GEOMATICS, NATURAL HAZARDS AND RISK
2021, VOL. 12, NO. 1, 2430–2461
https://doi.org/10.1080/19475705.2021.1960433

Advanced data mining techniques for landslide susceptibility mapping

Muhammad Bello Ibrahim[a,b], Zahiraniza Mustaffa[a], Abdul-Lateef Balogun[a], Indra Sati Hamonangan Harahap[c] and Mudassir Ali Khan[a]

[a]Department of Civil and Environmental Engineering, Universiti Teknologi PETRONAS Persiaran UTP, Seri Iskandar, Perak; [b]Department of Civil Engineering, Hussaini Adamu Federal Polytechnique, Kazaure, Jigawa State, Nigeria; [c]Civil Engineering Department, Universitas Islam, Krawitan, Umbulmartani, Ngemplak, Sleman Regency, Special Region of Yogyakarta, Indonesia

ABSTRACT

This paper describes the development and validation of landslides susceptibility models for mountainous regions using advanced data mining techniques. The investigation was carried out to ascertain the effectiveness of Naïve Bayes Multinomial (NBM) and Random Trees (RT) in landslide susceptibility mapping. The NBM is an advancement of the frequently used Naïve Bayes classifiers, while the RT was built to overcome the limitations of the traditional forest classifiers. A geospatial database for this investigation comprises 148 landslide locations influenced by ten (10) landslide conditioning factors. The factors (Slope Angle, Slopes Elevation, Slope Aspect, Plan curvature, Profile Curvature, Lithology, Soil type, Stream power index (SPI), Sediment transport index (STI), and Rainfall precipitation) were drawn using a Multi Collinearity Decision Making (MCDM) technique. A Frequency Ratio (FR) analysis was used to obtain the relative significance of the factors in the slides. Predictive models were also developed by quantifying these models using data mining techniques. A section of the entire geospatial data (70%) was used as training datasets, while the remaining part of the data (30%) was used to validate the trained datasets. SVM, RT, and NBM algorithms were used to produce predicted datasets from the training datasets. These predicted datasets were used to develop the Landslides Susceptibility Models. A comparative assessment between the two classifiers against the famous traditional learning algorithm, the Support vector machines (SVM), was conducted. Model performance evaluators such as the AUROC, RSME, F-measure, MAE, and ACC were employed to check the predictive capabilities and accuracies of the models. The indices indicated that the SVM model performed better than the other two algorithms in both training and validation datasets. Further analysis and comparison of the models reveal that the new data mining techniques are reliable for landslide susceptibility. Simultaneously, the traditional algorithm is also useful and remains relevant, especially with similar site conditions. This study has provided insights on better

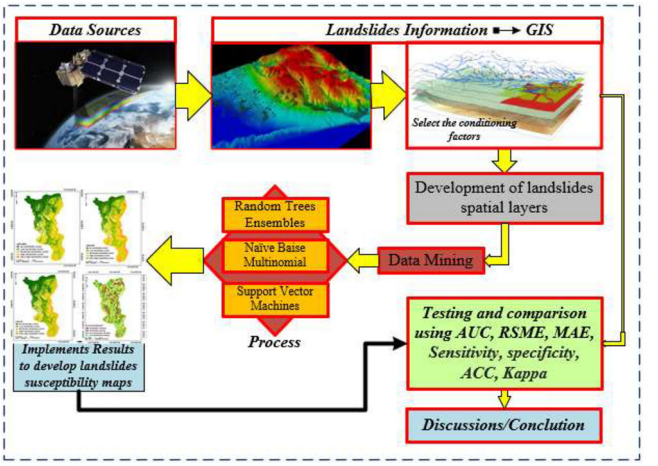
ARTICLE HISTORY
Received 3 March 2021
Accepted 21 July 2021

KEYWORDS
Landslides analysis; susceptibility mapping; soft computing; machine learning and GIS; data mining; conditioning factors; frequency ratio

CONTACT Muhammad Bello Ibrahim muhammad_17006885@utp.edu.my

planning and development and provision of mitigation strategies and further analysis on landslides in the study area, particularly in cases of limited data availability.

GRAPHICAL ABSTRACT



1. Introduction

Landslides are generally a form of geo-hazards because it affects human lives either directly or indirectly (Collins and Znidarcic 2004; Gue and Tan 2006; Ibrahim MB et al. 2020). It occurs when the soil or rock mass is displaced under the influence of gravity. This phenomenon causes great destruction of lives and properties (Figure 1a–f), and it is a cause of concern to many governments. It also posed fear to the people living in places that are susceptible to landslides. Generally, there are at least five primary forms of landslides based on the nature of material displacement. These include rock falls, topples, slides, spreads, and flows (USGS 2004). Landslides are believed to have multiple complex causative factors and triggering factors that are dependent on the nature of the environment. When loss of life is directly involved, then the phenomenon is labelled as hazardous. Landslides affect the environment, such as vegetation and farmlands, in areas that humans are not inhabited. It also affects the infrastructures built across rough terrains such as pipelines, roads, highways, earth dams, retaining walls, housing, small cities. These and many more are endangered by landslides (Bacha et al. 2020; Chang Z et al. 2020; Díaz et al. 2020; Li L et al. 2020; Prakash et al. 2020).

Landslides are now a global problem; according to world bank data, approximately 3.7 million km^2 of the earth's landmass is under serious threat from landslide activities (Mandal et al. 2021). The report also identified some 300 million people living within those high landslide susceptible areas, putting their lives in danger, and destructing massive economic activities within the areas. A recent study conducted by Balogun et al. (2021) reported a total death of 1370 and 784 injuries in 27 European countries alone from 1995 to 2014. In addition, a compensatory cost of about

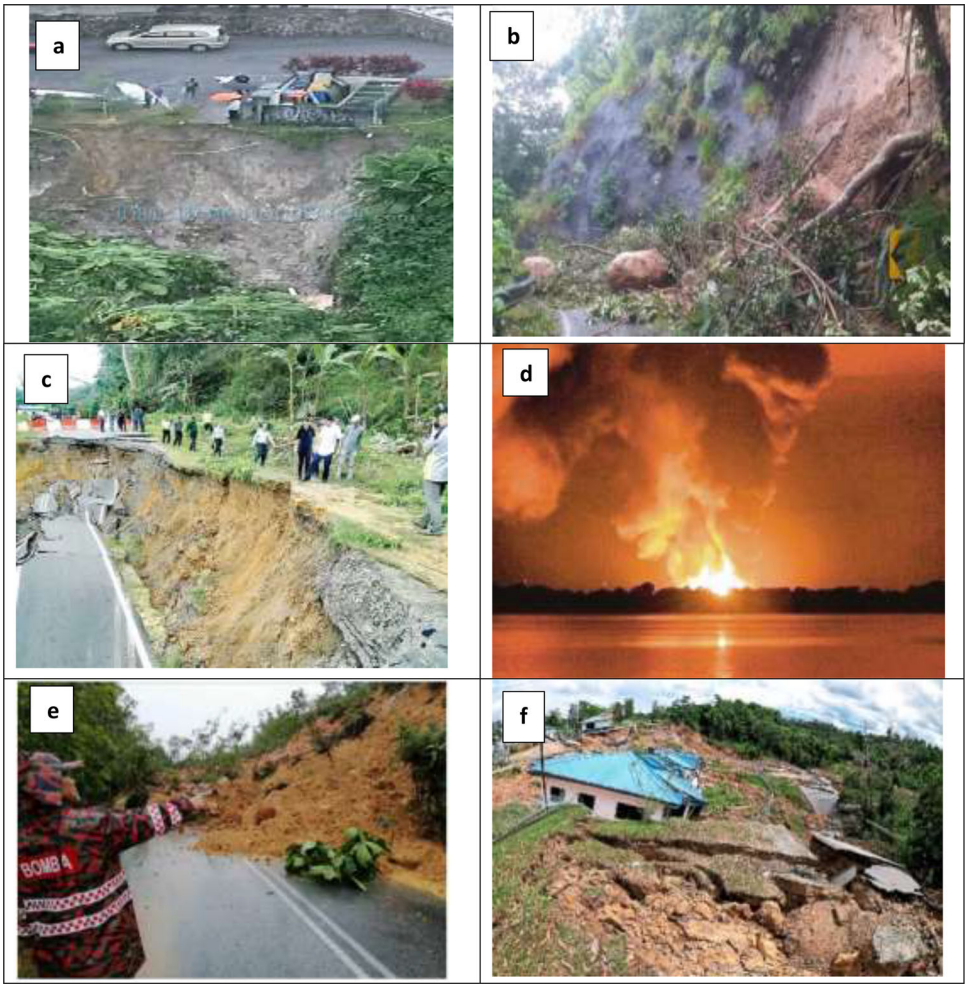


Figure 1. (a) Indicating the event of a landslide within a community in the study area; (b) A road linking two towns is completely cut out; (c) showing the devastating effects of landslides along a road section linking some communities in the state of Sarawak; (d) A gas pipeline in the study area that got ruptured and lit up the whole area, the event was caused by landslides activities triggered by nonstop rain that happened in the area; (e) Part of a road leading from Song bazaar to SMK Song, deep in the remote Kapit Division in Sarawak, has been cut off due to a landslide; (f) The condition of a house after a landslide which occurred near the water treatment plant in Paitan; Bernama pic, January 15, 2021 (internet sources).

4.7 billion euros to property loss. Another 200 million dollars was expended to restore damages from two landslide events (1974 and 1998) in Peru. From 1995 to 2004, the world has recorded some 163,658 deaths with 11,689 injuries to landslides hazards alone. While from 2004 to the year 2016, China alone has spent over one billion dollars on non-seismic landslides. Overall, the figures relating to landslide losses have been rising despite governments' and individuals' efforts to curtail the phenomenon (Remondo et al. 2003; Petley 2012; Pourghasemi and Rahmati 2018; Althuwaynee et al. 2021).

Due to the dynamic nature of landslides and their analysis, there have been improvements in studies that analyze landslides and provide early warnings of their occurrence in susceptible regions. Geotechnical and geological data obtained mainly from laboratory experimentations were used to produce physical models to determine susceptibility in early landslides studies (Pourghasemi and Rahmati 2018). Limitations and challenges encountered while analyzing landslides using the physical model (e.g. time wastage in actual data collection and analysis, huge experimental cost, especially in the analysis of larger areas) led to statistical analysis. The statistical models best correlate the dynamism in landslides between many predisposing factors and landslides (Balogun et al. 2021). To this effect, many statistical models such as frequency ratio (Guzzetti et al. 1999; Gorsevski et al. 2006), fuzzy logic (Akgun et al. 2012; Balal and Cheu 2018; Chen W et al. 2017a; Hauser-Davis et al. 2012; Irvin et al. 1997; Tien Bui et al., 2017a), the weight of evidence models (Chang M et al. 2020; Lee JH et al. 2018; Pamela et al. 2018; Polykretis and Chalkias 2018), logistic regressions (Bai et al. 2010; Chen W et al. 2017b; Chen W et al. 2019; Pradhan 2010), analytical hierarchy process (Althuwaynee et al. 2014; Saadatkhah et al. 2014; Mardani et al. 2015; Asadabadi et al. 2017) and many more have all been utilized to produce susceptibility models. More recently, soft computing procedures (Data mining) are increasingly being adopted for landslide susceptibility analysis due to their impressive performance. This approach combines GIS data and machine learning algorithms to analyze the landslide, predict and produce susceptibility maps statistically (Li DQ et al. 2021; Saravanan et al. 2021).

The technique uses approximate values to produce very accurate and valuable solutions (Tsangaratos and Ilia 2017b; Moayedi et al. 2019; Goel 2020). Overall, the accuracy of the models developed using this technique has shown high prediction performance and high success rates (Ayodele 2010; Oladipupo 2012; Goetz et al. 2015; Dickson and Perry 2016; Shirzadi et al. 2018; Ghorbanzadeh et al. 2019; Hegde and Rokseth 2020). The dynamic nature of landslides with their conditioning and triggering factors across different locations made researchers explore different algorithms to harness the maximum prediction rate from the soft computing techniques (Chen X and Chen W 2021; Diana et al. 2021; Saha et al. 2021; Youssef and Pourghasemi 2021). To date, many researchers are using machine learning algorithms to mine data and make valuable predictions of landslide occurrence effectively. For instance, (Chen W et al. 2019) used kernel logistic regression, naïve Bayes, and radial basis function network to produce landslide susceptibility maps. While (Oh and Lee 2017) utilized artificial neural networks and boosted trees to produce landslide susceptibility maps. Others like (Chen W et al. 2019; Dou et al. 2019; Fallah-Zazuli et al. 2019; Chen 2019; Hong et al. 2016; Lay et al. 2019; Lee S et al. 2017; Nhu et al. 2020a; Song et al. 2012; Tien Bui et al. 2016; Vafakhah et al. 2020; Vakhshoori et al. 2019; Liu Z et al. 2021; Oliva-González et al. 2019; Wang et al. 2021) have used different machine learning algorithms to mine data and produce susceptibility maps.

