Geoscience Frontiers

Spatial Landslide Susceptibility Assessment Using Machine Learning Techniques Assisted by Additional Data Created with Generative Adversarial Networks

--Manuscript Draft--

Manuscript Number:	GSF-D-20-00191R2
Article Type:	Research Paper
Keywords:	Landslide susceptibility; Machine learning; Generative adversarial networks; Convolutional neural networks; Geographic information systems
Corresponding Author:	Biswajeet Pradhan, PhD University Technology Sydney Sydney, AUSTRALIA
First Author:	Husam A. H. Al-Najjar
Order of Authors:	Husam A. H. Al-Najjar
	Biswajeet Pradhan, PhD
Abstract:	In recent years, landslide susceptibility mapping has substantially improved with advances in machine learning. However, there are still challenges remain in landslide mapping due to the availability of limited inventory data. In this paper, a novel method that improves the performance of machine learning techniques is presented. The proposed method creates synthetic inventory data using generative adversarial networks (GANs) for improving the prediction of landslides. In this research, landslide inventory data of 156 landslide locations were identified in Cameron Highlands, Malaysia, taken from previous projects the authors worked on. Altitude, slope, aspect, plan curvature, profile curvature, total curvature, lithology, land use and land cover, distance from the road, distance from the stream, stream power index, sediment transport index, terrain roughness index, topographic wetness index and vegetation density are geo-environmental factors considered in this study based on suggestions from previous works on Cameron Highlands. To show the capability of GANs in improving landslide prediction models, this study tests the proposed GAN model with artificial neural network (ANN), support vector machine (SVM), decision trees (DT), random forest (RF) and bagging ensemble models with ANN and SVM models. These models were validated using the area under the receiver operating characteristic curve (AUROC). The DT, RF, SVM, ANN and Bagging ensemble could achieve the AUROC values of (0.90, 0.94, 0.86, 0.69 and 0.82) for the training; and the AUROC of (0.76, 0.81, 0.85, 0.72 and 0.75) for the test, subsequently. When using additional samples, the models achieved the AUROC values of (0.92, 0.94, 0.88, 0.75 and 0.84) for the training and (0.78, 0.82, 0.82, 0.78 and 0.80) for the test, respectively. Without the use of additional samples created by the GAN model, SVM achieved the highest AUROC of 0.85, whereas ANN had the lowest AUROC of 0.72. RF and SVM achieved AUROC of 0.82 when the additional samples were used for training these models. Using the additional samples improved the test accuracy of all the models except SVM. As a result, in data-scarce environments, this research showed that utilizing GANs to generate supplementary samples is promising because it can improve the predictive capability of common landslide prediction models.

Spatial Landslide Susceptibility Assessment Using Machine Learning Techniques Assisted by Additional Data Created with Generative Adversarial Networks

Husam A. H. Al-Najjar [a] and Biswajeet Pradhan [a,b]*

[a] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, 2007 NSW, Australia; Husam.al-najjar@student.uts.edu.au); Biswajeet.Pradhan@uts.edu.au

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

*Corresponding author Email: biswajeet24@gmail.com or Biswajeet.Pradhan@uts.edu.au

Abstract

In recent years, landslide susceptibility mapping has substantially improved with advances in machine learning. However, there are still challenges remain in landslide mapping due to the availability of limited inventory data. In this paper, a novel method that improves the performance of machine learning techniques is presented. The proposed method creates synthetic inventory data using Generative Adversarial Networks (GANs) for improving the prediction of landslides. In this research, landslide inventory data of 156 landslide locations were identified in Cameron Highlands, Malaysia, taken from previous projects the authors worked on. Elevation, slope, aspect, plan curvature, profile curvature, total curvature, lithology, land use and land cover (LULC), distance to the road, distance to the river, stream power index (SPI), sediment transport index (STI), terrain roughness index (TRI), topographic wetness index (TWI) and vegetation density are geo-environmental factors considered in this study based on suggestions from previous works on Cameron Highlands. To show the capability of GANs in improving landslide prediction models, this study tests the proposed GAN model with benchmark models namely Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF) and Bagging ensemble models with ANN and SVM models. These models were validated using the area under the receiver operating characteristic curve (AUROC). The DT, RF, SVM, ANN and Bagging ensemble could achieve the AUROC values of (0.90, 0.94, 0.86, 0.69 and 0.82) for the training; and the AUROC of (0.76, 0.81, 0.85, 0.72 and 0.75) for the test, subsequently. When using additional samples, the same models achieved the AUROC values of (0.92, 0.94, 0.88, 0.75 and 0.84) for the training and (0.78, 0.82, 0.82, 0.78 and 0.80) for the test, respectively. Using the additional samples improved the test accuracy of all the models except SVM. As a result, in data-scarce environments, this research showed that utilizing GANs to generate supplementary samples is promising because it can improve the predictive capability of common landslide prediction models.

Keywords: Landslide susceptibility, Inventory, Machine learning, Generative adversarial network, Convolutional neural network, Geographic information system

1. Introduction

Natural hazards are major challenges worldwide, and many countries are spending a significant amount of their yearly budget to control and prevent them. Landslides pose a serious risk to human habitats. The risk of landslides is a major barrier to agricultural and urban development practices. In addition, ongoing urbanization is placing vast demands on infrastructure and escalating the threat to property and human lives. As a result, landslide hazard assessment has become a major step in planning the most suitable for risk mitigation measures. Experts frequently use the maps generated from this assessment to identify regions where thorough in-situ studies should be conducted. Landslide hazard assessment is a complex task that includes comprehension of the science of geotechnics, geomorphology, hydrology and statistics (Glade et al., 2012). This objective has motivated computational modeling studies, particularly the evaluation of landslide susceptibility. Statistical and physical models are often used to accomplish this task (Formetta et al., 2014).

Physical-based models combine susceptibility analysis with soil and rock mechanics, creating a physical basis for this method (Wang et al., 2019). They are appropriate at a local scale such as single slope, basin/ catchment and requires site-specific geotechnical data (Park et al., 2019). Generally, the infinite slope model is used in the analysis of slope stability with hydrological or earthquake models. Although reliable geotechnical parameters are essential for such models, lack of geotechnical data throughout a large scale area and the expensiveness remain the main obstacles in the physical-based models (Lee et al., 2014). Various landslides studies and assessments were carried out to develop landslide-prone areas in Malaysia (Fanos and Pradhan, 2019; Mezaal and Pradhan, 2018; Pradhan and Lee, 2010; Sameen et al., 2020).

Statistical models, also known as empirical models, use landslide inventories and other conditioning factors (e.g. terrain and land use), which can be extracted at large scales using remote sensing data and Geographical Information Systems (GIS). Such techniques have gained popularity in the field of landslide susceptibility assessment, especially when addressing the challenge of landslide mapping of prone areas at large scales, where enough geotechnics information is not available to perform physical-based models (Goetz et al., 2011).These models have also been supported by the latest progress in the availability and accessibility of remote sensing-based derived information, such as topography, land cover and precipitation products, thereby improving the application of the method at large scales.

Several scholars have evaluated various statistical models to assess landslide susceptibility (Akbar and Chen, 2018; Braun et al., 2018; Ciurleo et al., 2017; Goetz et al., 2015; Huang and Zhao, 2018; Kavzoglu et al., 2019; Süzen and Doyuran, 2004; Xiao et al., 2019; Zêzere et al., 2017). Early approaches to modeling landslide susceptibility are based on field investigations. Such techniques, however, are costly and site-specific, and they heavily involve extensive expertise in geology and geomorphology. Statistical approaches of landslide susceptibility modeling have become very popular during the last two decades. Recently, several scholars including (Kawabata and Bandibas, 2009; Lee and Sambath, 2006; Mandal and Mondal, 2019; Pradhan, 2013) evaluated several statistical models, such as frequency ratio (FR), logistic regression (LR), artificial neural network (ANN), certainty factor (CF), analytical hierarchy process (AHP) and fuzzy logic (FL). They suggested that ANN-, CF and FR-based FL are the most reliable techniques in assessing and predicting landslide susceptibility, at least for their case study. Regardless of the type of models and where they belong (statistical or machine learning), they are good for landslide susceptibility assessment of large areas. Statistical models can also be evaluated quantitatively at lower costs than evaluating a physical model. In addition, these models are computationally more efficient than physical models because the latter require simulations with numerous iterations to determine some geotechnical parameters that are used to prepare the susceptibility products. However, they have certain limitations, which include difficulties in explaining the results of the black box models and over-fitting in the case of limited training samples.

2. Related Works

Landslide susceptibility mapping has improved substantially during the last decade because of new data processing techniques such as sampling methods, machine learning models, and validation measures. Some studies have focused on sampling strategies, selection of training samples and addressing the effects of incomplete inventory datasets. In landslide susceptibility mapping, training data play a critical role in determining the accuracy and generalization of the model. The size of the training data has a significant effect on the accuracy of the susceptibility model. For example, in the training data under some sample threshold limits, Hussin et al. (2016) showed that model performance was very low, while the use of a large number of landslides above the threshold created a plateau effect, with no increase in model performances. Tsangaratos and Ilia (2016) also reported that the size of training data influences the prediction accuracy when using models such as LR and Naive Bayes.

Several studies have attempted to improve landslide susceptibility models by proposing new factors into the process including conditioning factors optimization (Al-Najjar et al., 2019; Canoglu et al., 2019; Dou et al., 2015; Kavzoglu et al., 2015; Kornejady et al., 2018; Samia et al., 2018; Soma et al., 2019). Moreover, model parameterization and integration methods have been studied to improve landslide susceptibility mapping. Statistical and machine learning models are often affected by the selection of proper hyper-parameters for a specific case study (Can et al., 2019; Feizizadeh et al., 2017). Moreover, the model's integration has also been active research for improving the landslide susceptibility in the last few years (Kalantar et al., 2018). Examples of model integration studies include ensemble models (Bragagnolo et al., 2020; Kadavi and Lee, 2018) and integration of data-driven and knowledge-based models (Ashournejad et al., 2019; Yan et al., 2019; Zhang et al., 2019).

