



Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China)

Wei Chen^{1,2} · Xusheng Yan¹ · Zhou Zhao¹ · Haoyuan Hong^{3,4,5} · Dieu Tien Bui⁶ · Biswajeet Pradhan^{7,8}

Received: 30 September 2017 / Accepted: 24 February 2018 / Published online: 13 March 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

The main goal of this study is to assess and compare three advanced machine learning techniques, namely, kernel logistic regression (KLR), naïve Bayes (NB), and radial basis function network (RBFNetwork) models for landslide susceptibility modeling in Long County, China. First, a total of 171 landslide locations were identified within the study area using historical reports, aerial photographs, and extensive field surveys. All the landslides were randomly separated into two parts with a ratio of 70/30 for training and validation purposes. Second, 12 landslide conditioning factors were prepared for landslide susceptibility modeling, including slope aspect, slope angle, plan curvature, profile curvature, elevation, distance to faults, distance to rivers, distance to roads, lithology, NDVI (normalized difference vegetation index), land use, and rainfall. Third, the correlations between the conditioning factors and the occurrence of landslides were analyzed using normalized frequency ratios. A multicollinearity analysis of the landslide conditioning factors was carried out using tolerances and variance inflation factor (VIF) methods. Feature selection was performed using the chi-squared statistic with a 10-fold cross-validation technique to assess the predictive capabilities of the landslide conditioning factors. Then, the landslide conditioning factors with null predictive ability were excluded in order to optimize the landslide models. Finally, the trained KLR, NB, and RBFNetwork models were used to construct landslide susceptibility maps. The receiver operating characteristics (ROC) curve, the area under the curve (AUC), and several statistical measures, such as accuracy (ACC), F-measure, mean absolute error (MAE), and root mean squared error (RMSE), were used for the assessment, validation, and comparison of the resulting models in order to choose the best model in this study. The validation results show that all three models exhibit reasonably good performance, and the KLR model exhibits

Highlights

- KLR, NB, and RBFNetwork models were compared in this study.
- The Chi-squared statistic was used to select conditioning factors.
- The ROC curve, ACC, MAE, and RMSE methods were used to assess the models' performances.
- KLR showed the most promising results.

✉ Wei Chen
chenwei.0930@163.com; chenwei0930@xust.edu.cn

✉ Haoyuan Hong
hong_haoyuan@outlook.com

¹ College of Geology & Environment, Xi'an University of Science and Technology, Xi'an 710054, China

² Shandong Provincial Key Laboratory of Depositional Mineralization & Sedimentary Minerals, Shandong University of Science and Technology, Qingdao 266590, China

³ Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210023, China

⁴ State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China

⁵ Jiangsu Center for Collaborative Innovation in Geographic Information Resource Development and Application, Nanjing, Jiangsu 210023, China

⁶ Department of Business Administration and Computer Sciences, Faculty of Arts and Sciences, Telemark University College, N-3800 Bø i Telemark, Norway

⁷ School of Systems, Management and Leadership, Faculty of Engineering and IT University of Technology Sydney, CB11.06.217, Building 11, 81 Broadway, PO Box 123, Ultimo, NSW 2007, Australia

⁸ Department of Energy and Mineral Resources Engineering, Choongmu-gwan, Sejong University, 209 Neungdong-ro Gwangjin-gu, Seoul 05006, Republic of Korea

the most stable and best performance. The KLR model, which has a success rate of 0.847 and a prediction rate of 0.749, is a promising technique for landslide susceptibility mapping. Given the outcomes of the study, all three models could be used efficiently for landslide susceptibility analysis.

Keywords Landslide · Kernel logistic regression · Naive Bayes · RBF network · China

Introduction

Landslide susceptibility is considered to be the spatial distribution of the probability of the occurrence of landslides by various investigators (Constantin et al. 2011; Guzzetti et al. 2006; Van Westen 2004). By using various models for predicting landslide, damage could be decreased to a certain extent (Pradhan 2010). In many areas of China, the lives and property of people have been seriously affected by landslides (Peng et al. 2015; Wu et al. 2015; Zhang et al. 2015). This is especially true in Long County, China, where thousands of people live in high-risk threat areas related to landslides. Therefore, it is necessary to prevent and assess landslide disasters for this area.

In the literature, various methods have been carried out to assess landslide susceptibility with the aid of GIS, such as probabilistic methods (Youssef et al. 2016), bivariate statistical methods (Constantin et al. 2011; Dou et al. 2014; Jaafari et al. 2014; Pradhan et al. 2014; Razandi et al. 2015; Regmi et al. 2014; Zhang et al. 2016), multivariate logistic regression methods (Park et al. 2013; Shahabi et al. 2015; Tien Bui et al. 2016), and knowledge-based methods (Althuwaynee et al. 2016; Kumar and Anbalagan 2016), etc.

Recently, machine-learning methods have been employed in landslide susceptibility mapping, such as support vector machines (Colkesen et al. 2016; Tien Bui et al. 2016), fuzzy logic (Kumar and Anbalagan 2015; Shahabi et al. 2015), neuro-fuzzy (Aghdam et al. 2016; Chen et al. 2017a; Dehnavi et al. 2015), neural network models (Gorsevski et al. 2016; Wang et al. 2016), kernel logistic regression (KLR) (Chen et al. 2017b; Tien Bui et al. 2016), naïve Bayes (NB) (Pham et al. 2017; Tsangaratos and Ilia 2016), random forest (Chen et al. 2017c; Trigila et al. 2015), and multivariate adaptive regression splines (Conoscenti et al. 2015; Felicísimo et al. 2013).

Although various methods have been used for landslide susceptibility assessment by different investigators throughout the world, some advanced machine-learning techniques, such as KLR and NB, have seldom been explored for landslide modeling, and a comparative study of the KLR, NB, and radial basis function network (RBFNetwork) models has never been used in the area of landslide susceptibility. Therefore, these three models were chosen to assess landslide susceptibility for the Long County area, China. This is an area where many landslides

have occurred during the last 10 years; however, the number of studies on these landslides is far from sufficient. The main difference between the current study and the approaches described in the aforementioned publications is that three state-of-the-art machine-learning techniques have been employed to produce landslide susceptibility maps, and their relative performance has been assessed for the first time.

General description of the study area

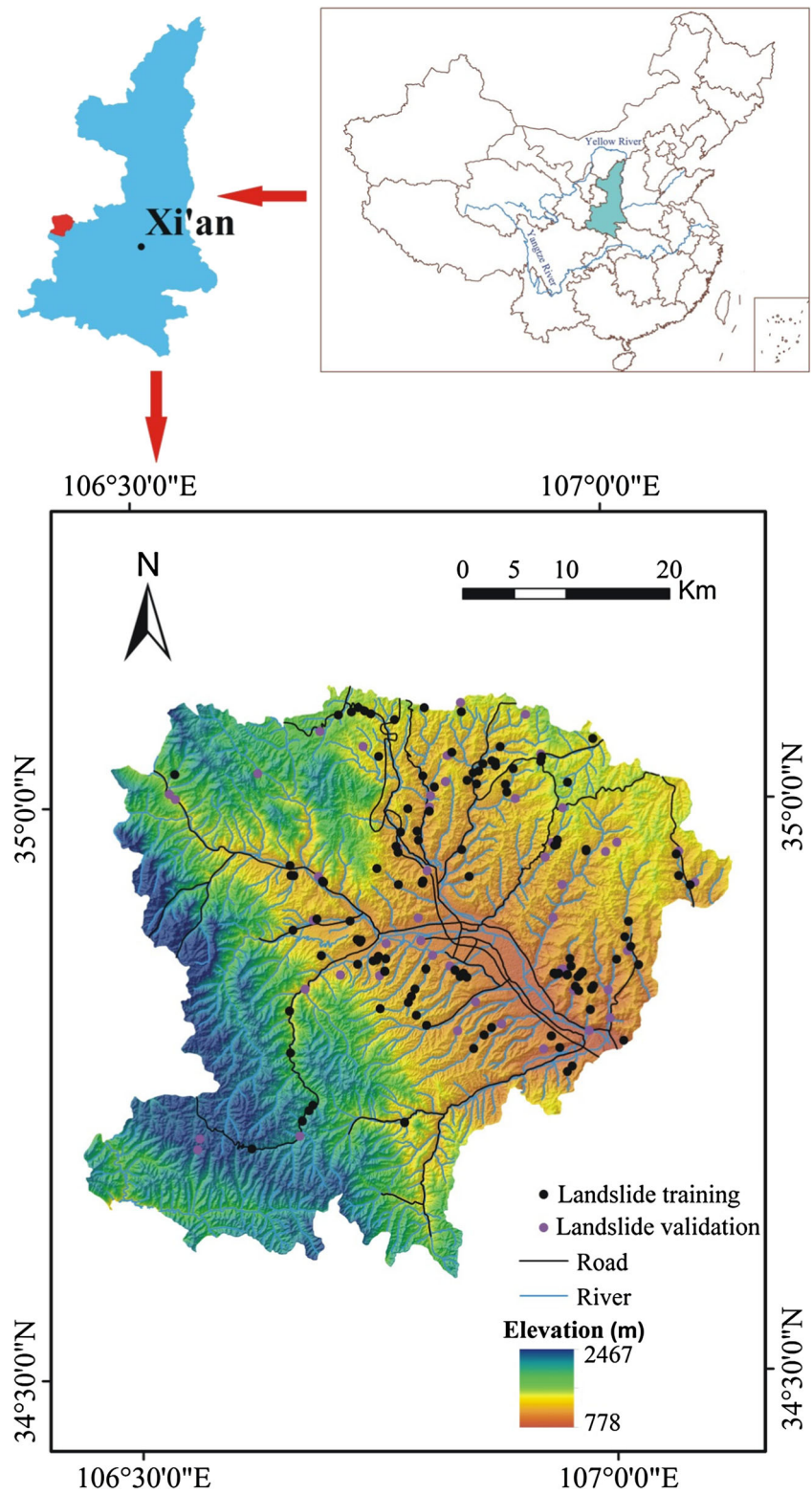
Long County is located in the northwest part of Baoji City, China (Fig. 1). The study area covers an area of 2400 km² and is located within 34°35'17"N to 35°6'45"N and 106°26'32"E to 107°8'11"E. The elevation of the highest point is 2467 m a.s.l, the lowest elevation is 778 m a.s.l, and the elevation decreases from the west to the east.

The Qian River and the Wei River are the main rivers in the study area, and they both belong to the Yellow River system. The Qian River is a tributary of the Wei River, and is the largest river in the region. The length of the Qian River in Long County is 68.8 km, the drainage area is 1957.9 km², the annual runoff is 3.54×10^8 m³, and the average flow is 5.6 m³/s.

The climate of Long County is affected by the Qinling Mountains, but it is mainly controlled by the Siberian High. According to the statistics of the Shanxi Meteorological Bureau (<http://www.sxmb.gov.cn>), it has obvious continental climate characteristics, which is a warm temperate continental monsoon climate zone. The annual average temperature is 10.7 °C. The hottest month is July, with an average temperature of 23.4 °C, and the coldest month is January, with an average temperature of −2.4 °C. The mean annual precipitation (MAP) is 600 mm, the maximum annual precipitation is 937 mm, and the minimum annual precipitation is 339.86 mm. In July–September, the total precipitation is 345 mm, which accounts for 58.2% of the annual precipitation.