Studies to assess the comparative performance of various algorithms have revealed that the study environment and conditioning factors under consideration affect the outcome of the analysis (Chen X and Chen W 2021; Hong et al. 2017; Mohammady

et al. 2019; Liang et al. 2020; Sun et al. 2021; Tsangaratos and Ilia 2017a; Xu et al. 2016; Yeshwanth et al. 2019). Different predictive results were recorded for the same machine learning algorithms used in different locations. For instance, (Merghadi et al. 2020) reported that the RF algorithm outperformed the conventional SVM. Even though both algorithms performed well by surpassing the 0.7 success rate benchmark. Tien et al. (2020) suggested that Deep learning neural networks perform better than the conventional learning algorithms followed by the SVM and then the RF models. Contrary to the previous investigation (Achour and Pourghasemi 2020), the RF model has performed better than both SVM and Boosted Regression Tree (BRT). However, most researchers have concluded that data mining techniques can be improved by exploring more study areas with different and enhanced newer machine learning algorithms (Balogun et al. 2021).

Thus, this paper explores relatively new machine learning algorithms, the Naïve Baise Multinomials (NBM) and Random Trees (RT). These algorithms are expected to perform better in developing landslides susceptibility maps in a mountainous region. Mountainous regions have often been characterized as data-scarce environments for soft computing analysis (Buijs et al. 2009; Lee JH et al. 2018; Marin et al. 2021). Data scarcity is when the needed data or information for a successful analysis is not readily available. Some forms of qualitative judgments, such as the weight of evidence WoE and the analytical hierarchy process AHP, are employed to augment the missing data (Ibrahim Sameen et al. 2019; Medwedeff et al. 2020; Marin et al. 2021). Although, the overall quality of the soft computing techniques depends on the quality and quantity of the data. Research has however, discovered that some of these algorithms could perform wonderfully well in such environments. For instance, the SVM algorithm is identified to evade overfitting by handling fewer training data over other algorithms (Rahmati et al. 2017; Ibrahim Sameen et al. 2019; Achour and Pourghasemi 2020). The NBM is an advancement of the frequently used Naïve Bayes classifiers, while the RT was built to overcome the limitations of the traditional forest classifiers with the potential of enhancing the accuracy of results. The final models were compared with the conventional data mining algorithm, the Support Vector Machines (SVM), to assess the performance of the new algorithms. The choice of SVM for validation is due to its impressive performance in previous studies, particularly with fewer learning data on many occasions (Ghorbanzadeh et al. 2019; Goetz et al. 2015; Ibrahim Sameen et al. 2019; Lee JH et al. 2018; Marin et al. 2021; Nhu et al. 2020b; Vakhshoori et al. 2019).

The random trees (RT) were obtained from the combination of the RF and DT, an additional step to computing the splits (Merghadi et al. 2020). Furthermore, the RT classifies the dataset decision by providing other subsets to tackle overlearning and overfitting, especially with insufficient data. Another advantage of considering this algorithm is that it can reduce the dataset variance more effectively because a total learning sample is used to establish the trees, replacing bootstrap. The Naïve Bayes Multinomial was the second ensemble used in this research. The algorithm is built based on the Bayesian theorems with improvements to the conventional Naïve Bayes (Chen W et al. 2017c). The idea is to see how well it will perform in terms of landslides predictions within its class compared with the other two algorithms.

The landslide conditioning factors in this work were not randomly selected. A technique that involves probability future selection technique was used to choose the factors influencing landslides in the study area (Cinelli et al. 2014; Yeshwanth et al. 2019; Jena et al. 2020). The factors were selected from a procedure called the weight of evidence, which is a part of the MCDM technique. The technique pairs the factors against each other, thereby placing the most influencing factor above the other based on supplied evidence (Pourghasemi et al. 2013b; Chen et al. 2016; Polykretis and Chalkias 2018). The selected factors were again scrutinized using frequency ratio FR. The FR further reduced the factors into a precise number of factors needed to influence landslides within the study area. The factors selection procedure will cause an improvement to the old ways of analyzing landslides using GIS data and machine learning. The discrepancies that are likely in this procedure, especially when allocating values, can be avoided through careful inputs of the weights and proper allocation of the discounts as the weight of evidence (Ghorbanzadeh et al. 2019; Wang Y et al. 2019; Mohan et al. 2020).

The rest of this paper is structured as follows: Section one (Introduction) talks about the importance of landslides analysis. Highlights of the various methods like the physical methods, statistical methods, and data mining methods were discussed. Limitations of existing techniques, the significance of data mining techniques, and the identification of the research gap are also discussed. Section two presents details of the study area. It describes the geology and the geomorphology of the study area. A detailed discussion on the methodology adopted follows in the third section. Results were discussed in section four. Finally, conclusions and recommendations are provided in the last section of the paper.

2. The study area

The study area for this research is a region of mountainous terrain with a gas pipeline that transports natural gas from Sabah to Sarawak in Malaysia (Figure 2). The site lies within latitudes $0^{\circ}02'45''N$ and $01^{\circ}32'45''N$ and longitudes $105^{\circ}24'05''E$ and $106^{\circ}10'45''E$ and it is covering some $3{,}811.9km^2$. The area has a population of about 35,300 from the year 2000 census report. Being a transit district, major roads that link the capitals of the two states Sabah and Sarawak, remain busy almost throughout the day. Economic activities in the area comprise some apple cultivation as well as palm plantations in the high lands. The region is well known for its rough mountainous terrains measuring over 1800 m above mean sea level at some points.

The climate in Lawas is that of a typical rain forest or simply an equatorial climate that is common to areas situated between 10° to 15° latitude to the equator. The temperatures are apparently high and very humid at some points in the year. An annual average temperature of about 30 °C remains the peak temperature, while an annual low temperature of 24.4 °C is also attainable within the study area. In contrast, the rainfall in this area is a typical northern Borneo monsoon rain that falls intensely from September to March. High annual rainfall of up to 4,178mm can be recorded within the mountainous regions of Lawas. These rains are sometimes characterized as continuous downpours because they usually fall continuously for several hours or even days.

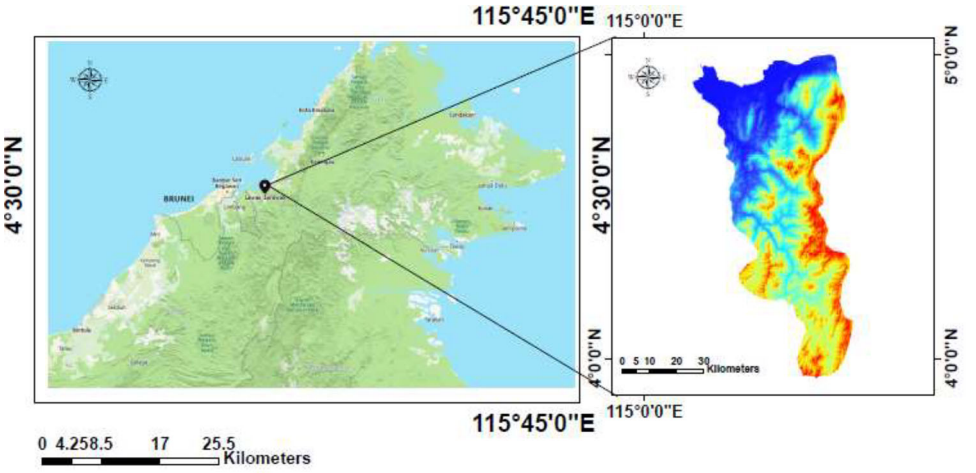


Figure 2. Study area of gas pipeline placed within Sabah and Sarawak area.

Lawas geologically is classified as stable, which means there are no seismic activities recorded yet in the area. The geological composition consists of thick and sequence layers of Eocene-Oligocene grey bluish fine to medium-grained sandstones. In addition, a formation of red/grey shales forms the soil beds traced to the 'Crocker Formation'. The Crocker formation has strata that extend towards the northeast forming steeper terrains as it moves to the east or west. Lithological units in the area comprise some thick to very thick-bedded rock units of sandstones and interbedded shales. Rock stratification in this area can be categorized into two categories: the 'sandy sequence' and the 'shale sequence'. An extension of low-lying flat areas that lies towards the coastal regions and extends in wetlands and swamps. At the same time, the hilly regions provide for most of the inhabited lands of the region. The geology and geomorphology of the study area have classified it as an area where landslides can quickly occur with a bit of trigger.

3. Data and methods

The methodology adopted for this research is as shown in (Figure 3). The work starts with data collection after an extensive literature review on landslides and their analysis. The identification of the landslide points leads to the development of an inventory map. This inventory is developed using landslides' history records, interpretation of satellite images, and site visitation reports for the past ten years. A total number of 148 landslide locations were identified from this study area. These landslide locations served as a preference for training and validation datasets. In other to avoid the bias associated with probabilities, the same number of non-landslides locations were identified. Digital Terrain Model (DTM) of the study area with a high resolution of 50×50 cm was used to derive all the spatial factors contributing to landslides in the area for the past ten (10) years. The datasets were randomly divided into a ratio (70%-30%) (Hegde and Rokseth 2020; Mohan et al. 2020). As stated earlier, 70% of the total data is used to train the algorithms. The remaining 30% is used as validation

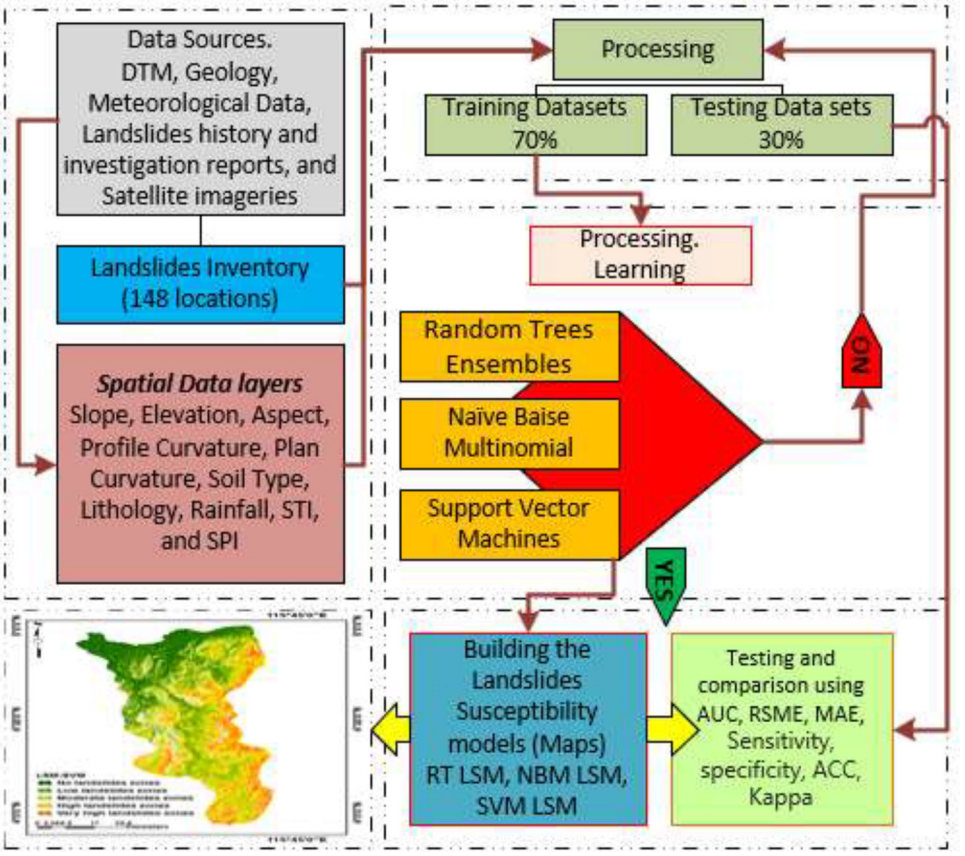


Figure 3. Study flow chart.

data. Splitting the datasets into training and validating or testing is necessary for data mining analysis. The idea is to use most of the data for training sets while fewer parts validate the trained data sets. Selection of landslides and non-landslides locations is conducted to ensure that there is a similarity in the data when it is split in tow. Validation datasets are being used to make predictions against the training datasets. The validation datasets have values already known in the attributes, making it easy to identify the correctness of the predictions.

