Applied Soft Computing 142 (2023) 110324

Contents lists available at ScienceDirect

Applied Soft Computing

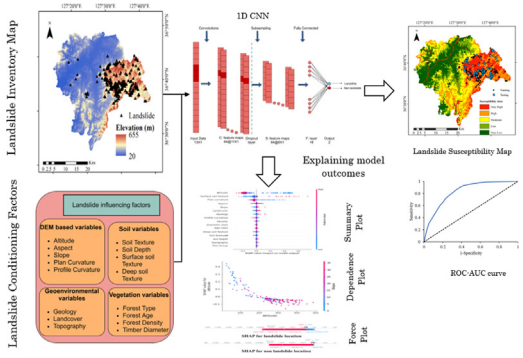
journal homepage: www.elsevier.com/locate/asoc

An explainable AI (XAI) model for landslide susceptibility modeling

Biswajeet Pradhan [a,b,*], Abhirup Dikshit [a], Saro Lee [c,d,**], Hyesu Kim [e]

[a] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007, Australia
[b] Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia
[c] Geoscience Data Center, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124 Gwahang-no, Yuseong-gu, Daejeon 34132, South Korea
[d] Department of Resources Engineering, Korea University of Science and Technology, 217 Gajeong-ro, Yuseong-gu, Daejeon 34113, South Korea
[e] Department of Astronomy, Space Science and Geology, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, South Korea

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 October 2021
Received in revised form 5 March 2023
Accepted 15 April 2023
Available online 25 April 2023

Keywords:
Landslide susceptibility
Convolutional neural networks
SHAP
Explainable AI

ABSTRACT

Landslides are among the most devastating natural hazards, severely impacting human lives and damaging property and infrastructure. Landslide susceptibility maps, which help to identify which regions in a given area are at greater risk of a landslide occurring, are a key tool for effective mitigation. Research in this field has grown immensely, ranging from quantitative to deterministic approaches, with a recent surge in machine learning (ML)-based computational models. The development of ML models, in particular, has undergone a meteoritic rise in the last decade, contributing to the successful development of accurate susceptibility maps. However, despite their success, these models are rarely used by stakeholders owing to their "black box" nature. Hence, it is crucial to explain the results, thus providing greater transparency for the use of such models. To address this gap, the present work introduces the use of an ML-based explainable algorithm, SHapley Additive exPlanations (SHAP), for landslide susceptibility modeling. A convolutional neural network model was used conducted in the CheongJu region in South Korea. A total of 519 landslide locations were examined with 16 landslide-affected variables, of which 70% was used for training and 30% for testing, and the model achieved an accuracy of 89%. Further, the comparison was performed using Support Vector Machine mode, which achieved an accuracy of 84%. The SHAP plots showed variations in feature interactions for both landslide and non-landslide locations, thus providing more clarity as to how the model achieves a specific result. The SHAP dependence plots explained the relationship between altitude and slope,

* Corresponding author at: Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007, Australia.
** Corresponding author at: Geoscience Data Center, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124 Gwahang-no, Yuseong-gu, Daejeon 34132, South Korea.
E-mail addresses: Biswajeet.Pradhan@uts.edu.au (B. Pradhan), leesaro@kigam.re.sr (S. Lee).

https://doi.org/10.1016/j.asoc.2023.110324

showing a negative relationship with altitude and a positive relationship with slope. This is the first use of an explainable ML model in landslide susceptibility modeling, and we argue that future works should include aspects of explainability to open up the possibility of developing a transferable artificial intelligence model.

1. Introduction

Landslides are frequently recurring natural disasters that can be triggered by rainfall, earthquakes, human activities, or a combination of these factors. Rainfall-induced landslides are the most common type and affect large parts of the world [1,2]. South Korea is highly susceptible to landslides triggered by heavy rainfall, with an increasing trend associated with changes in rainfall patterns, primarily due to climate change [3,4]. The risk associated with landslide events is particularly significant, given their occurrence in residential and transportation areas [4].

A key step in landslide mitigation is determining "where" landslides may occur, known as landslide susceptibility mapping [5]. A vast body of international work has explored different techniques for the development of a robust susceptibility model using various approaches [5,6]. Landslide susceptibility models can be broadly categorized into four types: (i) qualitative models [7], (ii) quantitative/data-driven models, (iii) semi-quantitative models [5], and (iv) deterministic models [8]. Irrespective of the model type, the basic procedure involves the collection of available landslide conditioning variables based on an appropriate mapping unit (e.g., pixel, landslide geometry), which serves as input to the model being used [5].

The use of machine learning (ML) models has recently gained traction owing to their ability to understand the complexities of interactions between the variables and target [9,10]. Among the most commonly used models are artificial neural networks [6,11], support vector machines [12,13], and random forests [14,15]. These models perform better than conventional statistical models. Moreover, several studies have looked at the use of ensemble models and improved the results yielded by simple ML models. However, these approaches directly classify the input data and do not uncover more representative features from these data, which would further improve the classification results [16]. With new feats achieved by deep neural networks, such as their defeat of humans in the game of Go [17], a two-hand poker game [18], and several others, they have also proven successful in different fields, including computer vision and natural language processing [19]. Various studies have examined different aspects of landslides, such as landslide detection [20], landslide mapping [21], and susceptibility modeling [22], using deep learning techniques.

Chen et al. [20] applied convolutional neural networks (CNN) to develop a change detection model capable of identifying landslide-prone regions using spatiotemporal images of mountainous regions in China. Prakash et al. [21] used a modified U-net architecture to map landslides in Douglas County, USA. The study compared their results with conventional techniques (pixel-based, object-based) and found that deep learning yielded better results. Wang et al. [23] were the first to use a CNN model for landslide susceptibility modeling. The study was conducted in Yanshan, China, and a comparison between CNN and other ML models' results revealed that the CNN outperformed the benchmark ML models. Sameen et al. [22] conducted a similar study for South Korea and found that deep neural networks performed better. The aforementioned studies demonstrate that deep learning models perform considerably better than traditional ML models. Therefore, in this study, we explored the use of a CNN to develop a robust landslide susceptibility model for South Korea's CheongJu region, which has witnessed frequent landslide recurrences in recent years.

However, all the above studies lack a key component: model explainability. Although hybrid and deep learning models have yielded superior and more accurate results, they are considered black boxes, and their use among stakeholders is minimal. A review article by Dikshit et al. [9] on the challenges of using artificial intelligence (AI) in the field of geohazards highlighted this key missing link in existing studies. More recently, Ozturk et al. [24] raised a similar question with the aim of exploring dynamic susceptibility and outcome interpretation in data-driven models. In this work, we attempt to interpret an ML model using an additive explainer, SHapley Additive exPlanation (SHAP) [25]. The algorithm has recently attracted interest owing to its additive properties, which provide various plots and help clarify the inter-dependencies among variables toward model outcomes. For example, Matin and Pradhan [26] used SHAP to explain the reliability of the ML model for mapping building damage after an earthquake event. Abdollahi and Pradhan [27] used SHAP to explain a deep learning model used for vegetation classification. García and Aznarte [28] used SHAP to analyze NO_2 predictions in Madrid using a long short-term memory (LSTM) model. In the field of geohazard studies, Dikshit and Pradhan [29] used SHAP to investigate how deep learning models achieve specific results under different drought conditions. Their study applied an LSTM model to predict droughts in Australia, comparing SHAP results with physical-based models, and found that SHAP yielded similar results. To the best of the authors' knowledge, this is the first paper to apply an explainable AI (XAI) model to investigate landslide susceptibility.

This study's key contribution lies in the application of an explainable model (SHAP) in landslide susceptibility modeling. The work explored model outcomes for the entire region as well as for specific landslide and non-landslide pixels. It further investigated dependence among the variables, which contributes to achieving the model's result. This can benefit the broader landslide study community by clarifying how ML models achieve their outcomes, so as to apply such models effectively for disaster management purposes. The objectives of the study were to apply a robust landslide susceptibility model based on a CNN, analyze and examine the results obtained from the CNN, and introduce the use of SHAP in landslide modeling followed by interpreting the model's outcomes.

The paper is organized as follows. The "Study Area" section (Section 2) describes the area of interest and the history of landslides in the region. Section 3 discusses the various datasets used, which are also landslide conditioning factors. Section 4 explains the CNN architecture and basics of the SHAP algorithm. Section 5 presents the CNN results, along with the landslide susceptibility map. The section also presents SHAP plots displaying model outcomes for the entire test dataset as well as landslide and non-landslide locations. Section 6 provides a comprehensive discussion of the results and highlights avenues for future research. Finally, Section 7 summarizes and concludes the study.