Approximately 62% of the study area has a slope gradient less than 20°, whereas areas with a slope gradient steeper than 40° account for 2% of the total study area. Areas that fall into the slope category of 20–40° account for 36% of the total study area.

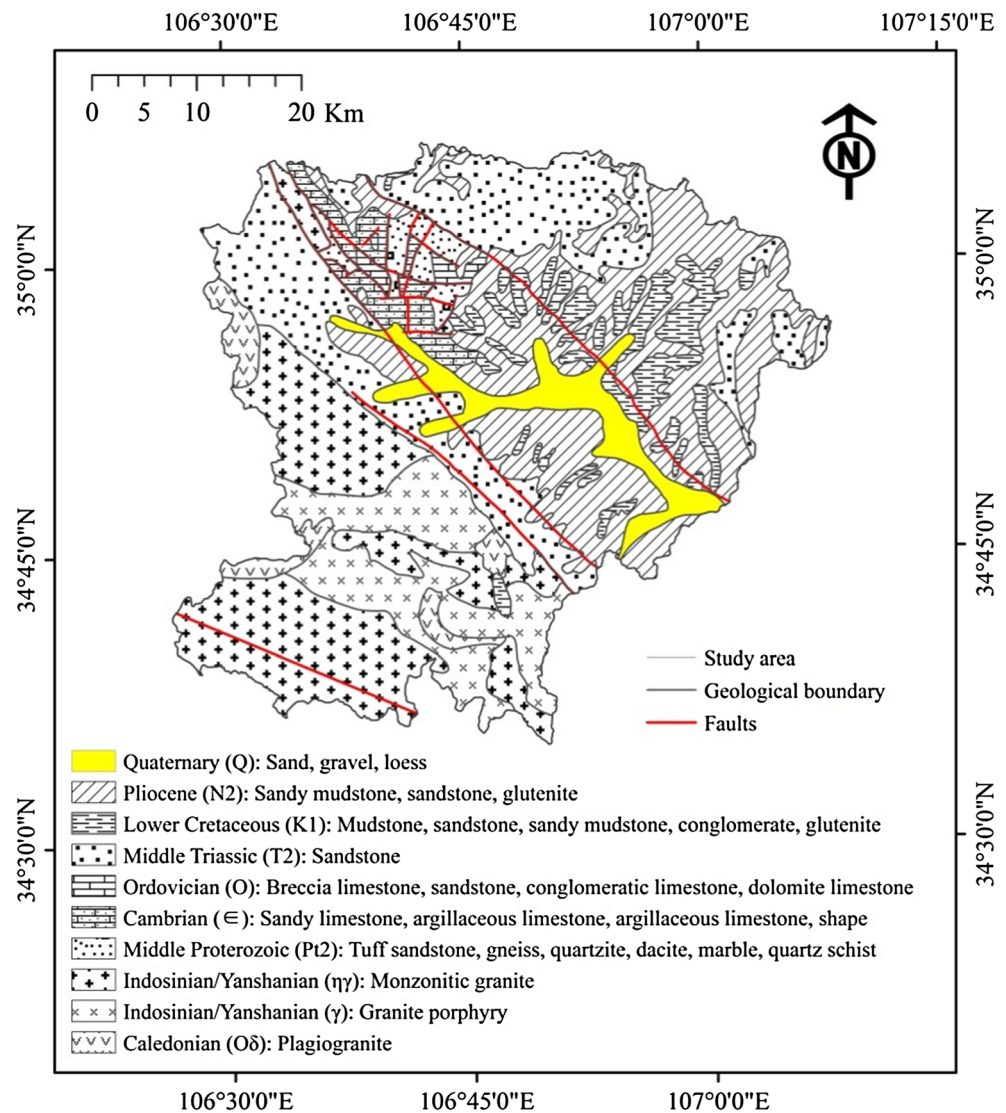
Fig. 1 Location of the study area and landslide inventory



The main lithologies are loess, mudstone, sandstone, conglomerate, glutenite, limestone, and igneous rocks. There are more than 20 geological units in the study area,

and these geological units can be reclassified into 10 lithological groups according the lithofacies and geologic ages (Fig. 2).

Fig. 2 Geology of the study area



Research methods and models

Methodology

A flowchart of the methodology used in this study is shown in Fig. 3. There are four main steps in the current study: (1) data preparation, (2) landslide conditioning factor analysis, (3) landslide susceptibility modeling, and (4) model performance evaluation.

Data preparation

Landslide inventory

Landslide inventory maps are considered to be essential for studying the relationship between landslides and landslide influencing factors (Booth et al. 2015; Mohammady et al. 2012). In the current study, a landslide inventory map has

been produced from various sources, including existing reports, interpretation of aerial photographs, and extensive field surveys in the study area. The landslide inventory map includes a total of 171 landslides, of which 120 (70%) landslides were randomly selected using Hawth's Analysis Tools for modeling, and the remaining 51 (30%) landslides were used for validating the models. An analysis of the landslide inventory map shows that the area of the smallest landslide is 30 m², the largest is 800,000 m², and the average is 60,000 m². Figure 1 shows the distribution of landslide locations in the study area.

Landslide conditioning factors

In previous studies, there has been no agreement in terms of the selection of reliable landslide conditioning factors because the root causes of landslides are very complex and difficult to confirm (Carlini et al. 2016; Cook et al. 2015).

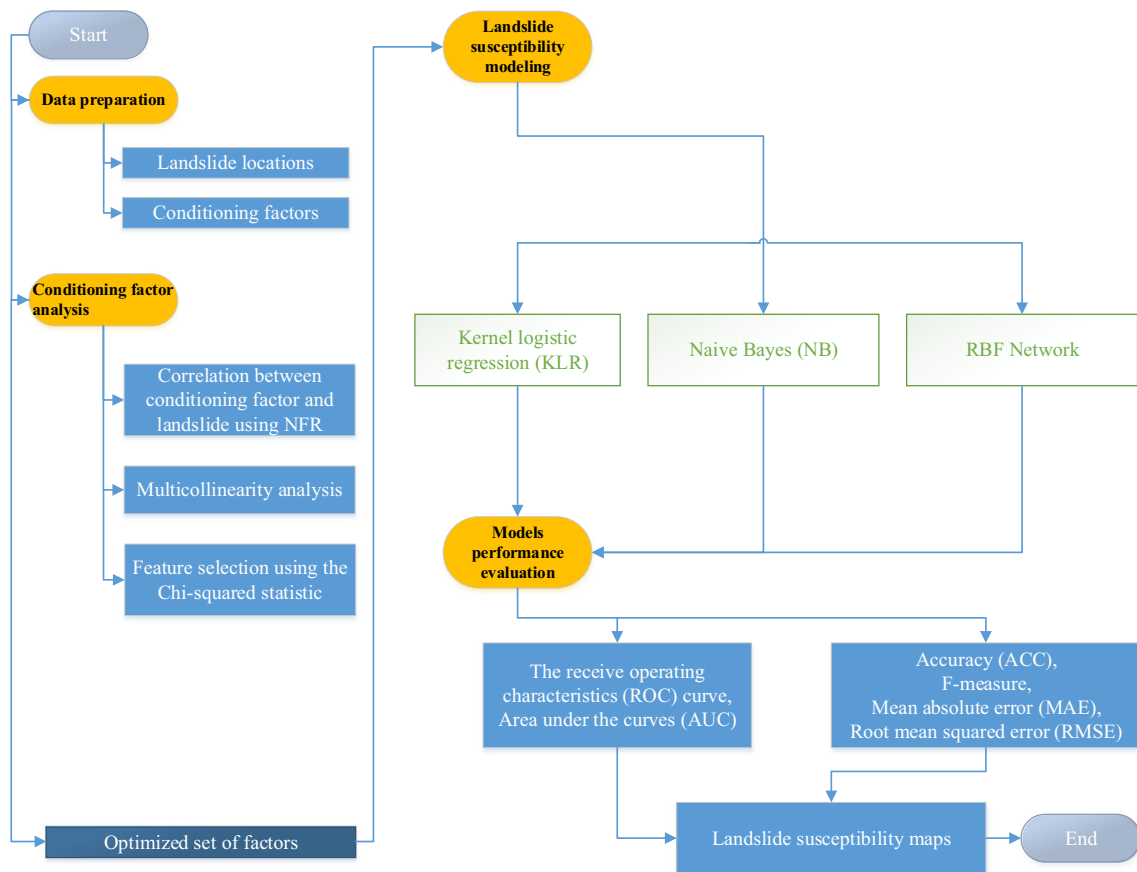


Fig. 3 Flowchart of the methods

Therefore, based on a review of the literature and conditions within the study area, a total of 12 landslide conditioning factors were selected, specifically slope aspect, slope angle, plan curvature, profile curvature, elevation, distance to faults, distance to roads, distance to rivers, lithology, normalized difference vegetation index (NDVI) values, land use, and rainfall.

Slope angle, slope aspect, plan curvature, profile curvature, elevation, and distance to rivers maps (Fig. 4a–e, g) for the study area were derived from the ASTER Global DEM with a spatial resolution of 30×30 m using ArcGIS 10.0 software. The ASTER Global DEM data were collected from the International Scientific & Technical Data Mirror Site of the Computer Network Information Center, Chinese Academy of Sciences (<http://www.gscloud.cn>). The distance to faults map and the lithology map (Fig. 4f, i) were extracted from geological maps at a scale of 1:200,000, which were collected from the local Land and Resources Bureau. The distance to roads map (Fig. 4h) was created using a 1:50,000 scale topographic map of the study area and ArcGIS 10.0 software. The NDVI and land use maps (Fig. 4j, k) were extracted from Landsat 8 OLI_TRIS images using ENVI 5.1 software. The Landsat 8 OLI_TRIS images with a spatial resolution of 30×30 m

were obtained from the same website with the ASTER Global DEM data. The rainfall map was constructed using rainfall data covering the study area, which were obtained and compiled from the Meteorological Bureau of Shaanxi Province (Fig. 4l). All 12 landslide conditioning factors were converted to the same spatial resolution (30×30 m). Detailed information on the classes of each conditioning factor is shown in Table 1.

