Landslide Spatial Modelling Using Unsupervised Factor Optimisation and Regularised Greedy Forests

Maher Ibrahim Sameen[1], Raju Sarkar[2], Biswajeet Pradhan[1,3*], Dowchu Drukpa[4], Abdullah M. Alamri[5], Hyuck-Jin Park[3]

[1]Center for Advanced Modeling and Geospatial System (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, CB11.06.106, Building 11, 81 Broadway, Ultimo, NSW 2007, Australia; maher.alzuhairi@uts.edu.au; Biswajeet.Pradhan@uts.edu.au

[2]Center for Disaster Risk Reduction and Community Development Studies, College of Science and Technology, Royal University of Bhutan, Rinchending, Phuentsholing, Bhutan, Email. rajusarkar.cst@rub.edu.bt

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

[4]Seismology and Geophysics Division, Department of Geology & Mines, Ministry of Economic Affairs, Thimphu, Bhutan; Email. ddukpa@moea.gov.bt

[5]Dept. of Geology & Geophysics, College of Science, King Saud Univ., P.O. Box 2455, Riyadh 11451, Saudi Arabia

*Corresponding author: Biswajeet.Pradhan@uts.edu.au

Abstract

This study evaluates the contribution of an unsupervised factor optimisation based on sparse autoencoders (SAEs) to spatial landslide modelling with regularised greedy forests (RGFs). A total of 952 landslides were identified by field surveys, equally divided and used for training and testing of the proposed model. Ten conditioning factors related to landslides, including geo-morphometrical (i.e. altitude, slope, aspect, curvature, slope length, topographic wetness index and sediment transport index) and geo-environmental (i.e. lithology, nearness to roads and nearness to streams), were used to investigate the spatial relationships between the variables and landslides. [1]The steps of the modelling were twofold. First, the factors were

---

Authorship statement: M.I.S. contributed to the methodology, formal analysis, investigation, visualisation, validation, and writing of the manuscript; R.S. contributed to the data acquisition, review and editing of the manuscript; B.P. contributed to the supervision, resources, methodology, project administration, funding support, visualisation, review and editing of the manuscript. D.P. contributed to the data acquisition and data curation. H-J.P. contributed to the editing of the manuscript.

optimised by SAE to reduce information redundancy and correlation in the data. Second, RGF was used to create landslide susceptibility maps with the optimised feature representations. The area under the receiver operating characteristic curve (AUROC) was used to assess the predictive ability of the proposed models. Experimental results show that the proposed SAE–RGF outperforms the RGF and random forest (RF) models in terms of prediction rate and is less sensitive to overfitting and underfitting. The highest prediction rate (AUROC = 0.892) was obtained with only seven features by the SAE–RGF model, which is better than the two other methods (RGF and RF). The unsupervised factor optimisation approach not only reduces computation time but also improves the prediction accuracy of tree-based models, including RGF. The generated landslide susceptibility maps can be implemented to mitigate landslide hazards and to designate land use by stakeholders (e.g. planners and engineers).

Keywords: landslide susceptibility; regularised greedy forests; unsupervised factor optimisation; GIS; Chukha Dzongkhag; Bhutan

1. Introduction

Landslide spatial modelling is a common study that offers helpful information to decision-makers and planners (Kocaman and Gokceoglu, 2019; Comert et al. 2019; Luo et al. 2019; Wang, Fang and Hong 2019; Ozturk et al. 2019). The outcomes of this process (e.g. susceptibility maps, the significance of conditioning factors) are helpful in assessing landslide hazards and in reducing their impact on human lives and infrastructure, such as road networks and agricultural systems. Conventional techniques for assessment of landslide susceptibility are based on field surveys and aerial photograph interpretation. These techniques have been proven to be time consuming and expensive in most cases. Alternative methods are based on automated algorithms that can learn from data on historical landslide events and generalise to other areas where no prior information is accessible. Various algorithms and modelling techniques, which have improved significantly over the last decades, are available. They have been developed to learn from a limited amount of data and to be generalised to areas other than the training area (s). Statistical methods, including bivariate (e.g. frequency ratio, certainty factors, statistical index and entropy index) (Liu and Duan 2018, Shirani et al. 2018) and multivariate (e.g. logistic regression) (Sun et al. 2018, Polykretis and Chalkias 2018), have been widely used by many researchers. Although these algorithms can learn from historical landslide data, they are sensitive to the selection of factors that contribute to landslides.

Algorithms that belong to machine learning family, such as artificial neural networks (Pradhan and Lee 2010, Zare et al. 2013, Xiao et al. 2018; Can et al. 2019), support vector machines (Hong et al. 2017), tree-based models (e.g. decision tree, extra trees, random forest) (Chu et al. 2019, Dou et al. 2019), fuzzy logic (Peethambaran et al. 2019) and neuro-fuzzy systems (Dehnavi et al. 2015, Aghdam et al. 2016, Polykretis et al. 2017, Chen et al. 2019; Ozer et al. 2018, 2019), have been introduced and applied to landslide susceptibility analysis to complement the limitations of statistical methods.

In statistical modelling, historical landslide data provide a clue to the selection of a stochastic model that acts as an abstraction for creating landslide predictions. In machine learning, however, data drive the selection of an algorithm to predict future landslides from the input data. That is, statistical models provide distributional assumptions about the nature of the true underlying relationships, whereas machine learning requires less or not a priori belief. Consequently, machine learning algorithms can discover novel relationships in the data, whereas statistical models can only find such relationships when guided by a human. For landslide susceptibility modelling, machine learning will be highly efficient if the input data are incomplete or difficult to understand in their raw form. Nevertheless, machine learning algorithms can overfit or find spurious correlations, which should be carefully designed and analysed to avoid such problems.

The recent trend in landslide susceptibility analysis using machine learning is hybrid models that combine the benefits of two or more machine learning models for good reasons. The combination of several algorithms into a single model is crucial because it offers higher generalisation ability than a single algorithm by reducing variance and bias or improving prediction. Significant research has developed and demonstrated the superiority of hybrid models to single models for landslide susceptibility (Huang and Zhao 2018, Pham et al. 2019). Hybrid models can be constructed by ensembles (Kadavi et al. 2018) or by integrating algorithms that do not belong to the same family or that aim at different processing stages (i.e. pre-processing, feature selection/extraction, optimisation, modelling). Examples of hybrid models recently developed for landslide susceptibility mapping include bivariate weights of evidence with multivariate logistic regression and RF (Chen et al. 2019); integrated ensemble fractal dimension with kernel logistic regression (Zhang et al. 2019); entropy and rotation forest-based credal decision tree classifier (He et al. 2019) and meta-optimisation of an adaptive neuro-fuzzy inference system with a grey wolf optimiser and biogeography-based optimisation algorithms (Jaafari et al. 2019).

Several other studies also include the development of hybrid susceptibility models using tree-based methods. Kutlug Sahin and Colkesen (2019) examined decision tree-based ensembles models such as canonical correlation forest and rotation forest. The former method outperforms the other on different ensemble techniques, including AdaBoost and bagging. Random forest also found superior to decision trees in (Dou et al. 2019). Kornejady et al. (2019) created a hybrid model which combined random forest and frequency ratio for the evaluation and efficiency of landslide susceptibility and found such models have a good performance (AUC value of 0.831). In another study. Nguyen et al. (2019) found that tree-based models such as best first decision trees-based rotation forest are superior to models created using an adaptive neuro-fuzzy inference system and artificial neural networks optimized by particle swarm optimization.

The selection of individual algorithms that form a hybrid model is often subjective. However, certain algorithm characteristics can be used to determine the elements of a hybrid model. For example, category (tree-based, probabilistic, neural networks), predictive ability based on previous work and computational performance are important properties to consider when selecting algorithms for a hybrid model. Tree-based models, such as decision trees, extra trees, RF and boosted trees, have demonstrated good performance, as presented in recent studies (Lee et al. 2018, Song et al. 2019, Meneses et al. 2019). Preparing high-quality spatial data and landslide inventories is also essential to enhance the performance of landslide susceptibility models (Huang and Zhao 2018, Soma et al. 2019).

This research aims at improving the performance of tree-based models such as regularised greedy forests (RGF) and random forest (RF) for landslide susceptibility modelling. To achieve this aim, an integrated model namely SAE-RGF which combines sparse auto-encoders as an unsupervised factor optimisation and RGF was developed and evaluated in Chukha Dzongkhag, Bhutan. To the best of our knowledge, RGFs have not yet been applied to landslide susceptibility mapping. In this sense, this study contributes to evaluating the effect of feature compression on susceptibility mapping. The proposed model was compared with RGF (without optimization) and other tree models such as RF.

2. Description of the study area

For the case study, Chukha Dzongkhag is chosen to evaluate the models suggested (Figure 1). This area lies between longitudes 89° 15′–89° 49′ and latitudes 26° 44′–27° 18′ in the southwestern part of Bhutan. It covers approximately 1,879.5 km^2 and has a population of

88,342 as of 2015. Its elevation ranges between 0 m to 4,413 m above mean sea level, with a mean elevation of 1,905 m. The slope angles vary from 0° to 89°.