A clustering method was applied to sample the landslides data from the non-landslides data. The K-means algorithm can place and group data to the specified cluster or centroid (Chang Z et al. 2020; Keyport et al. 2018). The Datasets were partitioned into the specified number of clusters (landslides and non-landslides). Simultaneously, the clustering is continued by grouping the datasets into the predefined clusters from the centroid. The landslides conditioning factors were selected using a modified future selection methods (Goetz et al. 2015; Huang and Zhao 2018; Ibrahim et al. 2020 MB). Ten landslide conditioning factors, including Slope Angle, Slope Elevation, Slope Aspect, Profile Curvature, Plan Curvature, Lithology, Soil Type, STI, SPI, and Rainfall, were used for the analysis. (Table 1) explains the data sources classification and the type of data or model obtained, while (Table 2) indicates details of soil and

Table 1. Spatial data types and data layers of the site.

S/No	Data source	Data format	Input scale and resolution	Data layer developed	Output format	Output resolution
1	Digital terrain model (DTM)	Raster	50 cm interval	Slope, aspect, elevation, plan curvature, profile curvature	Grid	0.5 × 0.5 m
2	Geology Map	Line	1:250,000	Soil type Lithology	Grid	0.5 × 0.5 m
3	Precipitation	Weather stations	10 years return period of annual rainfall data	Rainfall map	Grid	0.5 × 0.5 m

Table 2. Geological details of the study area.

Geological information		
S/No	Name of formation	Description
a.	Temburong formation	Predominantly argillaceous, composed of a laminate sequence of siltstones and shale, and is the laminate sequence of a staffer
b.	Belait formation	The Belait formation contains some conglomerate and pebbly sandstone at the base, passing upwards into alternating sandstone, shale, and coal. A more significant portion is occupied by medium- to very coarse-grained fluvial pebbly sandstone and conglomerate. Interbedded with the conglomerates are pebble-free medium to fine-grained sandstones and minor mudstones.
c.	Crocker formation	This formation comprises a few lithology types, including thick sandstone units, interbedded sandstone, shale unit, and thick shale unit.
d.	Kelalan formation	The Kelalan formation is characterized by having inter-bedded sandstone and hard grey shale, and rare limestone lenses. The sandstone beds of the Kelalan formation are thicker, and the strata are more metamorphic.
e.	Nyalau formation	The Nyalau formation (Middle Miocene) of the Bintulu area, Sarawak, occurred as offshore subtidal estuarine with sandstones, sandy shales, and shales.
f.	Meligan formation	The Meligan formation consists dominantly of white-grey, thick-bedded, well-cemented, frequently cross-bedded medium to coarse-grained sandstones.
g.	Setap shale formation	The Setap Shale formation consists of a thick, extensive, and monotonous shale succession with subordinate thin sandstone beds and a few thin limestone lenses. The typical lithology is grey shale, grey mudstone, sandstones, and a few limestones.
Soil		
No.	**Description of soil area**	**Description of soil composition**
1	Mountain Cuestas	Mudstone, Sandstones and miscellaneous rocks, cambisols/lithosols
2	Mountains and Hills	Ultrabasic igneous rock, rhodic and orthic ferrasols/lithosols
3	Pleau with Gentle Undulating surfaces	Colluvium Gleyic podzol, dysteric gleysols
4	Dissected terraces of 15-25 degrees	Alluvium sandstones and mudstones, orthic acrisols
5	Terraces	Alluvium, gleyic, and dysteric cambisols
6	Valley floor and terraces	Alluvium derived from ultrabasic rocks, orthic ferraisols
7	Swamps	Alluvium peat, humic and dytric histosols
8	Floodplains	Alluvium dystric/eutric regosols, humic dytric
9	Meander belts	Calcareous alluvium, calceric regosol/humic gleysol
10	Tidal Swamps	Sulphidic alluvium, thionic fluvisol/dysteric histosol

geological formations of the study area. Factors associated with the ground's surface are usually developed out of terrain models such as Digital Terrain Models DTM or Digital Elevation Models DEM. Similarly, the remaining landslides' spatial models were created from the topographical maps and charts. Landslides' spatial models that

have to do with the weather and climate are developed using detailed and up-to-date weather records of the study area.

3.1. Factor selection process

Various kinds of literature are yet to identify the number of factors to be used in a landslides analysis or how these factors can be drawn. The reason is that problems associated with landslides are always complex, depending on the nature of the environment (Chen W et al. 2019; Truong et al. 2018). Therefore, the model quality of any landslides analysis as observed depends on the quality of these factors. However, statistical interpolations in recent years were employed to help select relevant factors for the analysis, and results of such statistical selection have been overwhelmed, for example in (Gigović et al. 2019; Pham et al. 2020; Zhao and Chen 2020).

In this research, the weights of evidence (WofE) method was used to trim the number of conditioning factors identified based on the site visitation/investigation reports. The principles of this method are like that of Bayesian probability models (Chen X and Chen 2021; Ghorbanzadeh et al. 2019; Ibrahim M et al. 2019). Many researchers have used this principle to develop landslides susceptibility models for many study scenarios. The WofE technique calculates the weights of every landslide conditioning factor (B) in areas or locations of landslides or no landslides within the selected study area. Thus,

$$W_i^+ = \ln \frac{P\{BL\}}{P\{B\overline{L}\}} \tag{1}$$

$$W_i^- = \ln \frac{P\{\overline{B}L\}}{P\{\overline{B}\overline{L}\}} \tag{2}$$

Where P represents the probability, ln is the natural log, B is the potential landslides predictive factor, $\overline{B}$ potential non-landslides predictive factor, L is the locations of the landslides and $\overline{L}$ represents the non-landslides locations or points. W_i^+ Indicates the presence of a predictive variable within a landslide location with a magnitude that explains a positive correlation between the landslides and the predictive variable presence (Equation (1)). While W_i^- indicates the absence of a predictive variable with a negative correlation (Equation (2)).

A difference between the two weights W_i^+, W_i^- is defined by a factor called weight contrast W_f (Equation (3)) thus,

$$W_f = W_i^+ - W_i^- \tag{3}$$

This expression represents the entire spatial relationship between the predictor variable and landslides.

The second phase was to use the Frequency Ratio (FR) method to quantify the level of involvement of the factors in the slides. The method of FR in landslides susceptibility

Table 3. Frequency ratio of the landslide influencing factors.

Factor	Classes	Pixel count (z)	% Number of pixel count (x)	Landslides pixel count (y)	% Landslides pixel count (a)	Frequency ratio (FR)
Slope angle	0–10	845,817	17.15	4312	3.71	0.22
	10–20	914,561	18.54	12,480	10.73	0.58
	20–30	992,415	20.12	65,874	56.64	2.82
	30–40	623,458	12.64	7531	6.47	0.51
	40–50	514,689	10.43	12,549	10.79	1.03
	50–60	445,287	9.03	5324	5.01	0.55
	60–70	269,732	5.47	1283	1.1	0.2
	70–80	215,483	4.37	4218	3.62	0.83
	>80	110,269	2.24	2739	2.35	1.05
Elevation	<150	990,031	20.07	19,873	6.24	0.31
	150–500	679,763	13.78	98,652	30.96	2.25
	500–750	507,239	10.29	53,278	16.71	1.62
	750–1000	977,761	19.83	35,981	11.29	0.57
	1000–1200	960,387	19.47	25,698	8.06	0.41
	1200–1400	493,422	10	15,329	4.81	0.48
	>1400	323,108	6.55	69,875	21.93	3.35
Slope aspect	Flat	235,411	4.77	8641	7.43	1.58
	North	487,951	9.89	9861	8.48	0.86
	Northeast	565,073	11.46	6587	5.66	0.49
	East	621435	12.6	5698	4.9	0.39
	Southeast	569,832	11.55	43,259	37.19	3.22
	Southeast	548,763	11.13	23,478	20.19	1.81
	Southwest	689,753	13.99	1653	1.42	0.1
	West	589,712	11.96	14,586	12.54	1.05
	Northwest	623,781	12.65	2547	2.19	0.17
Profile curvature	Concave	2,143,114	43.46	458,785	78.41	1.8
	Flat	85,785	1.74	854	0.15	0.09
	Convex	2,702,812	54.8	125,496	21.45	0.39
Plan curvature	<(−0.001)	3,245,876	65.82	327,568	41.31	0.63
	(−0.01)–(0.01)	439,939	8.92	6572	0.83	0.09
	>(0.01)	1,245,896	25.26	458,796	57.86	2.29
Lithology	Temburong formation	654,791	13.28	84,561	27.46	2.07
	Belait formation	959,214	19.45	62489	20.29	1.04
	Croacker formation	621,487	12.6	31,562	10.25	0.81
	Kelalan formation	754,763	15.3	31,291	10.16	0.66
	Nyalau formation	570,137	11.56	48,561	15.77	1.36
	Meligan formation	658,741	13.36	36,946	12	0.9
	Setap Shale formation	712,578	14.45	12,548	4.07	0.28
Soil type	Cambisols/Lithosols	457,812	9.28	54,879	21.68	2.34
	Ferralsols/Lithosols	654,812	13.28	6451	2.55	0.19
	Podzols/Dysteric gleysols	345,874	7.01	25,548	10.09	1.44
	Orthic acrisol	327,812	6.65	4521	1.79	0.27
	Alluvium gleyic and cambisols	214,587	4.35	21,544	8.51	1.96
	Orthic ferralsol	439,361	8.91	2648	1.05	1.15
	dytric histosols	987,452	20.02	4712	1.86	0.09
	Humic/Dytric	687,452	13.94	18,753	7.41	0.53
	Calceric regosol/ humic gleysol	458,735	9.3	15,489	6.12	0.66
	Thoinic fluvisol/ dystric histosol	357,814	7.26	98,621	38.96	5.37
Rainfall	0–500	359,874	7.3	9875	3.27	0.45
	500–1000	687,458	13.94	58,746	19.47	1.4

(*continued*)

Table 3. Continued.

Factor	Classes	Pixel count (z)	% Number of pixel count (x)	Landslides pixel count (y)	% Landslides pixel count (a)	Frequency ratio (FR)
	1000–1500	865,743	17.55	65,781	21.8	1.24
	1500–2000	245,887	4.99	6987	2.32	0.46
	2000–2500	751,368	15.24	32,548	10.78	0.7
	2500–3000	434,149	8.8	32,578	10.79	1.23
	3000–3500	621,358	12.6	87,452	28.98	0.71
	>3500	965,874	19.58	7823	2.59	0.13
STI	5–10	565,229	11.46	62,547	22.01	1.92
	10–15	542,894	11	87,652	30.85	2.8
	15–20	794,344	16.11	24,589	8.65	0.54
	20–25	713,169	14.46	32,598	11.47	0.79
	25–30	637,306	12.92	42,583	14.99	1.16
	30–35	875,493	17.75	21,546	7.58	0.42
	35–40	803,276	16.29	12,587	4.43	0.27
SPI	−5–0	2,659,412	53.93	128,456	18.99	0.32
	0–3	2,272,299	46.07	547,832	81.01	1.76

analysis has been in use for quite some time now (Yan et al. 2019). Could be the first researchers to have reported the use of the technique for landslides analysis. The method solely provides the relationship between landslides in the area, the conditioning factors, and the interrelationship between the factors' variables. The FR is classified as a quantitative statistical approach that can relate the spatial distributions of the factors leading to landslides within their interdependencies and landslides. The FR as computed for this research (Table 3) indicates the probability of the ten (10) landslides conditioning factors as dependent variables and the inter-dependency within the component's pixel counts. The pixel counts signify the specific area of coverage by each factor occupied in the landslide and non-landslide locations. The percentage of the pixel counts is computed as the percentage of landslide pixel size to the variable corresponding to the percentage pixel count (Acharya and Lee 2019). The FR is interpreted as the values that specify the probability of involvement in the landslide occurrence shown by a particular factor. Those factors with higher probabilities indicate higher participation than those with lower probability values in the landslide occurrence. The FR respective probabilities can easily be compared to find factors contributing to the landslides more (higher probabilities) and those that contribute less (lower probabilities).