2. Study area

The study area is the CheongJu region (covering 939 km^2) of Chungcheongbuk-do Province located in the central part of South

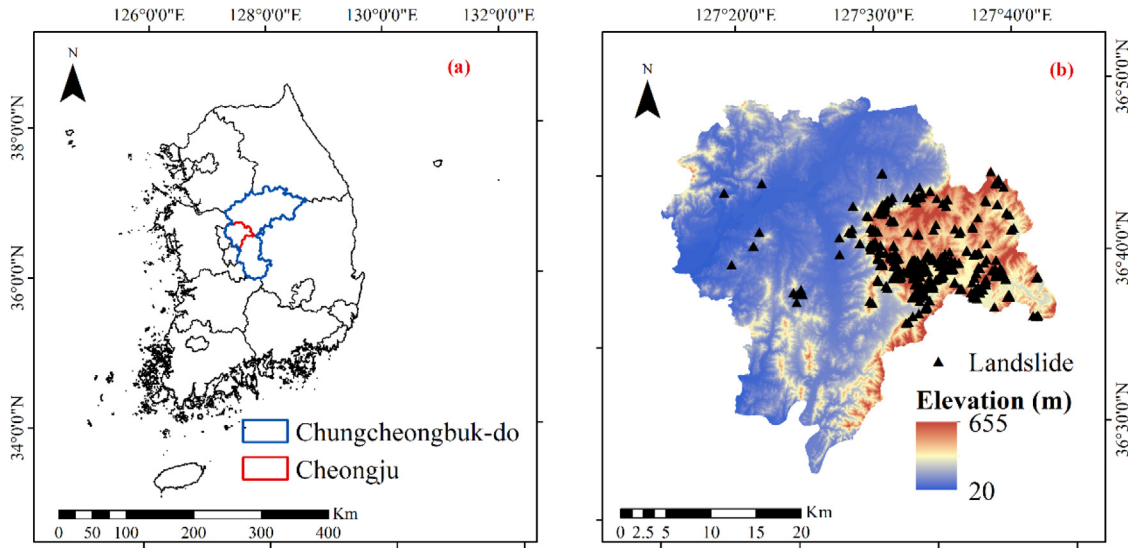


Fig. 1. (a) The blue boundary indicates Chungcheongbuk-do Province and the red boundary depicts the CheongJu region. (b) Elevation map of the CheongJu region with landslide locations.

Korea (Fig. 1). The country is situated in the northeast part of Asia next to China and Japan and has an area of 99,600 km^2, 70% of which is covered by mountains with elevations up to 1200 m [30]. Most landslides in South Korea occur due to intense rainfall, causing an average of 36 fatalities every year with annual damages of 500–1000M US$ [30,31]. CheongJu City has two rivers, around which a wide alluvial plain has formed. The east includes many mountains, such as the Sandang, Gunyeo, and Uam mountains. The study area has a distinctly temperate continental climate. Therefore, summers are hot and humid, whereas winters tend to be cold and dry. The average annual precipitation is greater than 1200 mm, and it rains intensively in summer. In July 2017, heavy rainfall of 290 mm a day associated with climate change was recorded. This heavy rainfall caused not only flooded rivers but also several landslides within a few days. This resulted in damage amounting to over 26M US$, of which 2.5M US$ was due to landslides while the remainder was due to flooding.

Chungcheongbuk-do Province is situated in a temperate climatic zone, with warm and humid summers and an annual rainfall of over 1500 mm. Rainfall occurs frequently during the months of June–September, with the highest rainfall occurring in July. As per Köppen classification, the region has two different climates—a hot humid continental climate and a warm humid continental climate—of which the former is the dominant type.

2.1. Data used

The occurrence of landslide events is influenced by multiple factors, including topography, vegetation, and soil [6]. The selection of variables for susceptibility modeling depends on the region of interest and the data available. Various researchers conducted investigations using different variables, and several review articles have provided comprehensive analyses of their variable selection [5,32]. Although several variables affect landslide incidence, variables are often selected based on the characteristics of the study area and the available data. Therefore, in the present study, we use 16 variables (details provided in Table 1) including topographical, geo-environmental, and soil variables, among others. Different landslide studies in South Korea have applied these factors [11,22]. Some of the variables had categorical values (soil attributes, vegetation attributes, and land cover), whereas the others had continuous values. No reclassification was performed for variables with continuous values to eliminate model sensitivity associated with reclassification.

Table 1
Landslide affecting variables used in the study.

Variable type	Variable name	Data source
Geomorpho-logical	Slope Topography Elevation Aspect Plan curvature Profile curvature	Ministry of Land, Infrastructure and Transport (MOLIT), Korea
Vegetation	Forest type Forest density Timber diameter Forest age	Korea Forest Service
Soil	Soil depth Soil drain Surface soil texture Deep soil texture	National Institute of Agricultural Sciences. Korea
Geology	Lithology	Korea Institute of Geoscience and Mineral Resources (KIGAM), Korea
Land use	Land use type	Ministry of environment, Korea

Most of the landslides occurred in mountainous regions, as is also evident from the elevation map (Fig. 1b), ranging from 20 to 655 m. According to Varnes [33] classification, this region is susceptible to slide and flow landslide types. The slides were single rotational types, and the flows commonly seen in the study area were debris flows. Detailed information on landslide occurrence locations was extracted from the landslide occurrence history provided by the Korea Forest Service. A spatial database was prepared by interpreting aerial images captured before and after each event (i.e., images from 2016 and 2018 were used to identify landslide occurrences in 2017). Multiple landslides occurred in 2018, with the largest landslide occurring at 36°41′52.14″N and 127°35′47.98″E, with a length of 962 m and an areal coverage of 6510 m^2, based on aerial photographs. Some of the damage that occurred as a result of landslides in 2018 is illustrated in Fig. 2(a, b). Fig. 3(a, b) shows an example of a landslide occurrence at 36°40′18.58″N, 127°42′3.28″E, with an area of 14 m^2, length of 4 m, and width of 3 m.

The region's geology is heterogeneous (Fig. 4). These variables affect landslide mechanisms, as rock types determine the susceptibility to slide activity [34]. The geology is divided by period into unknown metamorphic sedimentary rocks, Precambrian metamorphic rocks, Paleozoic metamorphic and sedimentary rocks,

Fig. 2. Landslide damage on July 16, 2018.

Fig. 3. Panoramic (a) and close-up view (b) of a landslide occurrence on July 16, 2018, in the CheongJu region (36°40′18.58″N, 127°42′3.28″E).

Mesozoic igneous rocks, and Quaternary alluvium. The metasedimentary rocks are known as the Ogcheng group [35]. They consist mainly of quartzite, schist, and phyllite. The Precambrian metamorphic rocks located in the northwest of the study area consist largely of gneiss—specifically, biotite angen gneiss, banded biotite gneiss, granitic gneiss, and some mica schist have been identified [35].

The Paleozoic metamorphic sedimentary rocks are widely distributed in the eastern part of the study area. Most exhibit a NE strike and NW dip. They are distributed from west to east in a layer composed of sandstone, shale, phyllite, and calcareous shale; a layer composed of sandy shale, phyllite, and calcareous shale; a limestone layer; schist layers; arenaceous phyllite; and layers including coal. The arenaceous phyllites are widely distributed in the westernmost region of the Paleozoic rocks. The Paleozoic rocks are folded, and thus, old formations are located in the syncline, with arenaceous phyllites repeated [36]. The Mesozoic igneous rocks include gneiss and are widely distributed over the CheongJu area. They consist of granites and dikes. Granites include porphyritic granite, biotite granite, and diorite [37]. There are acidic and basic dikes. Most acidic dikes are quartz veins and felsite dikes, whereas the basic dikes are andesitic in composition. The Quaternary alluvium is deposited along the rivers and consists of unconsolidated sediments, such as sand, gravel, and clay.

The CheongJu area can be divided into seven distinct topographical categories, of which more than 38% consist of mountains, followed by valleys (19%) (Fig. 5a). The land use patterns of the area affect landslide activity, as anthropogenic activities have caused changes, which in turn have led to an increase in such events (Fig. 5b). The global landslide inventory data gathered by Froude and Petley [38] also demonstrated that human activities linked to landslide activity have risen significantly and should thus be used as a variable in landslide modeling studies.

Geomorphological variables, such as the elevation, slope, aspect, and plan and profile curvature, were derived from a digital elevation model with a spatial resolution of 10 m. These factors are important for landslide modeling, as they have a direct correlation with slope failure [11,22]. The aspect (Fig. 6a) plays a

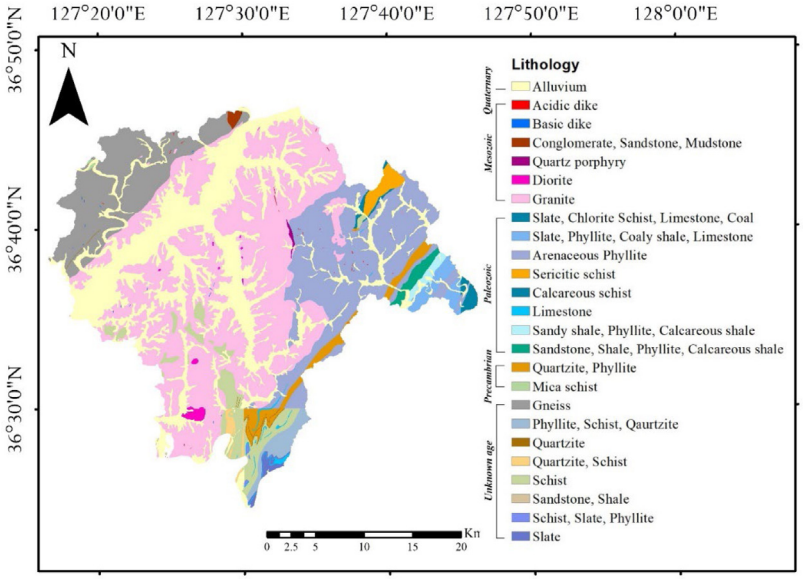


Fig. 4. Geological map of the study region.