Studies on sampling strategies for landslide susceptibility mapping have been active in recent years. Hussin et al. (2016) assessed different landslide sampling strategies (scarp centroid, points populating the scarp and entire scarp polygon) in a grid-based statistical model. These strategies achieve the highest performance when sampling shallow landslides as grid points and debris flow scarps as polygons. Yilmaz and Ercanoglu (2019) discussed the necessity of studying the selection of data mining techniques; they emphasized that sampling methods such as polygon features or seed cells representative pre-failure settings appear to be more genuine in obtaining truthful maps than other methods. Lai et al. (2019) also explored the influence of sampling strategies for improving landslide susceptibility mapping.

In addition to the size of training data and the sampling strategy, studies have investigated various ways of selecting training samples. Conoscenti et al. (2016) performed landslide susceptibility mapping through investigating the impact of landslide absence (negative samples) on the models; they extracted the landslide absence using randomly distributed circles that have a diameter equivalent to the mean width of the landslide source areas. Moreover, the individual grid cells were randomly distributed to distinguish the non-landslide zones (absence selection). Experiments from this study based on multivariate adaptive regression splines showed that absences selection using random circles are significantly better than the other method when learning and validation samples were extracted from the same area, and no significant difference was observed when testing the models outside the training area. Kalantar et al. (2018) evaluated the impact of landslide samples varieties on the SVM, LR and ANN methods; their investigation demonstrated that randomness in the training sample selection has a significant effect on the susceptibility models. The outcome showed that, in the section of training samples, the LR model is less sensitive than the SVM and ANN models. Zhu et al. (2019) proposed a method based on similarity sampling for absence selection; their experiments on a common machine learning models showed that this new method outperformed the existing methods, such as buffer control and target space exteriorization. Hong et al. (2019) assessed the impact of absence data selection on the RF model. Aktas and San (2019) developed a new automatic sampling method based on a two-level random sampling.

The impact of landslide inventory incompleteness on susceptibility mapping was also carried out in recent studies. Du et al. (2020) assessed landslide susceptibility in Tibet Chinese Himalayas, with a multinomial logistic regression model with reported average AUC of 0.867; however, there were some uncertainties in the landslide-prone areas defined by their AHP model. Steger et al. (2016) assessed the impact of spatially heterogeneous completeness of landslide information on statistical landslide susceptibility models (e.g. logistic regression) by artificially introducing two different mapping biases into available landslides and synthetically generated landslides. Although they reported AUROCs greater than 0.85, they suggested the method needed to be evaluated with other different models. In another study, Lee et al. (2018) employed optimized data mining and statistical methods for various scenarios considering limited inventories. In their model, SVM achieved the AUC of 0.85 when either the full or limited landslide inventories were used; however, generating additional inventories was not considered in their study. Steger et al. (2016) suggested that models directly associated with inventory-based incompleteness should be rejected regardless of their performance. Furthermore, they proposed using mixed-effects modeling if systematically missing landslide information can be attributed to a spatial variable (Steger et al., 2018).

The aforementioned studies indicate several ways to improve landslide susceptibility models, such as data-related methods and others that target the model construction and training process. This study aims to develop a new method for additional landslide sample creation with generative adversarial networks (GANs) which could be useful in the inventory-scarce environment. Several machine learning models, such as ANN, SVM, DT, RF and Bagging ensemble, with ANN and SVM as base classifiers, are used to evaluate the new method of landslide susceptibility mapping. These methods are compared in a case study in Cameron Highlands, Malaysia.

3. Study area and materials

3.1. Study area

The Cameron Highlands district, located in the state of Pahang, Malaysia (Fig. 1), was selected as a study area because it often experiences landslides and flash floods. These events are caused by heavy and prolonged rainfall causing significant damages to properties. In this tropical mountainous area, landslides are common as shown by government reports and past studies by (Matori and Basith, 2012; Pradhan and Lee, 2010).

From the geomorphology aspect, the region is characterized as hilly, and altitudes are in the range of 840–2110 meters (Sameen and Pradhan, 2019). The primary drainage characteristics of the area consist of two rivers, namely, the Bertam and the Telom. Considerable types of vegetation in Cameron Highlands include tropical forest and tea plantations, flower fields and temperate crops. Concerning lithology, the greater part of the region contains mega crystal biotite granites and phyllite as well as some schists layers (Pradhan and Lee, 2010). The area has a fair climate with an average annual rainfall starting from March to May and from November to December. The average nightly temperature of the study area is 14 °C, whereas the daily temperature reaches 24 °C. Approximately 8.0% (55 km^2) of the area is classified as cropland, 86% (600 km^2) is categorized as cultivated area, and 4.0% (27.5 km^2) represents as residential areas.

3.2. Landslide inventory map

Data-driven landslide susceptibility assessment requires landslide inventories for model training and validation. Landslide inventory can be prepared using field investigations, historical landslide events from news and government reports and remote sensing data analysis. In this investigation, landslide inventories were taken from the study compiled by (Mezaal and Pradhan, 2018; Pradhan and Lee, 2010; Sameen et al., 2020). Overall, 156 landslides were identified and verified in the study area.

3.3. Landslide conditioning factors

Fifteen conditioning factors including elevation, slope, aspect, plan curvature, profile curvature, total curvature (Fig. 2a-f), lithology, LULC, distance to road, distance to river, SPI, STI (Fig. 2g-l), TRI, TWI and vegetation density (Fig. 2m-o) were selected as geo-environmental factors because they have been widely used in landslide susceptibility studies (Al-Najjar et al., 2019; Can et al., 2019; Canoglu et al., 2019; Huang and Zhao, 2018; Lee and Sambath, 2006). The related data were obtained over the study area on 15 January 2015 by utilizing a light detection and ranging (LiDAR) airborne system with a specification of 25,000 HZ pulse frequency rate and a density of 8 points/m^2. Then, a one-meter spatial resolution of the digital elevation model was generated after removing non-ground points. Non-ground point removal was performed utilizing multi-scale curvature and inverse distance weighted interpolation approaches via ArcGIS Pro 2.4 software.

This study used six geomorphological factors, i.e. total curvature, plan curvature, profile curvature, slope, elevation and slope aspect in the susceptibility mapping given that landslides are influenced by terrain type. The elevation was included because it affects the extent of rock weathering and is used by many scholars for landslide susceptibility assessment (Ayalew and Yamagishi, 2005). The elevation of the investigation region was in the range from 690 to 1487 meters. The slope is another important factor, often included in landslide susceptibility studies (Kamp et al., 2008). The slope values ranged from 0° to 78.88°. We also included the slope direction (also known as slope aspect) because its task is to control concentrations of topographic wetness affected by precipitation and solar radiation. In addition, plan, profile and total curvature were also used (Ozdemir and Altural, 2013). In general, curvature affects slope instability. Plan curvature represents the curvature when it is vertical to the path of the highest slope. Profile curvature is parallel to the slope and designates the maximum slope orientation. It affects the speeding up and slowing down of stream movement (Lee et al., 2004). The total curvature is formed by combining the plane and profile curvatures (Romer and Ferentinou, 2016). If the surface is convex, the curvature is considered as positive; if it is concave, then it is considered as negative. The value of zero reveals a linear surface (Al-Najjar et al., 2019).

Lithology and LULC were also used as conditioning factors for the preparation of landslide susceptibility mapping. Lithology is important for landslide susceptibility assessment studies because it affects the nature and system of landslides as rocks vary in form of mineral structure besides internal formation (Kornejady et al., 2017). The lithology types in the study area are mostly granite. The study area also contains schist, phyllite and slate types of lithology (Pradhan and Lee, 2010). Human activities are also considered influential to landslides because they affect patterns of land use and land cover. The LULC map of the study area obtained from the Department of Survey and Mapping, Malaysia which shows that the area contains forest, agricultural areas, urban areas, water bodies, transportation, barren lands and others (industrial, infrastructure and utilities, institutions and community facilities). Also, the distance to the road and river were included in our analysis.

Moreover, four hydrological factors were used in this study. These factors are topographic wetness index (TWI), sediment transport index (STI), stream power index (SPI) and terrain roughness index (TRI). SPI represents the movement of solid particles when gravity plays its role on deposits (Rotigliano et al., 2012). STI represents slope failure and deposition. TRI describes the coarseness of the local terrain which affects the topographic and hydrological processes in the development of landslide occurrence. TWI reflects the direction and slope of the flow, which is considered as a measurement for mastering the hydrological processes. These factors were calculated using the following formulas (Yilmaz, 2009). Finally, vegetation density was also used as a landslide conditioning factor. The vegetation density was calculated using the normalized difference vegetation index variable (Pradhan, 2013) extracted from Landsat 8 images. A vegetation density map was classified under four types, i.e. high-density vegetation, medium density, poor density and non-vegetation.

$$\mathrm{SPI} = \mathrm{A_s} \times \tan\beta \qquad (1)$$

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

$$\mathrm{TRI} = \sqrt{Abs\,(\max^2 - \min^2)} \qquad (3)$$

$$\mathrm{TWI} = \ln\left(\frac{\mathrm{A_s}}{\tan\beta}\right) \qquad (4)$$

where, As is defined as a specific area of the catchment (m^2/m); (β) in radian, is a slope gradient (in °); min and max values represent the highest and lowest number of rectangular cells within nine DTM windows, respectively. The definition of the specific catchment is the area of the slope in the upper slide per unit of the length of a contour, which is the area of cells divided by the size of the cell (Kalantar et al., 2018).

4. Methodology

4.1. Overview

The proposed method creates synthetic inventory data using GANs for improving the prediction of landslides. Fig. 3 illustrates the overall workflow of the current study. First, a landslide inventory of 156 landslide locations and 15 conditioning factors were set as inputs for the models. The inventory dataset was split into 70% of training and 30% of testing samples. Then, five machine learning models (e.g. DT, RF, SVM, RF and Bagging ensemble) utilized to evaluate the landslide susceptibility without additional samples. Thereafter, the GAN method was used to create additional training samples with the existing inventory dataset; these new samples were combined with the original training dataset and used to train the same machine learning models again. Once the models were trained, they were tested with the same test dataset used in the first case (without additional samples). Finally, the landslide susceptibility maps were produced by the proposed models. Each map was classified into five susceptibility categorical classes. These models were then validated and assessed using the area under the receiver operating characteristic curve (AUROC).

