Natural Hazards (2022) 113:965–995
https://doi.org/10.1007/s11069-022-05332-9

ORIGINAL PAPER

Rainfall-induced landslide susceptibility mapping using machine learning algorithms and comparison of their performance in Hilly area of Fujian Province, China

Peng Ye[1] · Bin Yu[1] · Wenhong Chen[1] · Kan Liu[2] · Longzhen Ye[2]

Received: 5 September 2021 / Accepted: 21 March 2022 / Published online: 9 April 2022

Abstract

The rainfall can contribute significantly to landslide events, especially in hilly areas. The landslide susceptibility map (LSM) usually helps to mitigate disasters. However, how to accurately predict the susceptibility of landslides is still a difficult point in the field of disaster research. In this study, five advanced machine learning technologies (MLTs), including the Light Gradient Boosting Machine, extreme gradient boost, categorical boosting (CatBoost), support vector machine, and random forest, are utilized to landslide susceptibility modeling and their capabilities are compared through evaluation indicators. The northern part of Yanping, Fujian Province, China, is selected as the research object, because this area experienced mass landslide events due to extremely heavy rainfall in June 2010, resulting in many casualties and a large number of public facilities destroyed. The influencing factors for landslides, namely topographic, hydrological, geologic and human activities, are prepared from various data sources based on the availability. Through the analysis of the actual situation in the study area, 13 suitable landslide condition factors are considered and the availability of relevant factors is checked according to the multicollinearity test. The landslide inventory including 631 samples in this study area is obtained from historical information, satellite data in Google earth and performed field surveys. The landslide inventory is randomly divided into two datasets for model training and testing with a 7:3 ratio. The area under the curve of ROC, accuracy rate, Kappa index and F1 score are applied to compare the MLTs capabilities. In this paper, the results of factor importance analysis show that the first three important condition factors are the distance to faults, the distance to drainages and the slope. According to the LSMs, in the study area, the central and western regions are at high and very high landslide susceptibility levels, while almost all the eastern and northeastern regions are at medium and low landslide susceptibility levels. The CatBoost model is a very promising technology in landslide research according to the evaluation results, which means that for landslide susceptibility research, gradient boosting algorithms may get more accurate results and show better prospects in the future. Finally, the results of this paper will contribute to environmental protection to a certain extent.

Keywords CatBoost · LightGBM · Multicollinearity · Landslide susceptibility

Extended author information available on the last page of the article

1 Introduction

Landslide is a main geological event triggered by rainfall, earthquakes and human construction activities (Malamud et al. 2004; Li et al. 2020a, b; Yunus et al. 2020), and landslide disasters not only threaten the safety of human life and property, but also damage the ecological environment (Froude and Petley 2018; Li et al. 2020a, b). Therefore, many researchers have studied landslide susceptibility and its spatial distribution map (Korup and Stolle 2014). Mastering the mechanism of landslide occurrence and drawing a map of landslide-prone areas is of great significance to land management and urbanization construction.

Landslide susceptibility refers to the likelihood that a landslide event will occur under certain combinations of conditions (Guzzetti et al. 2007; Constantin et al. 2011). With landslide susceptibility mapping (LSM), it is possible to predict the areas most likely to experience landslides. In the past, GIS-based methods have been utilized by many researchers. However, some methods are subjective and can cause many human errors. The quantitative method is an objective judgment based on the geological conditions and influencing factors of the landslide, such as analytic hierarchy process (Yalcin et al. 2011), logistic regression (Sassa et al. 2017) and regression spline (Conoscenti et al. 2015). There are many factors that can cause landslides, which greatly increases the difficulty of explaining the causes and mechanisms of landslides. Therefore, accurate prediction of landslide susceptibility remains a difficult task, and an excellent modeling method is essential for the establishment of LSM. Machine learning technology (MLT) has been more and more widely utilized in LSM with the development of mathematical statistics theory and computer technology. Compared with traditional technical methods, MLT has the advantage of being able to handle complex and high-dimensional nonlinear data sets; for example, random forest (RF) and decision trees (DTs) have achieved good results in prevention of geological hazards and environmental monitoring applications. MLT has been utilized in LSM (Taalab et al. 2018; Pham et al. 2020), terrain classification (Are et al. 2015) and flood hazards research (Hong et al. 2018; Costache et al. 2020; Avand et al. 2021; Yariyan et al. 2020).

Although MLT performs well in disaster modeling, there is still a controversy about which method is most suitable for predicting landslide-prone areas (Lee et al. 2002; Carrara and Pike 2008). Different MLTs have different advantages and disadvantages. Therefore, it is very necessary to carry out research work on these advanced methods and techniques, including their quantitative and systematic comparison, in order to obtain reasonable conclusions; on the other hand, models with high prediction accuracy (even if the prediction accuracy increases by 1% or 2%) can indeed generate more accurate landslide susceptibility areas (Zhang et al. 2019a, b; Mason and Graham 2002; Jebur et al. 2014), so it is necessary to use more higher performance models to accurately predict landslide susceptibility areas. In the past few years, gradient boosting algorithms have become more and more popular among researchers due to their excellent predictive capability and robustness (Azizi and Hu 2020). The CatBoost, XGBoost and LightGBM, as for very novel methods, have been utilized in other fields. However, they are still rarely utilized in landslide susceptibility research and have not been widely studied.

The main purpose of this paper is to develop five models (i.e., RF, SVM, XGBoost, LightGBM and CatBoost) to map landslide susceptibility in northern Yanping, Fujian Province, China, which is most severely affected by the 618 event. The area under the ROC curve (AUC), the accuracy rate (ACC), the Kappa index and the F1 score are applied to evaluate the capabilities of the MLTs, and the main landslide conditioning factors are determined through the factor importance analysis. In this paper, these five MLTs are utilized for the first time in landslide susceptibility studies in the hilly area of Fujian Province, China. Furthermore, few related papers have dealt with the performance comparisons of novel gradient boosting algorithms (such as CatBoost and LightGBM) with the previously popular MLTs. This study also demonstrates that the gradient boosting algorithm may be more accurate for landslide susceptibility studies and may show better prospects in the future. Finally, the results of this paper can provide a certain degree of reference for local disaster risk management and control and can also provide positive inspiration for the future decision-making and planning of local government departments.

2 Study area and geological setting

On June 18, 2010, large-scale geological events caused by heavy rainfall mainly occur in the upper reaches of the Minjiang River. In this disaster events, especially the northern part of Yanping is the most severely affected. Since the number of geological events reaches its peak on June 18, 2010, and began to decrease after the 18, the local government and related media refer to this disaster event as the event of June 18, abbreviated as the 618 event. The study area is located in the north-central part of Fujian province between 26°41' to 26°46'N and 118°06' to 118°12'E (Fig. 1). Jianxi River is the only river in this region, which is an upper tributary of the Minjiang River, and it spans the entire Yanping area. The study area covers 34.54 km^2 of the most severely affected area in the 618 event in 2010.

2.1 Rainfall

The study area is located in the subtropical monsoon climate zone with abundant rainfall throughout the year. According to the 17-year record (1997 ~ 2013) of rainfall data from previous literature data (Yang 2016), the annual average rainfall in the study area is 1638 mm, and the average monthly rainfall in June exceeded 300 mm (346 mm). It is worth emphasizing that the maximum process rainfall amounts to 705.4 mm in June 2010, more than double the historical rainfall during the same period. The annual average temperature is about 19.3 °C, and average temperature of the hottest month and coldest month is 28 °C (July) and 9 °C (January), respectively (Fig. 2).

From June 14 to 25, 2010, due to the combined effects of high-altitude troughs, low vortex shear and ground stationary fronts, long-term continuous heavy rains occurred in the upper reaches of the Minjiang River, leading to regional group-occurring landslides and debris flow disasters. Based on incomplete statistics, the disaster causes 64 deaths, 7 missing and severe damage to water conservancy, electricity, roads and communication facilities. The direct economic loss amounted to 6.828 billion RMB.

In the study area, there are obvious regional differences in rainfall intensity. Figure 3 shows the daily rainfall records of the six rain gauge stations (Xiayang, Yanfu men, Shili an, Shaxi kou, Dafeng and Taiping) near the study area from June 8 to June 20. The daily rainfall peaked on June 14 and June 18, and the rainfall intensity began to decrease significantly after June 18. On June 14, 2010, the daily rainfall at Shaxi kou rain gauge station is 238.5 mm. Particularly within one hour from 9:00am to 10:00am, the rainfall reached 49.5 mm, which far exceed the extreme heavy rain level (36 mm), leading to regional disaster events occur in the Shaxi kou area at this time. The hourly rainfall measured by Xiayang rain gauge station reaches 37.0 mm (10:00am ~ 11:00am) and 45.5 mm

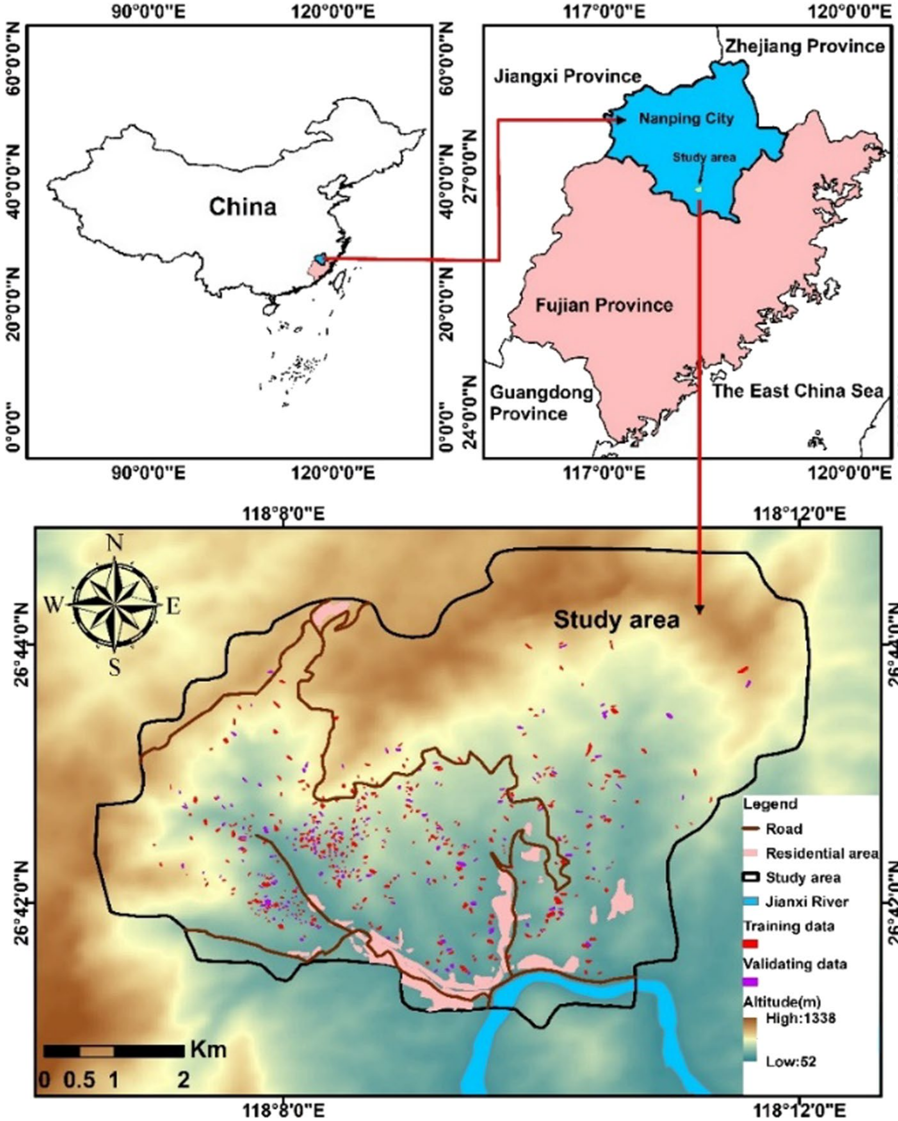


Fig. 1 Landslide inventory map and location of the study area

(11:00am ~ 12:00am), respectively, which reach the extreme heavy rain level, and large-scale disasters also occurred in the Xiayang area in the period. The occurrence of geological events and rainfall condition are positively correlated to a certain extent. According to documents, there are more landslide events in 1998, 2006, and 2010, for which the annual rainfalls are 2182 mm, 1897 mm, and 2814 mm, respectively. Particularly in 2010, the annual rainfall reaches the highest rainfall value in the past 20 years, which causes the group-occurring geological events flows in this region.