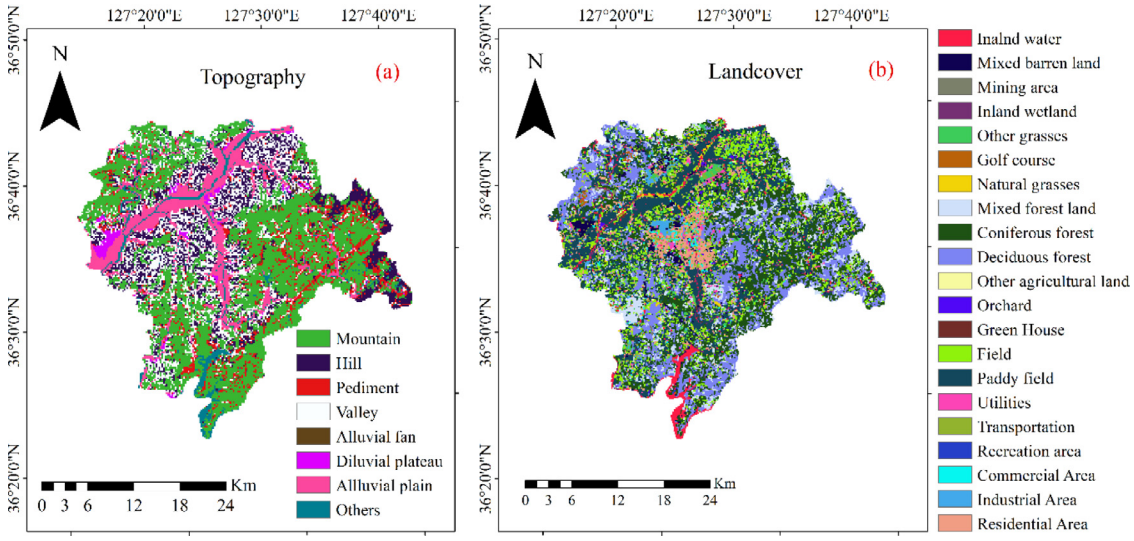


Fig. 5. (a) Topographical and (b) landcover maps of the CheongJu area.

critical role in landslide events, as it influences moisture conditions via rainfall and solar radiation [6]. The region's slope ranges from 0° to 78° (Fig. 6b) and is an important variable in landslide susceptibility modeling [22]. The curvature reflects a change in slope values along each slope's curve, which has the potential to influence slope stability [22]. Herein, two different variables are used to represent curvature: the plan and profile (Fig. 6c, d). The plan curvature is perpendicular to the direction of the peak slope. For convex surfaces, the curvature is positive and vice versa for concave surfaces [39]. By contrast, the profile curvature is parallel to the direction of the maximum slope, which affects flow across the surface [39].

Vegetation attributes are represented by four variables: forest type, forest density, forest age, and timber diameter, which were obtained from the Korean Forest Service (Fig. 7a–d). Owing to the overwhelming presence of hilly regions, these variables play a crucial role in landslide susceptibility modeling [22,40]. Similarly, forest density influences landslide occurrence as it is related to ground reinforcement, whereas the forest type has an influence as a denser canopy cover leads to higher interception and enhanced root penetration [40,41]. Moos et al. [40] studied the effect of forest structure on shallow landslides and found that forests in poor condition experienced more landslide occurrences.

Rainfall-triggered shallow landslides occur due to a complex process primarily caused by the interaction between hydrological processes and soil mechanical reactions toward hydrological loading [39]. Geology concerns not only the base of the region but also the soil materials. Therefore, soil characteristics, such as texture, hydraulic properties, and thickness, are affected by local lithology [42]. A cause investigation implemented after the CheongJu landslides found that the area was vulnerable to heavy rainfall owing to its concave water-collecting topography and thin soil layers. The variations in soil strength are dependent on pore water pressure, whereas soil hydration status during heavy rainfall is controlled by the surface topography, bedrock, and soil hydraulic properties [39]. Soil attributes also play a significant role, with clayey soil setting up a potential slip zone, particularly for shallow landslide events. In this study, we examined four different soil attributes: deep soil texture, soil depth, surface soil texture, and soil drainage (Fig. 8). Clayey soil plays a key

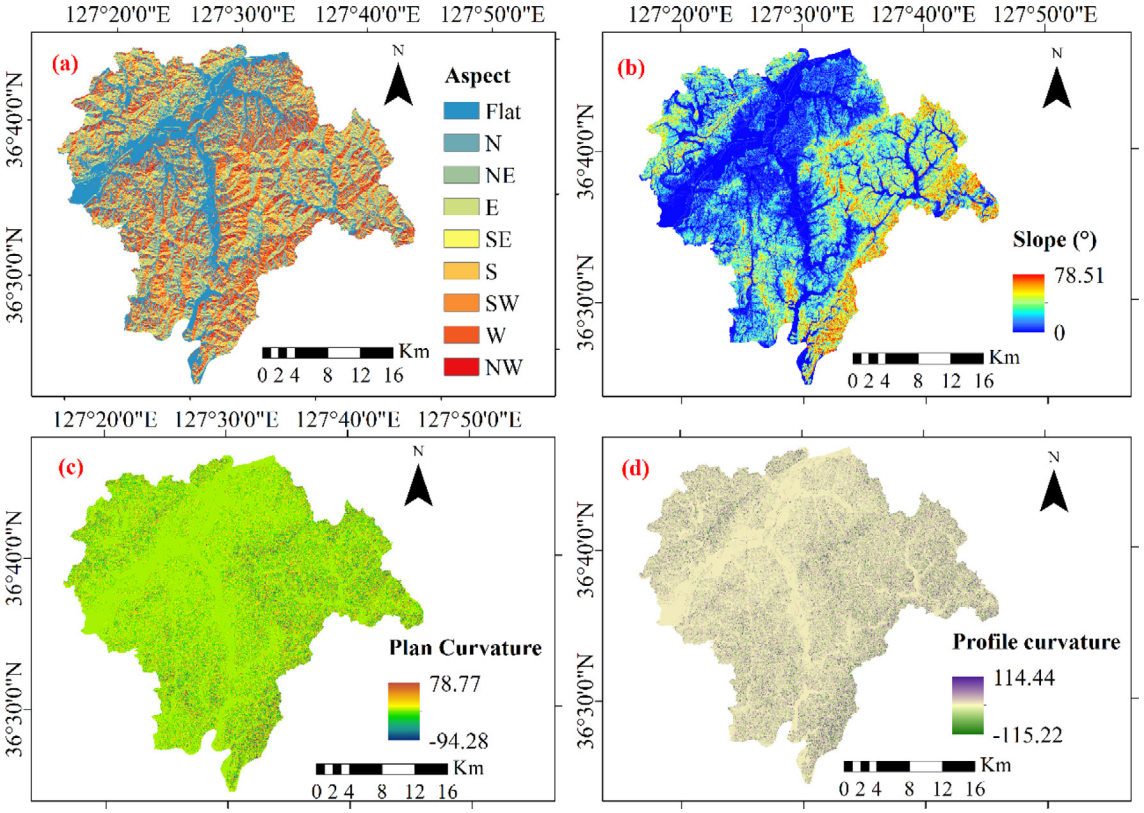


Fig. 6. Digital elevation model attributes used in the study: (a) aspect (b) slope (c) plan curvature, and (d) profile curvature.

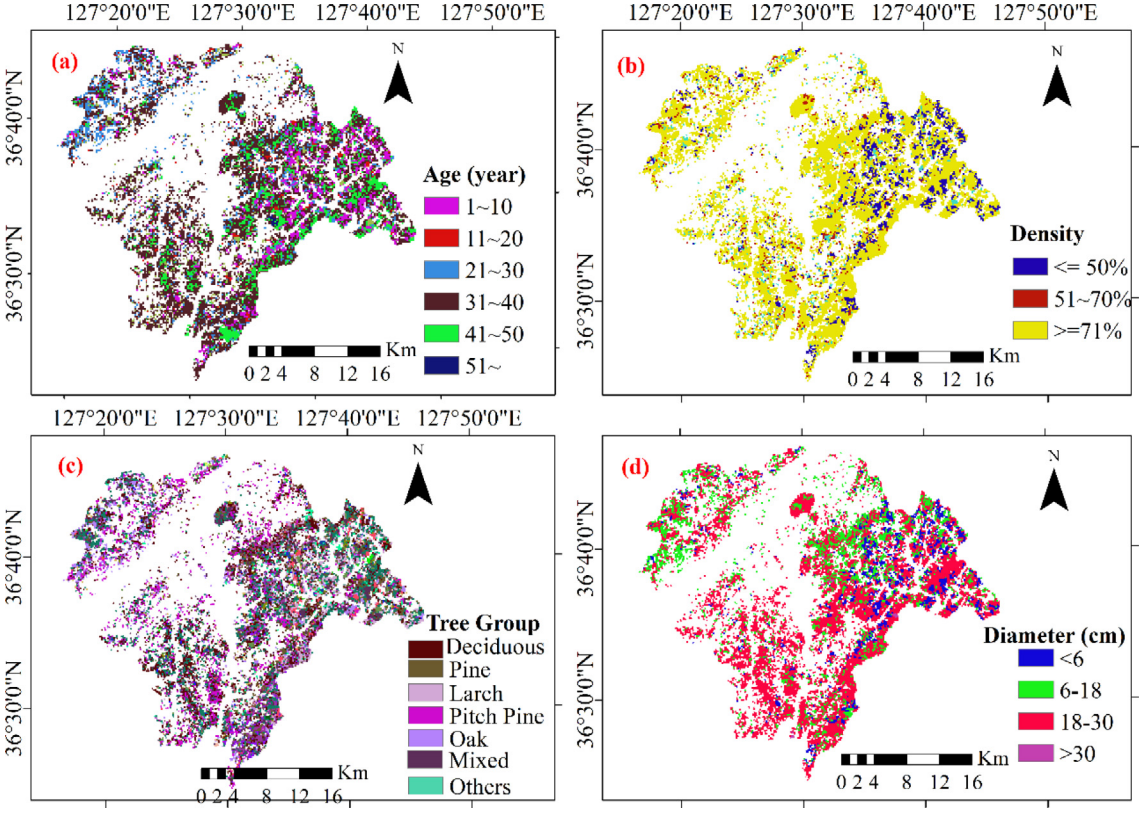


Fig. 7. Vegetation attributes used as variables in the study: (a) forest age, (b) forest density, (c) forest type, and (d) timber diameter.

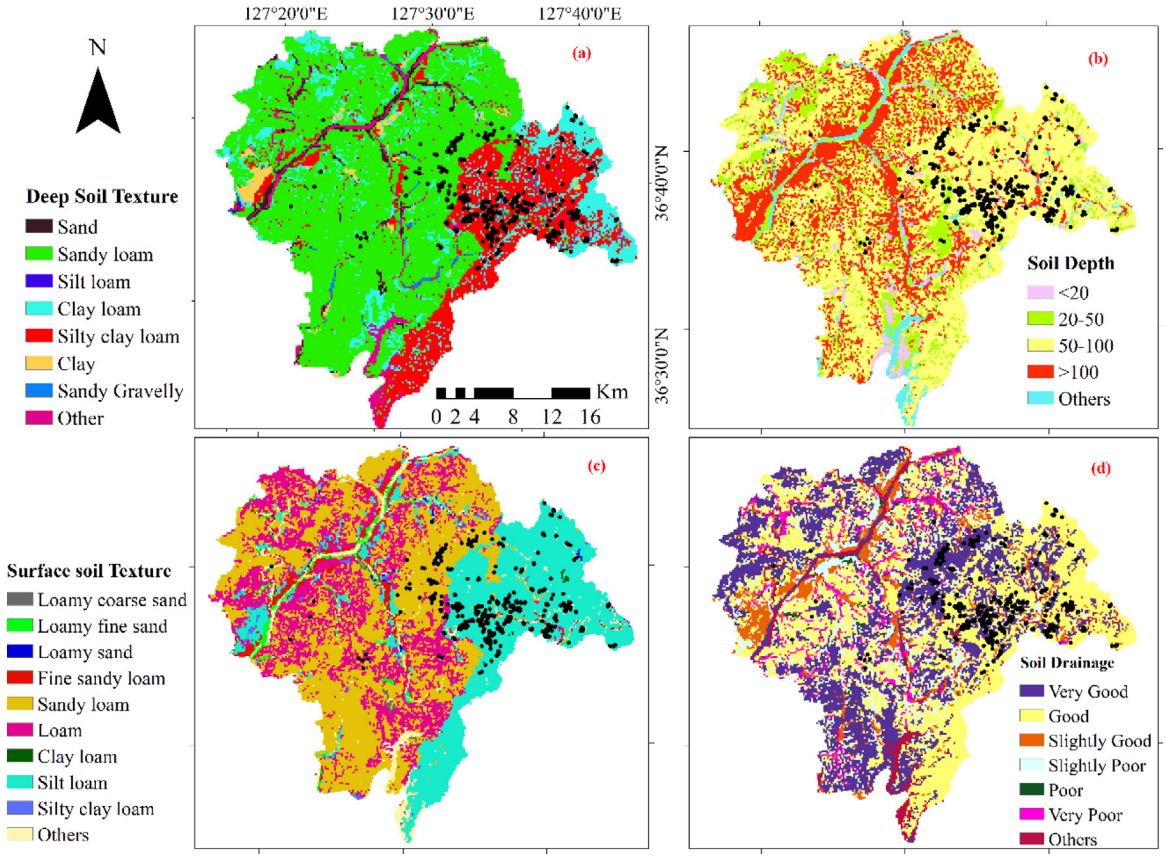


Fig. 8. Soil attributes used as variables in the study: (a) deep soil texture, (b) soil depth, (c) surface soil texture, and (d) soil drainage.

role in soil movement, as it has a strain softening behavior. Soil movement is influenced by basic soil properties, such as cohesion, shear strength, and the friction angle [43].

3. Models and techniques

This section discusses the application of a CNN and SHAP in the present study, details the architecture used, and demonstrates how SHAP helps with interpreting the results.

3.1. Convolutional Neural Network (CNN)

A CNN is a deep learning algorithm with characteristics that include hierarchical features, local connectivity, and shared weights [44]. Owing to these features, a CNN can hierarchically extract low-, middle-, and high-level image features. It comprises three main layers: a convolutional layer that reads inputted data sequences and automatically extracts relevant features, a pooling layer that extracts important features and focuses on important variables, and a fully connected layer that interpret the internal representation and puts out a vector representing multiple time steps [19]. Given the low number of features in the input data, a single convolution layer has been applied to avoid the problem of over smoothing [45].

The first layer behaves as a feature extractor that extracts the feature maps related to the target variable. It learns trainable filters to extract local information from the input matrix. The convolutional layer also uses an activation layer to add non-linearity. The selection of activation functions is vital in helping the network discern complex patterns in the data. The final layer conserves important information and reduces the number of parameters, particularly when large images are fed as input. Parameters can be pooled in several ways, such as the maximum, sum, average, etc, with maximum being used in this study.

3.2. Support Vector Machines (SVM)

SVM is a robust supervised technique derived from statistical learning theory and the principle of structural risk minimization [46]. The architecture was originally developed for classification works, and later extended to regression tasks. The motive of structural risk minimization is to minimize the upper bound of generalization error [46]. The objective of the model is to identify a hyperplane in a-dimensional plane, where a is the number of features, that clearly classifies the data points [47]. Given x as landslide affecting variables, Eq. (1) shows the separating hyperplane.

$$x_i\,(w * y_i + a) \geq 1 - \xi_i \tag{1}$$

where, w is the hyperplane orientation in the feature space, a is the hyperplane offset distance from the origin, and ξ_i is the positive slack variable.

SVM has been long used in landslide modeling and is one of the earliest used approaches in this field, often considered as a benchmark model [6,9,48]. In the SVM model, the separation of hyperplane formations from a training dataset is the first step. This separation is created in the original space with n coordinates (x_i is the variable of vector x) between the points of two different classes. Pixels are assigned values of ± 1 above or below the hyperplane, where class $(+1)$ are landslide pixels and class (-1) are non-landslide pixels.

3.3. SHapley Additive exPlanations (SHAP)

The SHAP concept emerged in game theory, in which an individual's contribution is calculated in a collective game [49]. The goal was to distribute the combined gain among the players,

depending on their contributions and outcomes. Based on Shapley values, meaningful rewards are provided in an unbiased manner to an individual rather than having players rewarded equally, based on consistency, local accuracy, and the null effect [49]. This concept has re-emerged in ML problems following work by Lundberg and Lee [25], owing to their black box nature, and has gained prominence in the field, as it can help with deciphering model outcomes, thus providing greater transparency and promoting trust in the use of ML models.

The unique aspect of the estimation of Shapley values can be categorized into three properties, which are local accuracy, missingness and consistency [25]. Local accuracy refers to the extent of relevant features captured in the explanation, missingness refers to the determination of explanation changes between runs and consistency refers to the degree of explanations and predictions being coupled [50]. Readers are referred to Lundberg and Lee [25] for an in-depth mathematical understanding of these attributes.

The Shapley value is the mean of the marginal contributions for all possible feature permutations. The mathematical expression is as follows:

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left[v\left(S \cup \{i\}\right) - v(S) \right] \tag{2}$$

where $\emptyset_i$ is the contribution of feature i, N is the set containing all features, n is the number of features in N, S is the subset of N that contains feature i, and v(N) is the base value, denoting the predicted outcome for each feature in N without knowledge of the feature values.

The model outcome for each observation is estimated by adding the SHAP value of every feature for observation. For a model f and feature vector z, the model is defined as:

$$g\left(z'\right) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z_i', \tag{3}$$

where g is the explanation model, $z' \epsilon \{0, 1\}^M$ is the simplified feature vector of z (so $z = h_z(z')$). M is the number of features and $\emptyset_i$ can be obtained from Eq. (2). $\emptyset_0$ is the model output when all the features are absent ($z' = h_z(0)$). The CNN model was trained once, and the analysis is performed on the test dataset. The Shapley values are based on the idea that the outcome of each possible combination of players should be considered to determine the importance of a single variable. This leads to a possibility of 2^F models to be trained in the SHAP formula. However, the library developed by Slundberg uses approximations and samplings to solve this issue.

Based on the ML architecture, various SHAP explainers are available. For an in-depth understanding of different explainers and their use cases, readers are referred to Molnar [51]. In this study, we used DeepExplainer, which is an enhanced version of the DeepLIFT algorithm that approximates the conditional expectations of SHAP values based on a selection of background samples. The sampling explainer is an extension of the Shapley sampling values explanation method [52].

4. Application of models

In the present work, one-dimensional (1D) CNN architecture was used, wherein the input data were considered to comprise an image with each pixel containing values of the landslide conditioning factors. Each input cell's data were represented via a column vector with the length equaling the number of variables. Each vector element corresponded to a variable's value. Landslide and non-landslide pixels were allocated values of 1 and 0, respectively, and the triggering factors were overlaid. Thereafter, all essential information was extracted for landslide and non-landslide locations and split into two parts—70% for training and 30% for testing—which is the most commonly applied split ratio for landslide modeling [6,11]. To compare the model results, SVM model was used where in, radial basis function (RBF) kernel was used, which has been proven to provide accurate results in classification tasks and is a popular choice in the field [48]. The kernel is expressed as:

$$K\left(\left(M_i, M_j\right)\right) = \exp\left(\left(-\Upsilon \left\| M_i - M_j \right\| 2\right)\right) \tag{4}$$

where $K\left(\left(M_i, M_j\right)\right)$ is the kernel function and Υ is a kernel parameter, which was set to 1.0.

Thereafter, landslide susceptibility map is produced, validated and assessed using the area under receiver operating characteristics curve. Based on the best performing model, the trained model is fed into SHAP algorithm, to identify the most impactful features and their interdependence. The methodology used in the present study is illustrated in Fig. 9.

One of the key steps in developing a neural network model is hyper parameter tuning [19]. The sizes of the convolutional and pooling layers reflect the scale of model operations. The activation function defines the weighted sum of the input and approximates any non-linear function. The ReLU function was used in the study to introduce non-linearity [53]. The loss function measures the inconsistency between the predicted and observed values; herein, we used the categorical cross-entropy loss function [53,54]. We further used the Adam optimizer with standard β values of 0.01 (momentum) and 0.001 (learning rate) for a momentumized gradient descent in our back propagation. To avoid overfitting, a dropout layer of 0.5 was added [55], and the model was applied with Keras [56] with TensorFlow as a backend. The number of epochs was 500. The CNN architecture used in the study is illustrated in Fig. 10.

The model's performance was examined based on receiver operating characteristics (ROC), an approach that is commonly applied in geohazard modeling [57]. It is based on the curve representing sensitivity plotted on the ordinate axis versus specificity plotted on the abscissa. Sensitivity refers to the landslide pixels correctly identified, whereas specificity refers to the non-landslide pixels correctly identified, which is based on a confusion matrix [57]. Accordingly, the area under the curve represents the overall model performance.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

where TP (true positive) and TN (true negative) are the number of grid cells that were correctly classified, and FP (false positive) and FN (false negative) are the number of grid cells that were incorrectly classified.

The area under the ROC curve (AUC) is used to assess the model's prediction quality by analyzing its ability to predict the occurrence or non-occurrence of events. Specifically, an AUC value of 1 indicates perfect agreement between actual and modeled data, whereas a value of 0.5 indicates inaccuracy in the model (random fit) [57,58]. After the CNN and SVM model were run, landslide susceptibility maps were developed, and the area was divided into five risk classes based on the natural breaks (Jenks) method: very high, high, moderate, low, and very low [23].

To understand the interactions and importance of variables in data-driven modeling, partial dependence plots or bar plots are typically generated. In the case of SHAP, dependence plots illustrate the impacts of variable relationships better than conventional approaches and may be considered a better alternative [27,28]. However, several plots can be constructed based on

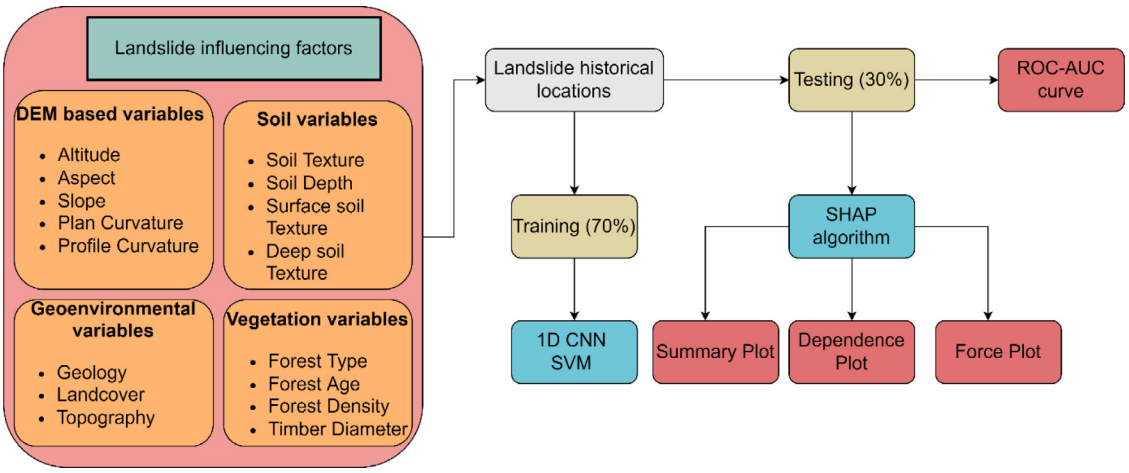


Fig. 9. XAI methodology used in the present study.

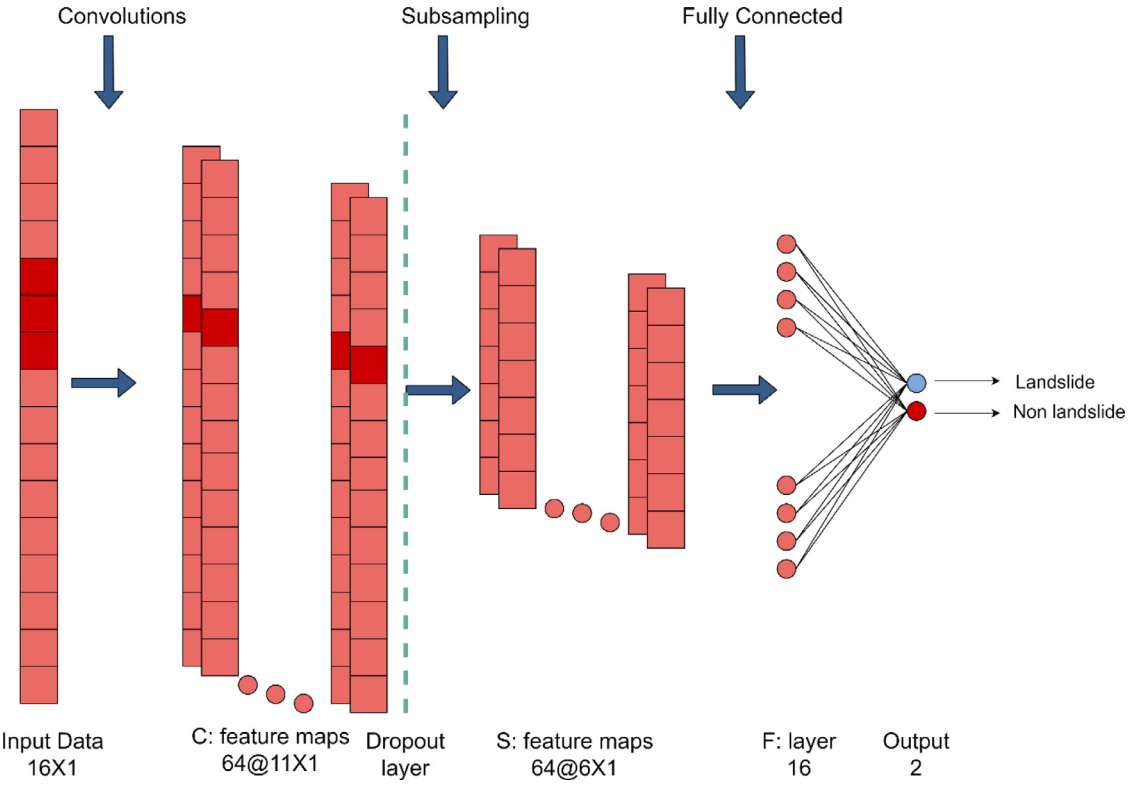


Fig. 10. Schematic illustration of the 1D CNN architecture used in the study.

Table 2
Confusion matrix of the models for the testing dataset.

Model		Landslide (1)	Non-landslide (0)	Accuracy
CNN	Landslide (1)	125	20	0.892
	Non-landslide (0)	31	298	
SVM	Landslide (1)	110	27	0.846
	Non-landslide (0)	46	291	

Shapley values. These include a summary plot that explains the cumulative effect of the variables, a dependence plot in which the effect of a single feature on the model predictions is plotted, an individual force plot that explains the effect of individual variables for a single observation, and a collective force plot that is a combination of all the force plots rotated at 90° and stacked horizontally to provide a single plot [25,51]. In the present study, we focused on summary plots and force plots. SHAP summary plots can replace conventional bar plots in examinations of global significance, whereas local explanations can be analyzed based on force plots [28].

5. Results

This section discusses the model results and the explainability of the CNN model with respect to achieving an accurate landslide susceptibility map. After model training, each grid was assigned a susceptibility index. Thereafter, weights were assigned to each factor class, and the susceptibility map was developed in an ArcGIS environment. For visualization purposes, the values were reclassified into five different categories using the natural breaks

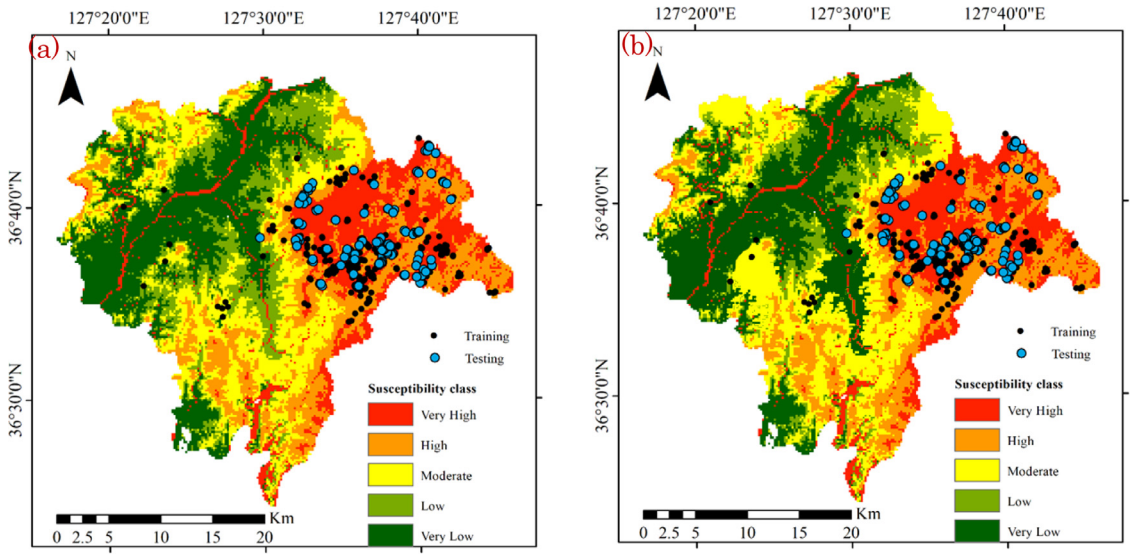


Fig. 11. Landslide susceptibility map developed using (a) CNN and (b) SVM model.

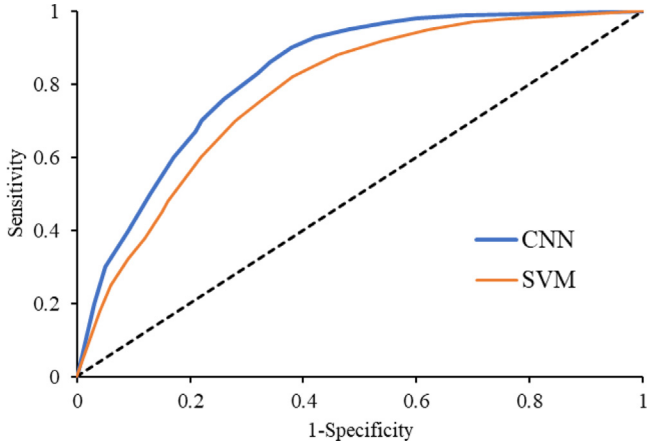


Fig. 12. ROC-AUC plot derived from the CNN model.

method (Fig. 11a, b). The model's ROC curve using the test set is illustrated in Fig. 12. The CNN model achieved an accuracy of 89.2%, whereas SVM model achieved an accuracy of 84.6% (see Table 2).

As the summary plot shows (Fig. 13), the altitude and surface soil texture are the most influential factors contributing to landslide incidence. In the case of altitude, the observations on the right-hand side of the plot rendered in blue indicate the negative correlation between altitude and landslide probability, i.e., the higher the altitude, the lower the chance of landslide occurrence in the study area. However, the relationship with slope was positive, meaning that as the slope increases, the probability of landslide occurrence increases as well. Altitude is widely known to be a key factor in landslide events. However, the SHAP plots suggest that not only altitude but also slope play important roles. To investigate this further, we examined the SHAP dependence plot for these variables (Fig. 14).

Soil texture was positively correlated with landslide occurrence. Considering the numbers assigned to each class, the probability of landslide occurrence is greater in areas with the clayey soil type. However, as the soil type changes to sand, the probability of landslide occurrences decreases. Regarding land cover, the probability of landslide incidence is lower in residential and built-up areas compared to natural grassland areas and scattered forest areas. The plan curvature ranked third among all potential landslide-causing factors, and as it increases, the probability of landslide occurrence decreases.

As the dependence plot illustrates, altitudes below 300 m have larger SHAP values, meaning that most landslides tend to occur within this altitude range. The landslide inventory data revealed that around 56% of the landslides occurred at altitudes of less than 300 m. Moreover, over 44% of landslide incidences occurred at altitudes ranging between 200 and 300 m. Less than 10% of the landslide events occurred at altitudes greater than 400 m. Upon examining the relationship between slope and landslide events, 18% of the landslides occurred at locations with slope values higher than 30°, with a maximum slope of 46°.

The plots below are individual force plots for landslide and non-landslide locations that clarify how the variables might affect the model outcome. Fig. 15 shows the SHAP force plot for individual landslide and non-landslide locations. These individual force plots illustrate three important characteristics: the output value (values in black and bold font under f(x)) indicating the prediction probability for an observation, the base value indicating the mean of the prediction probability for the entire test dataset, and variables that reduce (blue) or increase (pink) the probability of landslide occurrence.

The probability of a landslide occurrence for a landslide pixel (marked in green) is 85%, and variables such as altitude, aspect, and tree group are predicted to increase the probability (indicated in pink). Conversely, variables such as slope and forest age class cause the predicted probability to decrease. At another location (marked in brown), the probability of no landslide occurrence is around 98%. At this location, factors such as tree group and forest age class improve the model outcome, whereas altitude decrease the model outcome. It is important to clarify that the values of the variables mentioned here are the actual values.

6. Discussion and future work

The present study applied a CNN architecture to develop a landslide susceptibility model for South Korea's CheongJu region. The deep learning model was compared with SVM, which can be considered as a benchmark ML model. A total of 519 landslide locations along with 16 landslide conditioning variables, of which 70% were used for training and 30% for testing, were used to

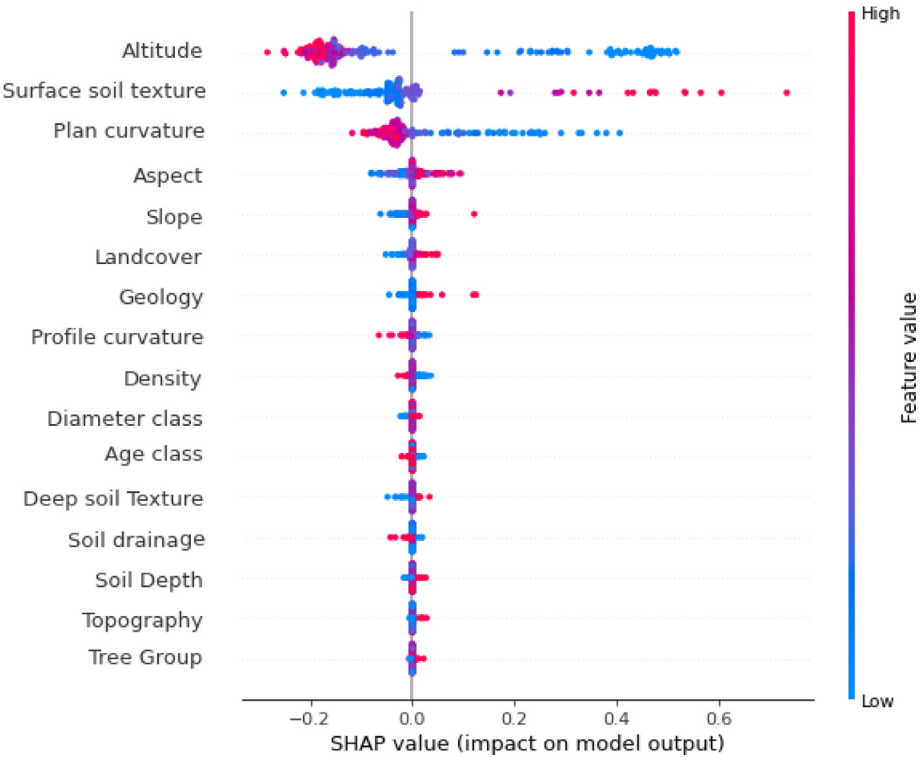


Fig. 13. SHAP summary plot of the test dataset.

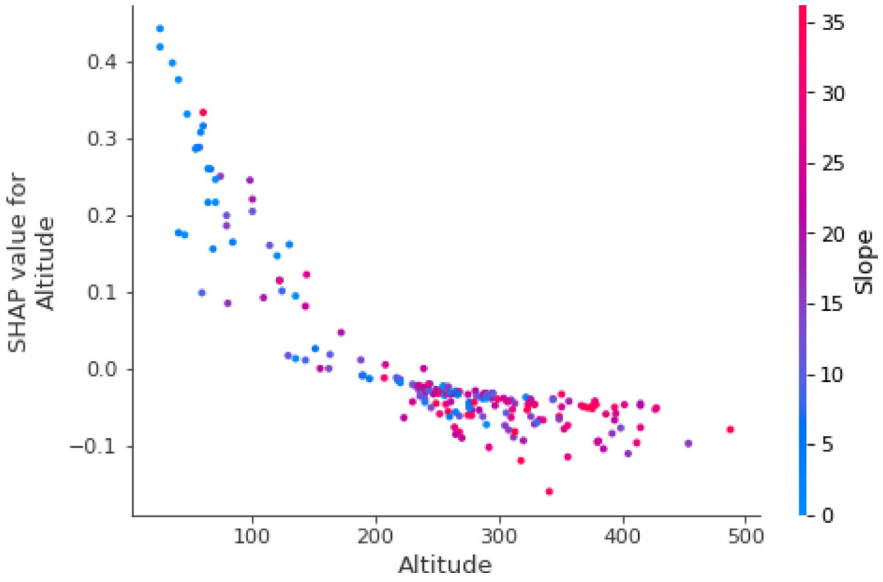


Fig. 14. SHAP dependence plot for altitude and slope.

develop the models. The model was validated using a ROC-based AUC approach, which revealed a model accuracy of 89.2% for CNN model and 84.6% for SVM model with the testing dataset. The results of the study show the capability of deep learning model to provide considerable improved results compared to benchmark models. Regarding previous studies in South Korea using CNNs, Lee et al. [59] examined the Mt. Umeyon region and achieved an accuracy of more than 99%. With regards to ML models, multiple studies have been conducted in South Korea. As an example, Kadavi et al. [60] studied landslide susceptibility using logistic regression and decision trees in Gangwon-do and achieved an accuracy of 90%. Lee et al. [11] achieved an accuracy of 80.1% using an artificial neural network for Inje, in the eastern part of South Korea, whereas Lee et al. [61] achieved an accuracy of 78.4% for the Seoul region. It is important to understand that variable selection was not conducted in this study to explore the importance of all the variables in the SHAP plots.

However, this paper's novel contribution is not limited to the development and application of the CNN model but also includes the introduction of the explainability concept for landslide studies. No mathematical definition for the concept of interpretability exists; rather, it can be explained as the degree to which a human can rationally understand the reasoning behind a decision and/or prediction based on a model output [62–64]. The more interpretable the model, the easier it is to contemplate why a specific decision is made. Ideally, the model would be self-explanatory,

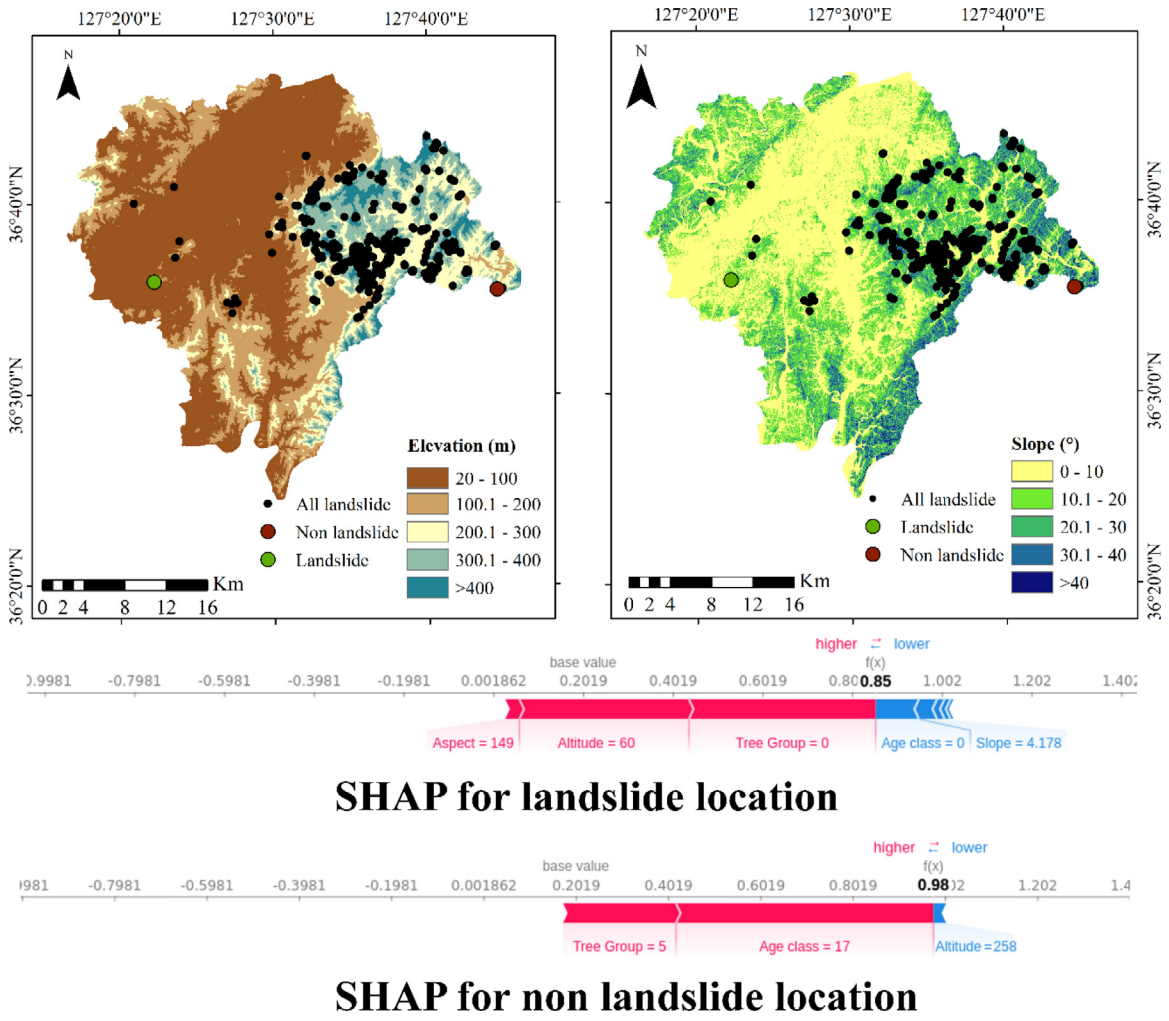


Fig. 15. SHAP individual force plots for landslide and non-landslide locations overlaid on the (a) elevation map and (b) slope map of the region.

which can be easily achieved for simple models, but with increasing complexity, particularly for ML-based models, this possibility is diminished [28]. Interpretable ML models are widely believed to be less accurate than complex deep neural networks. However, Rudin [65] discredited such notions and suggested that the notion of compensating for accuracy and explainability seems to be preventing users from developing XAI models. It must be borne in mind that a key difference exists between interpretable and explainable models. Hence, different explainable models exist (e.g., LIME [66], DeepLIFT [67], neural-backed decision trees [68], and SHAP [25]) that can provide explanations for model outcomes [25]. The SHAP framework provides values based on a co-operative game theory approach, which allows the model outcomes to be explained based on interactions among the variables used.

The study highlights the benefits of using SHAP by precisely characterizing model prediction outcomes at various geographic coordinates. As this is an introductory work, we have restricted ourselves to demonstrating SHAP's potential with respect to landslide susceptibility modeling. In the present work, we focused on the use of SHAP plots to clarify variations across the entire test dataset and individual locations for both landslide and non-landslide locations. Three different SHAP plots were developed for the testing dataset, which are (a) SHAP summary plot; (b) SHAP dependence plot and (c) SHAP force plot. The summary plot ranks the variables in terms of their importance and depicts the values which have contributed towards increasing or decreasing the model outcome. In the case of dependence plot, the variables considered were altitude and slope. The plots show that over 40% of landslides happened at altitudes between 200 m–300 m, whereas less than 10% of landslides occurred at heights more than 400 m. In case of relationship between slope and landslide events, less than 20% of landslides occurred with slopes more than 30°. The SHAP force plot reveals individual contribution for a specific geolocation, in this case we examined one known landslide pixel and non-landslide pixel. The results show the prediction probability and the contribution of different variables in increasing or decreasing the model outcome.

Most of the discussion was restricted to two variables—altitude and slope—which are typically considered the most important variables in landslide studies. However, researchers are encouraged to analyze different variables at various locations. Such detailed analysis of all the landslide variables lies beyond the scope of this study. As this constitutes an introduction to the use of SHAP in landslide studies, the study has some limitations. First, the study did not consider variable selection to showcase the full benefits of SHAP outcomes. Second, the present work does not explain outcomes from a true spatial context but rather, limits the explanation to certain locations.

The benefits of using SHAP are manifold: for example, it can be used for variable selection, as demonstrated by Matin and Pradhan [26], who used SHAP plots to eliminate non-influencing variables for earthquake damage mapping. Their study also revealed significant variation from conventional variable importance plots.

These plots explicitly examine how a model achieves a specific output based on each observation and, by using the plots, one can also examine spatial influence. With such benefits, we argue that future studies involving ML models should include explainability in several ways for landslide studies as well as in wider hydrological applications to promote trust among stakeholders.

7. Conclusions

An accurate and robust landslide susceptibility map is crucial for the implementation of effective landslide mitigation strategies. Despite considerable advancement in the use of ML models in susceptibility modeling, their use by government agencies or stakeholders remains minimal owing to their "black box" nature. Therefore, the use of explainable ML models to increase transparency in model outcomes has recently gained traction. This paper introduces the use of an explainable ML model for landslide susceptibility modeling in South Korea's CheongJu region. The study developed and compared CNN model with SVM model to develop a susceptibility map. The model outcomes of the CNN model was interpreted using SHAP, analyzing global and local interdependencies. A total of 519 points were used, of which 70% was used for training while the remainder was used for testing. This study's main findings are as follows:

- The CNN model achieved an accuracy of 89%, compared to 84% by SVM model proving the method's effectiveness in developing an accurate and robust landslide susceptibility model.
- The paper introduces the use of SHAP in landslide studies, and different plots were illustrated, including a summary plot, a dependence plot, and individual force plots. Based on the SHAP summary plot, the most important variables were altitude, surface soil texture, plan curvature, aspect, and slope.
- The study found that landslide occurrences were negatively correlated with altitude but positively correlated with slope.
- Future work involving ML models should consider the use of explainable models for landslide modeling.

CRediT authorship contribution statement

Biswajeet Pradhan: Conceptualization, Methodology, Modelling, Writing – review & editing, Funding, Writing – original draft. **Abhirup Dikshit:** Writing – original draft, Methodology, Modelling. **Saro Lee:** Supervision, Validation, Visualization, Data curation, Funding, Resources. **Hyesu Kim:** Validation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Funding

This research was supported by the Centre for Advanced Modelling and Geospatial Information Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, South Korea. Also, this research was supported by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM) and the National Research Foundation of Korea (NRF) grant funded by Korea government (MSIT) (No. 2023R1A2C1003095).

References

[1] D. Petley, Global patterns of loss of life from landslides, Geology 40 (10) (2012) 927–930.
[2] U. Haque, P. Blum, P.F. da Silva, P. Anderson, J. Pilz, S.R. Chalov, J.-P. Malet, M.J. Auflič, N. Andres, E. Poyiadji, P.C. Lamas, W. Zhang, I. Peshevski, H.G. Petursson, T. Kurt, N. Dobrev, J.C. Garcia-Davalillo, M. Halkia, S. Ferri, G. Gaprindashvili, J. Engstrom, D. Keellings, Fatal landslides in Europe, Landslides 13 (6) (2016) 1545–1554.
[3] S.L. Gariano, F. Guzzetti, Landslides in a changing climate, Earth Sci. Rev. 162 (2016) 227–252.
[4] H.G. Kim, D.K. Lee, C. Park, Assessing the cost of damage and effect of adaptation to landslides considering climate change, Sustainability 10 (5) (2018) 1628.
[5] P. Reichenbach, M. Rossi, B.D. Malamud, M. Mihir, F. Guzzetti, A review of statistically-based landslide susceptibility models, Earth-Sci. Rev. 180 (2018) 60–91.
[6] A. Merghadi, A.P. Yunus, J. Dou, J. Whiteley, B.T. Pham, D.T. Bui, R. Avtar, B. Abderrahmane, Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance, Earth-Sci. Rev. 207 (2020) 103225.
[7] C.J. van Westen, E. Castellanos, S.L. Kuriakose, Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview, Eng. Geol. 102 (3–4) (2008) 112–131.
[8] J. Mathew, V.K. Jha, G.S. Rawat, Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method, Landslides 6 (2009) 17–26.
[9] A. Dikshit, B. Pradhan, A.M. Alamri, Pathways and challenges of the application of artificial intelligence to geohazards modelling, Gondwana Res. 100 (2021) 290–301.
[10] A. Dikshit, R. Sarkar, B. Pradhan, S. Segoni, A.M. Alamri, Rainfall induced landslide studies in Indian himalayan region: A critical review, Appl. Sci. 10 (2020) 2466.
[11] S. Lee, S.W. Jeon, K.Y. Oh, M.J. Lee, The spatial prediction of landslide susceptibility applying artificial neural network and logistic regression models: a case study of Inje, Korea, Open Geosci. 8 (1) (2016) 117–132.
[12] C. Xu, F. Dai, X. Xu, Y.H. Lee, GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China, Geomorphology 145–146 (2012) 70–80.
[13] Y. Huang, L. Zhao, Review on landslide susceptibility mapping using support vector machines, Catena 165 (2018) 520–529.
[14] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, D.T. Bui, D. Zhao, J. Ma, A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility, Catena 151 (2017) 147–160.
[15] E.K. Sahin, I. Colkesen, T. Kavzoglu, A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping, Geocarto Int. (2018) 1–23.
[16] C. Andrieu, N. De Freitas, A. Doucet, M.I. Jordan, An introduction to MCMC for machine learning, Mach. Learn. 50 (2003) 5–43.
[17] D. Silver, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (2016) 484–489.
[18] M. Moravčík, et al., DeepStack: expert-level artificial intelligence in heads-up no-limit poker, Science 356 (2017) 508–513.
[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436.
[20] Z. Chen, Y. Zhang, C. Ouyang, J. Ma, Automated landslides detection for mountain cities using multi-temporal remote sensing imagery, Sensors 18 (2018) 821.
[21] N. Prakash, A. Manconi, S. Leow, Mapping landslides on EO data: Performance of deep learning models vs. traditional machine learning models, Remote Sens. 12 (3) (2020) 346.
[22] M.I. Sameen, B. Pradhan, S. Lee, Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment, Catena 186 (2020) 104249.
[23] Y. Wang, Z. Fang, H. Hong, Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China, Sci. Total Environ. 666 (2019) 975–993.
[24] U. Ozturk, M. Pittore, R. Behling, S. Rossner, L. Andreani, O. Korup, How robust are landslide susceptibility estimates? Landslides 18 (2021) 681–695.
[25] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017, arXiv preprint arXiv:1705.07874.
[26] S.S. Matin, B. Pradhan, Earthquake-induced building-damage mapping using explainable AI (XAI), Sensors 21 (2021) 4489.
[27] A. Abdollahi, B. Pradhan, Urban vegetation mapping from aerial imagery using explainable AI (XAI), Sensors 21 (2021) 4738.
[28] M.V. García, J.L. Aznarte, Shapley additive explanations for NO2 forecasting, Ecol. Inform. 56 (2020) 101039.
[29] A. Dikshit, B. Pradhan, Interpretable and Explainable AI (XAI) model for spatial drought prediction, Sci. Total Environ. 801 (2021) 149797.

[30] S.-G. Lee, M.G. Winter, The effects of debris flow in the Republic of Korea and some issues for successful risk reduction, Eng. Geol. 251 (2019) 172–189.
[31] S.G. Lee, S.R. Hencher, Slope safety and landslide risk management practice in Korea, in: K. Ho, V. Li (Eds.), Proceedings, 2007 International Forum on Landslide Disaster Management, Vol. 1, Hong Kong, 2007, pp. 125–168.
[32] M.E.A. Budimir, P.M. Atkinson, H.G. Lewis, A systematic review of landslide probability mapping using logistic regression, Landslides 12 (3) (2015) 419–436.
[33] D.J. Varnes, Landslide hazard zonation: a review of principle and practice Nat. Hazards, 3, UNESCO Press, Paris, 1984.
[34] C. Bartelletti, R. Giannecchini, G. D'Amato Avanzi, Y. Galanti, A. Mazzali, The influence of geological–morphological and land use settings on shallow landslides in the Pogliaschina T. basin (northern Apennines, Italy), J. Maps 13 (2) (2017) 142–152.
[35] H.-Y. Lee, H.-S. Yun, J.-D. Lee, Geological Report of the Pyongchon Sheet, Korea Institute of Energy and Resources, 1989, http://dx.doi.org/10.22747/data.20210702.4278.
[36] C.H. Lee, M.S. Lee, B.S. Park, Explanatory Texture of Geological Map of Miweon Sheet, Korea Research Institute of Geoscience and Mineral Resources, 1980, http://dx.doi.org/10.22747/data.20210702.4283.
[37] Y.-I. Kwon, M.S. Jin, Explanatory Texture of the Geological Map of Cheongju Sheet, Geological and Mineral Institute of Korea, 1974, http://dx.doi.org/10.22747/data.20210702.4277.
[38] M.J. Froude, D.N. Petley, Global fatal landslide occurrence from 2004 to 2016, Nat. Hazards Earth Syst. Sci. 18 (2018) 2161–2181.
[39] L. Fan, P. Lehmann, D. Or, Effects of soil spatial variability at the hillslope and catchment scales on characteristics of rainfall-induced landslides, Water Resour. Res. 52 (3) (2016) 1781–1799.
[40] C. Moos, P. Bebi, F. Graf, J. Mattli, C. Rickli, M. Schwarz, How does forest structure affect root reinforcement and susceptibility to shallow landslides? Earth Surf. Process. Landf. 41 (7) (2015) 951–960.
[41] B. Reubens, J. Poesen, F. Danjon, G. Geudens, B. Muys, The role of fine and coarse roots in shallow slope stability and soil erosion control with a focus on root system architecture: a review, Trees 21 (4) (2007) 385–402.
[42] C. Montzka, M. Herbst, L. Weihermüller, A. Verhoef, H. Vereecken, A global data set of soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity curves, Earth Syst. Sci. Data 9 (2017) 529–543.
[43] R.C. Sidle, H. Ochiai, Landslides Processes, Prediction, and Land Use, American Geophysical Union, Washington DC, ISBN: 978-0-87590-322-4, 2006.
[44] Y. Lecun, L. Bottou, Y. Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[45] C. Chen, X. Chen, H. Cheng, On the over-smoothing problem of CNN based disparity estimation, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8996–9004.
[46] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[47] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, New York, 2013.
[48] H.-J. Oh, P.R. Kadavi, C.-W. Lee, S. Lee, Evaluation of landslide susceptibility mapping by evidential belief function, logistic regression and support vector machine models, Geomat. Nat. Hazards Risk 9 (1) (2018) 1053–1070.
[49] L.S. Shapley, A value for n-person games, in: Contributions to the Theory of Games, Vol. 2, 1953, pp. 307–317.
[50] A. Amich, B. Eshete, Explanation-guided diagnosis of machine learning evasion attacks, in: J. Garcia-Alfaro, S. Li, R. Poovendran, H. Debar, M. Yung (Eds.), Security and Privacy in Communication Networks. SecureComm 2021, in: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 398, Springer, 2021.
[51] C. Molnar, Interpretable Machine Learning, Lulu Press, Morrisville, NC, USA, 2020.
[52] K. Zhang, P. Xu, J. Zhang, Explainable AI in deep reinforcement learning models: A SHAP method applied in power system emergency control, in: 4th IEEE Conference on Energy Internet and Energy System Integration. October 30-November 1, 2020, Wuhan, China, 2020.
[53] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117.
[54] J. Brownlee, Supervised and unsupervised machine learning algorithms, in: Machine Learning Mastery, 2016, 16.
[55] N. Srivastava, G. Hinton, A. Krizhevsky, et al., Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[56] C. Francois, Keras, Github, 2015, https://github.com/keras-team/keras.
[57] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874.
[58] B. Choubin, E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini, A. Mosavi, An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines, Sci. Total Environ. 651 (2) (2019) 2087–2096.
[59] S. Lee, W.-K. Baek, H.-S. Jung, S. Lee, Susceptibility mapping on urban landslides using deep learning approaches in Mt. Umyeon, Appl. Sci. 10 (2020) 8189.
[60] P.R. Kadavi, C.W. Lee, S. Lee, Landslide-susceptibility mapping in Gangwon-do, South Korea, using logistic regression and decision tree models, Environ. Earth Sci. 78 (116) (2019).
[61] S. Lee, M.-J. Lee, H.-S. Jung, Data mining approaches for landslide susceptibility mapping in Umyeonsan, Seoul, South Korea, Appl. Sci. 7 (7) (2017) 683.
[62] B. Kim, R. Khanna, O.O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Advances in Neural Information Processing Systems, 2016, pp. 2280–2288.
[63] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
[64] N. Choudhary, C.C. Aggarwal, K. Subbian, C.K. Reddy, Self-supervised short text modeling through auxiliary context generation, ACM Trans. Intell. Syst. Technol. 1 (2022) 1, http://dx.doi.org/10.1145/3511712.
[65] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.
[66] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[67] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, 2017, pp. 3145–3153.
[68] A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, H. Jin, S. Petryk, S.A. Bargal, J.E. Gonzalez, NBDT: Neural-backed decision trees, 2020, arXiv preprint arXiv:2004.00221.