3.2. Preparation of landslides spatial models

As stated above, we selected ten landslides predisposing factors for this analysis. The number of factors is decided after series of factor selection procedures are conducted. Again, the nature of the study area's terrain could also be responsible for the number of factors selected. Afterward, the landslides spatial models were prepared using the relevant data and methods as explained below.

3.2.1. Slope angle

The slope angle provides details of the surface steepness or inclination with the horizontal plane. Sloppy terrains with higher angles of inclination are more susceptible to

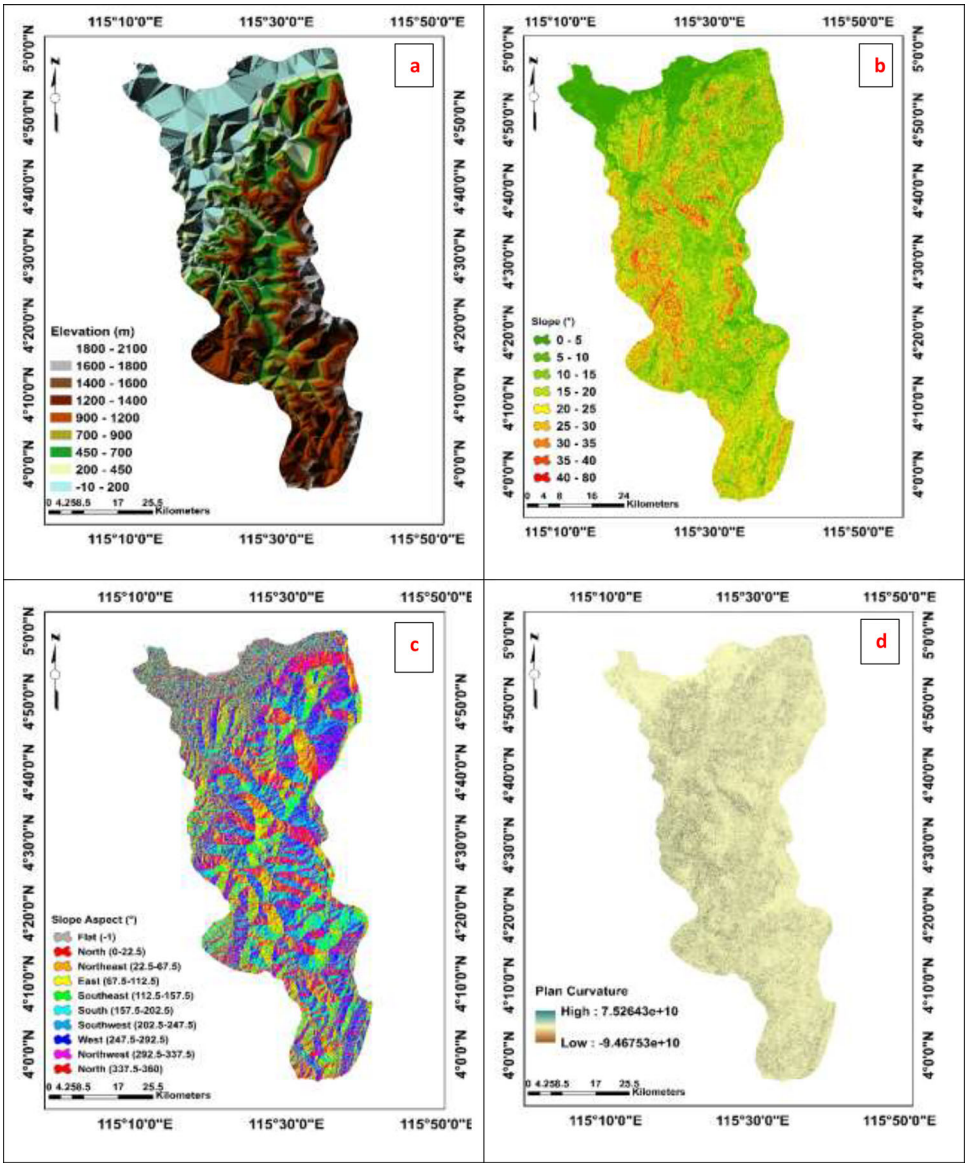


Figure 4. Showing the landslides visualization of conditioning factors (a) elevation; (b) slope; (c) aspect; (d) plan curvature; (e) profile curvature; (f) lithology; (g) soil; (h) SPI; (i) STI; (j) rainfall.

landslides. The slope angle has been used to date by many researchers for landslide prediction because of its relationship with gravitational forces that act on the detaching materials (Nath et al. 2020). Landslide occurs at specific critical slopes, usually termed unstable slopes. It is hard to single out and label a slope as safe or unsafe to landslides despite the size of its angle of inclination without considering other factors. From our study area, the slopes have ranged from 0° to about 82° (Figure 4b) which is a value too high for safe slopes under normal conditions.

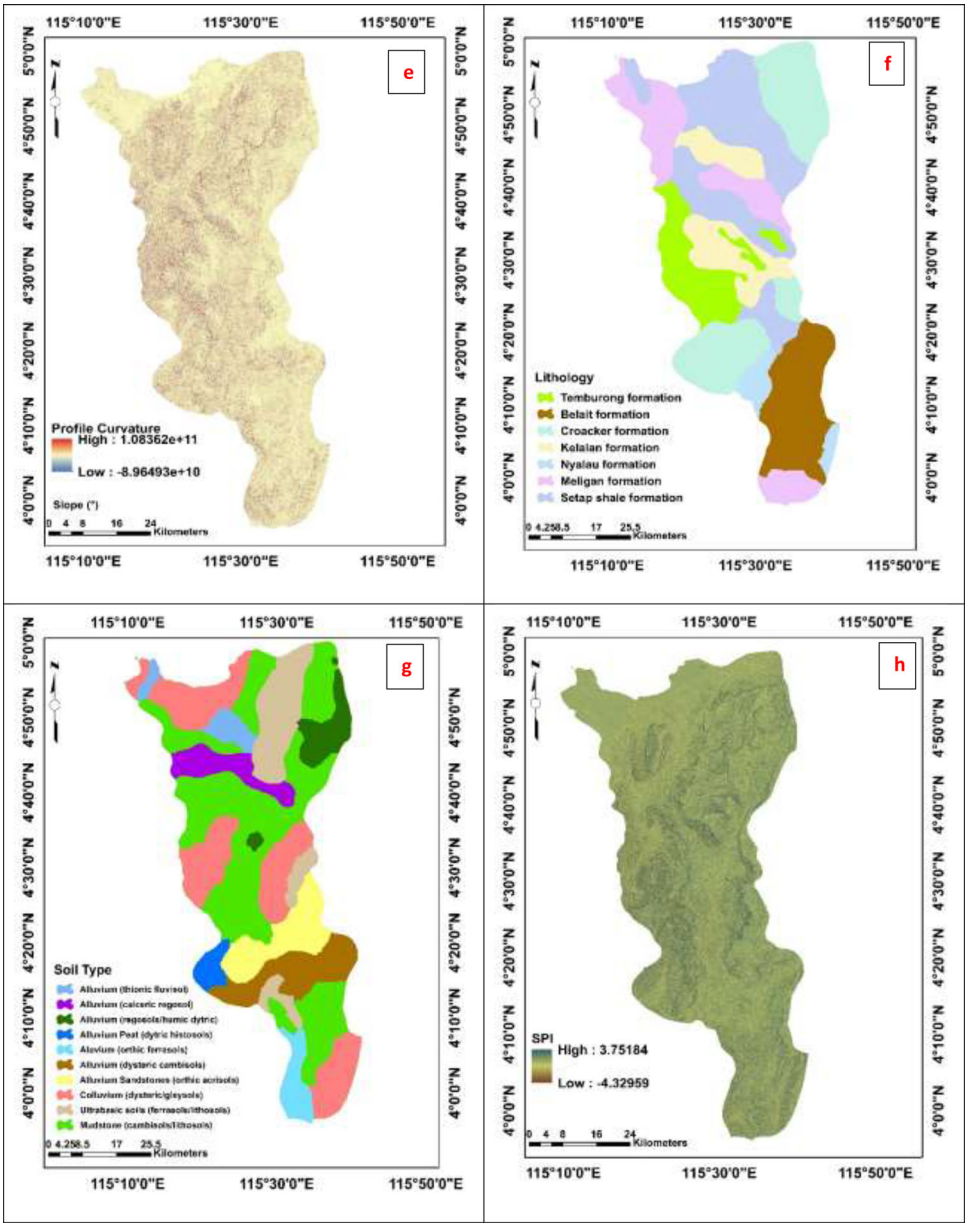


Figure 4. Continued

3.2.2. Profile curvature

Profile curvature (Figure 4e) affects running water flow velocity down slopes because it exits along the slopes' vertical plane. The slopes' rate of change can easily be measured with a change in the ground elevation (Pham et al. 2017)

3.2.3. Plan curvature

The curvatures specify the slopes' surface's nature; it is sometimes referred to as the 'slope of the slopes'. The curvatures originate from the intersections of planes with

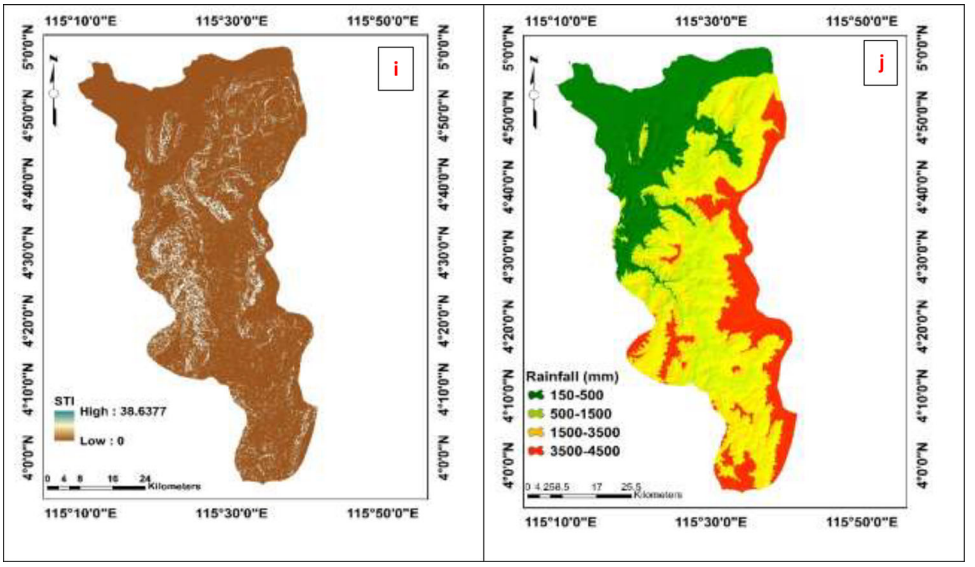


Figure 4. Continued

the direction they are situated on the earth's surfaces. Plan curvature (Figure 4d) has to do with the convergence of water or its divergence when there is a water flow down the slope. This process can quickly erode the sloped surfaces at some sections causing them to fail. Subsequently, landslides can quickly occur in the eroded sections. The nature of these uncertainties has made the plan curvature a critical factor for this analysis. The organic matter distribution in an area is greatly affected by this factor because it reflects the terrain's morphology (Ramakrishnan et al. 2013).

3.2.4. Slope aspect

The slope aspect is concerned with the orientation of the slopes in the study area. The slopes' exposure is critical because some slopes' faces orient to heavier rainfall directions than others. This action might subject such faces to weathering and degradation that will eventually trigger a landslide. Other parameters related to the nature of the slopes' orientation include exposure to direct sunlight, dry/wet heavy winds, saturation degrees, and other forms of discontinuities (Pradhan 2010; Gigović et al. 2019). From the study area (Figure 4c), the values translating the aspects range from -1° that represents the flat land areas to 360°.

3.2.5. Slope elevation

The slopes' elevation defines the slopes' height above the mean sea level (Figure 4a). Researchers have considered slope elevation an essential factor because it relates the detaching mass with slope stability conditions and variables. Unfortunately, the study area for this research has some unprecedented heights of over 800 m above the main sea level, making it more susceptible to landslides (Pourghasemi et al. 2013a).

3.2.6. Rainfall

Rainfall is a vital conditioning factor and a triggering mechanism. Many researchers emphasize the influence of rainfall above other factors on landslides occurrence, especially in areas with no seismic activities (Li WY et al. 2017). Therefore, the kriging method was used to develop the rainfall model (Figure 4j) using rainfall data for the past ten (10) years collected from 16 weather stations across the study area.

