


RESEARCH LETTER

Open Access



Evaluation of the landslide susceptibility and its spatial difference in the whole Qinghai-Tibetan Plateau region by five learning algorithms

Payam Sajadi¹, Yan-Fang Sang^{1,2*} , Mehdi Gholamnia³, Stefania Bonafoni⁴ and Saumitra Mukherjee⁵

Abstract

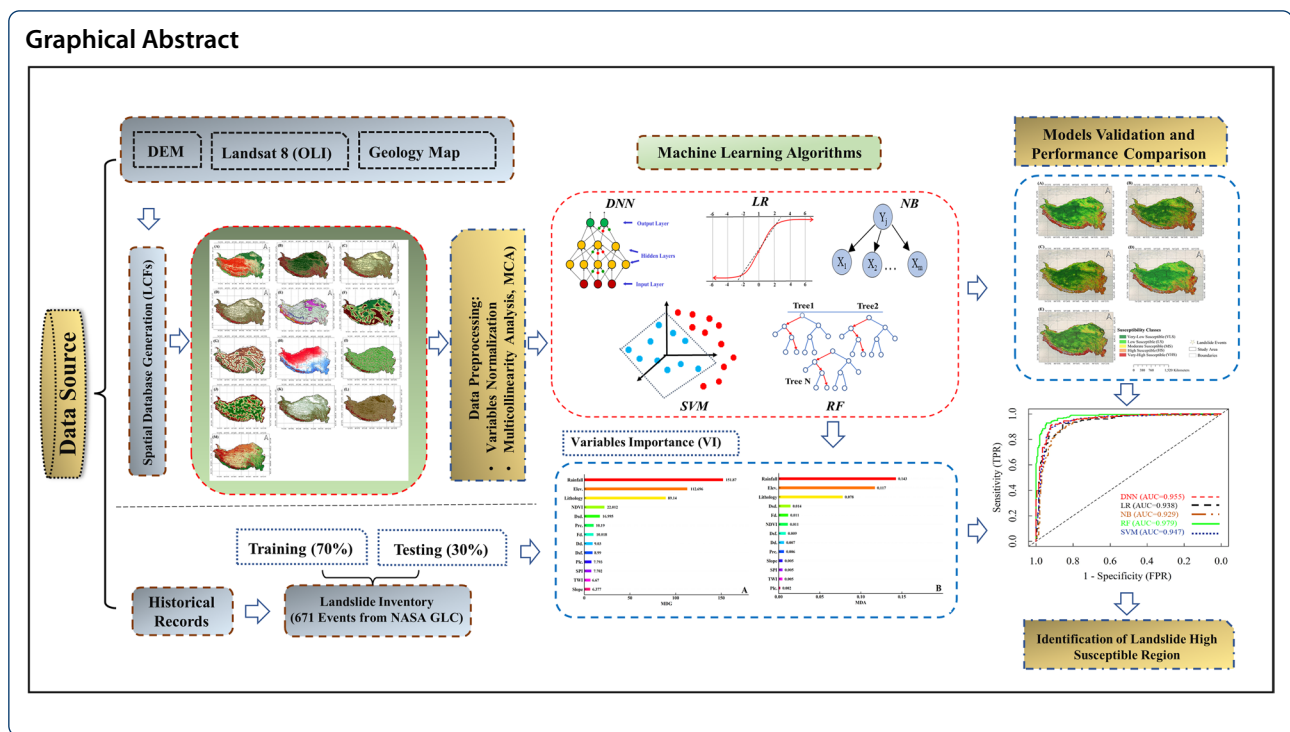
Landslides are considered as major natural hazards that cause enormous property damages and fatalities in Qinghai-Tibetan Plateau (QTP). In this article, we evaluated the landslide susceptibility, and its spatial differencing in the whole Qinghai-Tibetan Plateau region using five state-of-the-art learning algorithms; deep neural network (DNN), logistic regression (LR), Naïve Bayes (NB), random forest (RF), and support vector machine (SVM), differing from previous studies only in local areas of QTP. The 671 landslide events were considered, and thirteen landslide conditioning factors (LCFs) were derived for database generation, including annual rainfall, distance to drainage (D_{S_d}), distance to faults (D_{S_f}), drainage density (D_d), elevation (Elev), fault density (F_d), lithology, normalized difference vegetation index (NDVI), plan curvature (Pl_c), profile curvature (Pr_c), slope (S°), stream power index (SPI), and topographic wetness index (TWI). The multi-collinearity analysis and mean decrease Gini (MDG) were used to assess the suitability and predictability of these factors. Consequently, five landslide susceptibility prediction (LSP) maps were generated and validated using accuracy, area under the receiver operating characteristic curve, sensitivity, and specificity. The MDG results demonstrated that the rainfall, elevation, and lithology were the most significant landslide conditioning factors ruling the occurrence of landslides in Qinghai-Tibetan Plateau. The LSP maps depicted that the north-northwestern and south-southeastern regions (< 32% of total area) were at a higher risk to landslide compared to the center, west, and northwest of the area (> 45% of total area). Moreover, among the five models with a high goodness-of-fit, RF model was highlighted as the superior one, by which higher accuracy of landslide susceptibility assessment and better prone areas management in QTP can be achieved compared to previous results.

Keywords: Feature selection technique, Landslide susceptibility, Machine learning algorithm, Spatial differencing, Qinghai-Tibetan Plateau

*Correspondence: sangyf@igsnr.ac.cn; sunsangyf@gmail.com

¹ Key Laboratory of Water Cycle & Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

Full list of author information is available at the end of the article



Introduction

Landslide is a manifestation of the downslope movement of rocks, debris, and soil materials under the action of gravity force (Varnes 1984), representing pervasive and frequent natural threats in mountainous regions worldwide. Landslides are responsible for fatalities, significant damages to infrastructures, and economic losses at global scales (Dai et al. 2002; Jia et al. 2020). They are the products of complex synergetic interactions between various intrinsic, extrinsic and anthropogenic agents, known as landslide conditioning factors (Hutchinson 1995; Guzzetti et al. 1999). The combination of LCFs with human activities and climate change increases the frequencies and sizes of landslides worldwide. For instance, in China, various regions are suffering from landslide disasters, as such geologic hazards after an earthquake are the second most destructive natural hazards, causing economic losses of over 20 billion Yuan (CNY) every year (Hong et al. 2016). Therefore, landslide hazards studies are critical for early landslide predictions to reduce landslide-related disasters.

In this context, landslide susceptibility prediction has proven to be a fundamental and effective tool for predicting the spatial occurrence of landslide hazards in a susceptible region (Chang et al. 2020). LSP estimates the degree of vulnerability of a specific region to landslides considering both intrinsic (geology, topography,

geomorphology, etc.) and extrinsic (e.g., seismic activity, volcanos, and rainfall) factors (Guzzetti 2006; Wu et al. 2014). The performance of landslide susceptibility prediction is highly affected by the input reliability and implementation models (Tien Bui et al. 2016b).

To address this, various approaches with different sets of assumptions and procedures were developed to improve LSP accuracy, including heuristic, physically based, and traditional-statistical methods (Huang et al. 2020). The ranking/rating-based heuristic methods include expert knowledge systems (Zhu et al. 2014) such as analytical hierarchy process (Pawluszek and Borkowski 2017), and the gray relational method (Huang et al. 2019). Slope stability and mechanic laws are key concepts of physically based models that are not suitable for regional scales (Crosta et al. 2003) because of costly and time-consuming issues. Traditional-statistical methods have received much attention especially for the landslide susceptibility studies at large scales. These methods mainly include information value (Pasang and Kubiček 2020), weights of evidence (Mersha and Meten, 2020), frequency ratio (Mersha and Meten 2020), fractal theory (Hu et al. 2020), Dempster–Shafer (Tangestani 2009), and certainty factors (Wubalem and Meten 2020). Meanwhile, the reliability and accuracy of these approaches cannot meet the practical needs enough due to their inability to appraise the complex non-linear inter-relationship between the factors (Tien Bui et al. 2016a).

Recently, the emergence of machine learning brought new insights into landslide modeling, such as logistic regression (Yi et al. 2020), random forest (Dou et al. 2019), naïve bayes (Nhu et al. 2020b), support vector machine (Huang et al. 2020), maximum entropy (Kornejady et al. 2017), boosted regression trees (Song et al. 2020), artificial neural network (Sameen et al., 2020), deep neural network (Fang et al. 2020), and some hybrid models (Wang et al. 2020a). Generally, any machine learning model has its specific assumptions and applicable conditions, and thus cannot meet the needs of LSP in all situations (Reichenbach et al. 2018). Consequently, evaluating the performance of different learning algorithms is essential to clarify their strengths and weaknesses, and ultimately to identify the most efficient model in a particular area, because LSP with higher spatial accuracy (even 1% or 2%) has a great impact on identifying the spatial distribution of landslide-prone region (Jebur et al. 2014).

A review of the literature highlights the necessity of performance comparison of various models based on physiographic–topographic conditions of a certain region, to ascertain the most suitable model (Mahdadi et al. 2018; Xiao et al. 2018; Di Napoli et al. 2020; Yu and Chen 2020; Saha et al. 2021; Youssef and Pourghasemi 2021). As the probabilistic distribution functions in the models are different, their success and capabilities in prediction are different. It reflects that each model has its flexibility and specially applicability because of uncertainties of inputs and model selection. Therefore, the models should be tested and evaluated in the landslide-prone regions to acquire the most robust one for better management of these areas before landsliding in the future. In this context, Qinghai-Tibetan Plateau and its surrounding regions, especially Himalayas and its eastern areas, is highly affected by tectonic activities and faces numerous landslides and geo-hazards, mainly due to the presence of several active faults, geological structure, geomorphological evolution, and climatic effects (Deng et al. 2017; Yao et al. 2019; Qi et al. 2021; Zhao et al. 2021). Although a few studies compared the performances of different ML models for landslide susceptibility prediction in different local areas of QTP (Kumar et al. 2017; Pham et al. 2017b; Du et al. 2019; Peethambaran et al. 2020), these studies cannot reflect the very complex topographic structure of the entire QTP area, which is still unclear for the entire region. In addition, the most recent landslide sensitivity prediction studies are limited to performance comparison of specific machine learning algorithms such as DNN, RF, and SVM with non-complex surface topography compared to the entire QTP region (Al-Najjar and Pradhan 2021; Liu et al. 2021; Mandal et al. 2021; Wang et al. 2021; Youssef and Pourghasemi 2021). Therefore, a

comprehensive performance comparison between various advanced machine learning algorithms is essential to identify and illuminate the best learning algorithm for landslide susceptibility prediction and early landslide predictions across the entire QTP region, and to reduce the landslide-related disasters in the highly susceptible locations of the region.

To fill this gap, this study aims explicitly at performance and robustness comparison of five advanced and sophisticated benchmarks machine learning algorithms including logistic regression, deep neural network, support vector machine, naïve Bayes, and deep neural network for landslide susceptibility prediction, and to infer the most appropriate LSP model with highest predictive power in the entire Qinghai-Tibetan Plateau region. Besides, we not only validate the landslide susceptibility prediction maps by some performance metrics, but also assess the performance of the ML models along with the previous models performed in the QTP and adjacent areas to understand the future of landsliding situations. The abbreviations and acronyms used in this paper are summarized in Table 1.

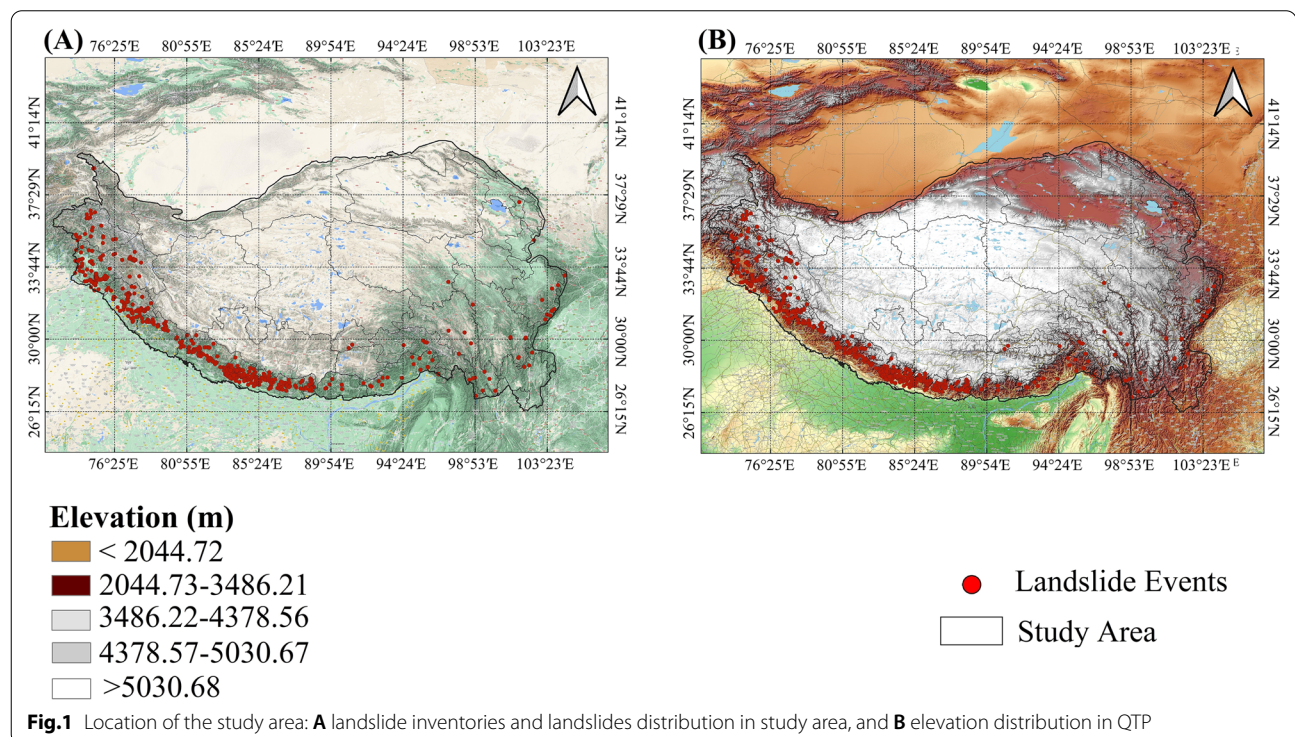
Study area and materials

Study area

The study area is located in the transition between in China, and its surrounding region (Himalayan regions), with a total area of about 3,038,856.96 km² (geographical coordination between 74° and 104° E, 25° and 40° N, see Fig. 1). The QTP is the highest (average altitude > 4000 m ASL) and largest plateau in the world, with unique topography (Huang et al. 2008). The altitude range in the study area varies from 100 to 8086 m, which is characterized by plains, valleys, and mountains. The surface topography changes from very gentle to highly steeped topography (0–78°). From the geological point of view, QTP is the product of the collision of the Eurasian and Indian plates (Molnar and Tapponnier 1975; Aiken and Brierley 2013). It is influenced by tectonic activity from the Himalayan region, resulting in the deformation and formation of complex structural features (Bartarya et al. 1996). Under the influence of major and minor thrust-faults, strike-faults, such as Himalayan main thrust fault (MF), Altyn-Tagh fault, Kunlun fault, Karakoram fault, Jiali fault, etc. (Taylor and Yin 2009; Elliott et al. 2010; Aiken and Brierley 2013; Zhang et al. 2020; Qi et al. 2021), the area is highly susceptible to significant geo-environmental hazards (landslide, earthquake, etc.). From the existing geological map of the study area, about 28 lithological structures were identified (see Fig. 2E). The most prominent geological formation consists of Mesoproterozoic crystalline metamorphic rocks and Mesozoic sedimentary rocks.

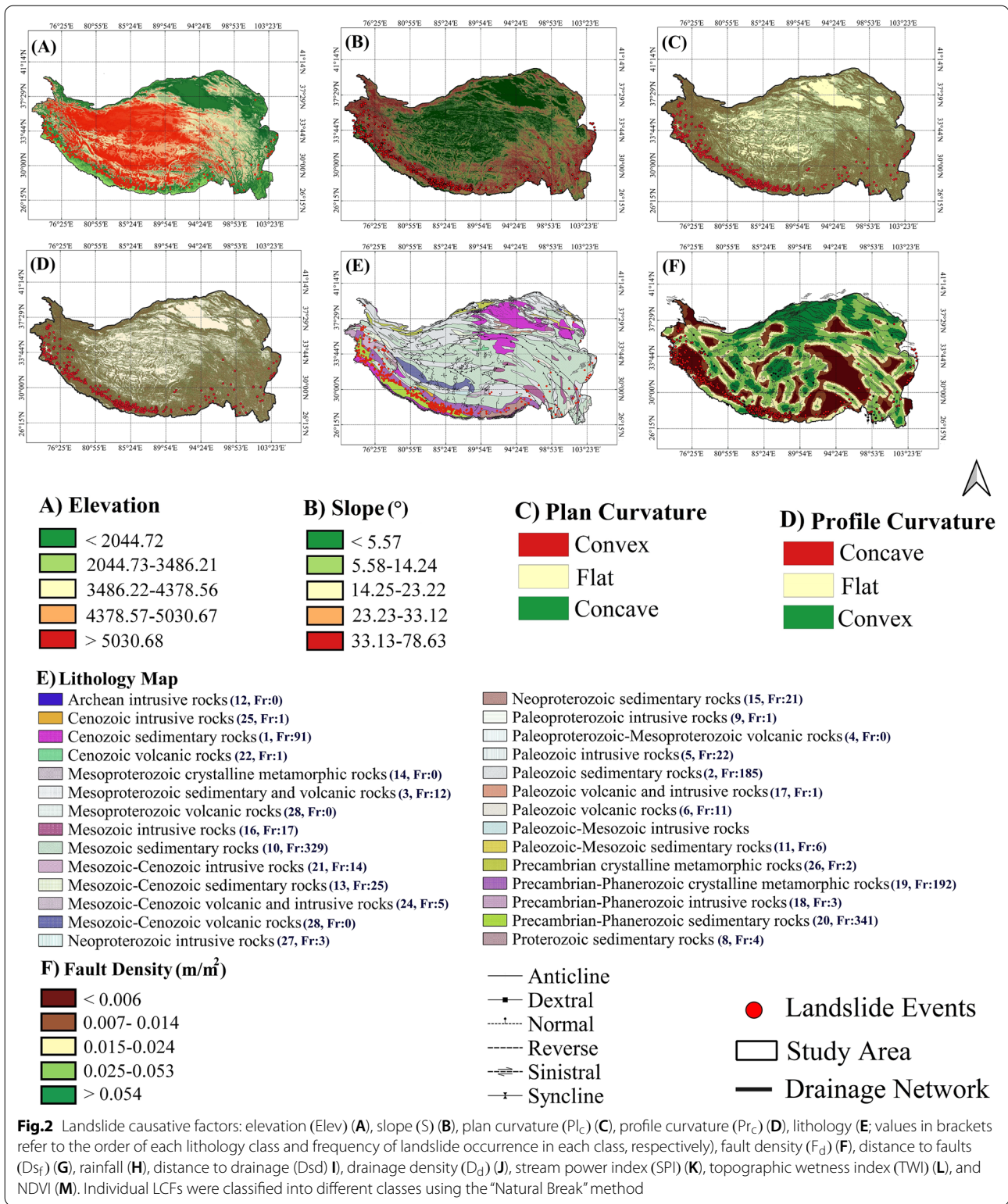
Table 1. List of abbreviations and their meanings mentioned in this study.

Abbreviations	Meaning	Abbreviations	Meaning
Acc	Accuracy	NB	Naïve Bayes
ANN	Artificial neural networks	OOB	Out-Of-Bag
AUC	Area under curve	Pl _c	Plan curvature
D _d	Drainage density	Pr _c	Profile curvature
DNN	Deep neural network	QTP	Qinghai-Tibetan Plateau
Ds _d	Distance to drainage	RBF	Radial basis function
Ds _f	Distance to fault	ReLU	Rectified linear unit
Elev	Elevation	ROC	Receiver operating curve
F _d	Fault density	S°	Slope
GPM	global precipitation measurement	SST	Sensitivity
HS	High susceptible	SPF	Specificity
LR	logistic regression	SGD	Stochastic gradient descent
LSP	Landslide susceptibility prediction	SVM	Support vector machine
NDVI	Normalized difference vegetation index	NPV	Negative predictive value
LS	Low susceptible	TOL	Tolerance
MCA	multi-collinearity analysis	TPI	Topographic wetness index
MDA	Mean decrease in accuracy	VIF	Variance inflation factor
MDG	Mean decrease Gini	VI	Variable importance (VI)
ML	Machine learning	VLS	Very-low susceptible
MS	Moderate susceptible (MS)		



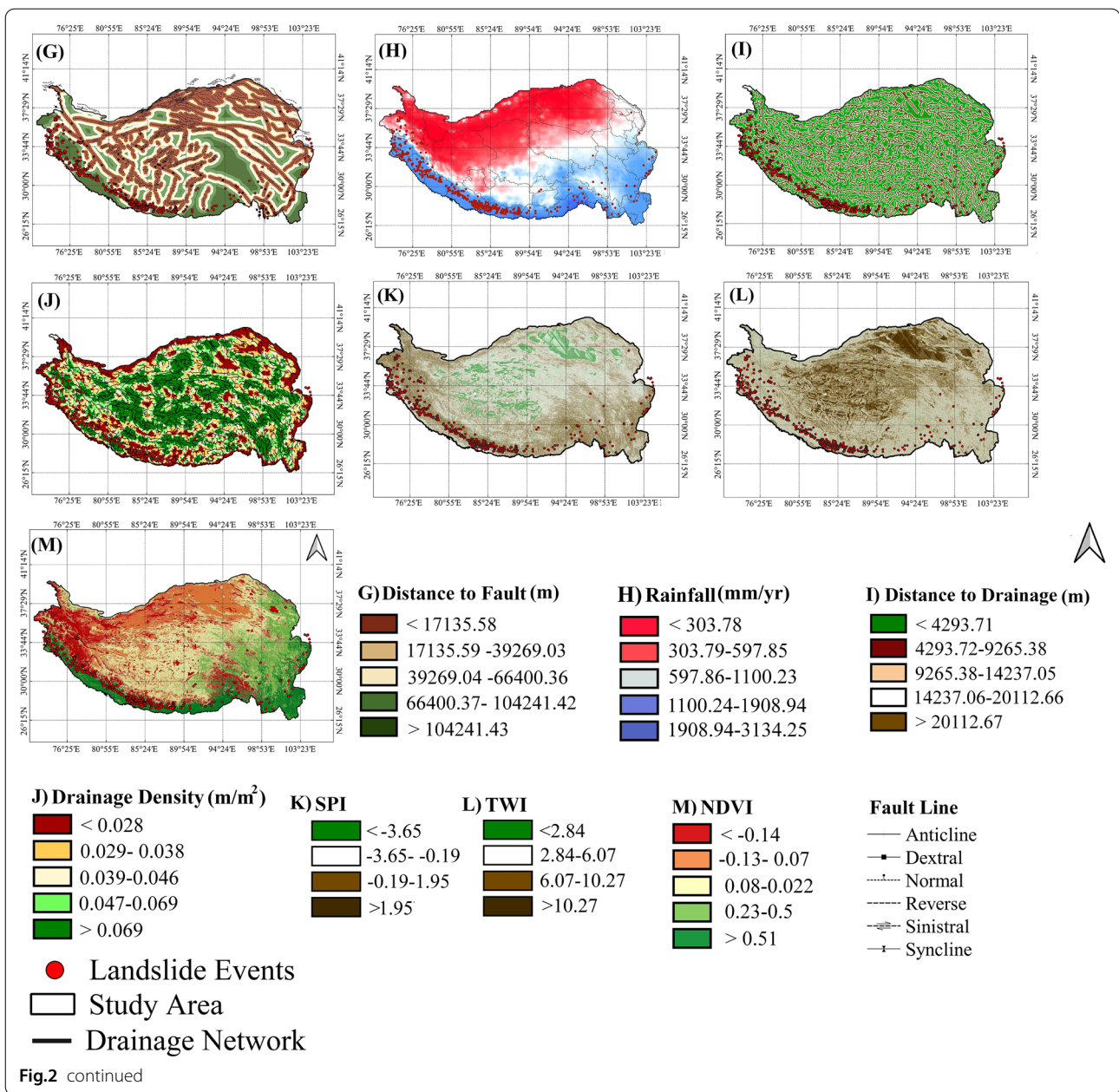
Due to its structural complexity, the QTP is characterized by low temperatures ranging from -15° to 10° $^{\circ}$ C and low precipitation, with cold and arid climatic conditions

in winter controlled by Siberian high and Mongolian high semi-arid climate in summer controlled by the South Asian monsoon system (Du et al. 2004; You et al. 2013).



The vegetation cover in QTP mainly includes steppe, shrub, desert, meadow, forest, barren areas (bare soil), and water bodies (ice and glacier) (Gillespie et al. 2019).

The QTP is the source of several major rivers, including the Yangtze River, the Yellow River, and the Ganges River, making it the “Water Tower of Asia” (Yao et al. 2019).



Furthermore, the study area covers a portion of Kosii, and Yarlung Tsangpo (Brahmaputra) transboundary rivers draining from the northern Himalayan slope into the QTP (Mukherjee 2008; Huang et al. 2011).

Spatial database construction

The landslide inventory generation is an essential step in any landslide susceptibility prediction modeling (Guzzetti et al. 1999), as it provides valuable information on different aspects of landslide events in the susceptible region (Rosi et al. 2018). Landslide inventories are based on a

key assumption that future landslides will occur under the same circumstances (same factors) that caused the previous landslides (Guzzetti et al. 2005). In this study, the landslide inventory was obtained from NASA Global Landslide Catalog (GLC) with the best resolution at 0.2° (Kirschbaum et al. 2010; Lin et al. 2017). The NASA-GLC contains information about landslide locations and characteristics with different triggering sources (chiefly rainfall, earthquake, downpour, snowfall-snow melt, etc.) from 2007 to 2016 (Stanley et al. 2020) obtained from numerous datacentres, including the International

Consortium on Landslides; International Landslide Centre, University of Durham; EM-DAT International Disaster Database; International Federation of Red Cross and Red Crescent Societies field reports; Reliefweb; humanitarian disaster information run by the United Nations Office for the Coordination of Humanitarian Affairs; and other online regional and national newspaper articles and media sources. A total of 671 landslide events obtained from GLC were confirmed in the study area (landslide, mudslide, rockfalls, debris flows, etc.). Also, a total of 671 non-landslide locations are randomly selected in non-landslide areas with an equal number of landslide events (Costanzo et al. 2014; Lin et al. 2017). The datasets were then randomly divided into two parts: 70% of the dataset was used for training, and the remaining 30% were used for test.

Landslide conditioning factors

The selection of influential landslide conditioning factors is vital for LSP modeling because the nature and evolution of landslide events are complex and require prior knowledge of significant LCFs in a particular region (Guzzetti et al. 1999; Chen et al. 2017; Wang et al. 2020a). Based on the study of landslide mechanisms, existing LSP-related literature, and data availability (Lin et al. 2017; Reichenbach et al. 2018; Abbaszadeh Shahri et al. 2019), 13 landslide conditioning factors were selected in this study and were classified into four main categories: surface topography, geologic, hydrologic, and landcover factors.

Surface topography factors include elevation (m) (Fig. 2A), slope (degree) (Fig. 2B), plan curvature (Fig. 2C), and profile curvature (Fig. 2D). They were derived from SRTM-DEM (90 m) (considering the scale of the analysis) available in the Google Earth Engine platform (CGIAR/SRTM90_V4). The elevation (m) is very sensitive to several geomorphological and geological processes that cause slope instability (Hu et al. 2020). Slope (degree) plays an indirect role in slope instability (Peethambaran et al. 2020) and has been used in many landslide studies (Aghdam et al. 2017; Pandey et al. 2019). Curvature is influenced by a change in altitude as the index of basin relief (Sajadi et al. 2021). The values of Pl_c and Pr_c represent concavity (negative values) /convexity (positive values) and flow velocity, respectively (Bordoni et al. 2020). Profile curvature and plan curvature with values close to 0 in both cases indicate a flat topography (Kornejady et al. 2017).

Geological factors include lithology (Fig. 2E), fault density (m/m^2) (Fig. 2F), and distance to fault (m) (Fig. 2G). Lithology is an important factor and provides valuable information on the degree of hardness, mineral composition, and associated bedrock structure (Ercanoglu 2005).

It has been widely used in various landslide hazard studies (Abbaszadeh Shahri et al. 2019; Arabameri et al. 2020; Peethambaran et al. 2020). The lithology map was extracted from the geologic map at a scale of 1:5,000,000 (Steinsouer et al. 1999). Structural lineaments, especially fault lines, are a type of discontinuity in the slope that increases the probability of slope failure and affects the magnitude and distribution of landslide events (Dou et al. 2015; Pham et al. 2016a). F_d , and Ds_f maps were generated by the Euclidean distance analysis and line density methods (Süzen 2002; Bordoni et al. 2020; Sajadi et al. 2020).

Annual mean rainfall (mm/year) (Fig. 2H), distance to drainage (m) (Fig. 2I), drainage density (m/m^2) (Fig. 2J), stream power index (SPI) (Fig. 2K), and topographic wetness index (Fig. 2L) were considered as the hydrological factors. Rainfall is an essential factor inducing slope failure in a particular area that does not have a uniform distribution in the region (Hu et al. 2020). The infiltration and liquefaction from rainfall reduce the suction rate of materials, lose the shear strength between soil material, and increase landslides probability (Pham et al. 2017a). The annual precipitation data from the gauged-adjusted version of the Integrated Multi-Satellite Retrievals for Global Precipitation Measurement product was used to produce the rainfall map of the study area. As one of the most recent precipitation datasets, this dataset provides precipitation data with high spatiotemporal resolution (30 min|0.1°) worldwide. Previous studies revealed the considerable potential of GPM precipitation in landslides studies (Kirschbaum and Stanley 2018; Thakur et al. 2020). Ds_d is another key factor describing the hydrological conditions of a given area that affect slope stability (Huang et al. 2020). The Ds_d map was produced using the Euclidean distance analysis method. The D_d was generated from the line density method to illustrate the spatial distribution of the drainage network in the study area (Sajadi et al. 2020). SPI has been widely used as a decisive hydrological factor in LSP studies because the erosive power of runoff directly affects slope toe erosion (Jebur et al. 2014). TWI measures the role of topography on the flow direction, indicating the slope condition, and determines the hydrological processes involved in surface runoff generation (Jebur et al. 2014; Sameen et al. 2020). SPI and TWI were derived from the specific catchment area (A_s) in meter by the slope ($\tan\beta$) as in the following equations:

$$TWI = \ln\left(\frac{A_s}{\tan\beta}\right), \quad (1)$$

$$SPI = \ln(A_s * \tan\beta). \quad (2)$$

NDVI as an index of vegetation distribution, plays a significant role in slope stability (Borga 2019) and depicts

the relationship between landslide events and vegetation cover (Choi et al. 2012). The annual average NDVI map was calculated from cloud-free Landsat OLI imagery using the GEE platform (LANDSAT/LC08/C01/T1_SR) (Fig. 2M). Since different landslide conditioning factors are produced in different scales and sizes, they were resampled into the corresponding original SRTM-DEM pixel grid (90*90 m) using nearest-neighbor interpolation techniques as a standard method to preserve the original characteristics of the dataset (Jamal and Mandal 2016).

Methodology

The methodological hierarchy in this study is based on four major steps: (1) spatial database construction as explained in “Spatial database construction” Section and “Landslide conditioning factors” Section; (2) data pre-processing (data normalization, and MCA); (3) landslide prediction modeling using five state-of-the-art machine learning algorithms, and LSP maps generation; and (4) model validation, performance comparison, and identification of the best model performance. The flowchart of the developed methodology is illustrated in Fig. 3. More details are explained below.

Data pre-processing

Before implementing machine learning algorithms, it is necessary to normalize (scale) all landslide conditioning factors to reduce the data dispersion and inconsistency (Ercanoglu 2005; Wang et al. 2020b). Because different variables (LCFs) have different ranges and types, it is necessary to scale all variables into a similar range to avoid any inconsistency among the variables. The normalization of

all variables was conducted considering the nature of the input variable using the following equation:

$$Z_{LCF} = \left(\frac{LCF_i - LCF_{min}}{LCF_{max} - LCF_{min}} \right), \tag{3}$$

where Z_{LCF} is the normalized value of LCF, LCF_i is the original variable and LCF_{max} and LCF_{min} is the minimum and maximum value of LCF_i , respectively (Pradhan and Lee 2010; Zare et al. 2013). In the next step, a multi-collinearity analysis (MCA) was performed to evaluate the collinearity rate among variables and avoid bias in the spatial differences between models (Arora et al. 2019). Variance inflation factor (VIF) and tolerance (TOL) are two important collinearity criteria. In general, $TOL < 0.1$ and $VIF > 10$ indicate high collinearity issues in the dataset (Dormann et al. 2013). In this study, MCA was performed using the “imcdiag” function available in the R package “mctest” (Ullah et al. 2019).

Finally, the capability of the five learning algorithms to predict the landslide susceptibility in the study area was evaluated according to their different principles and structures under similar condition (all LCFs are included in the modeling).

Landslide susceptibility prediction modeling and LSP maps generation

Deep neural network (structure, loss function, optimization and model implementation)

In essence, artificial neural networks are generic non-linear functions that resemble the human brain neural

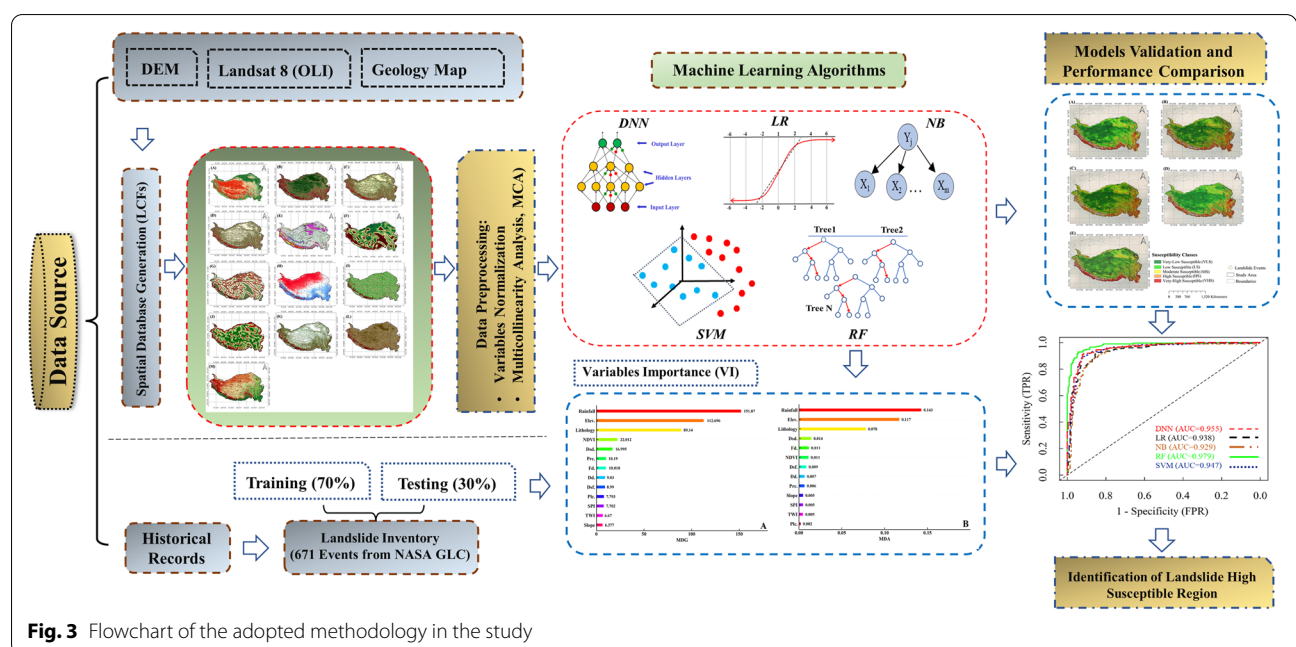


Fig. 3 Flowchart of the adopted methodology in the study

system (Zhu et al. 2018). The main advantage of ANN is the ability to handle all types of input data including binary, categorical, and continuous, which does not depend on the normal distribution of the input dataset (Kavzoglu and Mather 2003). A typical and widespread example of the ANN is the multi-layer feedforward neural network, which consists of an input layer, a hidden (learning phase), and an output layer (prediction) (Bishop 1995).

A neural network with a considerable number of hidden layers (depths) is known as deep learning (Nhu et al. 2020a). The number of neurons in the input and output layer is based on the application, while hidden layer neurons are often determined by trial and error (Peethambaran et al. 2020). Because rectified linear unit [Eq. (3)] benefits from the gradient descent algorithm for error (loss function) minimization, it can eliminate gradient vanishing phenomena and simplify the learning process (Singaravel et al. 2018; Nhu et al. 2020a). Hence, it is chosen as the activation function in the hidden layers of the proposed DNN for mapping the non-linear relationship between input and output.

$$\text{RELU}(\mathcal{E}) = \max(\mathcal{E}, 0), \quad (4)$$

where \mathcal{E} is the input variables (LCFs).

The sigmoid function is another popular activation function used as the transfer function to map the non-linear relationship in the output layer neurons for binary prediction (Bui et al. 2020). The network performance and convergence achievement were assessed using Log Loss for the landslide classification problem. In this study, Log Loss can be expressed as:

$$\text{Log}_{\text{Loss}}(L, \hat{L}) = \frac{-\sum_{i=1}^N \left(\log(\hat{L}_i)^{L_i} + \log(1 - \hat{L}_i)^{(1-L_i)} \right)}{N}, \quad (5)$$

where N is the number of samples, L_i is the actual output of sample i (landslide, non-landslide; 0, 1), \hat{L}_i is the predicted probability of sample i , L and \hat{L} is the vector of actual and predicted probabilities. The stochastic gradient descent optimization algorithm with optimal learning rate was used for weight adjustments in the hidden layer and to minimize the loss function (Nhu et al. 2020a).

In this study, 13 hidden neurons were found as optimal numbers based on several iterations (trial and error). A lower learning rate may increase the number of training epochs, and a higher learning rate is helpful to avoid local minimum. Because the mean and variance may change during the training and learning phase, the batch normalization and dropout models were configured to the DNN

structure to normalize data adaptively. This helps in regularization and improves the generalization capacity of the model to reduce the overfitting problem and accelerate the learning phase (Ioffe and Szegedy 2015; Carranza-García et al. 2019). A schematic of the proposed structure DNN development in this study is illustrated in Fig. 4. “Keras package” (<https://keras.rstudio.com>) was used to construct densely (deeply) connected neural networks (DNN) algorithm for LSP model in QTP region.

Logistic regression

A multivariate statistical model (generalized linear regression) based on non-linear function has been widely used for landslide susceptibility prediction in different regions (Park et al. 2013; Costanzo et al. 2014; Budimir et al. 2015). In contrast to typical linear regression, the dependent variables (prediction classes) can take categorical variables (landslide or non-landslide; 1, 0), the independent variables can be categorical, continuous, and binary (Atkinson and Massari 2011).

Besides, the model does not require the normal distribution assumption of independent variables (Wubalem 2020). The primary objective of LR is to model the linear relationship between the log odds (logit) of dependent variables and independent variables. For a binary response variable (landslide or non-landslide; 1, 0), this linear relationship can be shown as:

$$\text{logit}(L) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n + e, \quad (6)$$

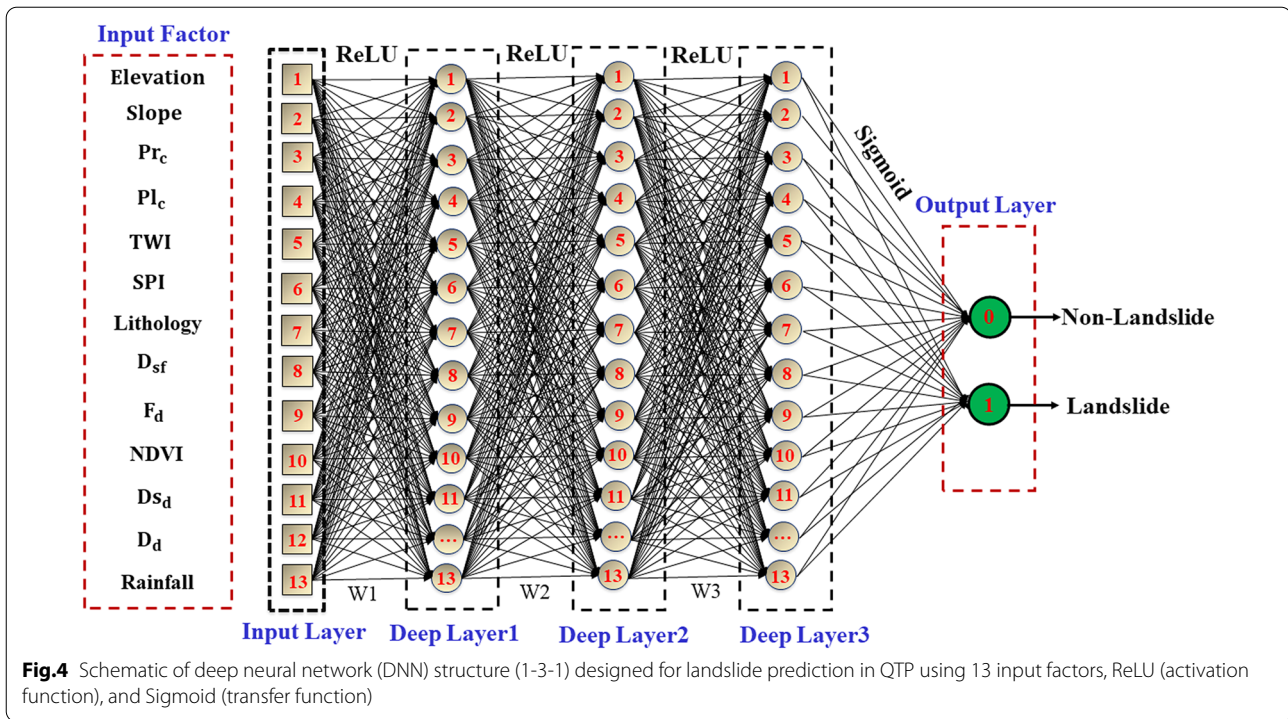
where L is the dependent variable (landslide) and α_0 is the intercept, $\alpha_1, \dots, \alpha_n$ are regression coefficients, x_1, \dots, x_n are the independent variables and e is error term. To convert the logit(L) into probability (P) the following equation is used:

$$P_{LR}(X) = \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n + \varepsilon)}{1 + \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n + \varepsilon)}, \quad (7)$$

where $P_{LR}(X)$ signifies the landslide probability for each input variable (X) between [0, 1]. Higher values of $P_{LR}(X)$ ($P_{LR}(X) > 0.5$) indicate a higher chance of slope failure, and lower values of P_{LR} ($P_{LR} < 0.5$) refer to higher slope stability (Zhao et al., 2019a, b). Logistic regression model was implemented using “glm” function in R-statistical software.

Naïve Bayes

NB is a classifier based on Bayesian law and maximum posterior hypothesis for statistical analysis based on the conditional independence assumption of variables (Tsanagaratos and Ilia 2016).



Because NB benefits from its simplicity, i.e., involves no complex iteration parameters during the model building (Wu et al. 2008), it becomes a widespread method for landslide studies (Tien Bui et al. 2012). In this study, consider $X = x_i (i=1,2,\dots,13)$ as the landslide 13 conditioning factors and $L_j (j=0(\text{Landslide, Non}), 1(\text{Landslide}))$ as the prediction class. The prediction class L_j using NB can be defined as:

$$L_j = \operatorname{argmax}\{P(L_j) \prod_{i=1}^{13} P(x_i|L_j)\}, \quad (8)$$

where $P(L_j)$ is the prior probability of L_j class, that is calculated from the ratio of the observed cases with actual class L_j in the training dataset. $P(x_i|L_j)$ is the conditional probability:

$$P(x_i|L_j) = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(x_i-\mu)^2}{2\delta^2}}, \quad (9)$$

where μ and δ are mean and standard deviation of x_i . “naivebayes” package (Majka 2019) was used to perform NB classification in this study.

Random forest

RF is an ensemble learning method of individual binary decision trees to produce higher predictions with broad

applications in regression, classification, and feature selection (Cutler et al. 2007). The basic principle behind random forest is to produce multiple uncorrelated decision trees ($h(X, \theta_k; K : 1, 2, 3 \dots n)$) to generate training subsets through the bootstrap aggregation model (Tibshirani 1996; Breiman 2001). Each decision tree predicts the sample classification individually, and the final result is decided based on the output of the individual tree (Sun et al. 2020). Not all training samples are included in the bagging process so that about two-thirds of the data are considered in-bag samples used for training the model. The remaining one-third of the samples are known as out-of-bag observations (OOBs), which evaluate the model's overall performance (accuracy) (Breiman 2001; Rodriguez-Galiano et al. 2012; Belgiu and Drăgu 2016). The predicted class or final RF result is determined by a majority votes or average of prediction derived from the growing trees (Breiman 2001; Cutler et al. 2007). RF also provides proximity index, and relative variable importance (VI) measures throughout the classification model building (Rodriguez-Galiano et al. 2012). VI measure is valuable for selecting the most influential features in a multidimensional dataset (Ghimire et al. 2010; Rodriguez-Galiano et al. 2012). The VI can be calculated using the mean decrease in Gini or the mean decrease in accuracy (Breiman 2001). In this study, the RF classification was performed through “randomforest” package (Liaw and Wiener 2002) available in R-statistical software.

There are two user-defined parameters for random forest implementation, including the number of growing trees (*n_{tree}*) to grow in the model and the number of variables (*m_{try}*) to split at each node (Sahin et al. 2020; Wang et al. 2020b). It is found that the RF model is susceptible to *m_{try}* parameter compared to the *N_{tree}* parameter (Ghosh et al. 2014; Belgiu and Drăgu 2016).

Support vector machines

SVM is a supervised learning algorithm based on statistical learning theory to delimit a hyperplane boundary that optimally classifies in multidimensional space (Cortes and Vapnik 1995; Kavzoglu and Colkesen 2009). As a powerful generalization and optimization model, SVM can transform the dataset into a high-dimensional space and preserve convergence. Consider a matrix of conditioning factors (*LCFs* : $X = X_{i(i=1,2,3,..13)}$) and a vector of landslide classes (landslide, non-landslide; $L_j = \{1, -1\}$). The optimal hyperplane can be derived as follow:

$$f(x) = \text{sign} \left[\sum_{i=1}^m a_i L_j K(X, X_i) + b \right], \tag{10}$$

where $f(x)$ is the SVM regression function, a_i is the positive real constant, m is total number of LCFs, b is the bias, and $K(X, X_i)$ is Kernel function that can be sigmoid, polynomial, linear, or radial basis function (Pham et al. 2016b). In a binary classification, the above equation (Eq. 10) can be solved as:

$$L_j \left[\omega^T \varphi(X_i) + b \right] \geq 1, \tag{11}$$

where $\varphi(X_i)$ is non-linear function that transforms the input space into high-dimensional space, and ω is weight vector. The classification and accuracy of SVM depend on the kernel function (Damaševičius 2010). The RBF kernel is the most potent and effective Kernel because of fewer parameters and excellent ability to reflect the non-linear relationship with high interpolation ability (Marjanović et al. 2011). The overall performance of the SVM model depends on the kernel parameters, such as the regularization parameter (C) and the kernel width (γ). In this study, we used the RBF with Gaussian kernel (Marjanović et al. 2011) to classify the non-linear characteristics of the landslide problem. The “e1071” package (Meyer et al. 2019) was used for LSP modeling using SVM. The SVM implementation includes a `tuneSVM` function that automatically selects the optimal regularization parameter (C) and kernel width (γ). Also, we used the “`tuneSVM`” function available in R to select the optimal kernel parameters ($\gamma = 0.5, C=10$).

Once the models were trained and validated, they were used to estimate the LSP for every pixel in the study area. The landslide susceptibility prediction models were then classified into five classes: very-low susceptible (VLS), low susceptible (LS), moderate susceptible (MS), high susceptible (HS), and very-high susceptible (VHS) regions using “equal interval” method, as a standard classification method for LSP modeling (Chen et al. 2018).

Models validation and performance evaluation

Performance evaluation is essential to evaluate the reliability and effectiveness of LSP models (Saha and Saha 2021) and to infer the most suitable model in the QTP region. The predictive capability of five LSP models was estimated using a confusion matrix (Youssef and Pourghasemi 2021). The confusion matrix provides the following four parameters: the true-positive (TP) and true-negative (TN), referring to the numbers of pixels correctly classified (landslide, non-landslide), and the false-positive (FP) and false-negative (FN), indicating the numbers of pixels classified incorrectly (Pham et al. 2020). Based on these derived parameters, three performance metrics including accuracy, sensitivity (SST), specificity (SPF), were calculated to compare the performance of the models and highlight the most suitable ML model for landslide susceptibility prediction for the entire QTP region (Table 2) (Pham et al. 2020; Wang et al. 2020a).

Besides, the receiver operating characteristic curve (ROC) and the area under ROC curve (AUC), widely used performance metrics, were also computed to compare the performance of the five models (Chen et al. 2020; Steger et al. 2021). In this study, ROC and AUC were calculated using “pROC” package (Robin et al. 2011) in R software.

Results and analysis

Suitability assessment of the factors for model training by MCA technique

The MCA analysis confirmed the suitability of all these factors (Table 3) for the ML modeling training. Although two factors of elevation and rainfall

Table 2. Statistical indexes used for evaluation of model performance and comparison.

No.	Metric	Equation
1	Accuracy (Acc)	$ACC = \frac{TP+TN}{TP+FP+TN+FN}$
2	Sensitivity (SST)	$SST = \frac{TP}{TP+FN}$
3	Specificity (SPF)	$SPF = \frac{TN}{TN+FP}$

Table 3 MCA of landslide conditioning factors (LCFs) using VIF and TOL

LCFs	Multicollinearity index	
	VIF	TOL
Prc	1.227	0.814
Plc	1.592	0.628
Elev	3.180	0.335
Dsf	2.243	0.445
Fd	2.280	0.438
NDVI	2.300	0.434
Rainfall	3.098	0.316
Dsd	1.324	0.754
Dd	1.109	0.901
Slope	1.214	0.823
SPI	1.179	0.848
TWI	1.076	0.929
Lithology	1.597	0.626

showed relatively higher VIF and lower TOL values ($VIF_{Elev.} = 3.180$, $VIF_{rainfall} = 3.098$, and $TOL_{Elev.} = 0.335$, $TOL_{rainfall} = 0.316$) than other factors, they did not exceed the critical threshold ($VIF > 10$, and $TOL < 0.1$) and thus can also be used for the model training.

The most important factors in modeling process

The relative importance of the landslide conditioning factors was assessed by MDG and MDA metrics obtained from the random forest model (Sahin et al. 2020). The mean decrease accuracy is measured from OOB error, and the mean decrease Gini indicates the role of individual variables in preserving uniformity of nodes and leaves throughout the model building (Ghosh et al. 2014; Belgii and Drăgu 2016). Higher values in both criteria indicate a higher role of the LCF in the LSP analysis (Williams 2011). The results in Fig. 5A, B concluded that rainfall with MDG (151.87) and MDA (0.143) was the most critical factor with a significant role in the distribution of landslides in the study area, followed by elevation (MDG = 112.696 and MDA = 0.117), and lithology (MDG = 89.140 and MDA = 0.078). Other factors including NDVI, distance to drainage, profile curvature, fault density, drainage density, distance to faults, plan curvature, SPI, TWI, and slope were ranked in consequent positions.

Application of machine learning models in landslide susceptibility prediction mapping

The five ML models were generated from training (LCFs:70%) and test datasets (LCFs:30%) based on the relationship between the LCFs and landslide probability

conditions to visualize the model's prediction capability in the study area. The susceptibility values are classified using equal interval methods of classifications, including very-low susceptible, low susceptible, moderate susceptible, high susceptible, and very-high susceptible. Results and outcomes from the five machine learning models are reported below.

Landslide susceptibility prediction by DNN

The DNN model, trained by the SGD and Log-Loss algorithms and hyperparameters, encompasses an input layer, three hidden layers, and an output layer (1-3-1 structure). A summary of the model structure and its hyperparameters is represented in Table 4. Figure 6 consists of the two basic metrics including loss function error (Loss) curve and accuracy curve (Acc) (ranging from 0 to 1, referring to non-landslide and landslide probabilities, respectively) from the training and validation phases which demonstrates the performance of the model over 300 epochs. Increasing in accuracy curve and decreasing the loss curve indicates that the learning speed of the model rapidly improves in both training and validation phase, implying that its ability to predict the landslides from the causative factors (input factors) improves over successive epochs. Throughout the 300 epochs, the minimum loss and maximum accuracy were 0.247 and 0.909 for training, and 0.255 and 0.895 for validation phases, respectively. The observed fluctuations in the loss and accuracy curves may be attributed to the effect of the dropout layer, which has been configured into the model to avoid overfitting (Sameen et al. 2020). In addition, this fluctuation can be related to the effect of the SGD algorithm, which causes additional fluctuations to the loss function (Nhu et al. 2020a). Overall, the model shows a reasonable range of learning during the training and validation phases.

The predicted landslide probabilities from the DNN model were used to produce the LSP map for the study area. The landslide susceptibility values (probabilities) obtained from the DNN ranged from 0.000 to 0.998 and were divided into five classes (Fig. 8A). The area percentage of each susceptible level is calculated and presented in Fig. 9. From a qualitative perspective, it is observed that most landslide events occurred in high-susceptible and very-high susceptible regions, which occupy about 30.83% of the total area. The highest area of the susceptible region belongs to the very-low susceptible region, which covers 36.23% of the total area, and very-high susceptible covers the minimum area with 14.90% (see Fig. 9).

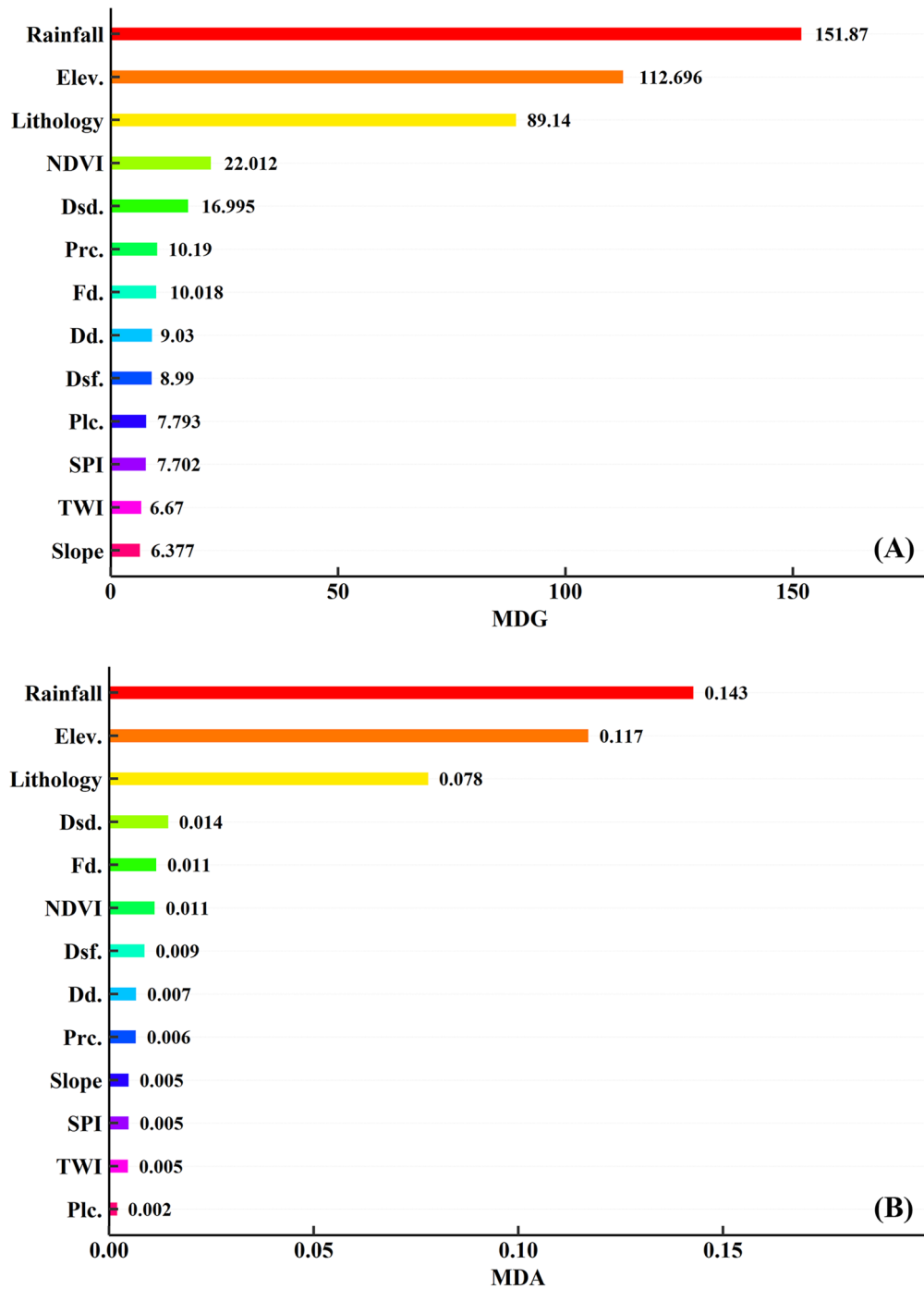


Fig. 5 RF model: variable importance (VI) highlighting the importance of individual landslide conditioning factors using mean decrease Gini (MDG, **A**), and mean decrease accuracy (MDA, **B**)

Table 4 Summary of model structure and hyper-parameter used for building a deeply connected neural network

Hyperparameters	Setting values
Number of inputs parameters	13
Number of densely connected units	13
Activation function	ReLU
Layer dropout rate	0.4,0.3,0.2
Transfer function	Sigmoid
Loss function	Log Loss
Optimizer	SGD
Learning rate	0.001
Metrics	Accuracy
Epoch	300
Batch size	32
Validation split	0.2

Landslide susceptibility prediction by LR

The spatial relationship between the conditioning factors and landslide events was calculated using the LR model,

and the regression coefficients and related statistics are provided in Table 5. The final LR model is determined as:

$$\begin{aligned}
 LR = & 0.076 + (0.016 * Plc) + (0.143 * Prc) \\
 & - (2.045 * elev) + (0.134 * Dsf) \\
 & - (0.550 * Fd) - (0.575 * NDVI) \\
 & + (0.989 * Rainfall) - (0.553 * Dsd) \\
 & + (0.114 * Dd) + (0.231 * Slope) \\
 & + (0.025 * SPI) - (0.197 * TWI) + (0.778 * Lithology).
 \end{aligned}
 \tag{12}$$

The estimated coefficients were used to predict the landslide probability map of the study area (Table 5). The high level of statistical significance of *p*-value (*p* > 0.1) for *Dsf*, *D_d*, *Plc*, *Prc*, *SPI*, and *TWI*, indicates that these variables had no statistically significant effects on landslide occurrence in the study area. In contrast, elevation, rainfall, *Dsd*, lithology, *NDVI*, *F_d*, and slope statistically significantly (*p* < 0.01) affect landslide susceptibility in the study area. The susceptibility index values from the LR model ranged from 0.000 to 0.999 and were

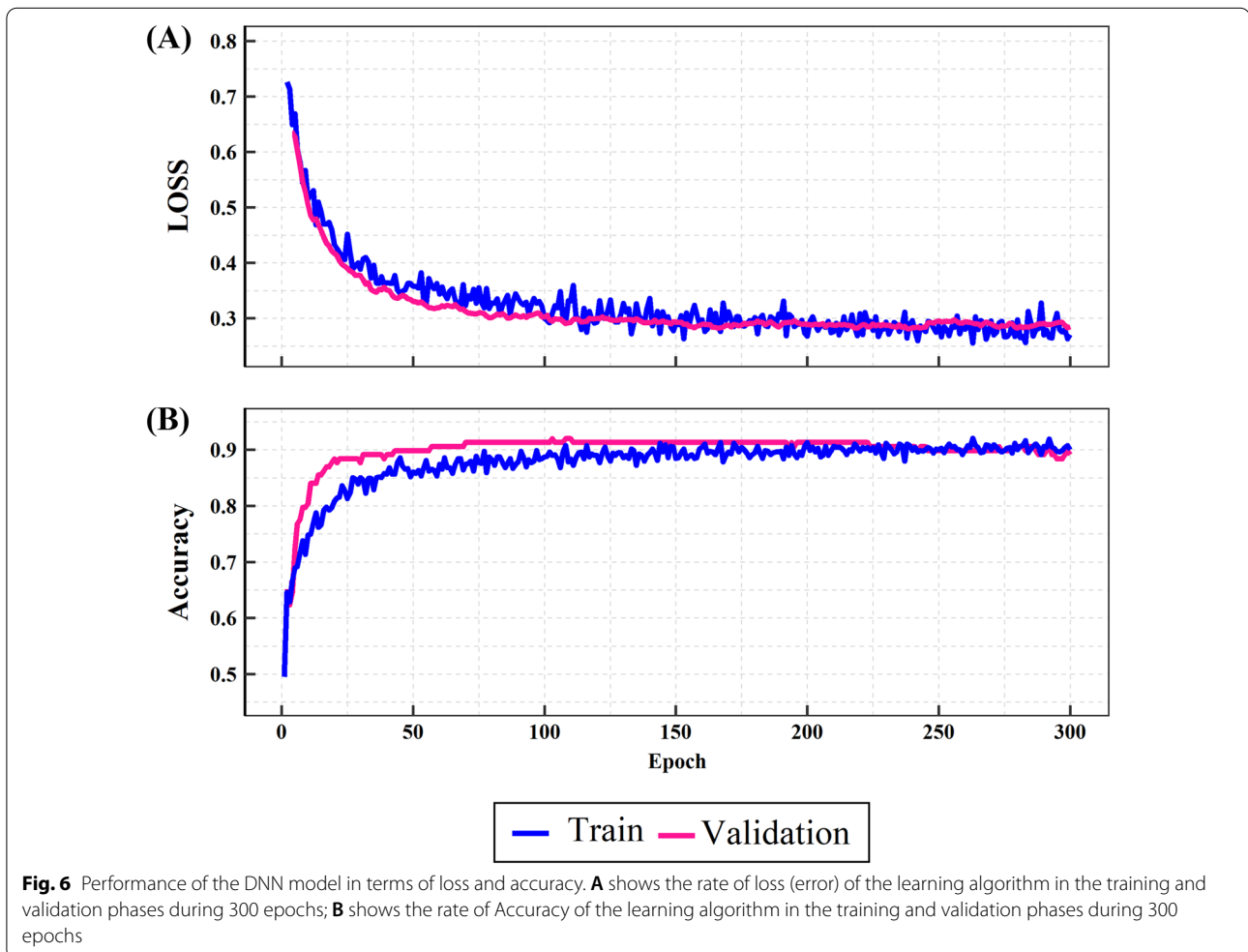


Fig. 6 Performance of the DNN model in terms of loss and accuracy. **A** shows the rate of loss (error) of the learning algorithm in the training and validation phases during 300 epochs; **B** shows the rate of Accuracy of the learning algorithm in the training and validation phases during 300 epochs