Chukha Dzongkhag is in the subtropical and temperate climatic zones. It experiences high annual rainfall (the highest being 4,000–6,000 mm) and nearly regular heavy rains up to 800 mm.day^{-1}, mostly during the southwestern monsoon between June and September. Consequently, the area is highly vulnerable to landslides, particularly during the rainy season. Most landslides occur alongside the Phuentsholing–Thimphu dual carriageway, a lifeline infrastructure that links the capital Thimphu with neighbouring nations. The vicinity is also characterised by closely fractured and weathered rocks, such as phyllites, slates and schists, which contain excessive quantities of clay minerals (Kuenza et al. 2010). The area contains steep slope terrain, which makes it highly at risk of slope failures brought by rainfall and associated disasters due to several road cuttings (Kuenza et al. 2010). Landslides frequently block the highway, thereby resulting in huge economic losses.

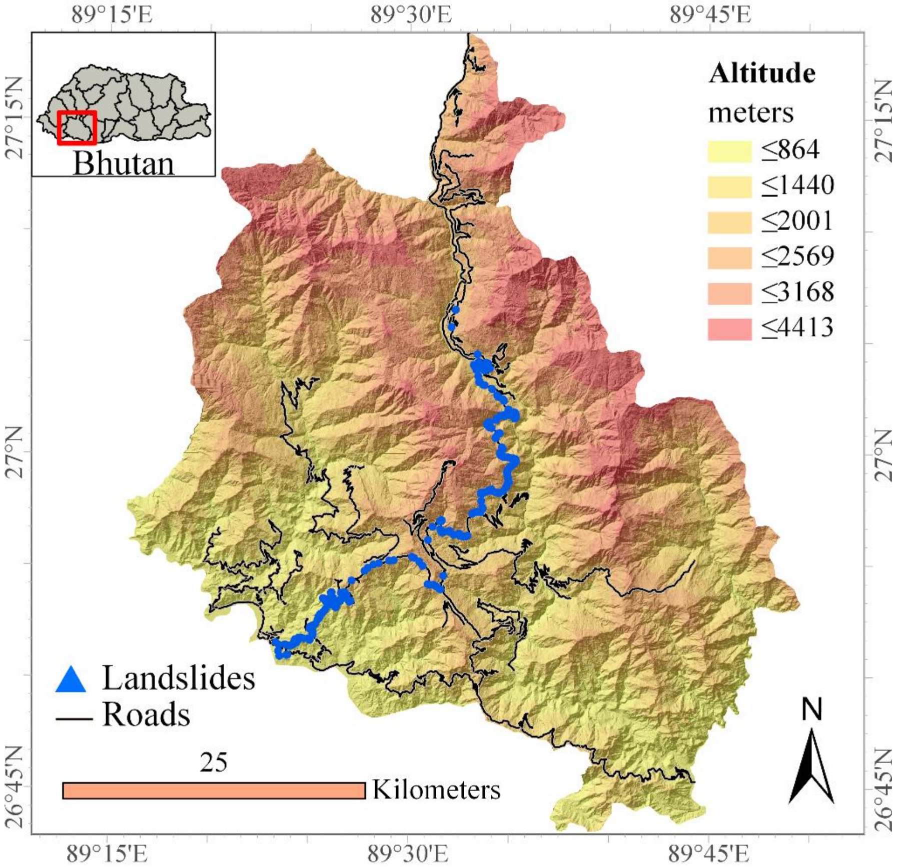


Figure 1: Location of the study area and landslide inventories provided in this research.

3. Methodology

3.1 Landslide inventories

For efficient mapping of landslide susceptibility, the first step used to train and validate machine learning methods is often regarded as a landslide inventory map. A standard landslide inventory map includes historic landslide records that consist of the location and areal coverage, prevalence facts, mass move type and landslide phenomenon volume in an area.

Field investigation and analysis of historical aerial photos and satellite images are two common methods used to prepare landslide inventory maps. In this study, 952 landslides were mapped, verified, and included in a spatial database. Figure 2 shows some photographs taken in the study area. Nearly all the landslides were caused by precipitation and occurred within less than 50 m from the Phuentsholing–Thimphu highway. The depths of the landslides in the study region range from several decimetres to a few metres based on visible and on-site intensity measurements. Landslides were mapped as single points (Gariano et al. 2018). The dataset was randomly divided into three subsets for training (70%, 666), validation (10%, 95) and final testing (20%, 191). The training dataset was used to train the proposed models, whereas the validation dataset was used to optimise the parameters of the same models. Finally, using the test datasets, the models were evaluated and compared with each other.

3.2 Conditioning factors related to landslides

In relation to a landslide inventory map, the modelling of landslide susceptibility requires conditioning factors that are representative, reliable and readily obtainable. These factors can be determined by field surveys (Oh and Pradhan, 2011) and inventory map analysis, landslide types and characteristics of the study area. In the present study, 10 landslide conditioning factors, including geo-morphometrical and geo-environmental factors, were selected based on the factors that were most commonly used in previous studies and the characteristics of the study area.

Figure 2: Sample of field photographs taken in the study area.

3.2.1 Geo-morphometrical factors

Seven geo-morphometrical factors were obtained with a resolution of 10 m from a digital elevation model generated from topographical maps. Subsequently, a raster resolution of 10 m was used to derive the conditioning factors. The factors were altitude, slope, aspect, curvature, slope length, topographic wetness index (TWI) and sediment transport index (STI). They were extracted using ArcGIS Pro 2.3. The factors with continuous values were reclassified into categorical classes using the Jenks natural breaks optimisation method (Hong et al. 2016) available in ArcGIS Pro 2.3, as recommended and defined by Hung et al. (2016).

a) **Altitude**: Elevated areas impact loading on the slope; elevation is therefore an important factor in landslide modelling. High-altitude areas increase the possibility of landslides, particularly if the sliding plain has an orientation close to an open excavation (Walker and Shiels, 2012). In this research, the elevation map was labelled into six classes (Figure 3).

b) **Slope**: Slope is a major factor in any analysis of landslide susceptibility and has often been used in past research (Hong et al. 2018, Lee et al. 2018, Sameen et al. 2018). Slope is an important topographical parameter, and landslide frequency is often high on steep slopes. The slope map was labelled into six classes (Figure 3).

c) **Aspect**: Slope aspect uses slope path and affects daylight, wind and precipitation exposure. Aspect also impacts vegetation and soil-related factors indirectly, such as vegetation cover, soil thickness and moisture. Slope aspect is therefore regarded as an important parameter in the evaluation of landslide susceptibility (Hong et al. 2017). In this study, aspect was divided into nine classes: flat, north-, northeast-, east-, southeast-, south-, southwest-, west- and northwest-facing classes (Figure 3).

d) **Curvature**: Is the curvature of a line formed by the intersection of a random plane with the terrain surface (Youssef et al. 2015). The curvature value can be positive or negative. A positive curvature represents an upwardly convex surface, whereas a negative curvature represents an upwardly concave surface on a point. If its value is near zero, the curvature can also have a flat shape. Curvature plays a key role in landslide modelling and in altering landform characteristics (Mandal and Maiti 2015). A convex surface immediately drains moisture, whereas for a long period a concave surface holds moisture. In this study, a curvature map was used after reclassifying it into six classes (Figure 3).

e) **Slope length**: This study considers slope length a landslide conditioning factor because it increases the capability of erosive agents to displace and transport materials downslope (Gomez and Kavzoglu 2005). Slope length was calculated using the digital elevation model and prepared for the modelling process with six classes (Figure 3).

f) **TWI**: This parameter is a hydrological factor that contributes to landslide occurrence; it combines local upslope contributing area and slope (Gallant 2000). High TWI values indicate low landslide occurrence probability. In this study, TWI was calculated using the following equation:

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{tan\beta}\right), \tag{1}$$

where α is the cumulative upslope area (per unit contour length), and β is the angle of slope at the calculation point. The TWI map for the study region was categorised into six classes (Figure 3).

g) **STI**: This parameter indicates the amount of sediment transportation through overland flow and is based mainly on the erosion of catchment evolution theories and transportation capacity that restricts sediment flux. In this study, the following equation was used to calculate STI. The generated values were classified into six classes (Figure 3):

$$\mathrm{STI} = \left(\frac{A_s}{22.13}\right)^{0.6} . \left(\frac{sin\beta}{0.0896}\right)^{1.3}, \tag{2}$$

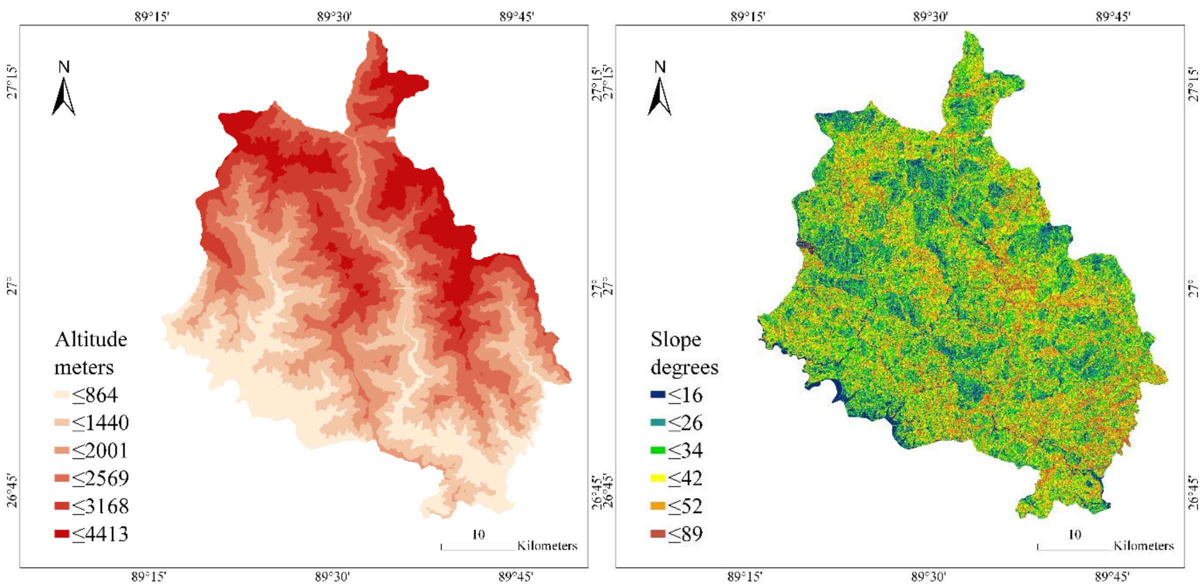
where A_s is the specific catchment area (m^2/m), and β is the slope gradient.

