Landslide Susceptibility Analysis using Gradient Boosting Models: A Case Study in Penang Island, Malaysia

Gao Han[1], Fam Pei Shan[1*], Tay Lea Tien[2] and Low Heng Chin[3]
1. School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, MALAYSIA
2. School of Electrical and Electronic Engineering, USM, Engineering Campus, Seberang Perai Selatan Nibong Tebal, Penang, 14300, MALAYSIA
3. Research and Innovation Unit, Universiti Sains Malaysia, 11800 USM, Penang, MALAYSIA
*fpeishan@usm.my

Abstract

Tree-based gradient boosting (TGB) models gain popularity in various areas due to their powerful prediction ability and fast processing speed. This study aims to compare the landslide spatial prediction performance of TGB models and non-tree-based machine learning (NML) models in Penang Island, Malaysia. Two specific instances of TGB models, eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) and two specific instances of NML models, artificial neural network (ANN) and support vector machine (SVM), are applied to make predictions of landslide susceptibility. Feature selection and oversampling techniques are considered to improve the prediction performance as well. The results are analyzed and discussed mainly based on receiver operating characteristic (ROC) curves as well as the area under the curves (AUC).

The results show that TGB models give better prediction performance compared to NML models, no matter what the sample size is. The TGB models' performances are improved when training with the dataset considering either feature selection or oversampling techniques. The highest AUC value of 0.9525 is obtained from the combination of XGBoost and SMOTE. The landslide susceptibility maps (LSMs) produced by XGBoost and LightGBM can provide valuable information in landslide management and mitigation in Penang Island, Malaysia.

Keywords: eXtreme Gradient Boosting, LightGBM, landslide susceptibility mapping, feature selection, oversampling techniques.

Introduction

Landslides are considered as one of the most hazardous natural disasters around the globe which may cause the losses of life and property.[9,44] Landslide susceptibility analysis (LSA) is a popular and effective way to determine the possibility of landslide occurrence in a specific area and further reduces the losses. LSA mainly works on studying the relationship between the landslide conditioning factors and the characteristics of the recorded landslides using various types of models such as logistic regression,[1] decision tree,[2] artificial neural network (ANN)[8,18] and support vector machine (SVM).[4,19] Other review papers on the machine learning models of LSA are provided by Guzzetti et al[23], Huabin et al[26] and Gao et al.[17]

In our previous landslide spatial research in Penang Island, various statistical and machine learning models are applied such as logistic regression, fuzzy logic, ANN and SVM.[18,19] The results showed that ANN and SVM outperformed among other models. Gradient boosting models are gaining more and more popularity in various areas due to its powerful prediction ability and fast processing speed.[12] Therefore, comparing the predicting performance between gradient boosting models and machine learning models such as ANN and SVM, is considered in this research.

The models used in this research are classified as tree-based gradient boosting (TGB) models and non-tree-based machine learning (NML) models. The TGB models can be considered as a type of ensemble machine learning models which work to combine several decision trees to produce better predictive performance than a single decision tree classifier.

The main idea of the ensemble models is to combine several weak learners which are slightly better than random guess to a strong learner, thus increasing the performance of the ensemble model.[48] The main principle behind gradient boosting can be interpreted as an optimization algorithm on a suitable cost function.[7] The TGB models such as eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) have shown powerful prediction performance in various fields as well as in machine learning competitions.[12,33,46] Moreover, the tree-based ensemble models are more and more popular and show more satisfactory performance in recent LSA studies.

Hong et al[24] compared three ensemble models, namely, AdaBoost, Bagging and Rotation Forest, in landslide susceptibility assessment and obtained promising results in the Guangchang area, China. Bandara et al[5] conducted the landslide research in two different study areas, Ratnapura district in Sri Lanka and Glenmalure in Ireland, using three tree-based ensemble models, namely, random forest, rotation forest and XGBoost. The results evaluated in terms of precision, recall and F-score were shown to be satisfactory. For the NML models, such as ANN and SVM, researchers have been using them to solve practical

applications for a comparatively long time and showed quite satisfactory results.[6, 13, 37] Compared to NML models, TGB models are more capable of handling larger-scale data and have faster training speed and lower memory usage.[31]

This study aims to compare the prediction performance between TGB and NML models in landslide spatial prediction area. Since each learning algorithm tends to suit some problem types better than others, the TGB and NML models are compared in this research to discover the most suitable model for the landslide data in Penang Island used to produce landslide susceptibility maps. Therefore, one of the objectives of this study is to discover which model among XGBoost, LightGBM, ANN and SVM is superior in predicting the landslide spatial prediction in Penang Island, Malaysia.

Feature selection is a key step in data mining which can help reduce the dimension of the datasets and improve the models' performance.[21] Therefore, two feature selection techniques, namely, extra trees classifier (ETC) and random forest classifier (RFC), are considered to measure the feature importance. The features with high importance score would be selected to train the models. Thus, the second objective of this study is to determine whether the models' prediction performances would be improved when combined with feature selection techniques. Furthermore, oversampling techniques are considered in this research as well. The third objective of this study is to assess the efficiency of the oversampling techniques.

Research Area and Data Preparation

The research area, Penang Island, is in the northwest of Peninsular Malaysia (Figure 1). To avoid duplicity, the detailed information about Penang Island such as the population, precipitation and geology can be referred to our previous research.[18,19]

The landslide inventory map is also described in figure 1. It provides the location of 382 previous landslide occurrences in Penang Island that are mainly collected from the landslide inventory database, Geographic Information System (GIS) images and field survey was conducted from 1995 to 2009. A landslide occurrence in Penang Island is shown in figure 2.

Landslide influencing factors including six categorical (Aspect, Curvature, Geology, Soil type, Landuse, Rainfall) and five continuous (Elevation, Slope, Distance to drainage, Distance to road, Distance to fault) variables are considered for landslide susceptibility analysis. The sources and formats of the available data as well as the figures of the eleven landslide influencing factors can be referred to our previous studies.[18,19]

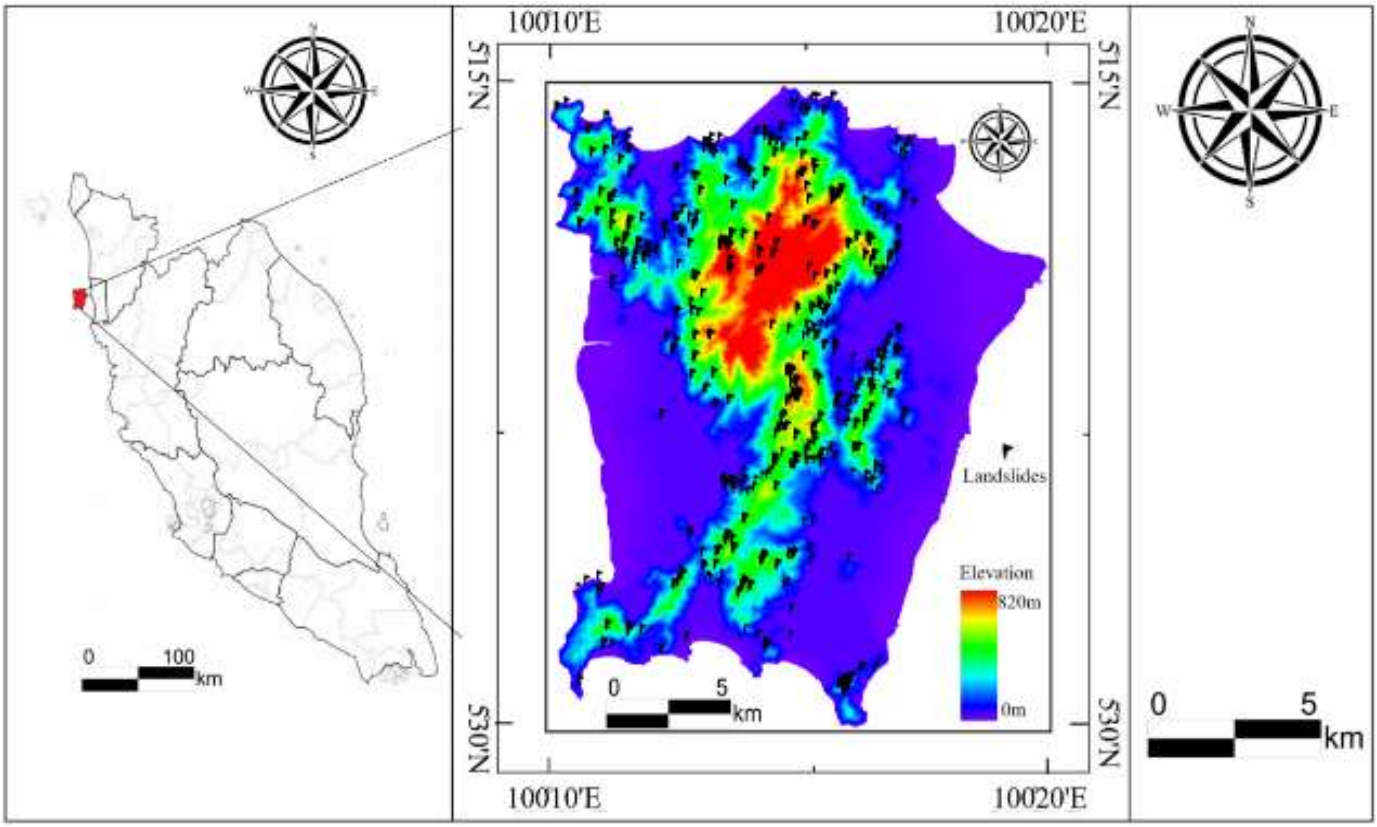


Figure 1: The map of Penang Island

Methodology

Two representative NML models, SVM and ANN and two representative TGB models, XGBoost and LightGBM, are considered to make predictions of landslide susceptibility in Penang Island, Malaysia. The description of landslide susceptibility modeling including SVM, ANN, XGBoost and LightGBM is shown in landslide susceptibility modelling. Three types of datasets are applied to train and validate the models with the sample size of 10,000, 20,000 and 37,546 where the total number of landslide samples accounts for half of it.