Table 5 Regression coefficients for the 13 landslide conditioning factors (LCFs) derived from the LR model

LCFs	β	S.E	z value	Pr(> z)
(Intercept)	0.076	0.137	0.552	0.580
Plc	0.016	0.153	0.103	0.9200
Prc	0.143	0.128	1.115	0.265
Elev	-2.045	0.248	-8.25	< 2e-16***
Dsf	0.134	0.199	0.675	0.500
Fd	-0.550	0.254	-2.161	0.031*
NDVI	-0.575	0.177	-3.236	0.001**
Rainfall	0.989	0.229	4.327	0.000***
Dsd	-0.553	0.147	-3.772	0.000***
Dd	0.114	0.144	0.793	0.4290
Lithology	0.778	0.147	5.302	0.0000***
Slope	0.231	0.133	1.732	0.083
SPI	0.025	0.150	0.168	0.867
TWI	-0.197	0.174	-1.136	0.256

Significant codes: 0****0.001***0.01**0.05*0.1"1"; β denotes the regression coefficient, and S.E. means standard error, β denotes the regression coefficient, and S.E. means standard error

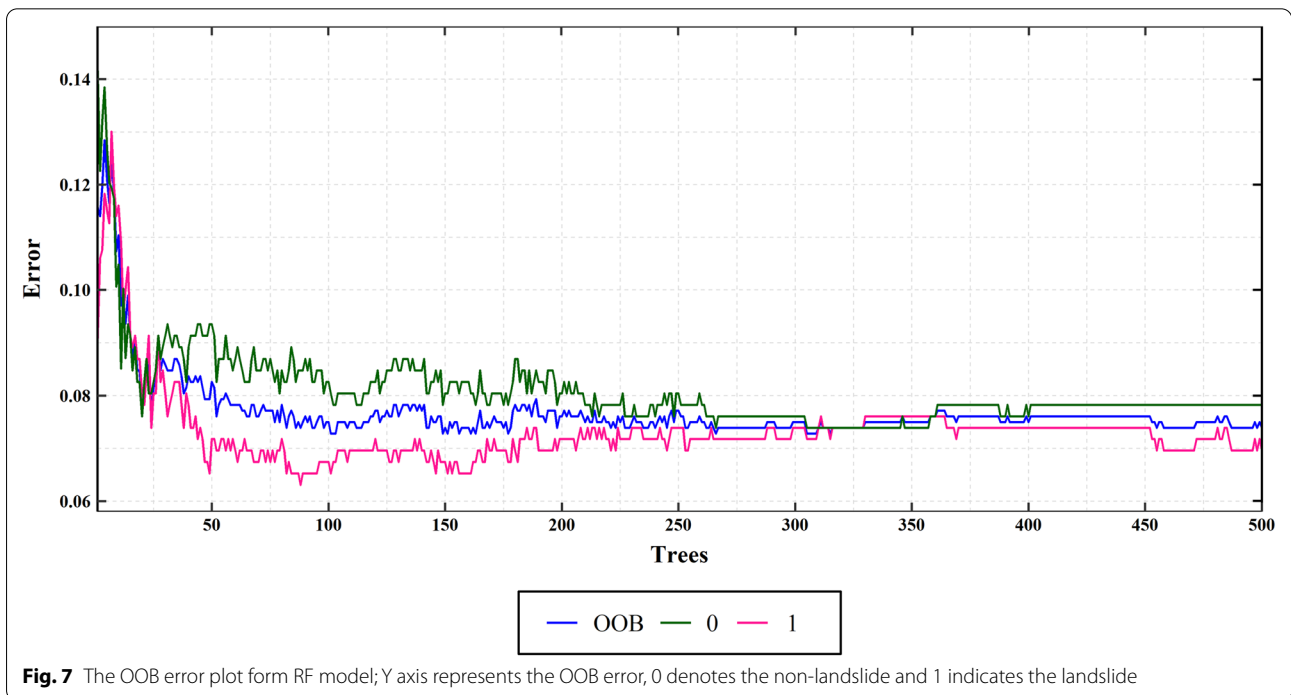
also classified into five classes (Fig. 8B). The area coverage of individual classes was 46.25, 14.32, 18.51, 12.04, and 8.87% for VLS, LS, MS, HS, and VHS, respectively (Fig. 9). The results showed that the highest percentage of the study area belonged to very-low susceptible and low susceptible regions (60.58%), whereas very-high susceptible regions occupied the minimum total area (8.87%).

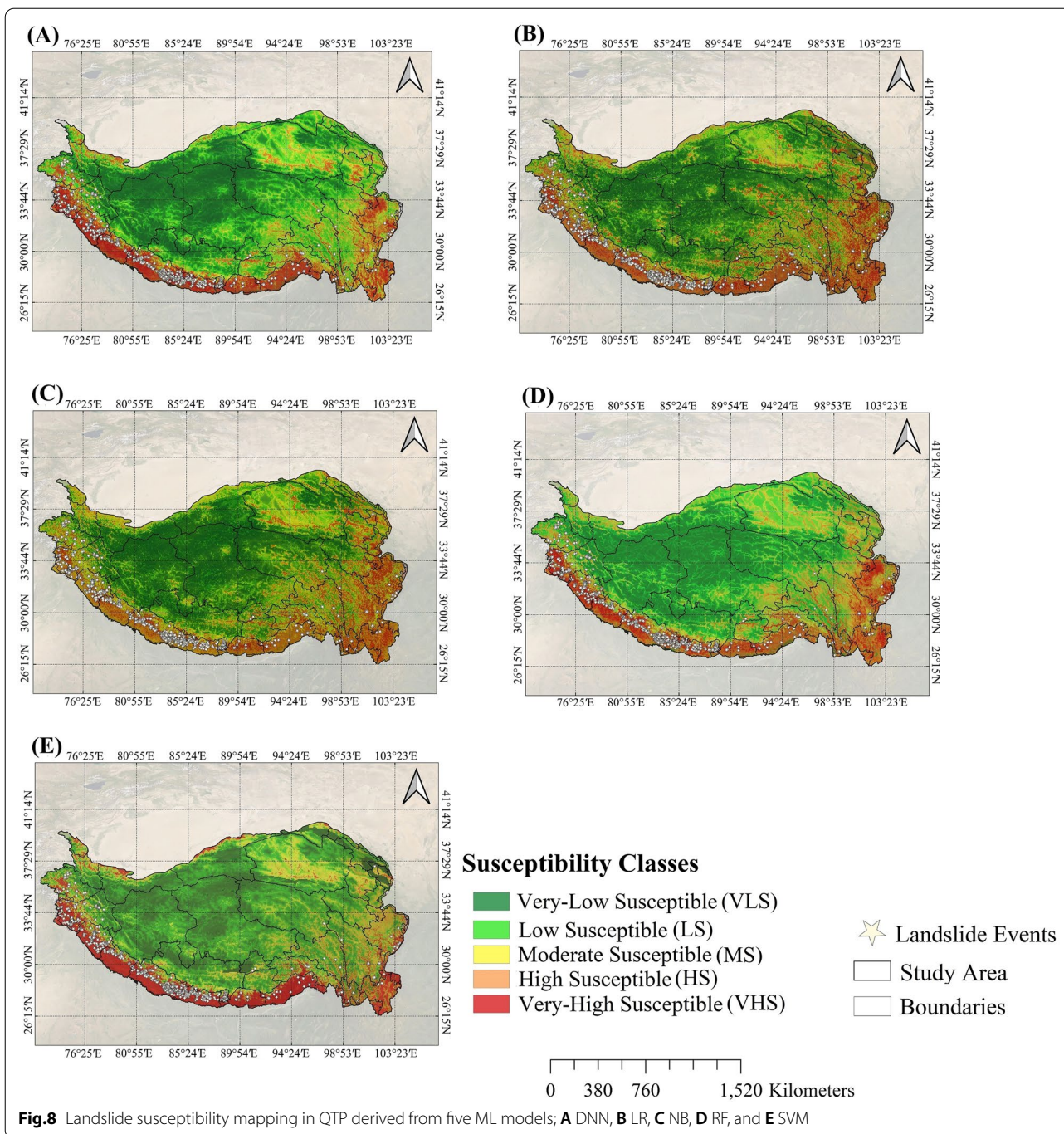
Landslide susceptibility prediction by NB

Using the training dataset, kernel density estimation (KDE), and additive smoothing (Laplace smoothing, Laplace=1), the NB model was constructed to predict the landslides in the study area. The susceptibility index values obtained from the NB model varied from 0.000 to 0.999, which is divided into VLS, LS, MS, HS, and VHS regions (Fig. 8C). The area coverage of each class is estimated at 42.18, 12.80, 21.50, 13.64, and 9.88%, respectively. The high and very-high susceptible regions cover a lower percentage than DNN, RF, and SVM models but are higher than the LR model. The moderate susceptible region in NB covers the highest percentage (21.50%) among all five models. A total percentage of 54.98% of the study area belongs to very-low susceptible and low susceptible regions (Fig. 9).

Landslide susceptibility prediction by RF

The *Ntree* and *mtry* were optimized (300 and 7, respectively) according to the out-of-bag observations error rate (Fig. 7). The OOB error analysis as a function of the number of trees (500 trees) illustrates that OOB decreases as the number of trees grow (see Fig. 7). This decreasing trend continues up to the value of 250 trees but becomes stable afterward. We used 300 trees as the optimal number of growing trees for the model implementation. Eventually, the RF model implemented from the training dataset (13 LCFs), with 300 trees and *mtry* parameter





of 7 (number of variables to split at each node), reached an OOB error of 7.28%, signifying that the algorithm attained the high capability to predict the landslide event from the input data. The LSP values predicted by the RF model for the entire region ranged from 0 to 1 (Fig. 8D) and were divided into five landslide-prone regions covering 33.84, 14.88, 19.83, 18.99, and 12.47% for VLS, LS,

MS, HS, and VHS, respectively (see Fig. 9). Overlaying of landslide locations on the LSP map confirmed the rationality of the produced map, showing a considerable match between the susceptible regions and landslide distribution in the study area. The minimum area covered belonged to the very-high susceptible class with only 12.47%.

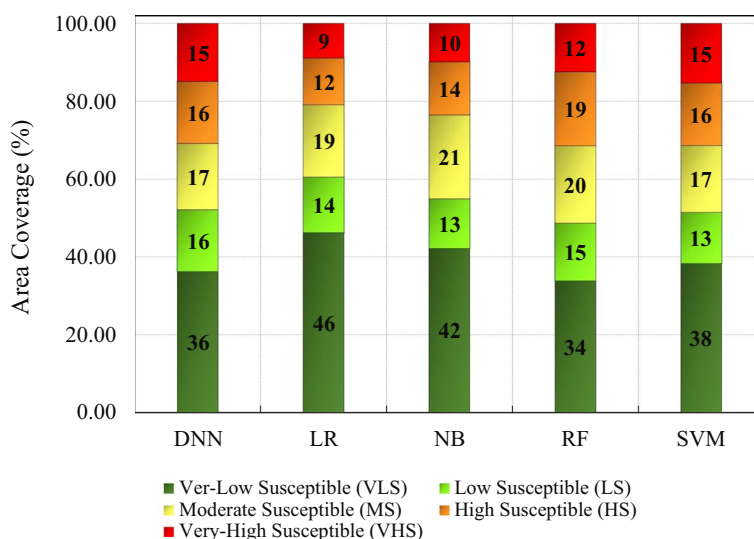


Fig. 9 Landslide susceptibility classes, area percentage (%) of the whole QTP for individual ML models; DNN, LR, NB, SVM, and RF

Table 6 Prediction performance comparison of five models using three criteria (test dataset)

Metrics	DNN	LR	Naïve Bayes	RF	SVM
TP	177	170	179	183	174
TN	181	181	165	181	180
FP	16	16	32	16	17
FN	20	27	18	14	23
Accuracy (Acc)	0.909	0.891	0.871	0.924	0.898
Sensitivity (SST)	0.898	0.863	0.838	0.929	0.883
Specificity (SPF)	0.919	0.919	0.909	0.919	0.914

Landslide susceptibility prediction by SVM

The SVM model was constructed for landslide prediction in the study area using the optimal hyperparameters and training dataset. We obtained the optimal kernel hyperparameters including γ and C equal to 0.5 and 10, respectively, by the “tune SVM” function. The predicted values from the SVM model ranged from 0.000 to 0.999. The calculated LSP values were divided into five landslide susceptibility classes (Fig. 8E). The LSP map obtained from the SVM model showed similar trends to the RF and DNN in landslide distribution, but with some differences. For example, SVM identified most of the total area as very-low and low susceptible regions, similar to predictions of DNN and RF models (about 51.47%). However, very-high susceptible locations in the SVM are relatively higher than the two models (15.30%) (see Fig. 9).

Performance comparison and validation

Model validation

The evaluation of the prediction capability of the models is critical for LSP modeling. As mentioned earlier, to assess the predictive capability of the five models, several performance criteria were applied using the test dataset, and the results are given in Table 6 and Fig. 10. The results showed relatively high variability in the performance of the five models across different criteria. The first measurement is accuracy ($0 < Acc < 1$), which indicates how the model accurately predicts landslides from input data in the training and test phases. The accuracy analysis revealed that RF had the highest accuracy ($Acc = 0.9239$), followed by DNN ($Acc = 0.9086$). The lowest accuracy was obtained in the NB ($Acc = 0.8731$). LR and SVM are considered the third and fourth accurate models with Acc values of 0.906 and 0.873. The second most important measure of the model capability is the sensitivity (SST), which indicates the number of landslide locations correctly classified as landslide (Shahabi et al. 2021). The SST analysis pointed to the higher performance of the RF model compared to the other four models. In terms of precision (specificity), three models including RF, DNN, and LR, showed a high value of precision ($SPF = 0.919$), while NB had the minimum precision ($SPF = 0.909$).

In addition, the overall performance of the five models was quantified using ROC-AUC analysis as a fundamental measure of model performance (Pham et al. 2019; Zhao et al. 2019a, b).

