Geoscience Frontiers 12 (2021) 101105

Contents lists available at ScienceDirect

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

Improved flood susceptibility mapping using a best first decision tree integrated with ensemble learning techniques

Binh Thai Pham [a], Abolfazl Jaafari [b], Tran Van Phong [c], Hoang Phan Hai Yen [d], Tran Thi Tuyen [e,*], Vu Van Luong [e], Huu Duy Nguyen [f], Hiep Van Le [g], Loke Kok Foong [h,i]

[a] University of Transport Technology, Ha Noi 100000, Viet Nam
[b] Research Institute of Forests and Rangelands, Agricultural Research, Education and Extension Organization (AREEO), Tehran, Iran
[c] Institute of Geological Sciences, Vietnam Academy of Sciences and Technology, 84 Chua Lang Street, Dong da, Ha Noi, Viet Nam
[d] Department of Geography, School of Social Education, Vinh University, Viet Nam
[e] Department of Resource and Environment Management, School of Agriculture and Resources, Vinh University, Viet Nam
[f] Faculty of Geography, VNU University of Science, 334 Nguyen Trai, Ha Noi, Viet Nam
[g] Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam
[h] Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[i] Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Viet Nam

ARTICLE INFO

Article history:
Received 14 April 2020
Received in revised form 13 August 2020
Accepted 4 November 2020
Available online 13 December 2020

Keywords:
Machine learning
Ensemble learners
Hybrid modeling

ABSTRACT

Improving the accuracy of flood prediction and mapping is crucial for reducing damage resulting from flood events. In this study, we proposed and validated three ensemble models based on the Best First Decision Tree (BFT) and the Bagging (Bagging-BFT), Decorate (Bagging-BFT), and Random Subspace (RSS-BFT) ensemble learning techniques for an improved prediction of flood susceptibility in a spatially-explicit manner. A total number of 126 historical flood events from the Nghe An Province (Vietnam) were connected to a set of 10 flood influencing factors (slope, elevation, aspect, curvature, river density, distance from rivers, flow direction, geology, soil, and land use) for generating the training and validation datasets. The models were validated via several performance metrics that demonstrated the capability of all three ensemble models in elucidating the underlying pattern of flood occurrences within the research area and predicting the probability of future flood events. Based on the Area Under the receiver operating characteristic Curve (AUC), the ensemble Decorate-BFT model that achieved an AUC value of 0.989 was identified as the superior model over the RSS-BFT (AUC = 0.982) and Bagging-BFT (AUC = 0.967) models. A comparison between the performance of the models and the models previously reported in the literature confirmed that our ensemble models provided a reliable estimate of flood susceptibilities and their resulting susceptibility maps are trustful for flood early warning systems as well as development of mitigation plans.

1. Introduction

As one of the most perilous natural hazards, floods bring serious damages to human property and life throughout the world. Floods account for tens of thousands of deaths and billions of dollars of economic losses annually (Aerts et al., 2018; Ahmadalipour and Moradkhani, 2019). With respect to the location, source and driving factors, floods are typically categorized into five different types, namely urban drainage, riverine flooding, ground failures, fluctuating lake levels, and coastal flooding and erosion (Wright, 2008). As a usual types of riverine flooding, flash floods appear once a considerable amount of water is discharged in a matter of a several minutes or hours (normally between two and five hours) of heavy rainfall. Flash floods also occur due to the sudden collapse of glaciers or a dam failure (Wang et al., 2019a). Flash floods are often characterized by a very high velocity that cause heavy damages to human lives and properties (Ahmadalipour and Moradkhani, 2019). The destructive impacts of flash floods are further exacerbated by climate change (Bubeck and Thieken, 2018), highlighting the importance of flood hazard and risk assessment and also identifying the most susceptible areas (Tehrany et al., 2013). However, the accurate prediction of future floods and identification of prone areas are challenging tasks that require accurate spatial and temporal data (Ouma and Tateishi, 2014) as well as robust predictive models (Bui et al., 2019; Rahmati et al., 2019).

In general, current modeling approaches to develop flood predictive models are categorized into three main categories, namely traditional statistical analysis, rainfall-runoff models, and pattern recognition

* Corresponding author.
E-mail address: tuyentt@vinhuni.edu.vn (T.T. Tuyen).

https://doi.org/10.1016/j.gsf.2020.11.003
1674-9871/

approaches. Utilizing data pertaining to long-term recorded time-series at gauged stations, statistical analysis is performed to develop the regression models that enable spatial discharge predictions over time (McCuen, 2016). Rainfall-runoff techniques which are typically classified into non-physically and physically-based models (Lee et al., 2019) have to do with runoff estimation from rainfall. Both statistical analysis and rainfall-runoff approaches need long-term data that often restrict their application in practice (Zhou et al., 2019). Within the concept of pattern recognition, machine learning methods have recently received attention because of their ability to handle complicated relationships between input variables and improve the quality of results (Bishop, 2006). Machine learning methods have been typically used under two general modeling types, namely simple single modeling and hybrid ensemble modeling (Choubin et al., 2019). The single applications of machine learning methods have recently upgraded to the hybrid ensemble modeling that has proven to provide more robust predictive models and more accurate estimations of future events (Jaafari et al., 2019; Pham et al., 2019a; Zidane et al., 2019; Nhu et al., 2020a). In the domain of flood modeling, the hybrid ensemble models include various models, such as weights-of-evidence and support vector machine (SVM) (Tehrany et al., 2014), frequency ratio and logistic regression (Tehrany et al., 2013), extreme learning machine and a particle swarm optimization (Bui et al., 2019), artificial neural network (ANN) and firefly algorithm (Ngo et al., 2018). Other models include different combinations of neuro-fuzzy with metaheuristic algorithms (Bui et al., 2018), weights-of-evidence and decision trees classifier (Costache, 2019), and ensemble learning techniques and different base methods (Costache and Tien Bui, 2019). Within these previous studies, the researchers have acknowledged the efficiency of the hybrid ensemble models to cope with the deficiencies of single models toward more reliable flood prediction. Therefore, we have motivated to develop ensemble flood predictive models that integrate Best First Decision Tree (BFT) classifier with Bagging, Decorate, and Random Subspace ensemble learning techniques for obtaining more accurate estimates of flash flood susceptibilities than a single model.

To the best of our knowledge, this is the first application of such ensemble models for flash flood modeling that seeks to (i) explore the capability of the ensemble modeling approach to elucidate the underlying patterns of flash floods, (ii) compare the utility of three ensemble learning techniques to improve the BFT classifier, and (iii) make a more accurate prediction of flood susceptibilities compared to the single BFT classifier. The development of these ensemble models is illustrated by actual data from a flood-prone area in the Nghe An Province, Vietnam. The performances of the models are presented and compared. Spatially explicit distribution maps of flash flood susceptibility are produced according to the results of each model and evaluated by several performance metrics.

