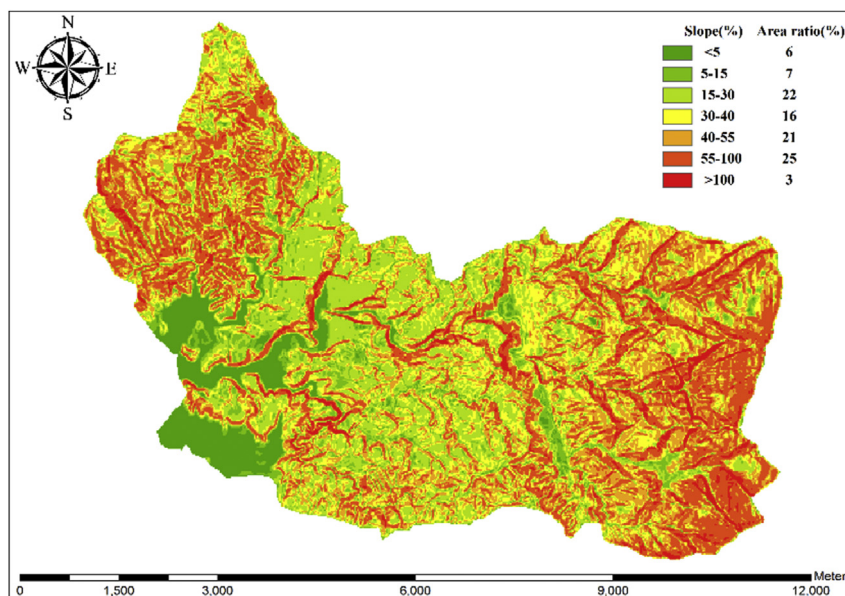

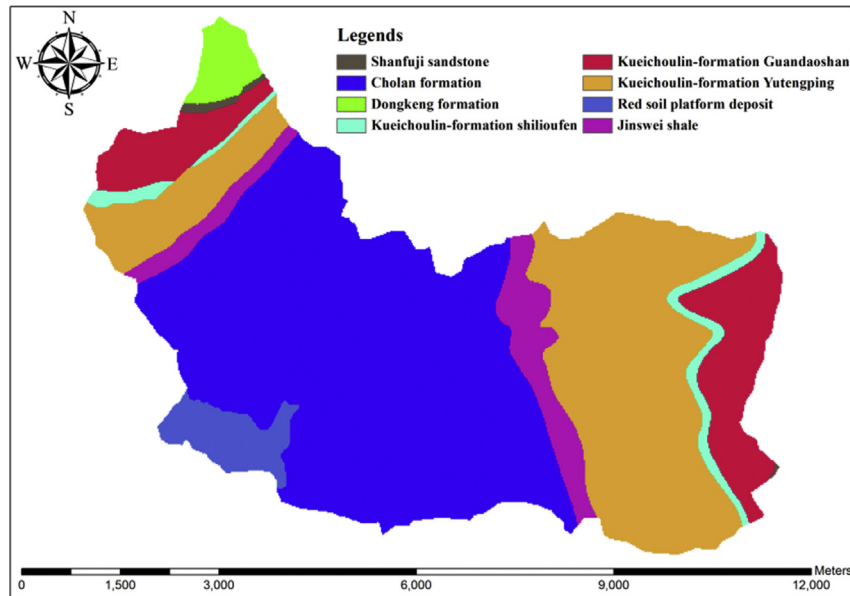*Fig. 3. Spatial distribution of slopes in the area of study.*

Fig. 4. Spatial distribution of lithological character in the area of study.

of nearly 84% of the study area are concentrated at the ranges of 15%~100%, revealing considerably steep terrain in the area (Fig. 3). The area is part of the Western Lushan terrain zone and features a northeastern−southwestern terrain distribution. Sandstone, mudstone, and shale are the primary lithological characters of the catchment area. The Cholan formation is the most widely distributed

lithological formation in the area (47.24%), followed by the Kueichoulin-formation Yutengping (28.47%); other formations featured in this area include the Tungkeng formation and Jinswei shale (Fig. 4). Table 1 lists the lithological composition of each formation.

3. Methodology

Suitable analysis units can enhance evaluation precision in analyzing landslide susceptibility. Numerous types of analysis units are currently available for landslide susceptibility evaluation. The most commonly applied units in landslide susceptibility analysis are slopes and grids [5,17]. The slope unit implies a topological continuity in a slope. However, the size of a slope unit is highly subjective and may require expert assistance. Grid units are simple in structure, quick to generate, and

Table 1. Lithological composition of the area of study.

