ZULFAQAR Journal of Defence Science, Engineering & Technology
e-ISSN: 2773-5281
Vol. 5, Issue 2 (2022)
DOI: https://doi.org/10.58247/jdset-2022-0502-16
Journal homepage: https://zulfaqarjdset.upnm.edu.my

FLASH FLOOD PREDICTION IN SELANGOR USING DATA MINING TECHNIQUES

Muhammad Hakiem Halim[a], Muslihah Wook[a*], Noor Afiza Mat Razali[a], Nor Asiakin Hasbullah[a], Hasmeda Erna Che Hamid[b]

[a] Department of Computer Science, Faculty of Science and Defence Technology, National Defence University of Malaysia, Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia
[b] Centre for Research and Innovation Management, National Defence University of Malaysia, Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia

ARTICLE INFO

ARTICLE HISTORY
Received: 16-05-2022
Revised: 01-07-2022
Accepted: 01-11-2022
Published: 31-12-2022

KEYWORDS
Artificial neural network
Data mining
Flash floods
Logistic regression

ABSTRACT

Flash floods are one of the most severe natural disasters, which pose a serious threat to infrastructure and human life, especially those in urban areas. As Selangor is one of Malaysia's most developed and progressive states, the occurrence of flash floods would have a serious impact on the economy and the survival of the people. Hence, the aim of this study was to predict the possibility of flash floods in Selangor by using data mining techniques, specifically logistic regression and artificial neural network. This study proposed six factors, namely, water level, rainfall, durations, weather, minimum temperature, and maximum temperature to be utilised in the flash flood prediction model. The prediction model was constructed based on data gathered from 32 locations in the state of Selangor from June 2020 till March 2021. The performance of the model was compared using the area under the receiver operating characteristic. The findings of this study could be considered as an alternative scientific tool for flash flood prediction and may help to effectively monitor flash flood occurrences in urban areas.

1.0 INTRODUCTION

Natural disasters are phenomena that happen naturally or are caused by humans that have severe impacts on people and the environment [1]. Floods are among the most hazardous natural catastrophes because of their high intensity and unpredictability, resulting in considerable property and societal destruction. Continuous rain or on-shore winds are common causes of flooding, while strong thunderstorms, snowmelt, ice jams, and dam failures might all contribute to flooding in a local area [2]. According to Ali [3], floods occur when water is submerged or overflows onto generally dry terrain, as a result of a combination of meteorological and hydrological conditions. Floods are generally classified as flash floods, river floods, coastal floods, and urban floods. Among these flood types, flash floods are significantly more destructive and catastrophic, occurring in a shorter period of time and possessing enormous destructive potential [4-5].

Generally, flash floods are caused by continuous heavy rains that aggravate rivers or drains to become unable to support high water density, leading water to overflow from their channels. In some cases, flash floods may occur as a result of ice melt in some places during certain seasons [6]. Previous studies have reported various factors for the occurrence of flash floods, such as rainfall [7-8], flood area [6], [9], water level [10-11], lithology [12-13], slope angle [14-15], duration [16-17], land use [18-19], topography index [18, 20], elevation [21-22], soil type [13, 15], distance from river [9, 13], and temperature [6, 23]. Although it is vital to study these factors to identify the exact causes of flash floods, a better understanding of the most relevant factors that relate to a particular area is also needed.

*Corresponding Author | Wook, M. | muslihah@upnm.edu.my

Since Selangor is one of the states most affected by flash floods in Malaysia [24], studies that investigate the susceptibility to this type of natural disaster are extremely useful for developing flood risk management plans, as well as for predicting and planning warning protocols. Bari et al. [25] asserted that flash floods that occurred in the last years have caused major material damages, especially to the infrastructure from the study area. The government's continuous efforts to reduce the possibility of flash floods have included channel improvements, levee building, flood by-pass construction, sediment trap construction, and improved hydrological data recording. However, these problems have continued to persist and have been getting worse rather than better [26].