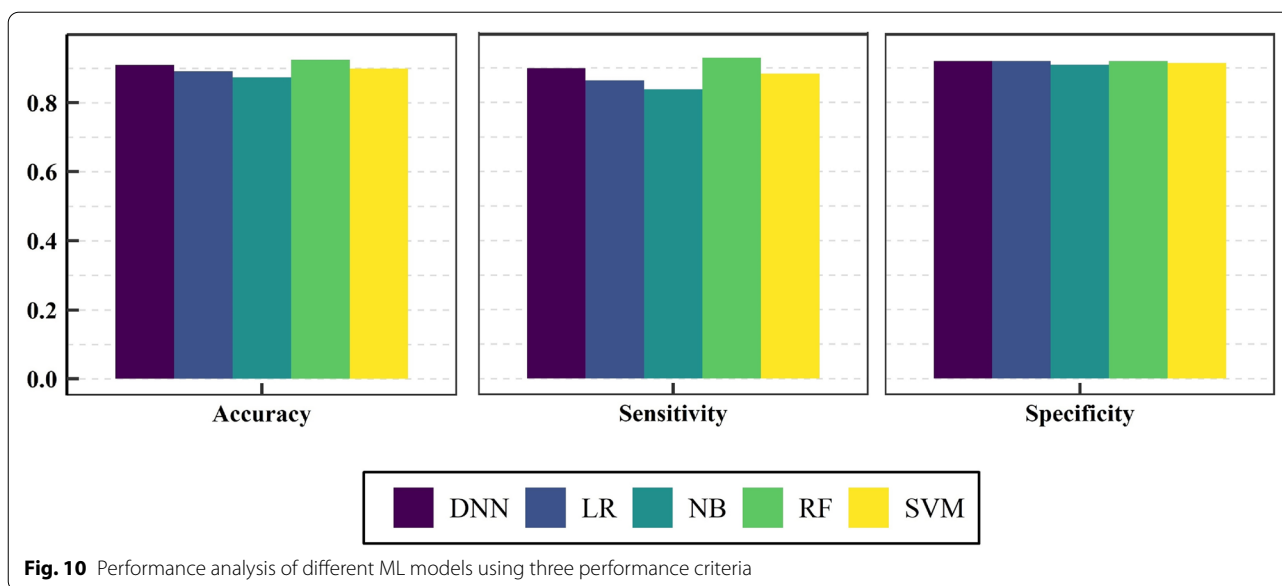


Fig. 10 Performance analysis of different ML models using three performance criteria

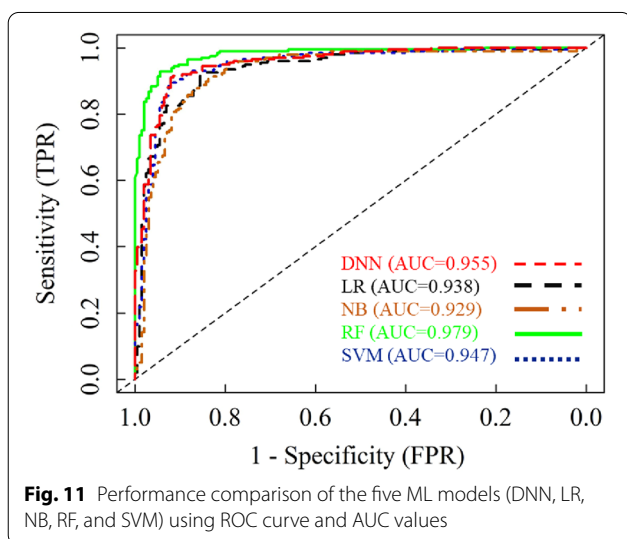


Fig. 11 Performance comparison of the five ML models (DNN, LR, NB, RF, and SVM) using ROC curve and AUC values

Validation and comparison of the LSP maps

We graphically evaluated the performances of the five models in preparing LSP maps by ROC curve and AUC based on the test dataset (Fig. 11). Although it shows that all the five ML models had a high predictive power (AUC > 0.930) (Yilmaz 2009), the RF model was highlighted as the most accurate and robust model (AUC = 0.980). The AUC values ranged from 0.930 (NB) to 0.980 (RF). The difference between the maximum and minimum AUC was only 4.99%. In terms of predictability, the RF model achieves an outstanding AUC value (nearly perfect performance), designating the highest level of agreement between predicted and observed landslide

events, followed by the DNN model (AUC = 0.9556) (see Fig. 11). It implies that RF and DNN have a high capability to predict possible future landslides in the study area. The AUC values for SVM, LR, and NB are 0.947, 0.947, and 0.930. NB showed the minimum AUC referring to the lowest model performance.

Discussion

Key factors controlling the landslide occurrence in QTP

The results presented the different performances of the five models in the QTP region, and it is well known that different causative factors have different contributions to the development of landslide events (Huang et al. 2020). Therefore, identifying variables that have an enormous impact on landslide occurrence is a high mandate for landslide susceptibility prediction, especially in the QTP region. Because the random forest model shows better performance than other models, the variable importance obtained from the RF model was used to highlight the most critical factors (see Fig. 5). The MDA and MDG considered the rainfall and elevation as the leading major conditioning factors followed by lithology, Dsd, and Fd. The high association of landslides with elevation and rainfall has been investigated in previous studies (Sun et al. 2020, 2021). The high inter-relationship between elevation and landslide becomes more tangible at the middle altitude regions (2000-3000 m) (see Fig. 2A), where the probability of slope failure decreases by elevation, as more resistant lithology is mainly formed at higher altitudes (Mohammady et al. 2012). The significant role of rainfall in the analysis can be attributed to the fact that rainfall events trigger most landslides in

the region. The analysis shows that the north-northwest and south-southeastern parts of the study area, which receive the most rainfall (> 1000 mm) (see Fig. 2G), have the highest concentration of landslides. Lithology is the third factor that played a significant role in the landslide events throughout the analysis ($MDA = 0.078$). Because different lithologic structures have different degrees of hardness, the permeability of bedrock and soil material (Ayalew and Yamagishi 2005) show different resistance rates to the landslide occurrence. The spatial distribution of landslides in the study area indicates that most landslides are concentrated in Precambrian–Phanerozoic (sedimentary rocks) and crystalline metamorphic rocks (Precambrian–Phanerozoic) (see Fig. 2E). The crystalline metamorphic and sedimentary rocks provide a highly propitious condition for the formation of joints, cracks, and faults, reducing the shear strength of the bedrock lithology, increasing the probability of slope failure (Yu and Chen 2020).

The role of D_{sd} can be attributed to the fact that a closer distance to the drainage network increases the slope cut by the drainage and reduces the shear strength of slope material and soil mass, which ultimately leads to slope failure (Yu and Chen 2020). The F_d , due to the presence of major, minor faults along the study area such as Himalayan main thrust fault (MF), Altyn-Tagh fault, Kunlun fault, Karakoram fault, Jiali fault, etc., showed a significant role in landslide development in the QTP region (Taylor and Yin 2009; Elliott et al. 2010; Aiken and Brierley 2013; Zhang et al. 2020; Qi et al. 2021). The role of faultiness (lineaments) has also been investigated in previous landslide studies, especially in different areas of QTP (Guo et al. 2015; Pham et al. 2017b; Zhao et al. 2019a; Qi et al. 2021). Overall, the significant roles of rainfall, elevation, lithology, and fault density in this study satisfy the two basic assumptions in landslide LSP modeling, which state: (1) landslides are a factor of surface topography (elevation, slope, geology, etc.), and (2) future landslides will occur under the impact of the same factors that caused the previous landslides (Guzzetti 2006).

Different performances of the five machine learning models

This study analyzed the performance and robustness of the five well-known machine learning models for landslide susceptibility prediction in QTP. The results showed that the landslide events are mainly distributed in high susceptible and very-high susceptible regions for all models, proving the reliability of the produced LSP maps (see Fig. 8A–E). The very-low susceptible and low susceptible locations have the maximum percentage of the total area ($A_{VLS} > 30\%$), whereas the very-high susceptible parts share the least percentage in every LSP map

produced by the five models ($A_{VHS} < 15\%$) (see Fig. 9). In addition, a similar trend of the area covered in each class was observed among all models, namely the area percentage decreases from very-low susceptible to low susceptible and then increases in moderate susceptible, and finally reaches the minimum percentage in the very-high susceptible region. From this similar pattern of LSP distribution, it is observed that the north-northwest and south-southeast direction of the QTP region, with hills and mountains, are the areas with high and very-high landslide-prone locations (see Fig. 8A–E), known as the frequent landslide sites reported in previous studies (Guo et al. 2015; Zhao et al. 2019b; Qi et al. 2021). Conversely, the central and eastern regions are considered low susceptible and very-low susceptible regions for landslide disasters, covering more than 45% of the total area in all studied models.

To highlight the differences between models' performance, a quantitative approach using several model validation criteria, including ACC, SST, and SPF was used in the test dataset (30% of the dataset). The performance metrics (see Fig. 10 and Table 6) showed that all models had reasonable goodness in their predictive performance with slight variations. In general, ML models delivered a high accuracy rate for landslide prediction because they are designed to obtain the optimal non-linear relationship between LCFs automatically (Achour and Pourghasemi 2020). However, the outperformance of a model can be quantified and identified when a model shows higher values in accuracy, sensitivity, and specificity compared to the other models (Nhu et al. 2020c). The analysis recorded the highest SST and SPF among five models for the RF model with 0.929 and 0.919, respectively. It indicates that 92.9% of the total landslide locations were correctly identified as landslide, and 91.9% of non-landslide locations were correctly identified as non-landslide (Nhu et al. 2020d). The classification performance of the DNN model in identifying landslide and non-landslide locations (SST, SPF) was highly comparable with the RF model with only 0.031 differences in SST, hence, the RF model slightly outperforms the DNN model. The lowest prediction capability in terms of SST and SPF belonged to the NB model with 0.838 and 0.909, respectively. Finally, SVM and LR model were showed a reasonable range of SST and SPF.

The ROC-AUC of the five models was also used to validate their reliability and compare their overall performance. The results were promising as the difference between the maximum and the minimum AUC was only 4.99%. However, even a tiny difference in AUC value may significantly impact landslide prediction in a particular area (Beguería 2006). Because landslide hazard maps with undesirable reliability and accuracy may lead to

severe consequences and socio-economic disasters (Thi Ngo et al. 2021), it is necessary to select models with a higher degree of reliability.

The analysis of ROC-AUC (see Fig. 11) indicated that the RF model ($AUC=0.9798$) is the most reliable and accurate compared to the other models, being consistent with finding in other related machine learning algorithm comparative studies (Goetz et al. 2015; de Oliveira et al. 2019; Achour and Pourghasemi 2020; Akinici et al. 2020). On the other hand, the DNN model showed a comparable result to the RF model, with a slight difference in AUC ($AUC_{RF} - AUC_{DNN} = 0.028$), ranking as the second in performance in this study. The best performance of the random forest model in many applications compared to the other classifiers was explained in the literature (Cutler et al. 2007; Sun et al. 2021). It also benefits from the OBB algorithm to optimize the membership probability and improve the model's overall performance (Breiman 2001; Rodriguez-Galiano et al. 2012). Concerning DNN, a major constraint in building it is finding the optimal parameter (hidden layers, neurons, etc.) to tune the structure of the model, which is a highly computational task and may affect the model performance (Bui et al. 2020). However, integrating random forest into deep neural network can significantly improve the model performance (Abbas 2018). Furthermore, the variable importance analysis by RF can select the most significant factors to reduce the dimensionality of the dataset and ultimately improve the model's performance (Sameen et al. 2020).

The analysis also revealed that the SVM model outperformed and outclassed the LR and NB models and ranked third among the five models. This finding was consistent with the previous studies that highlighted the superior performance of SVM compared to LR and NB models (Yao et al. 2008; Marjanović et al. 2011; Ballabio and Sterlacchini 2012; Tien Bui et al. 2012; Hu et al. 2020). The lower performance of LR and NB models can be attributed to their conceptual constraint, based on the independence assumption of variables that may be violated in different cases, reducing the models' predictive ability (Ballabio and Sterlacchini 2012).

Furthermore, the lower performance of SVM, LR, and NB compared to DNN and RF could be related to their lower generalization capability from the training dataset and increased error for actual data. This may be attributed to the overfitting issue as a common problem in landslide susceptibility mapping using machine learning models (Park and Kim 2019), especially at regional scale analysis. Data scarcity is a well-known constraint for LSP mapping at a regional scale (County and Bostjan 2021) regarding landslide inventory and conditioning factors data. Due to the scale of the analysis, despite applying

feature selection techniques, some noise may have been retained in the dataset and introduced to the model structures, which eventually reduced the model performance. In addition, because the landslide locations in the QTP regions occupy a significantly smaller portion than non-landslide regions, these models may not have optimally learned from input data to predict (generalize) the non-landslide locations. Accordingly, the relatively lower performance of DNN may be related to the remained noise in the dataset (Sameen et al. 2020). Hence, the combination of more advanced feature selection techniques such as principal component analysis (Sun et al. 2018) and variable importance can optimally reduce the dataset's noise and improve such models' predictive capability, especially DNN.

Comparison of landslide susceptibility prediction results with previous studies

Nevertheless, the accuracy of the LSP map obtained by the RF model in this study, which performed the best, was relatively higher than the previous similar studies in other parts of the region (e.g., Xianshuihe fault zone, eastern Himalayan syntax, and Pauri Garhwal in Uttarakhand, etc.) (Guo et al. 2015; Kumar et al. 2017; Du et al. 2019; Peethambaran et al. 2020). This comparison may be unfair because the previous studies have considered different areas of the region (local scales), and the landslide prediction results are susceptible to the spatial scale of the study, which determines the nature and the scale of the input dataset (Yesilnacar and Topal 2005; Achour and Pourghasemi 2020). Besides, it should be noted that different researchers use different data sources as another cause of the difference.

Conclusions

Landslide susceptibility prediction is a prerequisite for preventing and reducing landslide hazards, as one of the significant natural hazards that cause great damage to human life and the economy in QTP and surrounding areas. This study evaluated the performance of the five well-known and advanced machine learning models, including DNN, LR, NB, RF, and SVM, for LSP modeling in the entire QTP region with highly complex structure and surface topography. Before machine learning modeling implementation, data pre-processing (normalization and multi-collinearity analysis) was performed on thirteen landslide conditioning factors. After that, the five ML models were trained under similar conditions (input variables) using optimal hyperparameters to obtain the highest model performance according to the structural complexity of the study area. Finally, five landslide susceptibility prediction models were evaluated and

compared to highlight the most suitable LSP model using three state-of-the-art performances criteria.

Our finding revealed a reasonable range of goodness-of-fit in all the models with the superior performance of the RF model. The higher performance of the RF model confirmed its higher predictive capability for future landslides in the study area. This provided a comprehensive insight into the landslide prediction for future landslide hazards management and practices in the study area. However, the highly comparable result of the DNN model suggested that integrating the random forest model into DNN model can significantly improve the model performance and provide a more robust model for future landslide predictions.

The results of LSP maps from the five ML models depicted that the north-northwestern and south-south-eastern regions are highly susceptible) and very-high susceptible regions to the landslide events. Conversely, the low-susceptible and very-low susceptible regions are located in the study area's center, west, and northwest. The main limitation of this study area can be related to the data scarcity in terms of landslide historical events (triggering source) and landslide triggering factors (especially geological information) which is a well-known constraint for LSP modeling at the regional scale. Indeed, the availability of various sources of landslide events (snowmelt induced, earthquake-induced, etc.) is helpful for building a more comprehensive landslide inventory as the baseline information for improving the prediction capability of ML in LSP modeling. On the other hand, the results from the analysis highlighted that landslide events in the study area are affected by tectonics, topography, and rainfall. Therefore, a further seismic-related analysis, such as focal mechanism, peak ground acceleration, and peak ground velocity analysis, is highly recommended to improve the LSP modeling in the region. The combination of a comprehensive landslide inventory with earthquake-related analysis offers an excellent tool for early landslide predictions and landslide-related disasters mitigations in the study area. This eventually provides an unprecedented insight into the landslide causative factors in the entire QTP for further division of the area based on the landslide triggering factors.

Funding

This project was financially supported by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (No. 2019QZKK0903), the National Natural Science Foundation of China (No. 41971040), and the CAS Interdisciplinary Innovation Team (No. JCTD-2019-04).

Data and codes availability

Data generated for landslide susceptibility predicting using five ML models are mainly provided in the form of tables and figures in the content of manuscript documents. Further ML data and codes are available upon request from the corresponding author. Also, landslide inventories are obtained from the Global

Landslide Catalog (GLC). SRTM-DEM (90 m) is available at SRTM-DEM (90 m) available in the GEE platform (CGIAR/SRTM90_V4). The Landsat OLI dataset is available at (http://landsat.usgs.gov/CDR_LSR, and http://landsat.usgs.gov/CDR_LSR).

Declarations

Competing interests

The authors declare no conflict of interest.