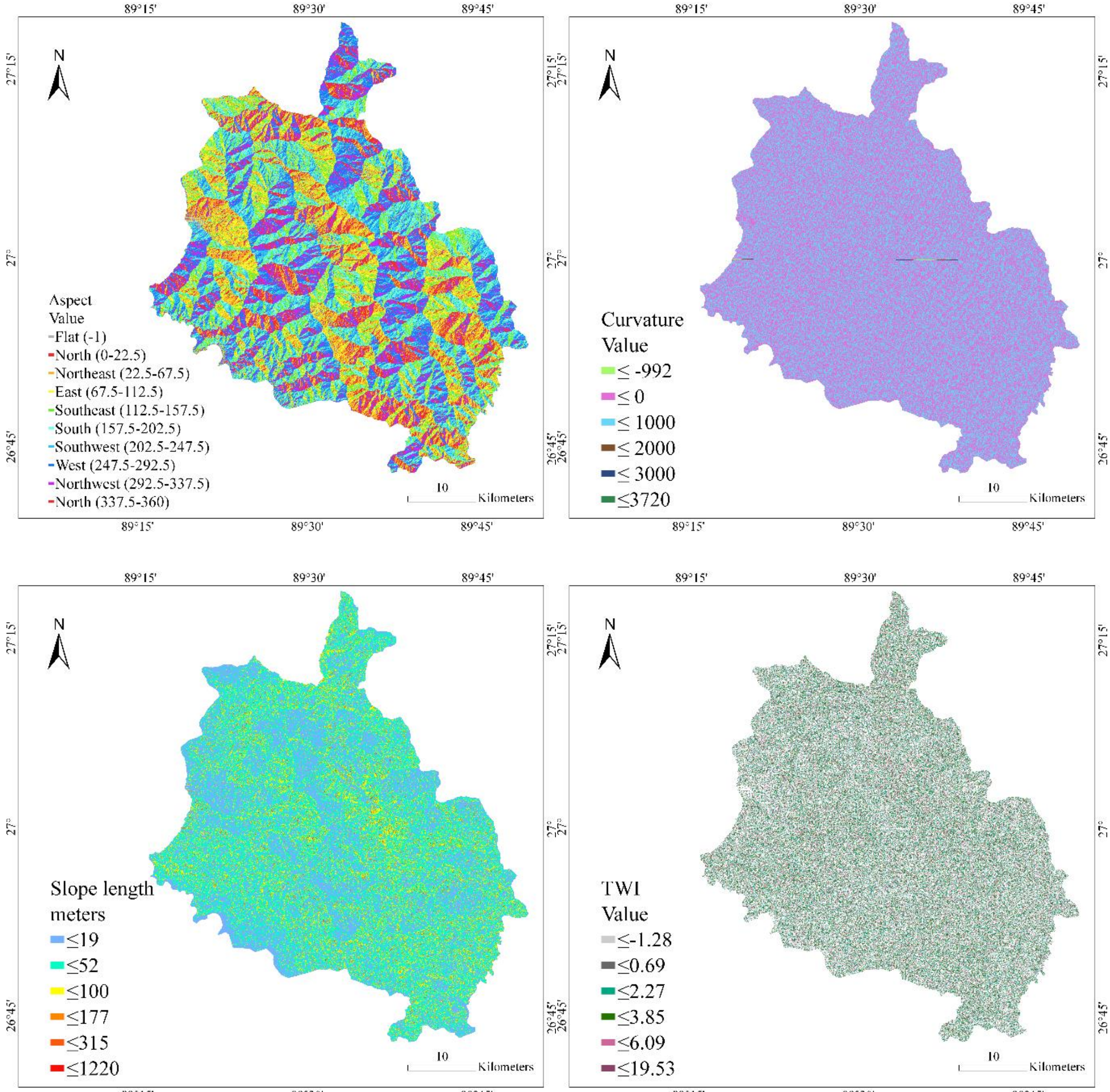
3.2.2 Geo-environmental factors

Three geo-environmental elements, namely, lithology, proximity to roads and proximity to streams, were used in this study, as explained in the following subsection.

a) **Lithology**: The area geologically belongs to the Lesser Himalayan formation. It includes sedimentary and low-grade metamorphic rocks. It consists primary of metasedimentary rocks like phyllite, schist, quartzite, and limestone that are tectonically active. The north part of the area is comprised of the Higher Himalayan crystalline rocks such as garnetiferous mica-shist, quartzite, and gneiss. Lithology is important in the analysis of landslide susceptibility because soft and weathered rocks are more vulnerable to landslides than hard unjointed rocks. The study area, Chukha Dzongkhag, is made up of various types of lithological units (Figure 3).

b–c) **Proximity to roads and streams**: Anthropogenic factors, including proximity to roads and streams, are regularly utilised in landslide susceptibility evaluation. Shallow to deep excavations, application of foreign loads and eviction of vegetative cover are common actions during construction along highways and roads. In addition, the intermittent flow regime of a hydrological community and gullies encompasses erosive and saturation processes. Subsequently, pore water pressure can be increased, which may lead to landslides in regions that adjoin drainage channels (Figure 3). Land use and land cover were not considered in this research because all the landslide points fall into one class (forest area) and no variance was in the data.




89°15'
89°30'
89°45'
27°15'
27°
26°45'
N
Aspect
Value
Flat (-1)
North (0-22.5)
Northeast (22.5-67.5)
East (67.5-112.5)
Southeast (112.5-157.5)
South (157.5-202.5)
Southwest (202.5-247.5)
West (247.5-292.5)
Northwest (292.5-337.5)
North (337.5-360)
10
Kilometers
Curvature
Value
≤ -992
≤ 0
≤ 1000
≤ 2000
≤ 3000
≤3720
Slope length
meters
≤19
≤52
≤100
≤177
≤315
≤1220
TWI
Value
≤-1.28
≤0.69
≤2.27
≤3.85
≤6.09
≤19.53

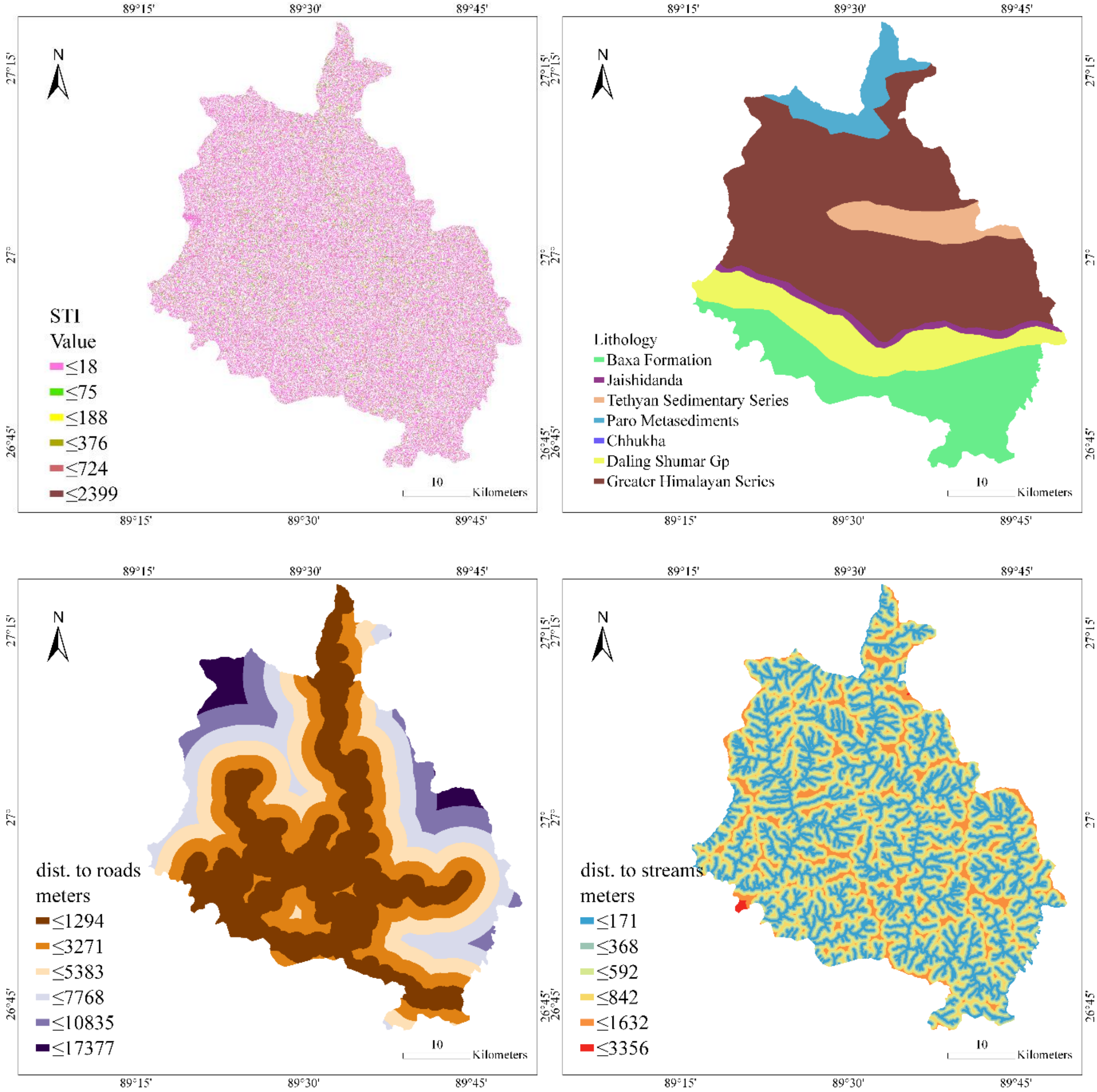


Figure 3: Landslide conditioning factors.

3.3 Proposed models

A modelling approach based on two machine learning algorithms, namely, SAEs and RGFs, was developed for landslide susceptibility assessment in Chukha Dzongkhag. The flowchart of this approach is presented in Figure 4. The landslide presence-absence samples were created after collecting and preparing the landslide inventory map, the spatial digital elevation model, and thematic layers. The landslide inventory samples were counted and used to randomly generate the absence samples. The final data combined the landslide presence and absence samples with a defined label (1 and 0, respectively) for each sample. Ten landslide conditioning elements were prepared from a spatial database. The values of the landslide conditioning elements at each sample location were utilised, and the derived information was prepared using a Microsoft Excel sheet.

The independent variables in the data were scaled (zero mean, unit variance) to improve the training process of SAE (only applied to the factors with continuous values). The dependent variable was converted with one-hot encoding. The data were then categorised into three subsets: for training (70%), validation (10%) and testing (20%). The SAE model was trained in an unsupervised manner, and a set of new features was generated. These new features were used to train the RGF model. In this study, the validation of the proposed models was based on a well-known area under the receiver operating characteristic curve (AUROC). Sensitivity analysis was also considered to assess the consequences of dimensionality reduction on the RGF model.

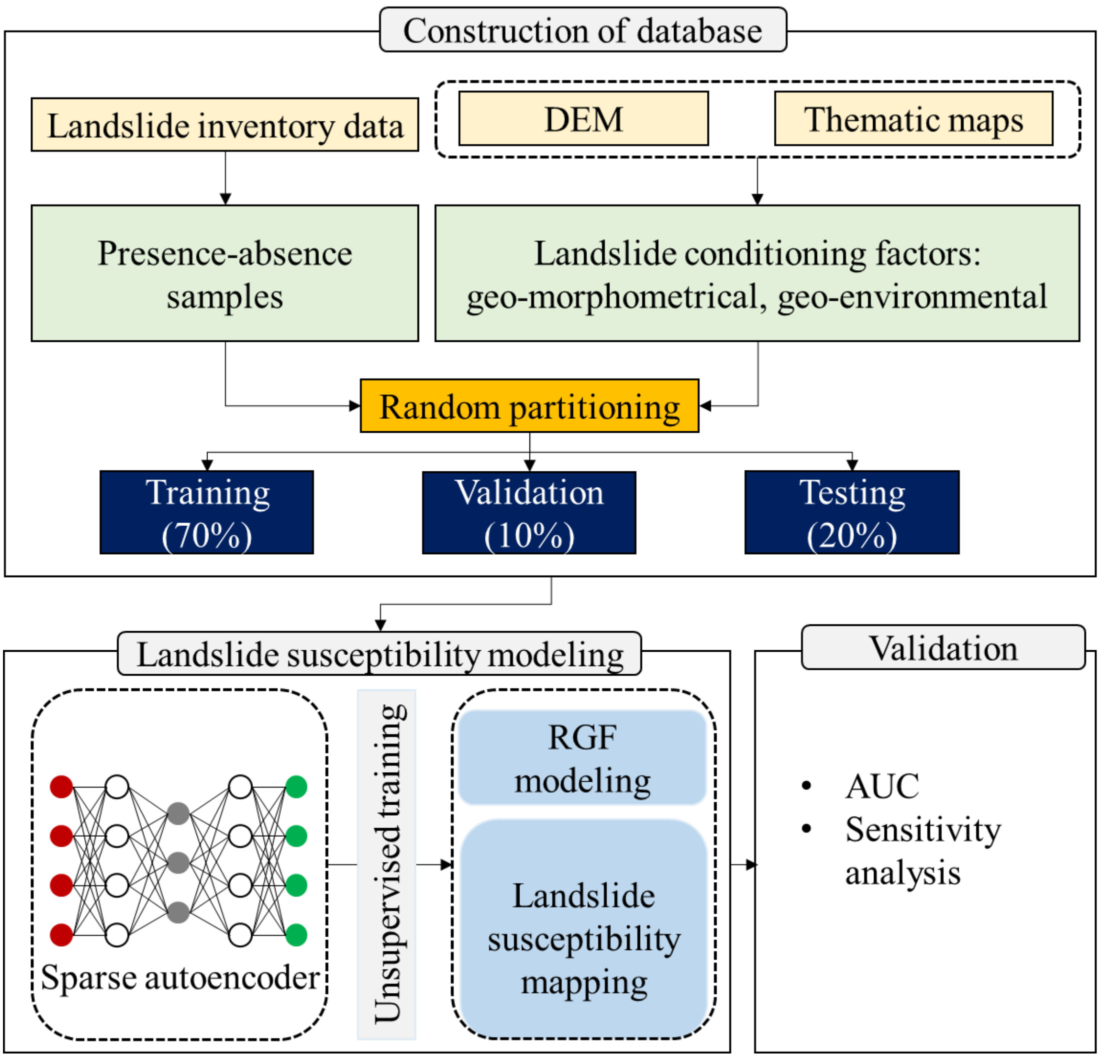


Figure 4: Flowchart of the proposed methodology.

3.3.1 Unsupervised factor optimisation by SAEs

Autoencoders are neural networks that can be used to learn features from a dataset in an unsupervised manner (Hinton and Salakhutdinov 2006). They can be shallow (i.e. with one hidden layer) or deep (i.e. with two or more hidden layers). The addition of more hidden layers depends on the complexity and amount of data. The proposed SAE model structure, as presented in Figure 4, has three hidden layers in addition to the input and output layers. The input $x \in \mathfrak{R}^N$ is mapped into a hidden representation $h^{(1)} \in \mathfrak{R}^N$ using $h^{(1)} = f_{\theta_1}(W^1 x + b^1)$, which is then used to learn another hidden illustration $h^{(2)} \in \mathfrak{R}^N$by $h^{(2)} = f_{\theta_2}(h^{(1)} + b^2)$. The output of this illustration is used to learn a third hidden illustration $h^{(3)} \in \mathfrak{R}^N$ by $h^{(3)} = f_{\theta_3}(h^{(2)} + b^3)$. The hidden representation $h^{(3)}$ is then utilised to regenerate an approximation $\hat{x}$ of the input. The hidden layer $h^{(2)}$ is considered the new feature representation of the input data. The dimension of the input layer x is 10, and each $h^{(1)}$ and $h^{(3)}$ has 14 hidden nodes, whereas the new feature representation $h^{(2)}$ has only 7 nodes. Therefore, the proposed SAE learns compressed representation, which can reduce the computational time of the RGF model. A sparsity constraint of L1 regularisation (10e-5) was enforced on the three hidden layers of the model to avoid overfitting in the model. The rectified linear unit activation function was used for the hidden layers (encoder), whereas the sigmoid function was used for the output layer (decoder).

The optimal parameters of the SAE model were selected by minimising the binary cross-entropy cost function using a backpropagation algorithm and stochastic gradient descent (i.e. Adamax). The model was trained for 1,000 epochs with a batch size of 32 and a learning rate of 0.002. The training was stopped when validation accuracy stopped improving (patience = 20 epochs). After the learning process, the SAE model learned a nonlinear function that mapped an input vector $x \in \mathfrak{R}^N$ into a new feature representation $h^{(2)}$.