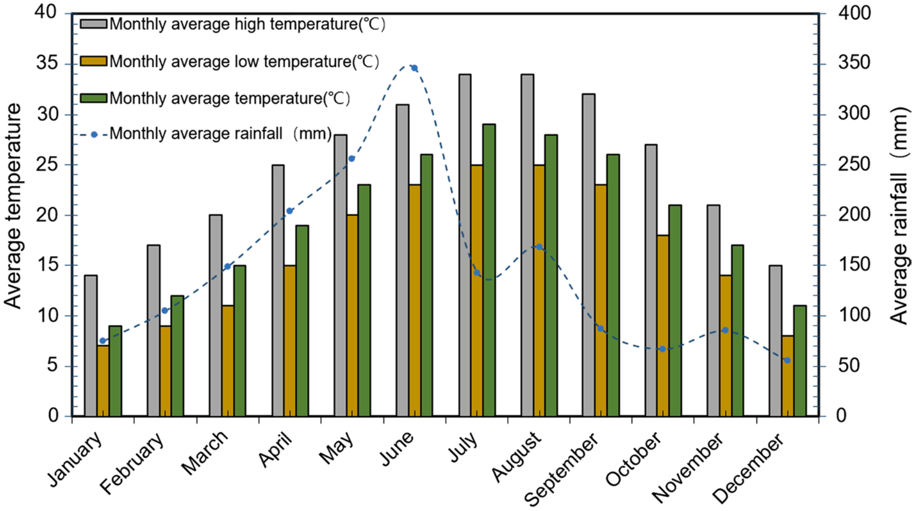


Fig. 2 Average rainfall and temperature in Yanping, Fujian, from 1997 to 2013

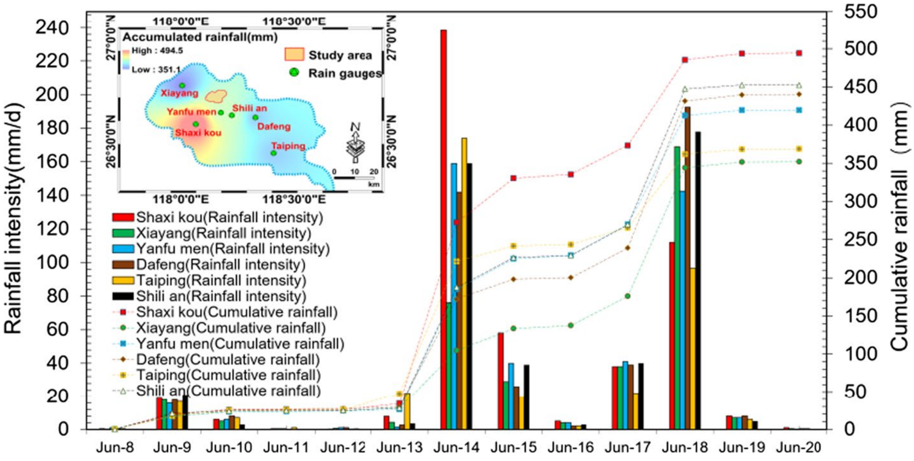


Fig. 3 Daily rainfall intensity and cumulative rainfall intensity measured by six rainfall gauge stations in Yanping from June 8 to 20, 2010

2.2 Geological setting

In the study area, the terrain in the northwest is higher than that in the southeast. It is located on the second uplift belt of the neocathaysian structural system. The tectonic movement is dominated by cathaysia structure and neocathaysian structural system. The neocathaysian structure system is composed of compound folds, compressive or compressive torsion fractures, such as Hulushan anticline, Xiqin syncline and Yanshan fault. The study area belongs to terrain of middle and low mountains hilly. The highest mountain is Mount Mangdang in the middle of the area, with an elevation of about 1363 m.

The main lithology in this area is mostly granite, volcanic rock, Pre-Sinian schist, and gneiss. Among them, the mineral components of igneous rock are mainly quartz, feldspar and biotite. The structure of this kind of rock mass is massive, and it has very high hardness and strong resistance to weathering. The metamorphic rocks in the study area are mostly granulite and schist, and biotite, muscovite, veinlet or banded feldspar are the main mineral components in the metamorphic rocks. Because there are many easily weatherable substances in this type of lithology, the strength of the lithology is generally poor, and it is easily weathered under the influence of temperature difference, rainfall and other factors, resulting in the development of internal fissures. With the continuous development of fissures, the rock mass is eventually decomposed into strongly weathered or completely weathered materials, which leads to the fact that the surface of the hillside in the study area is mostly covered with residual clay of the Quaternary period. The soil type of the local area is one and only. According to the results of field investigation and interviews with local people, the soil type is mainly red soil. The soil depth is generally 0.4 ~ 6 m. The local government has strict protection policies for vegetation. Therefore, most of the local area is covered with vegetation.

3 Methodology

Figure 4 shows the photographs of landslides in the study area. Figure 5 shows the overall flowchart of this study. Firstly, an accurate and reliable landslide inventory is established for study area based on multi-resource data. On the other hand, more detailed information related to the landslide events is collected through field surveys and interviews with local people. And then, 13 appropriate landslide conditioning factors are selected to conduct multicollinearity analysis. After that, five MLTs are used to generate LSMs, namely SVM, RF, CatBoost, XGBoost and LightGBM. At last, the area under the curve of ROC (AUC), accuracy rate (ACC), Kappa index and F1 score are applied to assess the MLTs capabilities. All the MLTs analysis in this study is performed in the Python (3.7.0) environment, and map compilation and production are obtained by ArcGIS software (10.5).

3.1 Preparation of datasets

The acquisition of accurate landslide data is the first step in predicting landslide-prone areas (Guzzetti et al. 2012). According to official data, the location of some landslide events in this study is determined; the other part is obtained by interpreting satellite images in Google Earth pro. Figure 4 shows some typical examples of shallow landslides in the local area. Figure 4a-c shows the field survey photographs after the landslide events. Figure 4d shows the drone photograph taken during the on-site investigation. According to official landslide data provided by Key Laboratory of Geohazard Prevention of Hill Mountains in March 2021, the data have not been made public before. The authors conduct a field survey in the northern part of Yanping in March of the same year. In order to ensure the accuracy of data in the landslide inventory, during the field investigation, the authors carry out drone aerial photography of each landslide point and record the soil type, lithology, vegetation type and soil depth of each landslide in detail. Furthermore, some of the landslide event locations are obtained by interpreting satellite imagery in Google Earth Pro. But it is difficult to judge whether the exposed traces circled are shallow landslides only from the Google satellite images, or it may be a rockfall or rockslide. For landslides

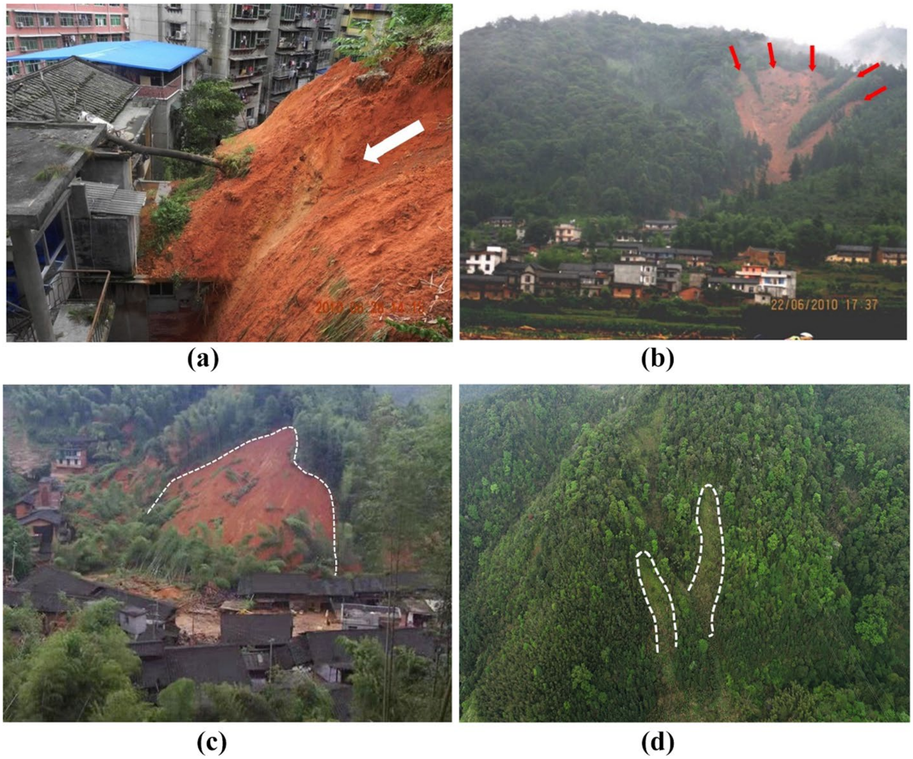


Fig. 4 Photographs of landslides in the study area. **a**, **b**, **c** Photographs of the scene of the landslide events. **d** Drone photographs taken during site investigation

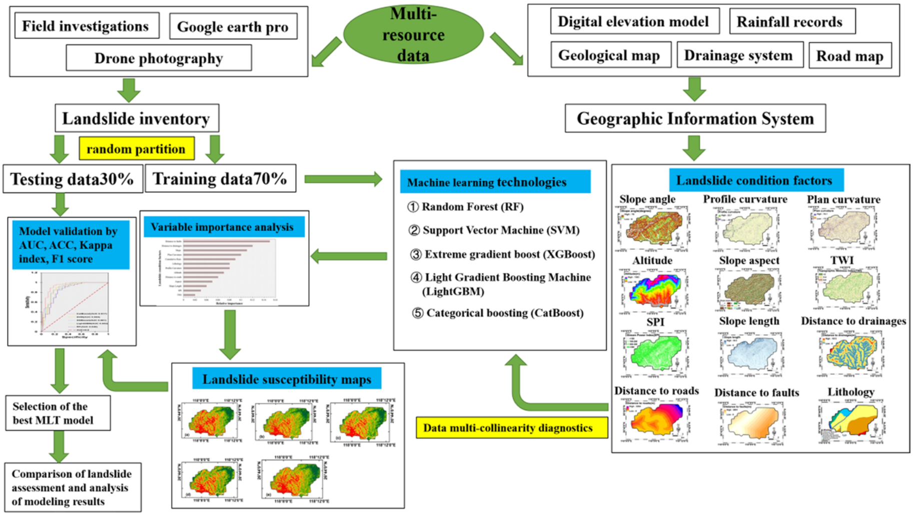

Multi-resource data
Field investigations
Google earth pro
Drone photography
Digital elevation model
Rainfall records
Geological map
Drainage system
Road map
Landslide inventory
Geographic Information System
random partition
Testing data30%
Training data70%
Machine learning technologies
① Random Forest (RF)
② Support Vector Machine (SVM)
③ Extreme gradient boost (XGBoost)
④ Light Gradient Boosting Machine (LightGBM)
⑤ Categorical boosting (CatBoost)
Landslide condition factors
Slope angle
Profile curvature
Plan curvature
Altitude
Slope aspect
TWI
SPI
Slope length
Distance to drainages
Distance to roads
Distance to faults
Lithology
Model validation by AUC, ACC, Kappa index, F1 score
Variable importance analysis
Landslide susceptibility maps
Data multi-collinearity diagnostics
Selection of the best MLT model
Comparison of landslide assessment and analysis of modeling results

Fig. 5 The flowchart of this study

that cannot be investigated at close range, drone photographs are undoubtedly a more reliable basis for judgment. As shown in Fig. 4d, although the vegetation has recovered to a certain extent after the 618 event, the traces of landslides can still be roughly seen from the drone photographs. For rockslides, vegetation is difficult to generate and restore due to the existence of exposed bedrock. Therefore, according to the situation of the on-site investigation and the analysis of the drone photographs, the samples that are mistaken for shallow landslides are excluded to ensure that all of the landslide types in the inventory are shallow landslides. The landslide conditioning factors are extracted from multi-resource data and stored in the spatial database utilizing a spatial analysis tool (ArcGIS software) with a pixel size of 12.5 m. The digital elevation model (DEM) data of the local area are mainly from the China Geographic Information Public Service Platform (https://www.tianditu.gov.cn/) with a resolution of 12.5 m × 12.5 m. The types of landslides in the local area are mainly shallow landslides with a depth of less than 5 m caused by extreme rainfall (Comert et al. 2019). After screening, the landslide inventory containing 631 samples is randomly divided into training and testing datasets at a ratio of 7:3. (Pourghasemi et al. 2013).