Author details

¹Key Laboratory of Water Cycle & Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. ²Key Laboratory of Compound and Chained Natural Hazards Dynamics, Ministry of Emergency Management of China, Beijing 100085, China. ³Department of Civil Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran. ⁴Department of Engineering, University of Perugia, Perugia, Italy. ⁵School of Environmental Sciences, Jawaharlal Nehru University, New-Delhi, India.

Received: 30 August 2021 Accepted: 25 January 2022

Published online: 14 February 2022

References

- Abbas MA (2018) Improving deep learning performance using random forest HTM cortical learning algorithm. <https://doi.org/10.1109/IWDRL.2018.8358209>
- Abbaszadeh Shahri A, Spross J, Johansson F, Larsson S (2019) Landslide susceptibility hazard map in southwest Sweden using artificial neural network. *CATENA* 183:104225. <https://doi.org/10.1016/j.catena.2019.104225>
- Achour Y, Pourghasemi HR (2020) How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? *Geosci Front* 11:871–883. <https://doi.org/10.1016/j.gsf.2019.10.001>
- Aghdam IN, Pradhan B, Panahi M (2017) Landslide susceptibility assessment using a novel hybrid model of statistical bivariate methods (FR and WOE) and adaptive neuro-fuzzy inference system (ANFIS) at southern Zagros Mountains in Iran. *Environ Earth Sci*. <https://doi.org/10.1007/s12665-017-6558-0>
- Aiken SJ, Brierley GJ (2013) Analysis of longitudinal profiles along the eastern margin of the Qinghai-Tibetan Plateau. *J Mt Sci* 10:643–657. <https://doi.org/10.1007/s11629-013-2814-2>
- Akinci H, Kilicoglu C, Dogan S (2020) Random forest-based landslide susceptibility mapping in coastal regions of Artvin, Turkey. *ISPRS Int J Geo-Information*. <https://doi.org/10.3390/ijgi9090553>
- Al-Najjar HAH, Pradhan B (2021) Spatial landslide susceptibility assessment using machine learning techniques assisted by additional data created with generative adversarial networks. *Geosci Front* 12:625–637. <https://doi.org/10.1016/j.gsf.2020.09.002>
- Arabameri A, Saha S, Roy J et al (2020) A novel ensemble computational intelligence approach for the spatial prediction of land subsidence susceptibility. *Sci Total Environ* 726:138595. <https://doi.org/10.1016/j.scitotenv.2020.138595>
- Arora A, Pandey M, Siddiqui MA et al (2019) Spatial flood susceptibility prediction in Middle Ganga Plain: comparison of frequency ratio and Shannon's entropy models. *Geocarto Int*. <https://doi.org/10.1080/10106049.2019.1687594>
- Atkinson PM, Massari R (2011) Autologistic modelling of susceptibility to landsliding in the Central Apennines, Italy. *Geomorphology* 130:55–64. <https://doi.org/10.1016/j.geomorph.2011.02.001>
- Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology* 65:15–31. <https://doi.org/10.1016/j.geomorph.2004.06.010>
- Ballabio C, Sterlacchini S (2012) Support vector machines for landslide susceptibility mapping: the Staffora River basin case study, Italy. *Math Geosci* 44:47–70. <https://doi.org/10.1007/s11004-011-9379-9>

- Bartarya SK, Viridi NS, Sah MP (1996) Landslide hazards: some case studies from the Satluj Valley, Himachal Pradesh. *J Him Geol* 17:193–207
- Beguería S (2006) Validation and evaluation of predictive models in hazard assessment and risk management. *Nat Hazards* 37:315–329. <https://doi.org/10.1007/s11069-005-5182-6>
- Belgiu M, Drăgu L (2016) Random forest in remote sensing: a review of applications and future directions. *ISPRS J Photogramm Remote Sens* 114:24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Bishop CM (1995) *Neural networks for pattern recognition*. Oxford University Press, Oxford
- Bordoni M, Galanti Y, Bartelletti C et al (2020) The influence of the inventory on the determination of the rainfall-induced shallow landslides susceptibility using generalized additive models. *CATENA* 193:104630. <https://doi.org/10.1016/j.catena.2020.104630>
- Borga M (2019) Hazard assessment and forecasting of landslides and debris flows: a case study in Northern Italy. *Extrem Hydroclimatic Events Multivar Hazards Chang Environ*. <https://doi.org/10.1016/b978-0-12-814899-0.00014-6>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1201/9780429469275-8>
- Budimir MEA, Atkinson PM, Lewis HG (2015) A systematic review of landslide probability mapping using logistic regression. *Landslides* 12:419–436. <https://doi.org/10.1007/s10346-014-0550-5>
- Bui DT, Tsangaratos P, Nguyen VT et al (2020) Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *CATENA* 188:104426. <https://doi.org/10.1016/j.catena.2019.104426>
- Carranza-García M, García-Gutiérrez J, Riquelme JC (2019) A framework for evaluating land use and land cover classification using convolutional neural networks. *Remote Sens*. <https://doi.org/10.3390/rs11030274>
- Chang Z, Du Z, Zhang F et al (2020) Landslide susceptibility prediction based on remote sensing images and GIS: Comparisons of supervised and unsupervised machine learning models. *Remote Sens*. <https://doi.org/10.3390/rs12030502>
- Chen W, Xie X, Peng J et al (2017) GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. *Geomatics Nat Hazards Risk* 8:950–973. <https://doi.org/10.1080/19475705.2017.1289250>
- Chen W, Zhang S, Li R, Shahabi H (2018) Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Sci Total Environ* 644:1006–1018. <https://doi.org/10.1016/j.scitotenv.2018.06.389>
- Chen W, Fan L, Li C, Pham BT (2020) Spatial prediction of landslides using hybrid integration of artificial intelligence algorithms with frequency ratio and index of entropy in Nanzheng county, China. *Appl Sci* 10:1–21. <https://doi.org/10.3390/app10010029>
- Choi J, Oh HJ, Lee HJ et al (2012) Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS. *Eng Geol* 124:12–23. <https://doi.org/10.1016/j.enggeo.2011.09.011>
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297
- Costanzo D, Chacón J, Cosocenti C et al (2014) Forward logistic regression for earth-flow landslide susceptibility assessment in the Platani river basin (southern Sicily, Italy). *Landslides* 11:639–653. <https://doi.org/10.1007/s10346-013-0415-3>
- County S, Bostjan I (2021) Regional-scale landslide susceptibility mapping using limited LiDAR-based landslide inventories for Sisak-Moslavina County, Croatia. *Sustainability*. <https://doi.org/10.3390/su13084543>
- Crosta GB, Imposimato S, Roddeman DG (2003) Numerical modelling of large landslides stability and runoff. *Nat Hazards Earth Syst Sci* 3:523–538. <https://doi.org/10.5194/nhess-3-523-2003>
- Cutler DR, Edwards TC, Beard KH et al (2007) Random forests for classification in ecology published by: Ecological Society of America. *Ecology* 88:2783–2792
- Dai FC, Lee CF, Ngai YY (2002) Landslide risk assessment and management: an overview. *Eng Geol* 64:65–87. [https://doi.org/10.1016/S0013-7952\(01\)00093-X](https://doi.org/10.1016/S0013-7952(01)00093-X)
- Damaševićius R (2010) Optimization of SVM parameters for recognition of regulatory DNA sequences. *TOP* 18:339–353
- de Oliveira GG, Ruiz LFC, Guasselli LA, Haetinger C (2019) Random forest and artificial neural networks in landslide susceptibility modeling: a case study of the Fão River Basin, Southern Brazil. *Nat Hazards* 99:1049–1073. <https://doi.org/10.1007/s11069-019-03795-x>
- Deng H, Wu LZ, Huang RQ et al (2017) Formation of the Siwanli ancient landslide in the Dadu River, China. *Landslides* 14:385–394. <https://doi.org/10.1007/s10346-016-0756-9>
- Di Napoli M, Carotenuto F, Cevasco A et al (2020) Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. *Landslides* 17:1897–1914. <https://doi.org/10.1007/s10346-020-01392-9>
- Dormann CF, Elith J, Bacher S et al (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography (cop)* 36:27–46
- Dou J, Bui DT, Yunus AP et al (2015) Optimization of causative factors for landslide susceptibility evaluation using remote sensing and GIS data in parts of Niigata, Japan. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0133262>
- Dou J, Yunus AP, Tien Bui D et al (2019) Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci Total Environ* 662:332–346. <https://doi.org/10.1016/j.scitotenv.2019.01.221>
- Du M, Kawashima S, Yonemura S et al (2004) Mutual influence between human activities and climate change in the Tibetan Plateau during recent years. *Glob Planet Change* 41:241–249
- Du G, Zhang Y, Yang Z et al (2019) Landslide susceptibility mapping in the region of eastern Himalayan syntaxis, Tibetan Plateau, China: a comparison between analytical hierarchy process information value and logistic regression-information value methods. *Bull Eng Geol Environ* 78:4201–4215. <https://doi.org/10.1007/s10064-018-1393-4>
- Elliott JR, Walters RJ, England PC et al (2010) Extension on the Tibetan plateau: recent normal faulting measured by InSAR and body wave seismology. *Geophys J Int* 183:503–535. <https://doi.org/10.1111/j.1365-246X.2010.04754.x>
- Ercanoglu M (2005) Landslide susceptibility assessment of SE Bartin (West Black Sea region, Turkey) by artificial neural networks. *Nat Hazards Earth Syst Sci* 5:979–992. <https://doi.org/10.5194/nhess-5-979-2005>
- Fang Z, Wang Y, Peng L, Hong H (2020) Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. *Comput Geosci* 139:104470. <https://doi.org/10.1016/j.cageo.2020.104470>
- Ghimire B, Rogan J, Miller J (2010) Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. *Remote Sens Lett* 1:45–54
- Ghosh A, Fassnacht FE, Joshi PK, Kochb B (2014) A framework for mapping tree species combining hyperspectral and LiDAR data: role of selected classifiers and sensor across three spatial scales. *Int J Appl Earth Obs Geoinf* 26:49–63. <https://doi.org/10.1016/j.jag.2013.05.017>
- Gillespie TW, Madson A, Cusack CF, Xue Y (2019) Changes in NDVI and human population in protected areas on the Tibetan Plateau. *Arctic, Antarct Alp Res* 51:428–439. <https://doi.org/10.1080/15230430.2019.1650541>
- Goetz JN, Brenning A, Petschko H, Leopold P (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput Geosci* 81:1–11. <https://doi.org/10.1016/j.cageo.2015.04.007>
- Guo C, Montgomery DR, Zhang Y et al (2015) Quantitative assessment of landslide susceptibility along the Xianshuihe fault zone, Tibetan Plateau, China. *Geomorphology* 248:93–110. <https://doi.org/10.1016/j.geomorph.2015.07.012>
- Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study. *Central Italy Geomorphol* 13:1995
- Guzzetti F, Reichenbach P, Cardinali M et al (2005) Probabilistic landslide hazard assessment at the basin scale. *Geomorphology* 72:272–299. <https://doi.org/10.1016/j.geomorph.2005.06.002>
- Guzzetti F (2006) *Landslide hazard and risk assessment*. Doctoral dissertation, Rheinische Friedrich Wilhelms-Universität Bonn
- Hong H, Pourghasemi HR, Pourtaghi ZS (2016) Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical

- models. *Geomorphology* 259:105–118. <https://doi.org/10.1016/j.geomorph.2016.02.012>
- Hu Q, Zhou Y, Wang S, Wang F (2020) Machine learning and fractal theory models for landslide susceptibility mapping: case study from the Jinsha River Basin. *Geomorphology* 351:106975. <https://doi.org/10.1016/j.geomorph.2019.106975>
- Huang X, Sillanpää M, Duo B, Gjessing ET (2008) Water quality in the Tibetan Plateau: Metal contents of four selected rivers. *Environ Pollut* 156:270–277. <https://doi.org/10.1016/j.envpol.2008.02.014>
- Huang X, Sillanpää M, Gjessing ET et al (2011) Water quality in the southern Tibetan Plateau: chemical evaluation of the Yarlung Tsangpo (Brahmaputra). *River Res Appl* 27:113–121
- Huang F, Wang Y, Dong Z et al (2019) Regional landslide susceptibility mapping based on grey relational degree model. *Earth Sci* 44:664–676
- Huang F, Cao Z, Guo J et al (2020) Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. *CATENA* 191:104580. <https://doi.org/10.1016/j.catena.2020.104580>
- Hutchinson JN (1995) Keynote paper: Landslide hazard assessment. In: International Symposium on Landslides, pp 1805–1841
- Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR, pp 448–456
- Jamal M, Mandal S (2016) Monitoring forest dynamics and landslide susceptibility in Mechi-Balason interflaves of Darjiling Himalaya, West Bengal using forest canopy density model (FCDM) and Landslide Susceptibility Index model (LSIM). *Model Earth Syst Environ* 2:1–17. <https://doi.org/10.1007/s40808-016-0243-2>
- Jebur MN, Pradhan B, Tehrany MS (2014) Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. *Remote Sens Environ* 152:150–165. <https://doi.org/10.1016/j.rse.2014.05.013>
- Jia H, Wang Y, Ge D et al (2020) Improved offset tracking for predisaster deformation monitoring of the 2018 Jinsha River landslide (Tibet, China). *Remote Sens Environ* 247:111899. <https://doi.org/10.1016/j.rse.2020.111899>
- Kavzoglu T, Colkesen I (2009) A kernel functions analysis for support vector machines for land cover classification. *Int J Appl Earth Obs Geoinf* 11:352–359. <https://doi.org/10.1016/j.jag.2009.06.002>
- Kavzoglu T, Mather PM (2003) The use of backpropagating artificial neural networks in land cover classification. *Int J Remote Sens* 24:4907–4938. <https://doi.org/10.1080/0143116031000114851>
- Kirschbaum D, Stanley T (2018) Satellite-based assessment of rainfall-triggered landslide hazard for situational awareness. *Earth's Futur* 6:505–523
- Kirschbaum DB, Adler R, Hong Y et al (2010) A global landslide catalog for hazard applications: Method, results, and limitations. *Nat Hazards* 52:561–575. <https://doi.org/10.1007/s11069-009-9401-4>
- Kornejady A, Ownegh M, Bahremand A (2017) Landslide susceptibility assessment using maximum entropy model with two different data sampling methods. *CATENA* 152:144–162. <https://doi.org/10.1016/j.catena.2017.01.010>
- Kumar D, Thakur M, Dubey CS, Shukla DP (2017) Landslide susceptibility mapping & prediction using Support Vector Machine for Mandakini River Basin, Garhwal Himalaya, India. *Geomorphology* 295:115–125. <https://doi.org/10.1016/j.geomorph.2017.06.013>
- Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2:18–22
- Lin L, Lin Q, Wang Y (2017) Landslide susceptibility mapping on a global scale using the method of logistic regression. *Nat Hazards Earth Syst Sci* 17:1411–1424. <https://doi.org/10.5194/nhess-17-1411-2017>
- Liu Z, Gilbert G, Cepeda JM et al (2021) Modelling of shallow landslides with machine learning algorithms. *Geosci Front* 12:385–393. <https://doi.org/10.1016/j.gsf.2020.04.014>
- Mahdadi F, Boumezeur A, Hadji R et al (2018) GIS-based landslide susceptibility assessment using statistical models: a case study from Souk Ahras province. N-E Algeria Arab J Geosci. <https://doi.org/10.1007/s12517-018-3770-5>
- Majka M (2019) naivebayes: High Performance Implementation of the Naive Bayes Algorithm. R package version 0.9.7. <https://CRAN.R-project.org/package=naivebayes>. Accessed 1 Jan 2021
- Mandal K, Saha S, Mandal S (2021) Applying deep learning and benchmark machine learning algorithms for landslide susceptibility modelling in Rorachu river basin of Sikkim Himalaya. *India Geosci Front* 12:101203. <https://doi.org/10.1016/j.gsf.2021.101203>
- Marjanović M, Kovačević M, Bajat B, Voženilek V (2011) Landslide susceptibility assessment using SVM machine learning algorithm. *Eng Geol* 123:225–234. <https://doi.org/10.1016/j.enggeo.2011.09.006>
- Mersha T, Meten M (2020) GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia. *Geoenviron Disasters*. <https://doi.org/10.1186/s40677-020-00155-x>
- Meyer D, Dimitriadou E, Hornik K, et al (2019) Libsvm e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. <https://CRAN.R-project.org/package=e1071>. Accessed 1 Jan 2021
- 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. <https://doi.org/10.1016/j.jseas.2012.10.005>
- Molnar P, Tapponnier P (1975) Cenozoic tectonics of Asia: effects of a continental collision. *Science* 189:419–426
- Mukherjee S (2008) Return of Kosi river induced by Tibet earthquake. *Nat Preced*. <https://doi.org/10.1038/npre.2008.2278.2>
- Nhu VH, Hoang ND, Nguyen H et al (2020a) Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area. *CATENA* 188:104458. <https://doi.org/10.1016/j.catena.2020.104458>
- Nhu VH, Shirzadi A, Shahabi H et al (2020b) Shallow landslide susceptibility mapping by Random Forest base classifier and its ensembles in a Semi-Arid region of Iran. *Forests*. <https://doi.org/10.3390/F111040421>
- Nhu VH, Shirzadi A, Shahabi H et al (2020c) Shallow landslide susceptibility mapping: a comparison between logistic model tree, logistic regression, naive bayes tree, artificial neural network, and support vector machine algorithms. *Int J Environ Res Public Health*. <https://doi.org/10.3390/ijerph17082749>
- Nhu VH, Zandi D, Shahabi H et al (2020d) Comparison of support vector machine, bayesian logistic regression, and alternating decision tree algorithms for shallow landslide susceptibility mapping along a mountainous road in the west of Iran. *Appl Sci*. <https://doi.org/10.3390/app10155047>
- Pandey VK, Sharma KK, Pourghasemi HR, Bandooni SK (2019) Sedimentological characteristics and application of machine learning techniques for landslide susceptibility modelling along the highway corridor Nahan to Rajgarh (Himachal Pradesh). *India Catena* 182:104150. <https://doi.org/10.1016/j.catena.2019.104150>
- Park S, Kim J (2019) Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. *Appl Sci*. <https://doi.org/10.3390/app9050942>
- 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. <https://doi.org/10.1007/s12665-012-1842-5>
- Pasang S, Kubiček P (2020) Landslide susceptibility mapping using statistical methods along the Asian highway, Bhutan. *Geosci* 10:1–26. <https://doi.org/10.3390/geosciences10110430>
- Pawluszek K, Borkowski A (2017) Impact of DEM-derived factors and analytical hierarchy process on landslide susceptibility mapping in the region of Rożnów Lake, Poland. *Nat Hazards* 86:919–952. <https://doi.org/10.1007/s11069-016-2725-y>
- Peethambaran B, Anbalagan R, Kanungo DP et al (2020) A comparative evaluation of supervised machine learning algorithms for township level landslide susceptibility zonation in parts of Indian Himalayas. *CATENA* 195:104751. <https://doi.org/10.1016/j.catena.2020.104751>
- Pham BT, Bui D, Prakash I, Dholakia M (2016a) Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. *J Geomat* 10:71–79
- Pham BT, Pradhan B, Tien Bui D et al (2016b) 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. <https://doi.org/10.1016/j.envsoft.2016.07.005>