The receiver operating characteristic (ROC) curves as well as the area under the curves (AUC) are considered as the main metrics to evaluate the prediction performance of the models. Scalar metrics such as accuracy, recall, precision and F1_score are considered to evaluate the models' performance as well.

Landslide susceptibility modeling

Support vector machine (SVM): As a popular supervised machine learning model, SVM is based on statistical learning theory and the structural risk minimization principle.[45] The objective of SVM is to find a hyperplane with the maximum margin which can be expressed mathematically as:

$$\text{maximize: } \frac{2}{\|w\|} \qquad (1)$$
$$\text{subject to: } y_i(w^T x + b) \geq 1\ ,\ i = 1,2,\ldots,n,$$

where $w = (w_1, w_2, \ldots, w_n)$ denotes the normal vector which determines the direction of the hyperplane and *b* is the displacement which determines the distance from the origin to the hyperplane. Figure 3 gives a typical SVM model.

There are two fundamental principles of SVM when dealing with non-linear classification problems which are the calculation of optimal hyperplane and the selection of kernel function.[47] A function can be regarded as a kernel function when it satisfies the Mercer's Theorem.[35] The commonly used kernel functions and their corresponding mathematical expressions are available in the research by Gao et al[18] and Gao et al.[19]. As one of the most popular kernel functions, radial basis function (RBF) is used in this research according to the performance of previous studies.[8,13,25,27,28,36,40]

Figure 2: Landslide occurrences in Penang Island

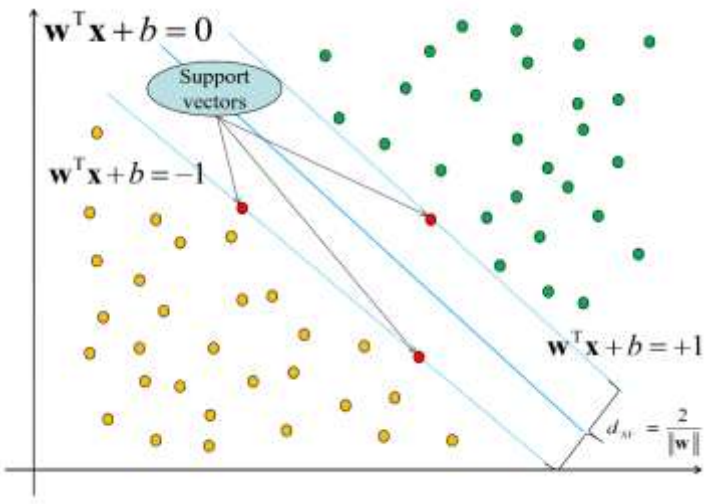


Figure 3: A typical SVM

Artificial neural network (ANN)

ANN is a popular technique used in regression and classification.[30] Back propagation (BP) is the most outstanding algorithm recently.[49] A typical ANN usually has three types of layers, namely, input layer, hidden layer(s) and output layer (Figure 4). The number of input neurons is usually determined by the number of input variables which are the landslide influencing factors in this research.[3] The number of hidden neurons is usually determined by trial and error.[22]

The main goal of an ANN is to build a data generating process which can generalize and predict outputs from inputs.[3] The activation function plays an essential role in ANN models. Rectified Linear Unit (ReLU) and sigmoid functions are selected as the activation functions in the hidden layer and output layer respectively. Batch gradient descent (BGD), a commonly used optimization algorithm in machine learning and deep learning, is used to update the parameters in this research.

eXtreme Gradient Boosting (XGBoost): XGBoost is short term for eXtreme Gradient Boosting, which is a scalable tree boosting model proposed by Chen et al.[12] It is an improved version of gradient boosting decision tree (GBDT).[16] Compared to GBDT, the users can specify the loss function in XGBoost. The base classifiers in XGBoost are CARTs with a single split, usually called a decision stump. XGBboost adopts level-wise learning algorithm to construct trees. The difference between level-wise and leaf-wise tree growing is displayed in figure 5.

The predictive model of XGBoost can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ i = 1,2,\ldots,n, \quad (2)$$

where x_i and $\hat{y}_i$ denote the *i*th sample and the prediction result respectively. The parameter *K* is the total number of tree models. The objective function can be displayed as:

$$O(\cdot) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

The left part in the objective function denotes the training loss. The less is the training loss, the better is the model performs. The right part denotes the complexity of the trees. The less is the model complexity, the higher is the general ability of the model. Since f_k denotes the *k*th tree instead of a numerical vector, the optimization methods such as SGD are unavailable here. To find the best trees, additive training method was applied in XGBoost.[11] Taylor expansion approximation is introduced in XGBoost to make user-defined loss function available and achieve a unity in the form. Taking Taylor expansion of the objective function, we can obtain the new expression as follows:

$$O(\cdot) = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + C \quad (4)$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ denote the first and second derivative of $l(y_i, \hat{y}^{(t-1)})$ with respect to $\hat{y}^{(t-1)}$ respectively. The complexity of a tree in XGBoost is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \quad (5)$$

where *w* and *T* denote the weight of the leaf node and the number of leaf nodes in the tree respectively.

Light Gradient Boosting Machine (LightGBM): LightGBM, proposed by Ke et al[31] is an improved version of gradient boosting decision trees (GBDTs) algorithm. The main idea behind GBTDs is to combine the predictions of multiple decision trees by adding them together. LightGBM adopts a leaf-wise leaf growth strategy with max depth limitation rather than level-wise, which is more prone to overfitting but is more flexible.[33] LightGBM implements a histogram-based algorithm to speed up the training process and reduce memory consumption.

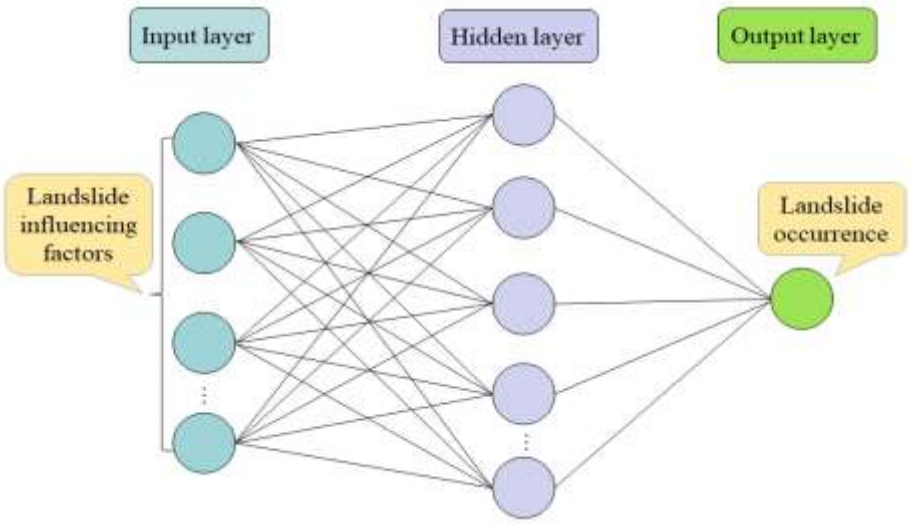


Figure 4: An ANN model

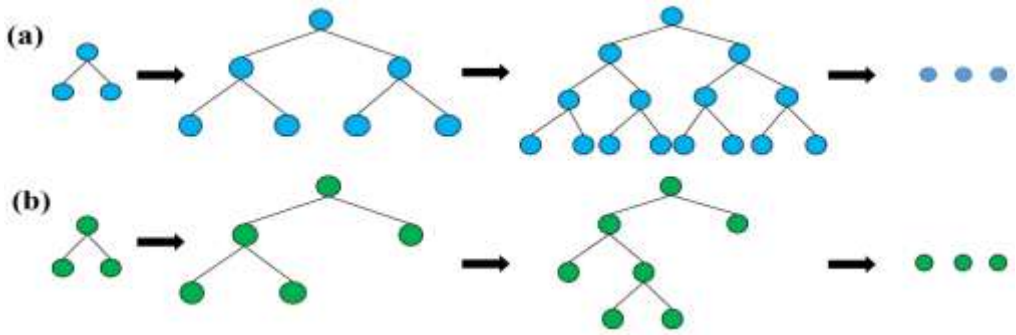


Figure 5: (a) level-wise and (b) leaf-wise algorithm

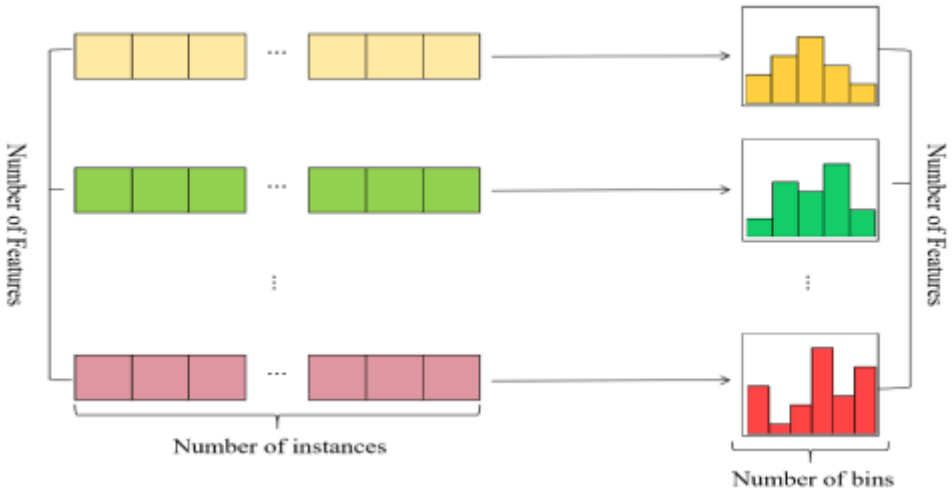


Figure 6: The histogram algorithm

The basic idea of the histogram algorithm is to discretize successive floating-point feature values into *k* integers and construct a histogram with *k* bins. The working process of histogram algorithm is displayed in figure 6. It works by traversing the original data and then accumulating them in the histogram in descending order.

