A Novel Hybrid Machine Learning-based Model for Rockfall Source Identification in Presence of Other Landslide Types Using LiDAR and GIS

Ali Mutar Fanos [1], Biswajeet Pradhan [2] *

[1] Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400, Serdang, Selangor Darul Ehsan, Malaysia; engalim87@gmail.com (A.M.F.); shattri@upm.edu.my (S.M).

[2] The Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, NSW, Australia; Biswajeet.Pradhan@uts.edu.au (B.P.)

Correspondence: Biswajeet.Pradhan@uts.edu.au Tel.: +61 2 9514 7937

Abstract

Rockfall is a common phenomenon in mountainous and hilly areas worldwide, including Malaysia. Rockfall source identification is a challenging task in rockfall hazard assessment. The difficulty rise when the area of interest has other landslide types with nearly similar controlling factors. Therefore, this research presented and assessed a hybrid model for rockfall source identification based on the best tested stacking ensemble model of random forest (RF), artificial neural network, Naive Bayes (NB) and logistic regression in addition to Gaussian mixture model (GMM) using high-resolution airborne laser scanning data. GMM was adopted to automatically compute the thresholds of slope angle for various landslide types. Chi-square was utilised to rank and select the conditioning factors for each landslide type. The best fit ensemble model (RF–NB) was then used to produce probability maps, which were used to conduct rockfall source identification in combination with the reclassified slope raster based on the thresholds obtained by the GMM. In the meantime, landslide potential area was structured to reduce the sensitivity and noise of the model to the variations in different conditioning factors for improving its computation performance. The accuracy assessment of the developed model indicates that the model can efficiently identify probable rockfall sources with receiver operating characteristic curve accuracies of 0.945 and 0.923 on validation and training datasets, respectively. In general, the proposed hybrid model is an effective model for rockfall source identification in the presence of other landslide types with a reasonable generalisation performance.

Keywords: Rockfall; Debris flow; Hybrid model; LiDAR; Gaussian Mixture Model

1 Introduction

Rockfalls are common natural hazard in many places worldwide, including Malaysia with high and steep terrain with presence of discontinuities (Simon et al. 2015). This phenomenon affects transportation ways, communication and urban areas that are situated near steep mountainous and hilly areas. The hazard of rockfall is increasing in mountainous regions due to the growth of population and economic activities (Fanos and Pradhan 2018). Rockfall can be defined as separate boulders released from a cliff with different motion modes: flying, bouncing, rolling, or sliding (Vernes 1984; Pradhan and Fanos 2017a). Such events can cause serious causalities because they are difficult to be predicted and can move rapidly depending on the geometric and geomorphologic characteristics of the moving block.

Considerable research has been performed on rockfall hazard around the world including identification of rockfall source areas (Fanos and Pradhan 2016; Losasso et al. 2017), prediction of rockfall trajectories (Pellicani et al. 2016; Fanos et al. 2016), probability assessment (Gigli et al. 2014), analysis of rockfall runout distance (Fanos et al. 2016), evaluation of rockfall bounce height and velocity (Giacomini et al. 2016) and risk analysis (Mitchell and Hungr 2016; Pradhan and Fanos 2017b).

In particular, identification of rockfall source areas is required in the assessment of rockfall probability and risk because it controls the trajectory of rockfall. Rockfall sources can be identified through in-situ survey or rockfall inventory dataset. Nevertheless, such techniques are costly, time consuming and require experts in this field who are only few in number. In-situ and inventory data are also usually unavailable or incomplete in space and time for several regions (Kromer et al. 2017). The availability of geographic information system (GIS) data and accurate 3D surface models has enabled the development of many approaches for rockfall source identification (Loye et al. 2009; Lan et al. 2010; Massey et al. 2014). Existing methods rely on the identification of slope angle threshold angles that are considered unstable. For example, a threshold of >49° was used by Lopez-Saez et al. (Lopez-Saez et al. 2016), whereas >60° was utilised in (Corona et al. 2013). Moreover, recently developed approaches rely on slope geometry derived from LiDAR point cloud and other conditioning elements, such as slope, aspect, curvature, block type and landuse, by using statistical, probabilistic and machine learning methods (Guzzetti et al. 2003). In Dickson et al. (2016), identification of unstable rocks was conducted using photogrammetric survey in composite construction regions. Many controlling factors of rockfall movement along slope were assessed by Agliardi et al. (2016). The results showed that rockfall source areas cannot be easily identified because they are controlled by different factors. More recently, Mote et al. (2019) proposed a method for rockfall risk assessment through the characterization of rockfall source areas. They considered the continuous cliff bands with slope steeper than 45° as rockfall source areas. Their result shows that rockfall sources are key element in rockfall risk assessment and designing a mitigation process. However, such method is critical to obtain a realistic result as it is restricted to cliff face and rockfall source areas are controlled by additional conditioning factors.

Landslides probabilities are controlled by various conditioning factors including morphological, hydrological, geological, and anthropogenic factors. However, each factor has different relative significance to landslide probability and considering a big number of conditioning factors could lead to a negative impact on landslide probability modelling thus producing an unrealistic result. On the other hand, structural and geotechnical, such the bedrock setting, the spatial frequency of discontinuities (fractures, cracks, and joints), the spatial orientation of the families of discontinuities also influence the landslide probability mapping. However, such information demands an extensive field geomechanical surveys which are costly and time consuming. In addition, such in-situ surveys are hard to be performed in regional scale study (wide area). This study focus on using LiDAR-based landslide conditioning to examine the performance of laser scanning data for landslide probability as alternative of structural and geotechnical factors.

Machine learning techniques, which have become common approaches for modeling landslide susceptibility over large regions. The basic assumption of the empirical approach is that future landslides are likely to occur in similar conditions of the past (Fanos and Pradhan 2016). Algorithms, such as random forest (RF) (Youssef et al. 2016; Chen et al. 2018), artificial neural network (ANN) (Pradhan et al. 2014; Truong et al. 2018), Naive Bayes (NB) (Pradhan et al. 2014; Pham et al. 2016) and logistic regression (LR) (Bui et al. 2016; Lombardo and Mai 2018) have been widely employed for landslide probability modelling. On the other hand, ensemble methods have been quite exercised in other fields, nevertheless, the application of these techniques in the assessment of rockfall issues is still rare (Truong et al. 2018). However, the use of ensemble models can improve the result of landslide probability mapping (Evans and Hudak 2007; Chen et al. 2018).

Kinta Valley is one of the main districts in Malaysia. The bedrock geology for Kinta Valley and surrounding areas are granitic hills, limestone bedrock, and mine. As a result, a lot of engineering geologic issues have been encountered Kinta Valley and its immediate surroundings, involving rockfalls, debris flow, and shallow landslides. The bedrock of limestone in Kinta Valley rises over the alluvial plains forming limestone hills with vertical to sub- vertical slopes (Simon et al. 2015).

The aforementioned studies have made remarkable attempts to propose approaches that can precisely allocate rockfall sources by photogrammetry or with LiDAR data. However, one issue still not considered which is where the analysis area includes other landslide types with nearly the same controlling conditioning factors such as shallow landslide, rockfall and debris flow. Although Fanos et al (2018) tried to identify rockfall source areas using an individual machine learning algorithm. Whereas, ensemble models can produce better accuracy. The optimization of the model hyper-parameters was nor performed. In addition, the slope thresholds were determined based on the inventory dataset not on the morphological units of the slope. Therefore, the current research proposes a hybrid model designed for rockfall source identification based on LiDAR dataset in such conditions (the presence of other landslide types). The proposed model uses three algorithms, namely, Gaussian mixture model (GMM) and stacking random forest (RF) coupled with Naive Bayes (NB) (RF–NB). Kinta Valley encountered several landslide incidents including roclfall, shallow landslide, and debris flow. Thus it was selected to evaluate the proposed hybrid model.

2. The Characteristics of the Study Area

The study area is located at Kinta Valley in the West of Malaysia (Figure 1), which is situated approximately 200 km north of the capital city, Kuala Lumpur. The study area is located approximately between the northeast corner (101°5'30" E, 4°34'50" N) and the southwest corner (101°10'45" E, 4°30'40" N). The study area consists of various landuse features, such as urban, grassland, peat swamp forest, oil palm forest and shrub. The extension of the study area is (5 * 5 km) with landslide density of (2.28 event/km^2).

The humidity at the study area is relatively high (approximately 82.3%) throughout the year, and the temperature lies between 23 °C and 33 °C (The Meteorological Service Department of Malaysia). The average annual rainfall in Kinta Valley is 323 mm.