2. Research area and data used

2.1. Research area

Our research area encompasses an approximately 2827 km^2 part of the Nghe An Province, located in north-central Vietnam (Fig. 1). The altitude of this area varies from 77 to 1551 m and about 72% of the research area has slopes varying from 10° to 30°. The characteristic of the climate is subtropical monsoonal enjoying two unique seasons: The wet season begins slowly in May, however, the accumulation of rain does not exceed 150 mm/month before August. The middle of the wet season is in September–October (> 400 mm/month), representing 60% to 70% of the annual precipitation. The dry season in January–April records accumulations of rain of about 30–40 mm/month. The research area is known as a storm center in north-central of Vietnam, where receives heavy and extreme rainfalls during the year. Historical records indicate that many regions of the Nghe An Province have experienced severe flash floods during the years 2007, 2010, 2011, 2017, and 2018 that caused significant loss of human life and property. Therefore, flash flood susceptibility mapping is important in this area for risk management and mitigating the consequences of floods.

2.2. Geospatial data

2.2.1. Historical floods

Knowledge of historical floods is essential to predict future floods (Bui et al., 2019; Choubin et al., 2019; Darabi et al., 2019). In this study, the spatial locations of 126 flash floods that have occurred in the research area were obtained from the historical archives of the Department of Natural Resources and Environment of the Nghe An

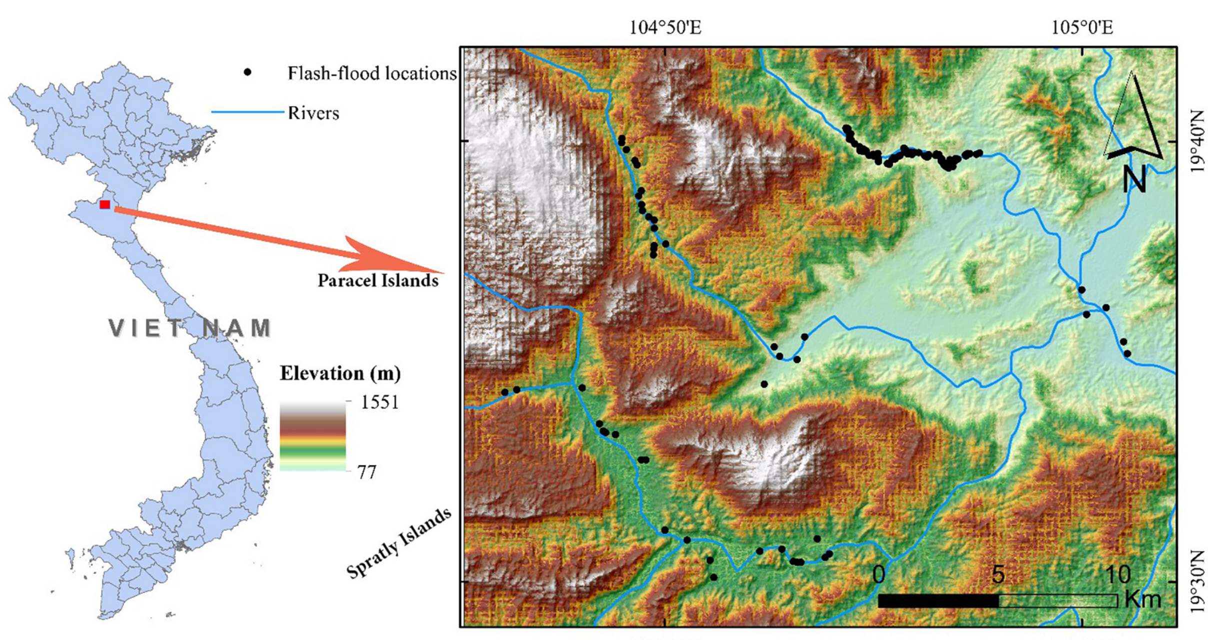

Fig. 1. Geographic locations of the research area and historical flash floods.

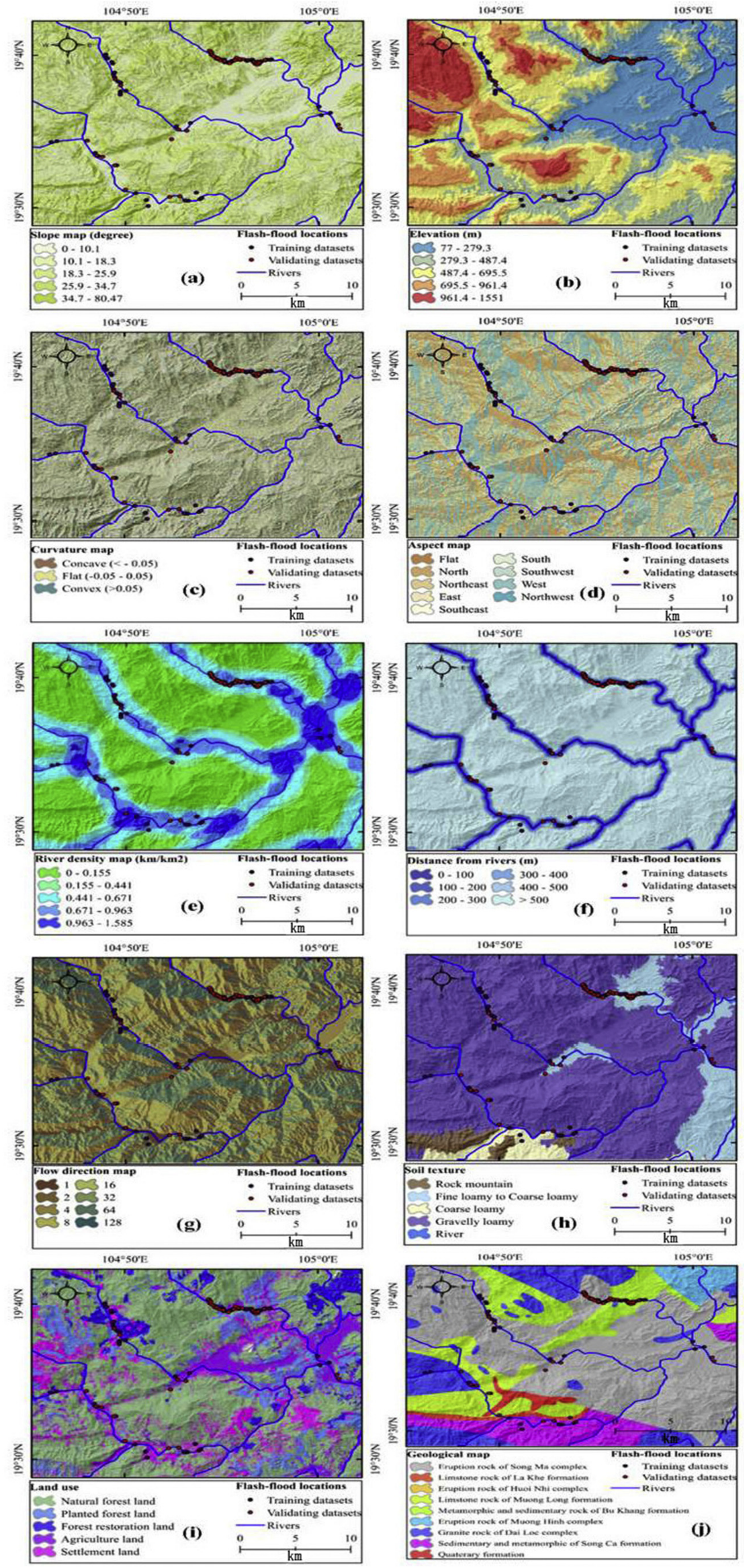

Fig. 2. Influencing factors used for flash flood susceptibility mapping: (a) slope, (b) elevation, (c) curvature, (d) aspect, (e) river density, (f) distance from rivers, (g) flow direction, (h) soil, (i) land use and (j) geology.

Province, and field investigations to generate an inventory map of historical floods. These flood locations were allocated to two different sets. One set includes 70% of data (88 floods), which was used to generate the training dataset, and the second set that consists of 30% remaining data (38 floods) and was used to generate the validation dataset. Along with the flood locations, 126 unflooded locations were sampled from the research area's unflooded portions and used for generating the final datasets. This led to the training and validation datasets that include 176 and 76 samples, respectively.

2.2.2. Influencing factors

Following a thorough review of the related literature (Marco and Cayuela, 1994; Ouma and Tateishi, 2014; Destro et al., 2018; Costache and Tien Bui, 2019; Darabi et al., 2019; Wang et al., 2019a, 2019b), investigating the characteristics of the floods that have occurred in the research area, and conducting several field surveys, we identified slope, elevation, aspect, curvature, river density, distance from rivers, flow direction, geology, soil, and land use as the flood influencing factors (Fig. 2). The topographic factors (i.e., elevation, curvature, slope, and aspect) were derived from a 10 m resolution digital elevation model (DEM) that was obtained from the Ministry of Natural Resources and Environment of Vietnam. The effect of topography factors on flood occurrences has been widely acknowledged in the literature. Since floods are more probable to occur in low-elevated, flat, and convergence areas (Sadler et al., 2018; Bui et al., 2019), we included the slope, elevation, and curvature into the modeling process to analyze the effects of these factors on future flood probability. We elected to use the aspect factor for flood modeling because this factor is associated with the convergence and directions of water flowing (Bui et al., 2019). River density that indicates the extent to which a watershed is drained by stream channels is another factor of the probability of flooding (Chapi et al., 2017). The landscapes with greater river density are more susceptible to flood occurrence. Due to the dependency of flood events to the terrestrial water storages, proximity to rivers affects flooding likelihood and magnitude considerably. Flow direction indicates how the overland flow is distributed over a watershed (Zhou et al., 2011) and is a key parameter when performing hydrological modeling for such flood prediction. Soil and land-use types were also selected as influencing factors because of their influence on the infiltration and runoff speed. The maps of soil (1:100,000 scale) and land-use (1:100,000 scale) types were provided by the Department of Natural Resources and Environment of the Nghe An Province. The last factor used in this study was geology, which represents underlying rock types and affects the infiltration and runoff in a watershed. The lithology map of the research area was obtained from the national geological and mineral resources maps (1,100,000 scale).

3. Modeling methodology

3.1. Methods used

3.1.1. Frequency ratio

To explore the spatial relationship between the historical record of flood events and each of the 10 influencing factors, we employed the frequency ratio (*FR*) method. *FR* is a straightforward and widely used statistical method for investigating the spatial relationships between a phenomenon and its influencing factors. The higher the *FR* value, the stronger the relationship between a phenomenon and a specific factor. Mathematically, the *FR* method is expressed as (Jaafari et al., 2019):

$$FR = \frac{N_{ij}}{A_{ij}} \Big/ \frac{N_r}{A_r} \quad (1)$$

where N_{ij} represents the flood pixels number in the class i of the influencing factor j, A_{ij} represents the pixels number in the class i of the influencing factor j. N_r and A_r are total pixels of floods and the total pixels of the research area.

3.1.2. Bagging

Bootstrap aggregating, also termed Bagging, is one of the first-introduced techniques for generating multiple sub-datasets and combining different base classifiers (Breiman, 1996). Bagging has the capability to decrease the classification variance toward improving the performance of the ensemble model (Breiman, 1996). Using an initial training dataset, Bagging generates n bootstrap subsets of which sizes equal to the original training dataset. These bootstrap subsets, also known as bootstrapped sub-datasets, are used to train a base classifier. The outputs are then combined using the majority voting procedure that can be expressed by:

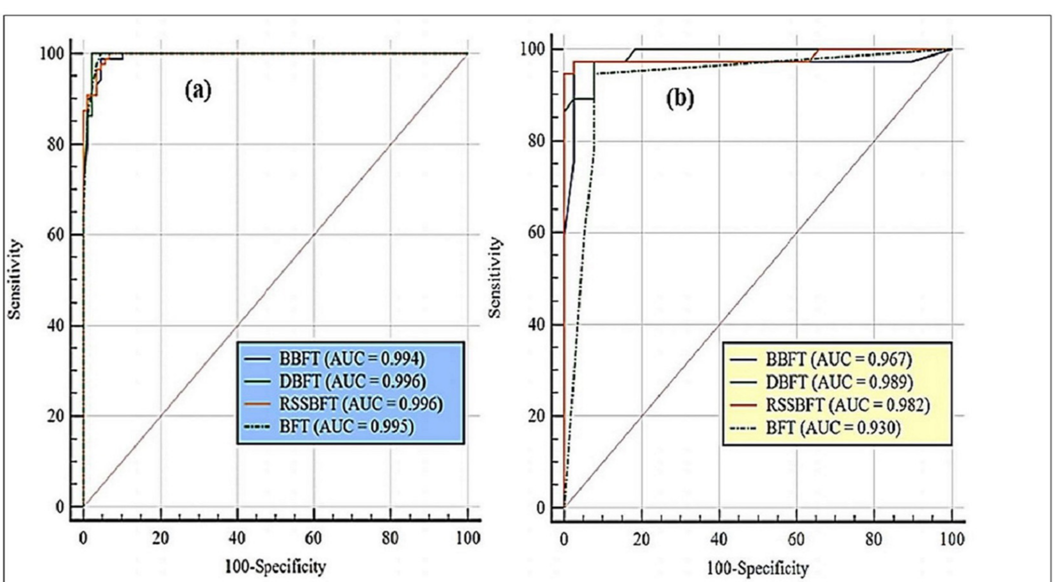

Fig. 3. AUC values in (a) training and (b) validation phases. AUC: area under curve. The diagonal line represents points where sensitivity = 1-specificity (i.e., AUC = 0.5). The closer the curve comes to the diagonal line, the worse the model performance.

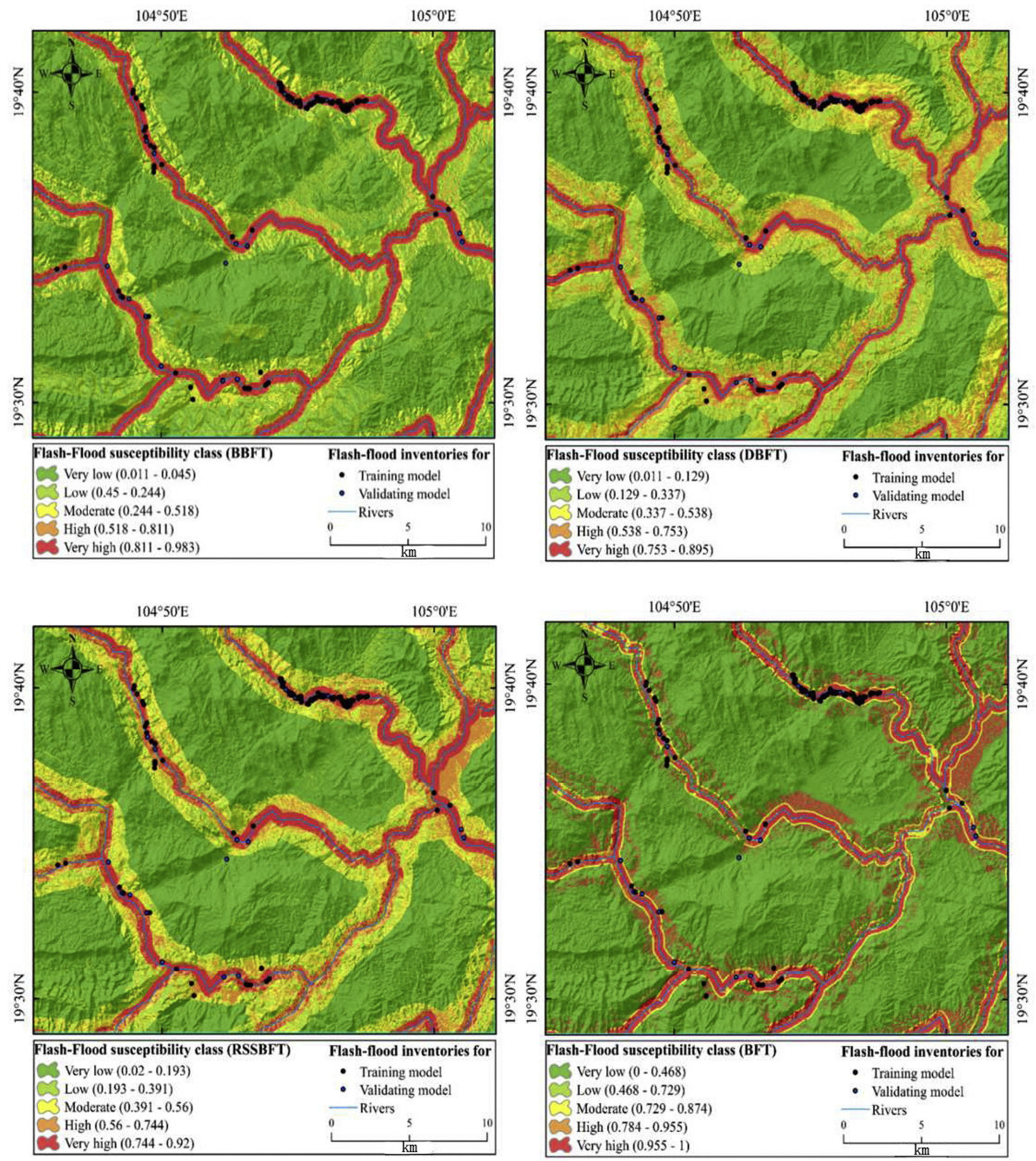

Fig. 4. Flash flood susceptibility maps.

$$\beta(x) = \underset{y \in \{-1,1\}}{\operatorname{argmax}} \sum_b \delta_{\operatorname{sgn}(C^b(x)),y} \tag{2}$$

where $\delta_{i,j}$ is the Kronecker symbol, $C^b(x)$ donates the constructed classifier, and $y \in \{-1,1\}$ is the class labels (i.e., flood and unflood).

3.1.3. Decorate

Decorate is an ensemble learner developed originally by Melville and Mooney (2005). As its name (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) suggests, this method produces various classifiers utilizing artificial training examples. The main

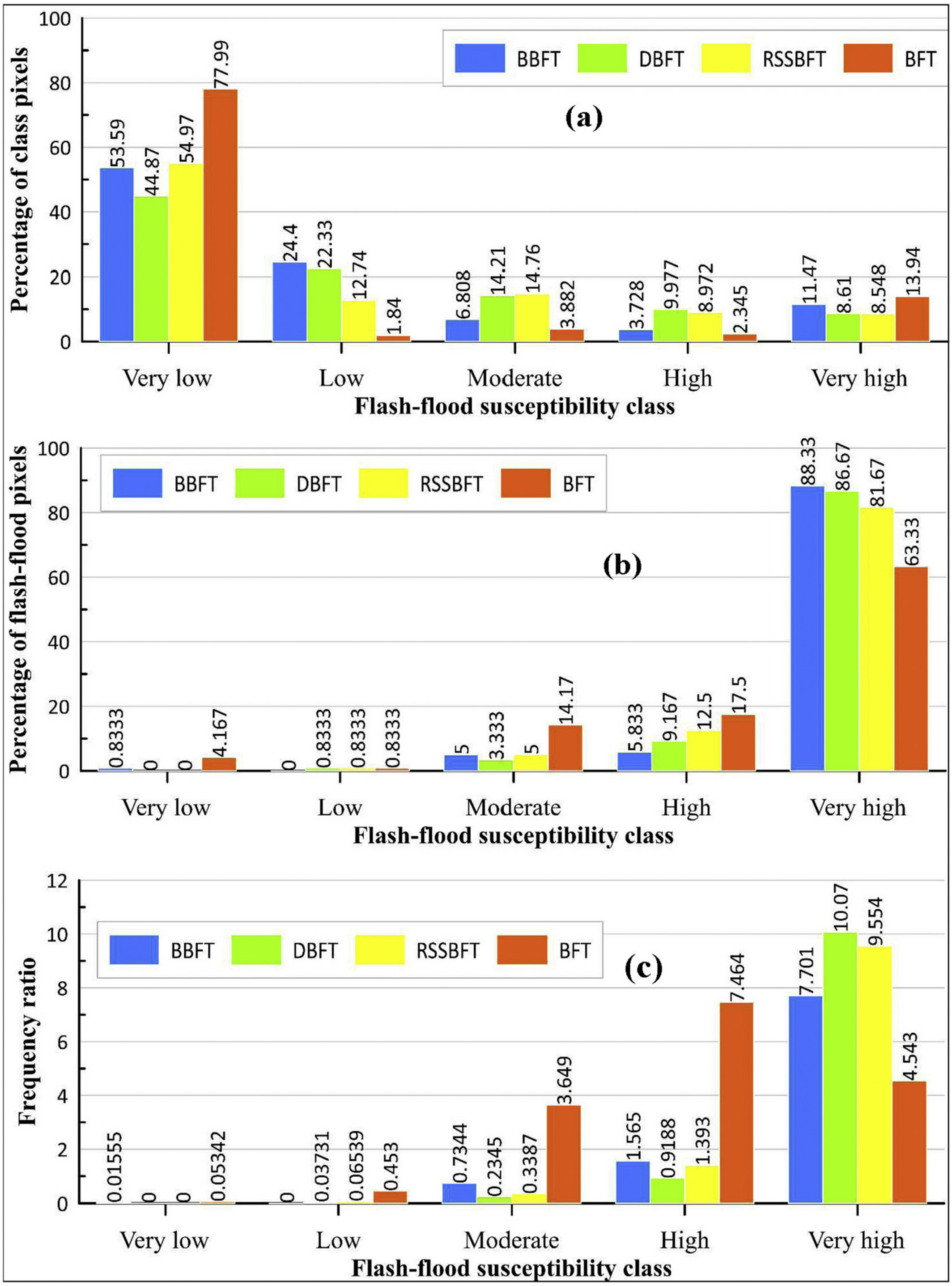

Fig. 5. Quantitative analysis of the flood susceptibility maps.

idea about Decorate is the expansion of the initial training set through usage of artificial training samples, which significantly differs from Bagging that just utilize a given training set to produce several classifiers. Using expanded training sets to train the base classifiers, Decorate offers an innovative approach for developing effective ensemble models with broad applications in the literature (Sun et al., 2015).

3.1.4. Random subspace (RSS)

RSS is an ensemble learning technique that uses a training process based on random sampling from a set of factors to reduce their correlations in an ensemble model (Barandiaran, 1998). Assuming each training sample M_i $(i = 1, \ldots, n)$ in a flood training set $M = (M_1, M_2, \ldots, M_n)$, a p-dimensional vector $K_i = (K_{i1}, K_{i2}, \ldots, K_{in})$ of p influencing factors, a random selection $r < p$ factors from the p-dimensional dataset M results in an r-dimensional random subspace of the initial p-dimensional feature space (Skurichina and Duin, 2002). Then, the base classifiers are created according to the r-dimensional random subspace and combined in a decision regulation based on a simple majority vote (Eq. (2)). Similar to Bagging, RSS uses bootstrapping and aggregation procedure. However, RSS bootstraps the feature space that differs from Bagging bootstrapping the training samples in the Bagging technique (Tao et al., 2006).

3.1.5. Best first decision tree (BFT)

BFT, developed by Shi (2007), is a supervised learning classifier for the classification problems. In contrast to the standard decision tree classifiers that expand nodes in a fixed depth-first order, in the BFT classifier, the best node is extended first. The best node refers to the node whose split greatly decreases impurity between the nodes existing for splitting (Shi, 2007). This advantage allows for exploring new pruning methods that utilize cross-validation to select the number of expansions. In general, BFT works in three main steps: (1) selecting an attribute to put at the root node and create a few branches for this attribute primarily based upon a number of parameters, (2) splitting training samples into several subsets, one for every branch extending from the root node, and (3) repeating this process and selecting the best subset from all available subsets for expansions till all nodes are pure or a predefined number of expansions is met.

3.1.6. Development of the ensemble models

To develop the ensemble flash-flood predictive models, we used the BFT classifier as the base method and used each one of the ensemble learning techniques to generate different training datasets for training the base BFT classifier. We then evaluated the results using the performance metrics. Each ensemble learning technique has a number of parameters, including the iteration number, percent of bag size, and threshold of weights, which ought to be properly adjusted for the best possible performance. In the present study, we manually adjusted the parameters until the most robust and accurate predictive models were obtained (Pham et al., 2019a, 2019b, 2020; Nguyen et al., 2020a, 2020b; Nhu et al., 2020b; Tran et al., 2020).

3.1.7. Performance metrics

In current study, we validated the performance of the developed flood models using the positive predictive value, negative predictive value, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), accuracy, root-mean-square error (RMSE), and Kappa metrics. Further descriptions can be found in the literature (Jaafari, 2018; Hong et al., 2019; Tran et al., 2020).

4. Results

4.1. Spatial relationship

The results of the FR method that measured and quantified the spatial relationship between influencing factors and historical floods indicated that the most flood-prone parts of our research area fall on river soil type (*FR* = 15.83), distance from rivers of 0–300 m (*FR* = 11.23, 6.38, and 5.52), river density of 0.671–0.963 km/km^2 (*FR* = 4.41), and geology formation of metamorphic and sedimentary rock (*FR* = 3.96). Conversely, the factor classes with *FR* = 0 were identified as the flood-proof portions of the research area (Table 1).

4.2. Model performance

The BFT model and its derived ensemble models were developed and validated using training and validation datasets. Based on the different performance metrics, we found all models successful in identifying the spatial pattern of flood susceptibilities (i.e., training performance) within the research area. Chief among them was the Decorate-BFT model that achieved the highest values of the PPV (97.73%), NPV (100%), SST (100%), SPF (97.78%), ACC (98.86%), Kappa (0.977) metrics and the lowest magnitude of error (RMSE = 0.174) (Table 2). According to these results, the DBFT model correctly classified ~98% of all pixels in flooded class, classified 100% of all pixels in unflooded class, classified 100% of the flooded pixels into the flooded class, classified ~98% of unflooded pixels into the unflooded class, classified ~99% of all training dataset pixels, which has an excellent agreement between the predicted and detected floods (Kappa = 0.977).

Table 1
Spatial relationship between influencing factors and historical floods measured using the FR method.

Factor	Class	No. of class pixels	No. of flood pixels	Class pixels (%)	Flood pixels (%)	FR
Soil type	Rock mountain	134,378	0	2.86	0.00	0.00
	Fine loamy to Coarse loamy	559,852	48	11.91	40.00	3.36
	Coarse loamy	168,722	3	3.59	2.50	0.70
	Gravelly loamy	3,811,898	59	81.11	49.17	0.61
	River	24,734	10	0.53	8.33	15.83
Land use	Natural forest land	2,683,277	33	57.10	27.50	0.48
	Planted forest land	590,296	19	12.56	15.83	1.26
	Forest restoration land	244,230	1	5.20	0.83	0.16
	Agriculture land	412,420	24	8.78	20.00	2.28
	Settlement land	762,109	43	16.22	35.83	2.21
Slope (degree)	0–10.1	997,434	73	21.22	60.83	2.87
	10.1–18.3	1,294,705	32	27.55	26.67	0.97
	18.3–25.0	1,278,205	10	27.20	8.33	0.31
	25.9–34.7	849,799	5	18.08	4.17	0.23
	34.7–80.47	281,829	0	6.00	0.00	0.00
Curvature	Concave (< −0.05)	1,867,273	51	39.73	42.50	1.07
	Flat (−0.05 to 0.05)	988,654	26	21.04	21.67	1.03
	Convex (<0.05)	1,846,045	43	39.28	35.83	0.91
Elevation (m)	77–279.3	1,152,317	83	24.52	69.17	2.82
	279.3–487.4	1,141,579	34	24.29	28.33	1.17
	487.4–695.5	1,288,408	3	27.42	2.50	0.09
	695.5–961.4	742,113	0	15.79	0.00	0.00
	961.4–1551	375,586	0	7.99	0.00	0.00
River density (km/km^2)	0–0.155	2,018,386	0	42.95	0.00	0.00
	0.155–0.441	595,327	1	12.67	0.83	0.07
	0.441–0.671	1,038,188	28	22.09	23.33	1.06
	0.671–0.963	745,565	84	15.86	70.00	4.41
	0.963–1.585	302,118	7	6.43	5.83	0.91
Dis. from rivers (m)	0–100	181,317	52	3.86	43.33	11.23
	100–200	178,135	29	3.79	24.17	6.38
	200–300	177,309	25	3.77	20.83	5.52
	300–400	175,688	7	3.74	5.83	1.56
	400–500	172,177	1	3.66	0.83	0.23
	>500	3,814,958	6	81.18	5.00	0.06
Aspect	Flat	81,687	6	1.74	5.00	2.88
	North	545,109	14	11.60	11.67	1.01
	Northeast	636,404	14	13.54	11.67	0.86
	East	633,862	10	13.49	8.33	0.62
	Southeast	639,522	16	13.61	13.33	0.98
	South	579,459	16	12.33	13.33	1.08
	Southwest	606,583	26	12.91	21.67	1.68
	West	492,965	10	10.49	8.33	0.79
	Northwest	486,381	8	10.35	6.67	0.64
Flow direction	1	837,549	32	17.82	26.67	1.50
	2	497,948	11	10.60	9.17	0.87
	4	750,544	33	15.97	27.50	1.72
	8	447,207	6	9.52	5.00	0.53
	16	591,869	17	12.59	14.17	1.12
	32	380,841	5	8.10	4.17	0.51
	64	654,173	11	13.92	9.17	0.66
	128	543,810	5	11.57	4.17	0.36
Geological	Eruption rock of Song Ma complex	2,393,888	52	50.94	43.33	0.85
	Limestone rock of La Khe formation	55,584	2	1.18	1.67	1.41
	Eruption rock of Huoi Nhi complex	42	0	0.00	0.00	0.00
	Limestone rock of Muong Long formation	405,545	3	8.63	2.50	0.29
	Metamorphic and sedimentary rock of Bu Khang formation	454,893	46	9.68	38.33	3.96
	Eruption rock of Muong Hinh complex	160,972	0	3.43	0.00	0.00
	Granite rock of Dai Loc complex	714,840	8	15.21	6.67	0.44
	Sedimentary and metamorphic of Song Ca formation	450,233	9	9.58	7.50	0.78
	Quaternary formation	65,975	0	1.40	0.00	0.00

Table 2
Training performance of the models.

Metric	BBFT	DBFT	RSSBFT	BFT
PPV (%)	95.45	97.73	93.18	95.45
NPV (%)	95.45	100.00	98.86	100.00
SST (%)	95.45	100.00	98.80	100.00
SPF (%)	95.45	97.78	93.55	95.65
ACC (%)	95.45	98.86	96.02	97.73
Kappa	0.909	0.977	0.921	0.954
RMSE	0.195	0.174	0.213	0.137

In the case of predicting future floods (i.e., validation performance), however, the BBFT and RSSBFT models had the highest performance (Table 3). With the exception of RMSE (BBFT = 0.187; RSSBFT = 0.236), the other metrics were identical for these two models. The single BFT model with the highest magnitude of RMSE (0.259) and lowest values of PPV (92.11%), NPV (94.74%), SST (94.59%), SPF (92.31%), ACC (93.42%), and Kappa (0.868) was identified as the least effective model to predict the future floods.

The overall model performance measured through the AUC metric (Fig. 3) demonstrated that the DBFT model ($AUC_{training} = 0.996$; $AUC_{validation} = 0.989$) was dominant over the other models. The second-best model was RSSBFT that achieved AUC values of 0.996 and 0.982 in the training and validation phases, respectively.

4.3. Susceptibility maps

When the BFT model and its ensembles were approved with respect to the training performance and their capability to predict future floods, they were employed for estimating the flood susceptibility values for the entire research area. After successful estimating the susceptibility values, the values were categorized into five susceptibility categories, i.e., very low, low, moderate, high, and very high, using the geometrical intervals classification scheme (Fig. 4). Among these maps, the BFT map represents the greatest portion (78%) of the landscape as very low susceptibility to flooding (Fig. 5a). While a majority portion (~58%) of our research area falls into very low susceptibility to flooding, regions of high and very high susceptibility to flooding covered 6%–10.4% of the research area. Despite the different performance of the models in classifying the research area to various levels of flood susceptibility, all four models agreed that the low-lying regions alongside the rivers are the most flood-prone portions of our research area.

Among the susceptibility classes the percentage of flood pixels varied and ranged from 0% for the very low susceptibility classes of the DBFT and RSSBFT models to 88.33% for the very high susceptibility class of the BBFT model (Fig. 5b). Further, in each susceptibility map, the very high susceptibility class has the greatest number of FR of the flood pixels (Fig. 5c), followed by the high susceptibility, moderate susceptibility, low susceptibility, and very low susceptibility classes, respectively, demonstrating the capability of the models to adequately demarcating different levels of flood susceptibilities within our research area.

Table 3
Validation performance of the models.

Metric	BBFT	DBFT	RSSBFT	BFT
PPV (%)	97.37	92.11	97.37	92.11
NPV (%)	97.37	97.37	97.37	94.74
SST (%)	97.37	97.22	97.37	94.59
SPF (%)	97.37	92.50	97.37	92.31
ACC (%)	97.37	94.74	97.37	93.42
Kappa	0.947	0.895	0.947	0.868
RMSE	0.187	0.236	0.230	0.259

5. Discussion

Despite various models that have been developed for different flood-prone areas around the world, accurate prediction of floods remains challenging that calls for investigating new methods and approaches. Here, we addressed this issue and evaluated the BFT classifier and its ensemble models that were derived from combining this classifier with three ensemble learning techniques. Our results showed that the Bagging, Decorate, and RSS ensemble techniques effectively improved the performance of the base BFT classifier and provided a more accurate estimate of the future floods compared to the single BFT classifier. Our finding is in line with the previous works reported in the literature that demonstrated the capability of the ensemble modeling approach for handling manifold information inherited in earth science. For example, Pham et al. (2017b) reported an improved landslide prediction using ANN integrated with the MultiBoost, AdaBoost, Dagging, Bagging, RSS, Rotation Forest ensemble techniques. In another study, Pham et al. (2019a) enhanced the performance of the decision stump classifier using the Rotation Forest, MultiBoost, and Bagging techniques for zoning groundwater potential. In the context of flood prediction modeling, Chapi et al. (2017) and Chen et al. (2019) illustrated that the Bagging and RSS techniques in combination of decision tree classifiers can produce effective ensemble models for flood prediction. In a recent study, Tran et al. (2020) proved that Hyperpipes algorithm coupled with the different ensemble techniques achieved the highest performance for landslide susceptibility mapping.

In our study, the Decorate ensemble leaning method performed better than the other ensemble techniques in improving the base BFT classifier for the prediction of flash floods. Compared to the Multi-Boost, AdaBoost, Bagging, Dagging, Rotation Forest, and RSS techniques that have been extensively used in the literature. Decorate is an ensemble learning technique with limited application. However, it seems that the application of this ensemble technique can result in developing effective predictive models with high-quality results. In a comparative study, Sun et al. (2015) reported that Decorate produce very good performance ensemble models that can provide improvements over the Bagging and AdaBoost techniques for both small and large datasets. In contrast to the Bagging and AdaBoost techniques that manipulate the original training sets to boost the quality of the base classifier performance (Pham et al., 2019b), Decorate utilizes artificial training examples to expand the training set that often results in more accurate and robust ensemble models than those obtained from using the Bagging and AdaBoost techniques (Sun et al., 2015). We showed that the RSS technique represented an ensemble model (RSSBFT) that was ranked as the second-best model. Shirzadi et al. (2017) and Pham et al. (2017a, 2019b) also reported on the improved accuracy of landslide prediction using the ensemble models derived by the RSS technique.

Overall, despite small differences between the models proposed in the current study, we found all three Bagging, Decorate, and RSS ensemble learners effective for boosting the performance of the base BFT classifier. From these results we can conclude that an ensemble modeling approach can efficiently decrease, noise, variance, and over-fitting problems of input data to achieve an improved and reliable modeling outcome. Some would argue that hybrid models may take extra time in generating final outputs compared to the simple single models. However, the accuracy of natural hazard prediction should not be compromised with the cost reduction strategies that end up causing cost increments in the long run. Instead, modelers should aim to innovate with the purpose of finding ways to improve prediction accuracy while costs are reduced. One way to do this is to utilize cloud computing services (e.g., Amazon WEB Services, Microsoft Azure, and Google Cloud) to increase the efficiency and effectiveness of the modeling process for obtaining reliable estimation on reasonable computation time.

5.1. Management plans to control floods

Risk management and preventive actions are the main concepts of flood susceptibility mapping. Today, evidence suggests that unregulated land-use planning and policy implementation along with and climate change have extremely intensified flood occurrences at an alarming rate worldwide. Risk management and preventive measures no longer concentrate on preventing flood but focus on the responsibilities of local authorities to reduce flood impacts. For instance, residents of flood-prone regions should be informed of probabilities and risks (Aakre et al., 2010; Hegger et al., 2017; Mafi-Gholami et al., 2020). Land-use planners and local authorities are expected to inform the local residents of the latest assessments of flood susceptibilities (e.g., the results presented here) and the regulations that prohibit new developments in the area with high susceptibility to flooding occurrences (Dai et al., 2002).

The susceptibility maps are highly valuable tools for reducing the residual losses by evacuating areas that are expected to have recurring and devastating floods (i.e., high and very high susceptible zones delineated by the flood susceptibility maps). The financial sectors also can use the susceptibility maps for the enforcement of land use regulations and the appropriate compensation for flood damages.

6. Concluding remarks

We proposed three ensemble models, which simultaneously took advantage of both decision tree classifiers and ensemble learning techniques, for flash flood prediction. The ensemble BBFT, DBFT, and RSSBFT models provided better predictions than the single BFT model in terms of the future flood probability. The efficacy of the ensemble modeling approach was demonstrated when the high-level training performance of the single BFT model decreased considerably in the validation phase to level much lower than those of the three ensemble models. Although the popularity of Decorate is not as much as Bagging and RSS, we illustrated the utility of this ensemble learning technique for the prediction of flash flood susceptibility.

The ensemble Decorate-BFT model delimited approximately 33% of the study area into the high and very high susceptibilities to flood occurrences. These results suggest that design, implementation, and verification of flood early warning systems should be directed to these portion of the Nghe An Province.

Highly accurate flood prediction models such as the Decorate-BFT model that was proposed in this study are useful tools for obtaining more informed and improved estimates of flash floods. Future studies may extend this modeling approach by exploring the use of the DBFT model for other regions with different geo-environmental setting, other types of floods (such as urban, river, and coastal floods), other types of natural hazards (such as landslides, land subsidence, and wildfires), and comparing the performance of other types of the base classifier (such as ANN, and SVM).

Declaration of competing interest

None.

Acknowledgments

This study received funding from the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant No. 105.08-2019.03. We thank two anonymous reviewers for very thoughtful comments that helped improve this paper.

References

Aakre, S., Banaszak, I., Mechler, R., Rübbelke, D., Wreford, A., Kalirai, H., 2010. Financial adaptation to disaster risk in the European union. Mitig. Adapt. Strateg. Glob. Chang. 15 (7), 721–736.

Aerts, J.C., Botzen, W.J., Clarke, K.C., Cutter, S.L., Hall, J.W., Merz, B., Michel-Kerjan, E., Mysiak, J., Surminski, S., Kunreuther, H., 2018. Integrating human behaviour dynamics into flood disaster risk assessment. Nat. Clim. Chang. 8 (3), 193–199.

Ahmadalipour, A., Moradkhani, H., 2019. A data-driven analysis of flash flood hazard, fatalities, and damages over the conus during 1996–2017. J. Hydrol. 578, 124106.

Barandiaran, I., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 1–22.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag, New York, p. 738.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Bubeck, P., Thieken, A.H., 2018. What helps people recover from floods? Insights from a survey among flood-affected residents in Germany. Reg. Environ. Chang. 18 (1), 287–296.

Bui, D.T., Panahi, M., Shahabi, H., Singh, V.P., Shirzadi, A., Chapi, K., Khosravi, K., Chen, W., Panahi, S., Li, S., Ahmad, B.B., 2018. Novel hybrid evolutionary algorithms for spatial prediction of floods. Sci. Rep. 8 (1), 1–14.

Bui, D.T., Ngo, P.T.T., Pham, T.D., Jaafari, A., Minh, N.Q., Hoa, P.V., Samui, P., 2019. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping. Catena 179, 184–196.

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.

Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X., Ahmad, B.B., 2019. Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. J. Hydrol. 575, 864–873.

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 (2), 2087–2096.

Costache, R., 2019. Flash-flood potential index mapping using weights of evidence, decision trees models and their novel hybrid integration. Stoch. Env. Res. Risk A. 33 (7), 1375–1402.

Costache, R., Tien Bui, D., 2019. Spatial prediction of flood potential using new ensembles of bivariate statistics and artificial intelligence: a case study at the Putna river catchment of Romania. Sci. Total Environ. 691, 1098–1118.

Dai, F.C., Lee, C.F., Ngai, Y.Y., 2002. Landslide risk assessment and management: An overview. Eng. Geol. 64 (1), 65–87.

Darabi, H., Choubin, B., Rahmati, O., Torabi Haghighi, A., 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.

Destro, E., Amponsah, W., Nikolopoulos, E.I., Marchi, L., Marra, F., Zoccatelli, D., Borga, M., 2018. Coupled prediction of flash flood response and debris flow occurrence: Application on an alpine extreme flood event. J. Hydrol. 558, 225–237.

Hegger, D.L., Mees, H.L., Driessen, P.P., Runhaar, H.A., 2017. The roles of residents in climate adaptation: a systematic review in the case of the Netherlands. Environ. Pol. Govern. 27 (4), 336–350.

Hong, H., Jaafari, A., Zenner, E.K., 2019. Predicting spatial patterns of wildfire susceptibility in the Huichang county, China: an integrated model to analysis of landscape indicators. Ecol. Indic. 101 (1), 878–891.

Jaafari, A., 2018. Lidar-supported prediction of slope failures using an integrated ensemble weights-of-evidence and analytical hierarchy process. Environ. Earth Sci. 77, 42. https://doi.org/10.1007/s12665-017-7207-3.

Jaafari, A., Razavi Termeh, S.V., Bui, D.T., 2019. Genetic and firefly metaheuristic algorithms for an optimized neuro-fuzzy prediction modeling of wildfire probability. J. Environ. Manag. 243, 358–369.

Lee, K.T., Ho, J.-Y., Kao, H.-M., Lin, G.-F., Yang, T.-H., 2019. Using ensemble precipitation forecasts and a rainfall-runoff model for hourly reservoir inflow forecasting during typhoon periods. J. Hydro Environ. Res. 22, 29–37.

Mafi-Gholami, D., Jaafari, A., Zenner, E.K., Kamari, A.N., Bui, D.T., 2020. Vulnerability of coastal communities to climate change: Thirty-year trend analysis and prospective prediction for the coastal regions of the Persian Gulf and Gulf of Oman. Sci. Total Environ. 741, 140305.