Kernel logistic regression (KLR)

KLR is defined as a kernel version of logistic regression. The main aim of this method is to find a discriminant function which can solve classification problems by means of transforming the original input space into a high-dimensional feature space using kernel functions (Cawley and Talbot 2008; Tien Bui et al. 2016). In landslide susceptibility evaluation, suppose a training dataset has n input samples (v_i, C_i) with $v_i \in R^n$, $C_i \in \{0, 1\}$, where v is a vector of the input space. In this case study, it indicates the selected landslide conditioning factors: slope aspect, slope angle, plan curvature, profile curvature, elevation, distance to faults, distance to rivers, distance to roads, lithology, NDVI values, land use, and rainfall. C_i takes one of two values: 0 indicates a non-

landslide pixel, and 1 indicates a landslide pixel. The aim of KLR is to construct a discriminant function that can separate the two classes, landslide and non-landslide. The discriminant function can be expressed as:

$$\text{logit}\{C(v)\} = w \cdot \varphi(v) + b \quad (1)$$

where w is a vector of model parameters, $\varphi(v)$ is a non-linear transformation of the output vectors, v is defined by a kernel function, and b is the bias term. $K: v \times v \rightarrow R$ evaluates the inner product between the image of input vectors in the feature space, i.e., $K(v, v') = \phi(v) \cdot \phi(v')$. For a kernel to support the interpretation as an inner product in a fixed

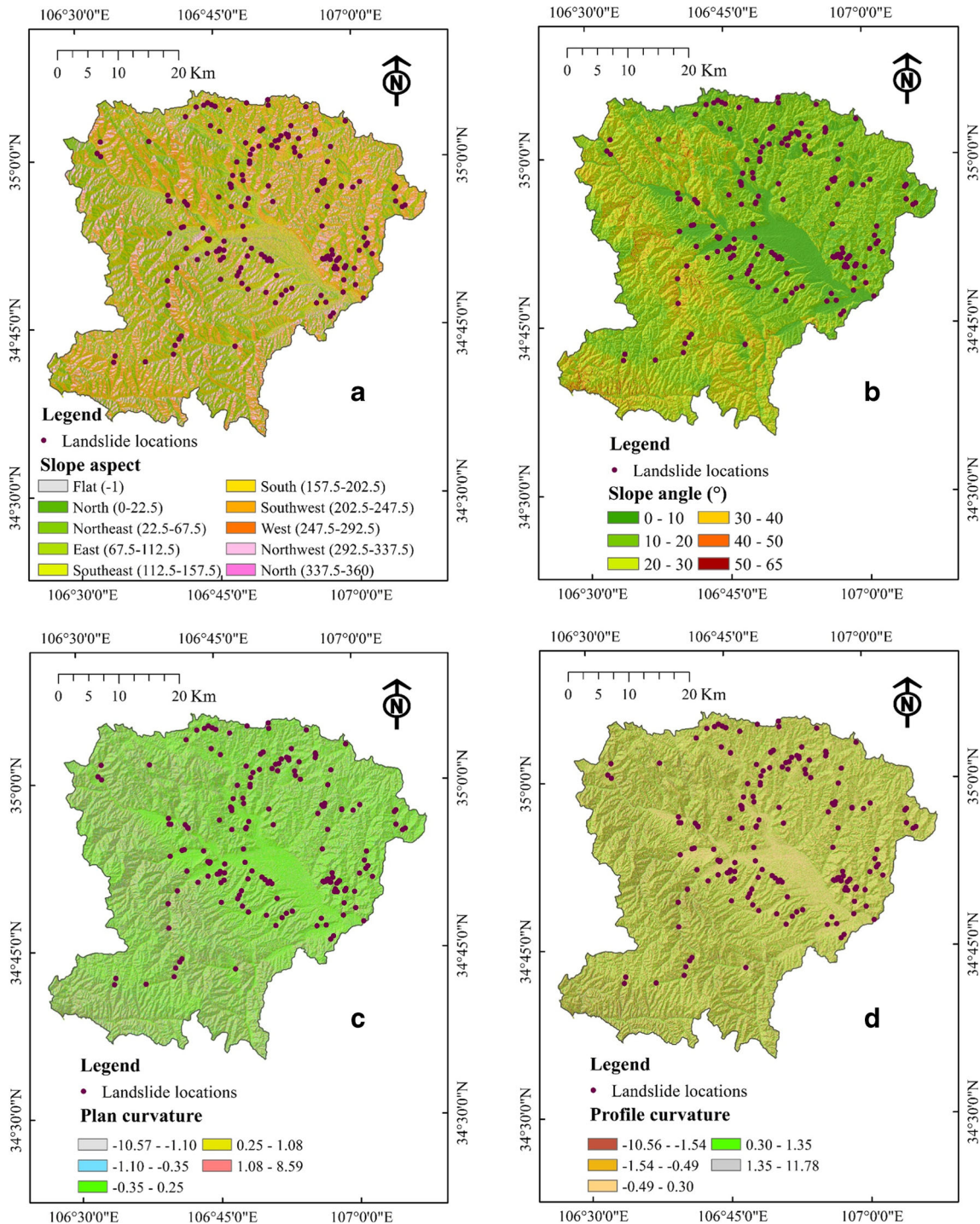


Fig. 4 Thematic maps of the study area: (a) slope aspect; (b) slope angle; (c) plan curvature; (d) profile curvature; (e) elevation; (f) distance to faults; (g) distance to rivers; (h) distance to roads; (i) lithology; (j) NDVI; (k) land use; (l) rainfall

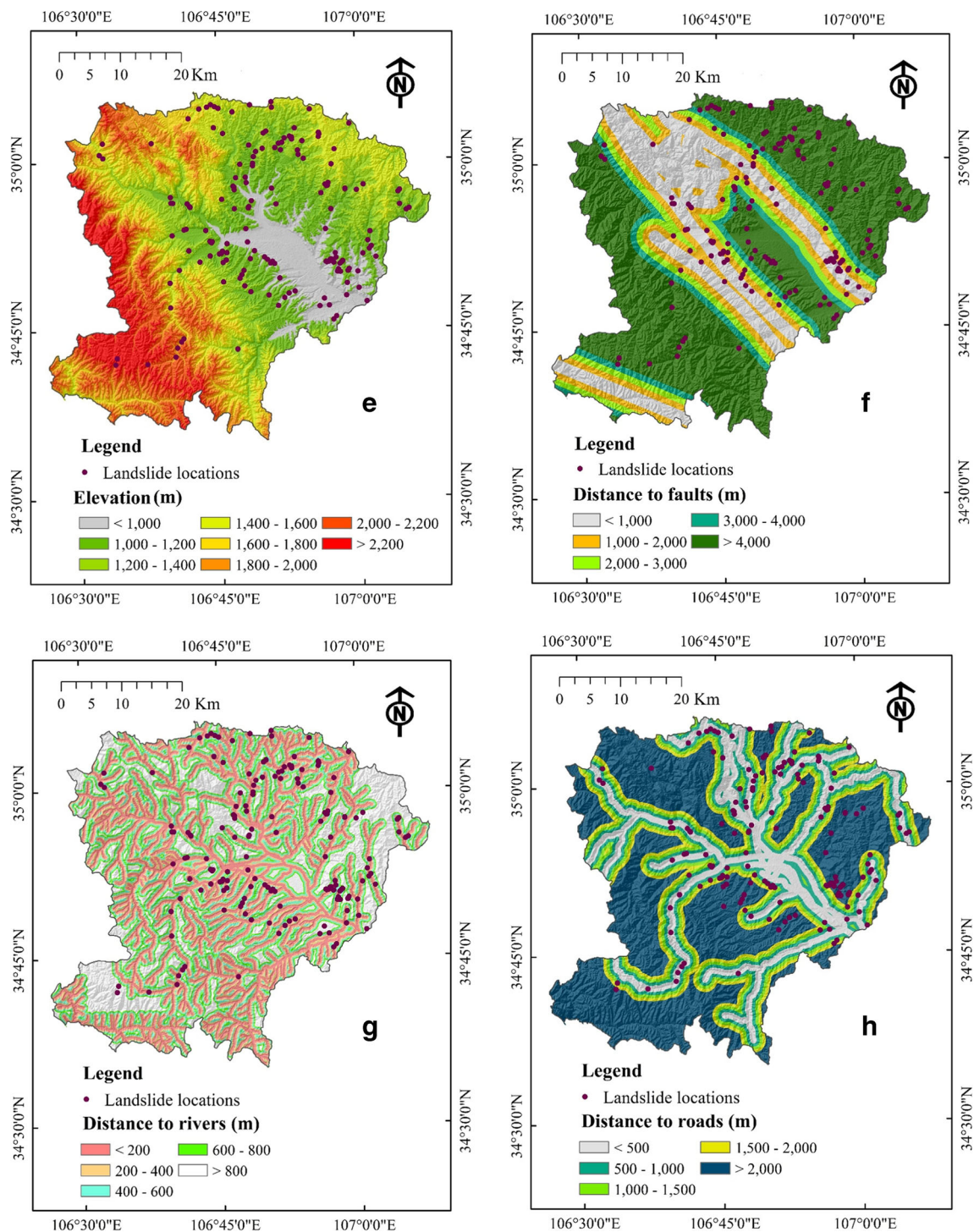


Fig. 4 (continued)

feature space, the kernel must obey Mercer’s condition (Mercer 1909); that is, the Gram matrix for the kernel, $K = [K_{ij} = K(v_i, v_j)]_{i,j=1}^l$, must be positive semi-definite. Provided that the training procedure can be formulated in such a way that the input vectors, v_i , appear only in the form of inner products, this construction allows the use of

very high dimensional feature spaces, resulting in very flexible, powerful models. The radial basis function (RBF) kernel may be the most commonly encountered kernel:

$$K(v, v') = \exp\left\{-\eta \|v - v'\|^2\right\}, \tag{2}$$

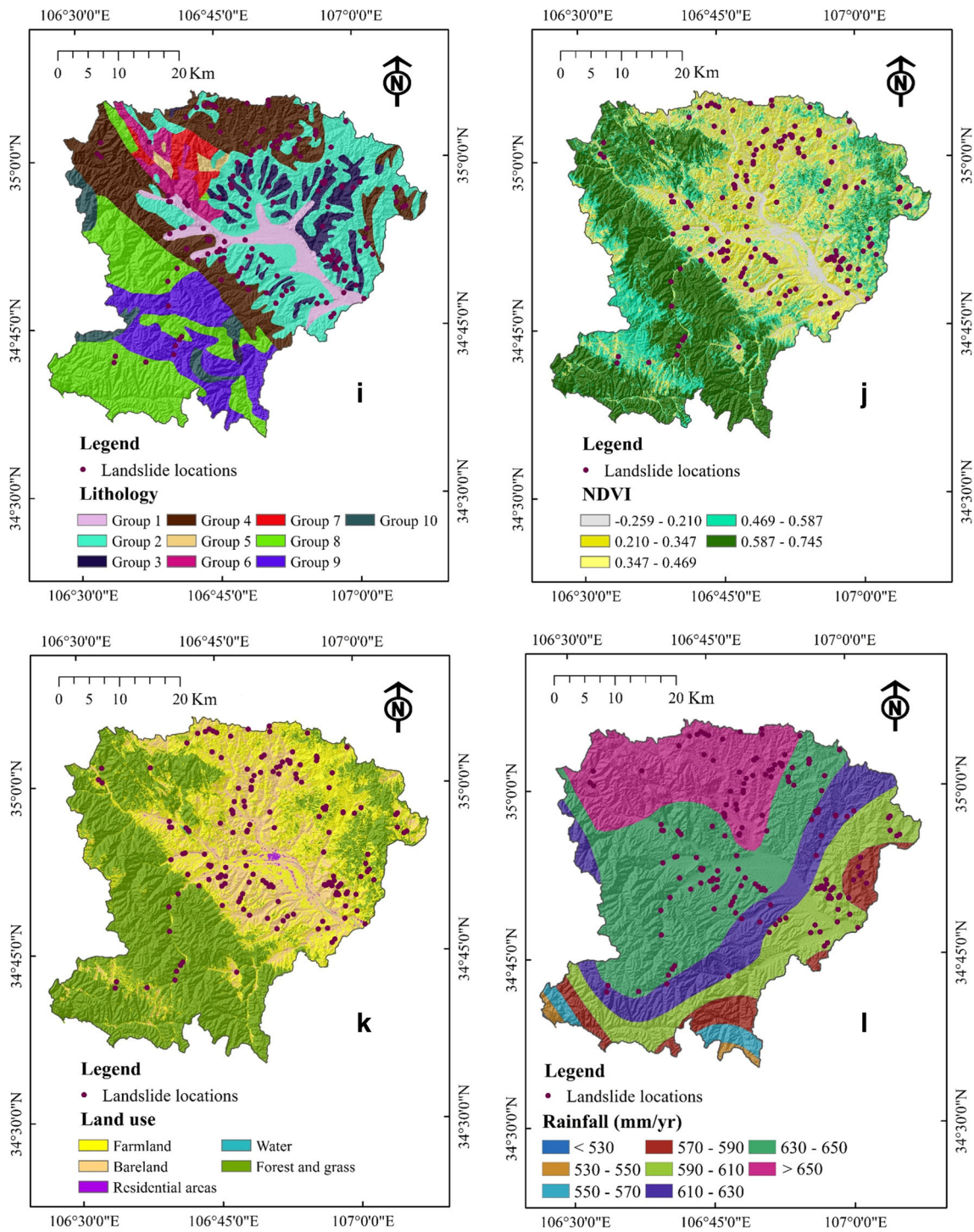


Fig. 4 (continued)

where η is a kernel parameter that controls the sensitivity of the kernel.