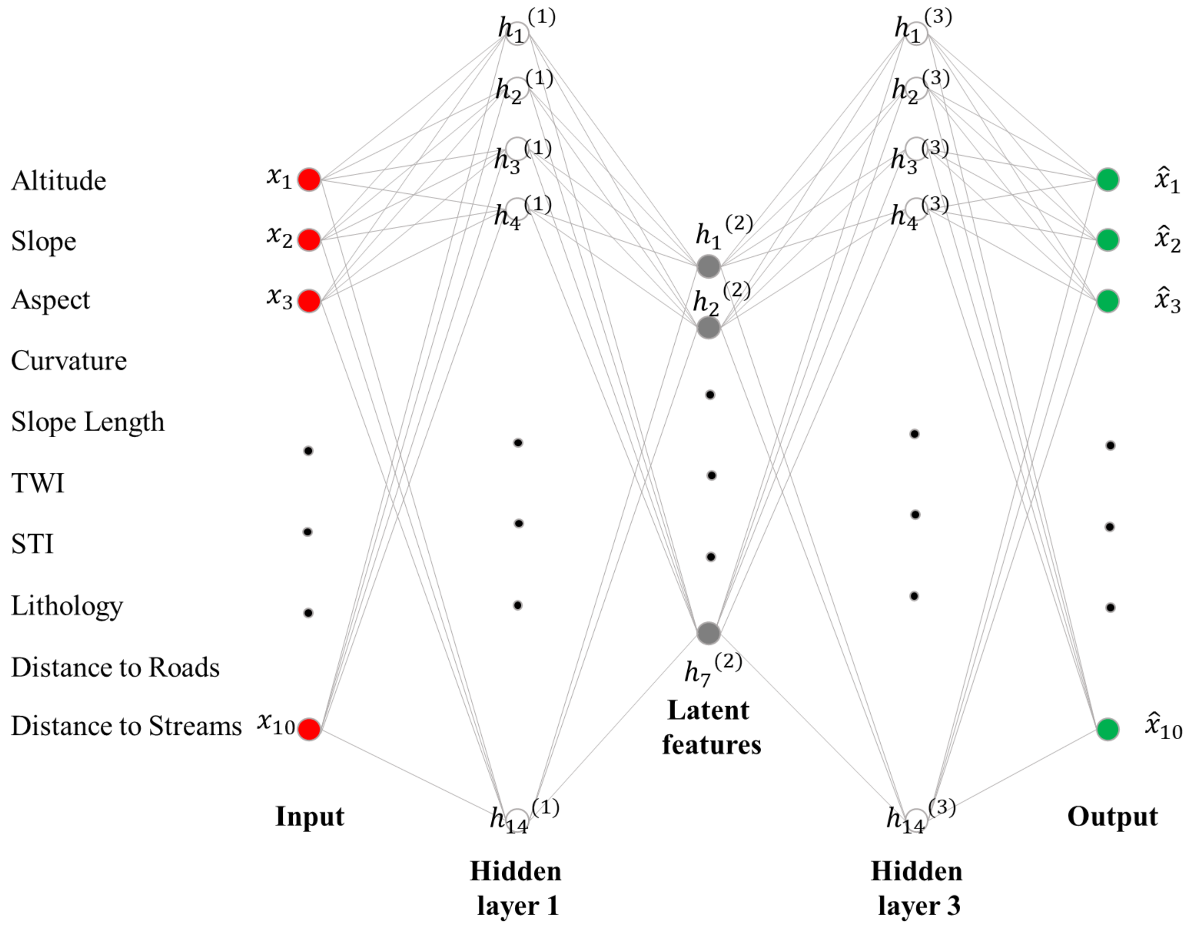


Figure 5: Structure of the proposed SAE model.

3.3.2 RGFs

Decision trees are also commonly used models in landslide susceptibility analysis and other applications. These models have a tree-like structure with terminal and nonterminal nodes. The former presents the decision outcomes, whereas the latter presents the attribute tests. The major advantages of these models are easy implementation and graphical presentation of the model structure. However, these models are susceptible to data noise and can overfit the training data if inaccurately validated. Researchers have proposed many improved versions of tree-based models, including boosted trees and their ensembles, such as RF and RGF, to overcome the limitations of decision trees. RGF combines several boosted trees and additively forms a forest as a single predictive model (Johnson and Zhang 2014). In boosted decision trees, the trees are locally optimised; in RGF, the trees are globally optimised. RGF utilises a tree structure because it uses fully corrective regularised steps. RGF is also faster and frequently more accurate than boosted trees, particularly for regression problems.

3.4 Model evaluation methods

AUROC was used to assess the predictive capability of the proposed model and compare it with other models. AUROC is widely adopted in landslide susceptibility studies (Pradhan et al. 2010, Shirzadi et al. 2017). The receiver operating characteristic (ROC) curve is constructed based on the sensitivity (the true positive rate) and specificity (the false-negative rate). AUROC is calculated using the following expression:

$$\mathrm{AUROC} = \frac{\sum \mathrm{TP} + \sum \mathrm{TN}}{\mathrm{P} + \mathrm{N}}, \tag{3}$$

where TP is the true positives, and TN is the true negative. A high AUROC value indicates an accurate model prediction. In general, values of 0.5–0.6, 0.6–0.7, 0.7–0.8, 0.8–0.9 and 0.9–1 indicate insufficient, moderate, good, very good and excellent performance, respectively (Bui et al. 2014).

4. Results and discussion

This section presents the major findings of the research and discusses the factor optimisation analysis, application of RGF and other tree-based landslide susceptibility modelling approaches.

4.1 Factor optimisation analysis

Table 1 lists the results of the unsupervised factor optimisation to RGF-based landslide susceptibility modelling. The feature space of the input data had a dimensionality of 10, which was then reduced to lower dimensions (9–2) by the proposed SAE model. A number of observations can be made from the findings of this experiment. The summary is presented in Table 1. Firstly, a high linear relationship exists between the SAE reconstruction errors on the training and testing datasets (R^2 = 0.991), which explains the efficiency of the SAE model in learning low representations of the input data without substantial overfitting and underfitting. The lowest reconstruction errors were 0.522 and 0.517 for the training and testing datasets, respectively, when the number of input features was reduced to eight. The largest reconstruction errors were 0.569 and 0.566 for the training and testing datasets, respectively, when the number of features was reduced to only two. These results suggest that reducing the dimensionality of the input data by using the SAE model requires careful analysis of the number of new representation features. Secondly, the success and prediction rates of the RGF model that was trained on the new representation features learned by the SAE indicate a linear relationship of R^2 = 0.85 between the reconstruction errors and the associated success rates. A considerably lower R^2 (0.53) was observed between the reconstruction errors and the associated prediction rates. These results indicate that low reconstruction errors by the SAE do not necessarily yield high success/prediction rates on the RGF model. The best success (AUROC = 0.931) and prediction (AUROC = 0.892) rates for the RGF model were observed when the input features were reduced to seven. Reducing the dimensionality of the input features into only two degraded the success and prediction rates of the RGF model by 17% and 16%, respectively. Furthermore, transforming the input features into a new set of feature representations with the same size as the input data (10 features) did not yield the best success and prediction rates for the RGF model.

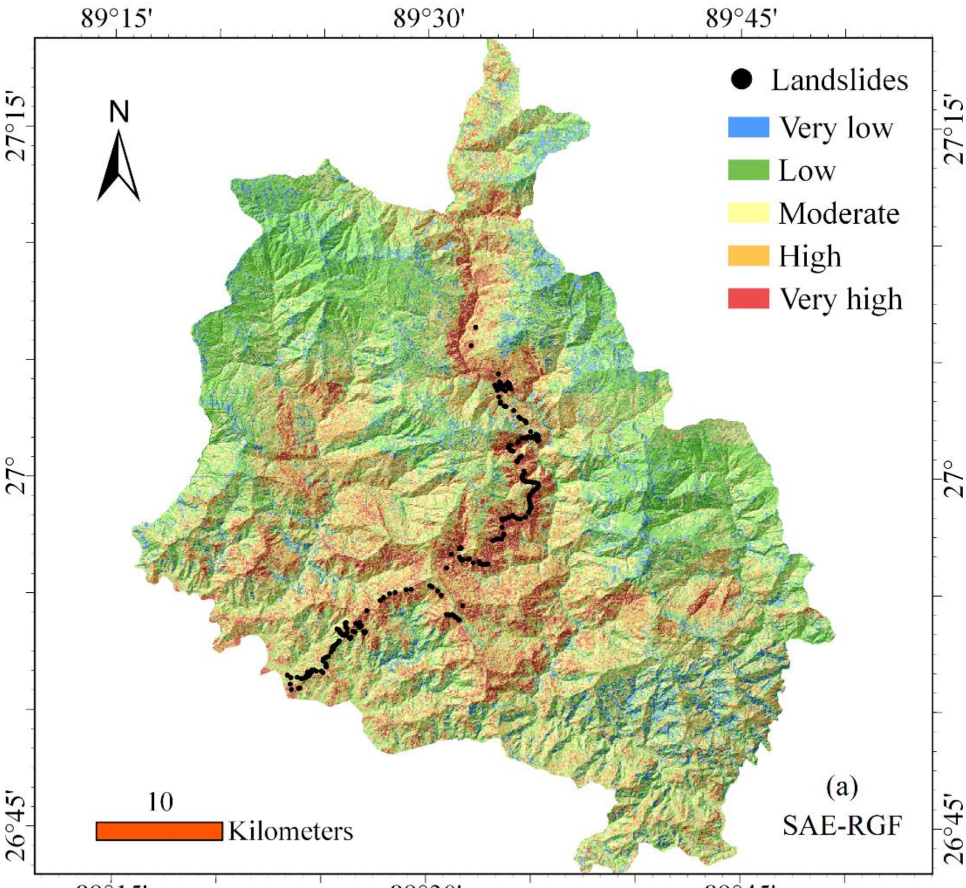
Table 1: Reconstruction errors estimated for the SAE and the associated success/prediction rates of the RGF model based on the training and testing datasets.

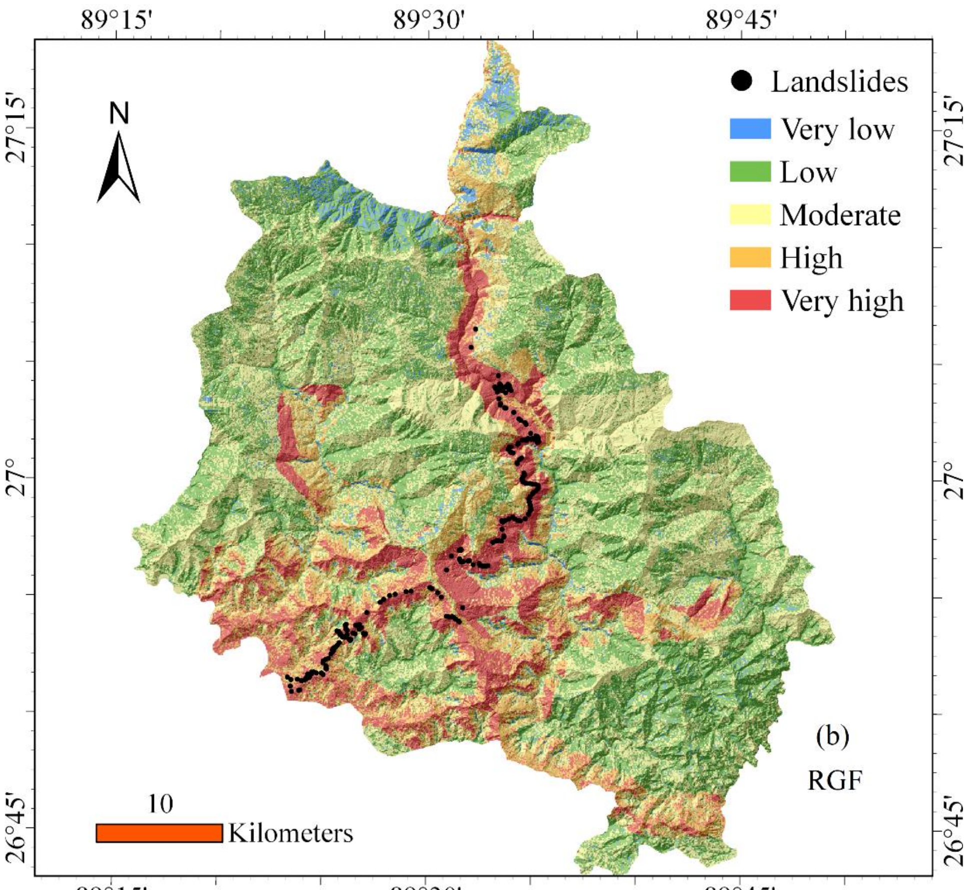
Number of compressed features	**Reconstruction error (training data)**	**Reconstruction error (testing data)**	**SAE–RGF success rate**	**SAE–RGF prediction rate**
2	0.569	0.566	0.765	0.736
3	0.553	0.548	0.819	0.715
4	0.545	0.539	0.880	0.829
5	0.532	0.529	0.904	0.826
6	0.541	0.538	0.896	0.763
7	0.540	0.534	**0.931**	**0.892**
8	**0.522**	**0.517**	0.928	0.848
9	0.533	0.527	0.919	0.889
10	0.534	0.528	0.908	0.830

4.2 Application of SAE–RGF, RGF and RF

Three susceptibility maps were generated for the study area using the proposed SAE–RGF model and the RGF model without unsupervised factor optimisation, and for comparison, with another tree-based model (RF) (Figure 6). The susceptibility maps were recategorized into five classes, namely, *very low*, *low*, *moderate*, *high* and *very high* area. The landslide inventories were overlaid with the susceptibility maps to support the visual interpretation of the maps. The SAE–RGF model divided the study area into the five susceptibility classes with percentages of 9%, 32%, 32%, 20% and 7%. The result indicated that 27% of the area, particularly along the Phuentsholing–Thimphu highway and nearby areas, is under high and very high risks due to landslides. Using the RGF model without applying SAE factor optimisation yielded a reduction in the very low and high susceptible zones by 6% and 7%, respectively. The RGF model predicted that 36% of the area is under low and moderate landslide susceptibility classes. The RGF model also predicted a higher percentage of the area (12%) than what the SAE–RGF model predicted. Significantly different results were observed for the RF model. The study area was divided into 44%, 38%, 10%, 3% and 4% susceptibility classes by the RF model. This model suggested that only 7% of the area is under high and very high susceptible zones.

Table 2 lists the success and prediction rates of the three models. The best success rate (0.972) was achieved by the RGF model, which outperformed the proposed SAE–RGF (0.931) and RF (0.876) models. However, the results regarding the prediction rates suggested that the proposed SAE–RGF model exhibited the best generalisation capability with a prediction rate of 0.892 compared with 0.865 and 0.824 for the RGF and RF models, respectively. Reducing the dimensionality of the input data from 10 to 7 helped improve the prediction capability of the RGF model. The percentage of landslides in the susceptibility classes was utilized by comparing landslide occurrences with the results of the landslide susceptibility maps (Figure 7). Approximately 43% and 33% of landslide inventories were identified in the very high and high susceptible zones, respectively, by the SAE–RGF model. The very high susceptible zone for the map produced by the RGF model contained 82% of the landslides. However, the map and AUROC values (success rate = 0.972, prediction rate = 0.865) implied that this phenomenon was due to the overfitting of the training data, and the model failed to predict the absence samples correctly. The very high and high susceptibility classes for the map produced by the RF model contained 27% and 13% of the landslides, respectively. This finding also suggested that the proposed SAE–RGF model helped identify numerous landslides located in the very high susceptible zone without considerable overfitting to the training data and produced reliable landslide susceptibility maps in the study region.


89°15'
89°30'
89°45'
27°15'
27°
26°45'
N
Landslides
Very low
Low
Moderate
High
Very high
10
Kilometers
(a)
SAE-RGF


89°15'
89°30'
89°45'
27°15'
27°
26°45'
N
Landslides
Very low
Low
Moderate
High
Very high
10
Kilometers
(b)
RGF

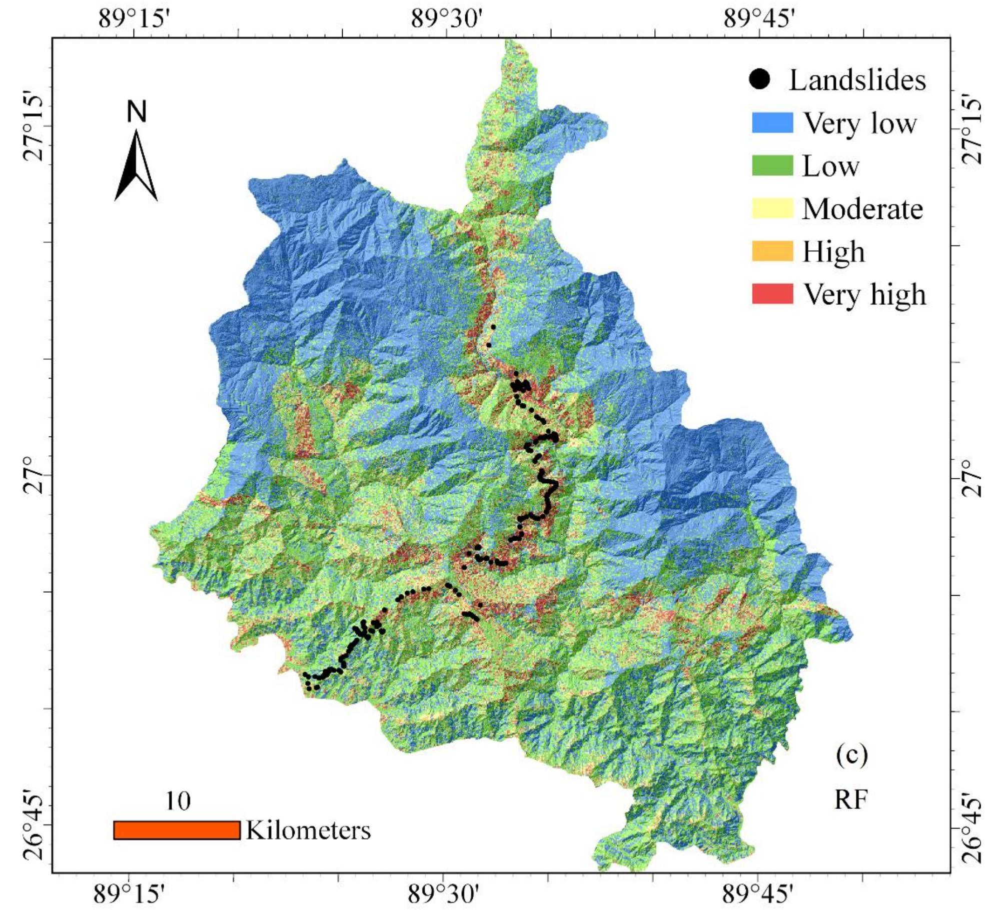


Figure 6: Landslide susceptibility maps produced by (a) SAE–RGF, (b) RGF and (c) RF.

Table 2: Success and prediction rates of the landslide susceptibility models.

Susceptibility model	Success rate	Prediction rate
SAE–RGF	0.931	**0.892**
RGF	**0.972**	0.865
RF	0.876	0.824

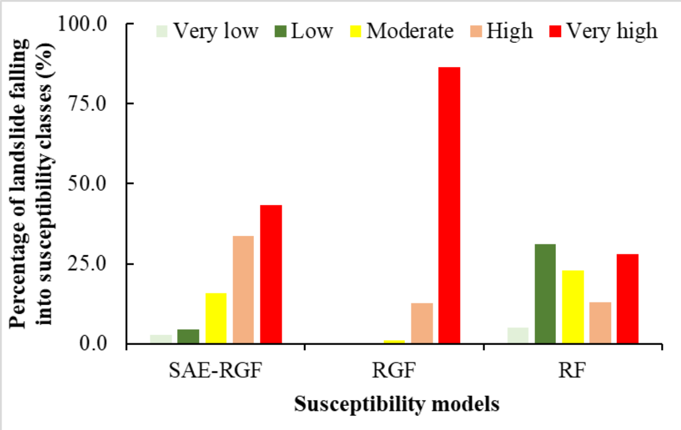


Figure 7: Graph showing landslides (in %) for different susceptibility classes for the three models.

Tree-based landslide susceptibility modelling methods, including their ensemble, such as RF, are frequently affected by variations and noise in data. The SAE model proposed in this study helps reduce information redundancy and noise in the data by learning a new set of nonlinear feature representations from the input data with a lower dimension than that of the original feature set. Previous methods on factor optimisation for landslide susceptibility mappings, such as the methods presented by Jebur et al. (2014) and Dou et al. (2015), are supervised and select a subset from the original data without any transformation to the input features. Although these methods help improve the prediction ability of statistical and machine learning methods, such as statistical index, logistic regression and support vector machines, they require high-quality training data and do not reduce noise nor improve input features in terms of information content. By contrast, the proposed SAE–RGF is unsupervised (no training data are required) and helps tree-based models that are highly sensitive to noise and data variations. Consequently, using the new representations learned by the SAE–RGF model can help improve tree-based models, such as RGF and RF, for landslide susceptibility within a study region. The proposed SAE–RGF model also helps reduce the training and inference prediction times of the RGF and RF models by reducing the input data dimension from 10 to 7 features.

However, the model also has several limitations at the current implementation. First, the selection of a new dimension, which is often lower than the original dimension of the landslide factors, can be challenging. It requires several experiments to evaluate different alternatives until the optimum one can be found. This challenge is getting harder when the original dimensionality is larger. Search methods can be used such as a grid or random search but that can be computationally expensive. To address this challenge, future implementations should focus on either automating this process within the workflow or developing a statistical measure that allows a good selection of this parameter. Second, after transforming the factors with a non-linear function learnt by the SAE model, the interpretation of the models is getting much harder than the original models. So, the current strategy is focused on prediction accuracy improvement rather than model interpretation and explainability. Those issues can be explored in future works by using interpretable models to perform factor optimisation.

6. Conclusions

This research demonstrated the use of an unsupervised factor optimisation approach based on sparse autoencoders (SA) to improve the performance of tree-based landslide susceptibility models in Chukha Dzongkhag, Bhutan. The model enables learning a new set of nonlinear feature representations with richer information and lower dimensionality. It is an important pre-processing step for landslide susceptibility modelling that requires neither additional training data nor human supervision. The success and prediction rates estimated based on AUROC indicated the prevalence of the proposed model over RGF and RF models, particularly in terms of generalisation to the test dataset.

Originally, the model used 10 landslide conditioning factors, including geo-morphometrical and geo-environmental. The performance of RGF was about 0.972 and 0.865 as for success and prediction rates, respectively. After transforming the factor values with a non-linear function learnt by the SAE, the accuracy of RGF has dropped to 0.908 and 0.830 as for the success and prediction rates, respectively. But interestingly, when the dimensionality of the factors was reduced to only 7 features, the prediction rate of RGF went up to become 0.892. As several landslide conditioning factors are often derived from a single source (DEM), those factors are statistically correlated to each other. Reducing the dimensionality of these factors is therefore useful and boosts the performance of the landslide susceptibility models. However, this comes with a challenge, which is that selecting a good dimension size to transform the factors requires additional experiments and statistical analysis. Further research is thus needed to improve our understanding of how these models should be applied to different geographical regions. Also, automating the selection of an optimised dimension to improve landslide susceptibility can be a good research direction.

The proposed model can be useful for disaster managers, urban planners and technicians in landslide-prone regions to improve landslide susceptibility evaluation procedures without raising information and computational resource expenses. Landslide susceptibility maps can be useful in enforcing reconstruction strategies in other geospatial apps and in choosing spatial sites.

Funding

This research is supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS) in the University of Technology Sydney (UTS) under Grants 321740.2232335 and 321740.2232357; Grant 321740.2232424, and Grant 321740.2232452. This research is also supported by Researchers Supporting Project number (RSP-2019 / 14, King Saud University, Riyadh, Saudi Arabia.

Conflicts of interest

The authors declare no conflict of interest.

Computer Code Availability

The Regularized Greedy Forest algorithm was implemented using the rgf-python library freely available at (https://github.com/RGF-team/rgf). The other bench-marked models were implemented using Sklearn (https://scikit-learn.org). Moreover, the complete notebook of the experiments presented in the paper is available on GitHub (https://github.com/malzuhairi/rgf_landslides).

References

Aghdam, I. N., Varzandeh, M. H. M., & Pradhan, B. (2016). Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). Environmental Earth Sciences, 75(7), 553.

Bui, D. T., Pradhan, B., Revhaug, I., and Tran, C. T. (2014). A comparative assessment between the application of fuzzy unordered rules induction algorithm and J48 decision tree models in spatial prediction of shallow landslides at Lang Son City, Vietnam. In Remote Sensing Applications in Environmental Research (pp. 87-111). Springer, Cham.

Can, R., Kocaman, S., & Gokceoglu, C. (2019). A Convolutional Neural Network Architecture for Auto-Detection of Landslide Photographs to Assess Citizen Science and Volunteered Geographic Information Data Quality. ISPRS International Journal of Geo-Information, 8(7), 300.

Chen, W., Panahi, M., Tsangaratos, P., Shahabi, H., Ilia, I., Panahi, S., Li, S., Jaafari, A. and Ahmad, B.B. (2019). Applying population-based evolutionary algorithms and a neuro-fuzzy system for modeling landslide susceptibility. Catena, 172, 212-231.

Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., Zhu, A.X., Pei, X. and Duan, Z. (2018). Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. Science of the Total Environment, 626, 1121-1135.

Chu, L., Wang, L. J., Jiang, J., Liu, X., Sawada, K., & Zhang, J. (2019). Comparison of landslide susceptibility maps using random forest and multivariate adaptive regression spline models in combination with catchment map units. Geosciences Journal, 23(2), 341-355.

Comert, R., Avdan, U., Gorum, T., & Nefeslioglu, H. A. (2019). Mapping of shallow landslides with object-based image analysis from unmanned aerial vehicle data. Engineering Geology, 105264.

Dehnavi, A., Aghdam, I. N., Pradhan, B., & Varzandeh, M. H. M. (2015). A new hybrid model using step-wise weight assessment ratio analysis (SWARA) technique and adaptive neuro-fuzzy inference system (ANFIS) for regional landslide hazard assessment in Iran. Catena, 135, 122-148.Aditian, A., Kubota, T., and Shinohara, Y. (2018). Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. Geomorphology, 318, 101-111.

Dou, J., Bui, D.T., Yunus, A.P., Jia, K., Song, X., Revhaug, I., Xia, H. and Zhu, Z., 2015. Optimization of causative factors for landslide susceptibility evaluation using remote sensing and GIS data in parts of Niigata, Japan. PloS one, 10(7), e0133262.

Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.W., Khosravi, K., Yang, Y. and Pham, B.T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. Science of the Total Environment, 662, 332-346.

Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.W., Khosravi, K., Yang, Y. and Pham, B.T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. Science of the Total Environment, 662, 332-346.

Gallant, J. P. W. J. C. (2000). Terrain analysis: principles and applications. John Wiley & Sons.

Gomez, H., and Kavzoglu, T. (2005). Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. Engineering Geology, 78(1-2), 11-27.

He, Q., Xu, Z., Li, S., Li, R., Zhang, S., Wang, N., Pham, B.T. and Chen, W. (2019). Novel Entropy and Rotation Forest-Based Credal Decision Tree Classifier for Landslide Susceptibility Modeling. Entropy, 21(2), 106.

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

Hong, H., Naghibi, S. A., Pourghasemi, H. R., & Pradhan, B. (2016). GIS-based landslide spatial modeling in Ganzhou City, China. Arabian Journal of Geosciences, 9(2), 112.

Hong, H., Pradhan, B., Sameen, M. I., Chen, W., and Xu, C. (2017). Spatial prediction of rotational landslide using geographically weighted regression, logistic regression, and support vector machine models in Xing Guo area (China). Geomatics, Natural Hazards and Risk, 8(2), 1997-2022.

Hong, H., Pradhan, B., Sameen, M. I., Kalantar, B., Zhu, A., and Chen, W. (2018). Improving the accuracy of landslide susceptibility model using a novel region-partitioning approach. Landslides, 15(4), 753-772.

Hong, H., Pradhan, B., Xu, C., & Tien Bui, D. (2015). Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. Catena, 133, 266–281. doi:10.1016/j.catena.2015.05.019

Huang, Y., and Zhao, L. (2018). Review on landslide susceptibility mapping using support vector machines. Catena, 165, 520-529.

Hung, L. Q., Van, N. T. H., Van Son, P., Khanh, N. H., & Binh, L. T. (2016). Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: a case study in the upper Lo River catchment (Vietnam). Landslides, 13(5), 1285-1301.

Jaafari, A., Panahi, M., Pham, B. T., Shahabi, H., Bui, D. T., Rezaie, F., and Lee, S. (2019). Meta optimization of an adaptive neuro-fuzzy inference system with grey wolf optimizer and biogeography-based optimization algorithms for spatial prediction of landslide susceptibility. Catena, 175, 430-445.

Jebur, M. N., Pradhan, B., and Tehrany, M. S. (2014). Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. Remote Sensing of Environment, 152, 150-165.

Johnson, R., and Zhang, T. (2014). Learning nonlinear functions using regularized greedy forest. IEEE transactions on pattern analysis and machine intelligence, 36(5), 942-954.

Kadavi, P., Lee, C. W., and Lee, S. (2018). Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sensing, 10(8), 1252.

Kaur, H., Gupta, S., Parkash, S., Thapa, R., Gupta, A., & Khanal, G. C. (2019). Evaluation of landslide susceptibility in a hill city of Sikkim Himalaya with the perspective of hybrid modelling techniques. Annals of GIS, 25(2), 113-132.

Kocaman, S., & Gokceoglu, C. (2019). A CitSci app for landslide data collection. Landslides, 16(3), 611-615.

Kornejady, A., Pourghasemi, H. R., & Afzali, S. F. (2019). Presentation of RFFR New Ensemble Model for Landslide Susceptibility Assessment in Iran. In Landslides: Theory, Practice and Modelling (pp. 123-143). Springer, Cham.

Kuenza, K., Dorji, Y. and Wangda, D. (2010) Landslides in Bhutan. In: Proceedings of the SAARC Workshop on Landslide Risk Management in South Asia, Thimphu, Bhutan, 11–12 May 2010, pp 73–80.

Kutlug Sahin, E., & Colkesen, I. (2019). Performance Analysis of Advanced Decision Tree-Based Ensemble Learning Algorithms for Landslide Susceptibility Mapping. Geocarto International, (just-accepted), 1-23.

Lee, J. H., Sameen, M. I., Pradhan, B., and Park, H. J. (2018). Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology, 303, 284-298.

Lee, S., Lee, M. J., and Lee, S. (2018). Spatial prediction of urban landslide susceptibility based on topographic factors using boosted trees. Environmental Earth Sciences, 77(18), 656.

Liu, J., and Duan, Z. (2018). Quantitative assessment of landslide susceptibility comparing statistical index, index of entropy, and weights of evidence in the Shangnan area, China. Entropy, 20(11), 868.

Luo, X., Lin, F., Zhu, S., Yu, M., Zhang, Z., Meng, L., & Peng, J. (2019). Mine landslide susceptibility assessment using IVM, ANN and SVM models considering the contribution of affecting factors. Plos One, 14(4), e0215134. doi:10.1371/journal.pone.0215134

Mandal, S., and Maiti, R. (2015). Semi-quantitative approaches for landslide assessment and prediction. Singapore: Springer.

Meneses, B. M., Pereira, S., and Reis, E. (2019). Effects of different land use and land cover data on the landslide susceptibility zonation of road networks. Natural Hazards and Earth System Sciences, 19(3), 471-487.

Nguyen, V.V., Pham, B.T., Vu, B.T., Prakash, I., Jha, S., Shahabi, H., Shirzadi, A., Ba, D.N., Kumar, R., Chatterjee, J.M. and Tien Bui, D. (2019). Hybrid Machine Learning Approaches for Landslide Susceptibility Modeling. Forests, 10(2), 157.

Oh, H. J., & Pradhan, B. (2011). Application of a neuro-fuzzy model to landslide-susceptibility mapping for shallow landslides in a tropical hilly area. Computers & Geosciences, 37(9), 1264-1276.

Ozer, B. C., Mutlu, B., Nefeslioglu, H. A., Sezer, E. A., Rouai, M., Dekayir, A., & Gokceoglu, C. (2019). On the use of hierarchical fuzzy inference systems (HFIS) in expert-based landslide susceptibility mapping: the central part of the Rif Mountains (Morocco). Bulletin of Engineering Geology and the Environment, 1-18.

Ozer, B. C., Mutlu, B., Nefeslioglu, H. A., Sezer, E. A., Rouai, M., Dekayir, A., & Gokceoglu, C. (2018, April). Expert-based landslide susceptibility modelling by using hierarchical fuzzy systems (HFS): an investigation from central part of Rif Mountains (Morocco). In EGU General Assembly Conference Abstracts (Vol. 20, p. 33).

Ozturk, H. S., Kocaman, S., & Gokceoglu, C. (2019). A low-cost approach for determination of discontinuity orientation using smartphone images and application to a part of Ihlara Valley (Central Turkey). Engineering Geology, 254, 63-75.

Peethambaran, B., Anbalagan, R., Shihabudheen, K. V., and Goswami, A. (2019). Robustness evaluation of fuzzy expert system and extreme learning machine for geographic information system-based landslide susceptibility zonation: A case study from Indian Himalaya. Environmental Earth Sciences, 78(6), 231.

Polykretis, C., and Chalkias, C. (2018). Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models. Natural Hazards, 93(1), 249-274.

Polykretis, C., Chalkias, C., and Ferentinou, M. (2019). Adaptive neuro-fuzzy inference system (ANFIS) modeling for landslide susceptibility assessment in a Mediterranean hilly area. Bulletin of Engineering Geology and the Environment, 78(2), 1173–1187.

Pradhan, B., & Lee, S. (2010). Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. Landslides, 7(1), 13-30.

Pradhan, B., Sezer, E. A., Gokceoglu, C., and Buchroithner, M. F. (2010). Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (Cameron Highlands, Malaysia). IEEE Transactions on Geoscience and Remote Sensing, 48(12), 4164-4177.

Sameen, M. I., Pradhan, B., and Lee, S. (2018). Self-learning random forests model for mapping groundwater yield in data-scarce areas. Natural Resources Research, 1-19. https://doi.org/10.1007/s11053-018-9416-1.

Shirani, K., Pasandi, M., and Arabameri, A. (2018). Landslide susceptibility assessment by Dempster–Shafer and Index of Entropy models, Sarkhoun basin, Southwestern Iran. Natural Hazards, 93(3), 1379-1418.

Shirzadi, A., Soliamani, K., Habibnejhad, M., Kavian, A., Chapi, K., Shahabi, H., Chen, W., Khosravi, K., Thai Pham, B., Pradhan, B. and Ahmad, A. (2018). Novel GIS based machine learning algorithms for shallow landslide susceptibility mapping. Sensors, 18(11), 3777.

Soma, A. S., Kubota, T., and Mizuno, H. (2019). Optimization of causative factors using logistic regression and artificial neural network models for landslide susceptibility assessment in Ujung Loe Watershed, South Sulawesi Indonesia. Journal of Mountain Science, 16(2), 383-401.

Song, Y., Niu, R., Xu, S., Ye, R., Peng, L., Guo, T., Li, S. and Chen, T. (2019). Landslide Susceptibility Mapping Based on Weighted Gradient Boosting Decision Tree in Wanzhou Section of the Three Gorges Reservoir Area (China). ISPRS International Journal of Geo-Information, 8(1), 4.

Sun, X., Chen, J., Bao, Y., Han, X., Zhan, J., and Peng, W. (2018). Landslide Susceptibility Mapping Using Logistic Regression Analysis along the Jinsha River and Its Tributaries Close to Derong and Deqin County, Southwestern China. ISPRS International Journal of Geo-Information, 7(11), 438.

Walker, L. R., & Shiels, A. B. (2012). Landslide ecology. Cambridge University Press.

Wang, Y., Fang, Z., & Hong, H. (2019). Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. The Science of the Total Environment, 666, 975–993. doi:10.1016/j.scitotenv.2019.02.263

Xiao, L., Zhang, Y., and Peng, G. (2018). Landslide susceptibility assessment using integrated deep learning algorithm along the china-nepal highway. Sensors, 18(12), 4436.

Youssef, A. M., Al-Kathery, M., & Pradhan, B. (2015). Landslide susceptibility mapping at Al-Hasher area, Jizan (Saudi Arabia) using GIS-based frequency ratio and index of entropy models. Geosciences Journal, 19(1), 113-134.

Zare, M., Pourghasemi, H. R., 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. Arabian Journal of Geosciences, 6(8), 2873-2888.

Zhang, T., Han, L., Han, J., Li, X., Zhang, H., and Wang, H. (2019). Assessment of Landslide Susceptibility Using Integrated Ensemble Fractal Dimension with Kernel Logistic Regression Model. Entropy, 21(2), 218.