The geological setting of the Kinta Valley is completely varied with a high percentage of igneous rocks. However, sedimentary (limestone) and metamorphic rocks (marble) are profusely present in the district. However, the selected area contains only limestone. Several faults exist in the study area (Pham et al. 2016). limestone hills prone to landslides incidents because of the presence of extensive fractures and joints that can be easily triggered by various factors, such as water saturation. The faults can also increase the potential of landslides occurrences as they triggering earthquakes. Consequently, Kinta Valley has encountered many landslide events including shallow landslide, rockfall and debris flow.

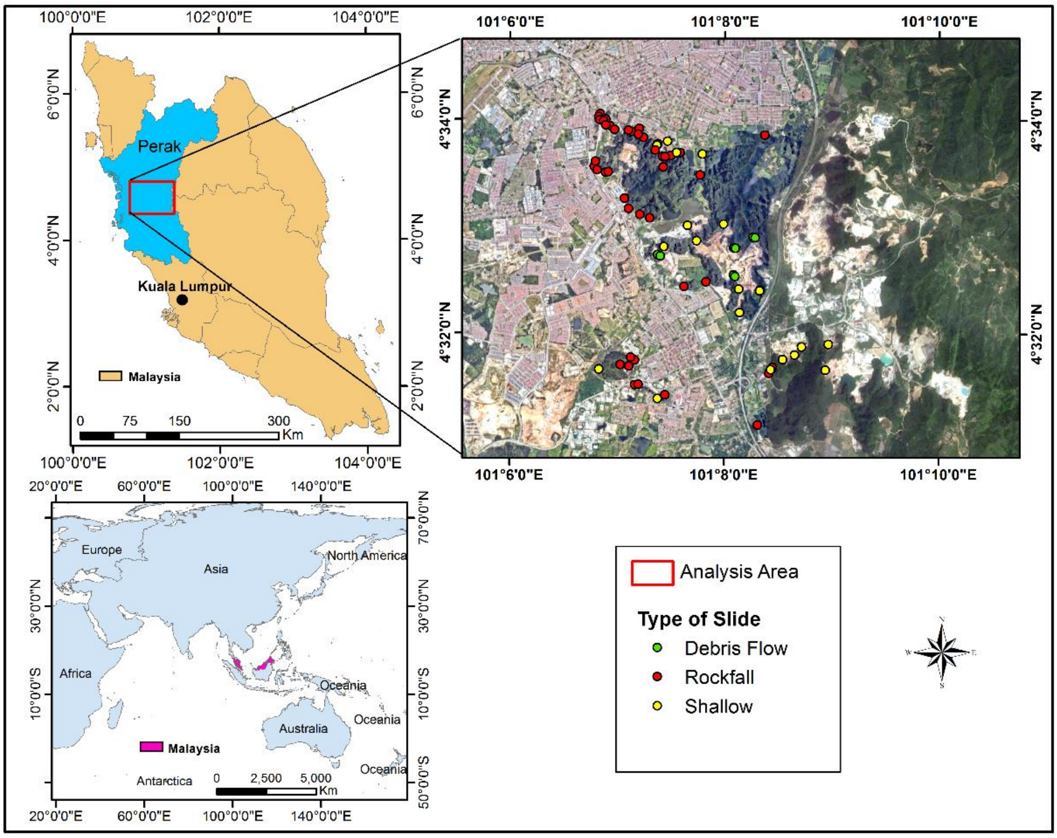


Fig. 1 Study area, Kinta Valley, Malaysia

3. Materials and Methods

3.1 The Used Datasets

The main dataset of this research contained laser scanning data. High-resolution LiDAR point clouds were gathered using an airborne LiDAR system (RIEGL) in 2015 with a flight height of 1000 m. Consequently, high-density point clouds were produced with around 10 pts./m^2. The collected dataset was processed through GIS to perform filtering and interpolation processes. Processing must be applied to the gathered point clouds to eliminate noises and outliers and produce a precise DTM for extracting the conditioning factors of rockfall.

The inventory dataset of landslides is a fundamental element in the assessment of rockfall source areas. This dataset was prepared from different sources including field surveys, remote sensing and historical records. High-resolution aerial photos (0.1 m) that captured during the collection of LiDAR data were utilised for the optical observation of previous landslide events in the study area. Field measurements were also performed using a GNSS system to gather the locations of landslides that occurred underneath vegetated areas or in regions invisible in the aerial photos. This process was conducted using a Global Navigation Satellite System with real-time corrections. Consequently, 87 landslides (28 shallow landslides, 39 rockfall and 20 debris flow), as well as their correlated attributes, were obtained for the assessment (Figure 1). The inventory dataset was divided into two groups (training and testing) to assess the accuracy of the proposed hybrid model. Thus, 70% of the inventory dataset was used to build the model, and the remaining data (30%) were used for validation. The dataset was divided into two group randomly insuring the distribution of each group on the whole study area and each group contains all landslide types.

3.2 Deriving of DTM

The collected raw data contained ground and up-ground points. Therefore, a filtering algorithm must be used to eliminate the up-ground points for obtaining an accurate DTM. The LiDAR-based DTM should be constructed accurately to extract accurate conditioning factors (Chen et al. 2017). Several approaches have been proposed to perform this process. The current study used an algorithm proposed by (Messenzehl et al. 2017) called multi-scale curvature algorithm (MCC) executed within GIS environment. This algorithm can derive an accurate DTM in urban areas with different natural and man-made features (Pham et al. 2017). The terrain details (sharply cut terrains) are essential to rockfall source identification; thus, the window size number should be selected carefully to retain these details (Brenning 2005). Therefore, a particular algorithm was developed to automatically update the number of window sizes for maintaining the details of terrain.

The optimal settings of MCC parameters rely on many elements, such as point cloud density, terrain characteristic and the slope interpolation resolution (Chen et al. 2017). Consequently, the MCC parameters of curvature tolerance threshold, scale domain number and convergence threshold were set to 0.3, 3 and 0.1, respectively. After the up-ground points were eliminated through filtering, the inverse weighted distance interpolation technique was used to generate the DTM from the remaining points. Given that the spacing of points was 0.4 m, the DTM was generated with a resolution of 0.5 m. The statistical analysis of the collected point clouds based on root mean square error revealed vertical and horizontal accuracies of 0.15 and 0.3 m, respectively.

3.3 Preparing of Landslide Conditioning Factors

The source areas of rockfall cannot be assessed on the basis of a certain factor (Agliardi et al. 2016). Thus, the present research used many conditioning factors such as hydrological, morphological, soil and anthropogenic factors to identify the rockfall sources in Kinta Valley.

Many factors were extracted from LiDAR dataset, aerial photos and the databases of government agencies.

Morphological factors (altitude, slope, aspect and curvature) were extracted from the produced 0.5 m DTM and GIS spatial analysis tools. The highest altitude in the current research was 375 m, whereas the lowest altitude was 37 m (Figure 2a). Slope, which is a major factor that controls rockfall, was utilised (Figure 2b). The aspect ratios were from 0° to 360°, which represent the direction of slope from the north in a clockwise direction (Figure 2c). The second derivative of the DTM was used to calculate the curvature factor (Figure 2d). The curvature controls the flow divergence and convergence across a terrain and the deceleration and acceleration of downslope flows. Therefore, this factor affects deposition and erosion.

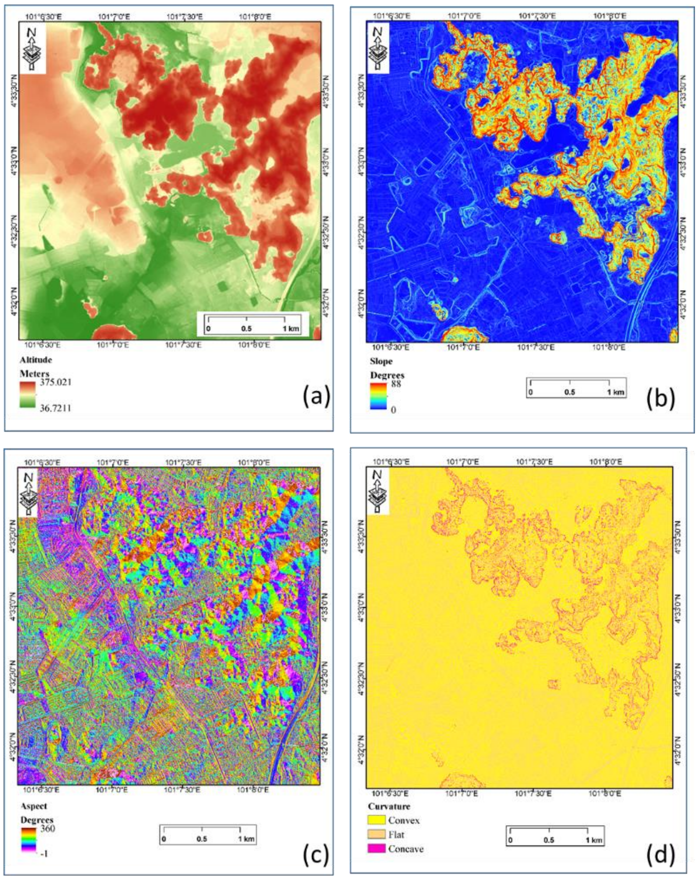
The topographic roughness index (TRI) is a key hydrological factor that affects landslides (Figure 2e). This factor can be calculated using Equation 1:

$$\mathrm{TRI} = \sqrt{\max^2 - \min^2}\,, \qquad (1)$$

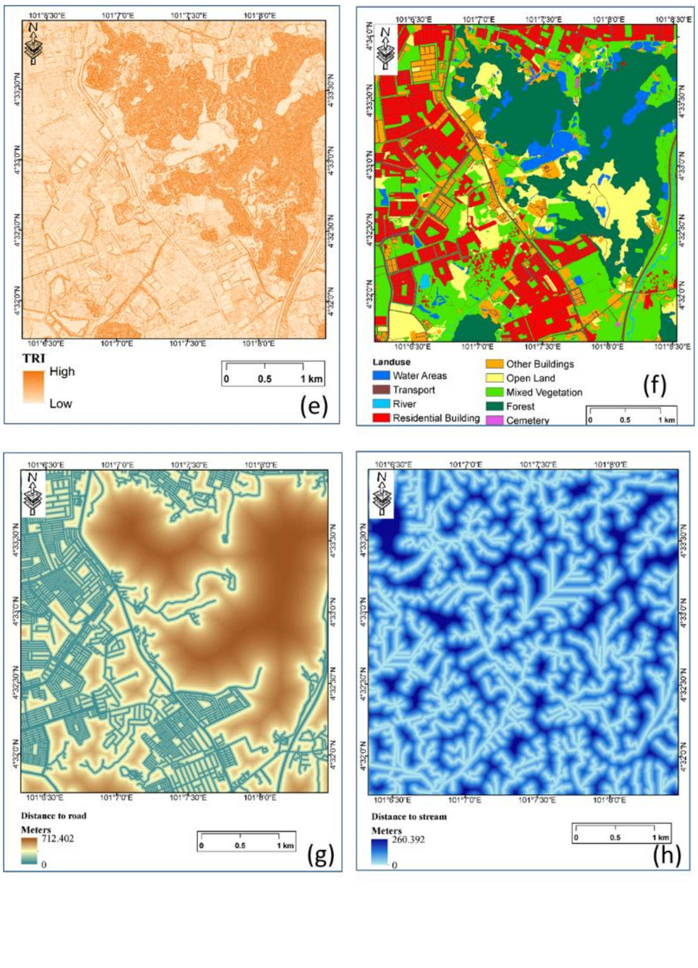
where max is the highest cell value in the nine rectangular neighbourhoods of altitude and min is the minimum value.

In the meantime, anthropogenic factors involve landuse/land cover (LULC) and distances to road. Other factors such as distances to stream (derived from a topographic layer) and lineament (derived from an existing map) were also considered in this study. Geological factor is not considered in this research because of the selected study area contains only one type (limestone). Thus, this factor has no impact on landslide probability mapping. In addition, the focus of the current research is on examining the performance of LiDAR deriving landslide conditioning factors. This can increase the generalization of the proposed methods and reduce the model sensitivity to the variation on the conditioning factors. The LULC layer was produced using classified SPOT 5 satellite images with supervised SVM approach (Department of Survey and Mapping Malaysia**)**. Field survey was performed to verify the LULC layer. The landuse map was classified into nine classes: water body, river, transportation, residential building, other buildings, cemetery, forest, mixed vegetation and open land (Figure 2f). Euclidean distance method was used to calculate the distances to road (Figure 2g), river (Figure 2h) and lineament (Figure 2i).

Sparsely vegetated areas are more prone to landslide incidents than forests. In the current research, vegetation density was utilised as one of the factors for the rockfall source identification. This factor was derived from SPOT 5 satellite images. Four classes were produced: dense vegetation, moderate vegetation, low vegetation and non-vegetation (Figure 2j). Overall, 10 conditioning factors were included in the modelling of rockfall source area identification. Soil texture (Figure 2k) consists of three different types (rocky loam, silt/clay, and loam). This factor is also considered in this research.



Altitude
Meters
375.021
36.7211
(a)
Slope
Degrees
88
0
(b)
Aspect
Degrees
360
-1
(c)
Curvature
Convex
Flat
Concave
(d)
0 0.5 1 km



TRI
High
Low
(e)
Landuse
Water Areas
Transport
River
Residential Building
Other Buildings
Open Land
Mixed Vegetation
Forest
Cemetery
(f)
Distance to road
Meters
712.402
0
(g)
Distance to stream
Meters
260.392
0
(h)
0 0.5 1 km

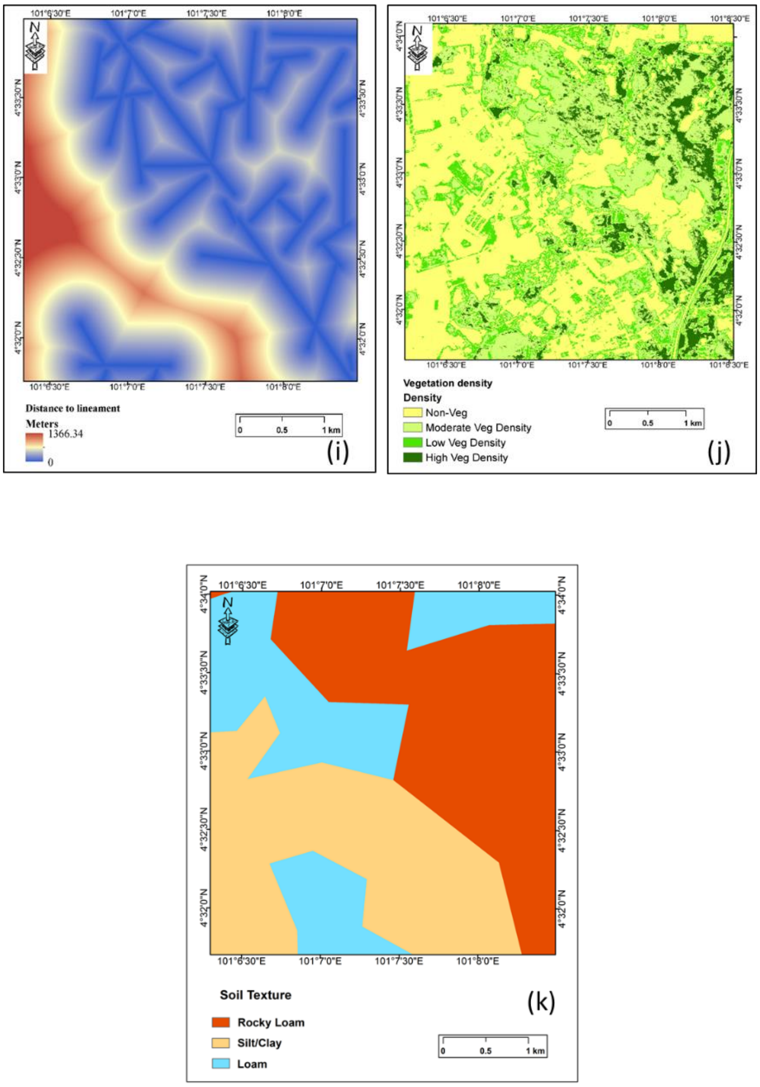


Fig. 2 Landslide conditioning factors

3.4 The Developing of the Proposed Hybrid Model

This research presents a hybrid model based on two algorithms, namely, ensemble stacking (RF–NB) and GMM, which involved many processing steps, as shown in Figure 3. The major datasets used in this research were landslide inventory map, GIS layers and a DTM derived from airborne LiDAR point clouds. The landslide inventory dataset was utilised to train various ensemble machine learning models and validate the hybrid model. GIS layers including LULC, vegetation density, soil texture, lineament, river and road were adopted to obtain the remaining conditioning factors. The high-resolution DTM was produced using LiDAR point clouds for extracting many factors such as slope, aspect, altitude curvature and TRI.

The first processing step is to determine the slope angle threshold of each landslide type automatically based on slope geomorphological units. GMM was run using the slope data that derived from the generated DTM to identify these thresholds. The second step is to determine the best conditioning factors that can identify variance landslide types, including rockfall. This process is performed using Chi-square model as a factor optimisation approach. Consequently, the relevant factors of each landslide type are determined. This process aims to reduce the number of factors for decreasing the time of computation and improving the generalisation capability of the proposed model. The use of only the best factors enables to improve the performance by eliminating redundant and noise information. Thereafter, stacking (RF–NB) model is trained with the inventory data and the selected factors. The stacking (RF–NB) model predicts the landslide probabilities in consideration of the landslide types in the study area. On the other hand, the landslide potential area was constructed. Consequently, a binary raster is generated to reflect the regions that are probable (class 1) and not probable (class 2) to encounter rockfall. This raster is produced through integrating two reclassified elements: slope and landuse. Considering that the study area has encountered many landslides types, the thresholds of slope angle obtained through GMM are used to reclassify the slope raster. The slope raster is reclassified accordingly after the thresholds are estimated automatically. In the meantime, the landuse raster is classified into two classes by integrating water bodies, stream, cemetery, residential building, transportation and other buildings in one class, and the other class contains the remaining classes (forest, vegetated area and open land). The two reclassified elements are integrated to produce the landslide potential area. This process is advantageous because it reduces the sensitivity of the model to the spatial variance in conditioning factors of landslides. In addition, it allows to filter-off the regions with no possibility of landslide. After the thresholds of slope angles are estimated by the GMM method and the likelihood landslide occurrence, the probable source regions can be identified through geoprocessing steps in ArcGIS. Lastly, the remaining data in the inventory dataset are used to validate the obtained results for demonstrating the performance of the proposed ensemble model. The stacking ensemble models were implemented using Python, whereas the GMM was run using Matlab R2016b. The proposed hybrid model was performed in ArcGIS 10.5 environment.

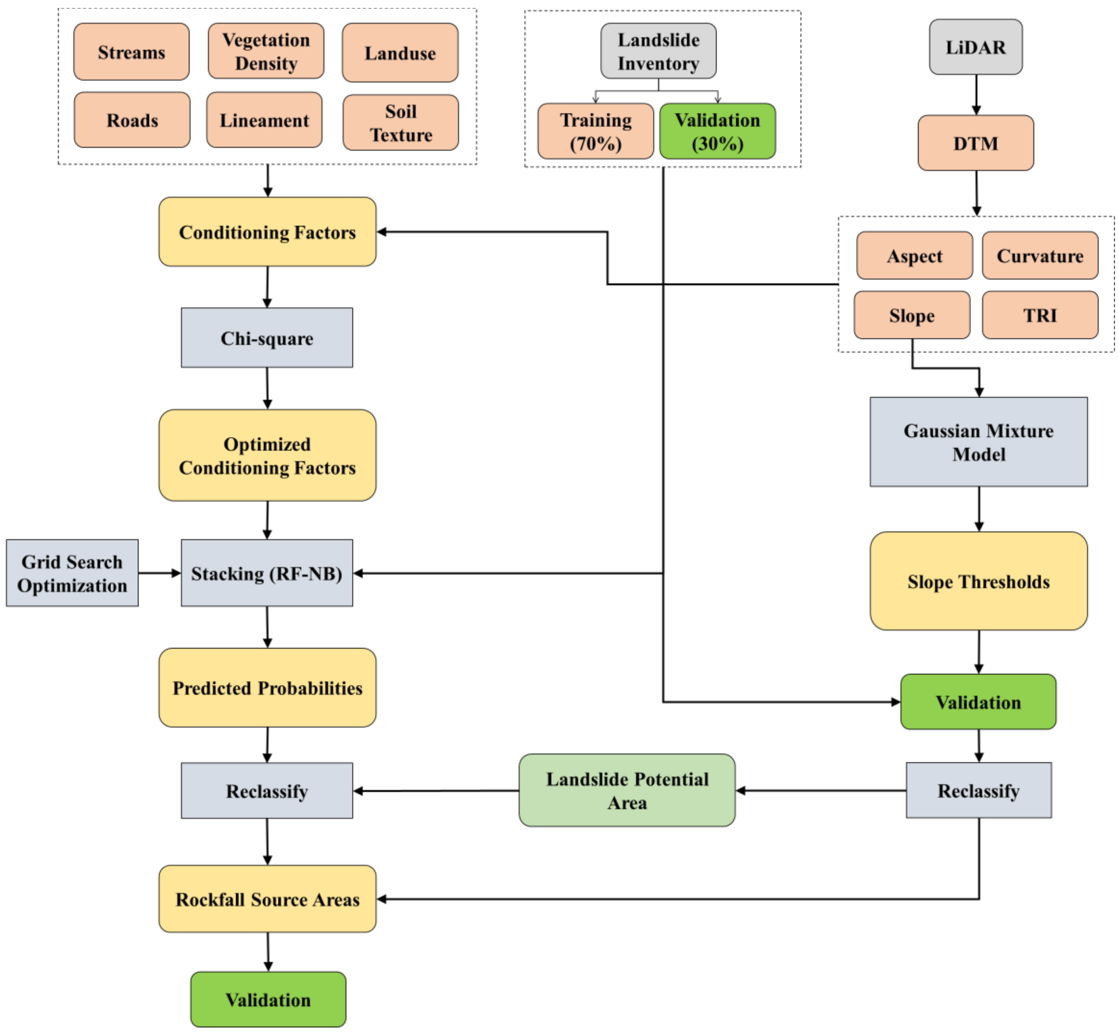


Fig. 3 Flowchart of the proposed hybrid model

3.5 Determination of slope thresholds

The distribution of slope angle can be represented in many Gaussian distributions that can reflect the morphological characteristics, such as rock cliff, steep slope, moderate steep, foot slope and plain. A slope is rated as a probable rockfall source area where the slope angle lies over a particular threshold of slope angle, which can be defined through the Gaussian distribution of the morphological unit (rock cliff becomes predominant over the steep slope). GMM comprises k multivariate components normally used as a parametric model for the distributions of landslide probability given by the following equation (Tien Bui et al. 2018):

$$p(x|\lambda) = \sum_{i=1}^{k} w_i \, g(x|\mu_i, \sum\nolimits_i), \qquad (2)$$

where x is d-dimensional features, $w_i, i = 1, \dots, k,$ are the mixture weights and $g(x|\mu_i, \sum_i), i = 1, \dots, k,$ are the component Gaussian densities. Each component density is a d-variate Gaussian function of the form

$$g(x|\mu_i, \sum\nolimits_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x - \mu_i)' \sum\nolimits_i^{-1} (x - \mu_i)\right\}, \qquad (3)$$

with mean vector μ_i and covariance matrix $\sum_i$. The mixture weights satisfy the constraint that $\sum_{i=1}^{k} w_i = 1$.

The GMM parameters were computed on the basis of the training dataset by using the iterative expectation–maximisation algorithm.

3.6 Ensemble Machine Learning Models

Machine learning algorithms provide better results for landslide identification than other probabilistic methods. In the last decades, machine learning algorithms have been used effectively in identifying probable landslide areas (Brenning 2005; Evans and Hudak 2007;

Scrucca et al. 2016). Methods, such as RF (Trigila et al. 2003; Chen et al. 2018; Segoni et al. 2018; Fanos et al. 2018), logistic regression (LR) (Catani et al. 2005; Pradhan et al. 2014; Bui et al. 2016), artificial neural network (ANN) (Manzo et al. 2013; Chen et al. 2017; Pham et al. 2017) and NB (Chen et al. 2017, Lombardo and Mai 2018), are popular and widely applied machine learning algorithms for landslide probability and produce high accuracy. However, existing methods for the modelling of landslide probability prove that the forecasting of landslide probability can be improved using hybrid machine learning algorithms (Fanos et al. 2018). Thus, new hybrid machine learning models for landslide probability should be developed.

The current research partially fills this gap in literature through proposing a new hybrid machine learning model for the probability modelling of different landslide types. Stacking is a machine learning ensemble approach. Contrary to other ensemble models, stacking can create a strong learner from weaker ones with better tuning in the search for landslide probability modelling processes. In comparison with other ensemble models, stacking also requires lesser running time and computational resources for training, optimisation and validation (Alves 2017). In this research, different stacking models, namely, (RF–ANN), (RF–NB), (RF–LR), (ANN–NB), (ANN–LR) and (NB–LR), were optimised and trained on the basis of the inventory data and the obtained conditioning factors. The hyperparameters of the used machine algorithms were firstly optimised using the grid search optimisation approach (Kotthoff et al. 2017). Then, the best fit stacking ensemble model (RF–NB) was utilised to derive the probability maps of different landslide types. The model was run with 174 samples of the inventory dataset (87 landslides and 87 non-landslides).

4 Results and Discussion

4.1 Slope thresholds

The slope angles distribution of various landslide types are presented in Figure 4 based on the inventory data. Various landslide types had occurred at various slope angles, which indicates the potential to identify and recognise the source areas of these types through the GMM. The figure also demonstrates that rockfall incidents had occurred at the highest slope angle range (45–75°). Shallow landslide incidents had occurred within the slope angle in the range from 23° to 43°. By contrast, debris flows had occurred at the lowest slope angle range (15°–25°). The thresholds of slope angle depend on the variation in slope angle distribution in a particular region. Thus, the GMM was used to evaluate the ability of determining the thresholds, and the slope angles were fine tuned in an unsupervised way via the GMM algorithm. Consequently, rockfall could be distinguished from other landslide types automatically on the basis of the slope angles.

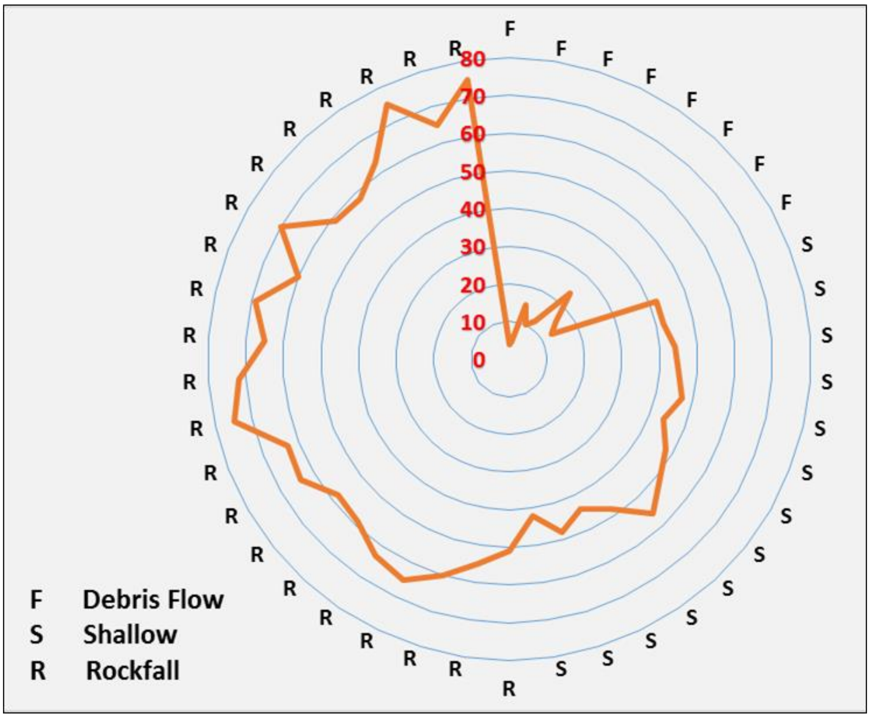


Fig. 4 Distribution of slope angle for various landslide types in the training dataset

The thresholds of slope angles derived via the GMM is illustrated in Figure 5. They included five components determined on the basis of the geometric unit of slope terrain. Thresholds were calculatedwithout the label (landslide type). In other words, it is unsupervised process. The mean values (μ_i) of the five components were obtained as follows: 1.46°, 6.23°, 16.43°, 43.21°, 66.31° and 47.22°. Thereafter, the normal values were defined depending on the μ_i values in consideration of the standard deviation and mean values of the dataset. This way could determine the efficient thresholds of slope angles. After the slope angles were plotted against the normal values, the effective thresholds of slope angles could be identified through the intersection of curves (slope terrain type), as illustrated in Figure 5. For example, the efficient threshold for debris flow was specified through intersecting the curves of foot slopes with moderate slopes and moderate slopes with steep slopes. This procedure resulted in an effective slope angle in the range from 9° to 23°. For shallow landslide, the effective slope angle threshold was determined by intersecting the curves of moderate slopes with steep slopes and steep slopes with cliffs. Consequently, the effective slope angle threshold ranged from 23° to 57°. By contrast, the efficient threshold of rockfall was identified via intersecting of steep slopes with cliffs and above. Therefore, the final threshold was chosen as > 57°.

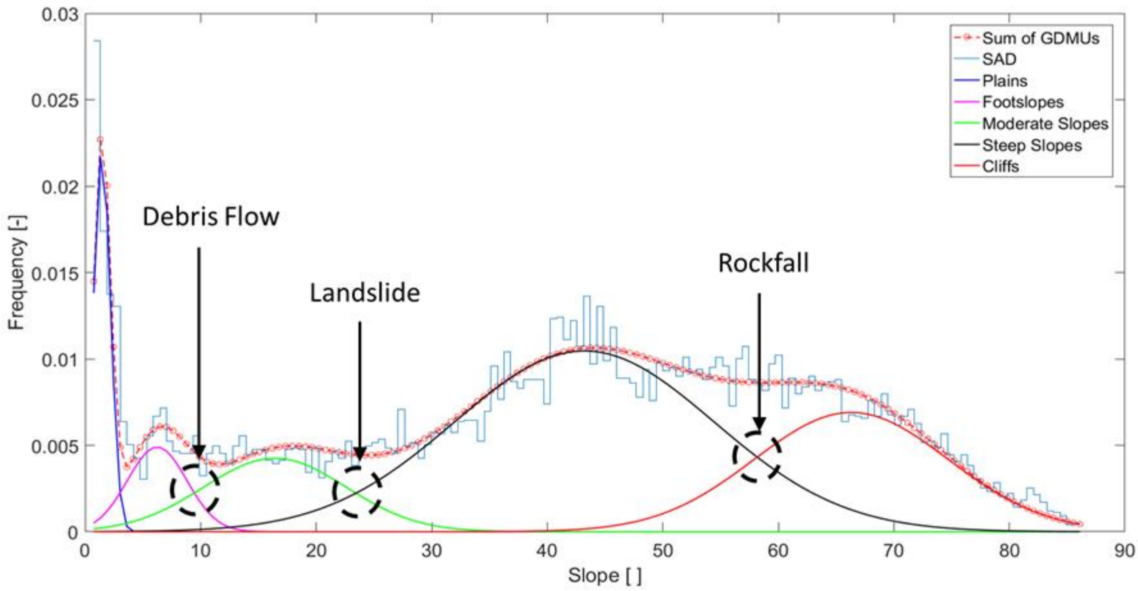


Fig. 5 Effective thresholds of slope angles determined through GMM

4.2 Results of Factors Optimisation

Table 1 shows the estimated ranks of the conditioning factors accounting for the different types of landslides, particularly the key factors ($\alpha < 0.05$) (aspect, slope, curvature, TRI, landuse, distance to lineaments, distance to streams, distance to roads and vegetation density). Chi-square model accuracies (areas under curve (AUC)) are shown with the best conditioning factors. Regarding rockfall, the best five conditioning factors were observed as slope, TRI and distances to lineament, road and stream. However, vegetation density, curvature and aspect were found less significant for the prediction of the rockfall occurrence probabilities in the study area.

Table 1 Factor ranking by Chi-square

Factor	Shallow	Rockfall	Debris Flow	Overall
Aspect	1	9	4	4
Slope	5	1	8	5
Curvature	8	8	7	8
TRI	6	2	3	2
Landuse	4	6	9	7
Distance to lineaments	9	3	6	3
Distance to streams	2	5	5	9
Distance to roads	3	4	1	1
Vegetation density	7	7	2	6
AUC	0.79	0.94	0.88	0.85

4.3 Results of Stacking Ensemble Models

The best conditioning factors were derived for each landslide type in the previous section. Consequently, different stacking ensemble models were developed on the basis of machine learning algorithms (RF, ANN, NB and LR) for the prediction of landslide occurrence probability in the study area. These models were trained with the best conditioning factors and the inventory dataset. The success rate curve (ROC) and the prediction rate curve (PRC) were used to assess the performance of each stacking ensemble model. The best fit stacking ensemble model (RF–NB) was used to derive the probability maps of each landslide type. Figure 6 illustrates the generated probability maps. The probability map is raster with spatial resolution of 0.5 m which is the same resolution of the generated DTM.The probability maps reflect that shallow landslides could occur in the east of the area. However, higher probability was observed in the steep terrain than in low-slope regions. Some portions in the south and northwest could experience shallow landslides. Figure 6a shows the highly susceptible regions for shallow landslides, which are marked in red colour. In the meantime, the northwest and northeast regions were predicted as highly prone to rockfall. The regions of steep cliffs with high slopes had high probability to encounter rockfall (Figure 6b). Furthermore, the middle towards eastern portions of the study area had high probability to encounter debris flow, particularly the areas with the low slope angle of < 23° (Figure 6c).

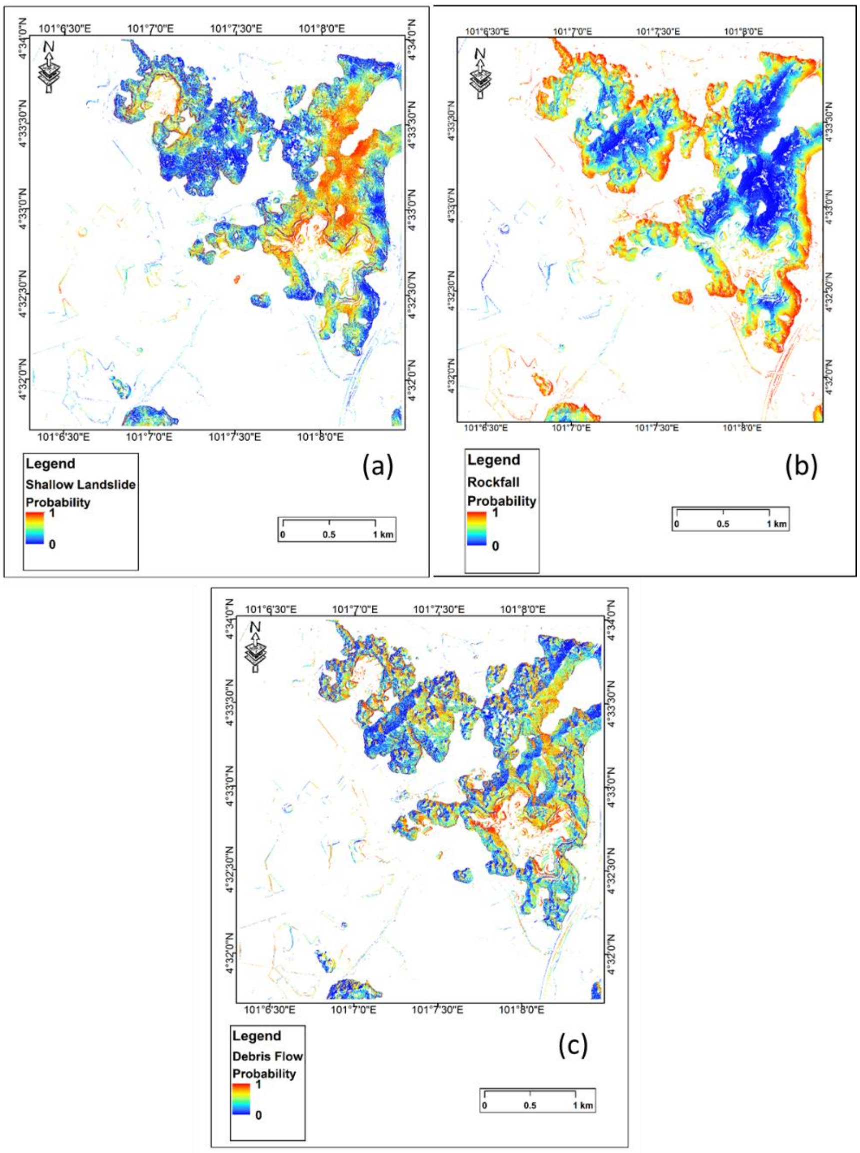


Fig. 6 Probabilities of different landslide types

Thereafter, the slope raster was reclassified using the effective thresholds of slope angles to create the landslide potential area raster. A raster with two classes, namely, high potential and less potential of encountering landslides, was obtained. The raster considered landuse and slope angle. The northeast portion, which has steep slopes, was more prone to landslides than others. In general, 24% of the study area could encounter landslides. The next sections demonstrate the results of the developed model to classify these regions depending on the landslide types and transform the probability raster into source areas by utilising the effective thresholds of slope angles.

4.4 Results of Accuracy Assessment of the Ensembles Models

The proposed ensemble model was validated using Receiver Operating Characteristic (ROC) and precision recall curve (PRC). ROC and PRC explain the known landslide percentage that lay on the rank of the probability level and show the graph of cumulative frequency (Evans and Hudak; Chen et al. 2018). The ROC was produced using the landslide inventory dataset for training, whereas the PRC was produced using the validation landslide dataset. Moreover, the area under curve (AUC) was adopted to assess the accuracy of the tested ensemble models for producing the landslide probability maps; high accuracy is achieved when the area is large (Pradhan et al. 2010; Hong et al. 2015; Wen et al. 2016; Park et al. 2018).

Amongst the tested stacking ensemble models, stacking (RF–NB) was found as a best fit model for producing landslide probabilities (Table 2). The highest ROC was found for rockfall (0.935), followed by that for debris flow (0.881). The highest PRC was obtained for rockfall (0.913), followed by that for debris flow (0.859). The model showed the lowest ROC and PRC of 0.805 and 0.797, respectively, for shallow landslides. In general, the proposed model showed weighted averages of 0.889 and 0.856 for ROC and PRC, respectively. The lowest performance accuracy was obtained from the stacking (NB–LR) model with three landslide types. In addition, the stacking (RF–LR) model also proved to be a good ensemble model for predicting landslide probabilities. However, the proposed stacking (RF–NB) ensemble model could be considered an efficient tool because the accuracy assessment revealed an excellent performance of the proposed model based on the validation and training data. Moreover, the model generalisation was expected to be excellent because the PRC of rockfall was higher than that of ROC accuracy, especially in areas with nearly the same characteristics as the tested area. Nevertheless, the accuracy of model performance is also affected by the number of the landslide inventory samples. A realistic model accuracy and result can be achieved with a big number of inventory samples for training and testing dataset. On the other hand, small number of inventory dataset can lead to unrealistic result even with high accuracy achieved through training process. Therefore, the better accuracy achieved in this study is with rockfall dataset due to the big number of inventory samples in comparison with other landslide types. In addition, the lack of the spatial frequency of discontinuities (fractures, cracks, and joints) did not affect the accuracy of the proposed model as it achieved a high accuracy especially with rockfall.

Table 2 Accuracy assessment of the proposed model

Stacking Model	**Debris Flow**		**Rockfall**		**Shallow Landslide**	
	ROC	**PRC**	**ROC**	**PRC**	**ROC**	**PRC**
RF–ANN	0.820	0.753	0.809	0.785	0.735	0.713
RF–NB	0.881	0.859	0.935	0.913	0.805	0.797
ANN–NB	0.795	0.813	0.754	0.739	0.705	0.689
RF–LR	0.857	0.839	0.874	0.853	0.743	0.755
NB–LR	0.703	0.675	0.734	0.715	0.659	0.627
ANN–LR	0.751	0.719	0.795	0.773	0.685	0.667

4.5 Identification of Rockfall Sources

The estimated landslide probabilities could be transformed into the source regions by using the efficient thresholds of slope angle derived through the GMM. Subsequently, the reclassified slope raster based on the obtained threshold (>57°) was intersected with the rockfall probability raster within GIS environment to create the probable rockfall source regions. Figure 7 shows the predicted areas of potential rockfall. These regions had steep cliff with other analysed elements (slope components). The model prediction accuracy could be evaluated by determining locations of the recorded rockfall incidents. Most of the historical rockfall incidents (91 %) were accurately predicted through the developed hybrid model. The model predicted that 3.5% (around 0.55 km^2) of the area is susceptible to rockfall. The regions that were predicted to be susceptible to rockfall were also investigated through in-situ survey. Many locations were observed to be sensibly predicted as high potential regions to rockfall. These regions were mainly formed by steep cliff surrounded by vegetated areas (Figure 7).

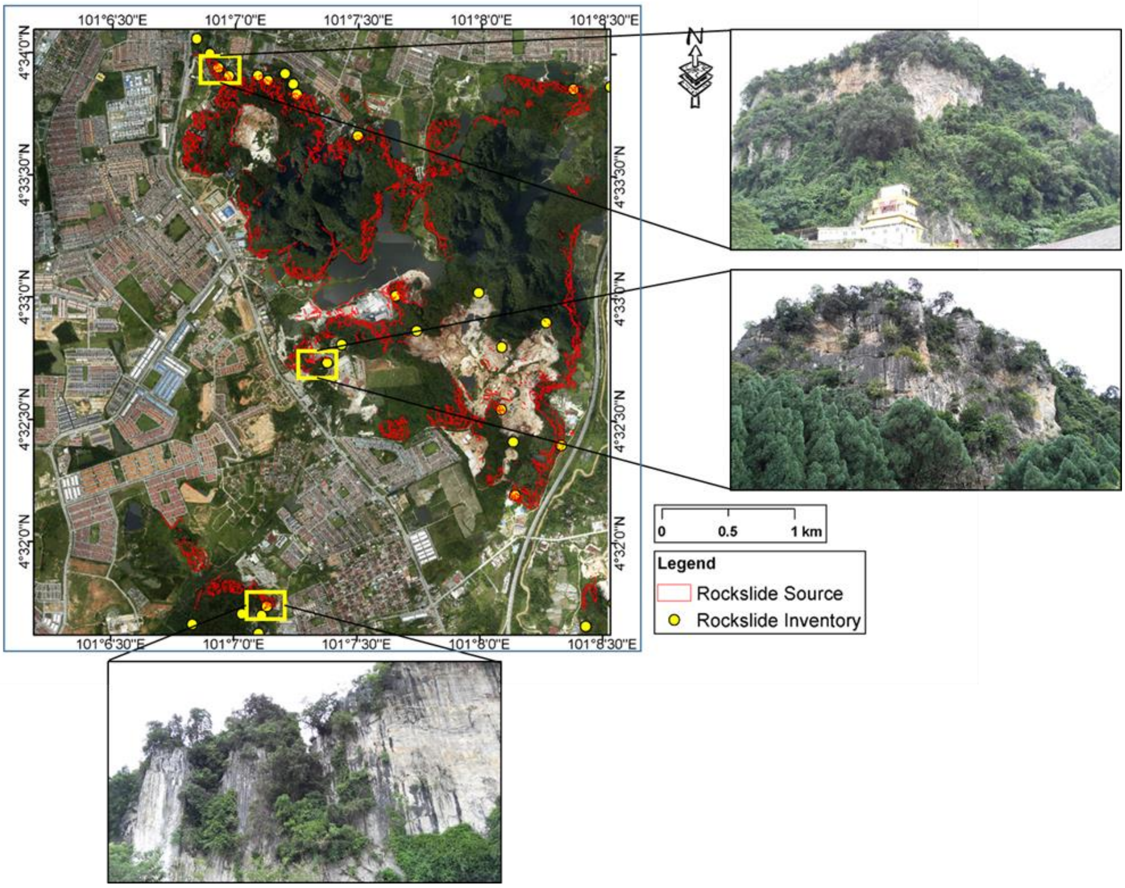


Fig. 7 Identified rockfall source areas using the proposed ensemble model

5 Conclusions

This research developed an ensemble model using two algorithms, namely, GMM and stacking ensemble model based on RF and NB, to identify rockfall source regions in the presence of other landslide types (shallow landslide and debris flow). The GMM model was used to determine the effective thresholds of slope angle for different landslide types and construct the landslide potential area raster. In the meantime, the best landslide conditioning factors were selected through the Chi-square method. Various ensemble models were developed on the basis of different machine learning algorithms (RF, ANN, NB and LR). The best fit ensemble model (stacking RF–NB) was used to produce the probability maps. The binary slope raster created through GMM was intersected with the rockfall probability map.

The developed ensemble model performed well with training and validation regions chosen at Kinta Valley. The model showed accuracies of 0.935 and 0.913 on training and validation datasets. For shallow landslide and debris flow, the proposed ensemble model provided accuracies of 0.805 and 0.881 on the training dataset and 0.797 and 0.859 on the validation dataset. Overall, the proposed ensemble model showed excellent average accuracy on all the landslide types in the inventory dataset. The model achieved weighted average accuracies of 0.889 and 0.856 on the training and validation datasets, respectively. Since the proposed model achieved a good accuracy, it proves that the conditioning factors derived from LiDAR can be used as an alternative of the geomechanical factors, such as discontinuity and fractures.

The major contribution of this study is the development of a hybrid model can predict the probable rockfall source regions accurately in the presence of other landslide types. However, additional assessment can be performed to improve the computing performance and accuracy of the proposed model for predicting a particular landslide type in the existence of other types in complex regions.

Funding
This research is supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS) in the University of Technology Sydney (UTS) under Grants 321740.2232335 and 321740.2232357.

Acknowledgments

The authors acknowledge and appreciate the provision of airborne laser scanning data orthophoto images from airborne laser scanning data (LiDAR) by the Department of Planning.

References

Agliardi F, Riva F, Galletti L, Zanchi A, Crosta GB (2016) Rockfall source characterization at high rock walls in complex geological settings by photogrammetry, structural analysis and DFN techniques. In EGU General Assembly Conference Abstracts 18: 13071

Alves A (2017) Stacking machine learning classifiers to identify Higgs bosons at the LHC. J INSTRUM 12(05): T05005.

Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat. Hazards Earth Syst. Sci. 5(6): 853-862. doi:10.5194/nhess-5-853-2005

Bui DT, Tuan TA, Klempe H, Pradhan, Revhaug I (2016) 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(2): 361-378. https://doi.org/10.1007/s10346-015-0557-6

Catani F, Casagli N, Ermini L, Righini G, Menduni G (2005) Landslide hazard and risk mapping at catchment scale in the Arno River basin. Landslides 2(4): 329-342. https://doi.org/10.1007/s10346-005-0021-0

Chen W, Pourghasemi HR, Zhao Z (2017) A GIS-based comparative study of Dempster-Shafer, logistic regression and artificial neural network models for landslide susceptibility mapping. Geocarto Int 32(4): 367-385. https://doi.org/10.1080/10106049.2016.1140824

Chen W, Yan X, Zhao Z, Hong H, Bui DT, Pradhan B (2019) Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China). B ENG GEOL ENVIRON 78(1): 247-266. https://doi.org/10.1007/s10064-018-1256-z

Chen Z, Gao B, Devereux, B (2017) State-of-the-art: DTM generation using airborne LIDAR data. Sensors 17(1): 150. https://doi.org/10.3390/s17010150

Corona C, Trappmann D, Stoffel M (2013) Parameterization of rockfall source areas and magnitudes with ecological recorders: when disturbances in trees serve the calibration and validation of simulation runs. Geomorphology 202: 33-42. https://doi.org/10.1016/j.geomorph.2013.02.001

Dickson ME, Perry GL (2016) Identifying the controls on coastal cliff landslides using machine-learning approaches. ENVIRON MODELL SOFTW 76: 117-127. https://doi.org/10.1016/j.envsoft.2015.10.029

Dou J, Yamagishi H, Zhu Z, Yunus AP, Chen CW (2018) TXT-tool 1.081-6.1 A Comparative Study of the Binary Logistic Regression (BLR) and Artificial Neural Network (ANN) Models for GIS-Based Spatial Predicting Landslides at a Regional Scale. In Landslide Dynamics: ISDR-ICL Landslide Interactive Teaching Tools (pp. 139-151). Springer, Cham. https://doi.org/10.1007/978-3-319-57774-6_10

Evans JS, Hudak AT (2007) A multiscale curvature algorithm for classifying discrete return LiDAR in forested environments. IEEE Trans Geosci Sens 45(4): 1029-1038. DOI: 10.1109/TGRS.2006.890412

Fanos AM, Pradhan B (2016) Multi-scenario rockfall hazard assessment using LiDAR data and GIS. Geotech. Geol. Eng. 34(5): 1375-1393. https://doi.org/10.1007/s10706-016-0049-z

Fanos AM, Pradhan B (2018) Laser scanning systems and techniques in rockfall source identification and risk assessment: a critical review. Earth Syst Environ 2(2): 163-182

Fanos AM, Pradhan B, Aziz AA, Jebur MN, Park HJ (2016) Assessment of multi-scenario rockfall hazard based on mechanical parameters using high-resolution airborne laser scanning data and GIS in a tropical area. ENVIRON EARTH SCI 75(15): 1129. https://doi.org/10.1007/s12665-016-5936-3

Fanos AM, Pradhan B, Mansor S, Yusoff ZM, bin Abdullah AF (2018) A hybrid model using machine learning methods and GIS for potential rockfall source identification from airborne laser scanning data. Landslides 15(9): 1833-1850. https://doi.org/10.1007/s10346-018-0990-4

Giacomini A, Ferrari F, Thoeni K, Lambert C (2016) A new rapid method for rockfall energies and distances estimation. In EGU General Assembly Conference Abstracts 18: 5323.