When constructing a statistical model of a landslide-prone area, as is the case here, it is prudent to take steps to avoid over-fitting the training data. As a result,

the KLR model was trained using a regularized cross-entropy loss function (Cawley and Talbot 2008):

$$E(w, b) = - \sum_{i=1}^n \{p_i \log C_i + (1-p_i) \log(1-C_i)\} + \frac{\mu}{2} \|w\|^2 \quad (3)$$

Table 1 Classes of landslide conditioning factors

Conditioning factors	Classes
Slope aspect	(1) Flat; (2) North; (3) Northeast; (4) East; (5) Southeast; (6) South; (7) Southwest; (8) West; (9) Northwest
Slope angle (°)	(1) 0–10; (2) 10–20; (3) 20–30; (4) 30–40; (5) 40–50; (6) > 50
Plan curvature	(1) -10.57 to -1.10; (2) -1.10 to -0.35; (3) -0.35 to 0.25; (4) 0.25–1.08; (5) 1.08–8.59
Profile curvature	(1) -10.56 to -1.54; (2) -1.54 to -0.49; (3) -0.49 to 0.30; (4) 0.30–1.35; (5) 1.35–11.78
Elevation (m)	(1) < 1000; (2) 1000–1200; (3) 1200–1400; (4) 1400–1600; (5) 1600–1800; (6) 1800–2000; (7) 2000–2200; (8) > 2200
Distance to faults (m)	(1) < 1000; (2) 1000–2000; (3) 2000–3000; (4) 3000–4000; (5) > 4000
Distance to rivers (m)	(1) < 200; (2) 200–400; (3) 400–600; (4) 600–800; (5) > 800
Distance to roads (m)	(1) < 500; (2) 500–1000; (3) 1000–1500; (4) 1500–2000; (5) > 2000
Lithology	(1) Group 1; (2) Group 2; (3) Group 3; (4) Group 4; (5) Group 5; (6) Group 6; (7) Group 7; (8) Group 8; (9) Group 9; (10) Group 10
NDVI	(1) -0.26 to 0.21; (2) 0.21–0.35; (3) 0.35–0.47; (4) 0.47–0.59; (5) 0.59–0.74
Land use	(1) Farmland; (2) Bareland; (3) Residential areas; (4) Water; (5) Forest and grass
Rainfall (mm/year)	(1) < 530; (2) 530–550; (3) 550–570; (4) 570–590; (5) 590–610; (6) 610–630; (7) 630–650; (8) > 650

$$\|w\|^2 = \alpha^T K \alpha \tag{4}$$

where μ is a regularization parameter controlling the bias–variance trade-off (Geman et al. 1992). The representer theorem (Kimeldorf and Wahba 1971; Schölkopf and Smola 2002) states that the solution of the optimization criterion described by Eq. (4) can be expressed in the form of an expansion over training patterns:

$$w = \sum_{i=1}^n \alpha_i \varphi(v_i) \tag{5}$$

where α_i , with $i = (1, 2, \dots, n)$, is the vector of the dual model parameters. Thus, the final form of KLR will be obtained:

$$\text{logit}\{C(v)\} = \sum_{i=1}^n \alpha_i K(v_i, v) + b \tag{6}$$

There are many kernel functions that can be found in the literature, such as the linear kernel function, the RBF kernel function, the polynomial kernel function, and the normalized polynomial kernel function. Among these, the RBF kernel function is the most popular and is applied in this study.

$$K(v_i, v_j) = \exp\left(-\frac{\|v_i - v_j\|^2}{2\delta^2}\right) \tag{7}$$

where δ is the tuning parameter that controls the sensitivity of the kernel.

Naive Bayes (NB)

The NB classifier is considered to be a classification system based on Bayes’ theorem. The NB classifier assumes

that all the attributes are fully independent given the output class, which is called the conditional independence assumption (Soria et al. 2011). The main advantage of this method is that it is very easy to construct, and complicated iterative parameter estimation schemes are not needed (Wu et al. 2008). Additionally, the NB classifier is robust to noise and irrelevant attributes. This method has also been applied in the research field of landslide susceptibility mapping (Pham et al. 2015, 2016; Tien Bui et al. 2012).

Given an observation consisting of n attributes x_i , $i = 1, 2, \dots, k$, where x_i represents the landslide conditioning factors of slope aspect, slope angle, plan curvature, profile curvature, elevation, distance to faults, distance to rivers, distance to roads, lithology, NDVI values, and land use, n represents the number of landslide conditioning factors, and $y \in \{m, n\}$, where m and n indicate landslide and non-landslide areas, respectively, is the output class. NB estimates the probability $P(y_j/x_i)$ for all possible output classes. The prediction is made for the class with the largest posterior probability as

$$y_{NB} = \underset{y_i \in \{\text{landslide, non-landslide}\}}{\text{argmax}_y} P(y)P(x_1, x_2, \dots, x_i|y) \tag{8}$$

In this case, for computational convenience, NB assumes that all attributes are conditionally independent (Kim et al. 2016), so Eq. (8) can be simplified as follows:

$$y_{NB} = \underset{y_i \in \{\text{landslide, non-landslide}\}}{\text{argmax}_y} P(y) \prod_{i=1}^k P(x_i|y) \tag{9}$$

To calculate each posterior probability $P(x_i|y)$, numerical attributes are usually assumed to follow a normal distribution,

with the value of the probability density function in Eq. (9) used for the posterior probability (Kim et al. 2016):

$$P(x_i|y) = \frac{1}{\sqrt{2\pi}\sigma_{(i|y)}} e^{-\frac{(x_i-\mu_{(i|y)})^2}{2\sigma_{(i|y)}^2}} \quad (10)$$

where μ is the mean and σ^2 is the variance of x_i .

RBFNetwork

RBFNetwork, a popular alternative to multilayer perceptron neural nets, is defined as a supervised neural network for modeling problems in polydimensional space (Hong et al. 2015b). The architecture of this network consists of three layers in this study: an input layer with 12 neurons, a hidden layer (referred to as the RBF units), and an output layer that contains one neuron (Fig. 5). The hidden layer is activated by a radially symmetric basis function, and it can be expressed by $\phi_i: R_n \rightarrow R$; typically, the Gaussian function is used for the activation function.

The input data are processed by the RBF units using the K-means algorithm to reduce its dimensionality and then to transform the data to a new space (Gil and Johnsson 2010). The learning procedure of the RBFNetwork is carried out in two phases: (1) the numbers of clusters (hidden neurons) are calculated using the K-means algorithm, and (2) optimal estimation of the kernel parameter is conducted (Tien Bui et al. 2016).

Landslide conditioning factor analysis

Multicollinearity analysis

Multicollinearity refers to the non-independence of landslide conditioning factors that may occur in datasets because of

their high correlation, thus resulting in erroneous system analysis (Dormann et al. 2013). Tolerances and variance inflation factor (VIF) methods are commonly used in the literature to check the multicollinearity of landslide conditioning factors (Hong et al. 2015a). In this study, multicollinearity among the conditioning factors was identified using tolerances and VIF methods. The formulas for calculating tolerances and the VIF are as follows:

$$\text{Tolerance} = 1 - R_j^2 \quad (11)$$

$$\text{VIF} = \left[\frac{1}{\text{Tolerance}} \right] \quad (12)$$

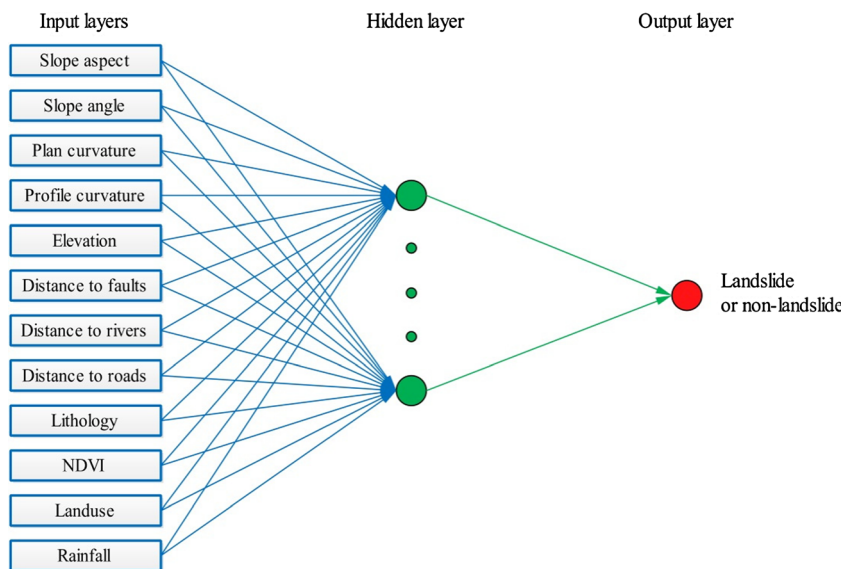
where R_j^2 is the coefficient of determination for the regression of the explanatory variable, j , on all the other explanatory variables (Sar et al. 2016).

Factor selection based on the chi-squared statistic

In landslide susceptibility analysis, the quality of input data should be as high as possible in order to reach an accurate and reliable result. In addition to this quality requirement, the selection of significant parameters (landslide conditioning factors) is another important step that must be carried out prior to landslide susceptibility modeling (Regmi et al. 2010; Tien Bui et al. 2016). Sometimes, several landslide conditioning factors may cause noise that reduces the prediction capability of the models (Lineback Gritzner et al. 2001). The chi-squared statistic-based feature selection method has been adopted in this study to select significant landslide conditioning factors for landslide susceptibility modeling (Andrews 1988; Rao and Scott 1987).

Among the techniques used to quantify the predictive capability of landslide conditioning factors, the chi-squared

Fig. 5 RBFNetwork used in this study



statistic is a common test for significance of the relationship between categorical variables (Press 1966). This method is based on the fact that one can compute the expected frequencies in a 2×2 contingency table (i.e., frequencies that we would expect if there were no relationship between the variables). To calculate the importance of landslide conditioning factors using the chi-squared statistic method, the null hypothesis states that knowing the level of a landslide conditioning factor does not help to predict the landslide occurrence (Satorra and Bentler 2001).

A goodness-of-fit statistic, χ^2 , is defined as following:

$$\chi^2 = \sum_{i=1}^n \frac{(F_{oi} - F_{ei})^2}{F_{ei}} \quad (13)$$

where F_{oi} represents the observed frequencies and F_{ei} represents the expected frequencies. A higher chi-squared statistic indicates higher performance of the landslide conditioning factor in the identification of landslides.