- Pham BT, Tien Bui D, Prakash I et al (2017a) A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS. *Environ Earth Sci*. <https://doi.org/10.1007/s12665-017-6689-3>
- Pham BT, Tien Bui D, Prakash I, Dholakia MB (2017b) Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA* 149:52–63. <https://doi.org/10.1016/j.catena.2016.09.007>
- Pham BT, Shirzadi A, Shahabi H et al (2019) Landslide susceptibility assessment by novel hybrid machine learning algorithms. *Sustain* 11:1–25. <https://doi.org/10.3390/su11164386>
- Pham BT, Nguyen-Thoi T, Qi C et al (2020) Coupling RBF neural network with ensemble learning techniques for landslide susceptibility mapping. *CATENA* 195:104805. <https://doi.org/10.1016/j.catena.2020.104805>
- Pradhan B, Lee S (2010) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environ Model Softw* 25:747–759. <https://doi.org/10.1016/j.envsoft.2009.10.016>
- Qi T, Meng X, Qing F et al (2021) Distribution and characteristics of large landslides in a fault zone: a case study of the NE Qinghai-Tibet Plateau. *Geomorphology* 379:107592. <https://doi.org/10.1016/j.geomorph.2021.107592>
- Reichenbach P, Rossi M, Malamud BD et al (2018) A review of statistically-based landslide susceptibility models. *Earth-Science Rev* 180:60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Robin X, Turck N, Hainard A, et al (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. <https://doi.org/10.1186/1471-2105-12-77>
- Rodriguez-Galiano VF, Ghimire B, Rogan J et al (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J Photogramm Remote Sens* 67:93–104. <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Rosi A, Tofani V, Tanteri L et al (2018) The new landslide inventory of Tuscany (Italy) updated with PS-InSAR: geomorphological features and landslide distribution. *Landslides* 15:5–19. <https://doi.org/10.1007/s10346-017-0861-4>
- Saha A, Saha S (2021) Application of statistical probabilistic methods in landslide susceptibility assessment in Kurseong and its surrounding area of Darjeeling Himalayan, India: RS-GIS approach. Springer, Netherlands
- Saha S, Arabameri A, Saha A et al (2021) Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. *Sci Total Environ* 764:142928. <https://doi.org/10.1016/j.scitotenv.2020.142928>
- Sahin EK, Colkesen I, Kavzoglu T (2020) A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. *Geocarto Int* 35:341–363. <https://doi.org/10.1080/10106049.2018.1516248>
- Sajadi P, Singh A, Mukherjee S et al (2020) Drainage network extraction and morphometric analysis in an Iranian basin using integrating factor analysis and geospatial techniques. *Geocarto Int*. <https://doi.org/10.1080/10106049.2020.1750060>
- Sajadi P, Singh A, Mukherjee S et al (2021) Multivariate statistical analysis of relationship between tectonic activity and drainage behavior in Qorveh-Dehgolan basin Kurdistan, Iran. *Geocarto Int* 36:540–562. <https://doi.org/10.1080/10106049.2019.1611948>
- Sameen MI, Pradhan B, Lee S (2020) Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment. *CATENA* 186:104249. <https://doi.org/10.1016/j.catena.2019.104249>
- Shahabi H, Shirzadi A, Ronoud S et al (2021) Flash flood susceptibility mapping using a novel deep learning model based on deep belief network, back propagation and genetic algorithm. *Geosci Front*. <https://doi.org/10.1016/j.gsf.2020.10.007>
- Singaravel S, Suykens J, Geyer P (2018) Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction. *Adv Eng Informatics* 38:81–90. <https://doi.org/10.1016/j.aei.2018.06.004>
- Song J, Wang Y, Fang Z et al (2020) Potential of ensemble learning to improve tree-based classifiers for landslide susceptibility mapping. *IEEE J Sel Top Appl Earth Obs Remote Sens* 13:4642–4662. <https://doi.org/10.1109/JSTARS.2020.3014143>
- Stanley T, Kirschbaum DB, Pascale S, Kapnick S (2020) Extreme precipitation in the Himalayan landslide hotspot. *Adv Glob Chang Res* 69:1087–1111. https://doi.org/10.1007/978-3-030-35798-6_31
- Steger S, Mair V, Kofler C et al (2021) Correlation does not imply geomorphic causation in data-driven landslide susceptibility modeling—benefits of exploring landslide data collection effects. *Sci Total Environ*. <https://doi.org/10.1016/j.scitotenv.2021.145935>
- Steinshouer DW, Qiang J, McCabe PJ, Ryder RT (1999) Maps showing geology, oil and gas fields, and geologic provinces of the Asia Pacific region. US Geol Surv Open-File Rep 97:470F. <https://doi.org/10.3133/ofr97470F>
- Sun X, Chen J, Bao Y et al (2018) Landslide susceptibility mapping using logistic regression analysis along the Jinsha river and its tributaries close to Derong and Deqin County, southwestern China. *ISPRS Int J Geo-Information* 7:1–29. <https://doi.org/10.3390/ijgi7110438>
- Sun D, Wen H, Wang D, Xu J (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. *Geomorphology* 362:107201. <https://doi.org/10.1016/j.geomorph.2020.107201>
- Sun D, Xu J, Wen H, Wang D (2021) Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: a comparison between logistic regression and random forest. *Eng Geol* 281:105972. <https://doi.org/10.1016/j.enggeo.2020.105972>
- Süzen ML (2002) Data driven landslide hazard assessment using geographical information systems and remote sensing. Doctoral dissertation, Middle East Technical University
- Tangestani MH (2009) A comparative study of Dempster-Shafer and fuzzy models for landslide susceptibility mapping using a GIS: an experience from Zagros Mountains, SW Iran. *J Asian Earth Sci* 35:66–73. <https://doi.org/10.1016/j.jseaes.2009.01.002>
- Taylor M, Yin A (2009) Active structures of the Himalayan-Tibetan orogen and their relationships to earthquake distribution, contemporary strain field, and Cenozoic volcanism. *Geosphere* 5:199–214. <https://doi.org/10.1130/GES00217.1>
- Thakur MK, Desamsetti S, Rajesh AN et al (2020) Exploring the rainfall data from satellites to monitor rainfall induced landslides—a case study. *Adv Sp Res* 66:887–894
- Thi Ngo PT, Panahi M, Khosravi K et al (2021) Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geosci Front* 12:505–519. <https://doi.org/10.1016/j.gsf.2020.06.013>
- Tibshirani R (1996) Bias, variance and prediction error for classification rules. Technical Report, Statistics Department, University of Toronto
- 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*. <https://doi.org/10.1155/2012/974638>
- Tien Bui D, Ho TC, Pradhan B et al (2016a) GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environ Earth Sci*. <https://doi.org/10.1007/s12665-016-5919-4>
- Tien Bui D, Tuan TA, Klempe H et al (2016b) 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. <https://doi.org/10.1007/s10346-015-0557-6>
- Tsangaratos P, Iliou I (2016) Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. *Landslides* 13:305–320. <https://doi.org/10.1007/s10346-015-0565-6>
- Ullah M, Aslam M, Ullah MDMI, LazyData T (2019) Package ‘mctest’. <https://doi.org/10.2352/ISSN.2470-1173.2019.7.IRIACV-466>
- Varnes DJ (1984) Landslide hazard zonation: a review of principles and practice. Series: commission on landslides of the IAGC, UNESCO. Nat Hazard 3:61
- Wang Y, Feng L, Li S et al (2020a) A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province. *China Catena* 188:104425. <https://doi.org/10.1016/j.catena.2019.104425>
- Wang Y, Sun D, Wen H et al (2020b) Comparison of random forest model and frequency ratio model for landslide susceptibility mapping (LSM) in Yunyang county (Chongqing, China). *Int J Environ Res Public Health* 17:1–39. <https://doi.org/10.3390/ijerph17124206>

- Wang H, Zhang L, Yin K et al (2021) Landslide identification using machine learning. *Geosci Front* 12:351–364. <https://doi.org/10.1016/j.gsf.2020.02.012>
- Williams G (2011) Data mining with Rattle and R: the art of excavating data for knowledge discovery. Springer Science & Business Media, Cham
- Wu X, Ren F, Niu R (2014) Landslide susceptibility assessment using object mapping units, decision tree, and support vector machine models in the Three Gorges of China. *Environ Earth Sci* 71:4725–4738. <https://doi.org/10.1007/s12665-013-2863-4>
- Wu X, Kumar V, Ross QJ, et al (2008) Top 10 algorithms in data mining. <https://doi.org/10.1201/9781420089653>
- Wubalem A, Meten M (2020) Landslide susceptibility mapping using information value and logistic regression models in Goncha Siso Eneses area, northwestern Ethiopia. *SN Appl Sci* 2:1–19. <https://doi.org/10.1007/s42452-020-2563-0>
- Wubalem A (2020) Landslide susceptibility mapping using statistical methods in Uatza Catchment Area, Northwestern Ethiopia. 1–21. <https://doi.org/10.21203/rs.3.rs-15731/v2>
- Xiao L, Zhang Y, Peng G (2018) Landslide susceptibility assessment using integrated deep learning algorithm along the china-nepal highway. *Sensors (switzerland)*. <https://doi.org/10.3390/s18124436>
- Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. *Geomorphology* 101:572–582. <https://doi.org/10.1016/j.geomorph.2008.02.011>
- Yao T, Xue Y, Chen D et al (2019) Recent third pole's rapid warming accompanies cryospheric melt and water cycle intensification and interactions between monsoon and environment: multidisciplinary approach with observations, modeling, and analysis. *Bull Am Meteorol Soc* 100:423–444. <https://doi.org/10.1175/BAMS-D-17-0057.1>
- Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Eng Geol* 79:251–266. <https://doi.org/10.1016/j.enggeo.2005.02.002>
- Yi Y, Zhang Z, Zhang W et al (2020) Landslide susceptibility mapping using multiscale sampling strategy and convolutional neural network: a case study in Jiuzhaigou region. *CATENA* 195:104851. <https://doi.org/10.1016/j.catena.2020.104851>
- Yilmaz I (2009) Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat-Turkey). *Comput Geosci* 35:1125–1138. <https://doi.org/10.1016/j.cageo.2008.08.007>
- You Q, Fraedrich K, Ren G et al (2013) Variability of temperature in the Tibetan Plateau based on homogenized surface stations and reanalysis data. *Int J Climatol* 33:1337–1347
- Youssef AM, Pourghasemi HR (2021) Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geosci Front* 12:639–655. <https://doi.org/10.1016/j.gsf.2020.05.010>
- Yu C, Chen J (2020) Landslide susceptibility mapping using the slope unit for southeastern Helong city, Jilin province, China: A comparison of ANN and SVM. *Symmetry (basel)* 12:1–23. <https://doi.org/10.3390/sym12061047>
- Zare M, Pourghasemi HR, Vafakhah M, Pradhan B (2013) Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. *Arab J Geosci* 6:2873–2888. <https://doi.org/10.1007/s12517-012-0610-x>
- Zhang J, Yun L, Zhang B et al (2020) Deformation at the Easternmost Altyn Tagh Fault: constraints on the growth of the Northern Qinghai-Tibetan Plateau. *Acta Geol Sin* 94:988–1006. <https://doi.org/10.1111/1755-6724.14555>
- Zhao B, Wang Y, Luo Y et al (2019a) Large landslides at the northeastern margin of the Bayan Har Block, Tibetan Plateau. *China R Soc Open Sci*. <https://doi.org/10.1098/rsos.180844>
- Zhao Y, Wang R, Jiang Y et al (2019b) GIS-based logistic regression for rainfall-induced landslide susceptibility mapping under different grid sizes in Yueqing. *Southeastern China Eng Geol* 259:105147. <https://doi.org/10.1016/j.enggeo.2019.105147>
- Zhao B, Zhao X, Zeng L et al (2021) The mechanisms of complex morphological features of a prehistorical landslide on the eastern margin of the Qinghai-Tibetan Plateau. *Bull Eng Geol Environ* 80:3423–3437. <https://doi.org/10.1007/s10064-021-02114-8>
- Zhu AX, Wang R, Qiao J et al (2014) An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic. *Geomorphology* 214:128–138. <https://doi.org/10.1016/j.geomorph.2014.02.003>
- Zhu AX, Miao Y, Wang R et al (2018) A comparative study of an expert knowledge-based model and two data-driven models for landslide susceptibility mapping. *CATENA* 166:317–327. <https://doi.org/10.1016/j.catena.2018.04.003>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