Marco, J.B., Cayuela, A., 1994. Urban flooding: The flood-planned city concept. In: Rossi, G., Harmancioğlu, N., Yevjevich, V. (Eds.), Coping with Floods. NATO ASI Series (Series E: Applied Sciences) 257. Springer, Dordrecht, pp. 705–721. https://doi.org/10.1007/978-94-011-1098-3_43.

McCuen, R.H., 2016. Modeling Hydrologic Change: Statistical Methods. Lewis Publishers, A CRC Press Company, Boca Raton, p. 448.

Melville, P., Mooney, R.J., 2005. Creating diversity in ensembles using artificial data. Inform. Fusion 6 (1), 99–111.

Ngo, P.T.T., Hoang, N.D., Pradhan, B., Nguyen, Q.K., Tran, X.T., Nguyen, Q.M., Nguyen, V.N., Samui, P., Bui, D.T., 2018. A novel hybrid swarm optimized multilayer neural network for spatial prediction of flash floods in tropical areas using sentinel-1 SAR imagery and geospatial data. Sensors. 18 (11), 3704.

Nguyen, P.T., Ha, D.H., Avand, M., Jaafari, A., Nguyen, H.D., Al-Ansari, N., Phong, T.V., Sharma, R., Kumar, R., Le, H.V., Ho, L.S., Prakash, I., Pham, B.T., 2020a. Soft computing ensemble models based on logistic regression for groundwater potential mapping. Appl. Sci. 10 (7), 2469.

Nguyen, P.T., Ha, D.H., Jaafari, A., Nguyen, H.D., Van Phong, T., Al-Ansari, N., Prakash, I., Le, H.V., Pham, B.T., 2020b. Groundwater potential mapping combining artificial neural network and real adaboost ensemble technique: the Daknong province case-study. Vietnam. Int. J. Environ. Res. Public Health 17 (7), 2473.

Nhu, V.-H., Janizadeh, S., Avand, M., Chen, W., Farzin, M., Omidvar, E., Shirzadi, A., Shahabi, H., Clague, J.J., Jaafari, A., Mansooryoor, F., Pham, B.T., Ahmad, B.B., Lee, S., 2020a. GIS-based gully erosion susceptibility mapping: a comparison of computational ensemble data mining models. Appl. Sci. 10 (6), 2039.

Nhu, V.-H., Shirzadi, A., Shahabi, H., Chen, W., Clague, J.J., Geertsema, M., Jaafari, A., Avand, M., Miraki, S., Asl, D.T., 2020b. Shallow landslide susceptibility mapping by random forest base classifier and its ensembles in a semi-arid region of Iran. Forests 11 (4), 421.

Ouma, Y., Tateishi, R., 2014. Urban flood vulnerability and risk mapping using integrated multi-parametric AHP and GIS: Methodological overview and case study assessment. Water 6 (6), 1515–1545.

Pham, B.T., Tien Bui, D., Pham, H.V., Le, H.Q., Prakash, I., Dholakia, M.B., 2017a. Landslide hazard assessment using random subspace fuzzy rules based classifier ensemble and probability analysis of rainfall data: a case study at Mu Cang Chai district, Yen Bai province (Vietnam). J. Ind. Soc. Remote Sens. 45 (4), 673–683.

Pham, B.T., Tien Bui, D., Prakash, I., Dholakia, M.B., 2017b. Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena 149, 52–63.

Pham, B.T., Jaafari, A., Prakash, I., Singh, S.K., Quoc, N.K., Bui, D.T., 2019a. Hybrid computational intelligence models for groundwater potential mapping. Catena 182, 104101.

Pham, B.T., Shirzadi, A., Shahabi, H., Omidvar, E., Singh, S.K., Sahana, M., Talebpour Asl, D., Bin Ahmad, B., Kim Quoc, N., Lee, S., 2019b. Landslide susceptibility assessment by novel hybrid machine learning algorithms. Sustainability 11 (16), 4386.

Pham, B.T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H.P.H., Phong, T.V., Nguyen, D.H., Le, H.V., Mafi-Gholami, D., Prakash, I., Thi Thuy, H., Tuyen, T.T., 2020. Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry 12 (6), 1022.

Rahmati, O., Darabi, H., Haghighi, A.T., Stefanidis, S., Kornejady, A., Nalivan, O.A., Tien Bui, D., 2019. Urban flood hazard modeling using self-organizing map neural network. Water 11 (11), 2370.

Sadler, J., Goodall, J., Morsy, M., Spencer, K., 2018. Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and random forest. J. Hydrol. 559, 43–55.

Shi, H., 2007. Best-First Decision Tree Learning. M.S. thesis. The University of Waikato, p. 121.

Shirzadi, A., Bui, D.T., Pham, B.T., Solaimani, K., Chapi, K., Kavian, A., Shahabi, H., Revhaug, I., 2017. Shallow landslide susceptibility assessment using a novel hybrid intelligence approach. Environ. Earth Sci. 76, 60.

Skurichina, M., Duin, R.P., 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern. Anal. Applic. 5 (2), 121–135.

Sun, B., Chen, H., Wang, J., 2015. An empirical margin explanation for the effectiveness of decorate ensemble learning algorithm. Knowl.-Based Syst. 78, 1–12.

Tao, D., Tang, X., Li, X., Wu, X., 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 28 (7), 1088–1099.

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.

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.

Tran, Q.C., Minh, D.D., Jaafari, A., Al-Ansari, N., Minh, D.D., Van, D.T., Nguyen, D.A., Tran, T.H., Ho, L.S., Nguyen, D.H., 2020. Novel ensemble landslide predictive models based on the hyperpipes algorithm: a case study in the Nam Dam commune, Vietnam. Appl. Sci. 10 (11), 3710. https://doi.org/10.3390/app10113710.

Wang, X., Kinsland, G., Poudel, D., Fenech, A., 2019a. Urban flood prediction under heavy precipitation. J. Hydrol. 577, 123984.

Wang, Y., Hong, H., Chen, W., Li, S., Pamučar, D., Gigović, L., Drobnjak, S., Bui, D.T., Duan, H., 2019b. A hybrid GIS multi-criteria decision-making method for flood susceptibility mapping at Shangyou, China. Remote Sens. 11 (1), 62.

Wright, J.M., 2008. Floodplain Management: Principles and Current Practices. The University of Tennessee–Knoxville, TN, USA https://training.fema.gov/hiedu/aemrc/courses/coursetreat/fm.aspx.

Zhou, Q., Pilesjö, P., Chen, Y., 2011. Estimating surface flow paths on a digital elevation model using a triangular facet network. Water Resour. Res. 47, W07522. https://doi.org/10.1029/2010WR009961.

Zhou, Q., Chen, L., Singh, V.P., Zhou, J., Chen, X., Xiong, L., 2019. Rainfall-runoff simulation in karst dominated areas based on a coupled conceptual hydrological model. J. Hydrol. 573, 524–533.

Zidane, I., Lhissou, R., Bouli, A., Mabrouki, M., 2019. An improved algorithm for mapping burnt areas in the Mediterranean forest landscape of Morocco. J. For. Res. 30 (3), 981–992.