Model performance evaluation

In landslide susceptibility modeling, it is important to perform a validation of the predicted results. Without validation, the predicted models and prepared maps are useless and without scientific significance (Chung and Fabbri 2003). In the literature, the receiver operating characteristic (ROC) curve is a useful tool that is commonly used for assessing the performance of the landslide susceptibility models.

The ROC curve is a plot of the true positive rate (also known as the sensitivity) and the false positive rate (also known as 1-specificity) for all possible cut-off values. In landslide susceptibility analysis, true positives represent the number of pixels that are correctly classified as landslides, and false positives are pixels in landslide-free areas that are misclassified as landslide points. The area under the ROC curve (AUC) is an indicator of the capability of a model to predict the landslide and non-landslide pixels. An AUC value of 1 indicates a perfect classification, while an AUC value of 0 indicates a non-informative model. For a random model, the AUC value is 0.5 (Tien Bui et al. 2016; Walter 2002). In addition to the AUC values, the standard error of the AUC values was also used, and a smaller standard error indicates a better model.

Several statistical measures, such as the accuracy (ACC), the F-measure, the mean absolute error (MAE) and the root mean squared error (RMSE), have also been used to assess the predictive capability of the landslide models:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (15)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (16)$$

F-measure

$$= 2 * \text{Sensitivity} * \text{Specificity} / (\text{Sensitivity} + \text{Specificity}) \quad (17)$$

Mean absolute error (MAE)

$$= \frac{|p_1 - a_1| + |p_2 - a_2| + \dots + |p_n - a_n|}{n} \quad (18)$$

Root mean squared error (RMSE)

$$= \text{Sqrt} \left(\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n} \right) \quad (19)$$

where TP and TN are the proportions of pixels that are correctly classified, FP and FN are the proportions of pixels that are incorrectly classified, the *F-measure* combines the sensitivity and specificity into their harmonic mean, and p_i is the predicted value, while a_i is the actual value (Tien Bui et al. 2014).

Results

Frequency ratio-based correlation between landslides and conditioning factors

The frequency ratios (FR) for the class or type of each conditioning factor were calculated by dividing the landslide occurrence ratio by the area ratio (Razavizadeh et al. 2017). In this study, all calculated FRs were normalized to the range of 0 to 1 (NFR). Figure 6 shows the spatial relationship between each landslide conditioning factor and landslide in terms of their NFRs. For slope aspect, southwest-facing slopes have the highest NFR value (0.198), followed by south-facing slopes (0.159), southeast-facing slopes (0.138), northwest-facing slopes (0.128), north-facing slopes (0.103), west-facing slopes (0.096), east-facing slopes (0.092), northeast-facing slopes (0.085), and flat areas (0). It can be seen that southwest-facing slopes are more prone to landslides. In terms of slope angle, the class of $<10^\circ$ has the highest NFR value (0.410); the other classes, such as the $10\text{--}20^\circ$ class (0.338), the $20\text{--}30^\circ$ class (0.228), the $30\text{--}40^\circ$ class (0.023), the $40\text{--}50^\circ$ class (0), and the $>50^\circ$ class (0) have lower NFR values. In the case of plan curvature, the 0.25–1.08 class is most prone to landslides because it has the maximum NFR value (0.261), followed by -0.35 to 0.25 (0.223), -1.10 to -0.35 (0.191), -10.57 to -1.10 (0.190), and $1.08\text{--}8.59$ (0.135). For profile curvature, the -10.56 to -1.54 class has the maximum NFR value (0.233), representing a high probability of landslide occurrence. In the case of elevation, the highest NFR value corresponds to the

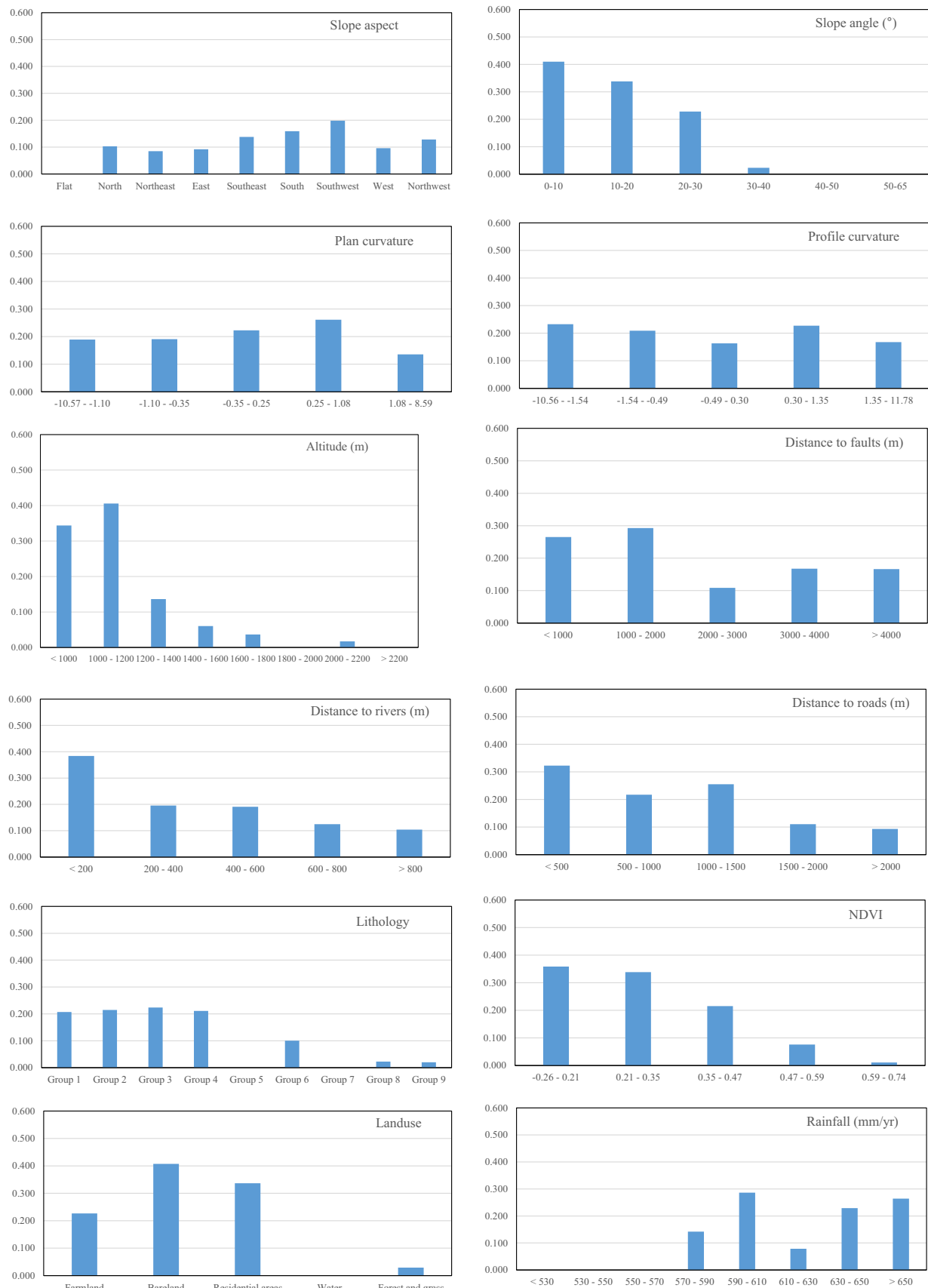


Fig. 6 Relationship between the landslides and the conditioning factors

elevation range between 1000 m and 1200 m (0.406), indicating that the highest probability of landslide occurrence is in this range. For distance to faults, < 1000 m, 1000–2000 m, 2000–3000 m, 3000–4000 m, and >4000 m have NFR values of 0.265, 0.293, 0.109, 0.168, and 0.166, respectively. It can be seen that the class of 1000–2000 m has the highest NFR value. In the case of distance to rivers, the highest value of NFR is seen in the <200 m class (0.384). In general, smaller distances to rivers correspond to greater landslide-occurrence probabilities. For the distance to roads, the class of <500 m has the highest NFR value (0.323); similarly, smaller distances to roads correspond to greater landslide-occurrence probabilities in this study. In the case of lithology, the highest NFR value corresponds to group 3 (0.224), whereas the lowest value of NFR (0) corresponds to groups 5, 7, and 10. For NDVI, the –0.26 to 0.21 class has the highest value of NFR (0.359), followed by 0.21–0.35 (0.339), 0.35–0.47 (0.215), 0.47–0.59 (0.076), and 0.59–0.74 (0.011). For land use, the highest NFR value corresponds to bare land (0.407), while the lowest NFR value corresponds to water (0). In the case of rainfall, the highest NFR value corresponds to the class of 590–610, whereas the lowest NFR values (0) corresponds to the classes of <530, 530–550, and 550–570.

Multicollinearity analysis of landslide conditioning factors

The calculated tolerances and VIF values for the 12 landslide conditioning factors are shown in Table 2. The results show that the lowest tolerance is 0.284 and the highest VIF value is 3.521 for the distance to rivers. All these values satisfy the critical values (tolerance <0.1 or VIF >10), which indicate that there is no multicollinearity among the 12 landslide conditioning factors.

Selection of landslide conditioning factors

After the multicollinearity analysis of landslide conditioning factors has been conducted, the next step is to select the best conditioning factors based on the chi-squared statistic. Table 3 shows the calculated average merit (AM) as the average chi-squared statistic and its standard deviation (SD) using 10-fold cross-validation for the 12 landslide conditioning factors. It can be seen that the highest AM is for elevation (72.092), which means that elevation has the highest predictive capability. This result is consistent with that of another study that was carried out by Tien Bui et al. (2016). Eight landslide conditioning factors have smaller AM values, such as NDVI (65.830), land use (64.661), lithology (45.291), distance to rivers (30.134), distance to roads (22.465), slope angle (10.660), rainfall (4.496), and distance to faults (1.962), which means these eight factors make smaller contributions to landslide

Table 2 Multicollinearity analysis for the landslide conditioning factors

Landslide conditioning factors	Collinearity statistics	
	Tolerance	VIF
Slope aspect	0.784	1.276
Slope angle	0.970	1.031
Plan curvature	0.980	1.020
Profile curvature	0.965	1.036
Elevation	0.459	2.177
Distance to faults	0.284	3.516
Distance to rivers	0.284	3.521
Distance to roads	0.936	1.069
Lithology	0.828	1.208
NDVI	0.941	1.062
Land use	0.423	2.363
Rainfall	0.913	1.095

modeling. Three landslide conditioning factors (slope aspect, plan curvature, and profile curvature) have AM values of 0, which means that they make no contribution to the landslide modeling. In consequence, the inclusion of these three factors may reduce the predictive accuracy of the resulting models, so they were discarded from the analysis in the study. Therefore, only nine conditioning factors were adopted in this study.

Application of the KLR model

The KLR model was constructed using the Weka software package (v.3.90). During the training process, a heuristic test was carried out to find the best kernel parameters for the RBF kernel using the training data and 10-fold cross-validation.