Data mining techniques employ machine learning for smartly generating nontrivial rules and patterns. The techniques have been applied in various fields of science, engineering, and business, such as retail, manufacturing, telecommunications, healthcare, insurance, and transportation. Data mining techniques have recently gained popularity for predicting natural disaster phenomena since these techniques can bring large-scale disaster data into real practice and become the necessary tools for natural disaster prediction, impact assessment, societal resilience [6], and disaster control measures [27]. Flash floods are one of the natural disasters that can be predicted and measured since large amounts of data are readily available, making the use of data mining techniques for this occurrence a perfect endeavour [13]. Additionally, the use of data mining for flash flood prediction would be beneficial because of their high predictive performance[28] and ability to handle complicated relationships between input variables [29].

The present research aims to predict potential occurrences of flash flood in Selangor. In this respect, techniques specific to data mining were used based on their reliability, and excellent performance and prediction capabilities when used with huge-scale data, which have become more common in hydrology and water management research activities [30]. The specific objectives of this study are as follows: (i) explore the factors affecting flash floods in Selangor; (ii) use data mining techniques for predicting flash flood occurrences; and (iii) validate the results using evaluation measurements.

2.0 FLASH FLOODS IN SELANGOR

The National Oceanic and Atmospheric Administration (NOAA) defines flash flood as an event caused by excessive rainfall in a short period of time. Meanwhile, the Malaysian Department of Irrigation and Drainage (DID) defines flash flood as an unexpected flood caused by heavy rainfall in the local and surrounding area [31]. Bui et al. [29] asserted that flash flood is a common natural hazard that occur rapidly and with high flow velocities, making them difficult to predict. A quick rise in water level, moves at a tremendous velocity, and enormous amounts of debris are all characteristics of flash floods. Moreover, the severity and duration of rainfall, as well as the steepness of watershed and stream gradients, are all elements that contribute to flash floods.

Flash floods are becoming more common in the western part of Peninsular Malaysia, which could be due to higher temperature, high rainfall, thunderstorm, string winds and extremely rainy hours [32]. According to Khalid and Shafiai [33], Malaysia receives an average annual rainfall of around 2,500 mm across all states, making it one of the wettest countries in the world. Flash flood has become a more serious problem in urban areas, as a result of the removal of vegetation, paving and replacement of ground cover with impermeable surfaces that accelerate runoff, and the development of drainage systems that accelerate runoff [34]. The destructive power of a flood combined with extraordinary speed and unpredictability makes flash floods the most deadly type of flood [35]. This type of flood occurs frequently and can be quite destructive, particularly to infrastructures, the environment, and the lives of those in the affected area. For example, the flash flood on 11 November 2018 had crippled most of the residential and urban areas in Kuala Lumpur, Malaysia [36].

Selangor is the most developed state in the western part of Peninsular Malaysia, which has experienced a rapid urbanisation process since Malaysia's independence in 1957. As shown in Fig. 1, flash floods in Selangor are obviously linked to rapid development, which has resulted in the loss of green and forested regions, as well as the replacement of natural surfaces with roofing and concrete [37]. As a result, the soil's ability to absorb rain water is reduced, potentially causing damage to the surrounding environment, particularly to flora and fauna [26].