4.2. Description of machine learning techniques

The following subsections describe the machine learning models used in this study.

4.2.1. ANN model

ANNs exhibit advantages over traditional computational methods (e.g. rule-based) because the model does not require a straightforward practice to estimate desired yields (Jain et al., 1996).After deciding on the number of hidden layers and the number of processing units in an individual layer, the ANN starts learning from the training samples (Aditian et al., 2018).

4.2.2. SVM model

The goal of SVM models is to find the widest margin between two classes in feature space, by a hyperplane (Vapnik, 1995). In landslide susceptibility, the aim is to discriminate between susceptible (1) and not susceptible (−1) pixels. Its main advantages include mapping the data to a high dimensional space where it is easier to classify with linear decision surfaces, also reformulating problems so that data is mapped implicitly into this space.

4.2.3. Decision tree (DT) model

The DT model is a supervised and nonparametric machine learning technique that is operable without prior knowledge about data distribution, with easy interpretation and capability to model as well as it handles the reduction of data complexity and the relationships between variables. Compared to other models, it is a flexible, fast, and robust algorithm that can be used to control the nonlinearity between the input features and discrete classes so that nonlinear relationships between parameters do not affect tree performance. Moreover, DT models are simple to construct and clarify for decision-makers (Kadavi et al., 2019; Saito et al., 2009; Yeon et al., 2010).

4.2.4. Random forest (RF) model

RF is a group of DTs that form an ensemble learning model used for classification and regression problems (Liaw and Wiener, 2002). These models are effective for prediction because they utilize the strength of each tree and their correlations and less sensitive to over-fitting problems. The difference between RF and DT is that a decision tree is built on a whole dataset, utilizing all the variables of interest, while a random forest randomly adopts observations and specific variables to construct multiple decision trees from, and then averages the results. In the present study, samples for landslide and non-landslide events were selected to construct the classification tree (30% of the samples were kept aside from the training and 500 nodes were set as a favorite value).

4.2.5. Bagging ensemble model

In machine learning, several classifiers sometimes are combined and trained to boost the prediction competence of a model (Polikar, 2012). Several combination methods, such as Bagging, AdaBoost, multi boost and stacking can be used such as averaging or majority voting (Breiman, 1996; Freund and Schapire, 1995; Kadavi and Lee, 2018; Webb, 2000). In landslide susceptibility, Bagging has shown superiority over the other methods. Bagging, which is also known as bootstrap aggregating, is a method of sub-dataset generation and combining learners. In this study, the bootstrap samples were employed to build base learners utilizing similar classification approaches, such as SVM and ANN. These based learners were then united by the dominant voting technique.

4.3. Additional data creation with GANs

The GAN which was introduced by Goodfellow et al., (2014) is a type of neural network that trained in an adversarial pattern to produce novel data mimicking specific divisions or distributions. Since their invention, numerous upgraded versions of GANs (concerning firmness of training and perceptual quality) have been developed, including Wasserstein, conditional, Laplacian pyramid and deep convolutional GANs. GANs have been applied for the generation of images, image in-painting, semi-supervised learning and image super-resolution in various domains.

The general design of a GAN consists of two functions (Goodfellow et al., 2014), i.e. a generator (G) and a discriminator (D) which its functionality is demonstrated in (Fig. 4). In consideration of a random uniform distribution, the G maps a sample from the data distribution. Meanwhile, the D is trained to discriminate whether the generated sample has a place in the genuine distribution of the data. The G and D are generally learned together following game theory, although they can be learned through other approaches and techniques.

For each duty, a sample from arbitrary noise z is created by the G to mislead D. Then, the real samples are presented by the D, as well as the samples created by the G, to categorize the samples as fake or real. By producing samples that can fool the D, the G is rewarded. By generating correct classification, D is also rewarded. Both tasks are continuously revised until a Nash equilibrium is obtained. Then, the repetition is paused. More particularly, let D (s) be the likelihood that s originates from genuine information (real data) rather than the generator. G and D play a minimax game with the following value function (Goodfellow et al., 2014).

$$\min_{G}\max_{D} V(D, G) = E_{s\sim p_{data}(s)}[\log D(s)] + E_{z\sim p_{z}(z)}[\log(1 - D(G(z)))] \quad (5)$$

4.4. Validation of susceptibility maps

For a given set of models, the validation was tested by calculating the area under the receiver operating characteristic curve (AUROC). The inventory dataset was split into 70% of training and 30% of testing samples. The ROC was created by plotting the sensitivity of the model versus 1-specificity. The values of AUROC ranged from 0.5 to 1.0, where a high value indicates the superiority of a model.

5. Results

5.1. Application of RFs in selecting factors for modeling

This study applied RFs to remove irrelevant factors from the analysis. The model was used with 180 base estimators and the entire inventory dataset. After the model was trained, the importance values of the 15 factors along with the standard deviation values were computed. Table 1 shows the results of this analysis. The results indicate that the slope factor has the greatest importance value (0.178), followed by LULC (0.171) and aspect (0.125). Most of the landslides have occurred in moderate to high steep areas (slope > 18°). This characteristic allowed the model to distinguish slides from non-slide pixels easily. Similarly, past landslides have occurred in certain land use areas, such as forest, agriculture and barren lands. Schist bedrock is more frequently exposed to slopes facing north through the southwest. The remaining factors, except SPI and STI, also have significant contributions to landslide occurrence. Thus, only SPI and STI were removed from the analysis in this study.

5.2. Evaluation of five applied models (without additional data)

The five models were evaluated by the most commonly used statistical measure, AUROC, where 70% and 30% of the inventory samples were used as training and test data, respectively. In all five models, the best values of the hyper-parameters as computed by the grid search over a specific search space were used, which is shown in (Table 2). Table 3 shows the results obtained for the studied models. The highest AUROC values for the training and test datasets were achieved by the RF (0.94) and SVM (0.85) models, respectively. Using either the training or test dataset, the ANN model has the lowest AUROC value compared with the other models. The Bagging ensemble model was disadvantageous in the current study when SVM was used as a base learner. The training and test accuracy of the SVM model was decreased by 0.04 and 0.1, respectively after the Bagging ensemble model was used. Therefore, the SVM model was a good choice for the study area. However, SVM still faces challenges. For example, it slows down with additional factors, its predictive capability can be degraded with a smaller training sample size and it requires careful optimization of the penalty parameter and the kernel function.

5.3. Evaluation of applied models (with additional data)

Additional training samples were generated by the proposed GAN model. These new samples were combined with the original training dataset and used to train the same models again. Once the models were trained, they were tested with the same test dataset used in the previous section (Section 5.2). Thus, a fair comparison was conducted to evaluate the proposed GAN model. Table 4 shows the AUROC values obtained for the five models using the training (with additional samples) and test datasets. The highest training accuracy was achieved by the RF model (0.94). The RF and SVM models achieved the same accuracy (0.82) using the test dataset. ANN has the lowest training accuracy of 0.75. However, ANN is as accurate as of the DT model on the test dataset.

The additional samples created by the GAN model contributed to increasing the training accuracy of the five models, except that the RF model that achieved the same accuracy in both cases. The ANN model gained the greatest benefit from the additional samples as its training accuracy increased by 0.06. Using the test dataset, the additional samples improved the predictive capability of the models, except that of the SVM model whose test accuracy was decreased by 0.03.

By employing the proposed models, five landslide susceptibility maps were generated from the study area using natural break methods (Fig. 5). Each map was classified into five categorical classes, i.e. very low, low, moderate, high and very high. The blue indicates a low susceptible area, whereas red indicates a highly susceptible area.

5.4. Influence of additional samples created by GANs on model performance

Various numbers of additional samples (5, 10, 20, 30, 40, 50, 100 and 500) were tested to analyze the influence of the number of generated samples on the performance of the models' prediction (Fig. 6). The analysis showed that the DT model performed the best using 10 additional samples on the training dataset, but performed worse using more than 50 additional samples. On the test dataset, the DT model performed the best with 40 additional samples. Meanwhile, the SVM model suffered from over-fitting on the training dataset using additional samples. With 500 additional samples, the SVM model achieved 0.97 AUROC on the training dataset, but it achieved only 0.72 AUROC on the test dataset. Similar results were observed for the Bagging ensemble model. With 500 additional samples, the model achieved 0.94 AUROC on the training dataset and 0.65 AUROC on the test dataset, thereby indicating over-fitting. Similarly, the ANN model also suffered from over-fitting on the training dataset. It achieves 0.75 AUROC with 5 additional samples and 0.91 with 500 additional samples. Among the models, the RF model was less sensitive to the number of additional samples. The best accuracy remained with the 50 additional samples on both datasets. The generation of samples with GANs does not always guarantee to improve model accuracy. Various tests should be evaluated before deciding on the final susceptibility models.

6. Discussion

Machine learning has been an effective landslide susceptibility mapping method. However, with insufficient data, these machine learning models often suffer from generalizing to areas other than the training area. Especially in landslide susceptibility mapping, gathering inventory data is expensive, and some areas have not experienced a large number of landslides. Nevertheless, many studies have attempted to develop models that work with insufficient data. For example, sampling strategy and validation methods have been validated to address the challenges of modeling with limited data effectively. Given that randomness of the training, data selection influences the model performance (Kalantar et al., 2018), sampling strategies that avoid model over-fitting to the training data have been proposed (Aktas and San, 2019; Conoscenti et al., 2016). More often than not, landslide inventory data are incomplete. Such incomplete data affect the selection of the absence samples. For this problem, Steger et al. (2016) suggested that models can correlate with landslide inventory incompleteness, and thus, they should be rejected regardless of their performance. Techniques such as factor optimization, development of new factors and model ensembling have also been extensively discussed in the recent literature.