3.2 Landslide conditioning factors

The choice of landslide condition factors (LCFs) affects the predictive capabilities of MLT. Therefore, it is necessary to comprehensively consider the type of landslide, landslide mechanism and the characteristics of the geological environment in the study area to prepare LCFs inventory (Merghadi et al. 2018; Wang et al. 2019). At present, researchers do not have a unified standard for the selection of LCF. The lithology, curvature, slope and aspect are widely used LCFs (Camilo et al. 2017). The 13 LCFs are selected from multi-resource data, including profile curvature, plane curvature, slope angle, altitude, aspect, stream power index (SPI), topographic wetness index (TWI), slope length (LS), distance to drainages, distance to roads, distance to faults, lithology and rainfall. These LCFs are utilized based on field investigation and studies (Tseng et al. 2015; Youssef 2016; Keesstra et al. 2016; Kornejady et al. 2017; Ghorbanzadeh et al. 2019; Sevgen et al. 2019). Among them, the morphological parameters such as the slope, aspect, altitude and curvature are extracted from 12.5 m DEM utilizing ArcGIS 10.5 software. The SPI, TWI and LS are calculated by editing formulas in ArcGIS 10.5 software. The lithology and faults are extracted from the geological map (1:50,000 scale) of Fujian Province. The distance to the drainages and the distance to the roads are extracted from the vector data of drainages and roads in Fujian Province. The Euclidean distance tool in ArcGIS 10.5 is utilized to generate distance to faults, drainages and roads map. The accumulated rainfall is obtained by interpolation of the ArcGIS 10.5 software using the rainfall data of the six rain gauge stations. There is no earthquake history record in the study area; therefore, this study does not consider the earthquake as a factor. The soil type in the local area is relatively single. According to the field investigation and the interviews with local people, the soil type is red soil, and the soil depth is not large (usually 0.4 ~ 6 m). Almost all landslide events are located in the area of this type of soil. Therefore, the soil type and texture factors are excluded from the selection process of the conditioning factors. On the other hand, the land-use types are also relatively single, which can be roughly divided into farmland, forest, grassland, shrub forest, bamboo forest and construction land. Most of the local area is covered by forests, and only a small amount of farmland, grassland, shrubs and bamboo forests are distributed around the construction land. The local government has strict protection policies for vegetation and the local forest coverage rate has been maintained at a relatively high level for many years. According to the on-site investigation, almost all landslide events occur in forest-covered areas. The farmland is usually located in an area with a slope of less than 20°, and landslides are difficult to occur under such slope conditions. The grassland, shrubs and bamboo forests are also usually distributed in relatively flat areas around construction land, making landslides almost impossible in such flat areas. Therefore, the location of landslide events in this study is not related to vegetation types and the LU/LC factor is not considered in this study.

3.2.1 Slope angle

The slope of the terrain affects the development of landslide events and determines the type of geological events, and it has an inseparable relationship with geological events. According to statistics, most shallow landslides mainly occur in areas with terrain slopes of 20° ~ 45° (Godt et al. 2008). Although the driving force of the slope movement increases with the increase in the slope angle (Guillard and Zezere 2012), if the slope is too large (> 45°), shallow landslides will hardly occur due to the thin soil layer. On the other hand, if the slope is too small (< 20°), shallow landslides may rarely occur due to insufficient gravity driving force (Godt et al. 2008). In addition, slope gradient greatly affects slope seepage and surface runoff and is directly related to soil moisture content (Magliulo et al. 2008). The slope extracted in local area ranges from 0° to 78° (Fig. 6a).

3.2.2 Curvatures

The susceptibility of landslides is affected by slope shape and topography (Haigh and Rawat 2012). The curvature can describe the contour of the terrain and is often utilized in various geomorphological surveys and studies (Evans 1979). The profile curvature has a direct effect on the surface runoff velocity, while the plane curvature affects the convergence and dispersion of the surface runoff. The plane curvature and profile curvature in this research are in the range of −13.2 to 18.1 and −19.5 to 23.7 (Fig. 6b,6c).

3.2.3 Altitude

According to former studies, the altitude is a vital element in landslide studies. (Ercanoglu and Gokceoglu 2004; Feizizadeh et al. 2014; Tien Bui et al. 2019). The different altitudes of the slopes reflect the different speeds of the crustal uplift in the area, that is, the different influences of geological structures, which will inevitably affect the stability of the slopes. In addition, wind speed, freezing and thawing, rainfall, temperature and other factors will change with the increase in altitude, which will have varying degrees of impact on slope stability. However, the effect of altitude on landslide susceptibility remains a topic for further study. In this paper, altitude map is extracted based on 12.5 m DEM. The altitude of the study area ranges from 70 to 1363 m above sea level, as shown in Fig. 6d.

3.2.4 Slope aspect

Although the aspect has no direct influence on the occurrence of landslides, various studies have shown that there is a correlation between the aspect and regional environmental factors (such as sunlight, soil moisture retention and vegetation cover), which may indirectly affect soil erosion processes on the hillside surface (Van Den Eeckhaut et al. 2009;

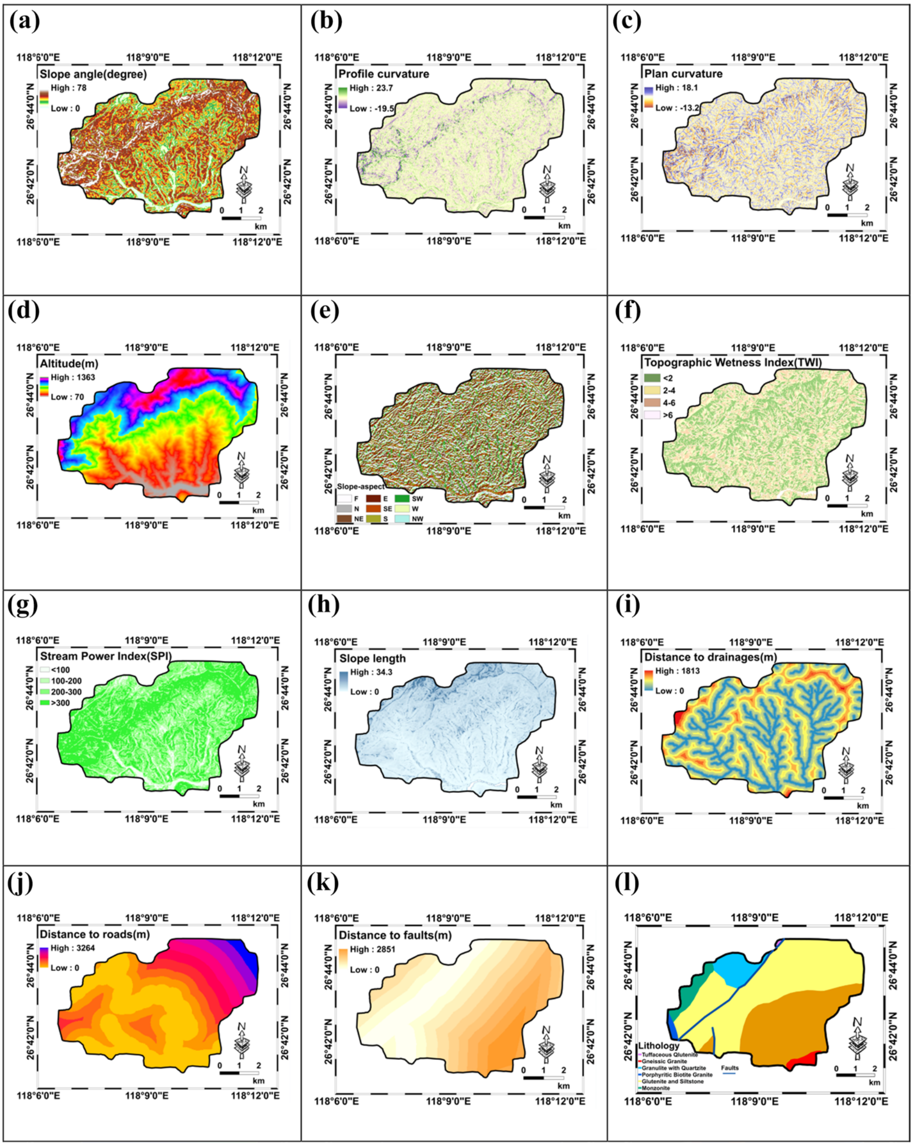


Fig. 6 The LCFs maps used in this study: **a** slope angle, **b** profile curvature, **c** plan curvature, **d** altitude, **e** slope aspect, **f** TWI, **g** SPI, **h** slope length, **i** distance to drainages, **j** distance to roads, **k** distance to faults and **l** lithology

Regmi et al. 2010; Quan and Lee 2012; Devkota et al. 2013). On the other hand, due to different slope directions, slopes have different wind-receiving surfaces, which may affect the weathering of rock and soil. The aspect in the study area is divided into flat (1), north (0 ~ 25; 335 ~ 360), northeast (25 ~ 70), east (70 ~ 115), southeast (115 ~ 160) and south (160 ~ 205), southeast (205 ~ 250), west (250 ~ 300) and northeast (300 ~ 335) (Fig. 6e).

3.2.5 TWI

The topographic wetness index (TWI), which is correlated with soil type and surface runoff (He et al. 2019), has been widely utilized to describe the influence of terrain conditions on the scale and location of a saturation sources. The pore water pressures and soil strength will be affected by the soil moisture, which directly affects the instability of the slope and increases the possibility of landslides. The TWI is shown in Eq. 1.

$$\mathrm{TWI} = \log\left(\frac{\alpha}{\tan\beta}\right) \tag{1}$$

where α is the cumulative upslope area, and β is the slope of that point. The TWI in the study area is shown in Fig. 6f.

3.2.6 SPI

The stream power index (SPI) is an index to quantify water erosion capacity under the assumption that the runoff velocity and flow are proportional to the specific catchment area and slope gradient (Gessler et al. 1995). The larger the specific catchment area and the slope, the greater the amount of water and flow velocity contributed by the upslope area. Therefore, the SPI value increases, which reflects the increased risk of slope erosion. The increased risk of slope erosion means that the possibility of landslides increases, as shown in Eq. 2:

$$\mathrm{SPI} = A_s \times \beta \tag{2}$$

where A_s is the specific catchment area, and β is the local slope. The SPI is shown in Fig. 6g.

3.2.7 Slope length (LS)

Slope length is the slope distance affected by uninterrupted surface water flow. The combination of slope steepness factor and slope length factor affects soil loss and hydrological processes in mountainous areas (Pourghasemi and Rahmati 2018). Slope length, which is considered as a sediment transport capacity index, can quantify the impact of surface runoff velocity on erosion. There are many approaches to determine the LS factor based on grid DEM. One of them is based on the uphill contribution area of each cell, which is extracted from DEM according to Eq. 3 (Moore and Birch 1986):

$$\mathrm{LS} = \left(\frac{A_s}{22.13}\right)^{0.4}\left(\frac{\sin\beta}{0.0896}\right)^{1.3} \tag{3}$$

where As is specific catchment area and β is in degree. The LS in the study area ranges from 0 to 34.3, as shown in Fig. 6h.

3.2.8 Distance to drainages

The proximity of the slope to the drainages has a potential influence on the stability of the slope. The stream may cause erosion of the slope or saturation of the lower part of the material, which adversely affects the stability of the slope (Gokceoglu and Aksoy 1996; Dou et al. 2019a). In order to evaluate the impact of this hydrological-related factor, distance to drainages is considered. The Euclidean distance operation is utilized to extract the distance to drainages in ArcGIS 10.5 software, and the range of result is from 0 to 1813 m (Fig. 6i).

3.2.9 Distance to roads

Human engineering activities (such as mountain road construction and urban construction) will excavate or cut slopes, which causes the original geological conditions to be changed and the natural stability of the slope to decrease. These ergonomics have significant negative effect to the slope stability (Wang et al. 2016). Therefore, distance to roads could be a potential indicator for evaluating slope stability. The distance to roads in the local area has a range from 0 to 3264 m and is developed utilizing the Euclidean distance tool in ArcGIS 10.5 software (Fig. 6j).

3.2.10 Distance to faults

The structural discontinuities, including faults, folds, cracks and joints, play an important role in reducing rock mass strength and increase the risk of landslides (Kanungo et al. 2006). Therefore, the distance to faults may be one of the main influencing factors of the landslide. In this paper, the geological map (1:50,000 scale) is utilized to extract the faults. The Euclidean distance tool in ArcGIS 10.5 is utilized to generate distance to faults map. Distance to faults has a range of 0 ~ 2851 m (Fig. 6k).

3.2.11 Lithology

The lithological characteristics have a certain degree of influence on landslide susceptibility, such as the degree of weathering of different rock masses (Duna et al. 2018). In this study, the geological map of Fujian Province (1:50,000 scale) is utilized to extract the lithology. Six lithological units are identified including, tuffaceous glutenite, porphyritic biotite granite, monzonite, granulite with quartzite, schist, gneissic granite, glutenite and siltstone (Fig. 6l).

3.3 Machine learning models

In recent years, the machine learning technology (MLT) has become an important means to solve modeling problems in the field of natural disasters (Sevgen et al. 2019). The MLT can automatically extract knowledge from huge databases and build classification and regression models. Compared with traditional methods, machine learning models are more efficient and accurate and can perform big data processing and analysis. In this study, five advanced MLTs with different levels of complexity are utilized to generate landslide susceptibility models, including the SVM, RF, CatBoost, XGBoost and LightGBM models.

3.3.1 Random forest (RF)

The RF is a machine learning algorithm that utilizes decision trees for classification and is widely utilized in landslide sensitivity modeling (Breiman 2001; Zhang et al. 2020a, b). Its core idea is to generate a large number of different decision trees through random sampling of samples and random extraction of classification attributes and use decision tree voting to improve the veracity of the model results. The algorithm does not regard all attributes as classification attributes when constructing decision trees, but randomly selects a part of the attribute set as classification attributes, so that there are differences between decision trees. The difference between decision trees is exactly the significance of the random forest algorithm using multiple weak classifiers to vote. The RF has some advantages including (a) it is not prone to the risk of overfitting and has strong ability of anti-noise, and (b) it can handle very high-dimensional data and does not need to make feature selection, and (c) it has strong adaptability to data sets and can handle both discrete data and continuous data, and (d) the training speed of the RF is very fast, and the importance of the output variables can be ranked.

3.3.2 Support vector machine (SVM)

The SVM is a binary classifier based on statistical theory that can be used to identify the best separating hyperplane for dividing two regions (Drucker et al. 1997). The classification function obtained by the SVM has the similar form as the neural network, and its output is a linear combination of several intermediate layer nodes, and each intermediate layer node corresponds to the inner product of the input sample and a support vector. Therefore, it is also called support vector network. Some researchers have applied the SVM method to LSM and provided good prediction results (Kumar and Anbalagan 2016; Kalantar et al. 2018; Colkesen et al. 2016). The four kernel function groups commonly utilized in the SVM include linear kernel (LN), polynomial kernel (PL), radial basis function (RBF) kernel and sigmoid kernel (SIG). In the current study, we use the SVM with RBF, which is the most commonly used kernel function for SVM models to construct landslide susceptibility model (Pourghasemi and Rahmati 2018).

3.3.3 Extreme gradient boost (XGBoost)

XGBoost is a gradient boosting machine (GBM), which grows a tree according to feature splitting and continuously adds trees (Zhang et al. 2020a, b). In fact, every time one tree is added, a new function will be obtained by fitting the residual value of the previous round of predictions and the performance of the model can be significantly improved through iterative calculations. The K trees are obtained through model training, and the leaf node of each tree corresponds to a score, and the predicted value of the sample is the sum of the scores of all trees. Therefore, the goal of the model is to make the predicted value of the tree group as close as possible to the true value through good generalization ability, which is a functional optimization problem from a mathematical point of view. Minimization of the loss function of the training data is a common method to construct the optimal model, as shown in Eq. 4 (Wang et al. 2020).

$$\varphi^t = \sum_{i=1}^{n} l\left(y_i, \left(\hat{y}_{\mathrm{i}}^{(t-1)} + f_t\left(X_i\right)\right)\right) + \Omega\left(f_t\right) \tag{4}$$

where i is the number of a given predicted value $\hat{y}_{\mathrm{i}}(\mathrm{i} = 1, 2, 3, \cdots, \mathrm{n})$; n represents the total number of y values; t is the iteration number; $l\left(y_i, \hat{y}_i\right)$ is the loss function between the actual value y_i and predictive value $\hat{y}_{\mathrm{i}}$; X_i is the features for the i^{th} sample, $f_t\left(X_i\right)$ is the base learner added at the t^{th} iteration; $\Omega\left(f_t\right)$ is the regularization term to prevent overfitting; and φ^t is the objective function at the t^{th} iteration.

As the winner of many Kaggle competitions, the XGBoost has become a reliable algorithm with its excellent capability. However, the XGBoost is rarely utilized in landslide susceptibility modeling. Therefore, in this paper, the XGBoost will be tried to generate landslide susceptibility model.

3.3.4 Light gradient boosting machine (LightGBM)

The LightGBM, which can be utilized for classification and regression tasks, is a distributed gradient boosting framework based on decision tree algorithm (Ke et al. 2017; Ma et al. 2018). Therefore, for the LightGBM which utilizes the optimal leaf-wise strategy to split leaf nodes, the leaf-wise algorithm reduces more loss than the level-wise algorithm when it grows to the same leaf node. On the other hand, the LightGBM uses the histogram algorithm that discretizes continuous floating point eigenvalues into k integers. The optimal split point can be found according to the discrete value of the histogram with width k. In summary, the LightGBM uses the histogram algorithm for feature selection and leaf-wise decision tree growth strategy, which makes its training more efficient, occupies less memory, has higher accuracy and supports parallel learning and large-scale data processing.

3.3.5 Categorical boosting (CatBoost)

The CatBoost, which is a GBDT framework based on symmetric oblivious trees, is originally developed by the Russian company Yandex in 2017 (Prokhorenkova et al. 2018). It has the following advantages: (a) It supports both numerical and categorical features, (b) it has faster GPU and multi-GPU support, (c) it contains visualization tools, and (d) it overcomes gradient bias. In CatBoost, the input sample set is sorted randomly and the average value of the labels is calculated. The average label value obtained through settlement has the same category value. (Huang et al. 2019; Zhang et al. 2020a, b). It is expressed as Eq. 5 (Prokhorenkova et al. 2018):

$$x_{p,k} = \frac{\sum_{j=1}^{p}\left[x_{j,k} = x_{i,k}\right]Y_i}{\sum_{j=1}^{n}\left[x_{j,k} = x_{i,k}\right]} \tag{5}$$

Suppose a set of data with samples $D = \left\{x_i, Y_i\right\}, i = 1, 2, ..n$. If a random permutation is $\sigma = \left(\sigma_1, \cdots, \sigma_2\right)$, the k^{th} feature of the p^{th} data is shown in Eq. 6 (Prokhorenkova et al. 2018):

$$x_{p,k} = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right]Y_{\sigma_j} + \beta P}{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] + \beta} \tag{6}$$

where $x_{i,k}$ is the k^{th} feature of the i^{th} training sample, Y_i is the target variable of the i^{th} sample, P is a prior value (i.e., the average target value in the dataset for a regression task), and β is the weight of the prior value.

3.4 Conditioning factors analysis

There are many factors used in the evaluation of landslide susceptibility. However, the conditioning factors may not be independent in the data set since they are highly correlated, which leads to erroneous results (Dormann et al. 2012). Therefore, some researchers have proposed several methods for the quantization problem of multicollinearity, such as the Pearson correlation matrix (Booth et al. 1994), conditional index (Belsley 1991), variance inflation factors (VIFs) and tolerate (TOL) (Liao and Valliant 2012).

Multicollinearity testing can be utilized to minimize the number of LCFs and reduce high-dimensional data. In this paper, the Pearson correlation matrix, the VIF and TOL are utilized to detect conditioning factors (i.e., multicollinearity) in statistical tests. The VIF and TOL pay attention to the change of the standard error of the landslide influencing factor, which means that the lesser the standard error value, the lower the possibility of multicollinearity, and the more reliable the conditioning factor is to use (Eqs. 7, 8).

$$\mathrm{TOL} = 1 - R^2 \tag{7}$$

$$\mathrm{VIF} = \frac{1}{\mathrm{TOL}} \tag{8}$$

where the R^2 is the proportion of the variance in the target variable.

The Pearson correlation matrix normalizes the contribution of each variable on the basis of covariance, in order to only measure the correlation of variables and not be affected by the scale of other variables. The output range is from −1 to 1. When one variable increases with the increase in another variable, it indicates that there is a positive correlation between them, and the coefficient is greater than 0; if one variable decreases with the increase in another variable, it indicates that there is a negative correlation between them, and the correlation coefficient is less than 0; if the correlation coefficient is equal to 0, it means that there is no correlation between the two variables. In the Pearson correlation coefficient matrix, when the correlation coefficient of two variables exceeds a certain threshold, it means that there is excessive correlation between the two variables, which will cause a large error in the results, so one of the variables should be discarded. The Pearson correlation matrix is defined as the quotient of the covariance and standard deviation between variables (Eq. 9):

$$r_{x.y} = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{\sqrt{\Sigma_{\mathrm{k}=1}^{n}\left(X_i - \overline{X}\right)^2}} \times \frac{y_i - \overline{y}}{\sqrt{\Sigma_{\mathrm{k}=1}^{n}\left(y_i - \overline{y}\right)^2}} \tag{9}$$

where n is the number of samples;x_i, y_i are conditioning factors indexed with i; and x is the mean of x_i, where: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and analogously same applies to $\overline{y}$. The VIF value is > 5, and TOL value < 0.2 or the Pearson correlation matrix is greater than 0.7 (Dou et al. 2020), which indicates that the predisposing factors have multicollinearity problems (O'Brien 2007; Dormann et al. 2012).

3.5 Model metrics

According to former research results, there are many statistical indicators that can evaluate the performance of MLTs and check the predictive accuracy of the developed LSM (Pham et al. 2019a; Tien Bui et al. 2019; Moradi et al. 2019; Yousefi et al. 2020; Ghasemian et al. 2020). In this study, the area under the curve of ROC (AUC), the accuracy rate (ACC), F1 score, Kappa index are utilized to assess the capabilities of MLTs. True positive (TP) and true negative (TN) are utilized to indicate the number of pixels correctly classified as landslide and non-landslide, respectively; the false positive (FP) and false negative (FN) are used to indicate the number of pixels that are misclassified as landslides and non-landslides.

The accuracy rate, which refers to the proportion of the number of correctly classified records to the total number of records, is the most commonly used indicator in the evaluation of MLTs (Eq. 10).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

F1 score refers to the harmonic mean of accuracy and recall. It is often used in statistics to measure the accuracy of a two-class model. The value of F1 score ranges from 0 to 1. A model is reliable if the value of F1 score is close to 1. It can be formulated as follows (Eqs. 11, 12, 13):

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$recall = \frac{TP}{TP + FN} \tag{13}$$

In this study, another used evaluation indicator is the Kappa index, which is a method used in statistics to assess consistency and accuracy of classification models. The value range of the Kappa index is [−1,1]; if the value is close to −1, the model is unreliable. While the closer the value is to 1, the more reliable the model is. The formula is as follows (Eqs. 14, 15, 16):

$$\text{Kappa} = \frac{P_a - P_{\exp}}{1 - P_{\exp}} \tag{14}$$

$$P_a = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{15}$$

$$P_{\exp} = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP}) + (\text{FP} + \text{TN})(\text{FN} + \text{TN})}{(\text{TP} + \text{TN} + \text{FN} + \text{FP})^2} \tag{16}$$

The range of Kappa calculation results is from −1 to 1, but usually the value is between 0 and 1. The Kappa index result is less than 0, indicating that the consistency of the model is poor. On the other hand, the values can be divided into five groups to indicate different levels of consistency: 0 ~ 0.2, 0.2 ~ 0.4, 0.4 ~ 0.6, 0.6 ~ 0.8 and 0.8 ~ 1 represent slight, fair, moderate, substantial and almost perfect, respectively.

In the studies of landslide susceptibility assessment, the ROC curve is a common method utilized to evaluate the capability of machine learning model (Beguería 2006; Mathew et al. 2009). The ROC is a comprehensive indicator reflecting susceptibility and specificity of continuous variables (Hanley et al. 1982; Avand et al. 2022). It calculates a series of susceptibility and specificity through setting many different critical values for continuous variables and then draws a curve with susceptibility as the y-axis and 1-specificity as the x-axis. The more the area under the curve, the better the presentation of the model becomes. The exactness group rankings are as follows: 0.9 ~ 1, 0.8 ~ 0.9, 0.7 ~ 0.8, 0.6 ~ 0.7, 0.5 ~ 0.6 represent excellent, good, fair, poor, fail, respectively (Eqs. 17, 18) (Hosmer and Lemeshow 2000).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{18}$$

The area under the curve (AUC) has been widely utilized as a popular tool in the classification performance. It is utilized to quantify the results of the ROC curve model (Mason and Graham 2002). The AUC value is positively correlated with the performance of the model. The larger the AUC value, the more excellent the performance of the model. But AUC is less than or equal to 0.5 means that the model is unreliable (Youssef 2016).

4 Results and analysis

4.1 Conditioning factors analysis

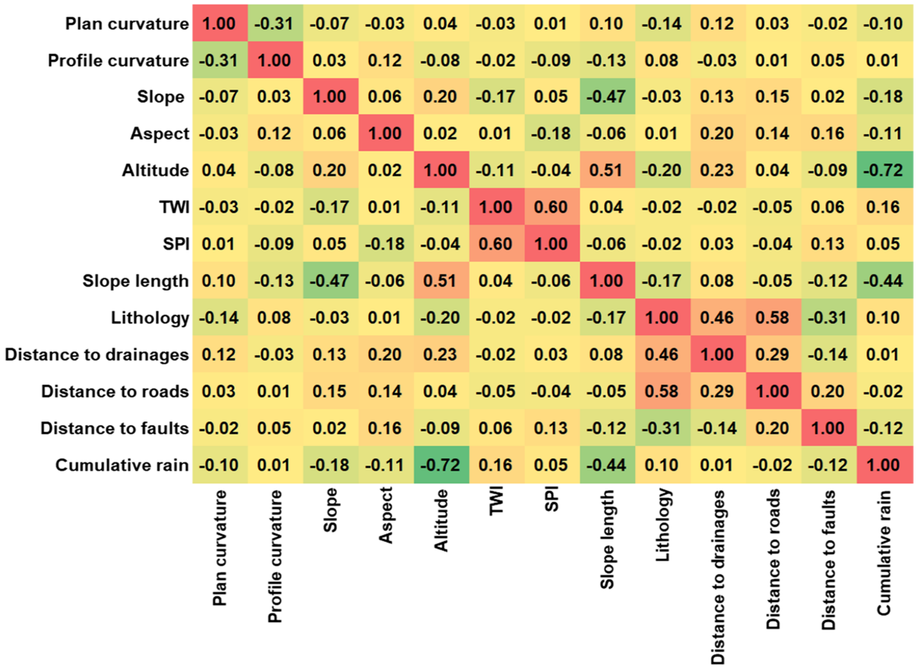
In this paper, the VIF, TOL and the Pearson correlation matrix are utilized to test whether there is multicollinearity between the conditional factors. According to the research results, the highest VIF value is 4.201 (< 5) and the lowest TOL value is 0.238 (> 0.2), which shows that there is no multicollinearity relationship between these 13 factors (Table 1).