Table 3 Average Chi-squared statistic for the landslide conditioning factors

Number	Landslide conditioning factors	Average merit	SD
1	Elevation	72.092	± 4.738
2	NDVI	65.830	± 3.960
3	Land use	64.661	± 5.420
4	Lithology	45.291	± 1.896
5	Distance to rivers	30.134	± 6.040
6	Distance to roads	22.465	± 2.832
7	Slope angle	10.660	± 1.179
8	Rainfall	4.496	± 4.577
9	Distance to faults	1.962	± 3.927
10	Slope aspect	0	0
11	Plan curvature	0	0
12	Profile curvature	0	0

The best values of the tuning parameter and the regularizing parameter are 0.02 and 0.02, respectively. The model was then applied to the whole study area. The calculated landslide susceptibility index (LSI) values were in the range of 0.000 to 0.955. Obviously, a higher landslide susceptibility index value indicates a higher probability of a landslide occurrence. The produced landslide susceptibility map (LSM) was reclassified into five classes using the LSI intervals such as very low (0.000–0.090), low (0.090–0.251), moderate (0.251–0.427), high (0.427–0.625), and very high (0.625–0.955) using the natural breaks method (Fig. 7). The results show that the very low class has the largest area percentage (51.75%), followed by the low (16.64%), moderate (13.61%), high (11.23%), and very high (7.13%) classes.

Application of the NB model

The NB model was constructed in a similar way as the KLR model. The calculated LSI values were in the range of 0.000–

0.999. Similarly, the produced landslide susceptibility map was also reclassified into five classes using the LSI intervals, such as very low (0.000–0.118), low (0.118–0.364), moderate (0.364–0.619), high (0.619–0.842), and very high (0.842–0.999), using the natural breaks method (Fig. 8). The results show that the very low class has the largest area percentage (57.19%), followed by the low (7.32%), moderate (6.63%), high (8.68%), and very high (20.55%) classes.

Application of the RBFNetwork model

Similarly, LSI values were also calculated using the RBFNetwork model. To determine the number of clusters in the hidden layer, a heuristic test was also carried out with varying numbers of clusters versus the area under the receiver operating characteristics curve (AUC) and the predictive accuracy (ACC) values using both the training and validation data. Finally, the optimal number of clusters in the hidden layer was determined to be 6 for the