Removing insignificant factors was useful to decrease the impact of model over-fitting due to the limited training. The RF model showed that SPI and STI were not influential and thus were removed from the analysis. Estimation of the factors also plays an important role in obtaining insights into the factors included in the model. Similar to previous studies, the present study found the slope to be a significant factor. The landslide inventory dataset showed that most of the landslides have occurred in moderate to high steep areas. A significant number of past landslides have occurred in certain land use areas, such as forest, agriculture and barren lands. The results of the RF model were also consistent with the inventory data, where LULC and aspect were found to be significant.

The evaluation of the models with and without additional samples showed that the proposed GAN can improve the performance of the susceptibility model. When the training data were used, the GAN model improved the accuracy of all the models except RF. Some models, such as ANN, performed better than others. Using the test data contributed to increasing the accuracy of all the models except SVM. Moreover, the number of additional samples significantly affected the modeling performance. The DT, SVM and ANN models over-fitted the training data when a large number of additional samples were included in the training set. The RF model was less sensitive to the number of additional samples than other models. Thus, adding newly generated samples to the training set may not always lead to an increase in model accuracy, especially on the test data. Therefore, the number of additional samples should be considered as a parameter and fine-tuned before training any machine learning model.

7. Conclusions

This study addressed the aforementioned problem with a GAN-based method. This model was used to create an additional training sample with the existing inventory dataset. The proposed method was evaluated on a dataset taken from Cameron Highlands, Malaysia. Five machine learning and statistical models were implemented to assess the proposed GAN model. The outcomes revealed that using additional samples created by the proposed GAN model can improve the predictive capability of the studied models, except SVM.

Generative models, such as GANs, can be useful for landslide susceptibility mapping, especially when the training data for the area under study are inadequate. However, the used models should be carefully analyzed to avoid over-fitting to the training samples. In addition, the hyper-parameters of the used models can be optimized to improve the overall performance of the landslide susceptibility models when samples created by generative models are used. Improvements in landslide susceptibility maps can help in the implementation of land use planning and the design of landslide mitigation strategies. Improvements in landslide susceptibility models also contribute towards improving landslide hazard and risk assessment. The proposed method, therefore, can be a useful tool for engineers, geoscientists and planners.

Funding: This research is funded by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, the University of Technology Sydney, Australia.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgments: The authors would like to thank the Faculty of Engineering and Information Technology, the University of Technology Sydney for providing all facilities during this research. We are also thankful to the Department of Mineral and Geosciences, the Department of Surveying Malaysia, and the Federal Department of Town and Country Planning.

References

Aditian, A., Kubota, T., 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. https://doi.org/10.1016/j.geomorph.2018.06.006

Akbar, A.Q., Chen, G., 2018. Comparison of major statistical methods and their combination using matrix validation for landslide susceptibility mapping. Lowland Technology International 20, 401–412.

Aktas, H., San, B.T., 2019. Landslide susceptibility mapping using an automatic sampling algorithm based on two level random sampling. Computers and Geosciences 133, 104329. https://doi.org/10.1016/j.cageo.2019.104329

Al-Najjar, H.A.H., Kalantar, B., Pradhan, B., Saeidi, V., 2019. Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms, in: In Earth Resources and Environmental Remote Sensing/GIS Applications. International Society for Optics and Photonics. p. 19. https://doi.org/10.1117/12.2532687

Ashournejad, Q., Hosseini, A., Pradhan, B., Hosseini, S.J., 2019. Hazard zoning for spatial planning using GIS-based landslide susceptibility assessment: a new hybrid integrated data-driven and knowledge-based model. Arabian Journal of Geosciences 12. https://doi.org/10.1007/s12517-019-4236-0

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65, 15–31. https://doi.org/10.1016/j.geomorph.2004.06.010

Bragagnolo, L., Silva, R.V. d., Grzybowski, J.M.V., 2020. Artificial neural network ensembles applied to the mapping of landslide susceptibility. Catena 184, 104240. https://doi.org/10.1016/j.catena.2019.104240

Braun, A., Urquia, Elias Leonardo Garcia Lopez, Rigoberto Moncada Yamagishi, H., 2018. Landslide Susceptibility Mapping in Tegucigalpa, Honduras, Using Data Mining Methods, in: Slope Stability: Case Histories, Landslide Mapping, Emerging Technologies. pp. 207–215.

Breiman, L., 1996. Bagging Predictors. Machine Learning 24, 123–140. https://doi.org/10.1007/BF00058655

Can, A., Dagdelenler, G., Ercanoglu, M., Sonmez, H., 2019. Landslide susceptibility mapping at Ovacık-Karabük (Turkey) using different artificial neural network models: comparison of training algorithms. Bulletin of Engineering Geology and the Environment 78, 89–102. https://doi.org/10.1007/s10064-017-1034-3

Canoglu, M.C., Aksoy, H., Ercanoglu, M., 2019. Integrated approach for determining spatio-temporal variations in the hydrodynamic factors as a contributing parameter in landslide susceptibility assessments. Bulletin of Engineering Geology and the Environment 78, 3159–3174. https://doi.org/10.1007/s10064-018-1337-z

Ciurleo, M., Cascini, L., Calvello, M., 2017. A comparison of statistical and deterministic methods for shallow landslide susceptibility zoning in clayey soils. Engineering Geology 223, 71–81. https://doi.org/10.1016/j.enggeo.2017.04.023

Conoscenti, C., Rotigliano, E., Cama, M., Caraballo-Arias, N.A., Lombardo, L., Agnesi, V., 2016. Exploring the effect of absence selection on landslide susceptibility models: A case study in Sicily, Italy. Geomorphology 261, 222–235. https://doi.org/10.1016/j.geomorph.2016.03.006

Dou, J., Bui, D.T., Yunus, A.P., Jia, K., Song, X., Revhaug, I., Xia, H., 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. https://doi.org/10.1371/journal.pone.0133262

Du, J., Glade, T., Woldai, T., Chai, B., Zeng, B., 2020. Landslide susceptibility assessment based on an incomplete landslide inventory in the Jilong Valley, Tibet, Chinese Himalayas. Engineering Geology 270, 105572. https://doi.org/10.1016/j.enggeo.2020.105572

Fanos, A.M., Pradhan, B., 2019. A novel hybrid machine learning-based model for rockfall source identification in presence of other landslide types using LiDAR and GIS. Earth Systems and Environment 3, 491–506. https://doi.org/10.1007/s41748-019-00114-z

Feizizadeh, B., Roodposhti, M.S., Blaschke, T., Aryal, J., 2017. Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. Arabian Journal of Geosciences 10. https://doi.org/10.1007/s12517-017-2918-z

Formetta, G., Rago, V., Capparelli, G., Rigon, R., Muto, F., Versace, P., 2014. Integrated Physically based System for Modeling Landslide Susceptibility. Procedia Earth and Planetary Science 9, 74–82. https://doi.org/10.1016/j.proeps.2014.06.006

Freund, Y., Schapire, R.E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting., in: In European Conference on Computational Learning Theory. pp. 23–37. https://doi.org/10.1007/3-540-59119-2

Glade, T., Anderson, M., Crozier, M.J., 2012. Landslide hazard and risk. https://doi.org/10.1002/9780470012659

Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers and Geosciences 81, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007

Goetz, J.N., Guthrie, R.H., Brenning, A., 2011. Integrating physical and empirical landslide susceptibility models using generalized additive models. Geomorphology 129, 376–386. https://doi.org/10.1016/j.geomorph.2011.03.001

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, in: In Advances in Neural Information Processing Systems (Pp. 2672-2680). pp. 2672–2680.

Hong, H., Miao, Y., Liu, J., Zhu, A.X., 2019. Exploring the effects of the design and quantity of absence data on the performance of random forest-based landslide susceptibility mapping. Catena 176, 45–64. https://doi.org/10.1016/j.catena.2018.12.035

Huang, Y., Zhao, L., 2018. Review on landslide susceptibility mapping using support vector machines. Catena 165, 520–529. https://doi.org/10.1016/j.catena.2018.03.003

Hussin, H.Y., Zumpano, V., Reichenbach, P., Sterlacchini, S., Micu, M., van Westen, C., Bălteanu, D., 2016. Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. Geomorphology 253, 508–523. https://doi.org/10.1016/j.geomorph.2015.10.030

Jain, A.K., Mao, J. and, Mohiuddin, K.M., 1996. Artificial neural networks: A tutorial. Computer 29, 31–44.

Kadavi, P.R., Lee, C., 2018. Application of Ensemble-Based Machine Learning Models to Landslide Susceptibility Mapping. Remote Sensing 10, 1–18. https://doi.org/10.3390/rs10081252

Kadavi, P.R., Lee, C.W., Lee, S., 2019. Landslide-susceptibility mapping in Gangwon-do, South Korea, using logistic regression and decision tree models. Environmental Earth Sciences 78, 1–17. https://doi.org/10.1007/s12665-019-8119-1

Kalantar, B., Pradhan, B., Amir Naghibi, S., Motevalli, A., Mansor, S., 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics, Natural Hazards and Risk 9, 49–69. https://doi.org/10.1080/19475705.2017.1407368

Kamp, U., Growley, B.J., Khattak, G.A., Owen, L.A., 2008. GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region. Geomorphology 101, 631–642. https://doi.org/10.1016/j.geomorph.2008.03.003

Kavzoglu, T., Colkesen, I., Sahin, E.K., 2019. Landslides: Theory, Practice and Modelling 50, 283–301. https://doi.org/10.1007/978-3-319-77377-3

Kavzoglu, T., Kutlug Sahin, E., Colkesen, I., 2015. Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Engineering Geology 192, 101–112. https://doi.org/10.1016/j.enggeo.2015.04.004

Kawabata, D., Bandibas, J., 2009. Landslide susceptibility mapping using geological data, a DEM from ASTER images and an Artificial Neural Network (ANN). Geomorphology 113, 97–109. https://doi.org/10.1016/j.geomorph.2009.06.006