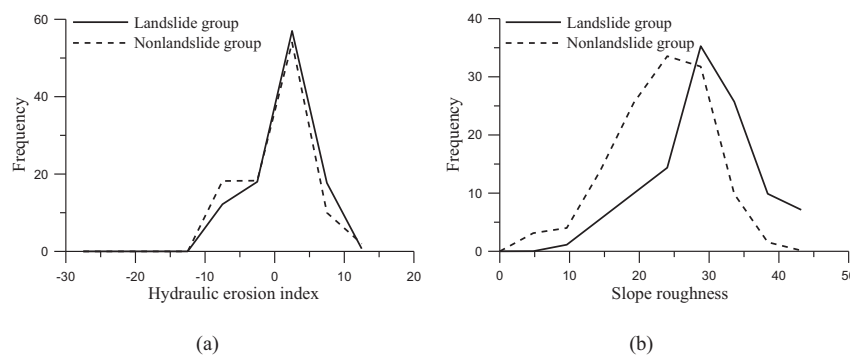| Formation	Composition
Cholan formation	Sandstone, mudstone, shale
Guandaoshan, Yutengping, Shangfuji sandstones	Sandstone, siltstone, shale
Dongkeng formation, Jinswei shale	Sandstone, shale
Red soil platform deposit	Red soil, gravel, sand, clay
Shilioufen	Shale

Fig. 5. Frequency distributions of landslide and nonlandslide groups for (a) hydraulic erosion index and (b) slope roughness.

Table 2. Results of factor selection analysis.

Factors	Selection Methods			Selected Factor
	Probability−Probability Plot	Frequency Distribution of Landslide and Nonlandslide Groups	Landslide Ratio Distribution	
hydraulic erosion index	○			
topographic humidity index	○		○	
slope roughness	○	○	○	✓
terrain roughness	○	○	○	✓
slope	○	○	○	✓
elevation	○	○		
plane curvature	○			
section curvature	○			
total curvature	○			
annual average rainfall	○	○	○	✓

consistent in size. Therefore, grid units were selected as the analysis units for comparing the landslide susceptibility maps created through the aforementioned three different models (viz., logistic regression, instability index, and SVM). These three methods were then used to calculate the landslide susceptibility values. Subsequently, the performance of the models was evaluated. The following subsection introduces the analysis units, data collection, factor selection, and landslide susceptibility analysis methods.

3.1. Analysis Units

The analysis units adopted in this study were grid units. Factors were established through the use of 20 m × 20 m grids for the follow-up landslide susceptibility analysis. A total of 116,564 grid units were applied in the study area.

3.2. Data Collection

Basic data must be collected and analyzed to examine the effect of environmental and external factors on landslide occurrence in the upstream catchment area of the Jingshan River. These data are listed as follows.

(1) Digital Elevation Model

The 5 m × 5 m digital elevation model (DEM) constructed by the Ministry of the Interior for the High-Precision and High-Resolution DEM Construction Program [11] provided the basic data. Considering computation capacity and the consistency of image resolutions, the DEMs were interpolated to 20 m × 20 m grid cells.

(2) Geological Map

This study adopted a 1:50,000 geological map of the study area, sourced from the Central Geological Survey, Taiwan.

(3) Annual Rainfall Data

The 2004−2013 annual average rainfall data from seven rainfall stations near the area of study, namely Dahu (1), Xinkai, Cholan (2), Sanyi (2), Pinglin Elementary School, Shuangchi (2), and Xiangbi, were collected for use in this study.

(4) Landslide Records

The records of six major landslide events in the Central Geological Survey, namely Typhoon Toraji (2001), Typhoon Mindulle (2004), Typhoon Kalmaegi (2008), Typhoon Sinlaku (2008), Typhoon Morakot (2009), and heavy rain on July 19, 2011, were selected for analysis.

3.3. Factor Selection

Nearly 60 factors have been employed in studies on landslide susceptibility [9]. These factors can be divided into three types: topographic, geologic, and hydrologic. Topologic factors, such as elevation, slope, aspect, slope roughness, and terrain roughness, have been frequently employed in studies on shallow landslides [19,20]. Geological factors, including lithological characters, reflect the geological structure of a specific area; shallow landslides occur easily in an environment with a fragile geological structure. According to previous studies, 10 commonly used landslide susceptibility factors, namely slopes, the hydraulic erosion index, the

topographic humidity index, total curvature, plane curvature, section curvature, terrain roughness, slope roughness, slope, elevation, and annual average rainfall, were preliminarily selected. The number of factors was then narrowed to enhance the accuracy of the landslide susceptibility models through the use of the probability−probability plot, frequency distribution of landslide and non-landslide groups, and landslide ratio distribution [10]. The methods for selecting related factors in the present study are detailed as follows:

(1) Probability−Probability Plot

Probability−probability plots of factors are used to examine whether specific factors are normally distributed and to display the relationship between the cumulative ratio of a variable and that of a standard normal distribution, which is represented by a diagonal line in the plot. When the data distribution of a factor is nearly normal, the data points approximate the diagonal line. All the preliminarily selected susceptibility factors were normally distributed and applicable for analyzing the frequency distributions of landslide and nonlandslide groups.