Based on the output of the Pearson correlation matrix, there are positive or negative correlations among the 13 selected variables, but none of their values exceed the allowable threshold of correlation 0.7 (Fig. 7) (Dou et al. 2020). The Pearson correlation coefficient matrix is designed to reflect the relationship between variables in terms of data, not mechanism. For example, parameters such as the altitude, cumulative rain, distance to roads, SPI, TWI and lithology have good relationships with each other. Among them, for the local area, the altitude is negatively correlated with cumulative rainfall, which may be related to the long-term rainfall type in the study area, while SPI and TWI, lithology and distance to roads, slope length and elevation, lithology and distance to drainages all have a certain degree of positive correlation. All in all, all the above results show that the above variables utilized in LSM studies are valid.

The novel method CatBoost is utilized to evaluate the importance of all selected influencing factors (Fig. 8). The results show that the distance to faults, distance to drainages and slope are the top three important conditioning factors. The fault is a geological

Table 1 Multicollinearity analysis for the 13 LCFs

Number	Landslide conditioning factor	Tolerance	VIF
1	Plan curvature	0.785	1.273
2	Profile curvature	0.773	1.294
3	Slope	0.382	2.616
4	Aspect	0.846	1.183
5	Altitude	0.238	4.201
6	TWI	0.565	1.771
7	SPI	0.564	1.772
8	Slope length	0.248	4.036
9	Lithology	0.943	1.061
10	Distance to drainages	0.416	2.405
11	Distance to roads	0.757	1.321
12	Distance to faults	0.493	2.028
13	Cumulative rain	0.358	2.793



	Plan curvature	Profile curvature	Slope	Aspect	Altitude	TWI	SPI	Slope length	Lithology	Distance to drainages	Distance to roads	Distance to faults	Cumulative rain
Plan curvature	1.00	-0.31	-0.07	-0.03	0.04	-0.03	0.01	0.10	-0.14	0.12	0.03	-0.02	-0.10
Profile curvature	-0.31	1.00	0.03	0.12	-0.08	-0.02	-0.09	-0.13	0.08	-0.03	0.01	0.05	0.01
Slope	-0.07	0.03	1.00	0.06	0.20	-0.17	0.05	-0.47	-0.03	0.13	0.15	0.02	-0.18
Aspect	-0.03	0.12	0.06	1.00	0.02	0.01	-0.18	-0.06	0.01	0.20	0.14	0.16	-0.11
Altitude	0.04	-0.08	0.20	0.02	1.00	-0.11	-0.04	0.51	-0.20	0.23	0.04	-0.09	-0.72
TWI	-0.03	-0.02	-0.17	0.01	-0.11	1.00	0.60	0.04	-0.02	-0.02	-0.05	0.06	0.16
SPI	0.01	-0.09	0.05	-0.18	-0.04	0.60	1.00	-0.06	-0.02	0.03	-0.04	0.13	0.05
Slope length	0.10	-0.13	-0.47	-0.06	0.51	0.04	-0.06	1.00	-0.17	0.08	-0.05	-0.12	-0.44
Lithology	-0.14	0.08	-0.03	0.01	-0.20	-0.02	-0.02	-0.17	1.00	0.46	0.58	-0.31	0.10
Distance to drainages	0.12	-0.03	0.13	0.20	0.23	-0.02	0.03	0.08	0.46	1.00	0.29	-0.14	0.01
Distance to roads	0.03	0.01	0.15	0.14	0.04	-0.05	-0.04	-0.05	0.58	0.29	1.00	0.20	-0.02
Distance to faults	-0.02	0.05	0.02	0.16	-0.09	0.06	0.13	-0.12	-0.31	-0.14	0.20	1.00	-0.12
Cumulative rain	-0.10	0.01	-0.18	-0.11	-0.72	0.16	0.05	-0.44	0.10	0.01	-0.02	-0.12	1.00

Fig. 7 The output results of the Pearson correlation matrix

structure that undergoes significant relative displacement along both sides under the action of the crustal force. The fault area tends to have strong crustal activity. Therefore, the tectonic breaks including faults play an important role in decreasing rock strength and causing landslides. In addition, runoff along drains enhances the undercut phenomenon, thereby increasing pore water pressure in the area near the drain, which plays an important role in triggering landslides. The slope factor is still the main factor affecting landslides. The

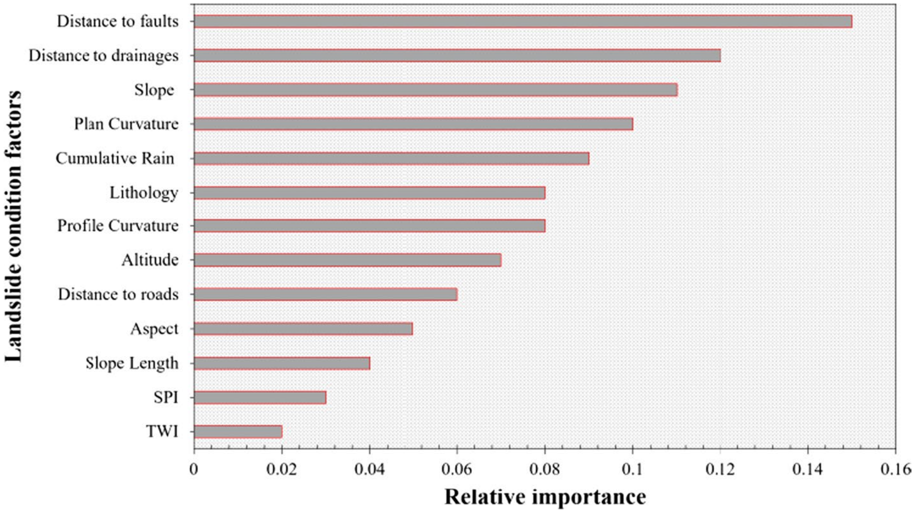


Fig. 8 Important order of conditioning factors according to CatBoost method

greater the slope angle, the greater the driving force for mass movement, which leads to an increased possibility of landslides. Other factors such as cumulative rainfall and lithology also have a great influence on the susceptibility model, which shows that the rainfall, as a triggering factor for landslides, plays an important role in controlling the occurrence of geological events. In the study area, landslide events are more likely to occur in metamorphic rock areas with relatively fragile lithology. The degree of weathering of different parent rocks roughly determines the type, probability and scale of landslides. The soil structure formed after weathering in metamorphic rock areas is relatively loose, which is conducive to rainwater infiltration and is easy to soften in contact with water. The above reasons lead to the reduction of shear strength. The interface between the upper loose soil mass and the underlying relatively stable bedrock is a potential structurally weak surface, which is very easy to slide along the rock–soil interface, resulting in landslides and collapses. However, compared to the other categories, SPI and TWI are the two least important categories. The SPI and TWI factors have strong control effects on soil erosion and material transport. In areas with barren or sparse vegetation, the soil physical and chemical properties are poor and the erosion resistance of the soil is significantly weakened, which leads to the most prone to soil erosion. The root system of vegetation can hold the soil, which can enhance the shear strength of the soil and improve the stability of the slope. On the other hand, vegetation with high coverage can improve hydrological conditions and soil properties. The tree canopy can intercept rainfall, which reduces the flow rate of surface water and groundwater on slopes, thereby reducing the occurrence of landslides. Fujian, as the province with the largest forest coverage area in China, is less affected by soil erosion on the hillside surface.

4.2 Landslide susceptibility models

The RF, CatBoost, SVM, XGBoost and LightGBM models are successfully generated utilizing the training landslide datasets, and susceptibility index values for every pixel (12.5 m × 12.5 m) in study area are obtained by calculation. The natural breaks classifier can identify breakpoints by picking the categorical breakpoints that best group similar values and maximize the differences between classes (Ayalew et al. 2004; Suzen and Doyuran 2004). In this paper, the natural breaks classifier method is utilized to reclassify the LSMs into four susceptibility levels including low, medium, high and extremely high. For comparative analysis, Fig. 9 shows the five LSMs provided by the RF, CatBoost, SVM, XGBoost and LightGBM models. The relative area ratio calculations for the susceptibility categories for each model are shown in Fig. 10.

When all maps based on MLTs are analyzed visually, the central and western regions of the study area are at high and very high susceptibility levels, while most of the eastern and northeastern regions are at medium and low susceptibility levels (Fig. 10). The reason may be that there are two large faults in the west and the terrain in the west of the study area is significantly steeper than in the east (Fig. 7a; Fig. 7l). All LSMs show that the southwestern of the local area has a very high susceptibility level. The LSM using the LightGBM model and the RF model classifies more northern and central regions into very high susceptibility level. The LSM of the XGBoost model and the SVM model divides more northwestern and southern regions into low susceptibility level. The LSM using the CatBoost model seems to show more detailed classification results, with more high and medium susceptibility areas in the central and eastern regions.

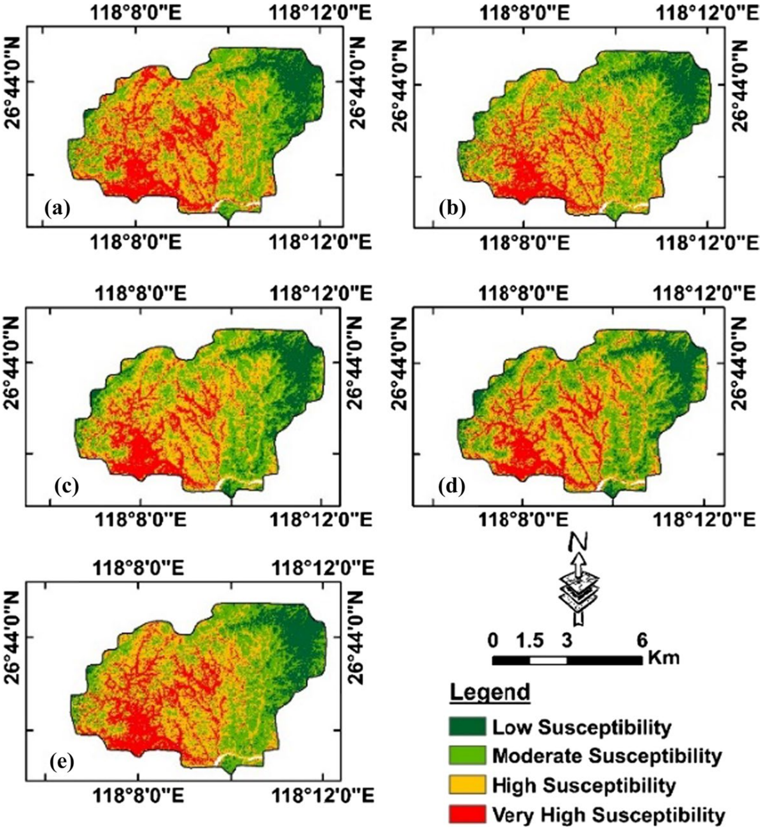


Fig. 9 Landslide susceptibility maps using: **a** RF, **b** CatBoost, **c** SVM, **d** XGBoost and **e** LightGBM

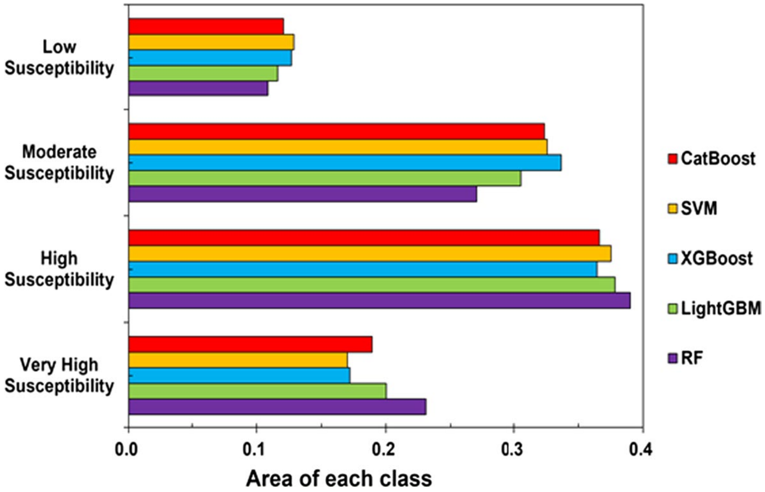


Fig. 10 Landslide susceptibility classes' areas percentage for all applied MLTs

Figure 10 shows the divided areas of the LSMs prepared by MLTs. Different LSMs have different proportions of grade areas. Among them, a total of 11 ~ 13% of the areas are classified as low susceptibility, 27 ~ 34% are moderately susceptible areas, and the remaining 54% ~ 62% are considered to have high and very high susceptibility. In the study area, the areas with high and very high susceptibility accounted for more than 50%, while areas with low susceptibility accounted for less than 15%, which also further explains why the study area is the most severely affected area in the 618 event.

4.3 Model performance and comparison

The AUC, ACC, Kappa index and F1 score are utilized to evaluate the model performance (Table 2). The LSMs generated by five models are compared and evaluated. The ROC curves and AUC values of the training and testing datasets are utilized to compare and evaluate the results of the process. Figure 11 shows the comparison and evaluation results of LSMs. The results from the prediction rate curve (Fig. 11b) show that the CatBoost (AUC = 0.917, ACC = 0.892) has the most excellent capability. Followed by the SVM (AUC = 0.894, ACC = 0.873), XGBoost (AUC = 0.881, ACC = 864), LightGBM (AUC = 0.852, ACC = 0.826) and RF (AUC = 0.848, ACC = 0.821) models. The performance results show that all models produce very good results (AUC > 0.750). The Kappa

Table 2 Performance results of the implemented models

Number	Landslide models	AUC	ACC	F1 score	Kappa index
1	SVM	0.894	0.873	0.885	0.744
2	RF	0.848	0.821	0.843	0.639
3	XGBoost	0.881	0.864	0.879	0.723
4	LightGBM	0.852	0.826	0.844	0.646
5	CatBoost	0.917	0.892	0.905	0.781

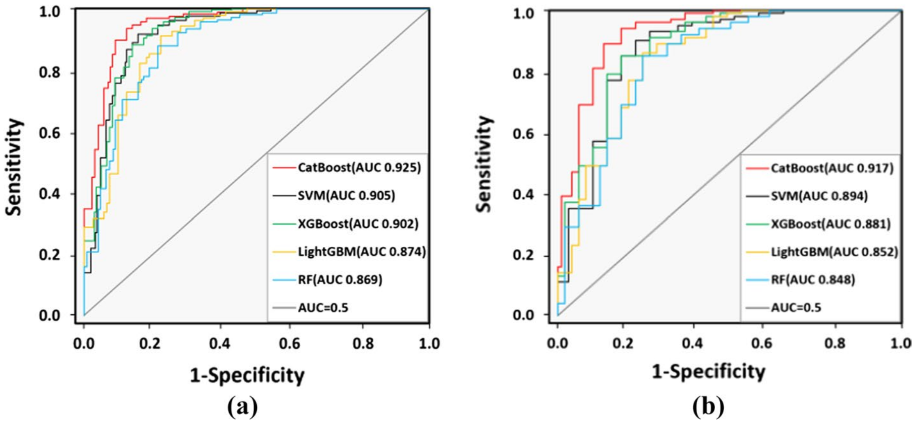


Fig. 11 **a** ROC curves and AUC for the five models with training dataset. **b** ROC curves and AUC for the five models with testing dataset

index varies from 0.639 to 0.781, and the F1 score varies from 0.843 to 0.905. Based on these results, there is a strong correlation between the investigated landslide areas and the predicted landslide areas. In general, the capability of the CatBoost is significantly outstanding than other models; followed by the SVM, XGBoost and LightGBM, RF achieved the lowest performance among the five implemented models.

5 Discussion

The selection of LCFs and the determination of their categories have a crucial impact on the quality of the landslide susceptibility model (Irigaray et al. 2007). Although many researchers have proposed some methods for selecting condition factors, such as Goodman–Kruskal and Kolmogorov–Smirnov test (Fernández et al. 2003) and GIS matrix combination method (Cross 2002), the criteria for the choice of condition factors are still controversial. In most of the previous studies, many researchers believe that the slope angle, altitude and curvature are the most important factors affecting the occurrence of landslides. (Pham et al. 2019b; Merghadi et al. 2018; Hadji et al. 2013). However, a growing body of research demonstrates that the relative importance of LCF in the model tends to correlate with characteristics of the study area (Park 2015). For example, in a study about landslide susceptibility in the mountainous area of Gangwon-do, Korea, researchers find that the SPI, slope length, slope aspect, slope and TWI have a positive impact on landslide susceptibility in the Inje region. However, in the Pyeongchang region, the aspect, land use, SPI, slope and TWI have relatively positive effects on the landslide susceptibility map (Lee et al. 2017). In the study about landslide susceptibility in the Abha Basin in the Asir region of Saudi Arabia, the results show that the conditioning factors such as slope, aspect, length, distance to roads and distance to rivers contribute significantly to landslide susceptibility, while the plane curvature, profile curvature and LU/LC have little effect on landslide susceptibility (Youssef et al. 2021). In a regional-scale landslide susceptibility case study on the volcanic islands of Izu-Oshima, Japan, the researchers find that the top two most important factors that predispose to landslides are lithology and rainfall (Dou et al. 2019b). From previous studies, it can be concluded that the importance of conditioning factors varies with different study areas. According to the results of this study, tectonic faults may have an extremely important influence on the occurrence of landslides due to the specific tectonic setting of the local area (Fig. 8). The terrain in the central and western regions of the study area is significantly steeper than that in the east, and there are two larger fault zones in the western region. The development degree of geological events is obviously affected by the fault structure, which directly affects the main section and surrounding rock masses. The lithology of the strata on both sides of the fault zone is subjected to strong mechanical action, so that the width of the fractured rock zone and the fractured development zone is several meters or even more than ten meters. The rock mass has low strength and strong water permeability and is prone to various geological events under rainfall conditions. The above reasons may cause landslides in the central and western regions to be more significant than in the eastern regions. Based on previous research (Pourghasemi and Rahmati 2018), the slope angle, distance to roads, distance to faults and distance to drainages are the first four important factors in landslide susceptibility prediction. The results obtained in the importance analysis of this paper are consistent with those of previous studies. Slope angle, distance to faults and distance to drainages are indeed the relatively most important categories. The slope angle factor is still the relatively main factor leading to the occurrence of landslides. The increase in the slope angle will lead to an increase in the driving force of the soil movement, resulting in a greater probability of landslides (Guillard and Zezere 2012; Tien Bui et al. 2017). In addition, the landslide susceptibility is also greatly affected by rainfall factors and lithology factors, which indicates that rainfall, as an inducing factor leading to the occurrence of landslides, plays a significant role in the occurrence of geological events. The landslide events mostly occur in glutenite, siltstone and schist areas, but rarely occur in granite areas. The reason may be that igneous rocks are generally more resistant to weathering than metamorphic rocks and sedimentary rocks. However, compared with other categories, the SPI and TWI are the two least important categories, and Pradhan et al. also obtain similar conclusions in landslide susceptibility studies in Deokjeokri and Karisanri watersheds (Pradhan et al. 2019). The reason for this may be that areas with high vegetation cover are less susceptible to soil erosion than areas with poor or sparse vegetation. The roots of vegetation can increase the shear strength of the soil by stabilizing the moisture in the soil to protect the soil. The tree canopy can intercept rainfall, which reduces the flow rate of surface water and groundwater on slopes, thereby reducing the occurrence of landslides.

On the other hand, in recent years, the gradient boosting algorithm has received extensive attention from more and more researchers. Although previous studies have shown that the RF algorithm performs well in some research fields (Shrestha et al. 2017; Pham et al. 2019b; Lagomarsino et al. 2017). However, with the increasing use of gradient boosting algorithms for landslide susceptibility studies (Kim et al. 2018; Lombardo et al. 2015; Song et al. 2019). The researchers find that the gradient boosting algorithms perform better than the RF model. Furthermore, the selection of model type depends on the specific study area. The landslide susceptibility model depends on the variables utilized for implementation. Due to the different implementation variables of the model, the applicability of the model may be different. For example, the results of related studies have shown that some models (such as RF) have excellent performance in certain domains, but mediocre performance in others (Hong et al. 2016). According to previous studies, in the multi-hazard susceptibility study of Jiuzhaigou, some researchers have proved that the capability of XGBoost is more excellent than RF through the results of the AUC evaluation index. (Cao et al. 2020). In the study of debris flow susceptibility in Shigatse, Tibet, the researchers compare the prediction performance of BPNN, XGBoost and RF algorithms and conclude that the XGBoost algorithms has the best prediction performance according to the results (Zhang et al. 2019a, b). In this study, results for the predicted value of RF (AUC = 0.853) are much lower than those obtained in the related literature (Youssef 2016). In addition, the predicted values of XGBoost and SVM are also slightly higher than those obtained in related studies (Cao et al. 2020; Lee et al. 2017), and we believe that these differences in prediction rates are likely due to the specific characteristics of each study area. Therefore, having a clear understanding of the differences between the various MLTs is very beneficial for selecting the best model for a specific region. The CatBoost and LightGBM, as new gradient boosting algorithms based on decision trees (Prokhornkova et al. 2018, Ke et al. 2017), have many advantages and have been utilized in some industries. In this paper, four indicators, AUC, ACC, Kappa index and F1 score, are used for model evaluation. The output results show that the performance of SVM in landslide susceptibility evaluation is slightly better than that of XGBoost, and the obtained findings are consistent with previous studies (Cao et al. 2020). But the calculation time of SVM algorithms is much longer than that of the XGBoost algorithms. Nevertheless, when the results of the two algorithms are not much different, the XGBoost model should still be the preferred algorithm, because the XGBoost with regularization has obvious advantages over SVM in solving the overfitting problem (Yao et al. 2008, Cheng et al. 2018). The prediction capability of the CatBoost model is the best, followed by SVM, XGBoost, LightGBM and RF. The CatBoost model significantly outperformed the other models in overall performance, while the RF model had the weakest predictive capability among these models. Similar conclusions are reached by previous research, the LightGBM and CatBoost are utilized for the first time to predict flash flood susceptibility (FFS) of the Wadi system (Hurghada, Egypt) (Saber et al. 2021). The results show that LightGBM and CatBoost prove to be more effective in flash flood prediction in arid regions. Compared with the RF algorithm based on the set of independently voted trees, the CatBoost model relies on the gradient boosting method to improve the model accuracy. Although different parameters may lead to different prediction accuracy, according to the research results, CatBoost, as a relatively novel machine learning algorithm, can generate LSMs more efficiently and robustly. And the LightGBM, by using histogram-esque algorithm to achieve extremely fast calculation speed and extremely low memory footprint may be more suitable for processing larger datasets. Therefore, the relatively small amount of data may be the main reason for the poor performance of this algorithm in this study. In this paper, these five MLTs are utilized for the first time in landslide susceptibility studies in the hilly area of Fujian Province, China. Furthermore, few related papers deal with the performance comparisons of novel gradient boosting algorithms (such as CatBoost and LightGBM) with the previously popular MLTs. As far as the current research results are concerned, due to the robustness of these novel gradient boosting algorithms, the gradient boosting algorithms may be more accurate for landslide susceptibility studies and may show better prospects in the future.

There are also some limitations in this paper, which need further research. Firstly, this study only selects the most severely affected area in the 618 event as the research object, and the area is relatively limited. Although this approach increases the pertinence of the results, whether the model is applicable to a wider range of areas is still worthy of further research. Secondly, changes in the geological environment in the region will exacerbate the evolution of unstable geological events. Therefore, it is necessary to regularly update the landslide list and adjust the factors. Thirdly, the results of factor importance analysis in this paper can be utilized as a reference for the evaluation of landslide susceptibility in other hilly areas in Fujian Province. But for other hilly areas, whether the reference factor is reliable remains to be further research. Finally, landslide susceptibility should be analyzed in the future with more detailed geological information or high-resolution datasets, which may lead to more accurate results.