Kornejady, A., Ownegh, M., Bahremand, A., 2017. Landslide susceptibility assessment using maximum entropy model with two different data sampling methods. Catena 152, 144–162. https://doi.org/10.1016/j.catena.2017.01.010

Kornejady, A., Ownegh, M., Rahmati, O., Bahremand, A., 2018. Landslide susceptibility assessment using three bivariate models considering the new topo-hydrological factor: HAND. Geocarto International 33, 1155–1185. https://doi.org/10.1080/10106049.2017.1334832

Lai, J.S., Chiang, S.H., Tsai, F., 2019. Exploring influence of sampling strategies on event-based landslide susceptibility modeling. ISPRS International Journal of Geo-Information 8. https://doi.org/10.3390/ijgi8090397

Lee, J.H., Sameen, M.I., Pradhan, B., Park, H.J., 2018. Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology 303, 284–298. https://doi.org/10.1016/j.geomorph.2017.12.007

Lee, M.L., Ng, K.Y., Huang, Y.F., Li, W.C., 2014. Rainfall-induced landslides in Hulu Kelang area, Malaysia. Natural Hazards 70, 353–375. https://doi.org/10.1007/s11069-013-0814-8

Lee, S., Ryu, J.H., Won, J.S., Park, H.J., 2004. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. Engineering Geology 71, 289–302. https://doi.org/10.1016/S0013-7952(03)00142-X

Lee, S., Sambath, T., 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environmental Geology 50, 847–855. https://doi.org/10.1007/s00254-006-0256-7

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Mandal, S., Mondal, S., 2019. Statistical approaches for landslide susceptibility assessment and prediction, in: Statistical Approaches for Landslide Susceptibility Assessment and Prediction. pp. 181–189. https://doi.org/10.1007/978-3-319-93897-4

Matori, A.N., Basith, A., 2012. Evaluation of landslide causative factors towards efficient landslide susceptibility modelling in the Cameron Highlands, Malaysia. WIT Transactions on Engineering Sciences 73, 207–218. https://doi.org/10.2495/DEB120181

Mezaal, M.R., Pradhan, B., 2018. An improved algorithm for identifying shallow and deep-seated landslides in dense tropical forest from airborne laser scanning data. Catena 167, 147–159. https://doi.org/10.1016/j.catena.2018.04.038

Ozdemir, A., Altural, T., 2013. A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan mountains, SW Turkey. Journal of Asian Earth Sciences 64, 180–197. https://doi.org/10.1016/j.jseaes.2012.12.014

Park, J.Y., Lee, S.R., Lee, D.H., Kim, Y.T., Lee, J.S., 2019. A regional-scale landslide early warning methodology applying statistical and physically based approaches in sequence. Engineering Geology 260, 105193. https://doi.org/10.1016/j.enggeo.2019.105193

Polikar, R., 2012. Ensemble Learning, In Ensemble machine learning. https://doi.org/10.1007/978-1-4419-9326-7

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Computers and Geosciences 51, 350–365. https://doi.org/10.1016/j.cageo.2012.08.023

Pradhan, B., Lee, S., 2010. Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. Landslides 7, 13–30. https://doi.org/10.1007/s10346-009-0183-2

Romer, C., Ferentinou, M., 2016. Shallow landslide susceptibility assessment in a semiarid environment - A Quaternary catchment of KwaZulu-Natal, South Africa. Engineering Geology 201, 29–44. https://doi.org/10.1016/j.enggeo.2015.12.013

Rotigliano, E., Cappadonia, C., Conoscenti, C., Costanzo, D., Agnesi, V., 2012. Slope units-based flow susceptibility model: Using validation tests to select controlling factors. Natural Hazards 61, 143–153. https://doi.org/10.1007/s11069-011-9846-0

Saito, H., Nakayama, D., Matsuyama, H., 2009. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. Geomorphology 109, 108–121. https://doi.org/10.1016/j.geomorph.2009.02.026

Sameen, M.I., Pradhan, B., 2019. Landslide detection using residual networks and the fusion of spectral and topographic information. IEEE Access 7, 114363–114373. https://doi.org/10.1109/access.2019.2935761

Sameen, M.I., Pradhan, B., Bui, D.T., Alamri, A.M., 2020. Systematic sample subdividing strategy for training landslide susceptibility models. Catena 187, 104358. https://doi.org/10.1016/j.catena.2019.104358

Samia, J., Temme, A., Bregt, A.K., Wallinga, J., Stuiver, J., Guzzetti, F., Ardizzone, F., Rossi, M., 2018. Implementing landslide path dependency in landslide susceptibility modelling. Landslides 15, 2129–2144. https://doi.org/10.1007/s10346-018-1024-y

Soma, A.S., Kubota, T., 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, 383–401. https://doi.org/10.1007/s11629-018-4884-7

Steger, S., Brenning, A., Bell, R., Glade, T., 2018. Incompleteness matters – An approach to counteract inventory-based biases in statistical landslide susceptibility modelling 20, 8551.

Steger, S., Brenning, A., Bell, R., Glade, T., 2016. The influence of systematically incomplete shallow landslide inventories on statistical susceptibility models and suggestions for improvements. Landslides 14, 1767–1781. https://doi.org/10.1007/s10346-017-0820-0

Süzen, M.L., Doyuran, V., 2004. A comparison of the GIS based landslide susceptibility assessment methods: Multivariate versus bivariate. Environmental Geology 45, 665–679. https://doi.org/10.1007/s00254-003-0917-8

Tsangaratos, P., Ilia, I., 2016. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena 145, 164–179. https://doi.org/10.1016/j.catena.2016.06.004

Vapnik, V.N., 1995. Constructing learning algorithms. In The nature of statistical learning theory.

Wang, H.J., Xiao, T., Li, X.Y., Zhang, L.L., Zhang, L.M., 2019. A novel physically-based model for updating landslide susceptibility. Engineering Geology 251, 71–80. https://doi.org/10.1016/j.enggeo.2019.02.004

Webb, G.I., 2000. MultiBoosting: a technique for combining boosting and wagging. Machine Learning 40, 159–196. https://doi.org/10.1023/A:1007659514849

Xiao, T., Yin, K., Yao, T., Liu, S., 2019. Spatial prediction of landslide susceptibility using GIS-based statistical and machine learning models in Wanzhou County, Three Gorges Reservoir, China. Acta Geochimica 38, 654–669. https://doi.org/10.1007/s11631-019-00341-1

Yan, F., Zhang, Q., Ye, S., Ren, B., 2019. A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model. Geomorphology 327, 170–187. https://doi.org/10.1016/j.geomorph.2018.10.024

Yeon, Y.K., Han, J.G., Ryu, K.H., 2010. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Engineering Geology 116, 274–283. https://doi.org/10.1016/j.enggeo.2010.09.009

Yilmaz, I., 2009. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat-Turkey). Computers and Geosciences 35, 1125–1138. https://doi.org/10.1016/j.cageo.2008.08.007

Yilmaz, I., Ercanoglu, M., 2019. Natural Hazards GIS-Based Spatial Modeling Using Data Mining Techniques, in: Landslide Inventory, Sampling and Effect of Sampling Strategies on Landslide Susceptibility/Hazard Modelling at a Glance. pp. 205–224.

Zêzere, J.L., Pereira, S., Melo, R., Oliveira, S.C., Garcia, R.A.C., 2017. Mapping landslide susceptibility using data-driven methods. Science of the Total Environment 589, 250–267. https://doi.org/10.1016/j.scitotenv.2017.02.188

Zhang, T. yu, Han, L., Zhang, H., Zhao, Y. hua, Li, X. an, Zhao, L., 2019. GIS-based landslide susceptibility mapping using hybrid integration approaches of fractal dimension with index of entropy and support vector machine. Journal of Mountain Science 16, 1275–1288. https://doi.org/10.1007/s11629-018-5337-z

Zhu, A.X., Miao, Y., Liu, J., Bai, S., Zeng, C., Ma, T., Hong, H., 2019. A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. Catena 183, 104188. https://doi.org/10.1016/j.catena.2019.104188

Figure caption

Fig. 1. Location of the study area and landslide inventory map

Fig. 2a-f. Maps of landslide conditioning factors: (a) Elevation, (b) Slope, (c) Aspect, (d) Plan curvature, (e) Profile curvature, and (f) Total curvature.

Fig. 2g-l. Maps of landslide conditioning factors: (g) Lithology, (h) LULC, (i) Distance to road, (j) Distance to river, (k) SPI, and (l) STI.

Fig. 2m-o. Maps of landslide conditioning factors: (m) TRI, (n) TWI, and (o) Vegetation density.

Fig. 3. Overall workflow used in this study.

Fig. 4. The general architecture of GANs.

Fig. 5. Landslide susceptibility maps produced by proposed (a) DT, (b) RF, (c) SVM, (d) Bagging ensemble, and (e) ANN models.

Fig. 6. Training and test AUROC values calculated for the five models trained with original training dataset and additional samples created by GANs.

Table caption

Table 1. Importance of affecting factors.

Table 2. Optimised parameters of five models and search spaces.

Table 3. AUROC values of five models using training and test datasets.

Table 4. AUROC values of models using training (with additional samples) and test datasets.