(2) Frequency Distribution of Landslide and Non-landslide Groups

The continuous values of the factors were segmented into multiple equidistant intervals, and two lines on the frequency of occurrence in the intervals, one for the landslide group and the other for the nonlandslide group, were illustrated. The larger the difference between the landslide and non-landslide groups, the more satisfactorily the factor discriminated the two groups. The results of the hydraulic erosion index, topographic humidity index, and terrain curvature overlapped in the frequency distribution chart, revealing that these factors could not clearly discriminate between the landslide and nonlandslide data.

(3) Landslide Ratio Distribution

The continuous values of the factors were segmented into multiple equidistant intervals, and the landslide ratio of each interval was separately calculated. Curves were plotted to examine whether the landslide ratios exhibited a specific trend. The landslide ratio was calculated as follows:

$$\text{Landslide ratio} = \frac{\textit{The landslide grids} \text{ of the interval}}{\textit{The total landslide grids}}$$

(12)

The frequency distribution of the landslide and nonlandslide groups as well as the landslide ratio of these factors were plotted, and the results were visually examined. For example, Fig. 5 presents the frequency distributions of landslide and non-landslide groups for the hydraulic erosion index and slope roughness. Slope roughness allowed successful division of the data into landslide and non-landslide data. However, the hydraulic erosion index could not clearly distinguish between the landslide and nonlandslide groups. The results are provided in Table 2. The adequate factors for the landslide susceptibility models were slope, terrain roughness, slope roughness, and annual average rainfall. However, aspect and lithological characters featured noncontinuous values, and thus the relationship between their numerical values and landslide occurrence could not be identified through the aforementioned selection process. Pearce and O'Loughlin [14] addressed the relationship between landslide characteristics and slope aspect. Chen et al. [4] maintained that geological structures typically determine the likelihood of landslide occurrence; for example, landslides are less likely to occur in rock formations with sturdier structures and a higher degree of consolidation. Because aspect and lithological characters contribute considerably to landslide occurrence, they were included in the models.

3.4. Logistic Regression

Logistic regression is a particular type of logarithmic linear model frequently applied in landslide susceptibility analyses [1,6]. This model identifies a set of regression equations capable of discriminating landslide and nonlandslide data through statistical regression. The values of the logistic distribution function range from 0 to 1 and are distributed in an S-shaped curve, similar to the cumulative distribution curve of a random variable. Equation (1) shows the aforementioned analysis model [13].

$$P(a_i = 1 | b_i) = \frac{1}{1 + e^{-(\alpha + \beta b_i)}}$$

(1)

where $P(a_i = 1 | b_i)$ represents the conditional probability of landslide occurrence; b_i is the column vector of the independent variable; and α and β are respectively the row vectors of the intercept and regression coefficient. When the number of independent variables is k, the logistic regression model is expressed as follows:

$$\ln\left(P_i/1 - P_i\right) = \alpha + \sum_{1-i}^{k} \beta_k b_{ki} \qquad (2)$$

where P_i represents the probability of event occurrence when a series of independent variables b_{1i}, $b_{2i,...}, b_{ki}$ are given.

3.5. Instability Index

The instability index model involves estimating the coefficient of variation in each factor, sequencing and determining the weight ratio of each factor according to its coefficient of variance, and thereby formulating a multivariate nonlinear mathematical model [8]. The instability index value is used to describe the level of instability of a slope, therefore not only enabling an explanation of the level of landslides caused by each factor but also integrally calculating the effect of all factors on landslide occurrence. The principles of this model are as follows:

(1) Calculate the density of landslide grids in each factor as a benchmark for evaluating the factor. The equation is expressed as follows:

$$G_i = S_i \Big/ \sum S_i \qquad (3)$$

where G_i represents the probability of landslide occurrence as a percentage; S_i represents the number of landslide grids in the specific class of the factor; $\sum S_i$ represents the total number of grids of the factor in each class.

(2) After the landslide percentage of each causative factor is identified, calculate the scoring value of each single factor in the specific class. Match the landslide percentage of each class to a value between 1 and 10 to obtain the scoring value of that class. The equation is expressed as follows:

$$D = 9(G_i - G_{min})/(G_{max} - G_{min}) + 1 \qquad (4)$$

where D represents the instability index value of each factor; and G_{max} and G_{min} represent the maximal and minimal landslide occurrence probability percentages of each factor, respectively.

(3) After the aforementioned scoring value is calculated, calculate the coefficient of variation of the landslide ratio in the class of the factor to obtain the weight comparison value of the factor. Use the coefficient to identify the sensitivity of the factor to the probability of landslide occurrence in the specific class. The equation is expressed as follows:

$$V = \frac{\sigma}{X} \times 100 \qquad (5)$$

where V represents the coefficient of variation, σ represents the standard deviation, and X represents the mean.

$$W_i = \frac{V_i}{V_1 + V_2 + \ldots + V_i} \qquad (6)$$

where W_i represents the weight of a single factor and V_i represents the coefficient of variation of the factor.

(4) Conduct a product superposition on the weighted and scoring values to calculate the landslide susceptibility value and thereby evaluate the level of stability in the grid. The equation is expressed as follows:

$$D_{total} = D_1^{W_1} \times D_2^{W_2} \times \ldots \ldots \times D_i^{W_i} \qquad (7)$$

where $D_1^{W_1}$, $D_2^{W_2}$, ... $D_i^{W_i}$ represent the susceptibility value in each factor; and D_{total} represents the total landslide susceptibility value, which is positively associated with the probability of landslide occurrence.

3.6. Support Vector Machine

The SVM method has been applied in numerous fields and seen substantial development [18]. According to its application, this method is further divided into two types: support vector classification and support vector regression (SVR). In this study, the SVR model was employed to establish a landslide susceptibility map. The process involved the use of a set of training data $[(x_1, y_1), (x_2, y_2), \ldots (x_i, y_i)]$, where x_i represents the input vector and y_i represents the output value. Thus, the SVR function was calculated as shown in Equation (8). The input vector mapped the nonlinear problem on a high-dimensional characteristic space through the nonlinear mapping function Φ, rendering the problem linear and establishing the optimal regression function fitting characteristic. The regression model featured the error tolerance zone ϵ.

$$f(x) = \text{sign}\left[\omega^T \cdot \Phi(x_i) + b\right] \qquad (8)$$

where $f(x)$ is a decision function used to categorize the test data input in the model; b represents the offset, without which the model calculation process must involve passing through the original point, limiting the assessment capability of the SVR model; and ω^T represents the level of model complexity,

which is positively associated with the over-fitting probability of the model. Therefore, the principle of structural risk minimization must be considered to prevent over-fitting when establishing the model. Equation (9) illustrates the model established according to this rule, where ξ_i and ξ_i^* respectively represent the upper and lower error values of a data point in the tolerance zone; i represents the ith data; and C is a penalty parameter, which is used to measure the penalty weight of an error value. The primary function form in an SVR model is a loss function ϵ as expressed in Equation (10).

$$\Omega^T \cdot \Phi(x_i) = \frac{1}{2}\omega^T \cdot \omega + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \quad (9)$$

$$y_i - (\omega \cdot x_i + b) \leq \epsilon + \xi_i \ (\omega \cdot x_i + b) - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, l$$

$$|y - f(x)|_\epsilon = \begin{cases} 0, if\, |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon, otherwise \end{cases} \quad (10)$$

Because the aforementioned calculation processes were overly complicated, the Lagrangian parameters α_i and α_i^* were converted to Lagrangian problems, which were then converted to quadratic problems to determine the solution. Equation (11) depicts the converted SVR model.

$$f(x) = \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i) \Phi(x_i)^T \cdot \Phi(x_j) + b \quad (11)$$

where each $(\alpha_i^* - \alpha_i)$ corresponds to a training data(x_i, y_i). If $(\alpha_i^* - \alpha_i) \neq 0$, then the data is involved

Table 3. Coefficient values of factors in the logistic regression model.

Parameter	Factor	Coefficient	Coefficient Values
L_1	Sandstone, mudstone, shale	β_1	−0.416
L_2	Sandstone, siltstone, shale	β_2	−0.097
L_3	Sandstone, shale	β_3	−
L_4	Red clay, gravel, sand, clay	β_4	−19.675
L_5	Shale	β_5	0.136
D_1	North	β_6	−1.03
D_2	Northeast	β_7	−0.521
D_3	East	β_8	−0.343
D_4	Southeast	β_9	0.121
D_5	South	β_{10}	0.013
D_6	Southwest	β_{11}	0.373
D_7	West	β_{12}	−
D_8	Northwest	β_{13}	0.367
F_1	Terrain roughness	β_{14}	4.61
F_2	Slope roughness	β_{15}	4.042
F_3	Slope	β_{16}	0.294
F_4	Annual average rainfall	β_{17}	0.127
C	Constant	C	−3.994

in the establishment of the regression model and is referred to as a support vector [3]. The mapping function is difficult to solve in the model, but the solution process can be simplified through the conversion of the function to the inner product $\Phi(x_i)^T \cdot \Phi(x_j)$, which is referred to as a kernel function in the SVM. A specific function that satisfies Mercer's condition [18] can be applied as a kernel function. There are numerous types of commonly applied kernel functions. In this study, a radial basis function, which is incorporated to process nonlinear and high-dimensional data and exhibits the parameter γ, was applied as the kernel function.