Gigli G, Morelli S, Fornera S, Casagli N (2014) Terrestrial laser scanner and geomechanical surveys for the rapid evaluation of rock fall susceptibility scenarios. Landslides 11(1): 1-14. https://doi.org/10.1007/s10346-012-0374-0

Guzzetti F, Reichenbach P, Wieczorek GF (2003) Rockfall hazard and risk assessment in the Yosemite Valley, California, USA. Nat. Hazards Earth Syst. Sci. 3(6): 491-503

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

Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K (2017) Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. J Mach Learn Res. 18(1): 826-830

Kromer R, Lato M, Hutchinson DJ, Gauthier D, Edwards T (2017) Managing rockfall risk through baseline monitoring of precursors using a terrestrial laser scanner. Can. Geotech. J 54(7): 953-967. https://doi.org/10.1139/cgj-2016-0178

Lan H, Martin CD, Zhou C, Lim CH (2010) Rockfall hazard analysis using LiDAR and spatial modeling. Geomorphology 118(1-2): 213-223. https://doi.org/10.1016/j.geomorph.2010.01.002

Lombardo L, Mai PM (2018) Presenting logistic regression-based landslide susceptibility results. ENG GEOL 244: 14-24. https://doi.org/10.1016/j.enggeo.2018.07.019

Lopez-Saez J, Corona C, Eckert N, Stoffel M, Bourrier F, Berger F (2016) Impacts of land-use and land-cover changes on rockfall propagation: Insights from the Grenoble conurbation. Sci Total Environ 547: 345-355. https://doi.org/10.1016/j.scitotenv.2015.12.148

Losasso, L, Jaboyedoff M, Sdao F (2017) Potential rock fall source areas identification and rock fall propagation in the province of Potenza territory using an empirically distributed approach. Landslides 14(5): 1593-1602. https://doi.org/10.1007/s10346-017-0807-x

Loye A, Jaboyedoff M, Pedrazzini A (2009) Identification of potential rockfall source areas at a regional scale using a DEM-based geomorphometric analysis. NAT HAZARD EARTH SYS 9(5): 1643-1653. https://doi.org/10.5194/nhess-9-1643-2009

Manzo G, Tofani V, Segoni S, Battistini A, Catani F (2013) GIS techniques for regional-scale landslide susceptibility assessment: the Sicily (Italy) case study. INT J GEOGR INF SCI 27(7): 1433-1452. https://doi.org/10.1080/13658816.2012.693614

Massey CI, McSaveney MJ, Taig T, Richards L et al (2014). Determining rockfall risk in Christchurch using rockfalls triggered by the 2010–2011 Canterbury earthquake sequence. Earthq Spectra 30(1): 155-181. https://doi.org/10.1193/021413EQS026M

Messenzehl K, Meyer H, Otto JC, Hoffmann T, Dikau R (2017) Regional-scale controls on the spatial activity of rockfalls (Turtmann valley, Swiss Alps)—a multivariate modeling approach. Geomorphology 287: 29-45. https://doi.org/10.1016/j.geomorph.2016.01.008

Mitchell A, Hungr O (2016) Theory and calibration of the Pierre 2 stochastic rock fall dynamics simulation program. Can. Geotech. J 54(1): 18-30. https://doi.org/10.1139/cgj-2016-0039

Mote TI, Skinner MD, Taylor ML, Lyons C (2019) Site-Specific Rockfall Risk Assessments and Rockfall Protection Structure Design Following the 2010/2011 Canterbury Earthquake Sequence. In IAEG/AEG Annual Meeting Proceedings, San Francisco, California, 2018-Volume 5 (pp. 143-152). Springer, Cham.

Olsen M, Wartman J, McAlister M, et al (2015) To fill or not to fill: sensitivity analysis of the influence of resolution and hole filling on point cloud surface modeling and individual rockfall event detection. Remote Sens. 7(9): 12103-12134. doi:10.3390/rs70912103

Park SJ, Lee CW, Lee S, Lee MJ (2018) Landslide susceptibility mapping and comparison using decision tree models: A Case Study of Jumunjin Area, Korea. Remote Sens. 10(10): 1545. https://doi.org/10.3390/rs10101545

Pellicani R, Spilotro G, Van Westen CJ (2016) Rockfall trajectory modeling combined with heuristic analysis for assessing the rockfall hazard along the Maratea SS18 coastal road (Basilicata, Southern Italy). Landslides 13(5): 985-1003. https://doi.org/10.1007/s10346-015-0665-3

Pham, BT, Bui DT, Pourghasemi HR, Indra P, Dholakia MB (2017) Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. THEOR APPL CLIMATOL 128(1-2): 255-273. https://doi.org/10.1007/s00704-015-1702-9

Pham BT, Bui DT, Prakash I, Dholakia, MB (2017) Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena 149: 52-63. https://doi.org/10.1016/j.catena.2016.09.007

Pham BT, Bui D, Prakash I, Dholakia M (2016) Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. J Geomatics 10: 71-79.

Pham BT, Pradhan B, Bui DT, Prakash I, Dholakia MB (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). ENVIRON MODELL SOFTW 84: 240-250. https://doi.org/10.1016/j.envsoft.2016.07.005

Pradhan B, Fanos AM (2017a) Application of LiDAR in rockfall hazard assessment in tropical region. In Laser Scanning Applications in Landslide Assessment (pp. 323-359). Springer, Cham. https://doi.org/10.1007/978-3-319-55342-9_16

Pradhan B, Fanos AM (2017b) Rockfall hazard assessment: an overview. In Laser Scanning Applications in Landslide Assessment (pp. 299-322). Springer, Cham.

Pradhan B, Abokharima MH, Jebur MN, Tehrany MS (2014) Land subsidence susceptibility mapping at Kinta Valley (Malaysia) using the evidential belief function model in GIS. NAT HAZARDS 73(2): 1019-1042. https://doi.org/10.1007/s11069-014-1128-1

Pradhan B, Sezer EA, Gokceoglu C, Buchroithner MF (2010) Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (Cameron Highlands, Malaysia IEEE Trans Geosci Remote Sens 48(12): 4164-4177. DOI: 10.1109/TGRS.2010.2050328

Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal 8(1): 289

Segoni S, Tofani V, Rosi A, Catani F, Casagli N. (2018) Combination of rainfall thresholds and susceptibility maps for dynamic landslide hazard assessment at regional

scale. Front. Earth Sci. 6: 85
Simon N, Ghani MFA, Hussin A et al (2015) Assessment of rockfall potential of limestone hills in the Kinta Valley. JSSM 10(2): 24-34
Tien Bui D, Shahabi H, Shirzadi A et al (2018) A novel integrated approach of relevance vector machine optimized by imperialist competitive algorithm for spatial modeling of shallow landslides. Remote Sens. 10(10): 1538. https://doi.org/10.3390/rs10101538
Trigila A, Frattini P, Casagli N et al (2013) Landslide susceptibility mapping at national scale: the Italian case study. In Landslide science and practice (pp. 287-295). Springer, Berlin, Heidelberg.
Truong X, Mitamura M, Kono Y et al (2018) Enhancing prediction performance of landslide susceptibility model using hybrid machine learning approach of bagging ensemble and logistic model tree. Appl Sci 8(7): 1046. https://doi.org/10.3390/app8071046
Varnes DJ (1984) Landslide hazard zonation: a review of principles and practice. Natural Hazards (No. 3)
Wen Z, He B, Xu D, Feng Q (2016) A method for landslide susceptibility assessment integrating rough set and decision tree: A case study in Beichuan, China. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 4952-4955). IEEE. DOI: 10.1109/IGARSS.2016.7730292
Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM (2016) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides 13(5): 839-856. https://doi.org/10.1007/s10346-015-0614-1