Figure 1

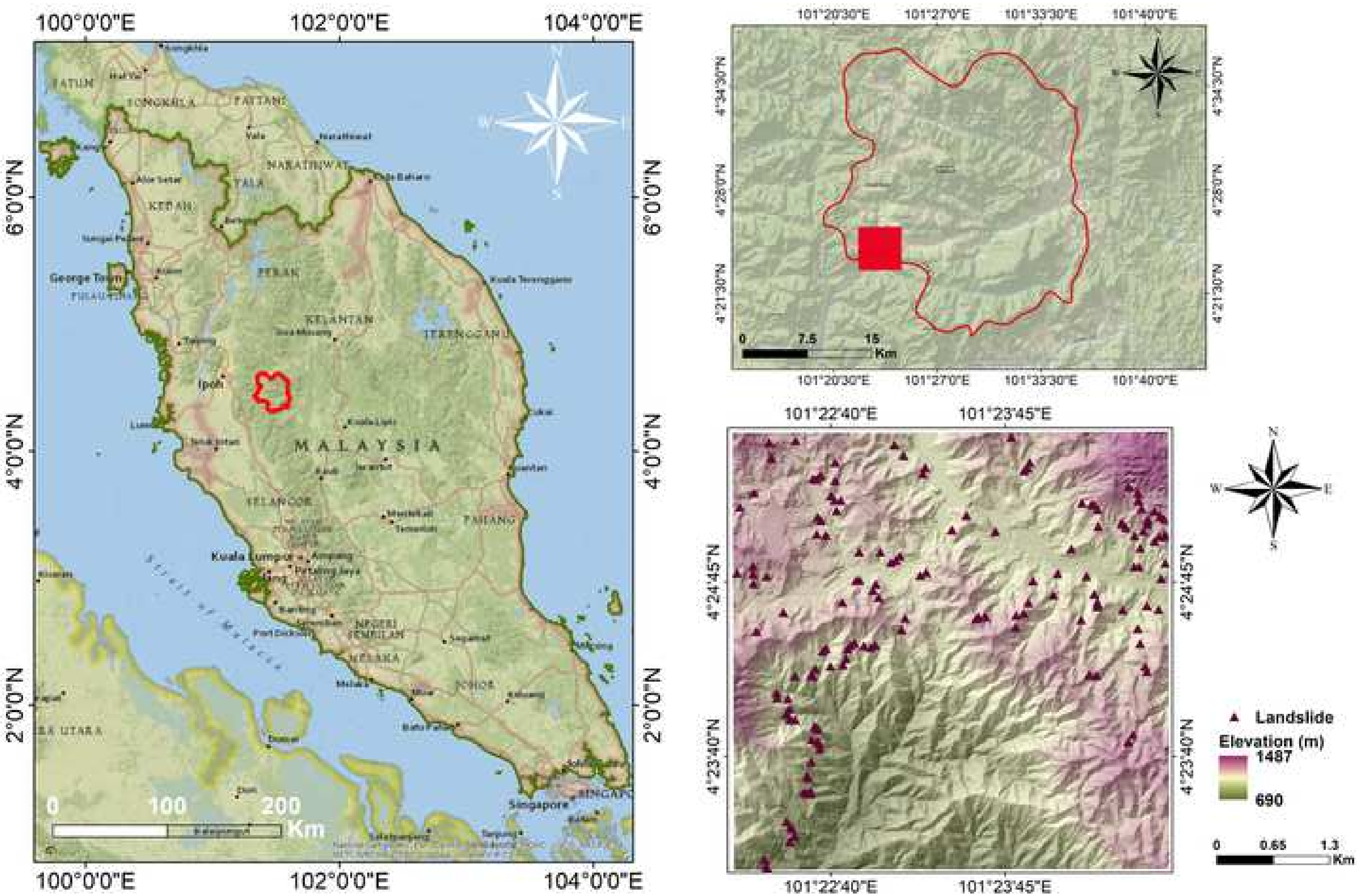

100°0'0"E
102°0'0"E
104°0'0"E
6°0'0"N
4°0'0"N
2°0'0"N
George Town
Ipoh
Kuala Lumpur
M A L A Y S I A
KELANTAN
TERENGGANU
PERAK
PAHANG
Singapore
0 100 200
Km
101°20'30"E
101°27'0"E
101°33'30"E
101°40'0"E
4°34'30"N
4°28'0"N
4°21'30"N
0 7.5 15
Km
101°22'40"E
101°23'45"E
4°24'45"N
4°23'40"N
Landslide
Elevation (m)
1487
690
0 0.65 1.3
Km