3.2.7. Lithology

This factor is vital in deciding landslides in the area because it reveals the type of rock formation. Furthermore, it is an essential factor because it relates to rocks' degradation, making it necessary to know the area's underlying rocks' properties. The model of the lithology of the study area was obtained by digitizing the details of the rock formation obtained from the geological department. So far, seven (7) classes of the formation were identified (Figure 4f), which brought about the formation of more hardened layers in the study area (Pham et al. 2017).

3.2.8. Soil type

The morphological changes made by the soil type when trying to establish the landslides' susceptibility are significant. Therefore, landslide intensity is mainly a function of the nature of the soil in that area. Ten (10) categories of soil classes (Figure 4g) were identified. The soil model was digitized from a detailed soil topography map of the study area obtained from the relevant authorities (Ramakrishnan et al. 2013).

3.2.9. Sediment transport index/stream power index (STI/SPI)

Sediment transport index STI (Figure 4i) characterizes the erosion rate within the study location and the rate flow of the erosion materials, and it was developed using the DTM of the study area. Another important factor determining water in an erosion flow scenario is the SPI (Figure 4h). The SPI for this research was also carved out from the high-resolution DTM of the study area.

3.3. The support vector machine

This algorithm is widely used to establish the landslide susceptibility maps of many places. It separates the linear case from the non-separable linear case, using a low-dimensional input space to map a nonlinear situation (Nath et al. 2020). An optimum hyperplane provides for the best separations in the classes, thus the expression;

$$y_i(w.x_i + b)/ \geq 1 - \xi_i \tag{4}$$

here $\boldsymbol{w}$ determines the position of the hyperplane in the future space, which is termed as the coefficient vector, $\boldsymbol{b}$ defines offsets that exist between the hyperplane and the origin and $\boldsymbol{\xi_i}$ is the positive slack variables.

To determine the optimum hyperplane, we have the expression below by solving Equation (1),

$$\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j (x_i x_j), \;\; subject \;\; to \;\; \sum_{i=1}^{n} a_i y_j = 0, \;\; 0 \le a_i \le C \tag{5}$$

$\boldsymbol{a_i}$ is the multiplier of the Lagrange, $\boldsymbol{C}$ is the constant called the penalty. The equation can be rewritten to give a classification decision function as,

$$g(x) = \sin\left(\sum_{i=1}^{n} y_i a_i x_i + b\right) \tag{6}$$

This classification decision function can be written to determine the separating hyperplane using the linear kernel function.

Thus,

$$g(x) = \sin\left(\sum_{i=1}^{n} y_i a_i K(x_i, y_i) + b\right) \tag{7}$$

The function $\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j})$ represent the kernel function.

The SVM algorithms under this expression provide four (4) different types of input or kernel functions. These include the radial basis function (RBF), Polynomial (PL), Sigmoid (SIG), and Linear (LN) functions.

3.4. Random trees classifiers

Random forest is a machine-learning algorithm that has been used in many landslide situations to make predictions and implement the predictions into maps with GIS

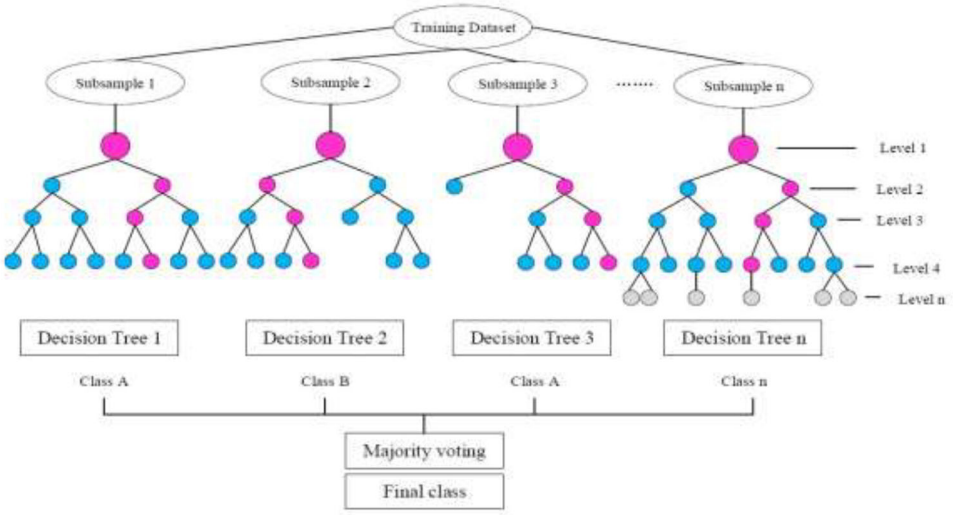


Figure 5. Random trees classifiers.

software. In this research, the random forest classifiers (Figure 5) address the classifiers' limitations. The random forest classifiers usually suffer from many high variances, which affects their accuracy compared to other classifiers (Ghorbanzadeh et al. 2019). Furthermore, random tree classifiers are introduced to overcome conventional forest algorithms' limitations when dealing with many variances. Landslide indeed has many variances when trying to analyze the phenomenon using soft computing procedures. In this situation, the random trees' capability to handle the variances was checked using the algorithm to train the datasets and monitor the outcomes' training processes.

Overall, unlike the support vector machines, this classifier can deal with mixed categorical and numerical variables. The classifier also has a lesser sensitivity when scaling the data, unlike the SVM that must normalize the data before the training process began. As reported by many scholars, the advantage of SVM over the random forest is that it performs with even a small data size or with an unbalanced data type (Ibrahim MB et al. 2020).

3.5. Naïve Bayes multinomial

The NBM algorithm belongs to the class of algorithms with the Bayesian theorems principles. This research considers using this algorithm to provide for the advancement of the frequently used Naïve Bayes classifiers. The improvement in the classification process can be viewed as a form of optimization to the Naïve Bayes performance. Multinomial Naïve Bayes classifiers compute a random variable's likelihood counts differently from Naïve Bayes (Chen W et al. 2019).

3.6. The landslide susceptibility modelling

In this paper, ten landslide-conditioning factors were used as the landslide predictors in the study area (Figure 4a–j). The relationship between landslide points as identified on the inventory map and the conditioning factors were extracted. The extracted data were then divided randomly into the mentioned ratio of 70% as training datasets and the remaining 30% as testing or validation datasets (Figure 6a). Next, the three algorithms discussed earlier were applied to the training datasets for classification. After the training procedures, the results now predicted values were used to develop the landslides susceptibility maps. Then, the susceptibility map was reclassified into five zones of landslides in the area based on the severity of the slides. Subsequently,

In developing the landslides susceptibility maps, landslides susceptibility indices (LSIs) were established from the training and testing datasets (Balogun et al. 2021). These values constitute the landslides susceptibility map of the area developed using the ArcMap software. In addition, generated maps were reclassified, meaning the maps were categorized according to the severity of the landslides' susceptibility. As a result, five categories were identified: regions of very high landslides susceptibility, high landslides susceptibility, moderate landslides susceptibility, low landslides susceptibility, and very low landslides susceptibility were classified from the original map. The landslides susceptibility maps are shown in (Figure 6b) for the SVM model, the Random trees (Figure 6c), and the Naïve Bayes Multinomial (Figure 6d) (Li WY et al. 2017).

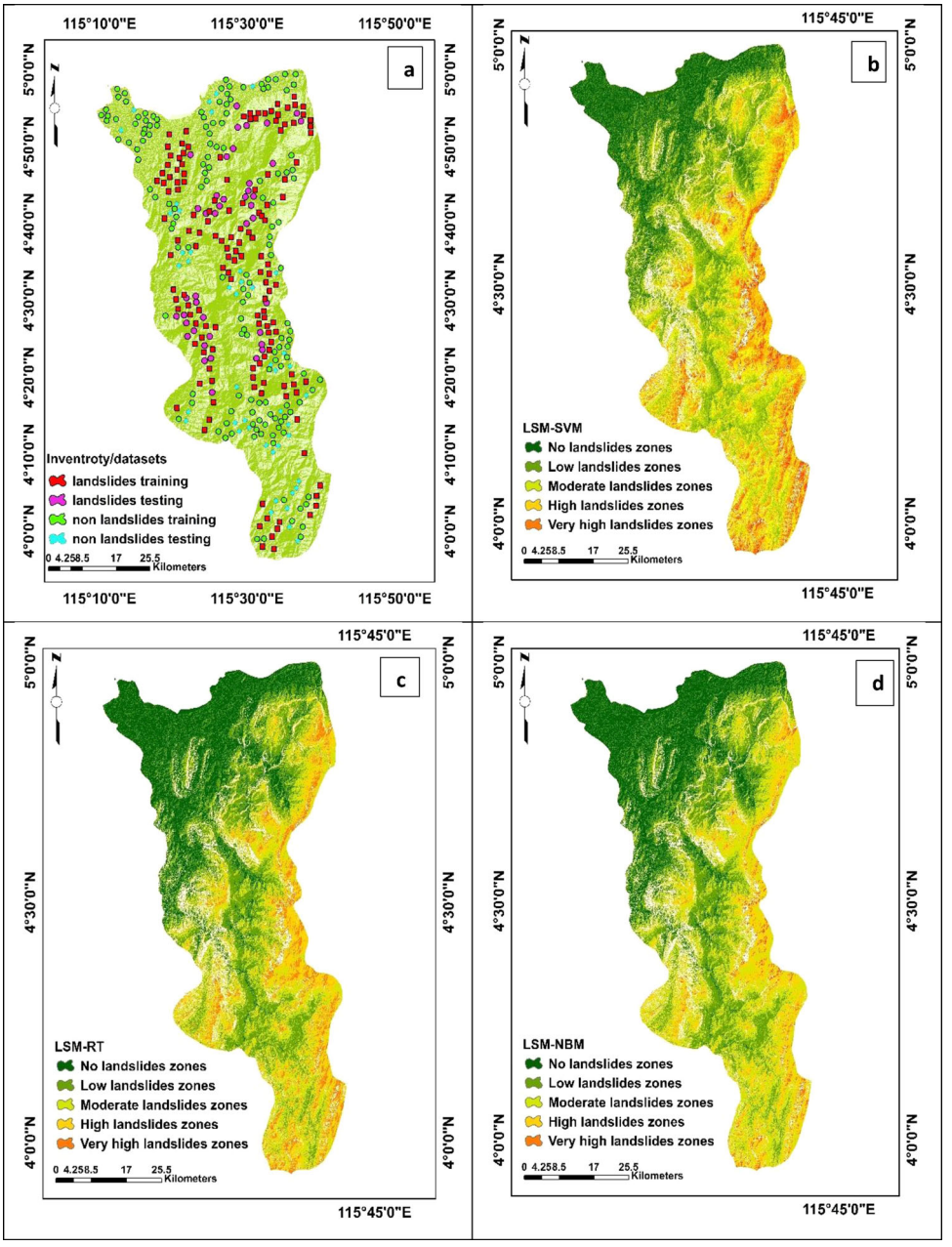


Figure 6. (a) Showing the landslides inventory and datasets; (b) landslides susceptibility map by SVM model; (c) landslides susceptibility map by RT model; (d) landslides susceptibility map by NBM model.

3.7. Evaluation of the model's performance

The ROC and AUC measure and visualizes the performance characteristics of our models in the multiclass classification. (Figure 7) shows the classes in a confusion matrix where the algorithms' performance on the datasets is put into classes. The classes include True Positive (TP), False Positive (FP), True Negative (TN), and False

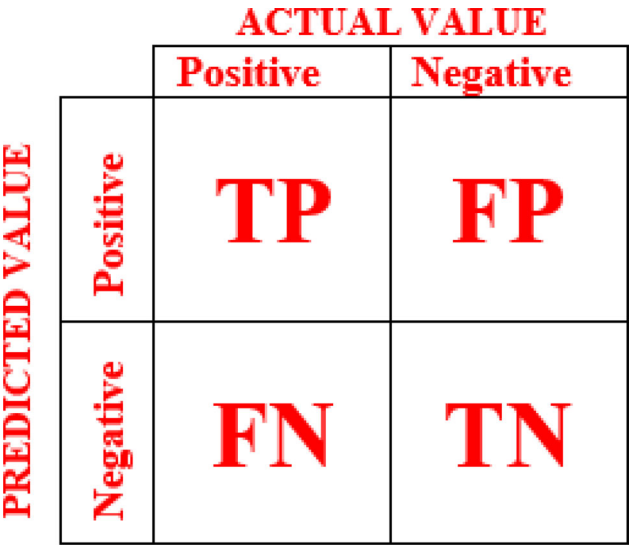


Figure 7. A confusion matrix.