3.7. Model Assessment

The predictive performance of the model was assessed using its receiver operating characteristic (ROC) curves. Area under the ROC curves (AUC) is a quantitative mesure used to assess the classification performance of response variables [12]. This study investigated the relationships of landslide susceptibility values with the landslide and nonlandslide data. Landslide and nonlandslide data are response variables. Therefore, AUC was selected to assess the model performance for binary classification.

To illustrate an ROC curve, a continuous value was segmented into multiple equidistant intervals, each of which was then cumulatively calculated from large to small. The horizontal axis represented the landslide susceptibility value, and the vertical axis represented the predicted success rate [2]. Subsequently, a curved line was plotted to determine the AUC, thereby quantifying the result reflected by the ROC curve. AUC ranges between 0 and 1; a large AUC value indicates that the predicted result of the model is satisfactory. Swets [16] maintained that a model exhibits almost no landslide occurrence discrimination capability when AUC = 0.5; the landslide occurrence discrimination capability of the model is acceptable when AUC = 0.5−0.7, excellent when AUC = 0.7−0.9, and outstanding when AUC > 0.9.

4. Results and Discussion

Three quantitative analysis models were employed for landslide susceptibility mapping by adopting the same potential factors, namely slope, terrain roughness, slope roughness, annual average rainfall, aspect, and lithological characters. Because the factors differed in their units and value distributions, they were normalized before the maps were constructed to prevent the generation of

Table 4. Factor classification result of each group in the instability index model.

Group	Slope	Aspect	Lithological characters	Annual average rainfall	Slope roughness	Terrain roughness	AUC
1	0.1202(7)	0.1440(8)	0.3294(5)	0.1268(8)	0.0673(9)	0.2122(10)	0.713
2	0.1119(7)	0.1258(8)	0.3300(5)	0.1346(5)	0.0714(6)	0.2252(7)	0.718
3	0.1210(7)	0.0978(8)	0.3200(5)	0.1434(4)	0.0761(4)	0.2400(5)	0.697
4	0.1256(7)	0.1238(8)	0.3260(5)	0.1325(6)	0.0703(6)	0.2217(7)	0.711
5	0.1116(7)	0.1256(8)	0.3305(5)	0.1344(4)	0.0713(6)	0.2248(7)	0.709
6	0.1090(7)	0.1488(8)	0.3224(5)	0.1311(5)	0.0695(9)	0.2193(7)	0.693
7	0.1159(7)	0.0950(8)	0.3427(5)	0.1393(5)	0.0739(4)	0.2331(7)	0.703
8	0.1099(7)	0.1235(8)	0.3432(5)	0.1322(5)	0.0701(6)	0.2211(10)	0.697
9	0.1153(7)	0.1296(8)	0.3110(5)	0.1386(5)	0.0736(6)	0.2320(5)	0.699

( ): numbers of classes.

abnormally large or small data, which would affect the mapping analysis results. The range of values of the normalized factors was shrunk to 0−1 to reduce the effects of the units and range values of the factors on the susceptibility analysis. The number of landslide grids in the study area was substantially smaller than that of the nonlandslide grids. Therefore, in establishing the logistic regression and SVM models, the model analysis errors caused by the difference between the amounts of landslide and nonlandslide data as well as the subjective error caused by human selection must be avoided. Through random selection, approximately 1,000 grids each of landslide and nonlandslide grid data were selected. The subsequent landslide susceptibility result of each model is presented as follows.

4.1. Logistic Regression

This model involved relating the probability of landslide occurrence to a series of independent variables. SPSS Statistics 20.0 was employed to calculate the logistic regression coefficients of the factors. In this analysis, the "continuous" data such as terrain roughness, slope roughness, slope, and annual average rainfall were treated as "scales" in SPSS, whereas aspect and lithological characters were considered as "nominal'' data. Aspects were divided into eight classes based on the eight points of a compass. Lithological characters were divided into five classes according to their formation distribution in the area of study. The coefficient values of the landslide susceptibility factors are summarized in Table 3. Then, Equation (2) with the calculated coefficients of the factors could be used to predict the landslide susceptibility value (P) of each grid. Although the developed logistic regression model does not indicate how landslides physically develop, it indicates that landslides are associated with specific factors. A positive regression coefficient implies that the corresponding factor increases the landslide susceptibility of the area, and a negative coefficient

value indicates a negative relationship with landslide occurrence. The regression coefficients revealed that among the investigated factors, terrain roughness and slope roughness had relatively high values (Table 3). This indicated that local changes in the terrain and slope profoundly and positively influenced the landslide susceptibility of the study area.

4.2. Instability Index

This model involved dividing each factor into several classes and calculating the density of landslide grids in each factor class, thereby obtaining its weight. However, an excessive number of classes in the instability index may lead to excessively high factor variance. In the present study, nine groups of factor classification were used to examine the performance of the instability index model. Aspect and lithological characters were category-based factors classified into eight and five classes, respectively. Slopes were divided into seven classes according to the Technical Regulations for Soil and Water Conservation [15]. Terrain roughness, slope roughness, and annual average rainfall were considered adjustable factors; Table 4 lists the classification and weighting of these three factors. The AUC analysis result of each group is also listed in Table 4.

The weighting results revealed the extents of which the factors affected the landslide occurrences in the study area. As indicated in Table 4, terrain roughness and lithological characters influenced the landslide susceptibility in the area considerably and constituted more than 50% of the total weighted value, whereas the effect of aspect on the susceptibility was smaller. Regarding the adjustable factors, because of its classification, slope roughness exhibited a larger variance in the weighted values than did annual average rainfall and terrain roughness. For clarifying the effect of adjusting the factor classes on the landslide susceptibility, two of the adjustable factors in each group were fixed, and
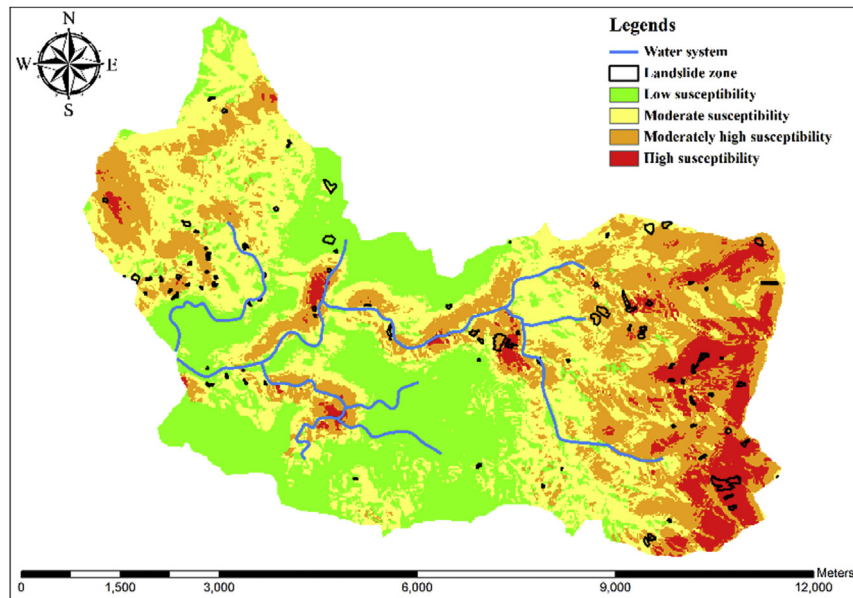
Fig. 6. The landslide susceptibility map predicted by Logistic regression.

the remaining adjustable factor was adjusted. For example, the classes of the slope and terrain roughness in one group were fixed, and that of the annual average rainfall was increased or reduced, to re-evaluate the landslide susceptibility. Table 4 lists the factors with the highest weights. Terrain roughness, which was easily affected by its level, exhibited considerable variability in its weight in different classes. Comparing the AUC values showed that when the factors of a model were excessively or insufficiently classified, the accuracy

of the model decreased. The AUC of the second group was the highest of all nine groups. Therefore, the factor classification of the second group was adopted for the follow-up analysis, the results of which were compared to those of the logistic regression and SVM analyses.

4.3. SVM

In the SVR analysis on landslide susceptibility, the SVR estimated values were designated as landslide
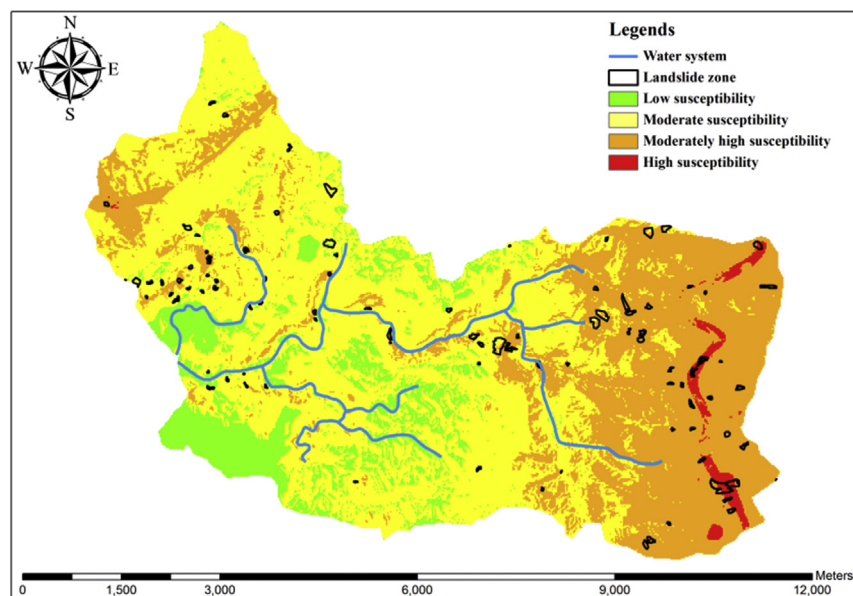