Figure 2a-f

Click here to access/download;Figure;Fig. 2 a-f.jpg

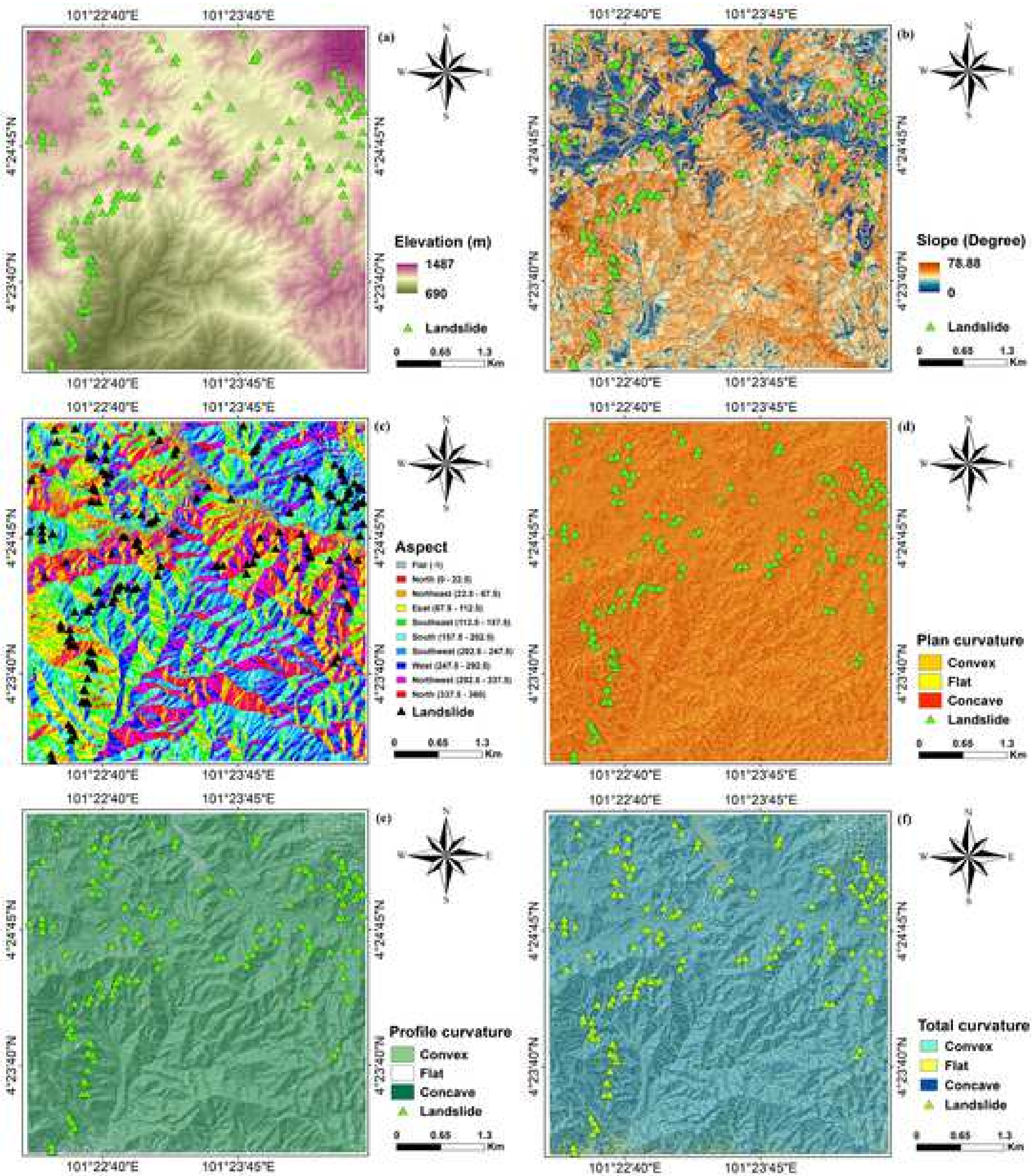


Figure 2g-l

Click here to access/download;Figure;Fig. 2 g-l.jpg

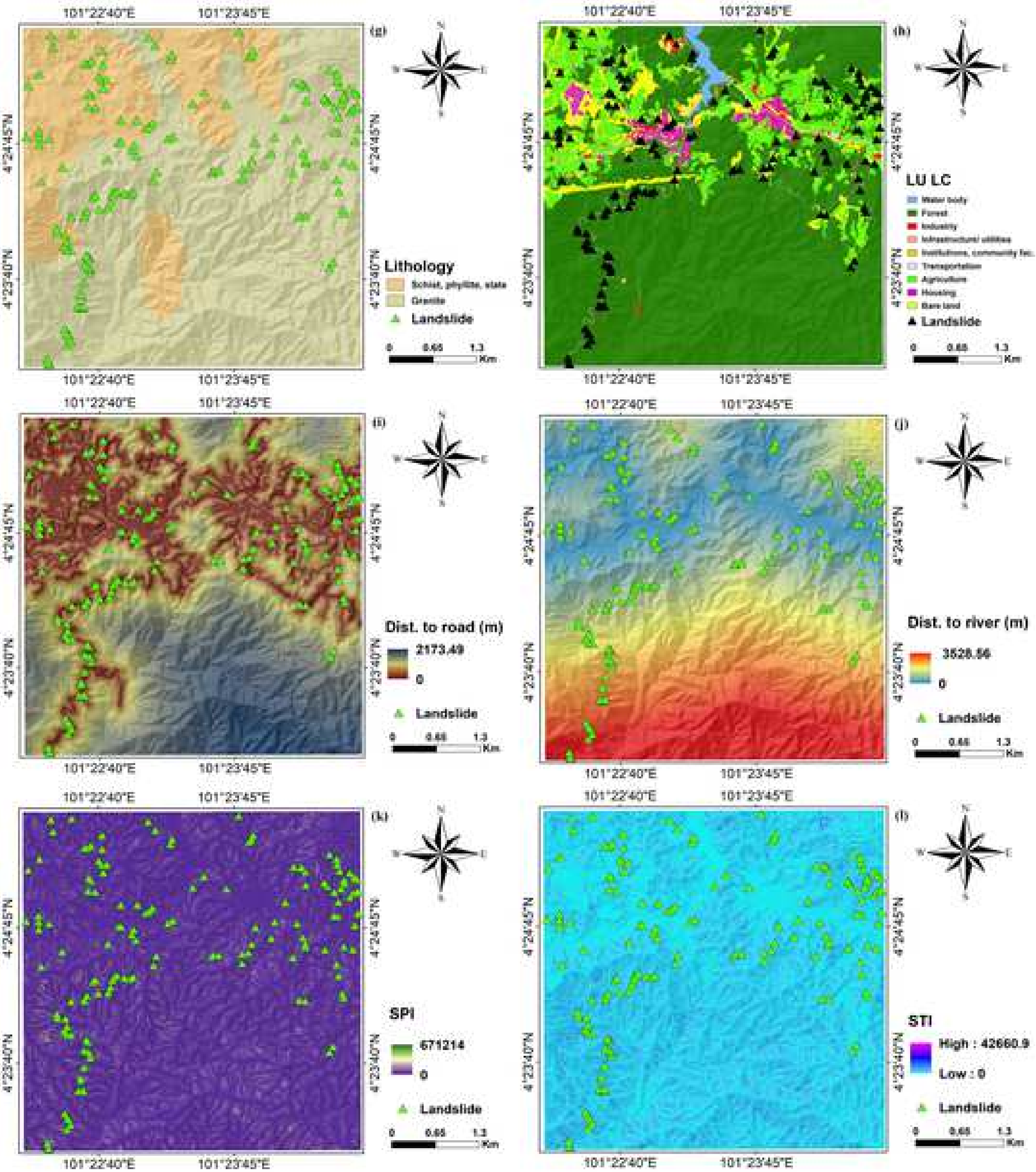


Figure 2m-o

Click here to access/download;Figure;Fig. 2 m-o.jpg

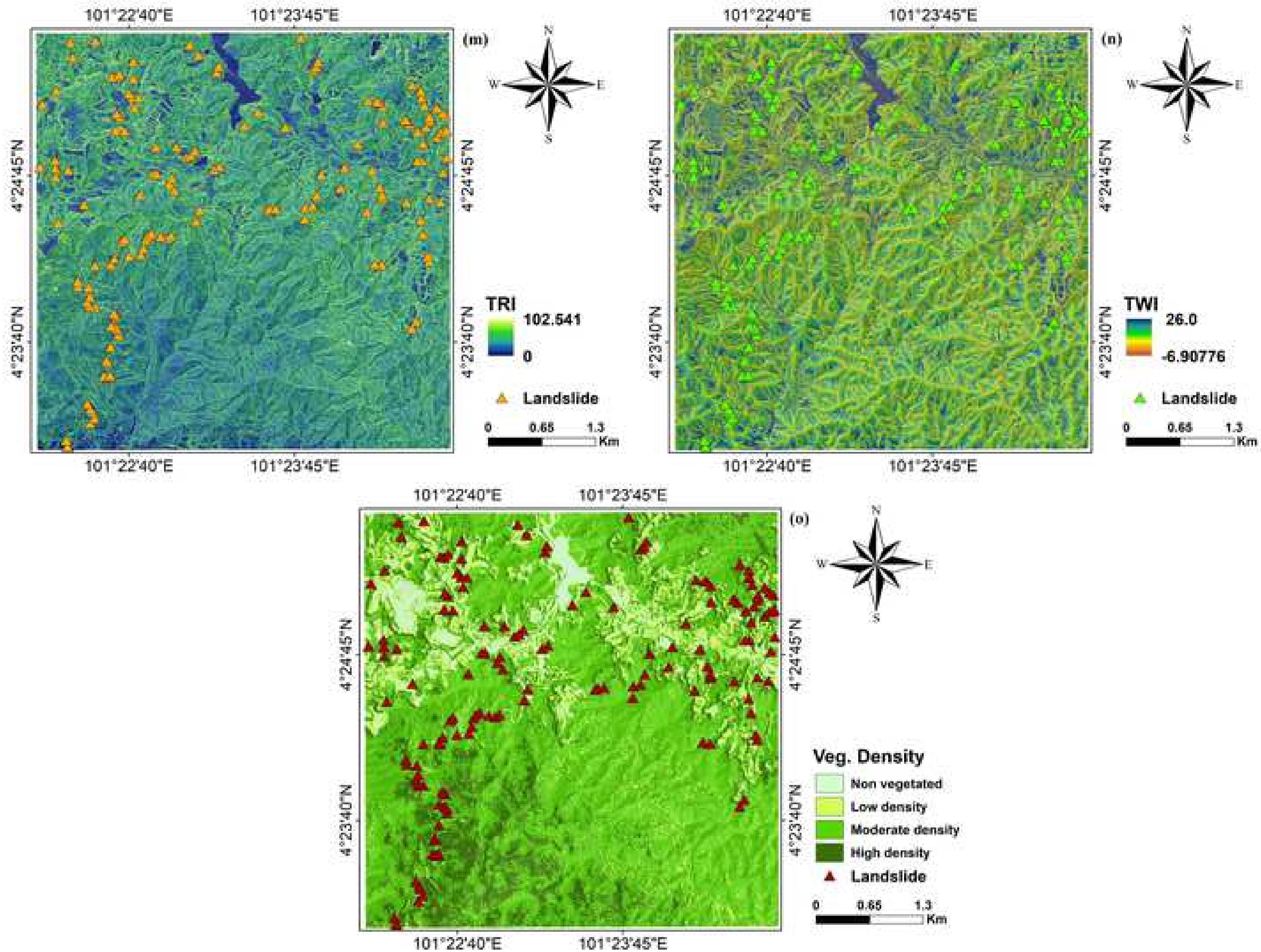


Figure 3

Click here to access/download;Figure;Fig. 3.jpg

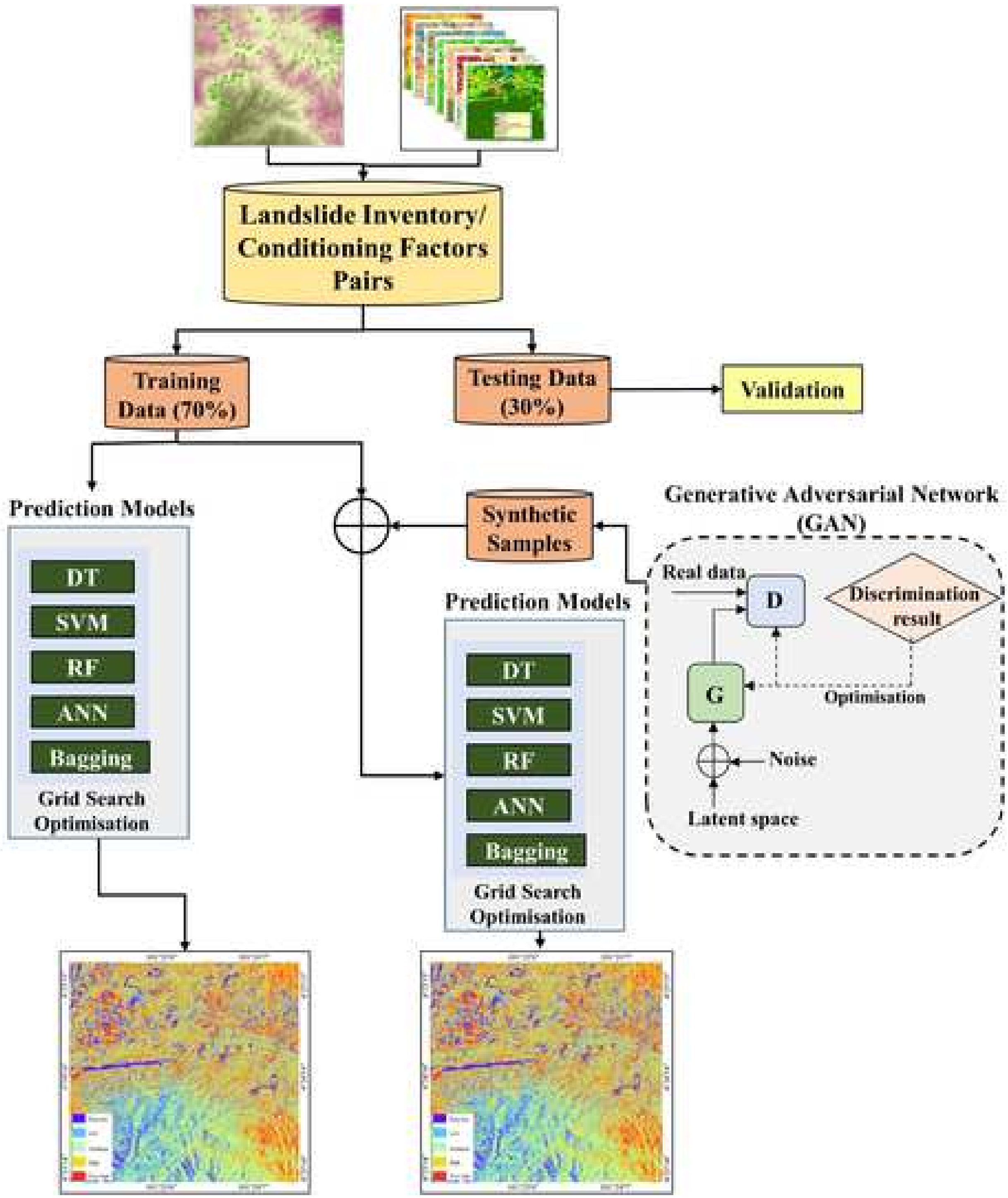
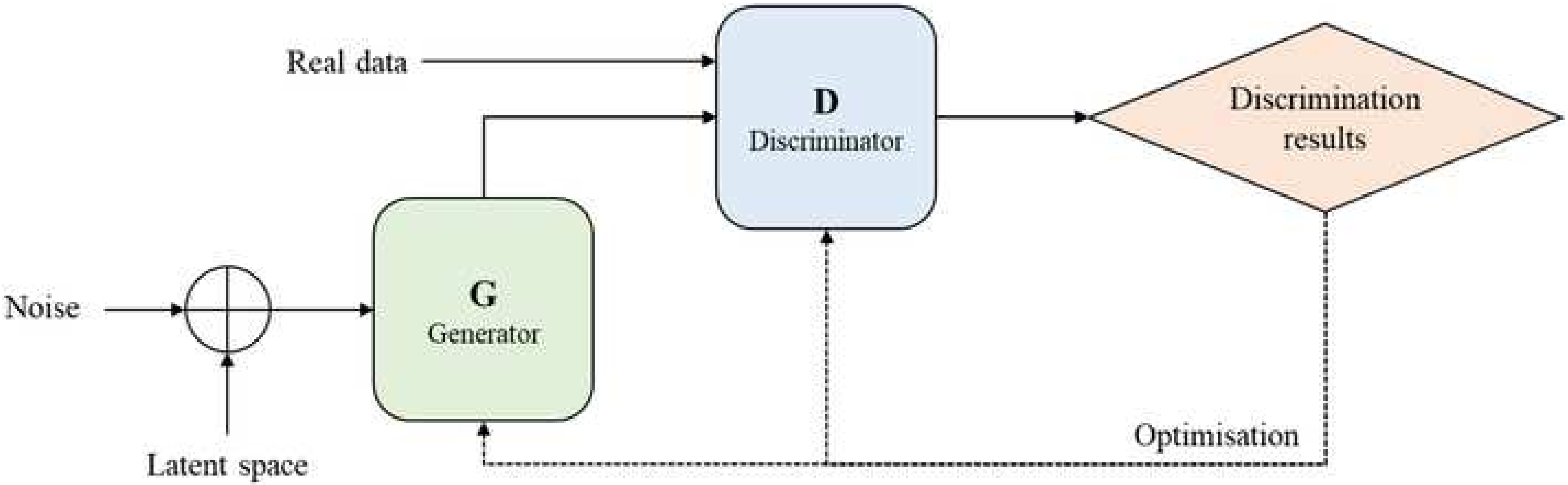