Fig. 7 Landslide susceptibility map derived from the KLR model

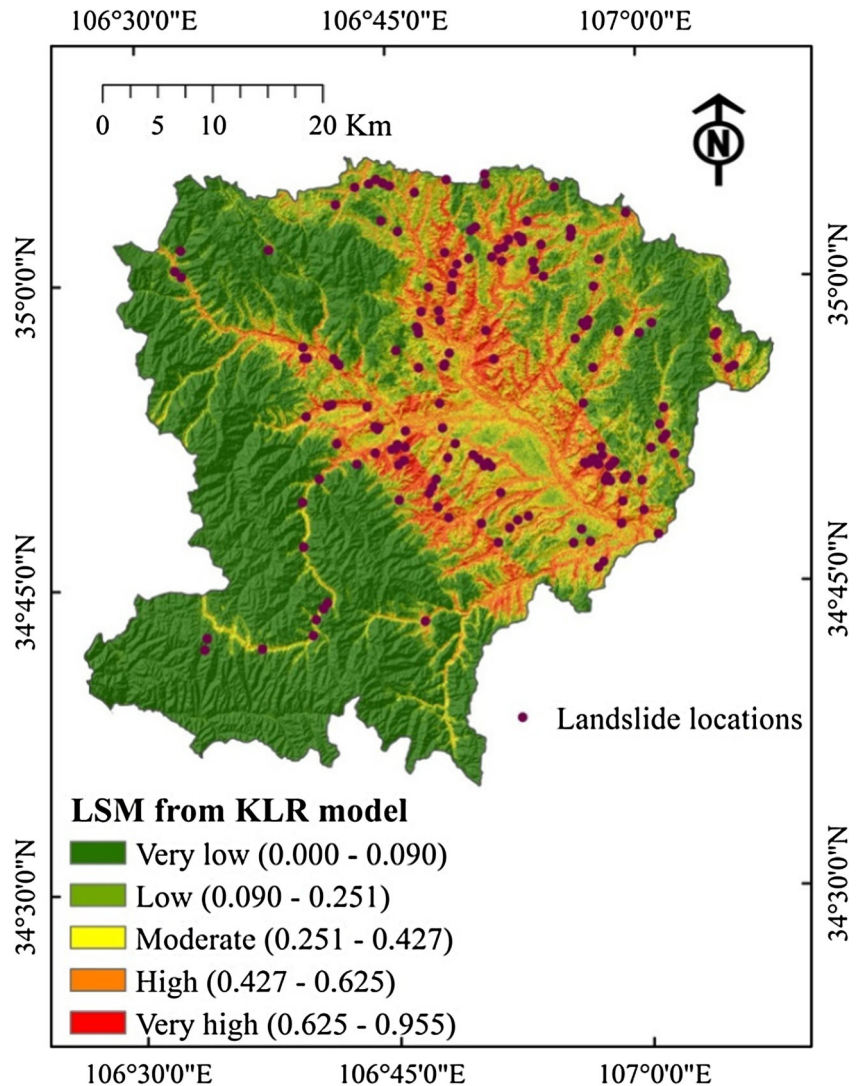


Fig. 8 Landslide susceptibility map derived from the NB model

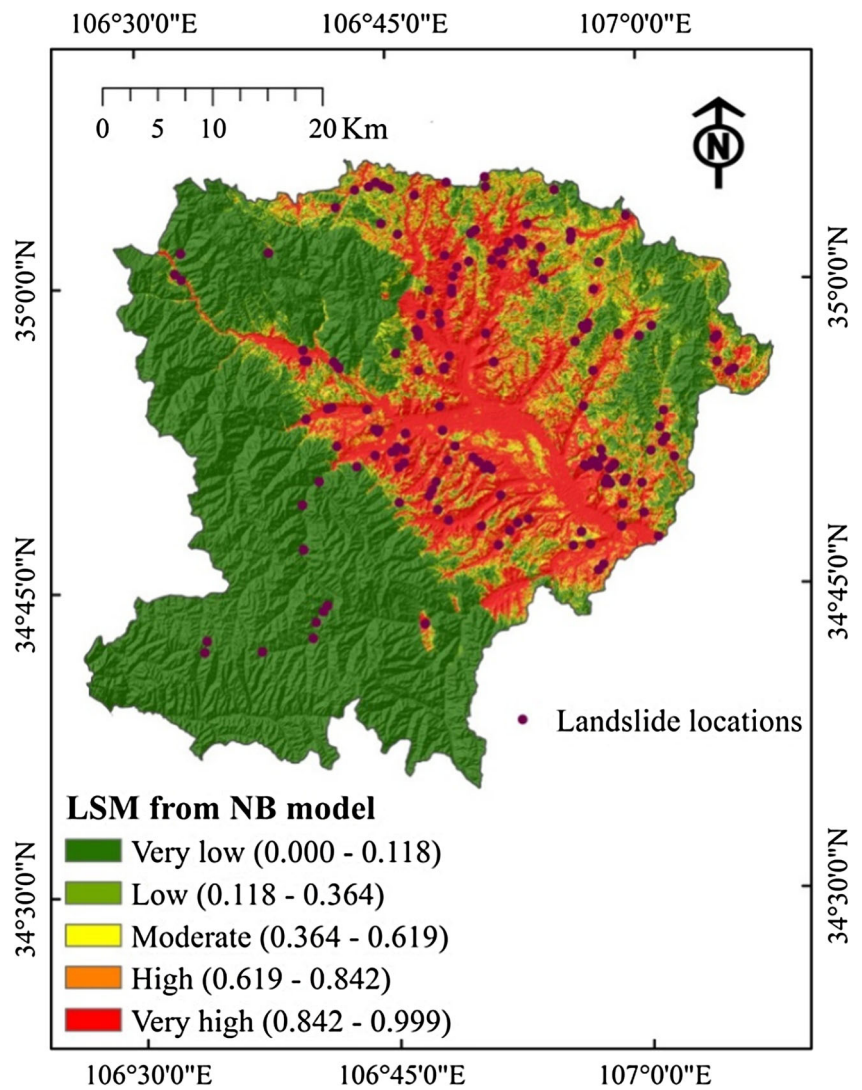


Fig. 9 Selection of clusters for the RBFNetwork model

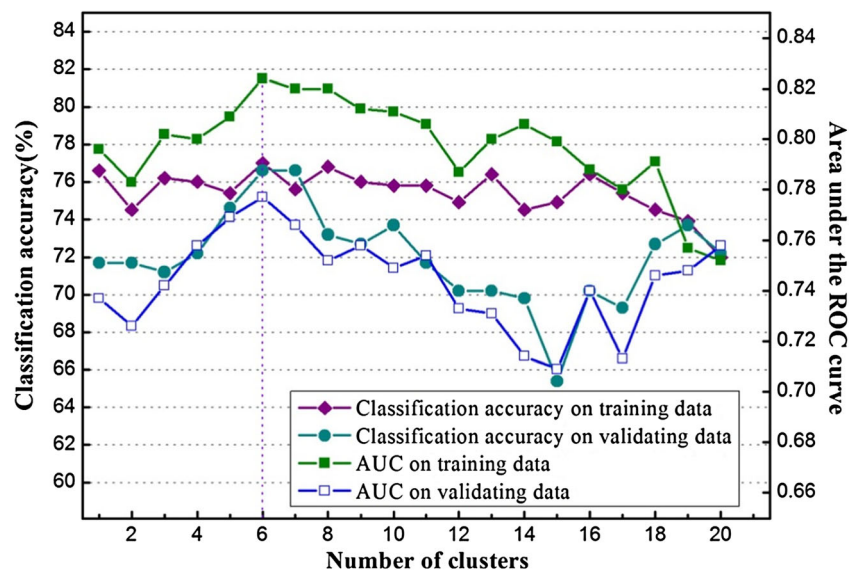


Fig. 10 Landslide susceptibility map derived from the RBFNetwork model

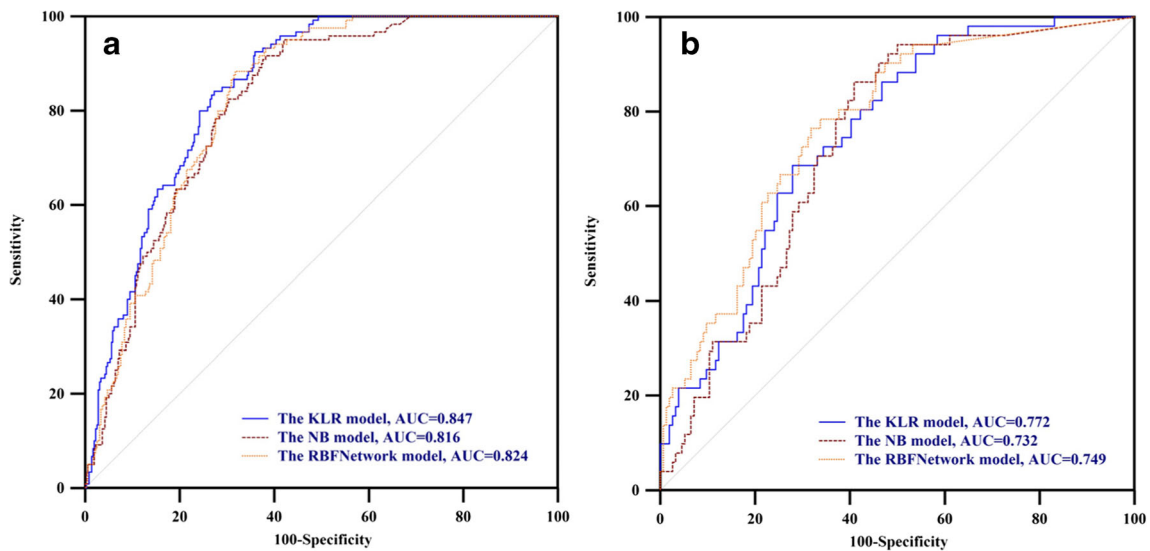
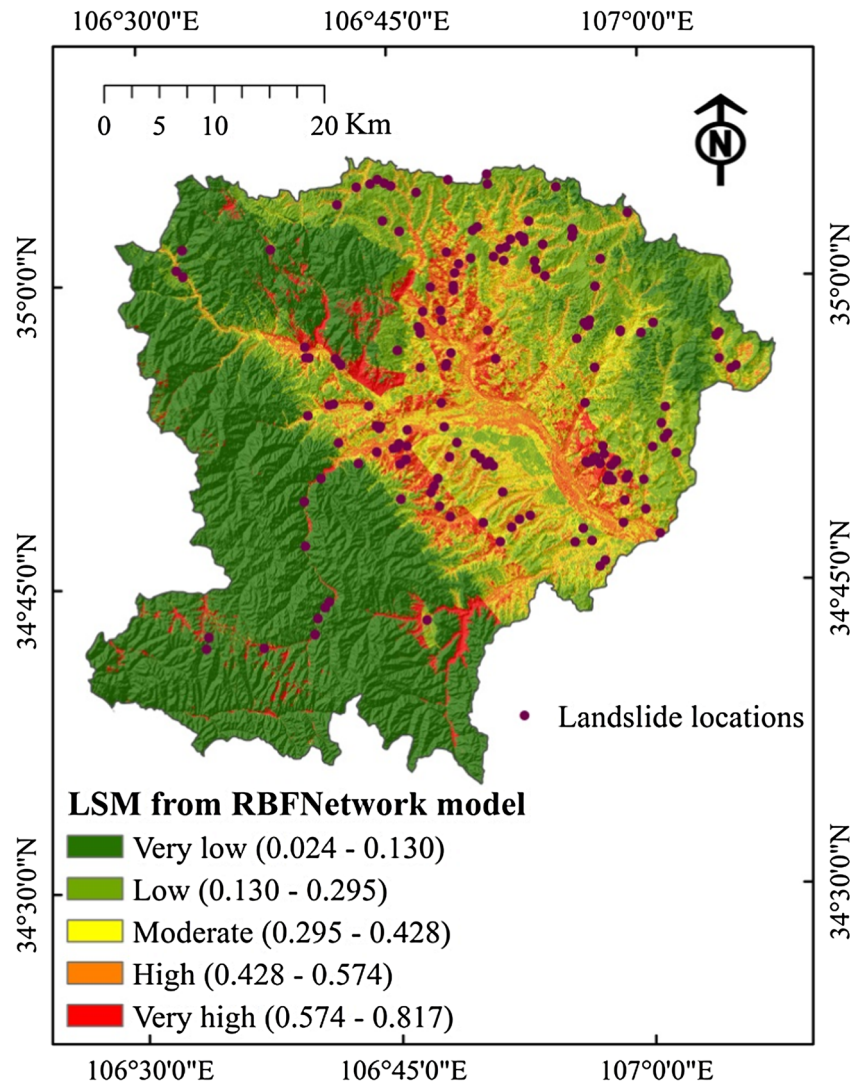


Fig. 11 ROC curves for the three landslide susceptibility maps produced by the KLR, NB, and RBFNetwork models: (a) success rate, (b) prediction rate

Table 4 Performance of the three models on the training dataset

Parameters	KLR	NB	RBFNetwork
AUC	0.847	0.816	0.824
ACC (%)	0.791	0.729	0.754
F-measure	0.787	0.746	0.745
MAE	0.259	0.279	0.268
RMSE	0.370	0.461	0.388

RBFNetwork model (Fig. 9). The constructed model was then applied the whole study area to calculate the LSI. The calculated LSI values were in the range of 0.024–0.817. All of the LSI values were imported into ArcGIS 10 to produce the landslide susceptibility map. The landslide susceptibility map was reclassified into five classes using the LSI intervals, such as very low (0.024–0.130), low (0.130–0.295), moderate (0.295–0.428), high (0.428–0.574), and very high (0.574–0.817) using the natural breaks method (Fig. 10). The results show that the very low class has the largest area percentage (47.98%), followed by the low (22.95%), moderate (12.38%), high (9.88%), and very high (7.18%) classes.

Model performance and validation

Using the selected nine landslide conditioning factors, the KLR, NB, and RBFNetwork models were constructed using the training data. The ROC curves and AUC values of these three models are shown in Fig. 11a and Tables 4 and 5. It can be observed that the KLR model has the highest success rate of 0.847, followed by the NB and RBFNetwork models, which have success rates of 0.816 and 0.824, respectively. In addition, four related evaluation statistics, the ACC, the F-measure, the MAE, and the RMSE, are included. It could be observed that the KLR model has the highest ACC value (0.791), the highest F-measure (0.787), the lowest MAE (0.259), and the lowest RMSE (0.370), and these results also indicate that the KLR model exhibits the best performance on the training dataset.

The prediction rates of the three landslide models were evaluated using the validation data, and the resulting ROC

Table 5 Parameters of AUC for the training dataset

Test result variable (s)	Area	SE	95% Confidence interval	
			Lower bound	Upper bound
KLR	0.847	0.0174	0.812	0.878
NB	0.816	0.0197	0.778	0.850
RBFNetwork	0.824	0.0186	0.787	0.857

Table 6 Performance of the three models on the validation dataset

Parameters	KLR	NB	RBFNetwork
AUC	0.772	0.732	0.749
ACC (%)	0.766	0.673	0.737
F-measure	0.747	0.692	0.719
MAE	0.269	0.326	0.298
RMSE	0.408	0.506	0.409

curves and AUC values are shown in Fig. 11b and Tables 6 and 7. It can be seen that all the models have good prediction capabilities, while the KLR model has the highest prediction rate of 0.772, followed by the RBFNetwork and NB models, with prediction rates of 0.749 and 0.732, respectively. In addition, the four related evaluation statistics, the ACC, the F-measure, the MAE, and the RMSE, also show that the KLR model has the highest ACC value (0.766), the highest F-measure (0.747), the lowest MAE (0.269), and the lowest RMSE (0.408), which also indicate that the KLR model exhibits the best performance on the validation dataset.

Finally, to check the statistically significant differences between the three landslide models, the Wilcoxon signed-rank test method with *p* and *z* values was used for pairwise comparisons of the three models. When *p* values are smaller than the 0.05 significance level and *z* values exceed critical values of *z* (−1.96 to +1.96), the performances of landslide models are different (Tien Bui et al. 2016). The results show that the performances of the three models in this study are significantly different (Table 8).

Overall, all three landslide models are acceptable for landslide susceptibility mapping in the study area, and the KLR model exhibits the most stable and best performance in this study.

Discussions and conclusions

This study assesses the effectiveness of KLR, NB, and RBFNetwork models for landslide susceptibility modeling of the Long County area in China. To construct the landslide models, the concept of binary classification from

Table 7 Parameters of AUC for the validation dataset

Test result variable (s)	Area	SE	95% Confidence Interval	
			Lower bound	Upper bound
KLR	0.772	0.0357	0.708	0.827
NB	0.732	0.0365	0.666	0.792
RBFNetwork	0.749	0.0358	0.684	0.807

Table 8 Pairwise comparison for the three landslide susceptibility models using Wilcoxon signed-rank test (two-tailed)

NO	Pairwise comparison	Z value	P value	Significance
1	KLR vs. NB	-8.909	0.000	Yes
2	KLR vs. RBFNetwork	-4.660	0.000	Yes
3	NB vs. RBFNetwork	-10.008	0.000	Yes

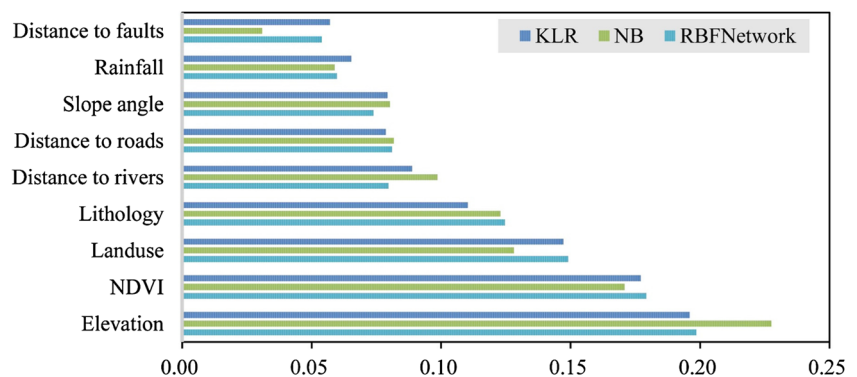
machine learning was adopted, in which the models were trained in order to classify the pixels examined in the study into two classes, landslide and non-landslide. In addition, in order to use these methods, it is necessary to create the training dataset and the validation dataset for the models, respectively (Tien Bui et al. 2017). Since the areas of the 171 landslides are very small compared to the total study area, the undersampling method (Pradhan 2013) was used to generate the same non-landslide points to avoid the problems associated with imbalanced distributions. This is the most widely used sampling method in landslide susceptibility; therefore, it may guarantee that the results from the three models are comparable with those in the literature. In addition, a total of 12 conditioning factors were selected based on analysis of the landslide inventory map, the characteristics of the study area, and the literature, though their predictive abilities are clearly different (Table 4). Slope aspect, plan curvature, and profile curvature revealed non-predictive ability values, and, therefore, they were excluded from this analysis. The distribution of landslides is relatively even across the classes of the above three factors, which resulted in low values for the chi-squared statistic. In contrast to those three factors, the other factors reveal predictive ability values that indicate that the selection and coding of these factors had been conducted successfully.

However, a standard guideline for the selection of landslide conditioning factors is still a topic of debate (Tien Bui et al. 2016). Thus, the classifier attribute evaluation method was further used to assess variable importance for the three models with the normalized predictive capabilities of

conditioning factors (Witten et al. 2011). It can be seen that, although all included 9 factors have positive predictive capabilities, they differ in contributions to different models (Fig. 12). In the case of the KLR model, elevation, NDVI, land use, and lithology have the highest contributions of 0.196, 0.177, 0.147, and 0.110, respectively. The other five conditioning factors have contributions less than 0.100. The contributions of 9 landslide conditioning factors are similar for the NB and RBFNetwork models. However, the contributions of elevation, NDVI, land use, and lithology are 0.228, 0.171, 0.128, and 0.123 for the NB model, while 0.199, 0.179, 0.149, and 0.125 for the RBFNetwork model, respectively. Therefore, it can be concluded that the predictive capability of a conditioning factor depends on the landslide model used (Tien Bui et al. 2016). Further studies are necessary to explore the selection methods of landslide conditioning factor.

The performance assessment shows that there is a difference between the three models in the classification of the training dataset. The classification accuracy of the KLR model is 79.1%, with an AUC of 0.847, indicating a good result (better than those of the other models). Regarding prediction capability, the NB model has the lowest capability compared to the KLR model and the RBFNetwork model. The KLR model with the RBF kernel has a better ability to adapt to the training data than the other models do. On the other hand, the performance of the NB model is influenced by its assumptions of conditional independence (Lee and Lee 2006), and these assumptions were violated in the training data of this study, resulting in the low classification rate. Compared to the NB model, the RBFNetwork performed better, due to its ability to better adjust the weights in the training phase. Based on the above analysis, we conclude that the KLR model is the best suited for this study.