LightGBM models aim to reduce the complexity of histogram building by undersampling the data including the number of instances and the number of features using the newly proposed techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) respectively.[31] The mathematical expressions of information gains of splitting feature *j* before and after applying GOSS are defined as equations (6) and (7) respectively.

$$V_{j|T}(d) = \frac{1}{n_T}\left(\frac{(\sum_{\{x_i \in T: x_{ij} \le d\}} g_i)^2}{n^j_{l|T}(d)} + \frac{(\sum_{\{x_i \in T: x_{ij} > d\}} g_i)^2}{n^j_{r|T}(d)}\right) \quad (6)$$

$$V_j(d) = \frac{1}{n}\left(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i)^2}{n^j_l(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i)^2}{n^j_r(d)}\right) \quad (7)$$

where T denotes the training dataset on a fixed note of the decision tree. The negative gradient of the loss function with respect to the output of the model is denoted as g_i. The datasets A_l and A_r denote the subsets with larger gradient instances in the left and right leaf nodes respectively. B_l and B_r are the subsets with small gradient instances in the left and right leaf nodes after a random sampling. The coefficient $\frac{1-a}{b}$ is applied to normalize the sum of the gradients over B where a and b denote the percentage of large gradient instances and the percentage of small gradient instances after removing the large gradient instances.[31]

The computation cost is greatly reduced and the training accuracy is not lost much after using GOSS.[31] The main idea behind EFB is to bundle exclusive features into a single feature, namely, EFB. Since the features never take non-zero values at the same time, EFB algorithm speeds up the model training process without hurting the accuracy.[31]

Feature selection: Feature selection plays an essential role in data mining which can help reduce the dimension of the datasets and further improve the models' performance in some cases.[21] Two methods of feature selection, RFC and ETC, are applied into the datasets in this research.

RFC is a commonly used feature selection method in various fields.[19,29,32,34,38,43] The working principle of RFC is made available by Gao et al.[19] ETC is a popular feature selection method as well which is like RFC. The major difference between ETC and RFC is the construction of the selection trees. On the one hand, ETC splits nodes by randomly choosing cut points while RFC chooses the optimum split points. On the other hand, all the available training samples are used to grow the tree for ETC rather than a bootstrap replica[14] which is used to draw samples without replacement in RFC.[20]

Oversampling Techniques: Oversampling techniques SMOTE[10] and SCOTE, are applied in this research. SMOTE is a popular oversampling technique to generate synthetic samples by operating in feature space instead of data space. SCOTE is a new oversampling technique proposed earlier.[19] The main objective of the two oversampling methods is to effectively augment the landslide occurrence samples.

Dataset building: The data used in this research is in pixel format. The total number of pixels is 3,004,631 with 2,984,386 non-landslide pixels and 20,245 landslide pixels where the number of 1,472 landslide pixels are considered as non-landslide pixels after a first-round sampling described in the previous research.[19] Therefore, the number of landslide pixels is being changed to 18,773. Three groups of datasets shown in table 1 are built to train models including the dataset with 10,000, 20,000 and 37,546, all with equal ratio, for landslide and non-landslide pixels. The random undersampling method is used to select landslide and non-landslide samples.

Evaluation methods: Scalar metrics such as accuracy, recall and precision are widely used in evaluating models' performances. Precision and recall can be combined into the F-score. The detailed definitions and mathematical expressions are provided by Gao et al.[18,19] ROC curve is a type of highly popular and effective method to evaluate the binary classifier's overall performance. It is a probability curve which shows the ability of the classification model to rank the positive samples relative to the negative samples. The value of the area under the ROC curve (AUC) is a commonly used indicator to measure the prediction performance ranging from 0.5 to 1. It measures how much the models can distinguish between different classes. The higher is the value of AUC, the better is the prediction performance. As a threshold-independent metric, AUC value is highly recommended in the model evaluation procedure in this research.[15]

Results and Discussion

The multicollinearity analysis is applied to discover redundant factors in this research which is an essential step in LSMs.[41] The results of multicollinearity analysis are displayed in table 2 which show that no factors should be removed in this research based on the metrics of tolerance (*Tol*) and variance inflation factor (*VIF*).[39] When the *Tol* value of an influencing factor is less than 0.2 or the *VIF* value is larger than 5, it indicates that the multicollinearity problem may exist and the factor should be removed from the model.[39]

All the experiments in this research are conducted using Python 3.60 in a Windows 10 server with an Intel Core i5 2.40 GHz processor. The datasets are split into training and validation datasets with the ratio of 80:20 for all models. The overall datasets with 3,004,631 pixels are considered to produce the LSMs using ArcGIS. The data are normalized using min-max normalization before entering the models. RBF function is selected as the kernel function for SVM models.

Table 1
The datasets used in this research

Dataset	No. of landslide samples	No. of non-landslide samples	No. of pixels	No. of factors
Data 1	5,000	5,000	10,000	11
Data 2	10,000	10,000	20,000	11
Data 3	18,773	18,773	37,546	11

Table 2
The results of multicollinearity analysis

Landslide factors	Data 1		Data 2		Data 3	
	Tol	***VIF***	***Tol***	***VIF***	***Tol***	***VIF***
Aspect	.707	1.415	.719	1.391	.715	1.400
Curvature	.670	1.493	.686	1.458	.679	1.472
Geology	.515	1.943	.524	1.908	.523	1.911
Soil	.686	1.459	.674	1.485	.678	1.475
Landuse	.844	1.184	.820	1.220	.824	1.214
Precipitation	.616	1.624	.615	1.626	.618	1.619
Height	.672	1.487	.665	1.505	.665	1.504
Distance to drainage	.752	1.329	.769	1.301	.760	1.316
Distance to road	.647	1.546	.649	1.541	.643	1.556
Distance to fault	.747	1.338	.745	1.343	.752	1.330
Slope	.631	1.584	.645	1.550	.642	1.559

The parameter gamma (γ) and penalty parameter *C* are determined using grid search method. The combination of (γ, *C*) is set to (3, 20). ReLU function is used in the hidden layer of the ANN models. The learning rate (α) is an important parameter to tune which is determined as 0.01 in this research. The hidden layer size is set to 25, 30 and 40 for data 1, data 2 and data 3 respectively. The parameter tuning for TGB models is a key procedure as well and is mainly based on 'trial and error' method. The learning rate and the maximum depth of the trees are considered to tune in XGBoost and LightGBM models. Learning rate (namely, eta for XGBoost) is used to control the weighting of new trees added to the model. The max_depth is used to control the depth of the tree to avoid overfitting. The parameter settings are shown in table 3.

The values for scalar metrics such as accuracy, recall, precision and F_1_score and AUC values for the original datasets and the newly generated datasets are shown in table 4. The corresponding ROC curves are displayed in figure 7. Even though they are considered as the less important metrics for model validation in this research, they provide some useful information as well, especially the F_1_score which can be interpreted as a weighted average of the precision and recall. In table 4, the highest value of F_1_score is 0.094 for the XGBoost model trained using data 3.

Based on the ROC values, the results show that the larger is the sample size, the better is the model performance which is suitable for all four models used in the research. For example, the ROC values for XGBoost are 0.9072, 0.9116 and 0.9227 for data 1, data 2 and data 3 respectively. The highest AUC values for XGBoost and LightGBM are 0.9227 and 0.9187 respectively. For ANN and SVM models, the AUC values are all less than 0.900, except for the SVM model trained using data 3 with the AUC of 0.9019. Overall, TGB models show better performance than NML models in this landslide research.

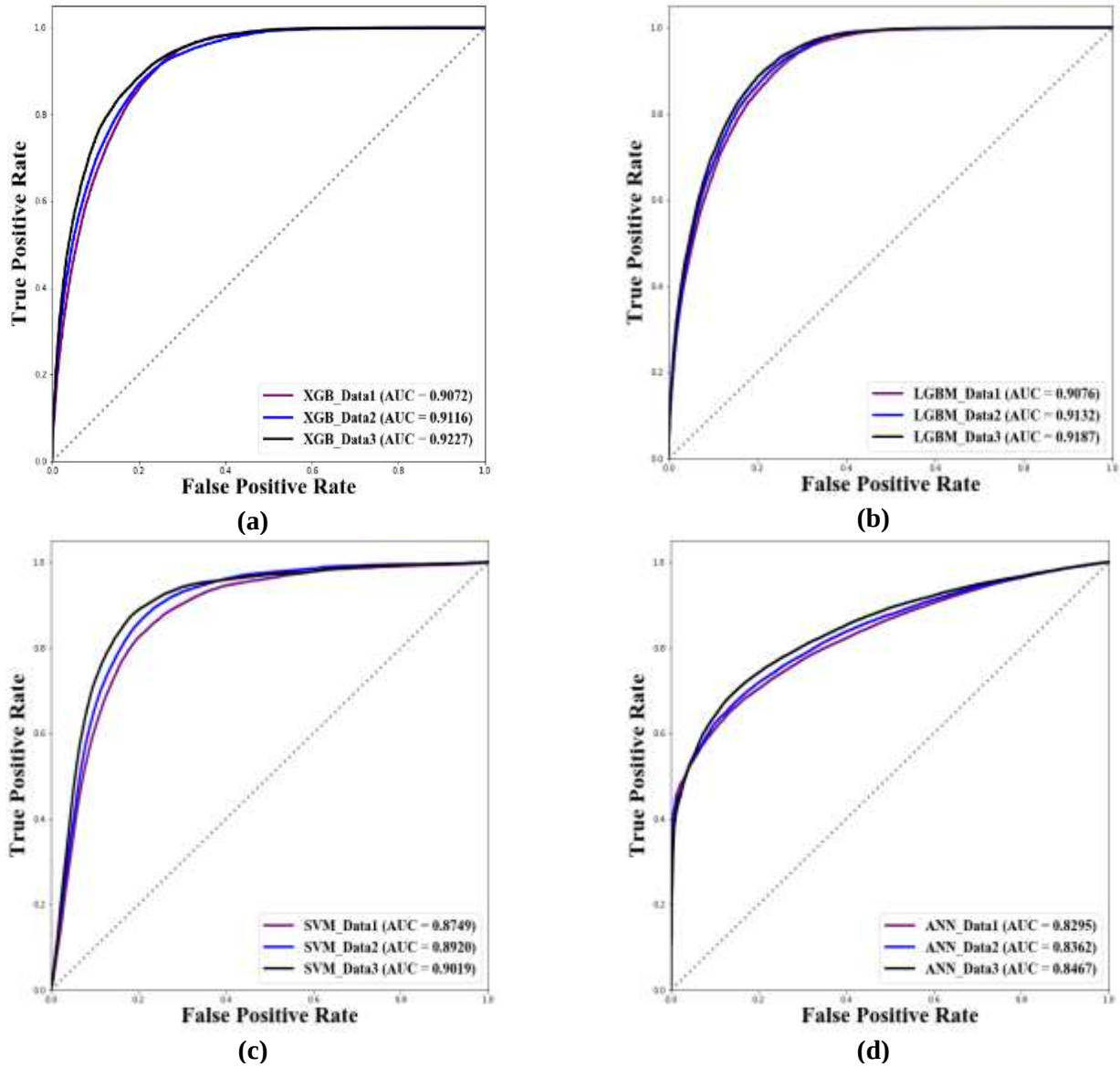


Figure 7: The ROC curves of (a) XGBoost; (b) LightGBM; (c) SVM; (d) ANN

Table 3
The parameter settings for TGB models

Model	Dataset used	Parameter setting
XGBoost	Data 1	eta=0.2; max_depth=10; others=default
	Data 2	eta=0.2; max_depth=10; others=default
	Data 3	eta=0.3; max_depth=10; others=default
LightGBM	Data 1	learning rate=0.2; max_depth=10; others=default
	Data 2	learning rate=0.2; max_depth=10; others=default
	Data 3	learning rate=0.2; max_depth=10; others=default

Table 4
The results of the models trained using the datasets without feature selection

Modelling	Dataset	Accuracy (%)			Recall	Precision	F_1_ score	AUC
		Training	Validation	Overall				
XGBoost	Data 1	95.31	88.10	87.67	0.718	0.038	0.073	0.9072
	Data 2	95.58	91.20	89.95	0.692	0.045	0.085	0.9116
	Data 3	93.89	91.45	90.65	0.723	0.050	**0.094**	**0.9227**
LightGBM	Data 1	92.38	86.45	86.77	0.739	0.037	0.070	0.9076
	Data 2	92.37	88.52	87.50	0.745	0.039	0.074	0.9132
	Data 3	90.55	88.89	88.10	0.752	0.041	0.078	0.9187
SVM	Data 1	85.18	85.05	81.23	0.809	0.028	0.055	0.8749
	Data 2	85.98	85.70	81.32	0.843	0.030	0.057	0.8920
	Data 3	86.06	86.75	83.10	0.856	0.033	0.064	0.9019
ANN	Data 1	71.93	70.30	75.87	0.720	0.020	0.039	0.8295
	Data 2	72.27	71.00	76.69	0.728	0.021	0.040	0.8362
	Data 3	74.87	73.18	76.63	0.764	0.022	0.042	0.8467

Feature selection methods RFC and ETC are considered to measure the importance of the eleven influencing factors. Moreover, sub-data sets constructed based on the results of feature selection are used to train and validate the TGB models. Before applying the feature selection methods, a total of eleven influencing factors are considered for model training and validation. The results of feature importance using RFC and ETC is displayed in tables 8 and 9 respectively.

Three experiments are conducted both for RFC and ETC respectively. The average values are considered to determine the feature importance. For example, the important scores of the factor Height obtained by RFC are 0.2201, 0.2204 and 0.2203 for the first, second and third time respectively. Therefore, the final importance score is 0.2203 which is obtained by calculating the average of the three scores.

According to the results shown in tables 5 and 6, the two least important variables are both curvature and soil type. The biggest difference between RFC and ETC is the order of the three most important variables. For RFC, the most three important features are height, fault and road in a descending order while for ETC, the order of the most three important features is fault, road and height. According to the feature importance results, three new datasets are generated with nine (Data3_F9), six (Data3_F6) and three (Data3_F3) factors from data 3 with the largest sample size respectively. The details of the newly generated datasets are shown in table 7.

The results for the TGB models trained with the subdatasets are shown in table 8. The ROC curves are displayed in figure 8. According to the results, the AUC values for XGBoost and LightGBM trained using Data3_F9 are 0.9494 and 0.9299 respectively which are both higher than the largest AUC value 0.9227 obtained from the XGBoost trained using data 3. For the results obtained by Data3_F6, the AUC values are comparable with those AUC values without feature selection.

Based on tables 5 and 6, the sum of the importance score of the first six important factors is around 92% for both ETC and RFC. It indicates that the removed five factors contribute only 8% feature importance. The models trained with the dataset with only three most important factors show poor performance with the AUC values of 0.8335 and 0.8400 for TGBoost and LightGBM respectively.

After combining the two TGB models with feature selection methods, the results are improved based on the AUC values. What to do next is to combine the TGB models with oversampling methods, SMOTE and SCOTE based on the original data 3. The new datasets generated with oversampling methods are described in table 9. The number of landslide samples doubled from 18773 to 37546. The total number of samples are 75192 after considering oversampling techniques.

Table 5
The feature importance using RFC

Factors	1st	2nd	3rd	Average	Top 3	Top 6	Top 9
Height	0.2201	0.2204	0.2203	0.2203	**0.6542**	**0.9175**	**0.9897**
Fault	0.2198	0.2185	0.2187	0.2190			
Road	0.2148	0.2152	0.2148	0.2149			
Drainage	0.1063	0.1062	0.1060	0.1062	0.3458		
Slope	0.0966	0.0961	0.0962	0.0963			
Aspect	0.0605	0.0607	0.0612	0.0608			
Rainfall	0.0319	0.0321	0.0324	0.0321		0.0825	
Geology	0.0227	0.0223	0.0227	0.0226			
Landuse	0.0169	0.0182	0.0175	0.0175			
Curvature	0.0090	0.0090	0.0089	0.0090			0.0103
Soil type	0.0014	0.0014	0.0015	0.0014			

*The three terms Fault, Road and Curvature stand for distance to fault, distance to road and distance to curvature, respectively.

Table 6
The feature importance using ETC

Factors	1st	2nd	3rd	Average	Top 3	Top 6	Top 9
Fault	0.2114	0.2119	0.2123	0.2119	**0.5953**	**0.9228**	**0.9880**
Road	0.1969	0.1962	0.1965	0.1965			
Height	0.1859	0.1877	0.1872	0.1869			
Drainage	0.1334	0.1344	0.1338	0.1339	0.4047		
Slope	0.1284	0.1284	0.1286	0.1285			
Aspect	0.0664	0.0647	0.0643	0.0651			
Rainfall	0.0327	0.0325	0.0330	0.0327		0.0772	
Landuse	0.0171	0.0171	0.0169	0.0170			
Geology	0.0157	0.0151	0.0157	0.0155			
Curvature	0.0094	0.0092	0.0090	0.0092			0.0120
Soil type	0.0028	0.0028	0.0028	0.0028			

*The three terms Fault, Road and Curvature stand for distance to fault, distance to road and distance to curvature, respectively.

Table 7
The new datasets after feature selection

Datasets	No. of Factors	No. of Samples	Features included	Sum of importance score (RFC/ETC)
Data3_F9	9	37546	Fault, Road, Height, Drainage, Slope, Aspect, Rainfall, Landuse, Geology	0.9897/0.9880
Data3_F6	6	37546	Fault, Road, Height, Drainage, Slope, Aspect	0.9175/0.9228
Data3_F3	3	37546	Fault, Road, Height	0.6542/0.5953

Table 8
The results of the models trained using the datasets with feature selection

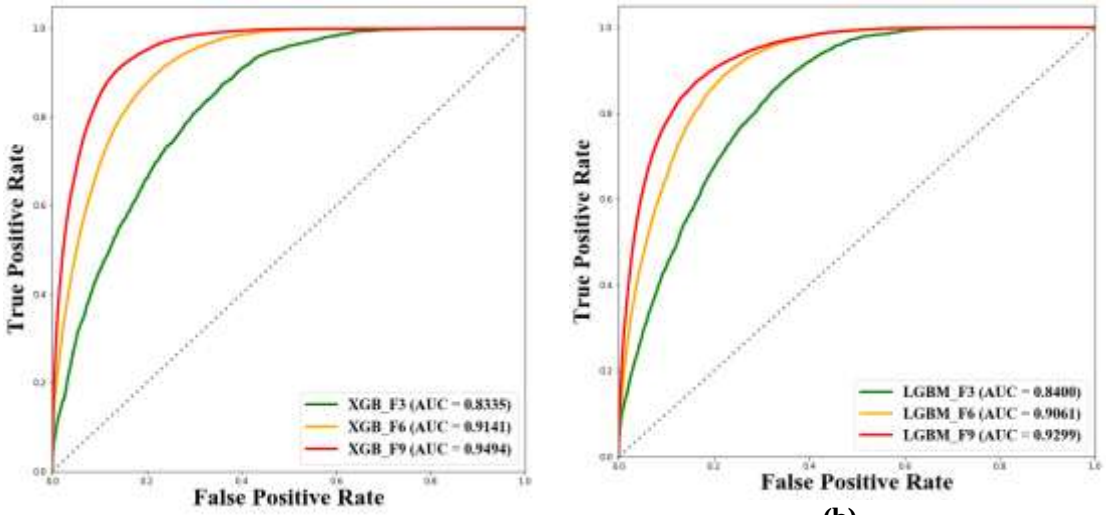
Modelling	Dataset	Accuracy (%)			Recall	Precision	F_1_ score	AUC
		Training	Validation	Overall				
XGBoost	Data3_F9	99.60	96.58	95.81	0.097	0.624	0.167	**0.9494**
	Data3_F6	96.70	92.81	91.97	0.052	0.627	0.095	0.9141
	Data3_F3	84.95	83.17	79.19	0.022	0.675	0.042	0.8335
LightGBM	Data3_F9	98.50	95.14	94.29	0.073	0.639	0.131	**0.9299**
	Data3_F6	97.23	93.17	92.37	0.049	0.564	0.091	0.9061
	Data3_F3	87.53	85.90	82.86	0.024	0.608	0.046	0.8400

Table 9
The new datasets after oversampling techniques

Dataset	No. of landslide samples	No. of non-landslide samples	No. of pixels	Data source
Data3_SMOTE	37,546	37,546	75,192	Original+SMOTE
Data3_SCOTE	37,546	37,546	75,192	Original+SCOTE

Table 10
The results of the models trained using the datasets with oversampling

Modelling	Dataset	Accuracy (%)			Recall	Precision	F_1_ score	AUC
		Training	Validation	Overall				
XGBoost	Data3_SMOTE	99.98	98.15	97.25	0.563	0.134	0.216	**0.9525**
	Data3_SCOTE	99.99	97.01	94.37	0.713	0.081	0.146	0.9446
LightGBM	Data3_SMOTE	96.60	95.40	94.05	0.676	0.074	0.133	0.9378
	Data3_SCOTE	95.07	93.22	91.27	0.761	0.056	0.105	0.9385



(a) **(b)**

Figure 8: The ROC curves of (a) XGBoost; (b) LightGBM based on Data3_F9.

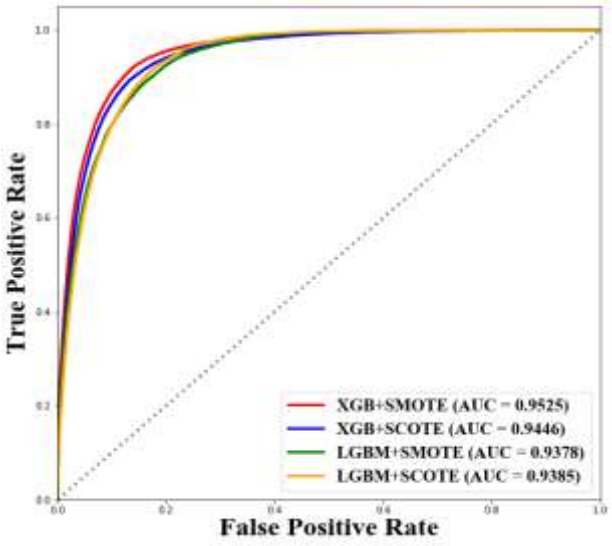


Figure 9: The ROC curves of LGB models with oversampling methods

The results after using oversampling techniques are displayed in table 10 and the ROC curves are displayed in figure 11. The results are further improved based on the AUC values after applying the oversampling methods to augment the landslide samples. The optimal performance (AUC=0.9525) occurred when combining the XGBoost and SMOTE method. The other AUC values around 0.94 are satisfactory as well.

XGBoost and LightGBM are applied to produce LSMs since they show better prediction performance. Figs. 10-12 display the LSMs produced based on the XGBoost and LightGBM model trained using the original datasets (Data 2 and Data 3) and newly generated data3_F9 respectively. The landslide susceptibility index (LSI) values for XGBoost and LightGBM models are sorted in descending order and then classified into four susceptibility categories: 'Very High [0-10%]', 'High [10-20%]', 'Medium [20-60%]' and 'Low [60-100%]' using natural breaks method. The susceptible area denotes the sum of the percentages of 'Very High' and 'High' category, namely, the first 20% of the LSI values of each dataset after being sorted in descending order.

The susceptible area in LSMs is mainly located in the middle mountainous region and the northwestern area, which is highly consistent with the landslide inventory map displayed in figure 1.

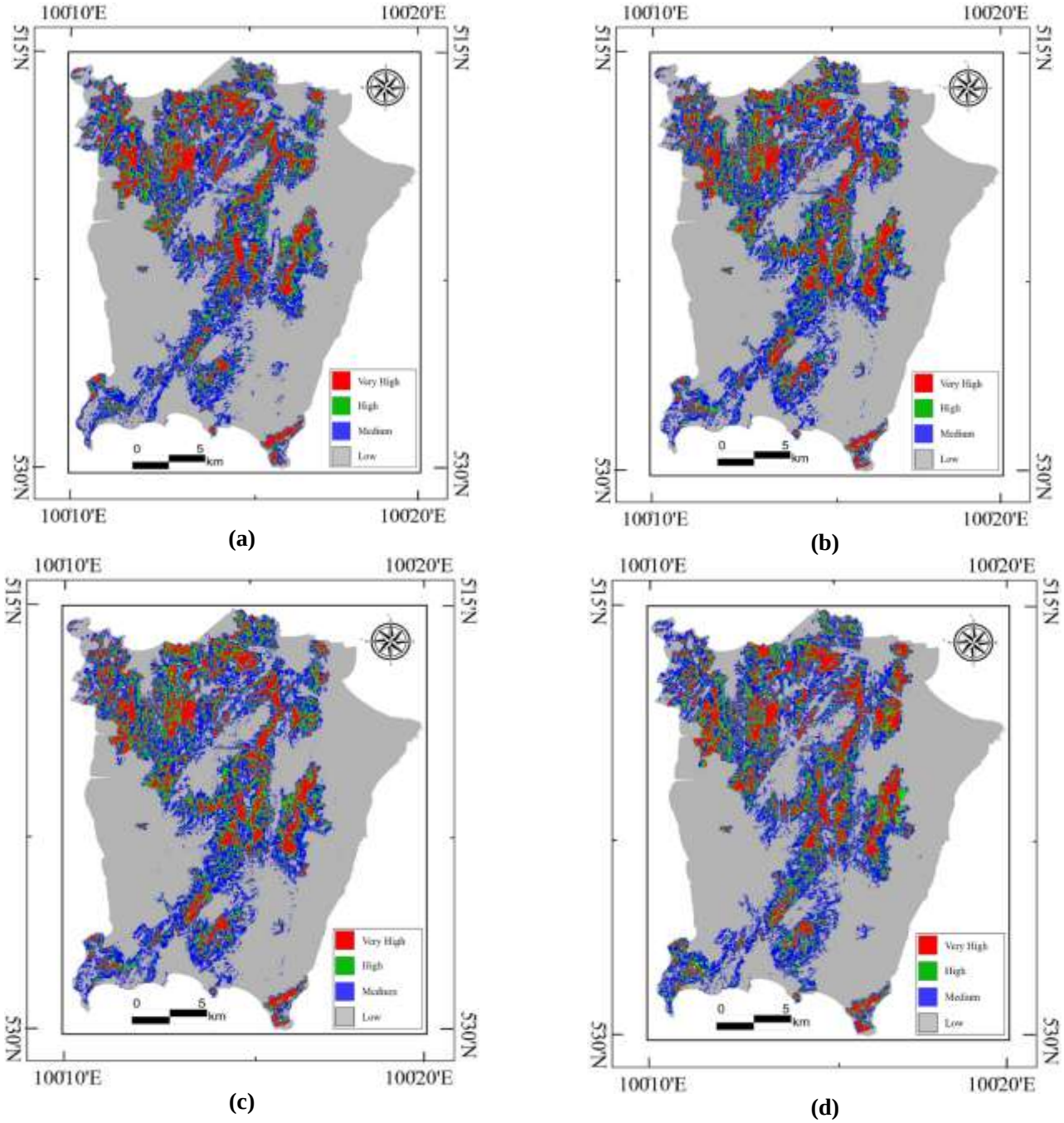
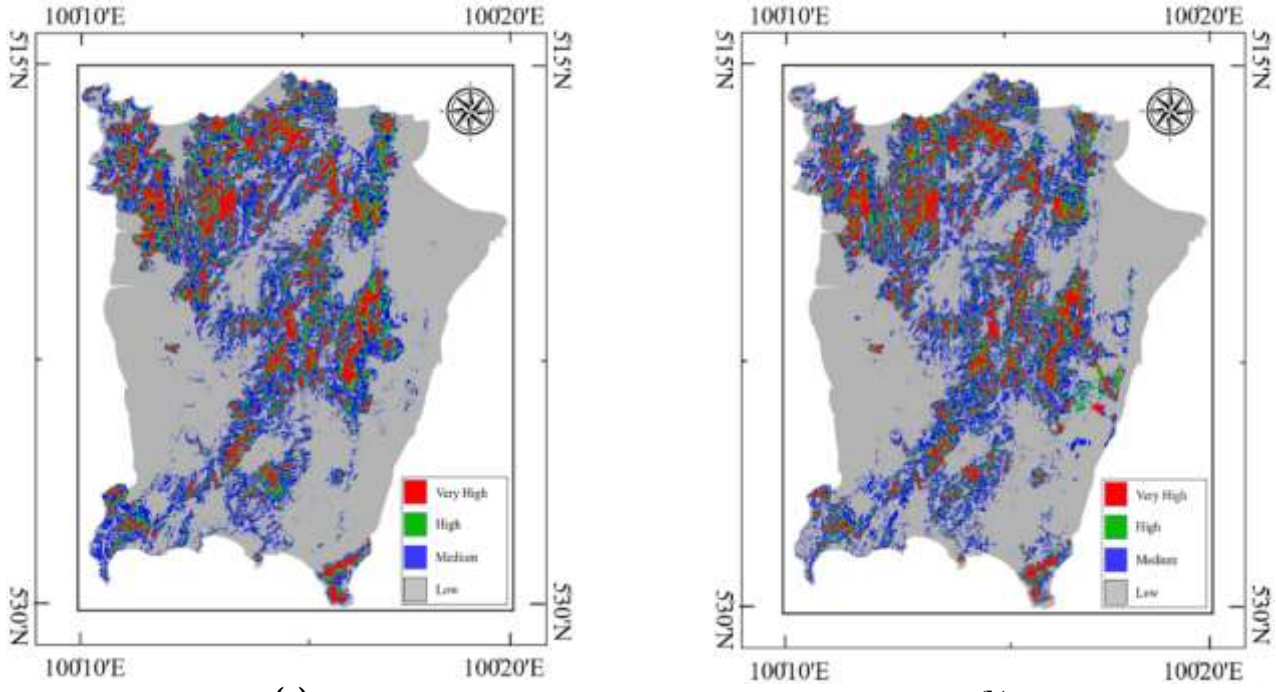


Figure 10: The LSMs produced by (a) XGBoost trained based on Data 2; (b) LightGBM trained based on Data 3; (c) LightGBM trained based on Data 2; (d) LightGBM trained based on Data 3



(a) **(b)**

Figure 11: The LSMs produced by (a) XGBoost trained based on Data3_F9; (b) LightGBM trained based on Data3_F9

The verification analysis of the LSMs is a key step in landslide spatial area which is usually conducted by comparing the prediction results with the existing landslide data based on two key assumptions.[42] The first one is that the landslide occurrences are related to the spatial distribution. Another one is that landslides are controlled by some influencing factors which can be analyzed statistically or empirically. The verification results for the two TGB models trained using the original datasets (Data 2 and Data 3), newly generated datasets (Data3_F3, Data3_F6 and Data3_F9) and the datasets (Data3_SMOTE and Data3_SCOTE) generated by oversampling techniques are displayed in figs. 13-15 respectively.

Based on the values of susceptible area in figure 14, the susceptible area XGBoost and LightGBM models both trained using data 3 successfully predicted account for 88.3% and 87.9% of the previous landslides respectively. When considering the newly generated samples to train the TGB models, the verification results are improved as well. The susceptible area for all models can cover more than 90% of the previous landslides. In particular, the susceptible area predicted by the XGBoost models trained using data3_F9 covers 94.8% of the landslide occurrences. Moreover, the susceptible area predicted by the XGBoost model trained using data3_SMOTE attains to 95.3% that is the highest value among the models. Two TGB models (XGBoost and LightGBM) and two NML models (SVM and ANN) are compared to produce the landslide susceptibility maps in Penang Island, Malaysia. XGBoost and LightGBM are both specific implementations of GBDT which is the initial version of gradient boosting model.[16] They share the advantage of GBDT models which are less likely overfitting and more likely generalizing data well.[31] The results suggest that XGBoost and LightGBM show better prediction performance compared to SVM and ANN.

Feature selection methods based on the feature importance are considered even though the feature dimension is not that high. However, the results are improved by removing two least important variables (Curvature and Soil type) from the dataset (Data 3). It indicates that the two variables may provide redundant information to the models. The total importance of the two variables accounts for 1% (1.03% for FRC and 1.2% for ETC). When continuing removing variables to six and three, the results show that the models' performance deteriorates. In other words, some useful information is removed. The results further got improved after considering the oversampling techniques which denote that the newly generated data can provide useful and efficient information for the model training.

There are several limitations of this research. First, the random undersampling technique is used to select landslide and non-landslide samples. For the non-landslide samples, only a minority of them are selected to train the models. During the undersampling, the distribution of the non-landslide samples is easily being changed. The stratified sampling deserves a try for future work. Secondly, the parameter tuning process for machine learning models plays a highly important role in the model training process. For XGBoost and LightGBM models, they have various types of parameters, both continuous and categorical, to tune. Theoretically, it is impossible to find out the optimum parameter combination for the models.

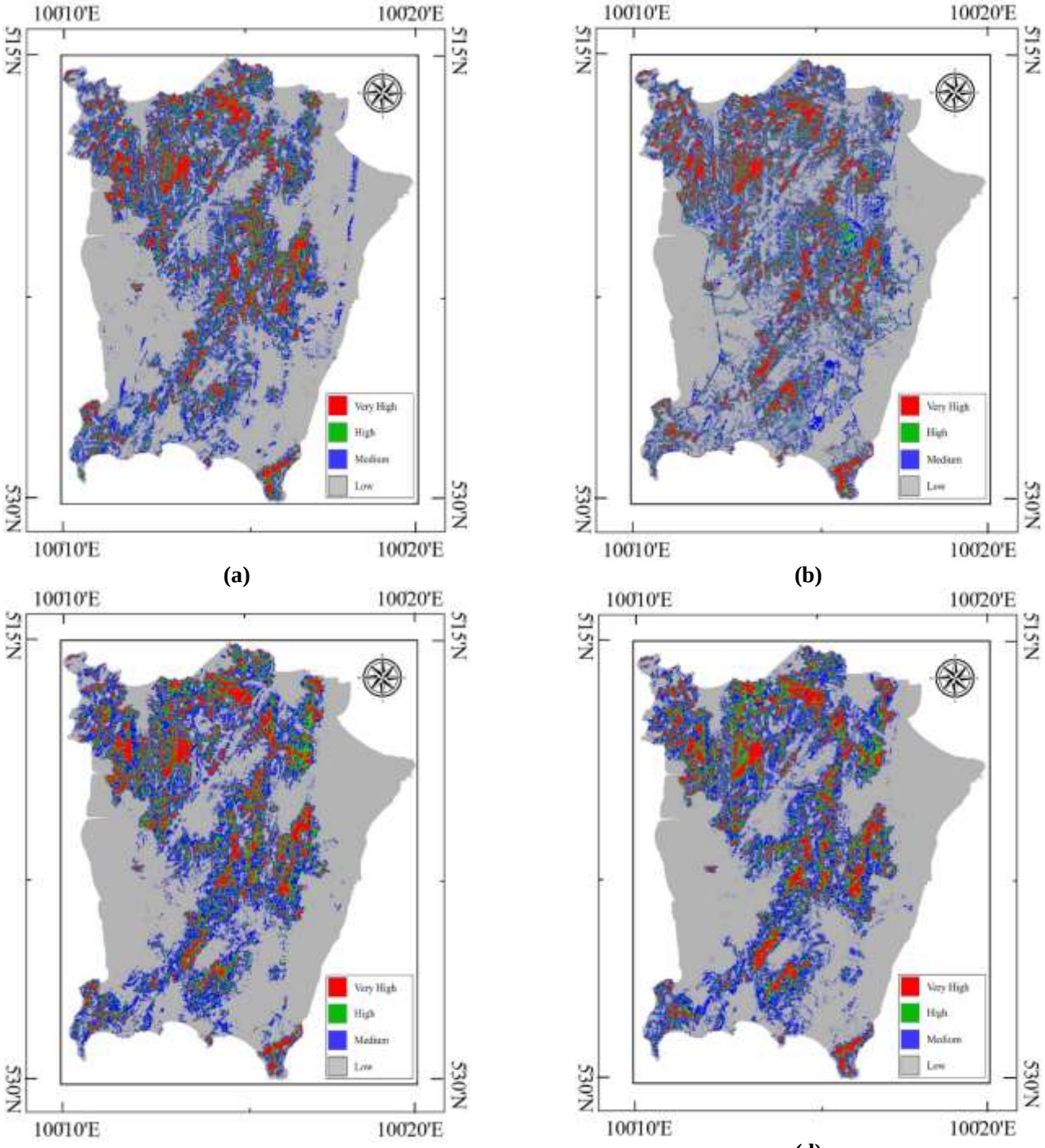
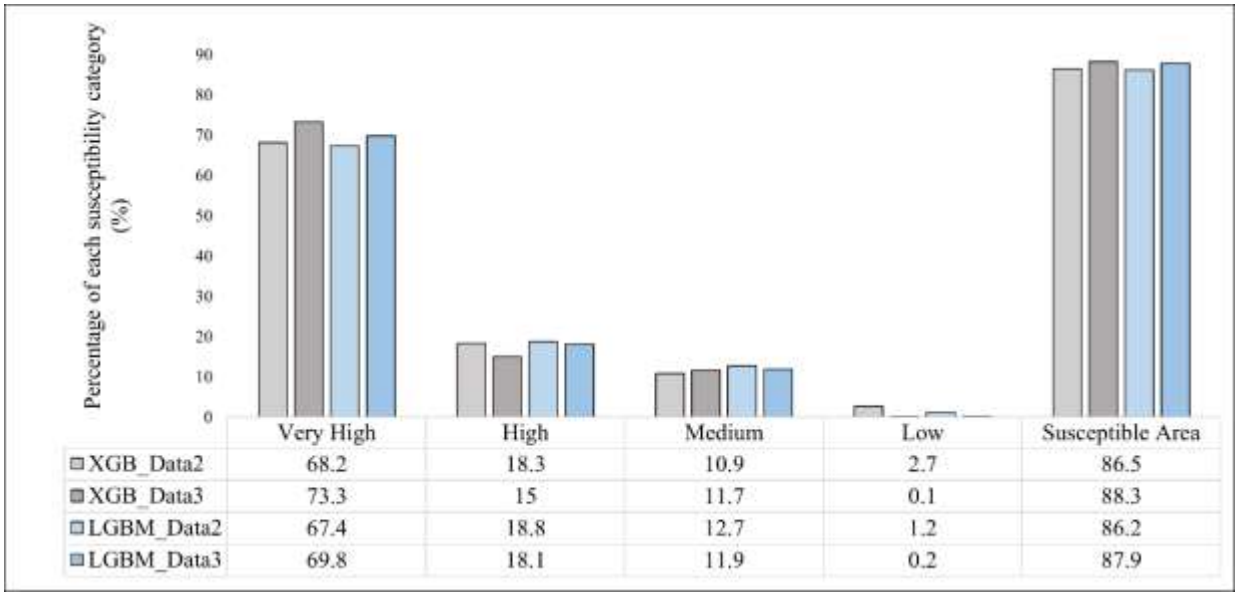


Figure 12: The LSMs produced by (a) XGB+SMOTE; (b) XGB+SCOTE; (c) LightGBM+SMOTE; (d) LightGBM+SCOTE

Conclusion

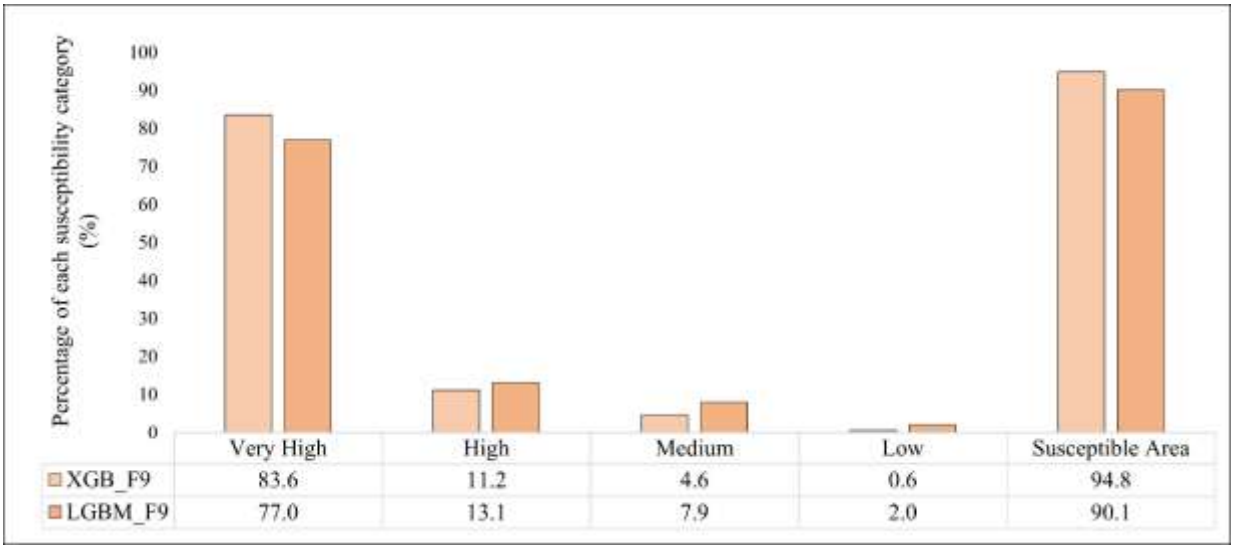
In this research, two tree-based gradient boosting models, namely, LightGBM and XGBoost and two non-tree-based machine learning models, namely, ANN and SVM are applied to the landslide spatial prediction research in Penang Island, Malaysia. Moreover, two feature selection methods, namely, ETC and RFC, are applied to gauge the feature importance. Two oversampling techniques are considered in this research as well. A total of six datasets with different sample sizes or generated using feature selection methods is used to train the models. Three of them, namely, data 1, data 2 and data 3, are made of original samples and the remaining three datasets, namely, data3_F9, data3_F6 and data3_F3, are part of data 3.

The results show that the TGB models outperform the NML models based on the AUC values, no matter which dataset is used to train.



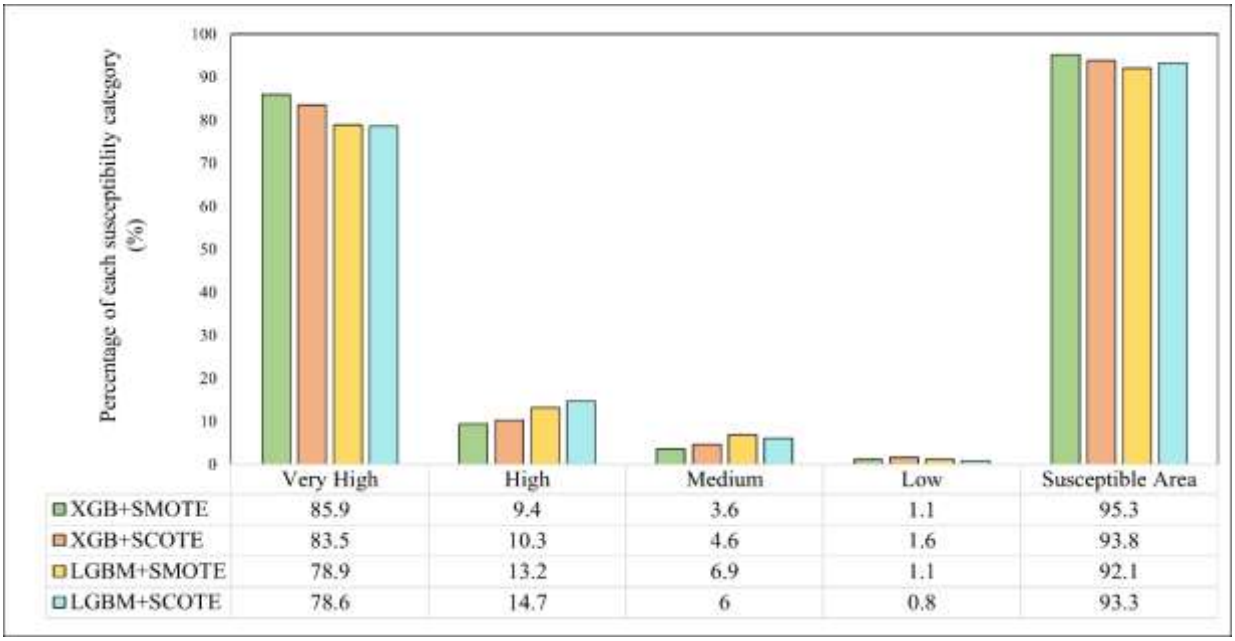
	Very High	High	Medium	Low	Susceptible Area
XGB_Data2	68.2	18.3	10.9	2.7	86.5
XGB_Data3	73.3	15	11.7	0.1	88.3
LGBM_Data2	67.4	18.8	12.7	1.2	86.2
LGBM_Data3	69.8	18.1	11.9	0.2	87.9

Figure 13: The verification results for XGBoost and LightGBM models based on Data 2 and Data 3



	Very High	High	Medium	Low	Susceptible Area
XGB_F9	83.6	11.2	4.6	0.6	94.8
LGBM_F9	77.0	13.1	7.9	2.0	90.1

Figure 14: The verification results for XGBoost and LightGBM models based on Data3_F9



	Very High	High	Medium	Low	Susceptible Area
XGB+SMOTE	85.9	9.4	3.6	1.1	95.3
XGB+SCOTE	83.5	10.3	4.6	1.6	93.8
LGBM+SMOTE	78.9	13.2	6.9	1.1	92.1
LGBM+SCOTE	78.6	14.7	6	0.8	93.3

Figure 15: The verification results for XGBoost and LightGBM models with oversampling methods

Moreover, the performance of the TGB models is improved after combining with the feature selection methods. Therefore, two research problems, namely, which model performs best among ANN, SVM, XGBoost and LightGBM and whether the feature selection methods can improve the performance of the models, are solved well. The LSMs are produced using LightGBM and XGBoost models trained using data 2, data 3 and data3_F9. The verification analysis results suggest that more than 85% and 90% of the previous landslide occurrences are successfully predicted by the maps produced by TBG models without and with feature selection techniques respectively. It indicates that all the maps can provide useful information to the decision makers and the feature selection and oversampling techniques can help improve the performance of model prediction in this research. Furthermore, the maps produced using XGBoost and SMOTE are highly recommended to be applied in the landslide mitigation and management in Penang Island, Malaysia.

References

1. Aditian A., Kubota T. and Shinohara Y., 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 **(2018)**

2. Alkhasawneh M.S., Ngah U.K., Tay L.T., Mat Isa N.A. and Al-Batah M.S., Modeling and testing landslide hazard using decision tree, *Journal of Applied Mathematics*, https://doi.org/10.1155/2014/929768 **(2014)**

3. Atkinson P.M. and Tatnall A.R.L., Introduction neural networks in remote sensing, *International Journal of Remote Sensing*, **18(4)**, 699-709 **(1997)**

4. Ballabio C. and Sterlacchini S., Support vector machines for landslide susceptibility mapping: the Staffora River Basin case study, Italy, *Mathematical Geosciences*, **44**, 47-70 **(2012)**

5. Bandara A. et al, A Generalized Ensemble Machine Learning Approach for Landslide Susceptibility Modeling, Data Management, Analytics and Innovation, Springer, 71-93 **(2020)**

6. Bhasin M. and Raghava G., Prediction of CTL epitopes using QM, SVM and ANN techniques, *Vaccine*, **22**, 3195-3204 **(2004)**

7. Breiman L., Arcing classifier (with discussion and a rejoinder by the author), *The Annals of Statistics*, **26(3)**, 801-849 **(1998)**

8. Bui D.T., Tuan T.A., Klempe H., Pradhan B. and Revhaug I., Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression and logistic model tree, *Landslides*, **13**, 361-378 **(2016)**

9. Chang Z., Du Z., Zhang F., Huang F., Chen J., Li W. and Guo Z., Landslide susceptibility prediction based on remote sensing images and GIS: Comparisons of supervised and unsupervised machine learning models, *Remote Sensing*, **12(3)**, 502 **(2020)**

10. Chawla N.V., Bowyer K.W., Hall L.O. and Kegelmeyer W.P., SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16**, 321-357 **(2002)**

11. Chen T., Introduction to boosted trees, *University of Washington Computer Science*, **22**, 115 **(2014)**

12. Chen T. and Guestrin C., Xgboost: A scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794 **(2016)**

13. Chen W.H., Hsu S.H. and Shen H.P., Application of SVM and ANN for intrusion detection, *Computers and Operations Research*, **32**, 2617-2634 **(2005)**

14. Efron B., The jackknife, the bootstrap and other resampling plans, Society for Industrial and Applied Mathematics **(1982)**

15. Fawcett T., An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861-874 **(2006)**

16. Friedman J., Greedy function approximation: A stochastic boosting machine, Technical Report, Department of Statistics, Stanford University **(1999)**

17. Gao H., Fam P.S., LowH.C., Tay L.T. and Lateh H., An overview and comparison on recent landslide susceptibility mapping methods, *Disaster Advances*, **12(12)**, 46-64 **(2019)**

18. Gao H., Fam P.S., Tay L.T. and Low H.C., Comparative landslide spatial research using statistical and machine learning models in Penang Island, Malaysia, *Bulletin of Engineering Geology and the Environment*, https://doi.org/10.1007/s10064-020-01969-7 **(2020a)**

19. Gao H., Fam P.S., Tay L.T. and Low H.C., Three oversampling methods applied in a comparative landslide spatial research in Penang Island, Malaysia, *SN Applied Sciences***, 2(9)**, 1-20 **(2020b)**

20. Geurts P., Ernst D. and Wehenkel L., Extremely randomized trees, *Machine Learning*, **63(1)**, 3-42 **(2006)**

21. Ghaemi M. and Feizi-Derakhshi M.R., Feature selection using forest optimization algorithm, *Pattern Recognition*, **60**, 121-129 **(2016)**

22. Gong P., Integrated analysis of spatial data for multiple sources: using evidential reasoning and artificial neural network techniques for geological mapping, *Photogrammetric Engineering and Remote Sensing*, **62(5)**, 513-523 **(1996)**

23. Guzzetti F., Carrara A., Cardinali M. and Reichenbach P., Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy, *Geomorphology*, **31**, 181-216 **(1999)**

24. Hong H., Liu J., Bui D.T., Pradhan B., Acharya T.D., Pham B.T., Zhu A.X., Chen W. and Ahmad B.B., Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China), *Catena*, **163**, 399-413 **(2018)**

25. Hong H., Pradhan B., Jebur M.N., Bui D.T., Xu C. and Akgun A., Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines, *Environmental Earth Sciences***, 75(1)**, 40 **(2016)**

26. Huabin W., Gangjun L., Weiya X. and Gonghui W., GIS-based landslide hazard assessment: an overview, *Progress in Physical Geography*, **29**, 548-567 **(2005)**

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

28. Jebur M.N., Pradhan B. and Tehrany M.S., Manifestation of LiDAR-derived parameters in the spatial prediction of landslides using novel ensemble evidential belief functions and support vector machine models in GIS, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **8**, 674-690 **(2014)**

29. Jiskoot H., Curran C.J., Tessler D.L. and Shenton L.R., Changes in Clemenceau Icefield and Chaba Group glaciers, Canada, related to hypsometry, tributary detachment, length–slope and area–aspect relations, *Annals of Glaciology*, **50(53)**, 133-143 **(2009)**

30. Kanungo D.P., Arora M.K., Sarkar S. and Gupta R.P., A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas, *Engineering Geology*, **85(3)**, 347-366 **(2006)**

31. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q. and Liu T.Y., Lightgbm: A highly efficient gradient boosting decision tree, In Advances in Neural Information Processing Systems, 3146-3154 **(2017)**

32. Lagomarsino D., Tofani V., Segoni S., Catani F. and Casagli N., A tool for classification and regression using random forest methodology: applications to landslide susceptibility mapping and soil thickness modeling, *Environmental Modeling and Assessment*, **22(3)**, 201-214 **(2017)**

33. Ma X., Sha J., Wang D., Yu Y., Yang Q. and Niu X., Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, *Electronic Commerce Research and Applications*, **31**, 24-39 **(2018)**

34. Menze B.H., Kelm B.M., Masuch R., Himmelreich U., Bachert P., Petrich W. and Hamprecht F.A., A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, *BMC Bioinformatics*, **10(1)**, 213 **(2009)**

35. Mercer J., Xvi, functions of positive and negative type and their connection the theory of integral equations, Philosophical Transactions of the Royal Society of London, Series A, containing papers of a mathematical or physical character, 415-446 **(1909)**

36. Micheletti N., Foresti L., Kanevski M., Pedrazzini A. and Jaboyedoff M., Landslide susceptibility mapping using adaptive support vector machines and feature selection, Geophysical Research Abstracts, EGU, 13 **(2011)**

37. Moraes R., Valiati J.F. and Neto W.P.G., Document-level sentiment classification: An empirical comparison between SVM and ANN, *Expert Systems with Applications*, **40**, 621-633 **(2013)**

38. Nguyen C., Wang Y. and Nguyen H., Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic, *Journal of Biomedical Science and Engineering*, doi: 10.4236/jbise.2013.65070, **6**, 551-560, **(2013)**

39. O'Brien R.M., A caution regarding rules of thumb for variance inflation factors, *Quality and Quantity*, **41**, 673-690 **(2007)**

40. Pham B.T., Bui D.T. and Prakash I., Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study, *Geotechnical and Geological Engineering*, **35**, 2597-2611 **(2017)**

41. Pradhan B., Laser scanning applications in landslide assessment, Springer **(2017)**

42. Pradhan B. and Lee S., Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling, *Environmental Modelling and Software*, **25**, 747-759 **(2010)**

43. Segoni S., Pappafico G., Luti T. and Catani F., Landslide susceptibility assessment in complex geological settings: Sensitivity to geological information and insights on its parameterization, Landslides, 1-11 **(2020)**

44. Shao X. et al, Planet image-based inventorying and machine learning-based susceptibility mapping for the landslides triggered by the 2018 Mw6. 6 Tomakomai, Japan Earthquake, *Remote Sensing*, **11**, 978 **(2019)**

45. Vapnik V., The support vector method of function estimation, In Nonlinear Modeling, Springer, Boston, MA, 55-85 **(1998)**

46. Wang K., Zuo P., Liu Y., Zhang M., Zhao X., Xie S., Zhang H., Chen X. and Liu C., Clinical and laboratory predictors of in-hospital mortality in patients with COVID-19: a cohort study in Wuhan, China. Clinical Infectious Diseases **(2020)**

47. Yao X., Tham L. and Dai F., Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China, *Geomorphology*, **101**, 572-582 **(2008)**

48. Zhou Z.H., Ensemble Learning, *Encyclopedia of Biometrics*, **1**, 270-273 **(2009)**

49. Zhou Z.H., Machine Learning, Tsinghua University Press **(2016)**.

(Received 14th April 2021, accepted 18th June 2021)