Figure 4



Real data
D
Discriminator
Discrimination results
Noise
G
Generator
Latent space
Optimisation

Figure 5

Click here to access/download;Figure;Fig. 5 .jpg

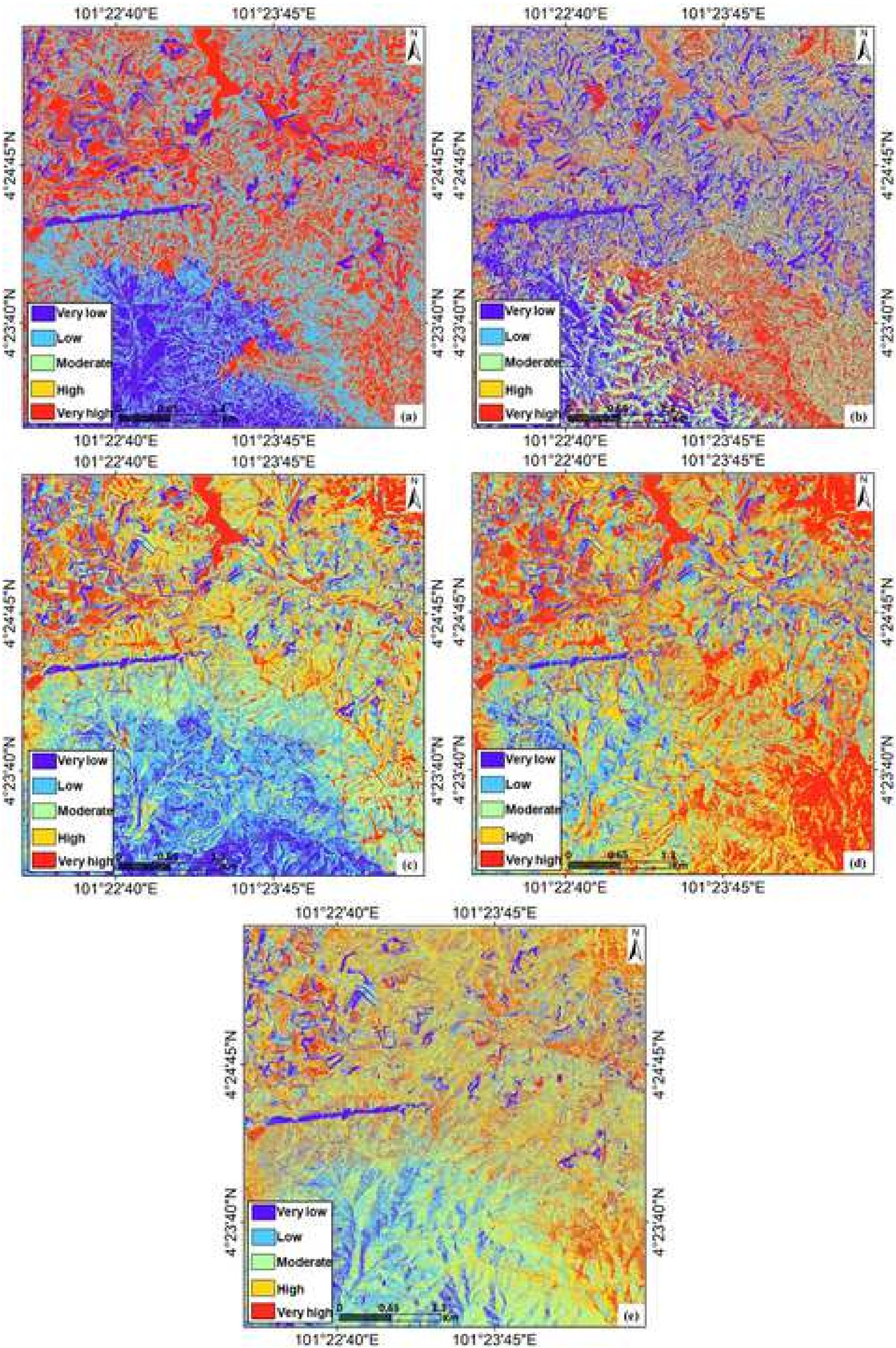


Figure 6 Click here to access/download;Figure;Fig. 6 .tif

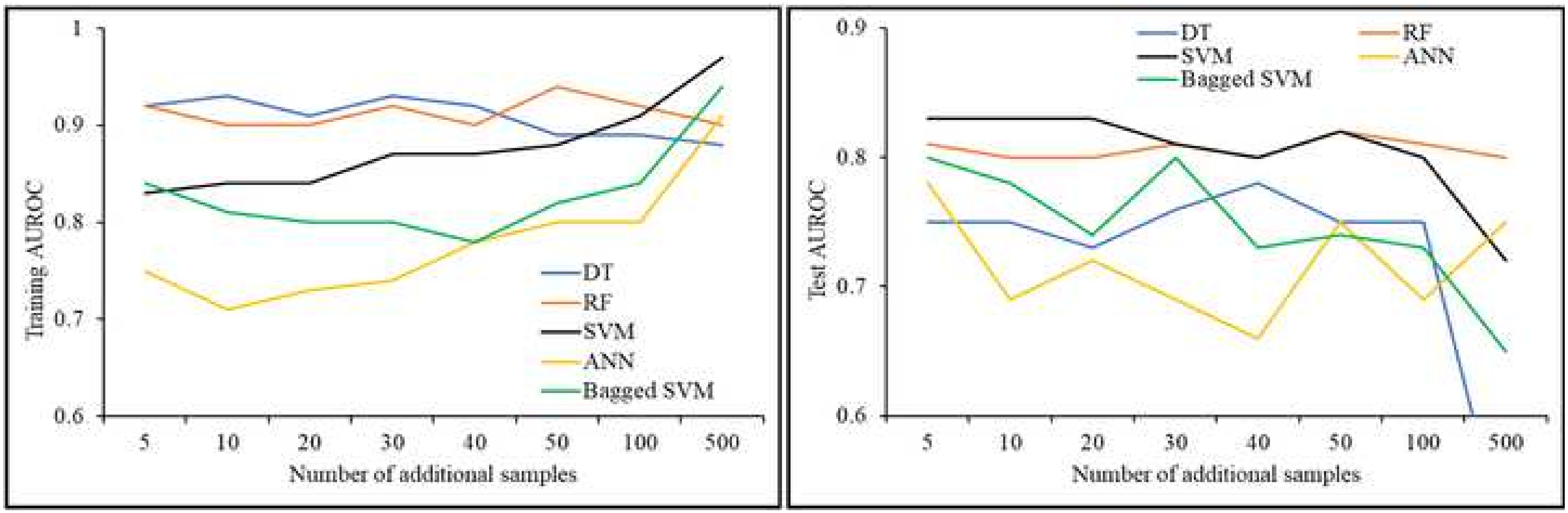


Table 1

Factor	RF importance value	Standard deviation
Slope	0.178	0.022
LULC	0.171	0.012
Aspect	0.125	0.013
Elevation	0.09	0.026
Vegetation density	0.073	0.017
TWI	0.063	0.023
Distance to river	0.061	0
TRI	0.06	0.018
Lithology	0.041	0.019
Total curvature	0.038	0.014
Plane curvature	0.034	0.039
Profile curvature	0.032	0.015
Distance to road	0.029	0.045
SPI	0	0
STI	0	0.015

Table 2

Model	Parameters	Search space	Best value (grid search with 10-fold cross validation)
DT	Maximum tree depth	[2–13]	5
RF	Number of base estimators	[10–1000]	180
SVM	C	[1–1000]	5
	Kernel function	[Linear, RBF, Sigmoid]	RBF
ANN	Learning rate	[10^{-5}–1.0]	0.01
	Activation function	[ReLU, Tanh, Sigmoid, Linear]	ReLU
	Number of hidden layers	[1–12]	1
	Number of hidden units in a hidden layer	[2–1024]	62
Bagging ensemble	Base learner	[DT, RF, SVM, ANN]	SVM

Table 3

Model	Training AUROC	Test AUROC
DT	0.9	0.76
RF	0.94	0.81
SVM	0.86	0.85
ANN	0.69	0.72
Bagging ensemble	0.82	0.75

Table 4

Model	Training AUROC	Test AUROC
DT	0.92	0.78
RF	0.94	0.82
SVM	0.88	0.82
ANN	0.75	0.78
Bagging ensemble	0.84	0.8

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: