Journal of Hydrology

Machine Learning Techniques as an Alternative Approach to Rainfall-Runoff Inundation Models for Flood Susceptibility Prediction

--Manuscript Draft--

Manuscript Number:	HYDROL46050
Article Type:	Research paper
Keywords:	machine learning; random forest; LightGBM; CatBoost; flooding susceptibility mapping; rainfall-runoff inundation model
Corresponding Author:	Mohamed Saber Uji, JAPAN
First Author:	Mohamed Saber
Order of Authors:	Mohamed Saber
	Tayeb Boulmaiz
	Mawloud Guermoui
	Karim I. Abdrado
	Sameh A. Kantoush
	Tetsuya Sumi
	Hamouda Boutaghane
	Tomoharu Hori
	Doan Van Binh
	Binh Quang Nguyen
	Thao T. P. Bui
	Ngoc Duong Vo
	Emad Mabrouk
Abstract:	Vietnam has experienced many natural disasters, particularly typhoons. This study aims to examine three machine learning (ML) approaches—random forest (RF), LightGBM, and CatBoost—for flooding susceptibility maps (FSMs) in the Vu Gia-Thu Bon (VGTB) River Basin of Vietnam. The results of ML are compared with those of the rainfall–runoff model, and different training dataset sizes are utilized in the performance assessment. Ten independent factors that influence the FSMs in the study area, namely, aspect, rainfall, curvature, DEM, horizontal distance from the river, geology, hillshade, land use, slope, and stream power index, are assessed. An inventory map that includes approximately 850 flooding sites is considered based on several post-flood surveys after the typhoons in 1999, 2006, 2007, 2009, 2013, and 2020. The inventory dataset is randomly divided into two sets: training (70%), and testing (30%). The AUC-ROC results are 97.9%, 99.5%, 99.5% for CatBoost, LightGBM, and RF, respectively. The FSMs developed by the ML methods show good agreement with flood inundation mapping developed using the rainfall-runoff model. The FSMs show that downstream areas (both urbanized and agricultural) are under “high” and “very high” levels of susceptibility. Additionally, different sizes of the input datasets (i.e., 30, 60, 90, 200, 400, 600, 800, 1000, and 1250 data points) are tested to determine the least number of data points having an acceptable reliability. The results show that the ML methods can reasonably predict FSMs, regardless of the number of training samples, although the final FSMs show some spatial differences when changes in susceptibility level are seen. The developed FSMs for such typhoon-prone regions can be used by decision-makers and planners in Vietnam to propose effective mitigation measures for community resilience and development.
Suggested Reviewers:	Hisham Eldardiry dardiry@uw.edu

	Expert in hydrology
	Hadir Abd-El Moneim hadir_eng@yahoo.com Expert in Hydrological Modeling
	Ashraf Elmoustafa elmoustafa010@yahoo.co.uk
	Ali Al-Maktoom ali4530@squ.edu.om
	Takahiro Sayama sayama.takahiro.3u@kyoto-u.ac.jp

Machine Learning Techniques as an Alternative Approach to Rainfall-Runoff Inundation Models for Flood Susceptibility Prediction

*Mohamed Saber[1, *], Tayeb Boulmaiz[2], Mawloud Guermoui[3], Karim I. Abdrado[4], Sameh A. Kantoush[1], Tetsuya Sumi[1], Hamouda Boutaghane[5], Tomoharu Hori[1], Doan Van Binh[6], Binh Quang Nguyen[1,7], Thao T. P. Bui[1], Ngoc Duong Vo[7], Emad Mabrouk[8,9]*

[1]*Disaster Prevention Research Institute (DPRI), Kyoto University, Kyoto 611-0011, Japan; *mohamedmd.saber.3u@kyoto-u.ac.jp, kantoush.samehahmed.2n@kyoto-u.ac.jp, sumi.tetsuya.2s@kyoto-u.ac.jp*

[2]*Materials, Energy Systems Technology and Environment Laboratory, Ghardaia University, Ghardaia, Algeria; boulmaiz.tayeb@univ-ghardaia.dz*

[3]*Unité de Recherche Appliquée en Energies Renouvelables, URAER, Centre de Développement des Energies Renouvelables, CDER, 47133 Ghardaïa, Algeria; gue.mouloud@gmail.com*

[4]*Faculty of Urban and Regional Planning, Cairo University, Giza 12613, Egypt; m.karim.ibrahim@cu.edu.eg*

[5]*Hydraulic Department, Badji Mokhtar-Annaba University, P.O. Box 12, Annaba, Algeria; hamouda.boutaghane@univ-annaba.dz*

[6]*Master Program in Water Technology, Reuse, and Management, Faculty of Engineering, Vietnamese German University, 2-Le Lai Street, Hoa Phu Ward, Thu Dau Mot City, Binh Duong Province 820000, Vietnam; binh.dv@vgu.edu.vn*

[7]*The University of Danang -University of Science and Technology, 54 Nguyen Luong Bang, Danang, Vietnam; nqbinh@dut.udn.vn*

[8]*College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait; emad.mabrouk@aum.edu.kw*

[9]*Department of Mathematics, Faculty of Science, Assiut University, Assiut 71516, Egypt; mabrouk@aun.edu.eg*

*__Corresponding:__ *mohamedmd.saber.3u@kyoto-u.ac.jp*

Abstract

Vietnam has experienced many natural disasters, particularly typhoons. This study aims to examine three machine learning (ML) approaches—random forest (RF), LightGBM, and CatBoost—for flooding susceptibility maps (FSMs) in the Vu Gia-Thu Bon (VGTB) River Basin of Vietnam. The results of ML are compared with those of the rainfall–runoff model, and different training dataset sizes are utilized in the performance assessment. Ten independent factors that influence the FSMs in the study area, namely, aspect, rainfall, curvature, DEM, horizontal distance from the river, geology, hillshade, land use, slope, and stream power index, are assessed. An inventory map that includes approximately 850 flooding sites is considered based on several post-flood surveys after the typhoons in 1999, 2006, 2007, 2009, 2013, and 2020. The inventory dataset is randomly divided into two sets: training (70%), and testing (30%). The AUC-ROC results are 97.9%, 99.5%, 99.5% for CatBoost, LightGBM, and RF, respectively. The FSMs developed by the ML methods show good agreement with flood inundation mapping developed using the rainfall-runoff model. The FSMs show that downstream areas (both urbanized and agricultural) are under "high" and "very high" levels of susceptibility. Additionally, different sizes of the input datasets (i.e., 30, 60, 90, 200, 400, 600, 800, 1000, and 1250 data points) are tested to determine the least number of data points having an acceptable reliability. The results show that the ML methods can reasonably predict FSMs, regardless of the number of training samples, although the final FSMs show some spatial differences when changes in susceptibility level are seen. The developed FSMs for such typhoon-prone regions can be used by decision-makers and planners in Vietnam to propose effective mitigation measures for community resilience and development.

Keywords

Machine learning, random forest, LightGBM, CatBoost, flooding susceptibility mapping, rainfall-runoff inundation model.

1. Introduction

Globally, floods are the most damaging natural disaster. Flash floods are more devastating than any other flooding type because of their short lag times (Bui et al., 2019b; Vinet, 2008). The disastrous impacts of flash floods have been documented in both developed and developing countries (Bisht et al., 2018); however, flood events are more destructive in developing countries, such as Vietnam. The observed increase in flash flood frequency is driven mainly by extreme changes in storm patterns and global climate change (Hirabayashi et al., 2013; Pachauri et al., 2014). Causes of severe flooding in Vietnam include tropical cyclones, typhoons, dense river networks, and extended coastal areas. The area is also highly vulnerable to floods caused by extreme storms. Vietnam is ranked eighth among the 10 countries with the highest number of weather events (Thao et al., 2020), where densely populated areas are more vulnerable to floods. Consequently, continuous risk in terms of human life and assets will always exist (Luu et al., 2021). Flash flood mitigation for risk reduction requires efficient monitoring measures for sustainable support of flood hazard management (Arora et al., 2020). Mapping flash flood susceptibility (FFS) areas is critical for scientists and governments worldwide (Ali et al., 2020).

Several studies have been performed to predict the likelihood of flood events. These studies can be clustered into three major groups: conventional analysis, rainfall-runoff, and pattern categorization (Tien Bui and Hoang, 2017). Conventional analysis uses time-series data for an extended period obtained from rainfall stations to produce regression models. The rainfall-runoff models (e.g., MIKE, PCSWMM 2D, HEC-RAS, etc.) focus on determining the relationship between rainfall and runoff to calculate temporal and spatial floods (Nguyen et al., 2015). This task is generally difficult because of difficulties in accessing affected areas, especially in developing countries; consequently, the performance of hydrological models can be affected, and detailed observational datasets are necessary for calibration and validation (Abdrabo et al., 2020; Abushandi and Merkel, 2011). Both groups have a major deficiency: the lack of data required often limits their applications and incurs a considerable cost for collecting the data (Fenicia et al., 2014). On the other hand, the last group (pattern classification) employs machine learning (ML) models that utilize historical geological, environmental, and flood data. Accordingly, flood-prone areas are defined as “flood” and “non-flood” classes (Bui et al., 2019b). However, comparative studies and integration between these groups are lacking (Demirel et al., 2009; Hsu et al., 1995; Humphrey et al., 2016; Kratzert et al., 2019; Yang et al., 2020).

Globally, the application of ML approaches for flood susceptibility prediction have been extensively assessed over the past 20 years. Therefore, the recent development of ML methods has resulted in substantial enhancements in flood modeling. Such practices have become widespread because of their ability to capture information without applying predefined assumptions and process complex datasets with high levels of accuracy in short periods (Arabameri et al., 2020; Costache et al., 2020b). Several studies have been conducted to develop reliable flooding susceptibility maps (FSMs) using GIS and remote sensing techniques. Currently, ML models are being adopted in combination with GIS to address many hydrological and environmental challenges (Akay and Taş, 2020). ML models used to predict flood susceptibility include artificial neural networks (ANNs), logistic regression (LR) (Arora et al., 2020; Shahabi et al., 2020), adaptive neuro-fuzzy inference system (ANFIS) (Costache, Hong, et al. 2020; Arora et al. 2020), genetic algorithms (GAs) (Darabi et al., 2019; Shirzadi et al., 2020), support vector machines (SVMs) (Choubin et al., 2019; Dodangeh et al., 2020), and random forest (RF) models. The RF model has been extensively used for flood risk assessments (Chen et al., 2020; Esfandiari et al., 2020). Several ML models and ensemble methods have been used to predict FFS (Shahabi et al., 2020). ML approaches involve multiple steps (Arora et al., 2020), including the preparation of inventory and influencing factors, as well as the assessment of the accuracy of the ML model. However, few studies have discussed the effect of inventory dataset size on the accuracy of results (Catal and Diri, 2009; Meadows and Wilson, 2021; Tiwari and Chatterjee, 2010).

Recently, ensemble and hybrid ML models have appeared, and they outperform single models in their prediction accuracy (Zenggang et al., 2021). Several ensemble ML techniques, such as the alternating decision tree, bagging, dagging, reduced-error pruning tree, logistic model tree, J48 decision tree, naïve Bayes tree, AdaBoost, and random subspace ensembles, have been applied to enhance the predictive accuracy of the FSM (Luu et al., 2021; Pham et al., 2021a; Tuyen et al., 2021). In Vietnam, several studies have developed FFS maps using ML, which can be categorized into three groups. The first group evaluates the utilization of new ML models and their ability to detect areas prone to floods. For instance, the AdaBoost, bagging, dagging, and random subspace ensemble learning techniques were combined with the Partial Decision Tree (PART) classifier to develop new GIS-based ensemble models for FSM in the Quang Binh Province (Luu et al., 2021). The second group attempts to overcome the limitations in the number of studies that utilize remote sensing data to generate input variables for FSM despite the merits of using such available data (Pham et al., 2019, pp. 2010–2018). As such, (Dhara et al., 2020; Ngo et al., 2021; V.-N. Nguyen et al., 2020; Nhu et al., 2020, p. 202)

suggested a hybrid approach using remotely sensed data with ML models for the FFS. The third group introduced a novel deep learning neural network (DLNN) algorithm for FSM (Tien Bui et al., 2020), integrating extreme learning machines (ELMs) and particle swarm optimization (PSO) (Bui et al., 2019b, 2020) along with a comparison between ML and deep learning techniques (Pham et al., 2021b) for the same study area.

In this study, we examined two methods, categorical boosting (CatBoost) and the light gradient boosting machine (LightGBM), for FSM for the first time in humid regions after successful application in arid regions (Saber et al., 2021). Previously, both methods have been applied to LightGBM and CatBoost. For instance, LightGBM has been used in previous studies owing to its accurate predictions, short computational time, and outstanding ability to avoid overfitting issues. Accordingly, our primary objectives are (1) to evaluate how practical the two ML approaches (LightGBM and CatBoost) are for predicting FFS in humid environments (Vu Gia-Thu Bon basin in Vietnam); (2) to compare the performance of these methods to the typical RF method; (3) to test the effect of the inventory datasets (number of points) on the accuracy of the results in the study area; and (4) to compare the rainfall-runoff inundation (RRI) 2D hydrological model with the proposed ML integrated models.

The subsequent sections of this study are as follows. In Section 2, a description of the study area is presented. In Section 3, an introduction to the datasets and methodology is developed. In Section 4, the results are presented and discussed. Finally, our conclusions are given in Section 5.

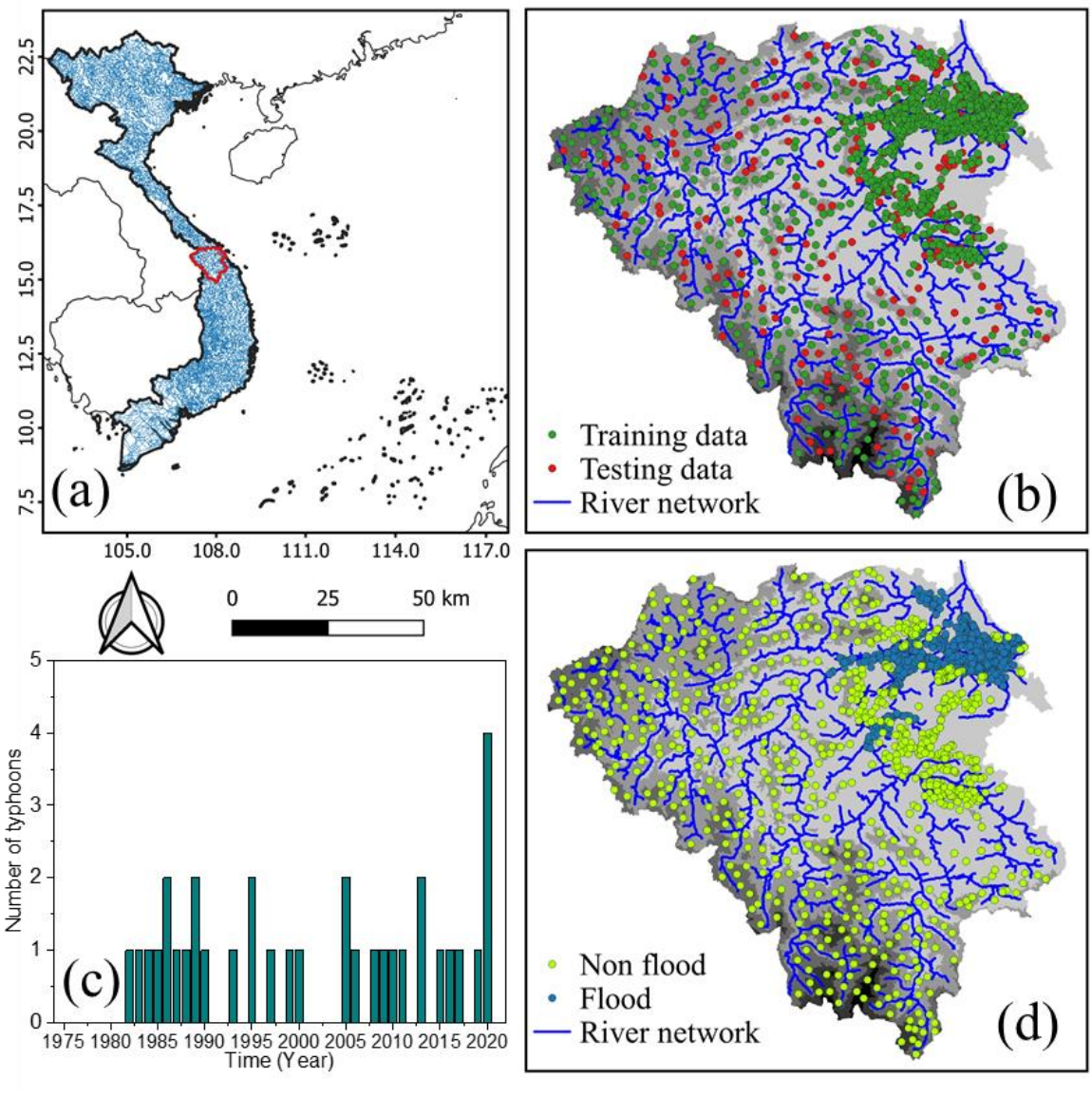
2. Study Area

The Vu Gia-Thu Bon (VGTB) River Basin (Fig. 1) is one of the largest river basins in Vietnam, with a surface area of 10350 km^2 (RETA, 2011). The land use types in the basin are forest (47%), cropland (26%), and grassland (20%) (Avitabile et al., 2016). The basin has a tropical monsoon climate with two seasons: dry summer (January–August) and rainy (September–December) seasons. The topographic features of the basin are hilly mountainous areas with approximately 60% of the basin having an elevation over 552 m. The average annual rainfall varies significantly, from 2000 mm in the downstream regions, to more than 4000 mm in mountainous areas. There are seasonal differences, with 65%–80% of the annual rainfall concentrated between September and December

(RETA, 2011). The rainfall in eight months of the dry season is only 20%–35% of the annual rainfall (Nauditt et al., 2017).

Owing to the difference in the spatial distribution of rainfall, the runoff in the VGTB basin varies significantly across seasons. The flooding season lasts from October to December and is accompanied by rains. River flow in this period accounts for approximately 62.5% to 69.2% of the total annual flow. The impacts of both intense rain and steep terrain usually lead to flooding with a high intensity and short occurrence times. Approximately 4–8 floods occur annually. The maximum flood peak occurs in October and November owing to different weather patterns, such as tropical depressions, cold air, and typhoons (Vu et al., 2011). The number of deaths and property damage caused by storms and floods is increasing over the years, especially in 2020, based on the report of the Commanding Committee for Disaster Prevention and Search and Rescue in Quang Nam Province (Fig. 2).

There are two main sub-basins in the VGTB river system: the Vu Gia Basin and Thu Bon Basin. The Quang Hue River is connecting the Vu Gia and Thu Bon Rivers. The Vu Gia River originates from the western slope in Kon Tum and flows towards Quang Nam province and Danang city. It connects with the sea at the Cua Han estuary. The length of the main river from the source to the Cua Han estuary was 204 km. The Thu Bon River originates from a mountain at an elevation of 1500 m in Kon Tum province. The length of the river from the source to the Cua Dai estuary was 198 km.



171

172 Fig. 1. Location map of the VGTB river basin: a) The Vietnam Map, b) flood inventory map
173 for training and validation datasets, (c) total annual precipitation of the entire basin, and (d)
174 flooded and non-flooded locations,

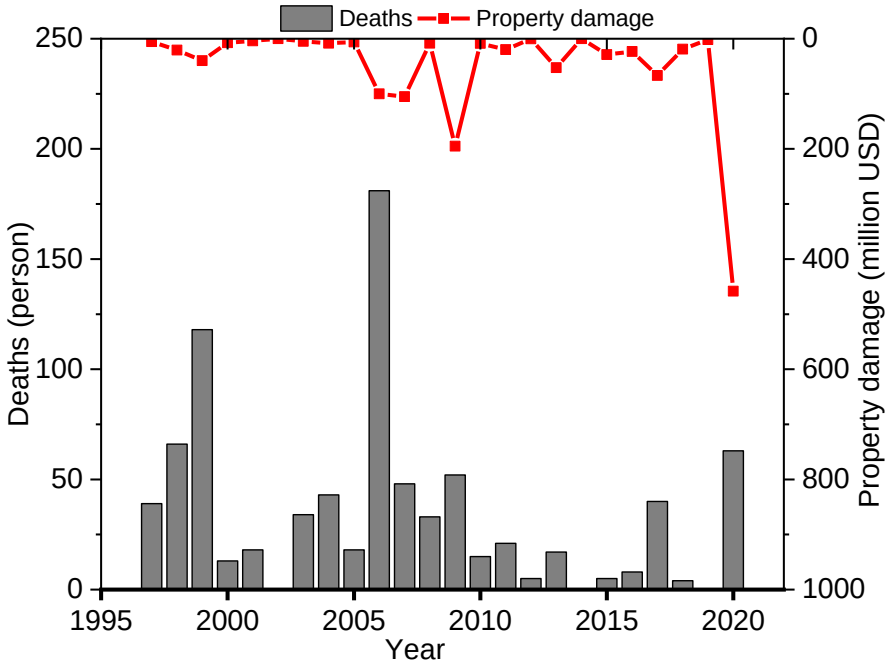


Fig. 2. Number of deaths and property damage caused by storms and floods from 1997 to 2020 in VGTB River basin (source: Commanding Committee for Natural Disaster Prevention and Control, Search and Rescue in Quang Nam Province).

3. Methodology

The methodology of this study consists of several steps, as illustrated in the flowchart in Fig. 3. There are two main parts to this methodology. First, a flood inventory map based on 850 flooded points is developed. These points were identified based mainly on post-flood surveys after typhoons in 1999, 2006, 2007, 2009, 2013, and 2020. Non-flooded points (850) over the catchment were randomly selected using GIS tools. Additionally, a total of 10 independent flood susceptibility factors (FSFs) were considered for modeling based on topographical, hydrological, geological, and landform characteristics, which have been used in several studies. The $FFSF_{S,}$ namely elevation, aspect, slope, hillshade, plan curvature, horizontal flow distance, stream power index (SPI), geology, rainfall, and land use/land cover were used to determine the linear relationship with other factors. In the later stages, the dataset was divided into two sets: (70%) for training and (30 %) for testing using the random

selection scheme. Spatial maps for each FFSF were produced using ArcGIS, considering the consistency of spatial resolution. Following that, two approaches, the information gain ratio (IGR) and variance inflation factor (VIF), were applied to examine the importance of the influencing factors in the FFS. Subsequently, the ML algorithms employed, namely, RF, LightGBM, and CatBoost were employed. The final results of the ML models were assessed for accuracy by using different statistical measures, including the most dominant, area under the curve (AUC). Moreover, as we have very high-quality observational flood locations, we tested the models to check the different sizes of the training datasets (Fig. 3). The final flood susceptibility maps developed by the ML models were then compared with the flood inundation maps from the physical hydrological model.

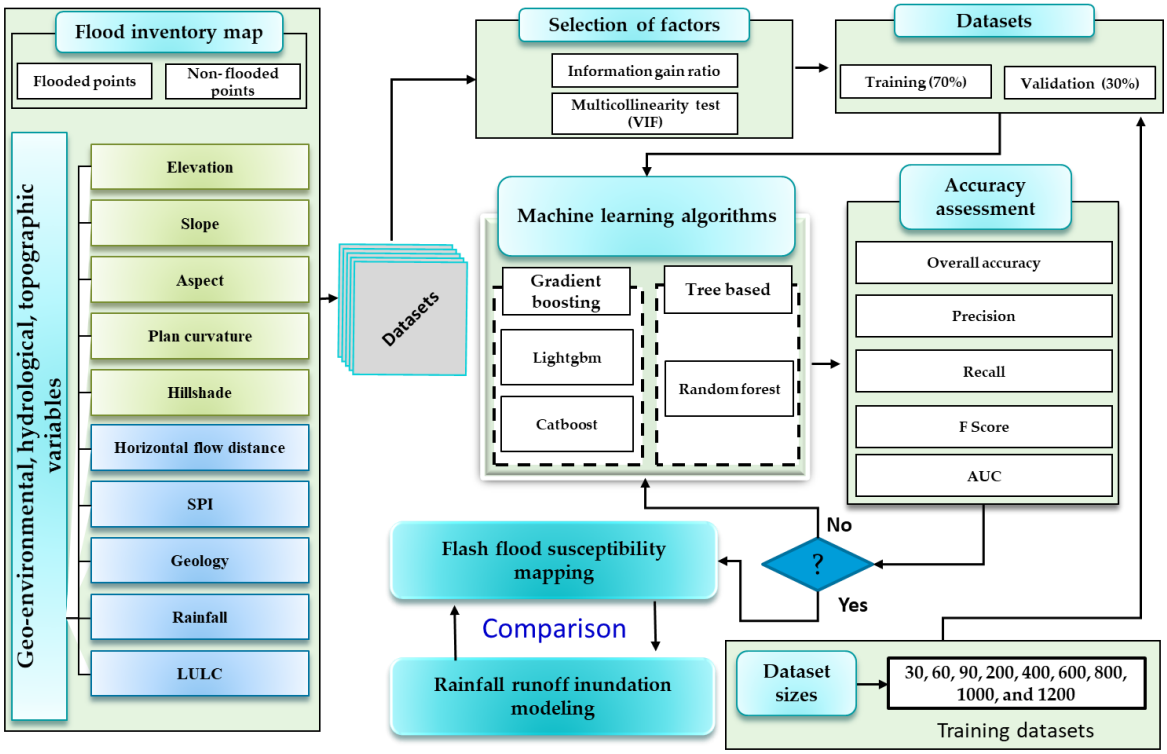


Fig. 3. Methodology flowchart for flash flood susceptibility mapping.

206 ## 3.1. Datasets

207 ### 3.1.1. Flash Flood inventory data

208

209 The first step in FFS mapping involves identifying flood points or locations based on
210 historical records of previous floods. This is the most important data input for FFS mapping
211 (Tehrany et al., 2014). From the records of past occurrences, the locations of future hazard
212 events could be estimated (Devkota et al., 2013; Tehrany and Kumar, 2018). Therefore,
213 analysis of previous historical flood events and their influencing factors is the primary step
214 in flood susceptibility analysis (Masood and Takeuchi, 2012). The flood inventory map
215 shows the sites of the flooded areas in any flood-prone basin (Bellu et al., 2016). An inventory
216 map can be developed from several sources, such as field surveys, flood forecasting records,
217 and remote sensing data (Band et al., 2020; Esfandiari et al., 2020; Wang et al., 2019). The
218 accuracy in selecting flood points will enhance the model accuracy for FSM (Arora et al.,
219 2019; Tehrany et al., 2013). In this study, 1700 ground control points (Fig. 1) were identified
220 for flooded (850) and non-flooded points (850). Approximately 1250 were used for training,
221 and 450 for testing the models. The flooded locations were compiled from historical flood
222 records and post-flood field surveys in 1999, 2006, 2007, 2009, 2013, and 2020 (Fig. 1). The
223 non-flood points were randomly selected. The flood points and non-floods were assigned
224 with values 1 and 0, respectively, and the points were divided into 70% and 30% for training
225 to set the flash flood prediction model and validation to verify the model performance and
226 generalization ability by using the random selection method.

227 ### 3.1.2. Geospatial database (flood controlling parameters)

228

229 The selection of flood controlling factors for the FFS mapping is important and affects the
230 accuracy of the models (Kia et al., 2012). During floods in a drainage system, runoff depends
231 on the characteristics of the watershed, topography, catchment area, and land use/land cover
232 types (Hölting and Coldewey, 2019). There are no standard and universal criteria for
233 selecting the controlling factors for FSM; therefore, according to the previous review and

study area characteristics, as well as the data availability, 10 flood triggering factors were prepared, including topographic, hydrological, geologic, and landform factors, namely, elevation, plan curvature, slope, aspect, horizontal flow distance, hillshade, SPI, rainfall, geology, and land use/land cover. Using ArcGIS, all data were resampled and prepared in spatial raster formats with a spatial resolution of 90 m (Fig. 4). All topographic factors were constructed based on the MERIT digital elevation model (Yamazaki et al., 2017). The spatial resolution of the terrain elevation was 3 s (~90 m at the equator). It was developed by eliminating the error components from existing digital elevation models (DEMs), such as Aw3D-30m v1 and SRTM3 v2.1. These data are freely available and accessible at http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/.

Below are the details of all the considered flood influencing parameters in this study.

Elevation: There is a direct relationship between altitude and flooding (Tehrany et al., 2013), which means that lowland surfaces are more vulnerable to floods than elevated areas (Khosravi et al., 2016). This means that the higher the topographic elevation, the lower the flood possibility (Tehrany et al., 2014; Youssef et al., 2016). The study area has complex topographic features, with very high elevations up to 2600 m, and the downstream region of the basin and coastal area has low elevations ranging from −3 m to 200 m; these regions are mainly residential and agricultural areas (Fig. 4a).

Slope: This is a significant factor influencing flooding (Khosravi et al., 2016; Meraj et al., 2018; Tien Bui et al., 2016), because of its effect on water velocity and surface flow (Torabi Haghighi et al., 2018). The steeper slopes contribute to a high velocity and increased volume of water in downstream areas (Chen et al., 2020). Slope influences hydrological features that directly affect water runoff (Tehrany et al., 2019). In the study area, the slope varied from 0° to 70° (Fig. 4b).

Aspect: Aspect is important for flooding (Choubin et al., 2019) as stated by Yates et al. (2002); the hydrological parameters are influenced by this aspect. When aspects receive low intensity of sunlight, which means more soil moisture, the moist slope will be more likely to increase runoff, leading to increased flooding risk (Yariyan et al., 2020). There is an

indirect relationship between the aspects of floods owing to their control over several geo-environmental factors, such as soils, rainfall, and vegetation (Rahmati et al., 2016). In this study, the aspect map was categorized into 10 classes ranging from flat to northwest (Fig. 4c).

Plan Curvature: This is considered to be an important and essential flood influencing factor by many researchers (Hong et al., 2018) and also affects hyporheics and heterogeneity (Cardenas et al., 2004). The different values of curvatures differentiate between the areas of the faster runoff from those with a slower runoff. While negative values cause an increase in runoff, and the positive values decrease it. The runoff is affected by the slope shape, as the flat form (zero curvature) and concave form (negative) have more potential for flooding than the convex form (positive) (Shahabi et al., 2020; Tehrany et al., 2015, 2014). For instance, concave slopes decelerate surface flow and increase filtration losses (Young and Mutchler, 1969), but convex slopes accelerate flow discharge and perhaps reverse filtration losses (Cao et al., 2016). The curvature map was developed from the DEM with three main classes (concave, flat, and convex), and the flat class was more dominant in the downstream area, as shown in Fig. 4d.

Hillshade: Hillshade or toposhade is directly related to the length and shade of hillslopes, which may affect the convergence of overland flow (Aryal et al., 2003). However, the toposhade is limited in previous studies (Bui et al., 2019b), and it was found to be the most important factor in predicting FFS mapping after slope and elevation (Bui et al., 2019a). Therefore, toposhade was selected as a flood influencing factor, as shown in Fig. 4e.

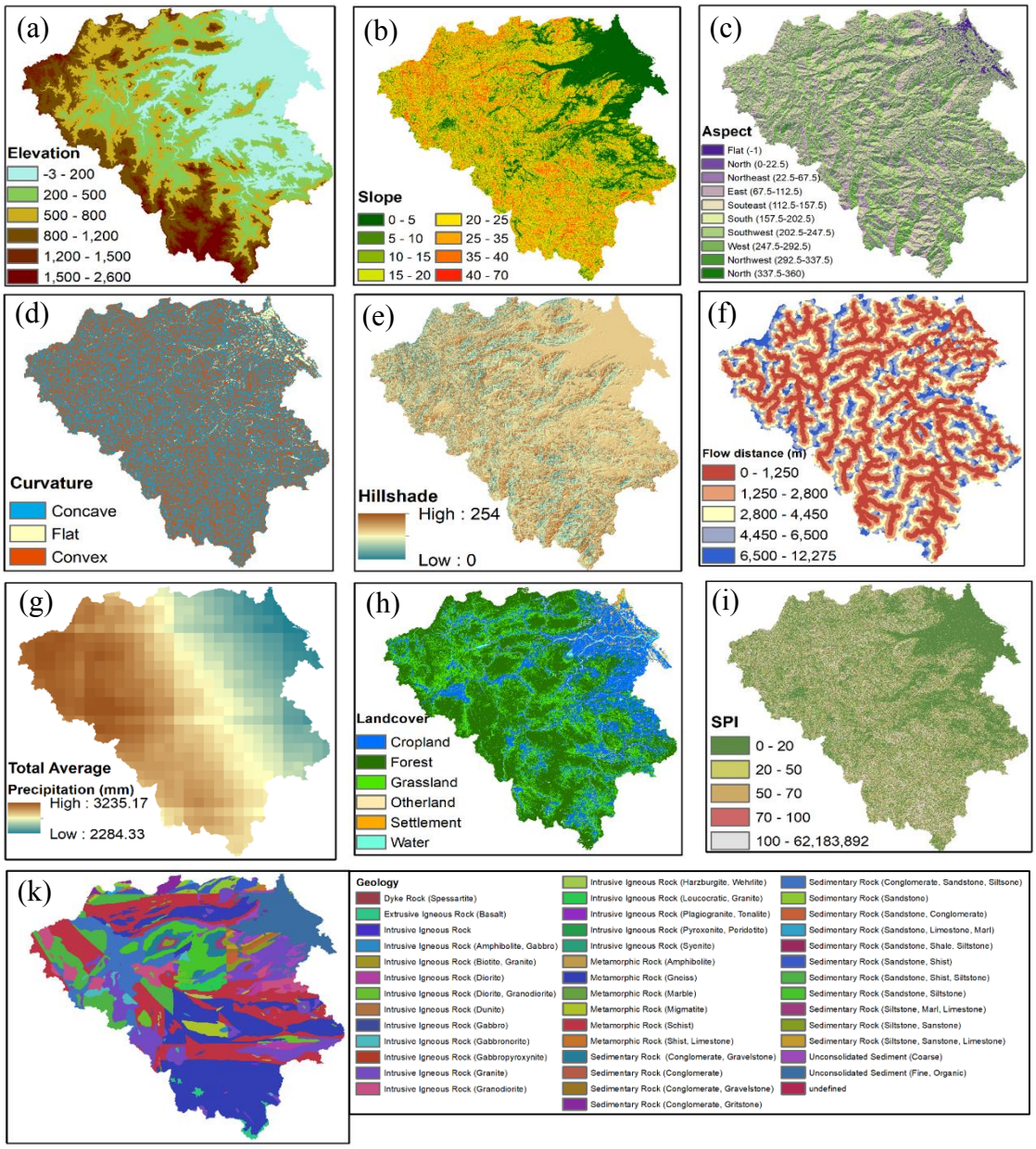


Fig. 4. Flood influencing factors: a) elevation, b) slope, c) aspect, d) plan curvature, e) hillshade, f) horizontal flow distance, g) rainfall, h) land use/land cover, i) SPI, j) geology.

Flow distance: The distance from the main rivers or streams has a significant impact on flooding occurrence in any area (Glenn et al., 2012). The areas adjacent to streams are

typically more prone to flooding(Chapi et al., 2017). The risk of flooding is proportional to the distance from rivers. It has been emphasized as a significant influencing factor for flooding (Predick and Turner, 2008), and flooding is frequent in areas adjacent to rivers (Bui et al., 2018; Darabi et al., 2019). The distance from streams signifies the distance from stream channels, which are the main means of overland flow (Gigović et al., 2017; González-Arqueros et al., 2018). In the present study, the horizontal flow distance was estimated (Fig. 4f.) using ArcGIS from the flow accumulation, flow direction, and DEM in ArcGIS.

Rainfall: Precipitation is one of the triggering factors for flooding as no rainfall indicates a lack of flooding. The total average rainfall was estimated between 2001 and 2019 using the PERSIANN Dynamic Infrared–Rain rate model (PDIR). Estimation of precipitation was done using remotely sensed information that utilizes ANNs (P. Nguyen et al., 2020, p.). It is a real-time global dataset with a high resolution of approximately (0.04° × 0.04°, or 4 km × 4 km, and it is freely available and accessible for download (https://chrsdata.eng.uci.edu/). The spatial maps show that the average precipitation ranges from 2235 mm in the downstream area to 3284 mm in the upstream region and mountainous area, as shown in Fig. 4g.

Land use and land cover: The influence of this factor was confirmed using the global cover map developed by the geospatial information authority of Japan (https://www.gsi.go.jp/kankyochiri/gm_global_e.html), and mainly from this website (https://globalmaps.github.io/glcnmo.html). Land use/land cover type was also considered as a controlling factor because of its influence on filtration and runoff velocity. The study area has approximately six classes (Fig. 4h), including cropland, forest, grassland, other land, settlement, and water. The forest is the more dominant type of land cover in the mountainous area, especially upstream of the basins, and agricultural land and urban areas are located in the downstream region.

Stream Power Index: This parameter indicates the power of erosion and discharge within a specific area of the river system (Poudyal et al., 2010). Several researchers have

considered the SPI as a flooding contributor because of its indication of surface flow. The highest values of SPI imply a fast flow of downstream water, which reveals lower flooding susceptibility, and low values imply slow flow leading to more inundation (Tehrany and Kumar, 2018). The SPI was calculated based on a method derived from Jebur et al. (2014). In the study area, SPI was classified into five classes, as shown in Fig. 4i.

Geology: This parameter is a crucial factor because of its effect on the infiltration and flow velocity. The lithology was obtained from the Land Use and Climate Change Interaction in Central Vietnam (LUCCI) (Nauditt and Ribbe, 2017). It has been classified into several geological types (Fig. 4k) with high variation in sedimentary, igneous, and metamorphic rock types.

3.1.3. Selection of flood influencing factors

Feature selection is a basic step in ML modeling for FSM. Removing redundant features may decrease in the training speed caused by a larger number of features and the loss of ill-conditioned landscape issues (Öztürk and Akdeniz, 2000). The latter occurs when highly correlated features exist in the training dataset, causing a problem in determining the learning-rate hyperparameter. Therefore, an incorrect choice of this hyperparameter value may affect the estimation ability of the model. Therefore, the feature selection process was based on three analyses to detect irrelevant factors, Spearman's rank correlation, the multicollinearity test, and the IGR.

3.1.4. Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is a nonparametric parameter used to evaluate the strength of the monotonic link between two parameters, X and Y. The value of the coefficient varies between −1 and 1, representing negative and positive degrees of association, respectively. The closer the coefficient value is to zero, the weaker the relationship between X and Y. A Pearson correlation coefficient value greater than 0.7 indicates a high level of collinearity (Tien Bui et al., 2016). The correlation coefficient is estimated as follows:

$$r(x,y) = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)} \quad (1)$$

where r is the correlation coefficient, x and y are the two variables, and n is the length of each variable.

3.1.5. Multicollinearity Test

In addition to the correlations between two features carried out using Spearman's coefficient, multicollinearity was checked in this study between all influencing factors. Multicollinearity analysis aims to detect the existing interrelatedness between variables and was performed (in this study) using the VIF. This factor is commonly used in flood susceptibility assessment studies (Bui et al. 2019; Khosravi et al. 2019; Rahman et al. 2019) which suggests a threshold > 5 to consider multicollinearity. However, in other studies, if the VIF value is greater than 10, the corresponding predictors are considered collinear; thus, it is recommended to exclude them from the models (Dou et al., 2019; Wang et al., 2019). Therefore, in this study, we considered a value of five as the threshold for selection. The independent predictors are defined as $X = \{X_1, X_2, ..., X_n\}$ and R_j^2 , and refer to the coefficient of determination when the jth independent predictor X_j is regressed on the other predictors. VIF is calculated using the following equation:

$$VIF = \frac{1}{1-R_j^2} \quad (2)$$

3.1.6. Information gain ratio

Conditioning factors were evaluated to identify their relative importance in flood occurrence using the IGR test (Quinlan, 1986; Xu et al., 2013). The latter is among the feature selection methods that have been considered in many classification studies (Khosravi et al., 2019; Shahabi et al., 2020). The use of an input with an IGR equal to zero means that no relationship exists between this factor and the output. This situation indicates that the use of such an input in the model will not add any information to the applied model; in contrast, it

will generate noise that will decrease its predictive capability. Therefore, removing these factors from the inputs is highly recommended. The IGR is calculated using Eq. (3):

$$IGR(x,Z) = \frac{Entropy(Z) - \sum_{i=1}^{n} \frac{|Z_i|}{|Z|} Entropy(Z_i)}{\sum_{i=1}^{n} \frac{|Z_i|}{|Z|} \log \frac{|Z_i|}{|Z|}} \quad (3)$$

3.2. Machine learning methods

ML approaches are the basic concept of employing algorithms to analyze and learn from the data to produce forecasting or classification systems. These techniques can be learned from previous experience or a given historical database. These methods can generalize the learning examples provided in the training phase to identify the main tasks that must be performed.

Several ML algorithms have been developed. These techniques can be classified according to their learning mechanisms (i.e., supervised, unsupervised, and semi-supervised learning). The choice of a suitable ML model and training method depends on the problems to be solved or the available data and its types. In the current study, we focused on using supervised ML techniques for FFS assessment. According to previous research, various ML techniques have been proposed recently to deal with FFS assessment (i.e., SVMs, ELMs, ANNs, Gaussian process regression (GPR), and classification and regression trees (CART)). In addition, few studies have addressed FFS using ensemble-learning approaches based on decision trees. These algorithms are based on boosting techniques that concentrate on misclassified data during the training phase. In this respect, the objective of this study is to evaluate the performance of two new modeling techniques, CatBoost and LightGBM, benchmarked with the conventional RF approach.

3.2.1. **Random forest**

RF models have proven efficiency when dealing with prediction and classification problems (Esfandiari et al., 2020; Schoppa et al., 2020). RF is an ensemble learning approach based on a decision tree model. It was developed by Breiman (2001), who combined bagging (Breiman, 2001) and random subspace (Ho 1998) techniques. This ML algorithm has proven to be reliable in many fields (Izquierdo-Verdiguier and Zurita-Milla, 2020; Pourghasemi et al., 2020; Zahedi et al., 2018). In this study, we aimed to predict flood or non-flood regions according to several conditioning factors; therefore, the RF model was used as a classifier method.

The weakness of decision trees is their sensitivity to training data, which may result in very different tree structures. In the RF method, the original training set is used to randomly generate several training sets, thereby allowing the creation of different trees (bagging method). The inputs of the decision trees have the same data size as the initial training, and because the data are randomly generated, the samples may be repeated two or more times. In addition, each tree in the RF is trained with a subset of features that allows the development of diversified trees that are not correlated. The final result (classification) was obtained by performing a majority voting method on each decision tree results (Pal, 2005).

Decision tree models are simple to use and easy to interpret; however, their performance is not always better than that of other classification methods (Malekipirbazari and Aksakalli, 2015). On the other hand, RF outperformed other ML algorithms, such as ANNs (Bachmair et al., 2017).

3.2.2. **Light Gradient Boosting Machine**

LightGBM is a variant of the gradient boosting decision tree (GBDT) that was developed by Microsoft (Ke et al., 2017). It uses a combination of weak learners to generate a robust model. The new variant includes algorithms such as histograms, leafwise tree growth, gradient-based one-side sampling (GOSS), and exclusive feature bundling (EFB).

In GBDT models, the commonly used algorithm for split operations is the presorted algorithm. To determine the optimal split, all possible split points are tested based on the information gain, which is a time-consuming operation. A new histogram algorithm was adopted in the LightGBM method. To reduce the time and complexity of the operation, the data are grouped into a histogram, and the split point is chosen based on it (Fig. 5).

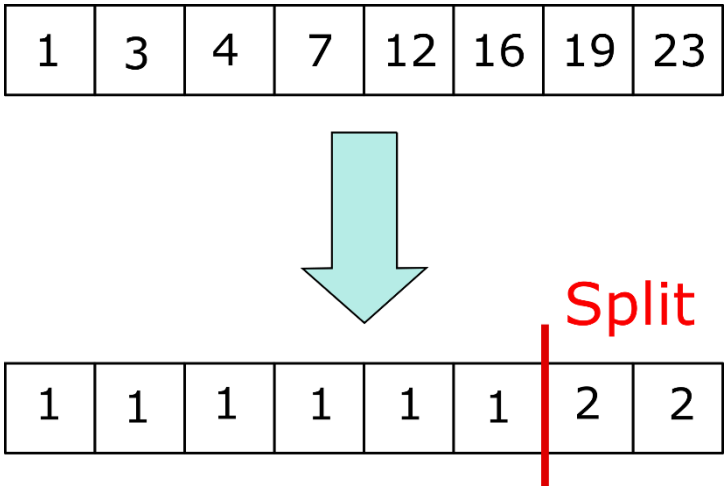


Fig. 5. Split operation example based on histogram algorithm.

In LightGBM, the decision tree growth strategy was changed by replacing the level-wise approach with the leafwise tree growth approach. When finding the best node to split, the former approach of the GBDT splits one level down, forming symmetric trees (Fig. 6). In LightGBM, only the leaves that reduced the maximum error were split (Fig. 6). Ge et al. (2019) recommended defining a maximum leaf-wise depth to avoid deep growth of trees and provoke overfitting of the model.

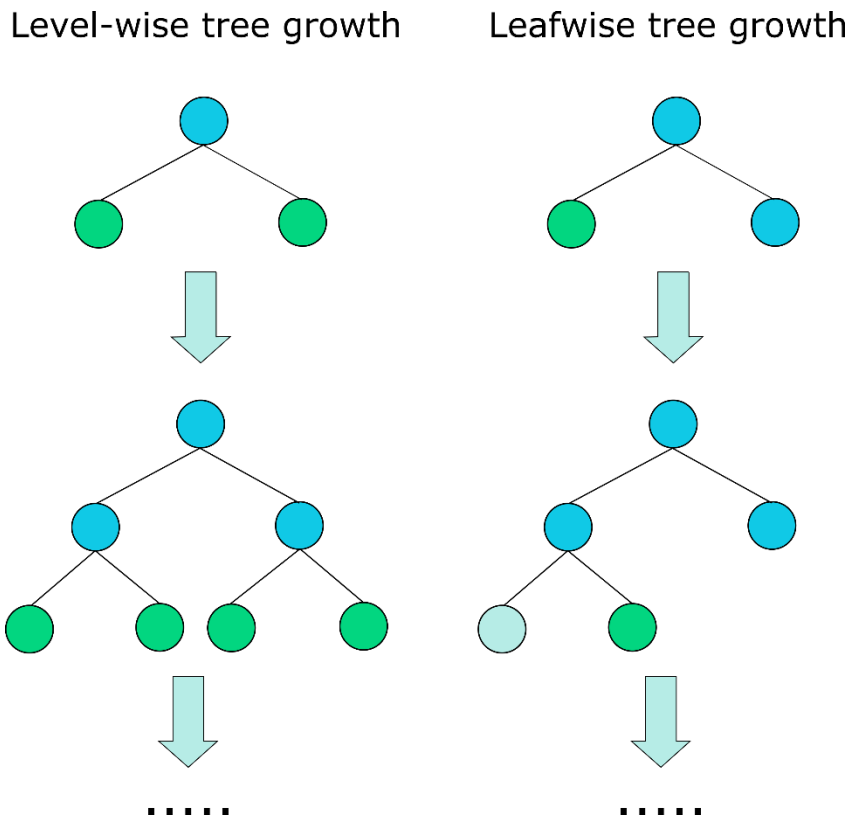


Fig. 6. Level-wise and leafwise tree growth strategies.

The LightGBM model also uses two algorithms (GOSS and EFB), making it faster than GBDT models while maintaining a high performance (Saber et al., 2021).

3.2.3. Categorical boosting

The CatBoost model is another enhanced boosting decision-tree learning technique, which was proposed by (Dorogush et al., 2018). It employs a gradient boosting scheme to construct a regression model through adjusted estimation. Furthermore, various refinements were performed to minimize overfitting of the model. The gradient boosting model is a useful ML tool that has yielded accurate results in many disciplines, including environmental parameter estimation, geospatial ecosystem factor dispersion, and meteorological forecasting. The CatBoost model operates well in terms of categorical attributes. Typically, the absence of

categorical characteristics increases the accuracy of the model. It is primarily dependent on the use of gradient boosting, which employs a binary-tree classification scheme. The following points outline the differences between CatBoost and the other boosting techniques.

- A sophisticated method was incorporated to convert category characteristics into numerical information. As mentioned by (Prokhorenkova et al. 2017), target statistics are very effective for dealing with categorical attributes with minimal information errors.
- CatBoost combines categorical variables to take advantage of the existing relationship between different parameters.
- To reduce the overfitting problem and improve the classification performance, a symmetrical tree strategy is used.

Let us suppose we have a dataset:

$$D = \{(X_J, Y_J)\} \qquad J = 1, \ldots\ldots, m \tag{4}$$

where $X_J = (x_j^1, x_j^2, \ldots, x_j^n)$ is a combination of attributes, and $Y_J \in R$, denotes the desired target. Input–output data are dispersed independently and identically, depending on an unknown function $\rho(\cdot,\cdot)$. The main objective of the learning scheme is to train a function $H: R^n \to R$ that can decrease information loss, that is, $L(H) := EL(y, H(X))$, where L is the smoothness error function and (X, y) denotes the testing samples from D. The gradient boosting approach builds a greedy series of approximations Ht:RmR, t = 0,1,2…, Ht = H((t−1)) + gt is the final function produced from prior approximation using an additive process Ht = H((t−1)) + gt.

$$g^t = arg\ min_{g \in G} L(H^{t-1} + g) = arg\ min_{g \in G} E\ L\left(y, H^{t-1}(X)\right) \tag{5}$$

In general, greedy techniques, such as Newton's method, employing a second-order approach of $L(H(t-1) + g)$ at $H(t-1)$ or adopting (negative) gradient stages, are used to address the optimization issue.