Negative (FN). The sensitivity (True positive or Recall) tells the proportion of positive class (landslides locations) that are correctly classified as landslides (Equation (8)). In contrast, the specificity (True Negative Rate) tells the proportion of negative class (non-landslides locations) that are correctly classified as non-landslides (Equation (9)). Between sensitivity and specificity lies False Negative Rate (FNR), which signifies the proportion of landslide points wrongly classified as landslides (Equation (10)). The False Positive Rate (FPR) tells the proportion of non-landslides incorrectly classified as non-landslides (Equation (11)).

$$sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$specificity = \frac{TN}{TN + FP} \tag{9}$$

$$FNR = \frac{FN}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{TN + FP} = 1 - specificity \tag{11}$$

Other statistical analyses reveal the model performance when it functions separately in the presence or absence of the datasets. These include the Root Mean Square Error (RSME), the Mean Absolute Error (MAE), Accuracy (ACC), and the F-Measure. For example, the RSME (Equation (12)) takes the square root of the difference between each observed data and predicted data per the total number of non-missing data points.

Table 4. Performance evaluation of the models for training and validation datasets.

	Training datasets			Validation datasets		
	SVM	RT	NBM	SVM	RT	NBM
Sensitivity	0.807	0.776	0.732	0.833	0.767	0.787
Specificity	0.782	0.778	0.741	0.791	0.797	0.803
ACC	0.795	0.777	0.736	0.814	0.763	0.793
AUC	0.833	0.814	0.792	0.841	0.822	0.814
RSME	0.224	0.241	0.274	0.475	0.512	0.489
MAE	0.445	0.521	0.327	0.481	0.573	0.349
F-measure	0.794	0.777	0.736	0.811	0.782	0.795
Kappa	0.589	0.553	0.579	0.590	0.564	0.625

$$RSME = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \quad (12)$$

$$where\ i = variabe_i,$$

$$N = number\ of\ non - missing\ data\ points$$

$$x_i\ actual\ obseved\ time\ series\ and$$

$$\hat{x}_i = estimated\ time\ series$$

The percentage of correctly predicted values to instance summation defines the (ACC) of the algorithm (Equation (13)).

$$\text{Accuracy(ACC)} = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

The MAE measures the acquired errors between the paired observations in the same class expression (Equation (14)).

$$MAE = \frac{|p_1 - a_1| + |p_2 - a_2| + \ldots + |p_n - a_n|}{n} \quad (14)$$

where p_i is the predicted value and a_i is the actual value.

The F-Measure is another statistical technique that measures the model's performance. It combines the precise values and sensitivity values to form a single measure that captures both properties with their exact weighting (Equation (15)).

$$F - measure = \frac{(2 \times Sensitivity \times Specificity)}{Sensitivity + Specificity} \quad (15)$$

Kappa index

The kappa index is denoted by the following relationship (Equation (16))

$$2\left(\frac{(TP*TN)-(FN*FP)}{(TP*FN)+(TP*FP)+(2*TP*TN)+(FN^2)+(FN*TN)+(FP^2)+(FP*TN)}\right) \quad (16)$$

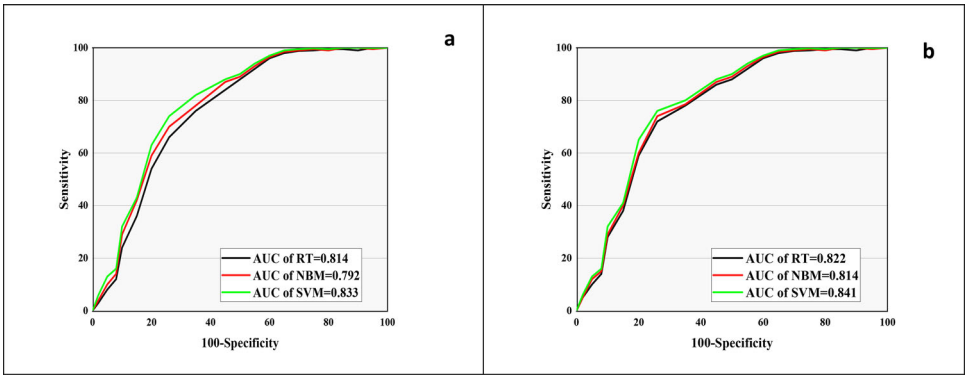


Figure 8. Sensitivity vs. Specificity graphs (a) training datasets; (b) testing datasets.

Table 5. Showing the significance level among the models.

Pairs	Z-value	P-value	Significance
SVM vs. RT	−3.678	0.000	Yes
SVM vs. NBM	−3.521	0.001	Yes
RT vs. NBM	−2.989	0.000	Yes

Table 4 expresses the results obtained from validating the three data mining algorithms' performances. The performance evaluation is conducted on both the training and validation datasets. As stated earlier, four (4) performance evaluators were used to check the prediction rate and the success rate of the models developed and the data used in developing them. Performance validation from the training datasets shows that the traditional data mining algorithm (SVM) is still significant in creating landslides susceptibility models from this study area (Figure 7). However, the two new models have also performed above the benchmark of 0.75 and could be used in landslides susceptibility analysis (Table 4).

3.8. Significance of the statistical evaluation

Landslides models obtained through mathematical simulations are evaluated for the model performance using statistical evaluation methods such as the AUC (Figure 8), RSME, MAE, F-Measure, Kapper, to mention a few. When two or more models are involved in an assessment, a statistical significance test is usually conducted to establish the best model and reduce the subjectivity level in the final report (Table 5). The P-value and Z-value test for the models were computed and explained using Wilcoxon signed-rank test technique (Tsangaratos and Ilia 2017a; Khosravi et al. 2018; Hong et al. 2020).

4. Results and discussion

The predicted models were built using WEKA software and data from ten (10) conditioning factors mentioned earlier (Figure 4a–j). The factors were selected using factor selection procedures and re-screened and quantified using frequency ratio FR. We compute the values of 8 performance indicators using various methods. These

performance indicators include the Sensitivity, Specificity, ACC, RSME, F-measure, MAE, AUC, and Kappa. As captured in (Table 4), the performance analysis was conducted to validate further the landslides predictions obtained from the SVM, RT, and NBM algorithms. Furthermore, these indices were established to further explains the models' maximum likelihood, for example, in the work of (Gholami et al. 2020; Balogun et al. 2021). Statistical significance test (Table 5) is conducted to check the level of significance among the models, which helps in reducing subjectivity (Ritter and Muñoz-Carpena 2013). The P-Value and the Z-values test were part of the statistical significance investigation on the models as reported similarly by (Chen W and Zhang S 2021; Mohammadifar et al. 2021).

The landslides susceptibility models developed from the SVM, RT, and NBM algorithms (Figure 5c,d) were subjected to the performance evaluation (Hong et al. 2020; Li L et al. 2020; Shin 2020). Conducting a performance analysis on results obtained through data mining techniques is crucial and cannot be over-emphasized (Remondo et al. 2003; Brock et al. 2020; Mohammadifar et al. 2021). In addition, the evaluations help verify and define the level of accuracy and performance of the landslides susceptibility models (Althuwaynee et al. 2021). Results from the performance evaluation show that the AUROC value for SVM on the training datasets is 0.833 against 0.814 and 0.792 for both RT and NBM. This means that the SVM models have higher accuracy over the remaining two algorithms. Thus, an area with similar environmental conditions with this study location can opt for the SVM algorithms even though the remaining two algorithms have performed wonderfully well. The SVM algorithm was observed to have higher strength in determining the probability of landslides pixels correctly classified as landslides (sensitivity). A 0.807 sensitivity value was observed for the SVM, while the remaining were computed to be 0.776 and 0.732 for both RT and NBM, respectively. The SVM recorded a specificity value of 0.782, and RT recorded 0.778, while NBM recorded 0.741 for the training datasets. Specificity values indicate the non-landslides regions or zones that are correctly classified or identified as non-landslides.

Another performance evaluator computed for this study is the Kappa index. This index is necessary to find a substantial agreement or disagreement between the prediction and observation outputs. For this study, kappa values obtained were 0.589, 0.553, and 0.579 for the SVM, RT, and NBM algorithms with the training data. Similarly, other statistical performance indicators computed were the RSME which recorded 0.224, 0.241, and 0.274 for the three algorithms (SVM, RT, and NBM). In addition, SVM recorded an accuracy (ACC) value of 0.795, 0.777 for the RT algorithm, and 0.736 was observed for the NBM. Similarly, these indicators were also computed on the validation or testing datasets (Table 4). In other words, the statistical evaluation of the testing datasets a way of validating the training process and the datasets used for the training (Chung and Fabbri 2003; Beguería 2006; Pradhan 2013; Truong et al. 2018).

The performance assessment on the three models (Table 4) also revealed the differences in classification accuracy for both the training and validation datasets. Although the models were observed to perform better with validation datasets, for instance, with an ACC value of 0.791, the SVM has recorded an AUC value of

0.841 higher than RT and NBM that recorded 0.822 0.814 respectively. This indicates a better prediction accuracy as well ahead of the RT and NBM for this study area, as confirmed by a similar study (Chen W et al. 2019). Furthermore, the observation made on the validation data for the three models indicated that the SVM outperforms the remaining two algorithms in prediction capabilities (Table 4). The trend in the results and the slight difference between the training datasets and the validation datasets may be attributed to the conditional independence assumptions (Chen W et al. 2019). These assumptions were specific to violations made in the training datasets, resulting in the variance and even the lower performance in some of the indicators recorded for the training datasets. Therefore, despite the NBMs' low classification rates in many of the indices or indicators, it displayed a better ability to make adjustments to the weights of some of the variables affected by the assumptions, this was also observed in similar studies (Chen X and Chen 2021; Liu X and Wang Y 2021).

In the case of the statistical significance, the Wilcoxon signed-rank test was computed for the $\boldsymbol{p}$ and $\boldsymbol{z}$ values. A comparison between the SVM, RT, and NBM conducted indicated the significant difference of the models. With the significance level in $\boldsymbol{p}$ values less than 0.05 and $\boldsymbol{z}$ values not exceeding the critical $\boldsymbol{z}$ (-1.96 to +1.96), the models are considered significantly different. The significant test results obtained in this research align with many findings using data mining techniques (Chen W et al. 2019). Thus, the results obtained from the Wilcoxon test shows that the susceptibility models developed from the three algorithms in this study are significantly different. Hence, based on this evaluation, the three models comprising SVM, RT, and NBM are acceptable statistically for landslides susceptibility analysis and mapping in this study area. Furthermore, the reliability of the models (Figure 6b–d), when compared, has been enhanced. The obtained differences are attributed to how well the training process was carried out, plus the sufficiency of the training datasets; these were also observed in many works of literature (Chen W and Zhang Y 2021; Mohammadifar et al. 2021). Lack of a considerable factor difference from model comparisons entails the absence of significant data overfitting in the training process (Hong et al. 2017). With the results analyzed so far, SVM models have outperformed the remaining algorithms by a small margin. However, the margin is significant enough to conclude that the SVM is the better algorithm for this study area among the three. This is in line with many findings, e.g. (Chang Z et al. 2020; Hong et al. 2017; Mohammadifar et al. 2021; Pradhan 2013) that reported SVM outperformed other traditional algorithms.

Landslides analysis using data mining techniques to produce regions of landslides susceptibility from GIS data has been an essential tool in regional planning and management (Hong et al. 2017). In addition, literature has proven that the data mining technique produces landslides susceptibility maps of high predictive accuracies that tackled real-life landslides scenarios (Ma and Xu 2019; Nhu et al. 2020c; Saha et al. 2021). Although, it is still challenging to produce high accuracy landslide models from the technique in various places due to the dynamism of landslides and the factors involved (Tien Bui et al., 2017b; Tien et al. 2020; Balogun et al. 2021). So far, no machine learning algorithm used in the data mining technique was observed to fit all

regions under all landslides conditioning factors perfectly. For instance, the SVM used in this analysis was discovered to perform better in many landslide incidences (Hong et al. 2017). With this in mind, we compare the SVM models with models from NBM and ensembles of DT and RF to find a higher-accuracy landslide model. The model will help manage landslides for this study area with substantial economic relevance that is often disrupted due to landslide activities.

5. Conclusion

This study has assessed the effectiveness of advanced data mining techniques to evaluate landslides in Lawas, an economic giant town in Sarawak, Malaysia. In achieving the said objectives of the study, three machine learning algorithms, namely the SVM, RT, and NBM, were used to train geospatial data extracted from various GIS sources (Bacha et al. 2020; Althuwaynee et al. 2021). The training was made to develop classification between identified landslides location in the area to non-landslides locations by examining the pixels of two classes. An equal number of non-landslides (148) locations was also identified to tackle the problems associated with imbalances in the probability distribution. The results obtained from this analysis are guaranteed comparable with results from the literature.

The ten (10) landslides conditioning factors drawn using the WoE technique were quantified using the FR method to establish the most influencing. The landslides spatial models were developed using respective data layers. The selection of the landslide location was evenly distributed across the study area to enhance the likeness in the split data and the training process. It was also observed from the performance evaluation test conducted (Table 4), the whole analysis was rightly executed and successfully analyzed. Although all three models displayed positive predictive capabilities, SVM turns out to work fine with the geological/geomorphological conditions of the study area. The geological factors were observed to have the highest contributions to landslides events in the area. Soil type, Slope angle, Elevation, and Curvatures were observed to higher FR values than the raining factors. According to reports, the latest landslide events happened after a continuous downpour event that lasted for several hours. With this information compared to our results, it can conclude that the rain in the first place serves as a triggering factor. The slopes could have survived the continuous rain that becomes the suspected most influencing factor.

This study can advise that despite the considerable rainfall intensity in parts of the study area, infrastructures like the pipeline can be protected using the SVM model to plan maintenance accordingly. Overall, the data mining technique looks very promising in managing the landslide mysteries from this study area. However, the study has now revealed more insight into the landslides' causative factors than just rain. When planning, those identified geomorphological factors such as the nature of the soil and the slope and height of the terrain should be given proper attention.

Acknowledgements

The authors would like to thank Universiti Teknologi PETRONAS for the support provided under Yayasan Universiti Teknologi PETRONAS YUTP (015LC0-196). The authors will also

like to extend their acknowledgments to the Editor in Chief, the associate editor, and two other anonymous reviewers for the time taken to review this paper to standard.

Disclosure statement

The authors of this manuscript have no competing statement to declare.

Data availability statement

The authors of this manuscript will state that numerous forms of data obtained from multiple sources were used in this research. High-resolution DEMs were confidentially obtained from the Geological Department of PETRONAS. We are yet to receive any PETRONAS approval so far to share part or whole of this data besides results findings. Rainfall and Topographical details were obtained from the Malaysian Meteorological Department and Geological Departments, respectively. Satellite images can be obtained from this website https://earthexplorer.usgs.gov.

References

Acharya TD, Lee DH. 2019. Landslide susceptibility mapping using relative frequency and predictor rate along Araniko Highway. KSCE J Civ Eng. 23(2):763–776.

Achour Y, Pourghasemi HR. 2020. How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? Geosci Front. 11(3):871–883.

Akgun A, Sezer EA, Nefeslioglu HA, Gokceoglu C, Pradhan B. 2012. An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. Comput Geosci. 38(1):23–34.

Althuwaynee OF, Aydda A, Hwang IT, Lee YK, Kim SW, Park HJ, Lee MS, Park Y. 2021. Uncertainty reduction of unlabeled features in landslide inventory using machine learning t-SNE clustering and data mining apriori association rule algorithms. Appl Sci (Switzerland). 11(2):1–17.

Althuwaynee OF, Pradhan B, Park HJ, Lee JH. 2014. A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. Catena. 114:21–36.

Asadabadi MR, Chang E, Saberi M. 2017. Are MCDM methods useful? A critical review of Analytic Hierarchy Process (AHP) and Analytic Network Process (ANP). Cogent Engineering. 6:1:1623153. doi: 10.1080/23311916.2019.1623153.

Ayodele TO. 2010. Types of machine learning algorithms, new advances in machine learning, Yagang Zhang (Ed.), InTech, Available from: http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms

Bacha AS, Van Der Werff H, Shafique M, Khan H. 2020. Transferability of object-based image analysis approaches for landslide detection in the Himalaya Mountains of northern Pakistan. Int J Remote Sens. 41(9):3390–3410.

Bai SB, Wang J, Lü GN, Zhou PG, Hou SS, Xu SN. 2010. GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology. 115(1–2):23–31.

Balal E, Cheu RL. 2018. Comparative evaluation of fuzzy inference system, support vector machine and multilayer feed-forward neural network in making discretionary lane changing decisions. NNW. 28(4):361–378.

Balogun A-L, Rezaie F, Pham QB, Gigović L, Drobnjak S, Aina YA, Panahi M, Yekeen ST, Lee S. 2021. Spatial prediction of landslide susceptibility in western Serbia using hybrid support vector regression (SVR) with with GWO, BAT and COA algorithms. Geosci Front. 12(3): 101104.

Beguería S. 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Nat Hazards. 37(3):315–329.
Brock J, Schratz P, Petschko H, Muenchow J, Micu M, Brenning A. 2020. The performance of landslide susceptibility models critically depends on the quality of digital elevations models. Geomat Nat Hazards Risk. 11(1):1075–1092.
Buijs FA, Hall JW, Sayers PB, Van Gelder PHAJM. 2009. Time-dependent reliability analysis of flood defences. Reliab Eng Syst Saf. 94(12):1942–1953.
Chang M, Zhou Y, Zhou C, Hales TC. 2020. Coseismic landslides induced by the 2018 Mw 6.6 Iburi, Japan, Earthquake: spatial distribution, key factors weight, and susceptibility regionalization. Landslides. 17(18): 755-772.
Chang Z, Du Z, Zhang F, Huang F, Chen J, Li W, Guo Z. 2020. Landslide susceptibility prediction based on remote sensing images and GIS: comparisons of supervised and unsupervised machine learning models. Remote Sensing. 12(3):502.
Chen W, Pourghasemi HR, Panahi M, Kornejady A, Wang J, Xie X, Cao S. 2017a. Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. Geomorphology. 297:69–85.
Chen W, Xie X, Peng J, Wang J, Duan Z, Hong H. 2017b. GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. Geomat Nat Hazards Risk. 8(2):950–973.
Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J. 2017c. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena. 151:147–160.
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). Bull Eng Geol Environ. 78(1): 247–266.
Chen W, Zhang S. 2021. GIS-based comparative study of Bayes network, Hoeffding tree and logistic model tree for landslide susceptibility modeling. Catena. 203:105344.
Chen X, Chen H, You Y, Chen X, Liu J. 2016. Weights-of-evidence method based on GIS for assessing susceptibility to debris flows in Kangding County, Sichuan Province, China. Environ Earth Sci. 75(1):1–16.
Chen X, Chen W. 2021. GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. Catena. 196:104833.
Chung CJF, Fabbri AG. 2003. Validation of spatial prediction models for landslide hazard mapping. Nat Hazards. 30(3):451–472.
Cinelli M, Coles SR, Kirwan K. 2014. Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment. Ecological indicators. 46 :138–148.
Collins BD, Znidarcic D. 2004. Stability analyses of rainfall induced landslides. J Geotech Geoenviron Eng. 130(4):362–372.
Diana MIN, Muhamad N, Taha MR, Osman A, Alam MM. 2021. Social vulnerability assessment for landslide hazards in Malaysia: a systematic review study. Land. 10(3):1–19.
Díaz SR, Cadena E, Adame S, Dávila N. 2020. Landslides in Mexico: their occurrence and social impact since 1935. Landslides. 17(2):379–394.
Dickson ME, Perry GLW. 2016. Identifying the controls on coastal cliff landslides using machine-learning approaches. Environ Modell Software. 76:117–127.
Dou J, Yunus AP, Bui DT, Sahana M, Chen C-W, Zhu Z, Wang W, Pham BT. 2019. Evaluating GIS-based multiple statistical models and data mining for earthquake and rainfall-induced landslide susceptibility using the LiDAR DEM. Remote Sens. 11(6):638.
Fallah-Zazuli M, Vafaeinejad A, Alesheykh AA, Modiri M, Aghamohammadi H. 2019. Mapping landslide susceptibility in the Zagros Mountains, Iran: a comparative study of different data mining models. Earth Sci Inform. 12(4):615–628.

Gholami H, Mohammadifar A, Pourghasemi HR, Collins AL. 2020. A new integrated data mining model to map spatial variation in the susceptibility of land to act as a source of aeolian dust. Environ Sci Pollut Res Int. 27(33):42022–42039.
Ghorbanzadeh O, Blaschke T, Gholamnia K, Meena SR, Tiede D, Aryal J. 2019. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sens. 11(2):196.
Gigović L, Drobnjak S, Pamučar D. 2019. The application of the hybrid GIS spatial multi-criteria decision analysis best–worst methodology for landslide susceptibility mapping. IJGI. 8(2): 79–29.
Goel L. 2020. An extensive review of computational intelligence-based optimization algorithms: trends and applications. Soft Comput. 24:16519–16549. doi: 10.1007/s00500-020-04958-w
Goetz JN, Brenning A, Petschko H, Leopold P. 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput Geosci. 81:1–11.
Gorsevski PV, Gessler PE, Boll J, Elliot WJ, Foltz RB. 2006. Spatially and temporally distributed modeling of landslide susceptibility. Geomorphology. 80(3–4):178–198.
Gue SS, Tan YC. 2006. Landslides: case histories, lessons learned and mitigation measures. IEM/JKR Geotechnical Engineering Conference 2006; (Ipoh, PERAK: March 6–7.)
Guzzetti F, Carrara A, Cardinali M, Reichenbach P. 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology. 31(1–4):181–216.
Chen W. et al. 2019. Spatial prediction of landslide susceptibility using GIS-based data mining techniques of ANFIS with Whale Optimization Algorithm (WOA) and Grey Wolf Optimizer (GWO). Appl Sci. 9: 3755.G.
Hauser-Davis RA, de Oliveira TF, da Silveira AM, Protázio JMB, Ziolli RL. 2012. Logistic regression and fuzzy logic as a classification method for feral fish sampling sites. Environ Ecol Stat. 19(4):473–483.
Hegde J, Rokseth B. 2020. Applications of machine learning methods for engineering risk assessment – a review. Saf Sci. 122:104492.
Hong H, Liu J, Zhu AX. 2020. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. Sci Total Environ. 718:137231.
Hong H, Pourghasemi HR, Pourtaghi ZS. 2016. Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology. 259:105–118.
Hong H, Pradhan B, Bui DT, Xu C, Youssef AM, Chen W. 2017. Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). Geomat Nat Hazards Risk. 8(2):544–569.
Huang Y, Zhao L. 2018. Review on landslide susceptibility mapping using support vector machines. Catena. 165:520–529.
Ibrahim MB, Harahap ISH, Balogun A-LB, Usman A. 2020. The use of geospatial data from GIS in the quantitative analysis of landslides. IOP Conf Ser: Earth Environ Sci. 540:012048.
Ibrahim M, Pradhan B, Lee S. 2019. Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment. Catena. 186: 104249. https://doi.org/10.1016/j.catena.2019.104249
Ibrahim Sameen M, Pradhan B, Tien Bui D, Alamri AM. 2019. Systematic sample subdividing strategy for training landslide susceptibility models.
Irvin BJ, Ventura SJ, Slater BK. 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma. 77(2–4):137–154.
Jena R, Pradhan B, Beydoun G, Alamri AM, Ardiansyah N, Sofyan H. 2020. Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia. Sci Total Environ. 749:141582.
Keyport RN, Oommen T, Martha TR, Sajinkumar KS, Gierke JS. 2018. A comparative analysis of pixel- and object-based detection of landslides from very high-resolution images. Int J Appl Earth Obs Geoinf. 64:1–11.

Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Prakash I, Tien Bui D. 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci Total Environ. 627:744–755.

Lay US, Pradhan B, Yusoff ZBM, Abdallah AFB, Aryal J, Park HJ. 2019. Data mining and statistical approaches in debris-flow susceptibility modelling using airborne LiDAR Data. Sensors (Switzerland). 19(16):3451.

Lee JH, Sameen MI, Pradhan B, Park HJ. 2018. Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology. 303: 284–298.

Lee S, Lee MJ, Jung HS. 2017. Data mining approaches for landslide susceptibility mapping in Umyeonsan, Seoul, South Korea. Appl Sci (Switzerland). 7(7) 683. https://doi.org/10.3390/app7070683.

Li DQ, Ding YN, Tang XS, Liu Y. 2021. Probabilistic risk assessment of landslide-induced surges considering the spatial variability of soils. Eng Geol. 283:105976.

Li L, Nahayo L, Habiyaremye G, Christophe M. 2020. Applicability and performance of statistical index, certain factor and frequency ratio models in mapping landslides susceptibility in Rwanda. Geocarto Int. 1–19. https://doi.org/10.1080/10106049.2020.1730451

Li WY, Liu C, Scaioni M, Sun WW, Chen Y, Yao DJ, Chen S, Hong Y, Zhang KH, Cheng GD. 2017. Spatio-temporal analysis and simulation on shallow rainfall-induced landslides in China using landslide susceptibility dynamics and rainfall I-D thresholds. Sci China Earth Sci. 60(4):720–732.

Liang Z, Wang CM, Zhang ZM, Khan KUJ. 2020. A comparison of statistical and machine learning methods for debris flow susceptibility mapping. Stochastic Environ Res Risk Assess. 34: 1887–1907. https://doi.org/10.1007/s00477-020-01851-8

Liu X, Wang Y. 2021. Probabilistic simulation of entire process of rainfall-induced landslides using random finite element and material point methods with hydro-mechanical coupling. Comput Geotech. 132:103989.

Liu Z, Gilbert G, Cepeda JM, Lysdahl AOK, Piciullo L, Hefre H, Lacasse S. 2021. Modelling of shallow landslides with machine learning algorithms. Geosci Front. 12(1):385–393.

Lombardo L, Opitz T, Ardizzone F, Guzzetti F, Huser R. 2020b. Space-time landslide predictive modelling. Earth Sci Rev. 209:103318.

Ma S, Xu C. 2019. Assessment of co-seismic landslide hazard using the Newmark model and statistical analyses: a case study of the 2013 Lushan, China, Mw6.6 earthquake. Nat Hazards. 96(1):389–412.

Mandal K, Saha S, Mandal S. 2021. Applying deep learning and benchmark machine learning algorithms for landslide susceptibility modelling in Rorachu river basin of Sikkim Himalaya, India. Geosci Front. 12(5):101203.

Mardani A, Jusoh A, Nor KMD, Khalifah Z, Zakwan N , Valipour A. 2015. Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014.Economic Research-Ekonomska Istraživanja. 28(1): 516–571. doi: 10.1080/1331677X.2015.1075139.

Marin RJ, Velásquez MF, Sánchez O. 2021. Applicability and performance of deterministic and probabilistic physically based landslide modeling in a data-scarce environment of the Colombian Andes. J South Am Earth Sci. 108:103175.

Medwedeff WG, Clark MK, Zekkos D, West AJ. 2020. Characteristic landslide distributions: an investigation of landscape controls on landslide size. Earth Planet Sci Lett. 539:116203.

Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Avtar R, Abderrahmane B. 2020. Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth Sci Rev. 207:103225.

Moayedi H, Mosallanezhad M, Safuan A, Amizah W, Jusoh W, Muazu MA. 2019. A systematic review and meta-analysis of artificial neural network application in geotechnical engineering: theory and applications. Neural Comput Appl. 32: 495–518.

Mohammadifar A, Gholami H, Comino JR, Collins AL. 2021. Assessment of the interpretability of data mining for the spatial modelling of water erosion using game theory. Catena. 200:105178.
Mohammady M, Pourghasemi HR, Amiri M. 2019. Assessment of land subsidence susceptibility in Semnan plain (Iran): a comparison of support vector machine and weights of evidence data mining algorithms. Nat Hazards. 99(2):951–971.
Mohan A, Singh AK, Kumar B, Dwivedi R. 2020. Review on remote sensing methods for landslide detection using machine and deep learning. Trans Emerg Telecommun Technol. 32(7): 1- 23. https://doi.org/10.1002/ett.3998.
Nath PK, Saikia CR, Bhattacharjee N. 2020. A spatio-temporal change detection. Analysis in Central Brahmaputra. Int J Adv Res Eng Technol. 11(10):714–726.
Nhu VH, Janizadeh S, Avand M, Chen W, Farzin M, Omidvar E, Shirzadi A, Shahabi H, Clague JJ, Jaafari A, et al. 2020a. GIS-based gully erosion susceptibility mapping: a comparison of computational ensemble data mining models. Appl Sci (Switzerland). 10(6):1–29.
Nhu VH, Mohammadi A, Shahabi H, Ahmad B, Bin Al-Ansari N, Shirzadi A, Clague JJ, Jaafari A, Chen W, Nguyen H. 2020b. Landslide susceptibility mapping using machine learning algorithms and remote sensing data in a tropical environment. Int J Environ Res Public Health. 17(14):1–23.
Nhu VH, Zandi D, Shahabi H, Chapi K, Shirzadi A, Al-Ansari N, Singh SK, Dou J, Nguyen H. 2020c. Comparison of support vector machine, Bayesian logistic regression, and alternating decision tree algorithms for shallow landslide susceptibility mapping along a mountainous road in the west of Iran. Appl Sci (Switzerland). 15(10): 5047; https://doi.org/10.3390/app10155047
Oh HJ, Lee S. 2017. Shallow landslide susceptibility modeling using the data mining models artificial neural network and boosted tree. Appl Sci (Switzerland). 7(10):1–14.
Oladipupo T. 2012. Types of machine learning algorithms. New Adv Mach Learn. 6: 19–49. https://doi.org/10.5772/9385.
Oliva-González AO, Ruiz-Pozo AF, Gallardo-Amaya RJ, Jaramillo HY. 2019. Landslide risk assessment in slopes and hillsides. Methodology and application in a real case1. DYNA (Colombia). 86(208):143–152.
Pamela Sadisun IA, Arifianti Y. 2018. Weights of evidence method for landslide susceptibility mapping in Takengon, Central Aceh, Indonesia. IOP Conf Ser: Earth Environ Sci. 118(1): 1–6.
Petley D. 2012. Global patterns of loss of life from landslides. Geology. 40(10):927–930.
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.
Pham BT, Phong TV, Nguyen-Thoi T, Trinh PT, Tran QC, Ho LS, Singh SK, Duyen TTT, Nguyen LT, Le HQ, et al. 2020. GIS-based ensemble soft computing models for landslide susceptibility mapping. Adv Space Res. 66(6):1303–1320.
Polykretis C, Chalkias C. 2018. Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models. Nat Hazards. 93(1):249–274.
Pourghasemi HR, Jirandeh AG, Pradhan B, Xu C, Gokceoglu C. 2013a. Landslide susceptibility mapping using support vector machine and GIS at the Golestan province, Iran. J Earth Syst Sci. 122(2):349–369.
Pourghasemi HR, Pradhan B, Gokceoglu C, Mohammadi M, Moradi HR. 2013b. Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arab J Geosci. 6(7):2351–2365.
Pourghasemi HR, Rahmati O. 2018. Prediction of the landslide susceptibility: which algorithm, which precision? Catena. 162:177–192.

Pradhan B. 2010. Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia. Adv Space Res. 45(10):1244–1256.
Pradhan B. 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput Geosci. 51:350–365.
Prakash N, Manconi A, Loew S. 2020. Mapping landslides on EO data: Performance of deep learning models vs. Traditional machine learning models. Remote Sens. 12(3):346.
Rahmati O, Tahmasebipour N, Haghizadeh A, Pourghasemi HR, Feizizadeh B. 2017. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. Geomorphology. 298:118–137.
Ramakrishnan D, Singh TN, Verma AK, Gulati A, Tiwari KC. 2013. Soft computing and GIS for landslide susceptibility assessment in Tawaghat area, Kumaon Himalaya, India. Nat Hazards. 65(1):315–330.
Remondo J, González A, Díaz de Terán JR, Cendrero A, Fabbri A, Chung CJF. 2003. Validation of landslide susceptibility maps; examples and applications from a case study in northern Spain. Nat Hazards. 30(3):437–449.
Ritter A, Muñoz-Carpena R. 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol. 480:33–45.
Saadatkhah N, Kassim A, Lee LM. 2014. Qualitative and quantitative landslide susceptibility assessments in Hulu Kelang area, Malaysia. Electron J Geotech Eng. 19C:545–563.
Saha S, Paul GC, Pradhan B, Abdul Maulud KN, Alamri AM. 2021. Integrating multilayer perceptron neural nets with hybrid ensemble classifiers for deforestation probability assessment in Eastern India. Geomat Nat Hazards Risk. 12(1):29–62.
Saravanan S, Istijono B, Jennifer JJ, Abijith D, Sivaranjani S. 2021. Landslide susceptibility assessment using frequency ratio technique – a case study of NH67 road corridor in the Nilgiris district, Tamilnadu, India. IOP Conference Series: Earth and Environmental Science, 2nd International Conference on Disaster and Management 30 September - 1 October 2020, Indonesia. 708: 012017.
Shin J. 2020. Random subspace ensemble learning for functional near-infrared spectroscopy brain-computer interfaces. Front Hum Neurosci. 14:236–239.
Shirzadi A, Soliamani K, Habibnejhad M, Kavian A, Chapi K, Shahabi H, Chen W, Khosravi K, Pham BT, Pradhan B, et al. 2018. Novel GIS based machine learning algorithms for shallow landslide susceptibility mapping. Sensors (Switzerland). 18(11):3777; https://doi.org/10.3390/s18113777
Song KY, Oh HJ, Choi J, Park I, Lee C, Lee S. 2012. Prediction of landslides using ASTER imagery and data mining models. Adv Space Res. 49(5):978–993.
Sun D, Xu J, Wen H, Wang D. 2021. Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: a comparison between logistic regression and random forest. Eng Geol. 281:105972.
Tien D, Tsangaratos P, Nguyen V, Van Liem N. 2020. Catena comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. Catena. 188:104426.
Tien Bui D, Ho TC, Pradhan B, Pham BT, Nhu VH, Revhaug I. 2016. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. Environ Earth Sci. 75(14). 1101. https://doi.org/10.1007/s12665-016-5919-4.
Tien Bui D, Nguyen QP, Hoang ND, Klempe H. 2017a. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. Landslides. 14(1). 1 - 17. https://doi.org/10.1007/s10346-016-0708-4.
Tien Bui D, Tuan TA, Hoang ND, Thanh NQ, Nguyen DB, Van Liem N, Pradhan B. 2017b. Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a

hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. Landslides. 14(2):447–458.
Truong X, Mitamura M, Kono Y, Raghavan V, Yonezawa G, Truong X, Do T, Tien Bui D, Lee S. 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.
Tsangaratos P, Ilia I. 2017a. Applying machine learning algorithms in landslide susceptibility assessments. In: Pijush S, Sanjiban SR, Valentina EB (Editors) Handbook of neural computation. 1st ed. Academic Press London: Elsevier Inc; p. 433–458.
Tsangaratos P, Ilia I. 2017b. Landslide assessments through Soft Computing Techniques within a GIS-based framework. Ameri J of Geogr Inform Sys. 6(1A): 40 - 42. https://doi.org/10.5923/s.ajgis.201701.
USGS. 2004. Landslide types and processes. Highway Research Board special report. U.S. Department of the Interior. Geological Survey Investigation report. USA. 2004-3072: https://doi.org/FactSheet2004-3072.
Vafakhah M, Mohammad Hasani Loor S, Pourghasemi H, Katebikord A. 2020. Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. Arabian J Geosci. 13(11):1–16.
Vakhshoori V, Pourghasemi HR, Zare M, Blaschke T. 2019. Landslide susceptibility mapping using GIS-based data mining algorithms. Water (Switzerland). 11(11):7–13.
Wang H, Zhang L, Yin K, Luo H, Li J. 2021. Landslide identification using machine learning. Geosci Front. 12(1):351–364.
Wang Y, Fang Z, Hong H. 2019. Science of the total environment comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. Sci Total Environ. 666:975–993.
Xu C, Shen L, Wang G. 2016. Soft computing in assessment of earthquake-triggered landslide susceptibility. Environ Earth Sci. 75(9): 767. https://doi.org/10.1007/s12665-016-5576-7.
Yan F, Zhang Q, Ye S, Ren B. 2019. A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model. Geomorphology. 327:170–187.
Yeshwanth M, Kumar PRS, Mathivanan G. 2019. Comparative study of machine learning algorithms for rainfall prediction. IJTSRD. 3(3):677–681.
Youssef AM, Pourghasemi HR. 2021. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. Geosci Front. 12(2):639–655.
Zhao X, Chen W. 2020. GIS-based evaluation of landslide susceptibility models using certainty factors and functional trees-based ensemble techniques. Appl Sci (Switzerland). 10(1): 16. doi:10.3390/app10010016.