Fig. 7. The landslide susceptibility map predicted by Instability index method.
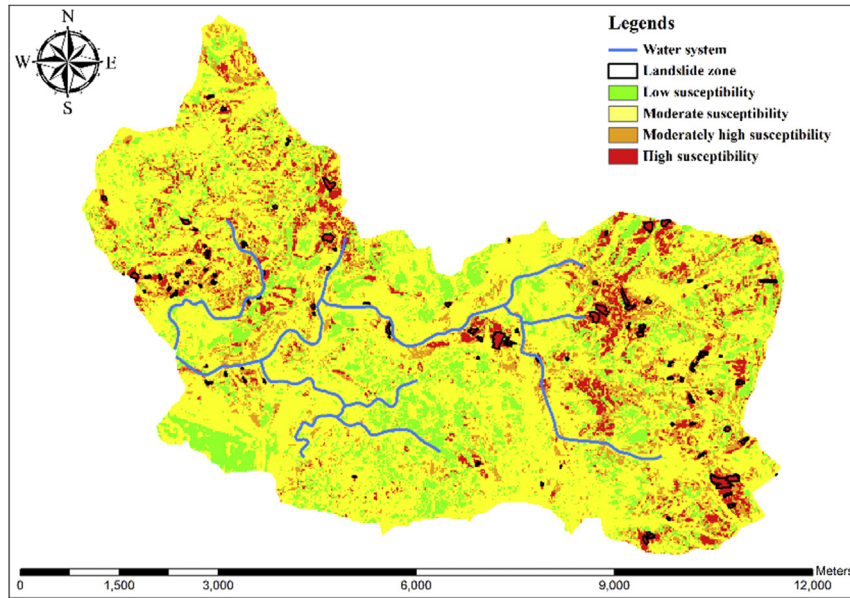
Fig. 8. Landslide susceptibility map as predicted by the SVM.

occurrence probability. A total of three parameters, namely penalty C, loss function ϵ, and kernel function parameter γ, must be determined and carefully prepared to establish an efficient SVR model. The two-stage grid search method was employed to identify the margin of error, penalty, and kernel function parameters, which were the pending parameters of the SVR model. A local search was performed to locate the ranges of the parameters, and a detailed search was then conducted to identify the optimal values of the parameters. The parameters were determined through LIBSVM [3]. Root-mean-square errors (RMSEs) were used as target functions; a smaller RMSE reveals that the parameter improves the accuracy of the SVR model. The optimal parameter results were identified as $(C,\epsilon,\gamma) = (55, 0.015, 210)$, and RMSE = 0.0007. This parameter set was applied to input in the SVR model and calculate the landslide occurrence probability of each grid.

4.4. Model Performance

Subsequently, the landslide susceptibility maps based on logistic regression, instability index, and SVM were illustrated for analysis and comparison. For the clarity of presentation, landslide susceptibilities were divided into four levels: <0.25, 0.25−0.5, 0.5−0.75, and >0.75, representing low, medium, medium-high, and high probability of landslide occurrence, respectively. A susceptibility of 0.5 was used as the cut-off value. Grids with a susceptibility above 0.5 were classified as a landslide grid (i.e., medium-high or high susceptibility), whereas those with lower probabilities were classified as non-landslide grids (i.e., low or medium susceptibility).

Figs. 6−8 reproduce the landslide susceptibility maps created using logistic regression, instability index, and SVM, respectively. The logistic regression map exhibited the most favorable prediction results for shallow landslides in areas near the river.

Table 5. Number of observed landslides that fall into the various susceptibilities in different models.

	Logistic regression		Instability index		SVM	
	Landslide occurrence grids	Landslide occurrence grids, %	Landslide occurrence grids	Landslide occurrence grids, %	Landslide occurrence grids	Landslide occurrence grids, %
Low susceptibility	94	9.39%	21	2.08%	2	0.002%
Medium susceptibility	235	23.48%	355	35.46%	11	1.10%
Medium-high susceptibility	393	39.26%	491	49.05%	15	14.99%
High susceptibility	279	27.87%	134	13.41%	973	83.91%
AUC	0.721		0.718		0.825	

This could be attributed to the topographic roughness and complexity of the riverine areas. The terrain roughness and slope roughness were prone to the landslide analysis using logistic regression. The instability index map underestimated susceptibility most frequently. Few of the observed landslide grids, located in the eastern part of the study area, were situated in predicted medium-high or high susceptibility zones. This was attributed to the fact that the landslide records used in this study involved extreme events (e.g., Typhoon Morakot). The data of extreme events increased the variation of the factors. The instability index was calculated using a linear combination of the factors and their weights. The weight of a factor was estimated using the variation of the factor. Therefore, the results of the instability index were highly affected by factors with high variation. This caused most of the observed shallow landslides to be located in the predicted low susceptibility zones. The SVM map could effectively differentiate the level of landslide susceptibility and prevent same landslide susceptibility in a certain area.