3.3. Rainfall-runoff inundation model

The RRI model was developed by the International Center for Water Hazard and Risk Management in Japan. It is a 2D model distributed hydrological capable of simulating the rainfall-runoff and flood inundation simultaneously (Sayama et al., 2012). The RRI model has been successfully applied in many studies worldwide (Abdel-Fattah et al., 2018; Perera et al., 2017; Saber et al., 2020; Tam et al., 2019; Try et al., 2020). In this study, the model was calibrated and validated based on the typhoon of 2020, showing acceptable results with the actual flood discharge and good agreement with flood inundation maps. The final flood inundation map produced by the model was used for comparison with the ML FSMs.

3.4. Evaluation of the model's performance validation

The receiver operating characteristic (ROC) curve measure is a commonly used and validated strategy for assessing the reliability of a model in geospatial research (Chen et al., 2020; Tehrany et al., 2013). The ROC curve is the most commonly used mechanism for analyzing flood susceptibility and landslide approaches. The classification performance of a given technique was evaluated using the AUC in several previous studies (Bui et al., 2012; Youssef et al., 2016; Youssef and Hegab, 2019). A high classification efficiency for a given classification model should have an AUC-ROC value of 0.5 to 1, and the model's performance is enhanced by boosting the AUC-ROC scores. When the AUC-ROC value was close to 1.0, the models offered the best rate of precision and consistency. This demonstrates the ability of the model to forecast the occurrence of disasters without bias (Bui et al. 2012). In this study, the ROC score was determined using the following formula (Chang et al., 2018):

Other quantitative metrics (accuracy, recall, precision, and F1-score) were employed to check the model performance and compare its classification ability with that of its counterpart models in the literature. Accuracy is defined as the ratio of correctly classified data to total observations [Eq. (6)]; precision can be defined by the ratio of properly positive classified data to total positive data [Eq. (7)]. Recall, which is also known as sensitivity, is defined by the ratio of positive observations to the total observations [Eq. (8)]. F1-score uses weighted averaging for both the precision and recall [Eq. (9)].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1\ score = \frac{2\ (Recall * Precision)}{Recall * Precision} \quad (9)$$

where true positive (TP) represents a properly categorized flooded pixel, true negative (TN) represents a correctly categorized non-flood pixel, false positive (FP) indicates the number of pixels categorized incorrectly as flood pixels, and false negative (FN) indicates the number of pixels categorized incorrectly as non-flood pixels.

4. Results and Discussion

4.1. Multicollinearity assessment and feature selection

According to (Chen et al., 2020), a value greater than 0.7 indicates a strong correlation between variables. This value was adopted in this study to detect the existence of a correlation

between the flood-influencing factors. Ten of the conditioning factors (DEM, NDVI, flow accumulation, vertical distance from the river, and slope) were identified as correlated with each other (Table 1). The VIF of the vertical distance from the river (= 12), DEM (= 10.5), SPI (= 7.7), and flow accumulation (= 7.4) factors were greater than the threshold value (> 5), which indicates a problem of multicollinearity (Fig. 7a).

To formulate an opinion on the importance of influencing factors in relation to flood generation, the IGR scores were computed and are illustrated in Fig. 7b. According to the results, most of the factors had an IGR greater than 0.05, and only four of them had an inferior IGR, that is, flow accumulation, flow direction, rainfall, and aspect.

The selection of conditioning factors was performed as follows:

1) Based on multicollinearity analysis, the vertical distance from the river, DEM, SPI, and flow accumulation factors were removed from the selection list.
2) Using the IGR as a selection criterion, flow direction, rainfall, and aspect were also removed because their IGR was almost equal to zero.
3) After removing the aforementioned factors, only the slope and the topographic wetness index (TWI) remained as correlated variables. By comparing the IGR (Fig. 7), we find that the slope factor is more important than the TWI in relation to flood generation. Therefore, the slope factor was selected for flood prediction based on the normalized difference vegetation index (NDVI), land use, curvature, geology, hillshade, and horizontal distance from the river.

Table 1. Spearman's correlation coefficients for FFS mapping.

	Aspect	**Rainfall**	**Curvature**	**DEM**	**NDVI**	**Flow acc.**	**Flow dir.**	**H. Dist.**	**V. Dist.**	**Geology**	**Hillshade**	**Land use**	**Slope**	**SPI**	**TWI**
Aspect	1.00														
Rainfall	-0.03	1.00													
Curvature	0.08	0.01	1.00												
DEM	0.03	0.20	0.27	1.00											
NDVI	0.04	-0.03	0.20	0.72	1.00										
Flow acc.	-0.07	0.01	-0.48	0.05	0.06	1.00									
Flow dir.	0.06	-0.07	0.11	0.06	0.06	-0.11	1.00								
H. Dist.	-0.04	0.03	0.02	0.49	0.39	0.08	0.02	1.00							
V. Dist.	0.04	0.17	0.28	0.97	0.72	0.00	0.06	0.52	1.00						
Geology	0.00	0.29	-0.02	0.08	-0.02	-0.01	-0.03	-0.09	0.06	1.00					
Hillshade	0.38	-0.05	0.08	0.03	0.02	-0.05	0.51	-0.06	-0.03	-0.06	1.00				
Land use	-0.03	-0.10	-0.11	0.59	-0.53	-0.13	-0.05	-0.25	-0.58	-0.07	0.00	1.00			
Slope	0.10	0.07	0.28	0.85	0.69	0.03	0.08	0.38	0.86	0.06	0.01	-0.53	1.00		
SPI	0.04	0.03	-0.41	0.32	0.27	0.86	-0.06	0.21	0.30	-0.01	-0.05	-0.27	0.38	1.00	
TWI	-0.10	-0.07	-0.49	0.76	-0.61	0.35	-0.12	-0.31	-0.79	-0.07	-0.04	0.45	-0.90	0.00	1.00

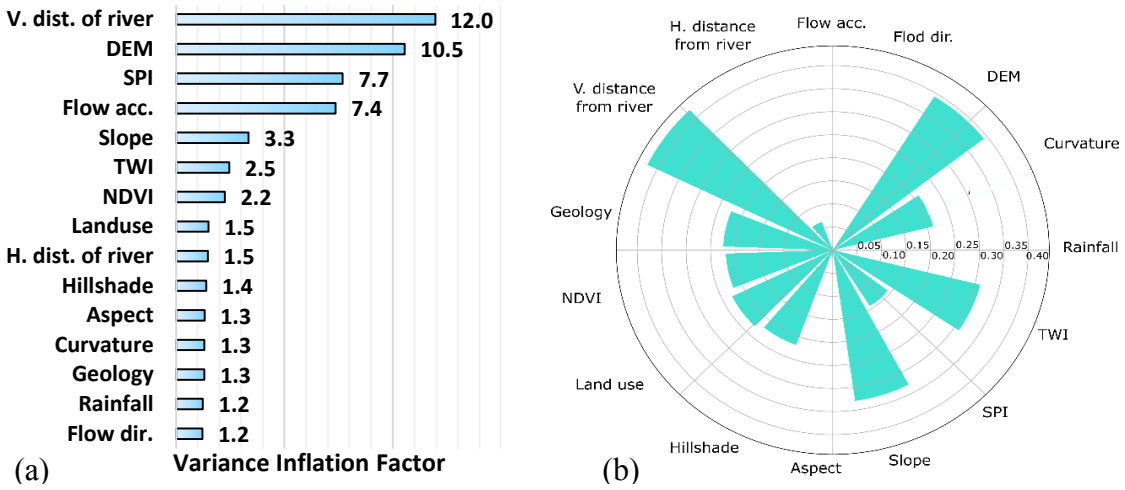


Fig. 7. Analysis of influencing factors: (a) VIF and (b) IGR for flood susceptibility.

4.2. Evaluation of the models

This section provides a detailed evaluation and comparison of all models developed in this study with respect to various classification criteria. The learning phase was performed using the K-fold cross-validation strategy. The examined data were partitioned into a learning set (60%), and the rest were deployed to evaluate the precision. The learning set was separated into two sets of training data (80%), which were used to adjust model weights and minimize classification errors, and validation data, which were used for hyperparameter tuning. Through the grid search approach, the appropriate hyperparameters for each classification technique were chosen. During the procedure, a wide variety of hyperparameter values was examined. Table 2 lists the optimal designs for each classifier.

The accuracy rates of all the studied models are listed in Table 3. As can be seen, all developed classification techniques achieved approximately identical results in terms of statistical metrics. The LightGBM model slightly outperformed the others in terms of speed convergence and classification metrics. The ROC curve of the generated models on the test

ensembles is displayed in Fig. 8, which reveals that the three suggested boosting strategies have similar qualities and provide significant accuracy. The maximum AUC was reached by LightGBM and RF models with the same score (99.5%), and CatBoost was ranked as the worst model with an AUC of 97.9%.

Furthermore, CatBoost scored the first rank in terms accuracy performance accuracy equal to 97.8 %, precision equal to 96%, accompanied by LightGBM classifier with an accuracy of 97.3% and precision of 95%. Finally, the RF model was ranked as the last classifier model with an accuracy equal to 95.5%, and precision of 96.2%. In comparison with previous studies, RF in this study outperformed many of the previous applications, including (e.g., AUC = 0.925, Chen et al. (2020); AUC = 0.886, Tang et al. (2020); AUC = 0.7878, Lee et al. (2017); AUC = 0.972, Achour and Pourghasemi (2020)).

The confusion matrix in Fig. 9. shows the performance of the used models in the study area, where CatBoost shows better prediction followed by LightGBM and, finally, the RF methods; however, all of them display acceptable prediction.

Two novel boosting classification models were examined in this study for FFS assessment in the VGTB River Basin. From the evaluation statistics, we can conclude that the LightGBM and CatBoost models proved their performance for FFS and can be used as essential tools for real-time application compared to their counterpart models because of their high performance and speed convergence.

Table 2. Parameter values of random forest, CatBoost, and LightGBM models.

METHOD	HYPERPARAMETER	GRID SEARCH VALUES	SELECTED VALUE
RF	Max depth	[2:2:40]	22
	Number of trees	[100:100:2000]	800

LIGHTGBM	Learning rate	[0.01:0.01:0.1]	0.09
	Max depth	[2:2:40]	32
	Number of leaves	[20:20:300]	220
	Min data	[5:5:100]	25
CATBOOST	Learning rate	[0.01:0.01:0.1]	0.07
	Max depth	[2:2:40]	6
	Leaf estimation iterations	[1:1:10]	8
	l2 leaf reg	Log [10(-21),10(-8)]	Log (10(-18))

582

583

584 Table 3. Statistical measures used for the model performance evaluation.

CRITERIA	**RANDOM FOREST**		**LIGHTGBM**		**CATBOOST**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
ACCURACY	0.999	0.955	0.999	0.973	0.999	0.978
PRECISION	0.998	0.962	0.998	0.950	0.998	0.960
RECALL	1.000	0.932	1.000	0.990	1.000	0.990
F1_SCORE	0.999	0.947	0.999	0.969	0.999	0.974
AUC	1.000	0.995	1.000	0.995	0.999	0.979

585

In this study, two new boosting classification techniques were investigated for flood susceptibility projection in the VGTB. This is the first work that investigates the use of CatBoost and LightGBM for flood classification in humid environments against the commonly used RF models. The obtained results revealed that LightGBM outperformed its counterpart ML models, especially in terms of classification metrics and processing time. This agrees with the findings of Saber et al. (2021) that LightGBM has proven its efficiency in flash flood prediction and outperforms the other two methods in terms of classification and processing time. In addition, it was stated that LightGBM outperformed other methods such as the RF, M5Tree, and other empirical models for estimating daily evapotranspiration in China as a humid subtropical region (Fan et al. 2019). Similarly, it was also found that LightGBM performed better than the others in terms of AUC (99.5%). The accuracy of CatBoost (97.9%) was also high compared to the previous studies in other fields. Among other methods, CatBoost, SVM, and RF have been applied to evapotranspiration modeling in China (Huang et al. 2019). They stated that CatBoost presented higher accuracy and lower computational cost than the other approaches (RF and SVM).

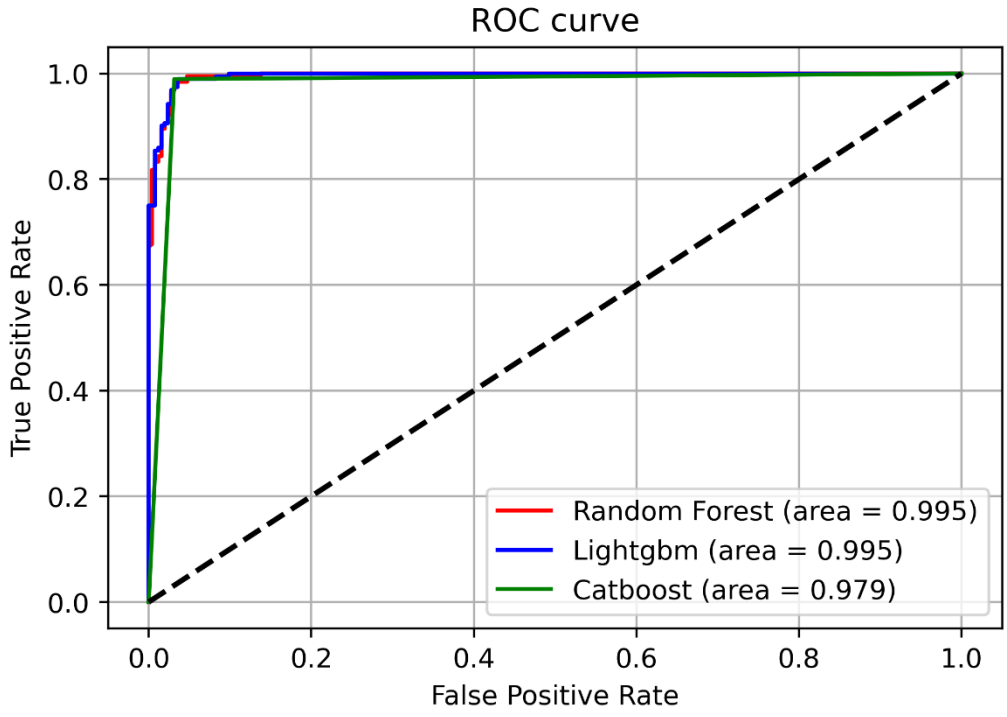


Fig. 8. Performance of random forest, CatBoost, and LightGBM models based on the area under the receiver operating characteristic (ROC) curves.