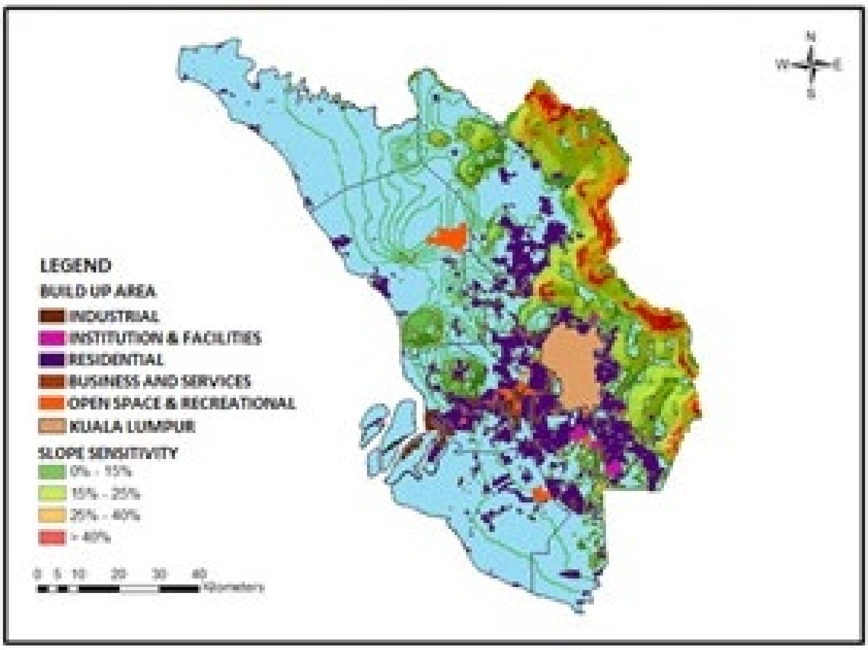


Fig. 1. Build up activities in Selangor [37]

3.0 DATA MINING IN FLASH FLOODS PREDICTION

Prior to delving into the main issue of the research, it is necessary to clarify the fundamental concepts of data mining. According to Kantardzic [38], data mining is a process of discovering various models, summaries, and derived values from a set of data. Data mining is also known as a computation approach for identifying historical patterns and trends from data, and for developing models for predicting future trends and patterns [39]. This approach comprises various techniques drawn from statistics, mathematics, and computer science, which can extract significant knowledge from extensive datasets. Data mining techniques can be classified into two categories, namely, predictive data mining and descriptive data mining. The goal of predictive data mining is to create a model that can be used for classification, prediction, estimation, and other similar tasks. Whereas descriptive data mining is concerned with identifying patterns and relationships in large datasets. Numerous techniques exist for performing both predictive and descriptive data mining. Multiple linear regression, logistic regression, decision tree, and artificial neural networks are some of the prominent techniques for predictive data mining. Meanwhile, descriptive data mining can be performed using association rules and cluster analysis techniques, to name a few.

Several research works have applied data mining techniques to process flash flood data. For instance, Panahi et al. [40] used data mining techniques, namely, convolutional neural networks and recurrent neural networks, to predict and map spatially explicit flash flood probability in northern Iran. These techniques were successful in capturing the heterogeneity of spatial patterns of flash flood probability in the flood area. Similarly, Shirzadi et al. [41] utilised the Bayesian belief network model to examine flash flood susceptibility mapping in Haraz, Iran. Their empirical work showed that the proposed techniques were promising for managing risk in flash flood-prone areas around the world. Likewise, Janizadeh et al. [13] highlighted five data mining techniques, namely, alternating decision tree, functional tree, kernel logistic regression, multilayer perceptron, and quadratic discriminant analysis, for predicting flash flood susceptibility in the Tafresh watershed, Iran. Their findings revealed that all five techniques are appropriate for mapping flash flood vulnerability in different places, thereby, able to protect people from catastrophic flooding. In the Malaysian context, Wardah et al. [42] used the artificial neural network technique for predicting flash floods in Klang River. Their findings indicated that this technique gave a satisfactory performance in predicting flash floods, specifically the area-averaged rainfall depth during convective rain events of a particular duration. Kia et al. [43] had also used artificial neural network with geographic information system (GIS) to model and simulate flood-prone areas in the southern part of Peninsular Malaysia. Their verification results revealed that the predicted and observed hydrological records were in reasonable agreement.

Thus, this study has concluded that previous studies in this field have mostly concentrated on the use of data mining techniques for predicting flash floods. Additionally, most of the reviewed studies that utilised data mining techniques have been conducted in Iran for predicting flash flood occurrences. To this end, studies that applied data mining techniques for predicting flash floods in Malaysia have been by Wardah et al. [42] and Kia et al. [43], specifically in the capital of Kuala Lumpur and the state of Johor, respectively. Both studies used artificial neural network for making flash flood predictions. Considering the frequent occurrences of flash floods in Selangor, only a limited number of studies has applied data mining techniques for predicting flash floods in this state. Therefore, in the subsequent sections, this study explored the prediction success of two data mining techniques, namely, logistics regression (LR) and artificial neural network (ANN) for assessing flash flood prediction. These two techniques were chosen because of the nature of data used in this study.

3.1 Logistic Regression (LR)

LR is a mathematical model formed based on linear regression, with the goal of predicting the relationship between the dependent variable (Y) and a set of independent variables (X). Linear regression is a technique for estimating data by establishing a straight-line equation to model or estimate data points. On the other hand, LR does not consider the relationship between the two variables as a straight line [6]. Instead, LR employs the natural logarithm function to determine the relationship between the variables, and test data is often used to determine the coefficients of the relationship. The natural logarithm function can then use these coefficients in the logistic equation to estimate future results. To calculate the probability, LR employs the concept of odds ratio, which is defined as the ratio of the probability of an event occurring against the probability of it not occurring. In contrast to other statistical techniques, LR does not require any assumptions to be established prior to conducting the analysis [44]. Additionally, it accepts all data types, including scale, nominal, and categorical data for the input, and only binary data for the output.

Nandi et al. [45] applied LR to quantify hazard potentials and to map 14 flood factors in Jamaica. They also applied two modelling techniques, namely, principal component analysis and frequency distribution analysis. Their findings revealed that these techniques yielded satisfactory validation results based on the receiver operating characteristic curve. Tehrany et al. [44] utilised LR with frequency ratio and statistical index for a 13-factor flood susceptibility mapping in Brisbane, Australia. Their validation results from area under the curve showed that the flood susceptibility maps generated from the LR, and statistical index were more reliable compared to the results obtained from the frequency ratio. Meanwhile, the prediction rate of LR was the highest compared to frequency ratio and statistical index. Based on these findings, they concluded that the performances of LR and statistical index were acceptable for making predictions.

A study was conducted in Beijing, China by Zhao et al. [46] to assess urban flood susceptibility using semi-supervised machine learning model (WELLSVM). Their study specifically compared the performance of WELLSVM against LR, ANN, and support vector machine (SVM) for identifying nine explanatory factors of urban flooding. The results showed that while the WELLSVM was well suited for assessing flood susceptibility, its performance was strongly dependent on the accuracy of the labelled flood inventories. Recently, Lopez and Rodriguez [47] applied LR to evaluate the probability of flash flood occurrences based on three factors in several watersheds within the city of São Paulo, Brazil. The results showed that the probability of a flash flood occurring in each number of watersheds was acceptable, whereas the probability of a flash flood not occurring has a better success rate.

3.2 Artificial Neural Network

ANN is a computational model of the human brain. It is inspired from the way actual brain systems work, making computers more capable of learning from information [6]. ANN is made up of artificial neurons called nodes that are always linked together. Each node represents a processing unit, and the links between the nodes specify the causal relationship between connected nodes. Typically, ANN consists of three layers: an input layer that comprises several nodes; one or more hidden layers; and an output layer where the output is displayed. ANN adapts and modifies its configuration in response to the patterns in the information that enters the network during the learning phase, and like humans, it learns by example.

The ANN model can be trained using a significant amount of data. Training entails providing input data and specifying the desired output. It is undeniable that an ANN draws its computing power from two factors, namely, its large parallel distributed structure and its capability for learning, and thus, generalisation. Generalisation is the process by which an ANN generates adequate outputs for new inputs that were not encountered during the learning phase [38]. Since it could learn and generalise, ANN has been successfully applied for pattern recognition and modelling across numerous fields.

In the case of flood domain, Kourgialas and Karatzas [48] have introduced a method for measuring flood hazards on a national scale by utilising ANN and a multi-criteria analysis in a GIS environment. The proposed method was examined in Greece, where flash floods are a common occurrence that have caused major damages in both rural and urban areas. They integrated seven factor maps that are directly related to flood generation in a GIS environment to identify the most flood-prone areas. The results indicated that the output of the ANN model exhibited a good performance. Nikoo et al. [49] proposed a new flood prediction system based on an ANN model coupled with a social-based algorithm for the flood-routing problem of the Kheir Abad River in Iran. The findings showed that the proposed ANN model was relatively efficient and can be extended to various ANN-based hydrological models. Kia et al. [43] applied ANN and GIS to develop a flood model for modelling and simulating flood occurrences involving seven flood causative factors in the Johor River Basin. Their findings indicated that the models were capable of simulating both peak and base river flows.

4.0 METHODOLOGY

The methodology for the current study consisted of several steps, as depicted in Fig. 2. The details of each step are explained in the subsequent sections.

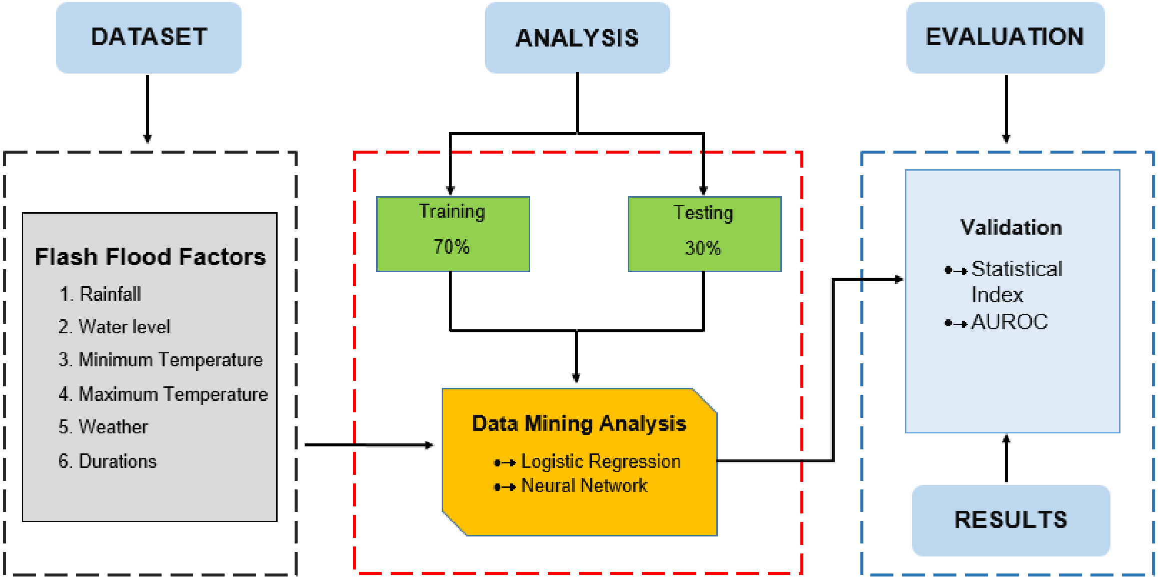


Fig. 2. Methodology of this study

4.1 Flash Flood Factors

According to previous studies, flash floods are essentially determined by four fundamental factors: precipitation, topography [46], geology, and human activity [50]. Six factors, including rainfall, water level, minimum temperature, maximum temperature, weather, and durations, were preliminarily determined based on selection criteria (e.g., objectivity, representativeness, and availability) and the process of flash flood generation. Table 1 summarises the characteristics of each flash flood factor considered in this study.

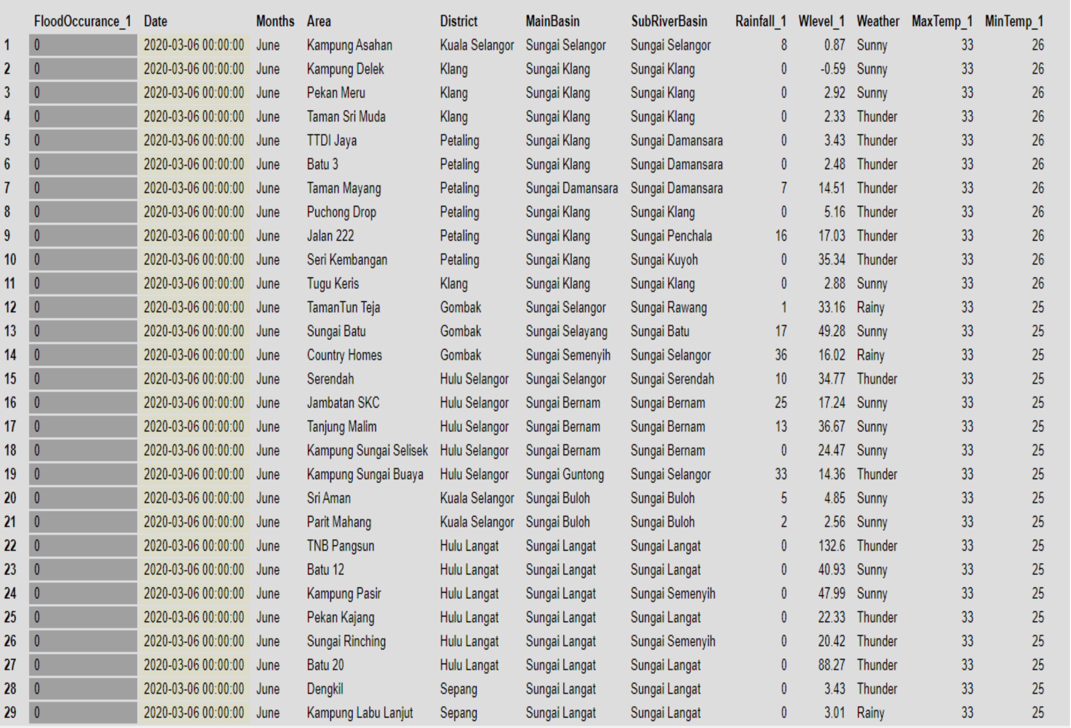
*Corresponding Author | Wook, M. | muslihah@upnm.edu.my

Table 1. Description of flash flood factors

Factor	Data type	Measurement
Rainfall	Double	mm
Water level	Double	m
Minimum Temperature	Integer	℃
Maximum Temperature	Integer	℃
Weather	String	Climate changes
Durations	Dates	Days

4.2 Data Description

The data for this study were gathered from the websites of the Selangor Department of Irrigation and Drainage (DID) (http://infobanjirjps.selangor.gov.my/rainfall.html) and the Malaysian Meteorological Department (MMD) (https://www.met.gov.my/?lang=en). The DID was used to gather data on water levels and rainfall, while the MMD was used to collect data on weather, as well as minimum and maximum temperature readings. The researchers compiled the data for the dates. Additionally, data on Selangor locales were collected, including the area, district, main basin, and sub-river basin. Between June 2020 and March 2021, a total of 9665 datasets were collected from 32 different locations throughout Selangor. These datasets were then partitioned into a training set and a testing set. The training and testing datasets are 70% and 30%, respectively, because these percentages are commonly used in previous studies on flash floods (e.g. [12], [51], [52]). The training dataset (6765) was used to develop the data mining models, while the testing dataset (2900) was utilised to validate the algorithms. Fig. 3 depicts a sample of the dataset obtained in Selangor.



	FloodOccurance_1	Date	Months	Area	District	MainBasin	SubRiverBasin	Rainfall_1	Wlevel_1	Weather	MaxTemp_1	MinTemp_1
1	0	2020-03-06 00:00:00	June	Kampung Asahan	Kuala Selangor	Sungai Selangor	Sungai Selangor	8	0.87	Sunny	33	26
2	0	2020-03-06 00:00:00	June	Kampung Delek	Klang	Sungai Klang	Sungai Klang	0	-0.59	Sunny	33	26
3	0	2020-03-06 00:00:00	June	Pekan Meru	Klang	Sungai Klang	Sungai Klang	0	2.92	Sunny	33	26
4	0	2020-03-06 00:00:00	June	Taman Sri Muda	Klang	Sungai Klang	Sungai Klang	0	2.33	Thunder	33	26
5	0	2020-03-06 00:00:00	June	TTDI Jaya	Petaling	Sungai Klang	Sungai Damansara	0	3.43	Thunder	33	26
6	0	2020-03-06 00:00:00	June	Batu 3	Petaling	Sungai Klang	Sungai Damansara	0	2.48	Thunder	33	26
7	0	2020-03-06 00:00:00	June	Taman Mayang	Petaling	Sungai Damansara	Sungai Damansara	7	14.51	Thunder	33	26
8	0	2020-03-06 00:00:00	June	Puchong Drop	Petaling	Sungai Klang	Sungai Klang	0	5.16	Thunder	33	26
9	0	2020-03-06 00:00:00	June	Jalan 222	Petaling	Sungai Klang	Sungai Penchala	16	17.03	Thunder	33	26
10	0	2020-03-06 00:00:00	June	Seri Kembangan	Petaling	Sungai Klang	Sungai Kuyoh	0	35.34	Thunder	33	26
11	0	2020-03-06 00:00:00	June	Tugu Keris	Klang	Sungai Klang	Sungai Klang	0	2.88	Sunny	33	26
12	0	2020-03-06 00:00:00	June	TamanTun Teja	Gombak	Sungai Selangor	Sungai Rawang	1	33.16	Rainy	33	25
13	0	2020-03-06 00:00:00	June	Sungai Batu	Gombak	Sungai Selayang	Sungai Batu	17	49.28	Sunny	33	25
14	0	2020-03-06 00:00:00	June	Country Homes	Gombak	Sungai Semenyih	Sungai Selangor	36	16.02	Rainy	33	25