Table 5 presents the number of observed landslides that fall into the various susceptibilities in different models. The AUC of each model attained an accuracy of more than 70%, revealing that all the three models effectively predicted the landslide occurrences in the area of study. However, the AUC of the SVR model was 0.825, substantially higher than those of the other two models. In the SVR map, approximately 15% and 84% of the observed landslide grids had medium-high and high susceptibility levels, respectively. In the logistic regression and instability index maps, the total percentages of observed landslide grids in medium-high and high susceptibility zones were lower (approximately 37% and 62%, respectively). This revealed that the SVR model to discriminate landslide and nonlandslide occurrences was considerably more satisfactorily than were the other two models. The SVR map is, hence, somewhat suitable to government or developers when formulating land utilization regulations and investing cost-effective engineering works.

5. Conclusion and Suggestions

In this study, the landslide susceptibility of upstream areas of the Jingshan River was evaluated. The factors influencing shallow landslide occurrence were lithological characters, aspect, slope, terrain roughness, slope roughness, and annual average rainfall. Subsequently, landslide susceptibility maps developed using logistic regression, instability index, and SVM were created and were compared and analyzed. The significance of the coefficients indicated that terrain roughness and slope roughness were the largest factors in landslide occurrence in the logistic regression model. In the instability index model, lithological characters and terrain roughness were the highest-weighted factors. The analysis thus indicated that geomorphological and geological factors influenced landslide occurrence in the study area more substantially than did hydrologic factors. The assessment of model performance indicated all three models had AUCs higher than 0.7. The presented methods statistically predicted landslide occurrence in the study area. In particular, the AUC of the SVM model was higher than 0.8, which was considerably higher than that of the other two models. The observed shallow landslides coincided with high percentages of high susceptibility areas in the SVM map. The results indicate that SVM was the most effective of the three models in assessing the spatial distribution of shallow landslides in the study area. The models employed in the study can complement each other. Complex data sets and a long calculation time are required for the logistic regression and SVM models, whereas the instability index model is simple to use and capable of rapidly analyzing data. Future studies may involve combining different models using various approaches to improve model accuracy and efficiency.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Agresti A. Categorical Data Analysis. 2nd ed. New York: John Wiley; 2002.

[2] Beguería S. Validation and Evaluation of Predictive Models in Hazard Assessment and Risk Management. Nat Hazards 2006;37:315—29.

[3] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] Chen H, Lin GW, Lu MH, Shih TY, Horng MJ, Wu SJ, Chuang B. Effects of topography, lithology, rainfall and earthquake on landslide and sediment discharge in mountain catchments of southeastern Taiwan. Geomorphology 2011;133:132—42.

[5] Erener A, Düzgün HSB. Landslide susceptibility assessment: what are the effects of mapping unit and mapping method? Environ Earth Sci 2012;66:859—77.

[6] Feinberg S. The analysis of cross-classified categorical data. 2nd ed. Cambridge, MA: MIT Press; 2007.

[7] 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;31:181—216.

[8] Jian LB. Application of geographic information system in the quantitative assessment of hillslope stability. Taiwan, ROC: National Chung Hsing University; 1992 (Master thesis, in Chinese).

[9] Koukis G, Ziourkas C. Slope Instability Phenomena in Greece: A Statistical Analysis. Bull Int Assoc Eng Geol 1991; 43:47−60.

[10] Lin HM, Chang SK, Wu JH, Juang CH. Neural network-based model for assessing failure potential of highway slopes in the Alishan, Taiwan Area: Pre- and post-earthquake investigation. Eng Geol 2009;104:280−9.

[11] Ministry of the Interior (MOI). Satellite Survey Center. Department of Land Administration, M.O.I; 2004.

[12] Mossman D. Assessing predictions of violence: Being accurate about accuracy. J Consult Clin Psychol 1994;62: 783−92.

[13] Ohlmacher GC, Davis JC. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas. Eng Geol 2003;69:331−43.

[14] Pearce AJ, O'loughlin CL. Landsliding during a M7.7 earthquake: influence of geology and topography. Geology 1985;13:855−8.

[15] Soil Water Conservation Bureau (SWCB). Technical Regulations for Soil and Water Conservation. 2010 (in Chinese).

[16] Swets JA. Measuring the accuracy of diagnostic system. Science 1988;240(4857):1285−93.

[17] Van Den Eeckhaut M, Reichenbach P, Guzzetti F, Rossi M, Poesen J. Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium. Nat Hazards Earth System Sci 2009;9:507−21.

[18] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer; 1995.

[19] Vasu NN, Lee SR. A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon, South Korea. Geomorphology 2016;263:50−70.

[20] Wang LJ, Guo M, Sawada K, Lin J, Zhang JC. Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. Catena 2015;135:271−82.