The outcomes of the study indicate that these three models could be useful tools for producing landslide susceptibility maps and could provide valuable information to local government agencies for natural hazard management and policy planning for the study area and other similar areas.

Fig. 12 Relative importance of conditioning factors for the three models

Acknowledgements The authors would like to express their gratitude to the Editor-in-Chief Martin Gordon Culshaw and two anonymous reviewers for their helpful comments on the manuscript.

Funding This research was supported by Project funded by China Postdoctoral Science Foundation (Grant No. 2017 M613168), Project funded by Shaanxi Province Postdoctoral Science Foundation (Grant No. 2017BSHYDZZ07), Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 17JK0511), the Open Fund of Shandong Provincial Key Laboratory of Depositional Mineralization & Sedimentary Minerals (Grant No. DMSM2017029), and the Open-ended Fund of the Key Laboratory for Geo-hazard in Loess Areas, Ministry of Land and Resources of China (Grant No. KLGLAMLR201603).

References

- Aghdam IN, Varzandeh MHM, Pradhan B (2016) Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). *Environ Earth Sci* 75:1–20
- Althuwaynee OF, Pradhan B, Lee S (2016) A novel integrated model for assessing landslide susceptibility mapping using CHAID and AHP pair-wise comparison. *Int J Remote Sens* 37:1190–1209
- Andrews DW (1988) Chi-square diagnostic tests for econometric models: introduction and applications. *J Econ* 37:135–156
- Booth AM et al (2015) Integrating diverse geologic and geodetic observations to determine failure mechanisms and deformation rates across a large bedrock landslide complex: the Osmundneset landslide, Sogn og Fjordane, Norway. *Landslides* 12:745–756
- Carlini M et al (2016) Tectonic control on the development and distribution of large landslides in the northern Apennines (Italy). *Geomorphology* 253:425–437
- Cawley GC, Talbot NL (2008) Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Mach Learn* 71: 243–264
- Chen W, Panahi M, Pourghasemi HR (2017a) Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling. *Catena* 157:310–324
- Chen W et al (2017b) GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. *Geomat Nat Haz Risk* 8: 950–973
- Chen W et al (2017c) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* 151:147–160
- Chung C-JF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. *Nat Hazards* 30:451–472
- Colkesen I, Sahin EK, Kavzoglu T (2016) Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *J Afr Earth Sci* 118:53–64
- Conoscenti C et al (2015) Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: a case of the Belice River basin (western Sicily, Italy). *Geomorphology* 242:49–64
- Constantin M, Bednarik M, Jurchescu MC, Vlaicu M (2011) Landslide susceptibility assessment using the bivariate statistical analysis and the index of entropy in the Sibiciu Basin (Romania). *Environ Earth Sci* 63:397–406
- Cook TL, Yellen BC, Woodruff JD, Miller D (2015) Contrasting human versus climatic impacts on erosion. *Geophys Res Lett* 42: 6680–6687
- Dehnavi A, Aghdam IN, Pradhan B, Varzandeh MHM (2015) A new hybrid model using step-wise weight assessment ratio analysis (SWAM) technique and adaptive neuro-fuzzy inference system (ANFIS) for regional landslide hazard assessment in Iran. *Catena* 135:122–148
- Dormann CF et al (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36: 27–46
- Dou J et al. (2014) GIS-based landslide susceptibility mapping using a certainty factor model and its validation in the Chuetsu Area, Central Japan. In: Sassa K, Canuti P, Yin Y (eds) *Landslide Science for a Safer Geoenvironment*. Springer, Cham, pp 419–424
- Felicísimo ÁM, Cuartero A, Remondo J, Quirós E (2013) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides* 10:175–189
- Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. *Neural Comput* 4:1–58
- Gil D, Johnsson M (2010) Supervised SOM based architecture versus multilayer perceptron and RBF networks. *Proceedings of the Linköping Electronic Conference*, pp 15–24
- Gorsevski PV, Brown MK, Panter K, Onasch CM (2016) Landslide detection and susceptibility mapping using LiDAR and an artificial neural network approach: a case study in the Cuyahoga Valley National Park, Ohio. *Landslides* 13:467–484
- Guzzetti F, Reichenbach P, Ardizzone F, Cardinali M, Galli M (2006) Estimating the quality of landslide susceptibility models. *Geomorphology* 81:166–184
- Hong H, Pradhan B, Xu C, Tien Bui D (2015a) Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. *Catena* 133:266–281
- Hong H, Xu C, Revhaug I, Tien Bui D (2015b) Spatial prediction of landslide hazard at the Yihuang area (China): a comparative study on the predictive ability of backpropagation multi-layer perceptron neural networks and radial basic function neural networks. In: Robbi Sluter C, Madureira Cruz CB, Leal de Menezes PM (eds) *Cartography – Maps Connecting the World*. Springer, Cham, pp 175–188
- Jaafari A, Najafi A, Pourghasemi H, Rezaeian J, Sattarian A (2014) GIS-based frequency ratio and index of entropy models for landslide susceptibility assessment in the Caspian forest, northern Iran. *Int J Environ Sci Technol* 11:909–926
- Kim T, Chung BD, Lee JS (2016) Incorporating receiver operating characteristics into naive Bayes for unbalanced data classification. *Computing* 99:1–16
- Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. *J Math Anal Appl* 33:82–95
- Kumar R, Anbalagan R (2015) Landslide susceptibility zonation in part of Tehri reservoir region using frequency ratio, fuzzy logic and GIS. *J Earth Syst Sci* 124:431–448
- Kumar R, Anbalagan R (2016) Landslide susceptibility mapping using analytical hierarchy process (AHP) in Tehri reservoir rim region, Uttarakhand. *J Geol Soc India* 87:271–286
- Lee C, Lee GG (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. *Inf Process Manag* 42:155–165
- Lineback Gritzner M, Marcus WA, Aspinnall R, Custer SG (2001) Assessing landslide potential using GIS, soil wetness modeling and topographic attributes, Payette River, Idaho. *Geomorphology* 37:149–165

- Mercer J (1909) Functions of positive and negative type, and their connection with the theory of integral equations. *Philos Trans R Soc Lond A* 209:415–446
- Mohammady M, Pourghasemi HR, Pradhan B (2012) Landslide susceptibility mapping at Golestan Province, Iran: a comparison between frequency ratio, Dempster–Shafer, and weights-of-evidence models. *J Asian Earth Sci* 61:221–236
- Park S, Choi C, Kim B, Kim J (2013) Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area, Korea. *Environ Earth Sci* 68:1443–1464
- Peng JB et al (2015) Heavy rainfall triggered loess-mudstone landslide and subsequent debris flow in Tianshui, China. *Eng Geol* 186:79–90
- Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). *Environ Model Softw* 84:240–250
- Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia M (2017) Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theor Appl Climatol* 128:255–273
- Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB (2015) Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theor Appl Climatol* 1–19
- Pradhan B (2010) Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. *J Indian Soc Remote Sens* 38:301–320
- Pradhan B (2013) 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* 51:350–365
- Pradhan B, Abokharima MH, Jebur MN, Tehrani MS (2014) Land subsidence susceptibility mapping at Kinta Valley (Malaysia) using the evidential belief function model in GIS. *Nat Hazards* 73:1019–1042
- Press SJ (1966) Linear combinations of non-central chi-square variates. *Ann Math Stat* 480–487
- Rao J, Scott A (1987) On simple adjustments to chi-square tests with sample survey data. *Ann Stat* 385–397
- Razandi Y, Pourghasemi HR, Neisani NS, Rahmati O (2015) Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Sci Inf* 8:867–883
- Razavizadeh S, Solaimani K, Massironi M, Kavian A (2017) Mapping landslide susceptibility with frequency ratio, statistical index, and weights of evidence models: a case study in northern Iran. *Environ Earth Sci* 76:499
- Regmi AD et al (2014) Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. *Arab J Geosci* 7:725–742
- Regmi NR, Giardino JR, Vitek JD (2010) Modeling susceptibility to landslides using the weight of evidence approach: western Colorado, USA. *Geomorphology* 115:172–187
- Sar N, Khan A, Chatterjee S, Das A, Mipun BS (2016) Coupling of analytical hierarchy process and frequency ratio based spatial prediction of soil erosion susceptibility in Keleghai river basin, India. *International Soil and Water Conservation Research*
- Satorra A, Bentler PM (2001) A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66:507–514
- Schölkopf B, Smola AJ (2002) *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge
- Shahabi H, Hashim M, Ahmad BB (2015) Remote sensing and GIS-based landslide susceptibility mapping using frequency ratio, logistic regression, and fuzzy logic methods at the central Zab basin, Iran. *Environ Earth Sci* 73:8647–8668
- Soria D, Garibaldi JM, Ambrogi F, Biganzoli EM, Ellis IO (2011) A ‘non-parametric’ version of the naïve Bayes classifier. *Knowl-Based Syst* 24:775–784
- Tien Bui D, Nguyen QP, Hoang N-D, Klempe H (2017) A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. *Landslides* 14:1–17
- Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012) Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and Naive Bayes Models. *Math Probl Eng* 2012
- Tien Bui D, Pradhan B, Revhaug I, Tran CT (2014) A comparative assessment between the application of fuzzy unordered rules induction algorithm and J48 decision tree models in spatial prediction of shallow landslides at Lang Son City, Vietnam. In: Srivastava PK, Mukherjee S, Gupta M, Islam T (eds) *Remote Sensing Applications in Environmental Research*. Springer, New York, pp 87–111
- Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I (2016) 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* 13:361–378
- Trigila A, Iadanza C, Esposito C, Scarascia-Mugnozza G (2015) Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* 249:119–136
- Tsangaratos P, Ilia I (2016) Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. *Catena* 145:164–179
- Van Westen C (2004) Geo-information tools for landslide risk assessment: an overview of recent developments, *Proceedings 9th International Symposium on Landslides*. Balkema, Amsterdam, pp 39–56
- Walter S (2002) Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 21:1237–1256
- Wang L-J, Guo M, Sawada K, Lin J, Zhang J (2016) A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. *Geosci J* 20:117–136
- Witten IH, Frank E, Mark AH (2011) *Data mining: practical machine learning tools and techniques*. 3rd edn. Morgan Kaufmann, Burlington
- Wu X et al (2008) Top 10 algorithms in data mining. *Knowl Inf Syst* 14:1–37
- Wu YM, Lan HX, Gao X, Li LP, Yang ZH (2015) A simplified physically based coupled rainfall threshold model for triggering landslides. *Eng Geol* 195:63–69
- Youssef AM, Pourghasemi HR, El-Haddad BA, Dhahry BK (2016) Landslide susceptibility maps using different probabilistic and bivariate statistical models and comparison of their performance at Wadi Itwad Basin, Asir region, Saudi Arabia. *Bull Eng Geol Environ* 75:63–87
- Zhang G et al (2016) Integration of the statistical index method and the analytic hierarchy process technique for the assessment of landslide susceptibility in Huizhou, China. *Catena* 142:233–244
- Zhang M, Yin YP, Huang BL (2015) Mechanisms of rainfall-induced landslides in gently inclined red beds in the eastern Sichuan Basin, SW China. *Landslides* 12:973–983