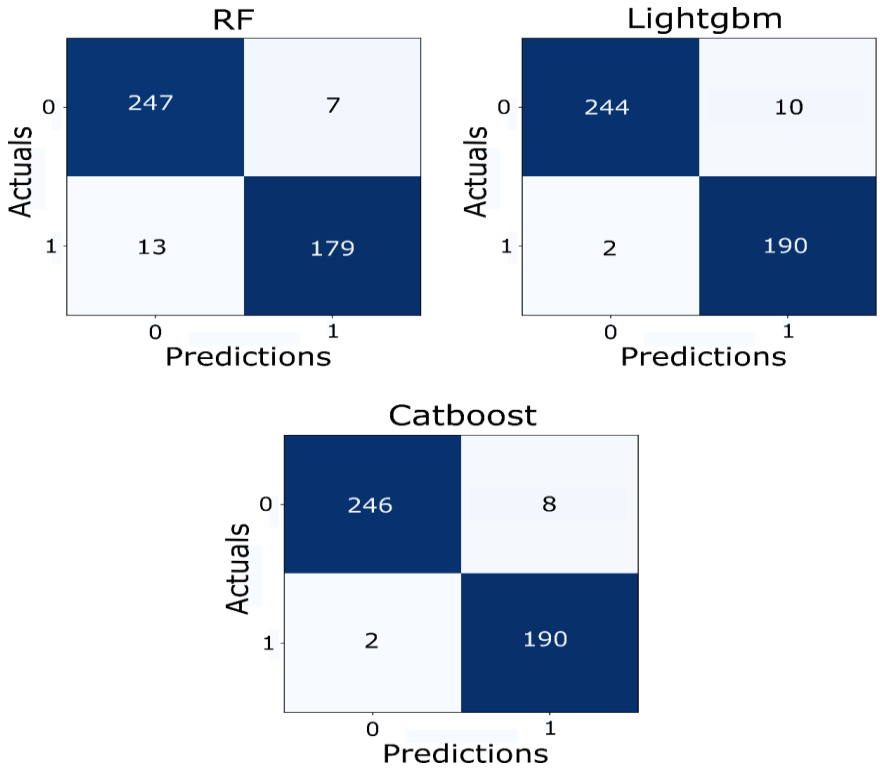


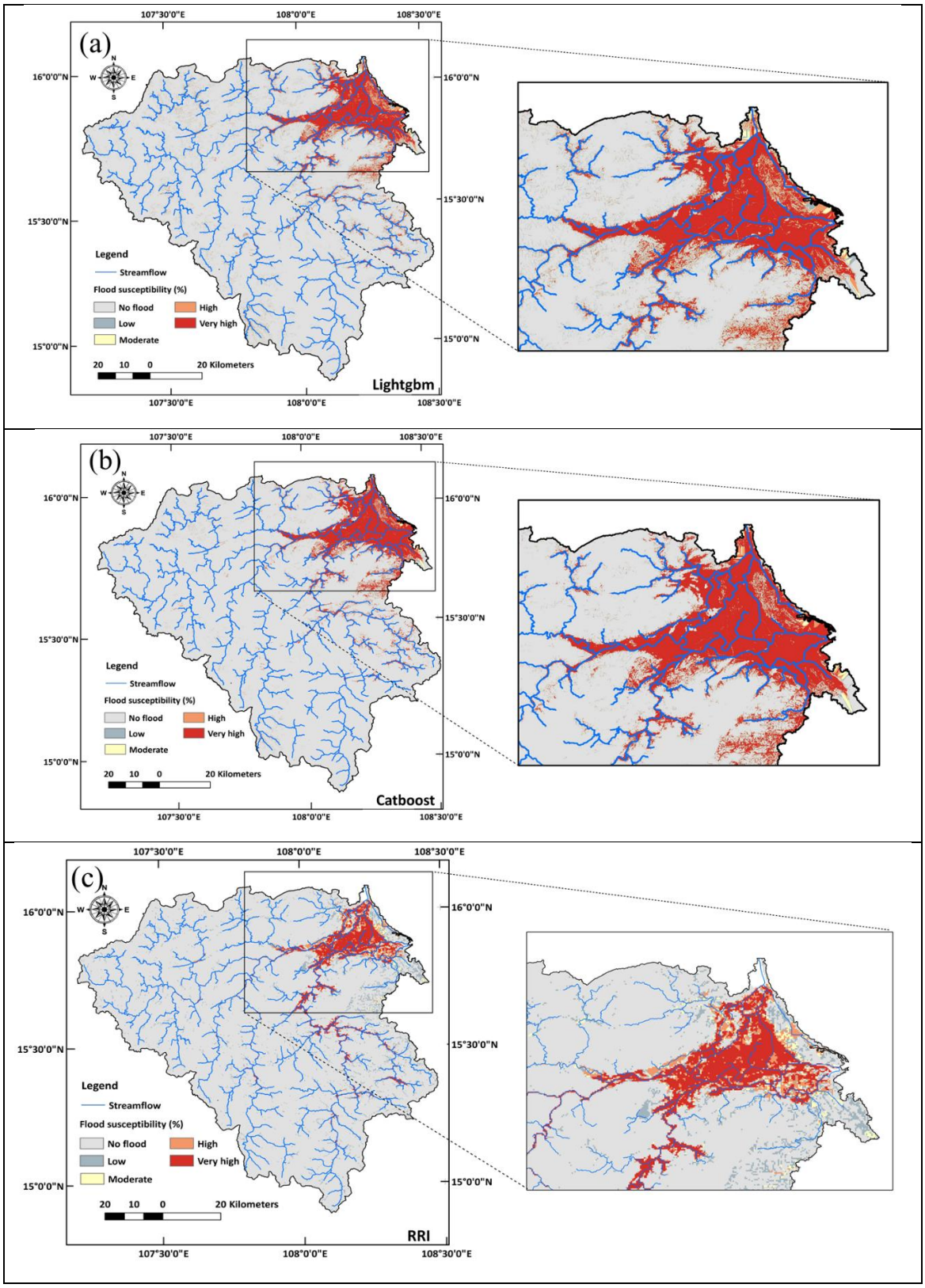
Fig. 9. Confusion matrix showing the performance of the used models in VGTB River Basin.

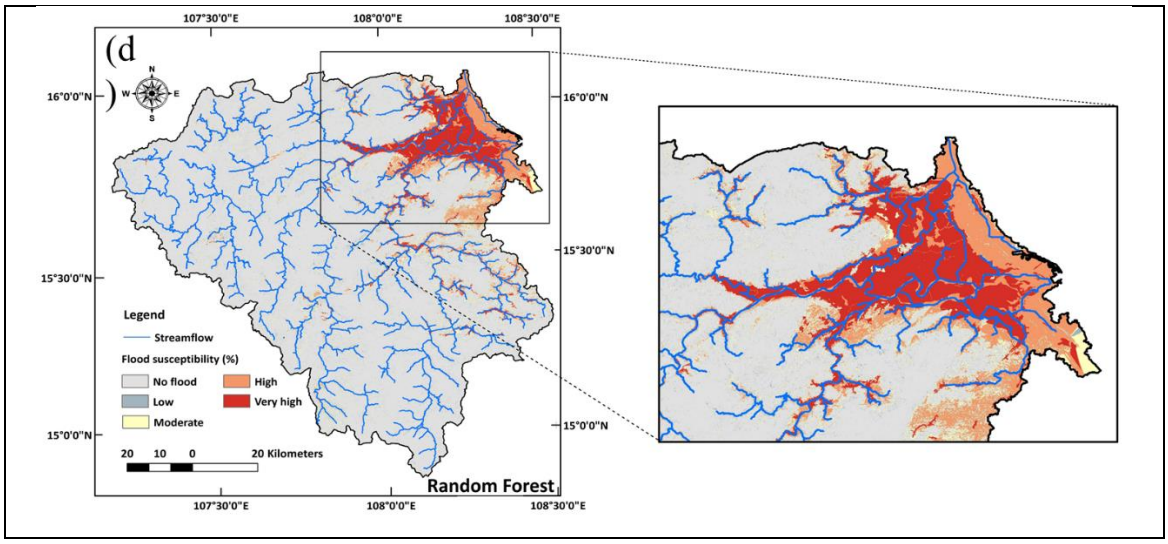
4.3. Flash Flood Susceptibility Modeling

The evaluation metrics of the newly applied boosting methods (CatBoost and LightGBM) along with RF demonstrated their high performance in predicting flooding in a humid climate environment. Accordingly, these methods were employed to estimate the flood susceptibility maps for the entire VGTB river basin. The three FSMs developed using the three methods

(RF, LightGBM, and CatBoost) were compared with the FIM of the RRI model, as shown in Fig. 10. The flooding susceptibility values were then mapped under five levels of susceptibility classes: no flood, low, moderate, high, and very high. The regions affected by these different susceptibility levels varied slightly depending on the model.

The FSMs by the three models showed that the areas of high and very high susceptibility to flooding to be 13% (RF), 11% (LightGBM), and 10% (CatBoost) of the study area, which agrees with the FIM developed by RRIat approximately 11%. This level of susceptibility is predominant in the coastal and plane areas along the Vu Gia and Thu Bon Rivers (Fig. 10 and 11). The spatial distributions of the high and very high levels were similar in all the maps produced by the ML and RRI models. The areas affected by a moderate level of susceptibility to flooding (Fig. 11) were estimated at 10% (RF), 0% (LightGBM), and 1% (CatBoost), indicating that both LightGBM and CatBoost are more similar to the RRI model which shows a value of approximately 1%. The areas affected by the low level of susceptibility to flooding (Fig. 11) were estimated at 36% (RF), 1% (LightGBM), and 2% (CatBoost), also revealed that both LightGBM and CatBoost are performed better, with good agreement with the RRI model showing the value of approximately 3%. It was also found that the areas that were not subjected to the flash floods were approximately 42% (RF), 83% (LightGBM), and 87% (CatBoost) of the total study area (Fig. 11), showing good agreement with RRI model that shows approximately 90%. However, the performances of the employed models are almost the same and the two new methods of LightGBM and CatBoost outperform RF in terms of the spatial coverage of the flood susceptibility levels in comparison with the RRI model. The RF overestimated the low flood susceptibility in the study area. All methods show an agreement in terms of the same spatial pattern of FSM, highlighting that the coastal areas are the areas prone to flooding where most of the residential and agricultural regions are located.


(a)
107°30'0"E
108°0'0"E
108°30'0"E
16°0'0"N
15°30'0"N
15°0'0"N
Legend
Streamflow
Flood susceptibility (%)
No flood
Low
Moderate
High
Very high
20 10 0 20 Kilometers
Lightgbm
(b)
Catboost
(c)
RRI



640

641 Fig. 10. Flood susceptibility maps by LightGBM (a), CatBoost (b), RF (c), and RRI (d),
642 respectively, from the top to the bottom.

643

644

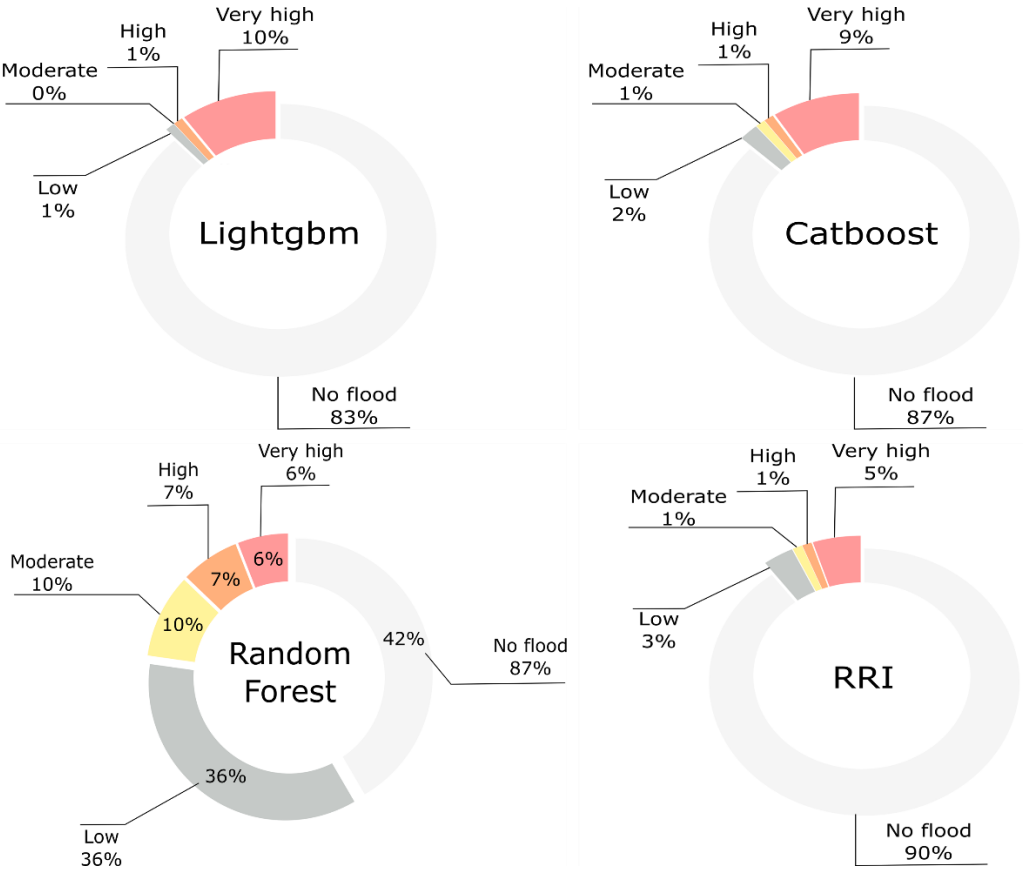


Fig. 11. Affected area of the flood susceptibility levels for the three applied ML methods and flood inundation map of RRI model.

4.4. Testing different sizes of the datasets

In this section, we tested different sizes of the datasets, including flooded and non-flooded points (1250, 1000, 800, 600, 400, 90, 60, and 30) of the training model (Fig. 12). The training datasets classified as 50% and 50% for flooded and non-flooded points, respectively; however, the testing datasets were the same during the simulation (Fig. 12). We found that accuracy scores for all the models and all the tested cases were greater than 90% (Fig. 13), except for dataset sizes of 60 and 30 points in LightGBM. The accuracy score slightly decreases with the decrease in the datasets in both the LightGBM and CatBoost models, but

is not consistent in the RF model. This implies that the ML approaches employed in this study can effectively work with very limited training datasets with a slight decrease in accuracy, which will be applicable for ungauged or regions with very low monitoring and observations of flooding occurrences and impacts. The FSMs developed based on different training datasets show that most of the spatial maps are acceptable as overall spatial coverage; however, there are some small spatial differences in the susceptibility flooding levels (Figs. 14 and 15). For instance, the affected areas (Figs. 14 and 15) in the very high flood susceptibility flood category are almost the same about 6% for all datasets (1250, 1000, 800, 600, 400, 90, 60, and 30), except for dataset of 200, which was approximately 5%. On the other hand, the affected areas by the high flood susceptibility level are also varying but the highest percentage was 9% for the 200 and 60 datasets and the lowest was 6% for 30, 90, 600, and 800 datasets. The variation in moderate flood occurrence ranged from 17% to 9%. The dataset size of 30 was the highest among the others, about 17%. The range of the low flood susceptibility category was highly variable from 20% to 41%, the lowest was for the dataset of 30, and the highest was for the dataset of 800. The reasons for such variation are probably the random selection of the flooded samples, which in some cases are not representative of all the influencing factors. In general, we noticed that the spatial coverage was not extremely different, but some differences based on the categories were observed. The areas with no flood levels are also changeable by about 42%, 42%, 38%, 45%, 44%, 45%, 42%, 44%, and 51% for the datasets of 1250, 1000, 800, 600, 400, 200, 90, 60, and 30, respectively. Interestingly, the highest percentage was recorded by the dataset of 30 points and the lowest was recorded by dataset 800. The results of the analysis of different data sizes for ML training show that ML can effectively predict the flood susceptibility maps in the study area regardless of the number of samples, with the condition of the used data being observational flooded sites.

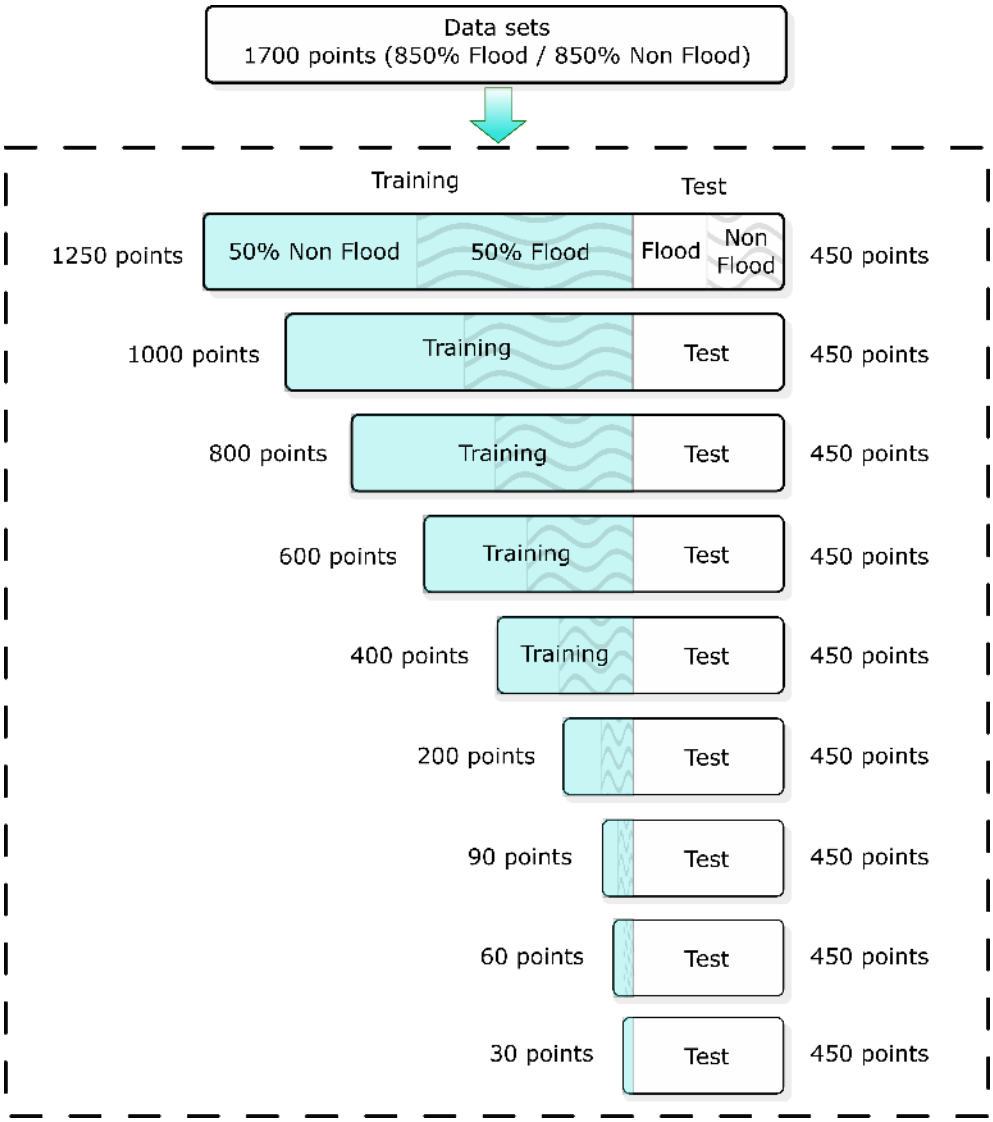
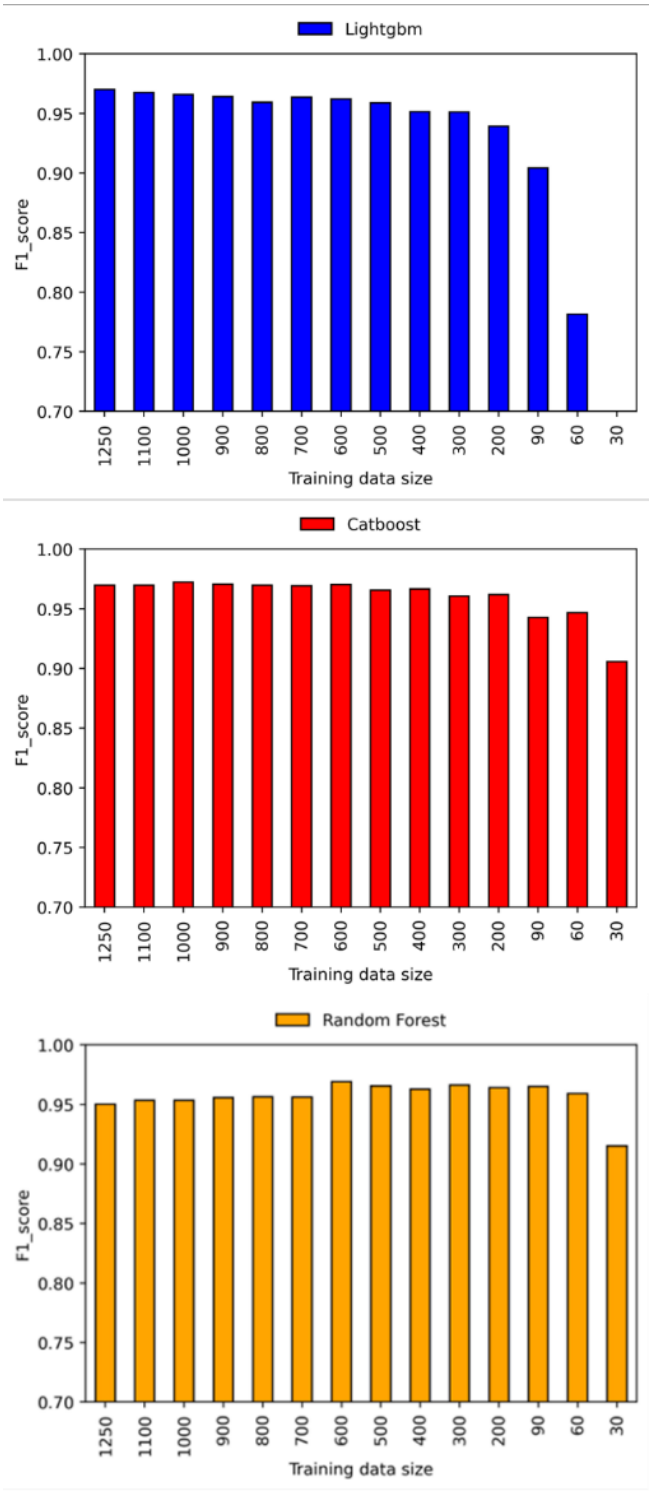


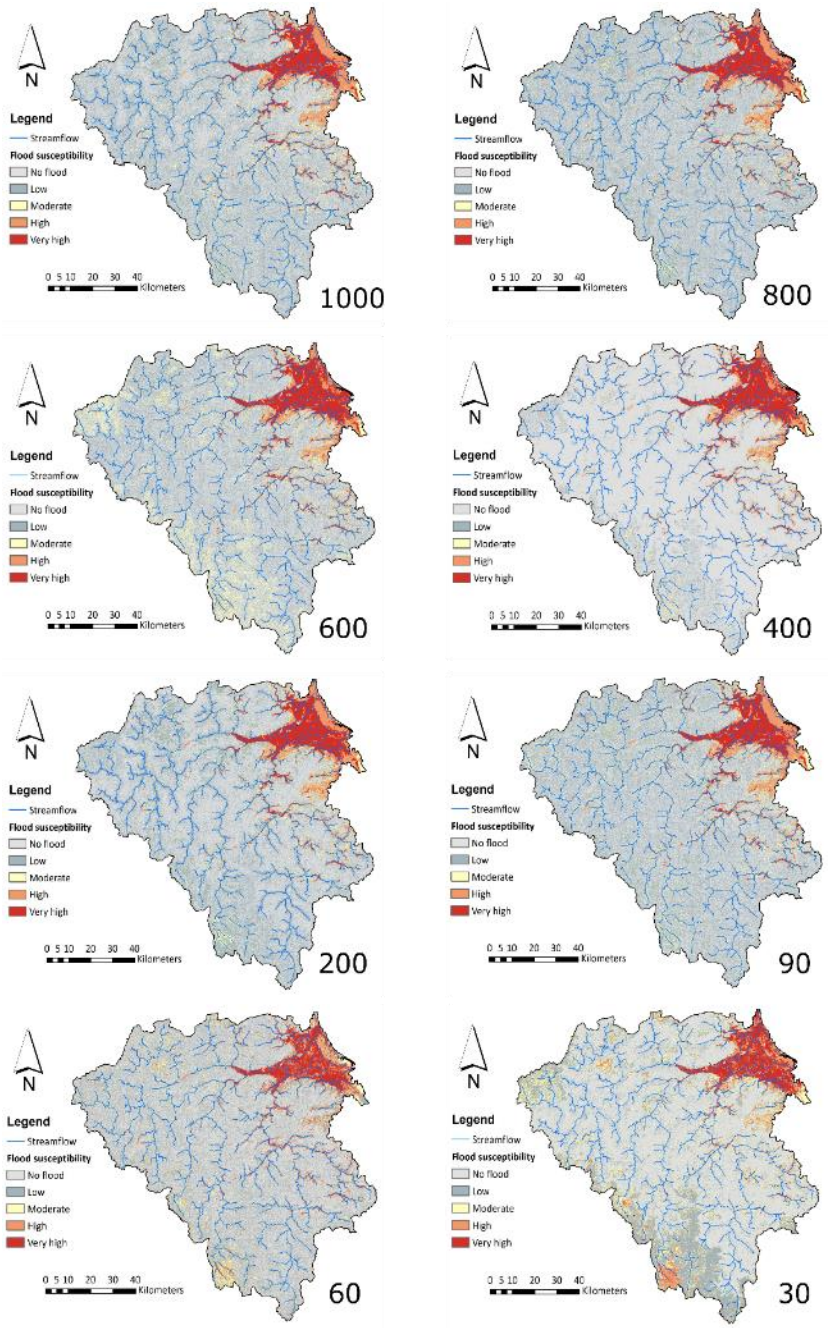
Fig. 12. Datasets used in the training and testing of the ML models.



687

688 Fig. 13. Accuracy of the models based on different training datasets.

689



690

691 Fig. 14. Impact of data size on flood susceptibility map.

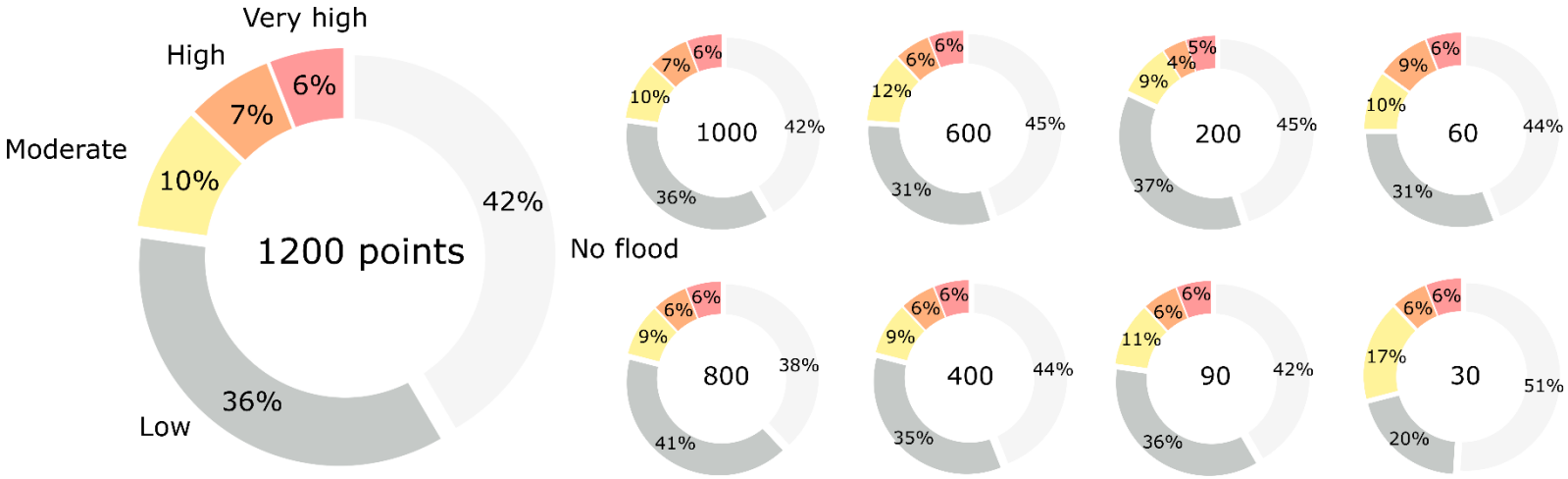


Fig. 15. Percentage of the affected areas under different flood susceptibility classes using different dataset sizes (RF method).

4.5. Discussion and comparison with results of RRI model

Currently, there is increasing interest in the use of data-driven methods as appropriate alternatives to hydrological and hydraulic models. Therefore, the scientific community of flood risk assessment is endeavoring to develop much more logical and mathematical methods to predict flood-susceptible areas at different catchment scales (Arora et al., 2020). In the study area, there have been some previous studies on flood susceptibility mapping that use ML approaches and deep learning. Therefore, testing different methods is highly recommended, particularly in regions where data are limited and hydrological models are challenging. This study provides the application of three ML methods: RF, LightGBM, and CatBoost. The latter two methods were tested for FFS mapping for the first time in this humid region with a high frequency of typhoon events. The results of the FFS maps prove that both methods can predict flood-prone areas with acceptable accuracy in comparison with the RF method, which is widely applicable. RFs are applied in many other related studies with different accuracies, e.g., AUC = 78% (Band et al., 2020), 99.3% (Li et al., 2019), 94.5% (Talukdar et al., 2020), 93.8% (Park and Lee, 2020), and 89.4% (Nguyen et al., 2018). In this study AUC = 99% for RF was higher than most of the previous studies. Additionally, the newly applied methods of LightGBM and CatBoost showed almost the same accuracy of 99% and 98%, respectively, revealing better performance than most of the previous studies. These three methods have been tested in Hurghada, Egypt (Saber et al., 2021), stating that LightGBM has the advantage of better classification metrics and fast processing time, and outperforms other methodologies such as CatBoost and RF. In addition, their results showed that LightGBM and CatBoost have proven their efficiency in flash flood prediction in arid regions.

Additionally, the three methods exhibited better performance than the average of the previously applied methods for flood susceptibility mapping, which was 90% as an average of approximately 140 previous applications from more than 30 publications that have been reviewed. The performance of the previous methods applied for FSM based on AUC varies from 64% (Shafizadeh-Moghadam et al., 2018) to 99.3% (Li et al., 2019). LightGBM has

also been applied in other applications, such as the real-time prediction of regional urban runoff (https://doi.org/10.1016/j.jhydrol.2021.127124), showing high accuracy and fast processing predictions. CatBoost was also applied in Germany, with better performance in comparison with other methods, showing good accuracy with an AUC of 0.816 (Kaiser, M. H. E. (2021)).

The maps of flood susceptibility developed using ML techniques (Fig. 16) showed an acceptable fit with the developed flood inundation map by the RRI model, showing that the ML approaches are promising for flood prediction and can be used without detailed observations and challenges of model calibrations as alternative tools for hydrological models. The results of LightGBM and CatBoost are more comparable to the flood inundation map developed by the physical RRI model, indicating that they are more acceptable than RF, which overestimates the low flood susceptibility level in the study area.

Furthermore, we tested different datasets for the training of the three ML models, with the conclusion that datasets greater than 90 points can be sufficiently accurate for reasonable prediction of the FSM. LightGBM and CatBoost showed a slightly declining trend in the accuracy of the results based on the dataset sizes; however, RF did not show such a trend. These results are highly valuable for the application of ML to ungauged basins with very limited datasets.

5. Conclusions

Flooding resulting from typhoons is one of the most threatening disasters in Asian countries and worldwide. Therefore, the present study introduced three ML methods to accurately predict flooding susceptibility in humid areas in Vietnam. The first method is RF, which is well known and widely applicable in many applications including FSM, and the other two methods (LightGBM and CatBoost) were examined for the first time for FSM in this humid region. The methods were trained and validated based on a flood inventory map and 10 influencing flood factors. Owing to the availability of high-quality observations, we

754 also tested different datasets for the training (i.e., 30, 60, 90, 200, 400, 600, 800, 1000, and
755 1250 data points) to determine the least data points that provided acceptable reliability, as
756 well as to understand the differences in the spatial FSMs in the study area. Interestingly, we
757 found that the accuracy of results based on all the tested datasets was higher than 90%,
758 indicating that a limited number of observations can be used efficiently in terms of model
759 accuracy, although the final FSMs showed some differences spatially from one susceptibility
760 level to the others. This finding is highly important to demonstrate that ML methods can
761 work efficiently with an acceptable level of accuracy within a small number of actual training
762 datasets. The conclusions of this study can be summarized as follows:

763 - We applied three ML models—RF, LightGBM, and CatBoost—to predict flood
764 susceptibility in humid areas that experienced successive extreme typhoons.
765 - The LightGBM and CatBoost models were tested for the first time in this specific
766 climatic region, and showed high performance in comparison with the RF method.
767 - The results of the ML methods showed good agreement with the rainfall-runoff model
768 for flood inundation mapping, especially the LightGBM and CatBoost models in
769 terms of coverage areas of the flood susceptibility levels.
770 - Different training datasets were examined to determine the lowest acceptable number
771 of observations for flood susceptibility in ML.
772 - The FSMs demonstrated that downstream areas with high residential and agricultural
773 activity are highly susceptible to flooding.
774 - The outcomes of this study could be used as guidance and reference for flood risk
775 mitigation and management in this region and, consequently, to assist managers,
776 decision makers, and planners in effectively managing and mitigating floods in high-
777 flood-susceptible areas.

778 As an extension of this research, we are examining ML for predicting flood depth, as a
779 comparison between ML approaches and RRI models. The study concludes that the ML
780 approach can be an alternative for hydrological models in terms of flood susceptibility
781 mapping; however, we are still working to prove this with flood depth prediction.

Acknowledgment:

Data sources, such as the Geospatial Information Authority of Japan, Chiba University, and collaborating organizations.

Funding:

This work was funded by the Asia-Pacific Network for Global Change Research (APN) under project reference number CRRP2020-09MYKantoush (Funder ID: https://doi.org/10.13039/100005536).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

References

[1] Abdel-Fattah, M., Kantoush, S.A., Saber, M., Sumi, T., 2018. Rainfall-runoff modeling for extreme flash floods in wadi samail, oman. J. Jpn. Soc. Civ. Eng. Ser B1 Hydraul. Eng. 74, I_691-I_696.

[2] Abdrabo, K.I., Kantoush, S.A., Saber, M., Sumi, T., Habiba, O.M., Elleithy, D., Elboshy, B., 2020. Integrated Methodology for Urban Flood Risk Mapping at the Microscale in Ungauged Regions: A Case Study of Hurghada, Egypt. Remote Sens. 12, 3548. https://doi.org/10.3390/rs12213548

[3] Abushandi, E.H., Merkel, B.J., 2011. Application of IHACRES rainfall-runoff model to the Wadi Dhuliel arid catchment, Jordan. J. Water Clim. Change 2, 56–71.

[4] Akay, A.E., Taş, İ., 2020. Mapping the risk of winter storm damage using GIS-based fuzzy logic. J. For. Res. 31, 729–742.

[5] Ali, S.A., Parvin, F., Pham, Q.B., Vojtek, M., Vojteková, J., Costache, R., Linh, N.T.T., Nguyen, H.Q., Ahmad, A., Ghorbani, M.A., 2020. GIS-based comparative assessment of flood susceptibility mapping using hybrid multi-criteria decision-making approach, naïve Bayes tree, bivariate statistics and logistic regression: A case of Topľa basin, Slovakia. Ecol. Indic. 117, 106620.

[6] Arabameri, A., Saha, S., Mukherjee, K., Blaschke, T., Chen, W., Ngo, P.T.T., Band, S.S., 2020. Modeling Spatial Flood using Novel Ensemble Artificial Intelligence Approaches in Northern Iran. Remote Sens. 12, 3423. https://doi.org/10.3390/rs12203423

[7] Arora, A., Arabameri, A., Pandey, M., Siddiqui, M.A., Shukla, U.K., Bui, D.T., Mishra, V.N., Bhardwaj, A., 2020. Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga

Plain, India. Sci. Total Environ. 750, 141565. https://doi.org/10.1016/j.scitotenv.2020.141565
[8] Arora, A., Pandey, M., Siddiqui, M.A., Hong, H., Mishra, V.N., 2019. Spatial flood susceptibility prediction in Middle Ganga Plain: comparison of frequency ratio and Shannon's entropy models. Geocarto Int. 1–32.
[9] Aryal, S.K., Mein, R.G., O'Loughlin, E.M., 2003. The concept of effective length in hillslopes: assessing the influence of climate and topography on the contributing areas of catchments. Hydrol. Process. 17, 131–151.
[10] Avitabile, V., Schultz, M., Herold, N., De Bruin, S., Pratihast, A.K., Manh, C.P., Quang, H.V., Herold, M., 2016. Carbon emissions from land cover change in Central Vietnam. Carbon Manag. 7, 333–346.
[11] Bachmair, S., Svensson, C., Prosdocimi, I., Hannaford, J., Stahl, K., 2017. Developing drought impact functions for drought risk management. Nat. Hazards Earth Syst. Sci. 17, 1947–1960.
[12] Band, S.S., Janizadeh, S., Chandra Pal, S., Saha, A., Chakrabortty, R., Melesse, A.M., Mosavi, A., 2020. Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms. Remote Sens. 12, 3568. https://doi.org/10.3390/rs12213568
[13] Bellu, A., Fernandes, L.F.S., Cortes, R.M., Pacheco, F.A., 2016. A framework model for the dimensioning and allocation of a detention basin system: The case of a flood-prone mountainous watershed. J. Hydrol. 533, 567–580.
[14] Bisht, S., Chaudhry, S., Sharma, S., Soni, S., 2018. Assessment of flash flood vulnerability zonation through Geospatial technique in high altitude Himalayan watershed, Himachal Pradesh India. Remote Sens. Appl. Soc. Environ. 12, 35–47.
[15] Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
[16] Bui, D.T., Hoang, N.-D., Pham, T.-D., Ngo, P.-T.T., Hoa, P.V., Minh, N.Q., Tran, X.-T., Samui, P., 2019a. A new intelligence approach based on GIS-based multivariate adaptive regression splines and metaheuristic optimization for predicting flash flood susceptible areas at high-frequency tropical typhoon area. J. Hydrol. 575, 314–326.
[17] Bui, D.T., Ngo, P.-T.T., Pham, T.D., Jaafari, A., Minh, N.Q., Hoa, P.V., Samui, P., 2019b. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping. CATENA 179, 184–196. https://doi.org/10.1016/j.catena.2019.04.009
[18] Bui, D.T., Panahi, M., Shahabi, H., Singh, V.P., Shirzadi, A., Chapi, K., Khosravi, K., Chen, W., Panahi, S., Li, S., 2018. Novel hybrid evolutionary algorithms for spatial prediction of floods. Sci. Rep. 8, 1–14.
[19] Bui, D.T., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., 2012. Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models. Catena 96, 28–40.
[20] Bui, Q.-T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., Pham, V.-M., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. J. Hydrol. 581, 124379. https://doi.org/10.1016/j.jhydrol.2019.124379
[21] Cardenas, M.B., Wilson, J., Zlotnik, V.A., 2004. Impact of heterogeneity, bed forms, and stream curvature on subchannel hyporheic exchange. Water Resour. Res. 40.

[22]Catal, C., Diri, B., 2009. Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. Inf. Sci. 179, 1040–1058.
[23]Chang, M.-J., Chang, H.-K., Chen, Y.-C., Lin, G.-F., Chen, P.-A., Lai, J.-S., Tan, Y.-C., 2018. A support vector machine forecasting model for typhoon flood inundation mapping and early flood warning systems. Water 10, 1734.
[24]Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T., Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ. Model. Softw. 95, 229–245.
[25]Chen, T.-H.K., Qiu, C., Schmitt, M., Zhu, X.X., Sabel, C.E., Prishchepov, A.V., 2020. Mapping horizontal and vertical urban densification in Denmark with Landsat time-series from 1985 to 2018: A semantic segmentation solution. Remote Sens. Environ. 251, 112096. https://doi.org/10.1016/j.rse.2020.112096
[26]Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci. Total Environ. 651, 2087–2096.
[27]Costache, R., Hong, H., Pham, Q.B., 2020a. Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. Sci. Total Environ. 711, 134514. https://doi.org/10.1016/j.scitotenv.2019.134514
[28]Costache, R., Popa, M.C., Tien Bui, D., Diaconu, D.C., Ciubotaru, N., Minea, G., Pham, Q.B., 2020b. Spatial predicting of flood potential areas using novel hybridizations of fuzzy decision-making, bivariate statistics, and machine learning. J. Hydrol. 585, 124808. https://doi.org/10.1016/j.jhydrol.2020.124808
[29]Darabi, H., Choubin, B., Rahmati, O., Haghighi, A.T., Pradhan, B., Kløve, B., 2019. Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. J. Hydrol. 569, 142–154.
[30]Demirel, M.C., Venancio, A., Kahya, E., 2009. Flow forecast by SWAT model and ANN in Pracana basin, Portugal. Adv. Eng. Softw. 40, 467–473.
[31]Devkota, K.C., Regmi, A.D., Pourghasemi, H.R., Yoshida, K., Pradhan, B., Ryu, I.C., Dhital, M.R., Althuwaynee, O.F., 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, 135–165.
[32]Dhara, S., Dang, T., Parial, K., Lu, X.X., 2020. Accounting for Uncertainty and Reconstruction of Flooding Patterns Based on Multi-Satellite Imagery and Support Vector Machine Technique: A Case Study of Can Tho City, Vietnam. Water 12, 1543. https://doi.org/10.3390/w12061543
[33]Dodangeh, E., Choubin, B., Eigdir, A.N., Nabipour, N., Panahi, M., Shamshirband, S., Mosavi, A., 2020. Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. Sci. Total Environ. 705, 135983. https://doi.org/10.1016/j.scitotenv.2019.135983
[34]Dorogush, A.V., Ershov, V., Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. ArXiv Prepr. ArXiv181011363.
[35]Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.-W., Khosravi, K., Yang, Y., Pham, B.T., 2019. Assessment of advanced random forest and decision tree

907 algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic
908 Island, Japan. Sci. Total Environ. 662, 332–346.
909 [36]Esfandiari, M., Jabari, S., McGrath, H., Coleman, D., 2020. FLOOD MAPPING USING RANDOM
910 FOREST AND IDENTIFYING THE ESSENTIAL CONDITIONING FACTORS; A CASE STUDY IN
911 FREDERICTON, NEW BRUNSWICK, CANADA. ISPRS Ann. Photogramm. Remote Sens. Spat.
912 Inf. Sci. 5.
913 [37]Fenicia, F., Kavetski, D., Savenije, H.H., Clark, M.P., Schoups, G., Pfister, L., Freer, J., 2014.
914 Catchment properties, function, and conceptual model representation: is there a
915 correspondence? Hydrol. Process. 28, 2451–2467.
916 [38]Gigović, L., Pamučar, D., Bajić, Z., Drobnjak, S., 2017. Application of GIS-interval rough AHP
917 methodology for flood hazard mapping in urban areas. Water 9, 360.
918 [39]Glenn, E.P., Morino, K., Nagler, P.L., Murray, R.S., Pearlstein, S., Hultine, K.R., 2012. Roles of
919 saltcedar (Tamarix spp.) and capillary rise in salinizing a non-flooding terrace on a flow-
920 regulated desert river. J. Arid Environ. 79, 56–65.
921 [40]González-Arqueros, M.L., Mendoza, M.E., Bocco, G., Castillo, B.S., 2018. Flood susceptibility
922 in rural settlements in remote zones: The case of a mountainous basin in the Sierra-Costa
923 region of Michoacán, Mexico. J. Environ. Manage. 223, 685–693.
924 [41]Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim,
925 H., Kanae, S., 2013. Global flood risk under climate change. Nat. Clim. Change 3, 816–821.
926 [42]Hölting, B., Coldewey, W.G., 2019. Surface water infiltration, in: Hydrogeology. Springer, pp.
927 33–37.
928 [43]Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.-X., Chen, W., 2018. Application of fuzzy
929 weight of evidence and data mining techniques in construction of flood susceptibility map
930 of Poyang County, China. Sci. Total Environ. 625, 575–588.
931 [44]Hsu, K., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall-
932 runoff process. Water Resour. Res. 31, 2517–2530.
933 [45]Humphrey, G.B., Gibbs, M.S., Dandy, G.C., Maier, H.R., 2016. A hybrid approach to monthly
934 streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial
935 neural network. J. Hydrol. 540, 623–640.
936 [46]Izquierdo-Verdiguier, E., Zurita-Milla, R., 2020. An evaluation of Guided Regularized Random
937 Forest for classification and regression tasks in remote sensing. Int. J. Appl. Earth Obs.
938 Geoinformation 88, 102051.
939 [47]Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A
940 highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3146–
941 3154.
942 [48]Khosravi, K., Nohani, E., Maroufinia, E., Pourghasemi, H.R., 2016. A GIS-based flood
943 susceptibility assessment and its mapping in Iran: a comparison between frequency ratio
944 and weights-of-evidence bivariate statistical models with multi-criteria decision-making
945 technique. Nat. Hazards 83, 947–987.
946 [49]Khosravi, K., Shahabi, H., Pham, B.T., Adamowski, J., Shirzadi, A., Pradhan, B., Dou, J., Ly, H.-
947 B., Gróf, G., Ho, H.L., Hong, H., Chapi, K., Prakash, I., 2019. A comparative assessment of
948 flood susceptibility modeling using Multi-Criteria Decision-Making Analysis and Machine
949 Learning Methods. J. Hydrol. 573, 311–323. https://doi.org/10.1016/j.jhydrol.2019.03.073

[50]Kia, M.B., Pirasteh, S., Pradhan, B., Mahmud, A.R., Sulaiman, W.N.A., Moradi, A., 2012. An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. Environ. Earth Sci. 67, 251–264.
[51]Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. Hydrol. Earth Syst. Sci. Discuss. 1–32.
[52]Li, X., Yan, D., Wang, K., Weng, B., Qin, T., Liu, S., 2019. Flood Risk Assessment of Global Watersheds Based on Multiple Machine Learning Models. Water 11, 1654. https://doi.org/10.3390/w11081654
[53]Luu, C., Pham, B.T., Phong, T.V., Costache, R., Nguyen, H.D., Amiri, M., Bui, Q.D., Nguyen, L.T., Le, H.V., Prakash, I., Trinh, P.T., 2021. GIS-based ensemble computational models for flood susceptibility prediction in the Quang Binh Province, Vietnam. J. Hydrol. 599, 126500. https://doi.org/10.1016/j.jhydrol.2021.126500
[54]Malekipirbazari, M., Aksakalli, V., 2015. Risk assessment in social lending via random forests. Expert Syst. Appl. 42, 4621–4631.
[55]Masood, M., Takeuchi, K., 2012. Assessment of flood hazard, vulnerability and risk of mid-eastern Dhaka using DEM and 1D hydrodynamic model. Nat. Hazards 61, 757–770.
[56]Meadows, M., Wilson, M., 2021. A Comparison of Machine Learning Approaches to Improve Free Topography Data for Flood Modelling. Remote Sens. 13, 275.
[57]Meraj, G., Khan, T., Romshoo, S.A., Farooq, M., Rohitashw, K., Sheikh, B.A., 2018. An integrated geoinformatics and hydrological modelling-based approach for effective flood management in the Jhelum Basin, NW Himalaya. Multidiscip. Digit. Publ. Inst. Proc. 7, 8.
[58]Nauditt, A., Firoz, A.B.M., Trinh, V.Q., Fink, M., Stolpe, H., Ribbe, L., 2017. Hydrological drought risk assessment in an anthropogenically impacted tropical catchment, Central Vietnam, in: Land Use and Climate Change Interactions in Central Vietnam. Springer, pp. 223–239.
[59]Nauditt, A., Ribbe, L., 2017. Land use and climate change interactions in central Vietnam. Springer.
[60]Ngo, P.-T.T., Pham, T.D., Hoang, N.-D., Tran, D.A., Amiri, M., Le, T.T., Hoa, P.V., Bui, P.V., Nhu, V.-H., Bui, D.T., 2021. A new hybrid equilibrium optimized SysFor based geospatial data mining for tropical storm-induced flash flood susceptible mapping. J. Environ. Manage. 280, 111858. https://doi.org/10.1016/j.jenvman.2020.111858
[61]Nguyen, H.Q., Degener, J., Kappas, M., 2015. Flash flood prediction by coupling KINEROS2 and HEC-RAS models for tropical regions of Northern Vietnam. Hydrology 2, 242–265.
[62]Nguyen, P., Ombadi, M., Gorooh, V.A., Shearer, E.J., Sadeghi, M., Sorooshian, S., Hsu, K., Bolvin, D., Ralph, M.F., 2020. PERSIANN Dynamic Infrared–Rain Rate (PDIR-Now): A Near-Real-Time, Quasi-Global Satellite Precipitation Dataset. J. Hydrometeorol. 21, 2893–2906.
[63]Nguyen, V.-N., Tien Bui, D., Ngo, P.-T.T., Nguyen, Q.-P., Nguyen, V.C., Long, N.Q., Revhaug, I., 2018. An Integration of Least Squares Support Vector Machines and Firefly Optimization Algorithm for Flood Susceptible Modeling Using GIS, in: Tien Bui, D., Ngoc Do, A., Bui, H.-B., Hoang, N.-D. (Eds.), Advances and Applications in Geospatial Technology and Earth Resources. Springer International Publishing, Cham, pp. 52–64. https://doi.org/10.1007/978-3-319-68240-2_4
[64]Nguyen, V.-N., Yariyan, P., Amiri, M., Dang Tran, A., Pham, T.D., Do, M.P., Thi Ngo, P.T., Nhu, V.-H., Quoc Long, N., Tien Bui, D., 2020. A New Modeling Approach for Spatial Prediction of

Flash Flood with Biogeography Optimized CHAID Tree Ensemble and Remote Sensing Data. Remote Sens. 12, 1373. https://doi.org/10.3390/rs12091373
[65] Nhu, V.-H., Thi Ngo, P.-T., Pham, T.D., Dou, J., Song, X., Hoang, N.-D., Tran, D.A., Cao, D.P., Aydilek, İ.B., Amiri, M., Costache, R., Hoa, P.V., Tien Bui, D., 2020. A New Hybrid Firefly–PSO Optimized Random Subspace Tree Intelligence for Torrential Rainfall-Induced Flash Flood Susceptible Mapping. Remote Sens. 12, 2688. https://doi.org/10.3390/rs12172688
[66] Öztürk, F., Akdeniz, F., 2000. Ill-conditioning and multicollinearity. Linear Algebra Its Appl. 321, 295–305.
[67] Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R., Church, J.A., Clarke, L., Dahe, Q., Dasgupta, P., 2014. Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Ipcc.
[68] Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222.
[69] Park, S.-J., Lee, D.-K., 2020. Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. Environ. Res. Lett. 15, 094052. https://doi.org/10.1088/1748-9326/aba5b3
[70] Perera, E.D.P., Sayama, T., Magome, J., Hasegawa, A., Iwami, Y., 2017. RCP8. 5-based future flood hazard analysis for the lower Mekong river basin. Hydrology 4, 55.
[71] Pham, B.T., Jaafari, A., Nguyen-Thoi, T., Van Phong, T., Nguyen, H.D., Satyam, N., Masroor, M., Rehman, S., Sajjad, H., Sahana, M., 2021a. Ensemble machine learning models based on Reduced Error Pruning Tree for prediction of rainfall-induced landslides. Int. J. Digit. Earth 14, 575–596.
[72] Pham, B.T., Luu, C., Phong, T.V., Trinh, P.T., Shirzadi, A., Renoud, S., Asadi, S., Le, H.V., von Meding, J., Clague, J.J., 2021b. Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling? J. Hydrol. 592, 125615. https://doi.org/10.1016/j.jhydrol.2020.125615
[73] Pham, T.D., Xia, J., Ha, N.T., Bui, D.T., Le, N.N., Tekeuchi, W., 2019. A review of remote sensing approaches for monitoring blue carbon ecosystems: Mangroves, seagrassesand salt marshes during 2010–2018. Sensors 19, 1933.
[74] Pourghasemi, H.R., Kariminejad, N., Amiri, M., Edalat, M., Zarafshar, M., Blaschke, T., Cerda, A., 2020. Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. Sci. Rep. 10, 1–11.
[75] Predick, K.I., Turner, M.G., 2008. Landscape configuration and flood frequency influence invasive shrubs in floodplain forests of the Wisconsin River (USA). J. Ecol. 96, 91–102.
[76] Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106.
[77] Rahman, M., Ningsheng, C., Islam, M.M., Dewan, A., Iqbal, J., Washakh, R.M.A., Shufeng, T., 2019. Flood susceptibility assessment in Bangladesh using machine learning and multi-criteria decision analysis. Earth Syst. Environ. 3, 585–601.
[78] Rahmati, O., Pourghasemi, H.R., Zeinivand, H., 2016. Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. Geocarto Int. 31, 42–70.
[79] RETA, 2011. Investment, Managing water in Asia's river basins: Charting progress and facilitating - The Vu Gia-Thu Bon Basin.

[80]Saber, M., Abdrabo, K.I., Habiba, O.M., Kantosh, S.A., Sumi, T., 2020. Impacts of Triple Factors on Flash Flood Vulnerability in Egypt: Urban Growth, Extreme Climate, and Mismanagement. Geosciences 10, 24.
[81]Saber, M., Boulmaiz, T., Guermoui, M., Abdrado, K.I., Kantoush, S.A., Sumi, T., Boutaghane, H., Nohara, D., Mabrouk, E., 2021. Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. Geocarto Int. 1–27.
[82]Schoppa, L., Disse, M., Bachmair, S., 2020. Evaluating the performance of random forest for large-scale flood discharge simulation. J. Hydrol. 590, 125531.
[83]Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., Shirzadi, A., 2018. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. J. Environ. Manage. 217, 1–11. https://doi.org/10.1016/j.jenvman.2018.03.089
[84]Shahabi, H., Shirzadi, A., Ronoud, S., Asadi, S., Pham, B.T., Mansouripour, F., Geertsema, M., Clague, J.J., Bui, D.T., 2020. Flash flood susceptibility mapping using a novel deep learning model based on deep belief network, back propagation and genetic algorithm. Geosci. Front. S1674987120302401. https://doi.org/10.1016/j.gsf.2020.10.007
[85]Shirzadi, A., Asadi, S., Shahabi, H., Ronoud, S., Clague, J.J., Khosravi, K., Pham, B.T., Ahmad, B.B., Bui, D.T., 2020. A novel ensemble learning based on Bayesian Belief Network coupled with an extreme learning machine for flash flood susceptibility mapping. Eng. Appl. Artif. Intell. 96, 103971.
[86]Talukdar, S., Ghose, B., Salam, R., Mahato, S., Pham, Q.B., Linh, N.T.T., Costache, R., Avand, M., 2020. Flood susceptibility modeling in Teesta River basin, Bangladesh using novel ensembles of bagging algorithms. Stoch. Environ. Res. Risk Assess. 34, 2277–2300.
[87]Tam, T.H., Abd Rahman, M.Z., Harun, S., Hanapi, M.N., Kaoje, I.U., 2019. Application of Satellite rainfall products for flood inundation modelling in Kelantan River Basin, Malaysia. Hydrology 6, 95.
[88]Tehrany, M.S., Jones, S., Shabani, F., 2019. Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. CATENA 175, 174–192. https://doi.org/10.1016/j.catena.2018.12.011
[89]Tehrany, M.S., Kumar, L., 2018. The application of a Dempster–Shafer-based evidential belief function in flood susceptibility mapping and comparison with frequency ratio and logistic regression methods. Environ. Earth Sci. 77, 1–24.
[90]Tehrany, M.S., Pradhan, B., Jebur, M.N., 2015. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stoch. Environ. Res. Risk Assess. 29, 1149–1165.
[91]Tehrany, M.S., Pradhan, B., Jebur, M.N., 2014. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J. Hydrol. 512, 332–343.
[92]Tehrany, M.S., Pradhan, B., Jebur, M.N., 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J. Hydrol. 504, 69–79.
[93]Thao, N.T.P., Linh, T.T., Ha, N.T.T., Vinh, P.Q., Linh, N.T., 2020. Mapping flood inundation areas over the lower part of the Con River basin using Sentinel 1A imagery. Vietnam J. Earth Sci. 42, 288–297.

[94]Tien Bui, D., Hoang, N.-D., 2017. A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1. 1) for spatial prediction of floods. Geosci. Model Dev. 10, 3391–3409.
[95]Tien Bui, D., Hoang, N.-D., Martínez-Álvarez, F., Ngo, P.-T.T., Hoa, P.V., Pham, T.D., Samui, P., Costache, R., 2020. A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area. Sci. Total Environ. 701, 134413. https://doi.org/10.1016/j.scitotenv.2019.134413
[96]Tien Bui, D., Pradhan, B., Nampak, H., Bui, Q.-T., Tran, Q.-A., Nguyen, Q.-P., 2016. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. J. Hydrol. 540, 317–330. https://doi.org/10.1016/j.jhydrol.2016.06.027
[97]Tiwari, M.K., Chatterjee, C., 2010. Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs). J. Hydrol. 382, 20–33.
[98]Torabi Haghighi, A., Menberu, M.W., Darabi, H., Akanegbu, J., Kløve, B., 2018. Use of remote sensing to analyse peatland changes after drainage for peat extraction. Land Degrad. Dev. 29, 3479–3488.
[99]Try, S., Tanaka, S., Tanaka, K., Sayama, T., Oeurng, C., Uk, S., Takara, K., Hu, M., Han, D., 2020. Comparison of gridded precipitation datasets for rainfall-runoff and inundation modeling in the Mekong River Basin. PLoS One 15, e0226814.
[100] Tuyen, T.T., Jaafari, A., Yen, H.P.H., Nguyen-Thoi, T., Phong, T.V., Nguyen, H.D., Van Le, H., Phuong, T.T.M., Nguyen, S.H., Prakash, I., Pham, B.T., 2021. Mapping forest fire susceptibility using spatially explicit ensemble models based on the locally weighted learning algorithm. Ecol. Inform. 63, 101292. https://doi.org/10.1016/j.ecoinf.2021.101292
[101] Vinet, F., 2008. Geographical analysis of damage due to flash floods in southern France: The cases of 12–13 November 1999 and 8–9 September 2002. Appl. Geogr. 28, 323–336. https://doi.org/10.1016/j.apgeog.2008.02.007
[102] Vu, T.T.L., Nguyen, L.D., Hoang, T.S., Bui, T.A., Nguyen, M.T., Nguyen, T.H., 2011. Solutions for flood and drought prevention and mitigation in Quang Nam.
[103] Wang, Y., Hong, H., Chen, W., Li, S., Panahi, M., Khosravi, K., Shirzadi, A., Shahabi, H., Panahi, S., Costache, R., 2019. Flood susceptibility mapping in Dingnan County (China) using adaptive neuro-fuzzy inference system with biogeography based optimization and imperialistic competitive algorithm. J. Environ. Manage. 247, 712–729.
[104] Xu, Y., Dai, Y., Dong, Z.Y., Zhang, R., Meng, K., 2013. Extreme learning machine-based predictor for real-time frequency stability assessment of electric power systems. Neural Comput. Appl. 22, 501–508.
[105] Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S., Bates, P.D., 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44, 5844–5853.
[106] Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., Zhao, B., 2020. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. J. Hydrol. 590, 125206.
[107] Yariyan, P., Janizadeh, S., Van Phong, T., Nguyen, H.D., Costache, R., Van Le, H., Pham, B.T., Pradhan, B., Tiefenbacher, J.P., 2020. Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. Water Resour. Manag. 34, 3037–3053.

[108] Young, R.A., Mutchler, C.K., 1969. Soil movement on irregular slopes. Water Resour. Res. 5, 1084–1089.
[109] Youssef, A.M., Hegab, M.A., 2019. Flood-hazard assessment modeling using multicriteria analysis and GIS: a case study—Ras Gharib area, Egypt, in: Spatial Modeling in GIS and R for Earth and Environmental Sciences. Elsevier, pp. 229–257.
[110] Youssef, A.M., Pradhan, B., Sefry, S.A., 2016. Flash flood susceptibility assessment in Jeddah city (Kingdom of Saudi Arabia) using bivariate and multivariate statistical models. Environ. Earth Sci. 75, 12.
[111] Zahedi, P., Parvandeh, S., Asgharpour, A., McLaury, B.S., Shirazi, S.A., McKinney, B.A., 2018. Random forest regression prediction of solid particle Erosion in elbows. Powder Technol. 338, 983–992.
[112] Zenggang, X., Zhiwen, T., Xiaowen, C., Xue-min, Z., Kaibin, Z., Conghuan, Y., 2021. Research on image retrieval algorithm based on combination of color and shape features. J. Signal Process. Syst. 93, 139–146.

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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

Highlights (3 to 5 bullet points (maximum 85 characters including spaces per bullet point)

Highlights

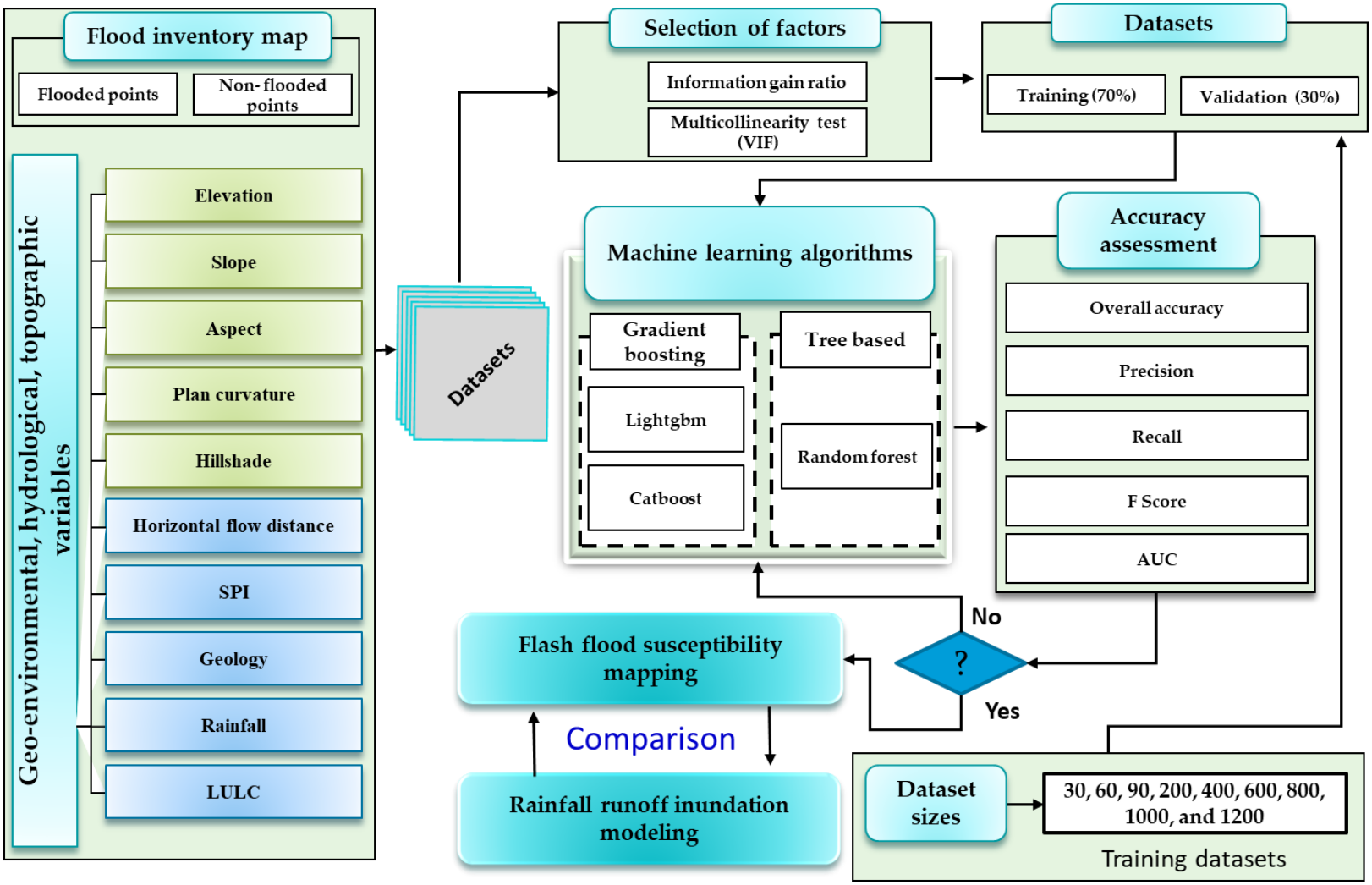
- Random forest, LightGBM, and CatBoost machine learning models were applied to predict flood susceptibility in a humid area that experienced successive extreme typhoons
- LightGBM and CatBoost were tested for the first time in this specific climatic region, with acceptable agreement with the RRI model
- Different training datasets were examined to find the lowest acceptable number of observations for flood susceptibility in machine learning
- The FSMs showed that downstream areas with high residential and agricultural activities were highly susceptible to flooding

Abstract

Vietnam has experienced many natural disasters, particularly typhoons. This study aims to examine three machine learning (ML) approaches—random forest (RF), LightGBM, and CatBoost—for flooding susceptibility maps (FSMs) in the Vu Gia-Thu Bon (VGTB) River Basin of Vietnam. The results of ML are compared with those of the rainfall–runoff model, and different training dataset sizes are utilized in the performance assessment. Ten independent factors that influence the FSMs in the study area, namely, aspect, rainfall, curvature, DEM, horizontal distance from the river, geology, hillshade, land use, slope, and stream power index, are assessed. An inventory map that includes approximately 850 flooding sites is considered based on several post-flood surveys after the typhoons in 1999, 2006, 2007, 2009, 2013, and 2020. The inventory dataset is randomly divided into two sets: training (70%), and testing (30%). The AUC-ROC results are 97.9%, 99.5%, 99.5% for CatBoost, LightGBM, and RF, respectively. The FSMs developed by the ML methods show good agreement with flood inundation mapping developed using the rainfall-runoff model. The FSMs show that downstream areas (both urbanized and agricultural) are under "high" and "very high" levels of susceptibility. Additionally, different sizes of the input datasets (i.e., 30, 60, 90, 200, 400, 600, 800, 1000, and 1250 data points) are tested to determine the least number of data points having an acceptable reliability. The results show that the ML methods can reasonably predict FSMs, regardless of the number of training samples, although the final FSMs show some spatial differences when changes in susceptibility level are seen. The developed FSMs for such typhoon-prone regions can be used by decision-makers and planners in Vietnam to propose effective mitigation measures for community resilience and development.

Keywords

Machine learning, random forest, LightGBM, CatBoost, flooding susceptibility mapping, rainfall-runoff inundation model.



Graphical Abstract

4/26/2022

Prof. Andras Bardossy
Editor
Journal of Hydrology

Dear Prof. Andras Bardossy

I wish to submit an original article for publication in Journal of hydrology, titled "Machine Learning Techniques as an Alternative Approach to Rainfall-Runoff Inundation Models for Flood Susceptibility Prediction".

This study elaborates on the machine learning (ML) techniques that can be used to predict flood susceptibility as an alternative for the rainfall-runoff inundation model (RRI). Two machine learning methods (CatBoost and LightGBM) along with conventional Random forest (RF) were examined for the first time in a typhoon-prone area in humid environment in Vietnam. We believe that our study makes a significant contribution to the literature because it is the first study that addresses this application with varying sizes of data sets and analyses the affecting on the model accuracy of the used ML techniques, as well as comparing with RRI model.

Further, we believe that this paper will be of interest to the readership of your journal because the employed two ML methods (CatBoost and LightGBM) superior the conventional Method of RF in terms of accuracy and processing speed. Additionally, the flood susceptibility maps developed by the ML techniques agree with the flood inundation map developed by RRI Model.

This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. We have read and understood your journal's policies, and we believe that neither the manuscript nor the study violates any of these. There are no conflicts of interest to declare.

Thank you for your consideration. I look forward to hearing from you.

Sincerely,
Mohamed Saber, on behalf of the coauthors
Disaster Prevention Research Institute
Kyoto University
Goka-sho, Uji City, Kyoto 611-0011, Japan
Tel.: +81-70-3600-6556
Email: mohamedmd.saber.3u@kyoto-u.ac.jp