6 Conclusions

LSM maps can be used as information maps for environmental management, land-use planning and infrastructure development. Therefore, it is important to generate a robust and accurate model to reduce errors in landslide susceptibility assessment. The novelty of this study is that these machine learning methods (RF, SVM, CatBoost, LightGBM and XGBoost) are utilized for the first time in landslide susceptibility studies in the hilly area of Fujian, China. Moreover, there are few related papers that incorporate such novel gradient boosting algorithms such as CatBoost and LightGBM which are compared with previous studies. This study is based on the landslide events that occur in the 618 event and 13 landslide conditioning factors. The CatBoost algorithm is utilized to analyze the importance of the conditioning factors leading to the landslide events. According to the results of the factor importance analysis, the slope, distance to faults and distance to drainages are the top three most important categories. The results of various evaluation indicators show that the prediction capability of CatBoost is the best, followed by SVM, XGBoost, LightGBM and RF models. The Kappa index ranges from 0.639 to 0.781, and the F1 score ranges from 0.843 to 0.905. On the other hand, these results indicate that the observed landslides are in strong agreement with the predicted landslides. While all five models are suitable for landslide modeling, the CatBoost model has significantly better prediction accuracy, which can improve the reliability of the landslide susceptibility map. The landslide susceptibility maps generated utilizing high-accuracy MLT are critical for policymakers, planners and engineers to identify landslide-prone areas, prevent and mitigate landslide risks, identify suitable land-use planning areas and establish early warning systems. According to the distribution of the landslide susceptibility map in this paper, the high- prone areas of landslide are mainly located in the southwest, especially near the two faults. Therefore, the local government should organize local residents to relocate to low-risk areas in the east. Moreover, they should also pay close attention to these high-risk areas and take positive and effective measures, such as rainstorm monitoring and early warning, to reduce losses caused by landslides in the future. Although CatBoost, XGBoost and LightGBM have been utilized to some extent in other fields, they are still rarely utilized in landslide susceptibility assessment research, so further research is needed to verify the capability of these algorithms on different types of landslide datasets. In addition, it requires further investigation on whether these new models can be successfully utilized in other similar geological and rainfall-triggered cases or regions around the world. However, as far as the current research results are concerned, we can be sure that the CatBoost method is significantly better than the other four. As for whether this result depends on the datasets used, more follow-up research is needed. In conclusion, the CatBoost model is a very promising method in landslide research, which also means that for landslide susceptibility research, the gradient boosting algorithm may achieve better results and show better prospects in the future.

Acknowledgements This work is supported by the National Natural Science Foundation of China (NSFC, contract number: U21A2032) and Fujian Science and technology disaster prevention project (Fujian Department of Natural Resources [2020], No. 46).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Are K, Tien Bui D, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic 52:394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

Avand M, Kuriqi A, Khazaei M, Ghorbanzadeh O (2022) DEM resolution effects on machine learning performance for flood probability mapping. J Hydro-Environ Res 40:1–16. https://doi.org/10.1016/j.jher.2021.10.002

Avand M, Moradi H, Lasboyee MR (2021) Spatial prediction of future flood risk: an approach to the effects of climate change. Geosciences 11(1):25–45. https://doi.org/10.3390/geosciences11010025

Ayalew L, Yamagishi H, Ugawa N (2004) Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. Landslides 1:73–81. https://doi.org/10.1007/s10346-003-0006-9

Azizi V, Hu GP (2020) Multi-product pickup and delivery supply chain design with location-routing and direct shipment. Int J Prod Econ 226:107648. https://doi.org/10.1016/j.ijpe.2020.107648

Beguería S (2006) Validation and evaluation of predictive models in hazard and risk assessment. Nat Hazards 37:315–329. https://doi.org/10.1007/s11069-005-5182-6

Belsley D (1991) A guide to using the collinearity diagnostics. Comput Sci Econ Manage 4:33–50. https://doi.org/10.1007/BF00426854

Booth GD, Niccolucci MJ, Schuster EG (1994) Identifying proxy sets in multiple linear regression:an aid to better coefficient interpretation. US Dept of Agric For Serv Ogden. https://doi.org/10.2170/jjphysiol.50.463

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Camilo DC, Lombardo L, Mai PM, Dou J, Huser R (2017) Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSOpenalized Generalized Linear Model. Environ Model Softw 97:145–156. https://doi.org/10.1016/j.envsoft.2017.08.003

Carrara A, Pike RJ (2008) GIS technology and models for assessing landslide hazard and risk. Geomorphology 94:257–260. https://doi.org/10.1016/j.geomorph.2006.07.042

Cao J, Zhang Z, Du J, Zhang LL, Song Y, Sun G (2020) Multi-geohazards susceptibility mapping based on machine learning-a case study in Jiuzhaigou. China Nat Hazards 102(3):851–871. https://doi.org/10.1007/s11069-020-03927-8

Cheng S, Zhang S, Li L, Zhang D (2018) Water quality monitoring method based on TLD 3D fish tracking and XGBoost. Math Probl Eng 7:1–12. https://doi.org/10.1155/2018/5604740

Colkesen I, Sahin EK, Kavzoglu T (2016) Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. J Afr Earth Sci 118:53–64. https://doi.org/10.1016/j.jafrearsci.2016.02.019

Comert R, Avdan U, Gorum T, Nefeslioglu HA (2019) Mapping of shallow landslides with object-based image analysis from unmanned aerial vehicle data. Eng Geol 3:105264. https://doi.org/10.1016/j.enggeo.2019.105264

Conoscenti C, Ciaccio M, Caraballo-Arias NA, Gomez-Gutierrez A, Rotigliano E, Agnesi V (2015) Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: a case of the Belice River basin (western Sicily, Italy). Geomorphology 242:49–64. https://doi.org/10.1016/j.geomorph.2014.09.020

Constantin M, Bednarik M, Jurchescu MC, Vlaicu M (2011) Landslide susceptibility assessment using the bivariate statistical analysis and the index of entropy in the Sibiciu Basin (Romania). Environ Earth Sci 63:397–406. https://doi.org/10.1007/s12665-010-0724-y

Costache R, Ţîncu R, Elkhrachy I, Pham QB, Popa MC, Diaconu DC, Avand M, Costache I, Arabameri A, Tien Bui D (2020) New neural fuzzy-based machine learning ensemble for enhancing the prediction accuracy of flood susceptibility mapping. Hydrol Sci J 65(16):2816–2837. https://doi.org/10.1080/02626667.2020.1842412

Cross M (2002) Landslide susceptibility mapping using the matrix assessment approach: a Derbyshire case study. Eng Geol Spec Publ 15:247–261. https://doi.org/10.1144/GSL.ENG.1998.015.01.26

Devkota KC, Regmi AD, Pourghasemi HR, Yoshida K, Pradhan B, Ryu IC, Dhital MR, Althuwaynee OF (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling-Narayanghat road section in Nepal Himalaya. Nat Hazards 65(1):135–165. https://doi.org/10.1007/s11069-012-0347-6

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2012) Collinearity: a reviewof methods to deal with it and a simulation study evaluating their performance. Ecography (Cop) 36:27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dou J, Yunus AP, Tien Bui D (2019a) Evaluating GIS-based multiple statistical models and data mining for earthquake and rainfall-induced landslide susceptibility using the LiDAR DEM. Remote Sens 11:638. https://doi.org/10.3390/rs11060638

Dou J, Yunus AP, Tien Bui D, Merghadi A, Mehebub S, Zhu ZF, Chen CW, Khosravi K, Yang Y, Pham BT (2019b) Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. Sci Total Environ 662:332–346. https://doi.org/10.1016/j.scitotenv.2019.01.221

Dou J, Yunus AP, Merghadi A, Shirzadi A, Nguyen H, Hussain Y, Avtar R, Chen YL, Pham BT, Yamagishi H (2020) Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. Sci Total Environ 720:137720. https://doi.org/10.1016/j.scitotenv.2020.137320

Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: Advances in Neural Information Processing Systems, pp 155–161

Duna CR, D'Arcy M, McDonald J, Whittaker CA (2018) Lithological controls on hillslope sediment supply: Insights from landslide activity and grain size distributions. Earth Surf Process Landf. https://doi.org/10.1002/esp.4281

Ercanoglu M, Gokceoglu C (2004) Use of fuzzy relations to produce landslide susceptibility map of a landslide prone area (West Black Sea Region, Turkey). Eng Geol 75:229–250. https://doi.org/10.1016/j.enggeo.2004.06.001

Evans IS (1979) An integrated system of terrain analysis and slope mapping. Zeitschrift Fur Geomorphologie 36:274–295. https://doi.org/10.3987/R-1985-01-0033

Feizizadeh B, Blaschke T, Nazmfar H (2014) GIS-based ordered weighted averaging and Dempster-Shafer methods for landslide susceptibility mapping in the Urmia Lake Basin. Iran Int J Digital Earth 7(8):688–708. https://doi.org/10.1080/17538947.2012.749950

Fernández T, Irigaray C, El Hamdouni R, Chacón J (2003) Methodology for landslide susceptibility mapping by means of a GIS. Application to the Contraviesa Area (Granada, Spain). Nat Hazards 30:297–308. https://doi.org/10.1023/B:NHAZ.0000007092.51910.3f

Froude MJ, Petley DN (2018) Global fatal landslide occurrence from 2004 to 2016. Nat Hazards Earth Syst Sci 18:2161–2181. https://doi.org/10.5194/nhess-18-2161-2018

Gessler PE, Moore ID, McKenzie NJ, Ryan PJ (1995) Soil-landscape modelling and spatial prediction of soil attributes. Int J Geogr Inf Syst 9:421–432

Ghasemian B, Asl DT, Pham BT, Avand M, Nguyen HD, Janizadeh S (2020) Shallow landslide susceptibility mapping: A comparison between classification and regression tree and reduced error pruning tree algorithms. Vietnam J Earth Sci 42(3):208–227. https://doi.org/10.15625/0866-7187/42/3/14952

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. Rem Sens. 11(2):196–217. https://doi.org/10.3390/rs11020196

Godt JW, Baum RL, Savage WZ (2008) Transient deterministic shallow landslide modeling: requirements for susceptibility and hazard assessments in a GIS framework. Eng Geol 102:214–226. https://doi.org/10.1016/j.enggeo.2008.03.019

Gokceoglu C, Aksoy H (1996) Landslide susceptibility mapping of the slopes in the residual soils of the Mengen region (Turkey) by deterministic stability analyses and image processing techniques. Eng Geol 44:147–161. https://doi.org/10.1016/S0013-7952(97)81260-4

Guillard C, Zezere J (2012) Landslide susceptibility assessment and validation in the framework of municipal planning in Portugal: the case of Loures Municipality. Environ Manage 50:721–735. https://doi.org/10.1007/s00267-012-9921-7

Guzzetti F, Peruccacci S, Rossi M, Stark CP (2007) Rainfall thresholds for the initiation of landslides in central and southern Europe. Meteorol Atmos Phys 98(3):239–267. https://doi.org/10.1007/s00703-007-0262-7

Guzzetti F, Mondini AC, Cardinali M, Fiorucci F, Santangelo M, Chang K (2012) Landslide inventory maps: new tools for an old problem. Earth-Science Rev 112:42–66. https://doi.org/10.1016/j.earscirev.2012.02.001

Hadji R, Boumazbeur AE, Limani Y, Baghem M, Chouabi AEM, Demdoum A (2013) Geologic, topographic and climatic controls in landslide hazard assessment using GIS modeling: a case study of Souk Ahras region, NE Algeria. Quat Int 302:224–237. https://doi.org/10.1016/j.quaint.2012.11.027

Haigh M, Rawat JS (2012) Landslide disasters: seeking causes–a case study from Uttarakhand, India. Management of Mountain Watersheds, pp 218–253. https://doi.org/10.1007/978-94-007-2476-1-18

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. https://doi.org/10.1148/radiology.143.1.7063747

He Q, Shahabi H, Shirzadi A, Li S, Chen W, Wang N, Chai H, Bian H, Ma J, Chen Y (2019) Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms. Sci Total Environ 663:1–15. https://doi.org/10.1016/j.scitotenv.2019.01.329

Hong HY, Pourhasemi 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. https://doi.org/10.1016/j.geomorph.2016.02.012

Hong HY, Tsangaratos P, Ilia I, Liu J, Zhu AX, Chen W (2018) Application of fuzzy weight of evidence and data mining techniques in the construction of flood susceptibility map of Poyang County, China. Sci Total Environ, https://doi.org/10.1016/j.scitotenv

Hosmer DW, Lemeshow S (2000) Applied Logistic Regression, 2nd edn. Wiley-Blackwell, Hoboken, NJ, USA

Huang G, Wu L, Ma X, Zhang W, Fan J, Yu X, Zeng W, Zhou H (2019) Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J Hydrol 574:1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

Irigaray C, Fernández T, El Hamdouni R, Chacón J (2007) Evaluation and validation of landslide-susceptibility maps obtained by a GIS matrix method: examples from the Betic Cordillera (southern Spain). Nat Hazards 41:61–79. https://doi.org/10.1007/s11069-006-9027-8

Jebur MN, Pradhan B, Tehrany MS (2014) Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. Remote Sens Environ 152:150–165. https://doi.org/10.1016/j.rse.2014.05.013

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

Kanungo DP, Arora MK, Sarkar S, Gupta RP (2006) A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. Eng Geol 85(3):347–366. https://doi.org/10.1016/j.enggeo.2006.03.004

Keesstra SD, Quinton JN, van der Putten WH, Bardgett RD, Fresco LO (2016) The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. Soil 2(2):111–128. https://doi.org/10.5194/soil-2-111-2016

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154

Kim JC, Lee S, Jung HS, Lee S (2018) Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. Geocarto Int 33:1000–1015. https://doi.org/10.1080/10106049.2017.1323964

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

Korup O, Stolle A (2014) Landslide prediction from machine learning. Geol Today 30(1):26–33. https://doi.org/10.1111/gto.12034

Kumar R, Anbalagan R (2016) Landslide susceptibility mapping using analytical hierarchy process (AHP) in Tehri reservoir rim region. Uttarakhand J Geol Soc India 87:271286. https://doi.org/10.1007/s12594-016-0395-8

Lagomarsino D, Tofani V, Segoni S, Catani F, Casagli NA (2017) Tool for classification and regression using random forest methodology: applications to landslide susceptibility mapping and soil thickness modeling. Environ Model Assess 22:201–214. https://doi.org/10.1007/s10666-016-9538-y

Lee S, Chwae U, Min K (2002) Landslide susceptibility mapping by correlation between topography and geological structure: the Janghung area. Korea Geomorphol 46(3–4):149–162. https://doi.org/10.1016/S0169-555X(02)00057-0

Lee S, Hong SM, Jung HS (2017) A support vector machine for landslide susceptibility mapping in Gangwon Province, Korea. Sustainability 9(1):48–63. https://doi.org/10.3390/su9010048

Liao D, Valliant R (2012) Variance inflation factors in the analysis of complex survey data. Surv Methodol 38:53–62

Li H, Xu Y, Zhou J, Wang X, Yamagishi H, Dou J (2020a) Preliminary analyses of a catastrophic landslide occurred on July 23, 2019, in Guizhou Province, China. Landslides 17:719–724. https://doi.org/10.1007/s10346-019-01334-0

Li Y, Liu X, Han Z, Dou J (2020b) Spatial proximity-based geographically weighted regression model for landslide susceptibility assessment: a case study of Qingchuan area. China Appl Sci 10:1107. https://doi.org/10.3390/app10031107

Lombardo L, Cama M, Conoscenti C, Märker M, Rotigliano E (2015) Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Nat Hazards 79:1621–1648. https://doi.org/10.1007/s11069-015-1915-3

Magliulo P, DiLisio A, Russo F, Zelano A (2008) Geomorphology and landslide susceptibility assessment using GIS and bivariate statistics: a case study in southern Italy. Nat Hazards 47:411–435. https://doi.org/10.1007/s11069-008-9230-x

Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004) Landslides, earthquakes, and erosion. Earth Planet Sci Lett 229:45–59. https://doi.org/10.1016/j.epsl.2004.10.018

Mason SJ, Graham NE (2002) Areas beneath the relative operating characteristics(ROC) and relative operating levels (ROL) curves: statistical significance andinterpretation. Q J R Meteorol Soc 128:2145–2166. https://doi.org/10.1256/003590002320603584

Mathew J, Jha VK, Rawat GS (2009) Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. Landslides 6(1):17–26. https://doi.org/10.1007/s10346-008-0138-z

Ma X, Sha J, Wang D, Yu Y, Yang Q, Niu X (2018) Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron Commer Res Appl 31:24–39. https://doi.org/10.1016/j.elerap.2018.08.002

Merghadi A, Abderrahmane B, Tien Bui D (2018) Landslide susceptibility assessment at Mila Basin (Algeria): a comparative assessment of prediction capability of advanced machine learning methods. ISPRS Int J Geo-Inf 7(7):268–298. https://doi.org/10.3390/ijgi7070268

Moore ID, Burch GJ (1986) Physical basis of length–slope factor in the universal soil loss equation. Soil Sci Soc Am J 50:1294–1298. https://doi.org/10.2136/sssaj1986.03615995005000050042x

Moradi H, Avand M, Janizadeh S (2019) Landslid susceptibility survey using modeling methods. Spat Model GIS R Earth Environ Sci 2019:259–275. https://doi.org/10.1016/B978-0-12-815226-3.00011-9

O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41:673–690. https://doi.org/10.1007/s11135-006-9018-6

Park NW (2015) Using maximum entropy modeling for landslide susceptibility mapping with multiple geoenvironmental data sets. Environ Earth Sci 73(3):937–949. https://doi.org/10.1007/s12665-014-3442-z

Pham BT, Prakash I, Dou J, Singh SK, Trinh PT, Tran HT, Le TM, Van Phong T, Khoi DK, Shirzadi A, Tien Bui D (2019) A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. Geocarto Int. https://doi.org/10.1080/10106049.2018.1559885

Pham BT, Prakash I, Singh SK, Shirzadi A, Shahabi H, Tran T, Tien Bui D (2019b) Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: hybrid machine learning approaches. CATENA 175:203–218. https://doi.org/10.1016/j.catena.2018.12.018

Pham BT, Phong TV, Avand M, Al-Ansari N, Singh SK, Le HV, Prakash I (2020) Improving voting feature intervals for spatial prediction of landslides. Math Problems Eng. https://doi.org/10.1155/2020/4310791

Pradhan AMS, Kim YT (2019) Rainfall-induced shallow landslide susceptibility mapping at two adjacent catchments using advanced machine learning algorithms. ISPRS Int J Geo-Inf 9(10):569–591. https://doi.org/10.3390/ijgi9100569

Pourghasemi HR, Goli Jirandeh A, Pradhan B, Xu C, Gokceoglu C (2013) Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province. Iran J Earth Syst Sci 122(2):349–369. https://doi.org/10.1007/s12040-013-0282-2

Pourghasemi HR, Rahmati O (2018) Prediction of the landslide susceptibility: which algorithm, which precision? CATENA 162:177–192. https://doi.org/10.1016/j.catena.2017.11.022

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 31:6638–6648

Quan H, Lee B (2012) GIS-based landslide susceptibility mapping using analytic hierarchy process and artificial neural network in Jeju(Korea). KSCE J Civil Eng 16:1258–1266. https://doi.org/10.1007/s12205-012-1242-0

Regmi NR, Giardino JR, Vitek JD (2010) Modeling susceptibility to landslides using the weight of evidence approach: western Colorado, USA. Geomorphology 115:172–187. https://doi.org/10.1016/j.geomorph.2009.10.002

Saber M, Boulmaiz T, Guermoui M, Abdrado KI, Kantoush SA, Sumi T, Boutaghane H, Nohara D, Mabrouk E (2021) Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. Geocarto Int 2021:1974595. https://doi.org/10.1080/10106049.2021.1974959

Sassa K, Guzzetti F, Yamagishi H, Arbanas Ž, Casagli N, Tiwari B, Liu KF, Strom A, McSaveney M, McSaveney E, Dang K, Setiawan H (2017) Landslide dynamics: ISDR-ICL landslide interactive teaching tools (LITT). Advancing Culture of Living With Landslides. Springer International Publishing Cham, pp193–218. https://doi.org/10.1007/978-3-319-59469-9_17

Sevgen E, Kocaman S, Nefeslioglu HA, Gokceoglu C (2019) A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression. ANN Random For Sens 19(18):3940. https://doi.org/10.3390/s19183940

Shrestha S, Kang TS, Suwal M (2017) An ensemble model for co-seismic landslide susceptibility using GIS and random forest method. ISPRS Int J Geo-Inf 6:365–387. https://doi.org/10.3390/ijgi6110365

Song Y, Niu R, Xu S, Ye R, Peng L, Guo T, Li S, Chen T (2019) Landslide susceptibility mapping based on weighted gradient boosting decision tree in Wanzhou section of the three gorges reservoir area (China). ISPRS Int J Geo-Inf 8(1):4–23. https://doi.org/10.3390/ijgi8010004

Suzen ML, Doyuran V (2004) Data driven bivariate landslide susceptibility assessment using geographical information systems: a method and application to Asarsuyu catchment. Turkey Eng Geol 71(3–4):303–321. https://doi.org/10.1016/S0013-7952(03)00143-1

Taalab K, Cheng T, Zhang Y (2018) Mapping landslide susceptibility and types using random forest. Big Earth Data 00:1–20. https://doi.org/10.1080/20964471.2018.1472392

Tien Bui D, Tuan TA, Hoang ND, Thanh NQ, Nguyen DB, Liem NV, Pradhan B (2017) 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. https://doi.org/10.1007/s10346-016-0711-9

Tien Bui D, Shirzadi A, Shahabi H, Geertsema M, Omidvar E, Clague JJ, Thai Pham B, Dou J, Talebpour Asl D, Bin Ahmad B, Lee S (2019) New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. Forests 10(9):743–765. https://doi.org/10.3390/f10090743

Tseng CM, Lin CW, Hsieh WD (2015) Landslide susceptibility analysis by means of event-based multitemporal landslide inventories. Nat Hazard Earth Syst 3(2):1137–1173. https://doi.org/10.5194/nhessd-3-1137-2015

Van Den Eeckhaut M, Moeyersons J, Nyssen J, Abraha A, Poesen J, Haile M, Deckers J (2009) Spatial patterns of old, deep-seated landslides: a case study in the northern Ethiopian highlands. Geomorphology 105:239–252. https://doi.org/10.1016/j.geomorph.2008.09.027

Wang F, Fan X, Yunus AP, Siva Subramanian S, Alonso-Rodriguez A, Dai L, Xu Q, Huang R (2019) Coseismic landslides triggered by the 2018 Hokkaido, Japan (Mw 6.6), earthquake: spatial distribution, controlling factors, and possible failure mechanism. Landslides 16:1551–1566. https://doi.org/10.1007/s10346-019-01187-7

Wang Y, Song C, Lin Q, Li J (2016) Occurrence probability assessment of earthquaketriggered landslides with Newmark displacement values and logistic regression: the Wenchuan earthquake, China. Geomorphology 258:108–119. https://doi.org/10.1016/j.geomorph.2016.01.004

Wang Z, Hong T, Piette MA (2020) Building thermal load prediction through shallow machine learning and deep learning. Appl Energy 263:114683. https://doi.org/10.1016/j.apenergy.2020.114683

Yalcin A, Reis S, Aydinoglu AC, Yomralioglu T (2011) A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey. CATENA 85:274–287. https://doi.org/10.1016/j.catena.2011.01.014

Yang CF (2016) Critical rainfall amount of geological hazards at Northern Mountains in Fujian Province (Yanping District). J Water Resour Archit Eng 14(5):219–223. https://doi.org/10.3969/j.issn.1672-1144

Yariyan P, Avand M, Abbaspour RA, Haghighi AT, Costache R, Ghorbanzadeh O, Janizadeh S, Blaschke T (2020) Flood susceptibility mapping using an improved analytic network process with statistical models. Geomat Nat Haz Risk 11(1):2282–2314. https://doi.org/10.1080/19475705.2020.1836036

Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. Geomorphology 101:572–582. https://doi.org/10.1016/j.geomorph.2008.02.011

Yousefi S, Avand M, Yariyan P, Pourghasemi HR, Keesstra S, Tavangar S, Tabibian S (2020) A novel GIS-based ensemble technique for rangeland downward trend mapping as an ecological indicator change. Ecol Ind 117:106591. https://doi.org/10.1016/j.ecolind.2020.106591

Youssef AM (2016) Landslide susceptibility delineation in the Ar- Rayth area, Jizan, Kingdom of Saudi Arabia, using analytical hierarchy process, frequency ratio, and logistic regression models. Environ Earth Sci 73:8499–8518. https://doi.org/10.1007/s12665-014-4008-9

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 12:639–655. https://doi.org/10.1016/j.gsf.2020.05.010

Yunus AP, Fan X, Tang X, Jie D, Xu Q, Huang R (2020) Decadal vegetation succession from MODIS reveals the spatio-temporal evolution of post-seismic landsliding after the 2008 Wen chuan earthquake. Remote Sens Environ 236:111476. https://doi.org/10.1016/j.rse.2019.111476

Zhang RH, Wu CZ, Goh ATC, Böhlke T, Zhang WG (2020a) Estimation of diaphragm wall deflections for deep braced excavation in anisotropic clays using ensemble learning. Geosci Front 12(2021):365–373

Zhang WG, Zhang RH, Wu CZ, Goh ATC, Lacasse S, Liu ZQ, Liu HL (2019a) State-of-the-art review of soft computing applications in underground excavations. Geosci Front 11:1095–1106. https://doi.org/10.1016/j.gsf.2020.03.003

Zhang Y, Zhao Z, Zheng J (2020b) CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. J Hydrol 588:125087. https://doi.org/10.1016/j.jhydrol.2020.125087

Zhang YH, Ge TT, Tian W, Liou YA (2019b) Debris flow susceptibility mapping using machine-learning techniques in Shigatse Area, China. Remote Sens Basel 11(23):2801–2827. https://doi.org/10.3390/rs11232801

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Peng Ye[1] · Bin Yu[1] · Wenhong Chen[1] · Kan Liu[2] · Longzhen Ye[2]

✉ Bin Yu
yubin08@cdut.cn

Peng Ye
zzyepeng_105@stu.cdut.edu.cn

Wenhong Chen
625006176@qq.com

Kan Liu
100311216@qq.com

Longzhen Ye
FJDKYYLZ@163.com

[1] State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China

[2] Key Laboratory of Geohazard Prevention of Hill Mountains, Ministry of Natural Resources, Fuzhou 350002, China