15	0	2020-03-06 00:00:00	June	Serendah	Hulu Selangor	Sungai Selangor	Sungai Serendah	10	34.77	Thunder	33	25
16	0	2020-03-06 00:00:00	June	Jambatan SKC	Hulu Selangor	Sungai Bernam	Sungai Bernam	25	17.24	Sunny	33	25
17	0	2020-03-06 00:00:00	June	Tanjung Malim	Hulu Selangor	Sungai Bernam	Sungai Bernam	13	36.67	Sunny	33	25
18	0	2020-03-06 00:00:00	June	Kampung Sungai Selisek	Hulu Selangor	Sungai Bernam	Sungai Bernam	0	24.47	Sunny	33	25
19	0	2020-03-06 00:00:00	June	Kampung Sungai Buaya	Hulu Selangor	Sungai Guntong	Sungai Selangor	33	14.36	Thunder	33	25
20	0	2020-03-06 00:00:00	June	Sri Aman	Kuala Selangor	Sungai Buloh	Sungai Buloh	5	4.85	Sunny	33	25
21	0	2020-03-06 00:00:00	June	Parit Mahang	Kuala Selangor	Sungai Buloh	Sungai Buloh	2	2.56	Sunny	33	25
22	0	2020-03-06 00:00:00	June	TNB Pangsun	Hulu Langat	Sungai Langat	Sungai Langat	0	132.6	Thunder	33	25
23	0	2020-03-06 00:00:00	June	Batu 12	Hulu Langat	Sungai Langat	Sungai Langat	0	40.93	Sunny	33	25
24	0	2020-03-06 00:00:00	June	Kampung Pasir	Hulu Langat	Sungai Langat	Sungai Semenyih	0	47.99	Sunny	33	25
25	0	2020-03-06 00:00:00	June	Pekan Kajang	Hulu Langat	Sungai Langat	Sungai Langat	0	22.33	Thunder	33	25
26	0	2020-03-06 00:00:00	June	Sungai Rinching	Hulu Langat	Sungai Langat	Sungai Semenyih	0	20.42	Thunder	33	25
27	0	2020-03-06 00:00:00	June	Batu 20	Hulu Langat	Sungai Langat	Sungai Langat	0	88.27	Thunder	33	25
28	0	2020-03-06 00:00:00	June	Dengkil	Sepang	Sungai Langat	Sungai Langat	0	3.43	Thunder	33	25
29	0	2020-03-06 00:00:00	June	Kampung Labu Lanjut	Sepang	Sungai Langat	Sungai Langat	0	3.01	Rainy	33	25

Fig. 3. A sample of dataset in Selangor

4.3 Logistic Regression Analysis

The probability of flash flood occurrences was constructed using the LR model, as shown by Equation (1) and (2). This technique was chosen because it can incorporate all the data types for the dependent and independent variables in this study, which consisted of scale, nominal, and categorical data. Like other regression analyses, the LR is useful when the dependent variable is dichotomous or has binary values, such as 1 or 0, yes or no, success or failure, presence or absence, and flooding or no flooding [53]. This model was also found to be effective for predicting the presence or absence of features based on values of predictor variables [54]. This type of values is commonly interpreted as the probability of one state of the dependent variable, as they are limited to fall between 0 and 1 [55]. In this study, the dependent variable was a binary variable representing the occurrence or absence of a flash flood. Quantitatively, the relationship between flash flood occurrence and its dependency on several variables can be based on the logistic function, $f(z)$, which is expressed as follows:

$$p = \frac{1}{1 + e^{-z}} \tag{1}$$

where p represents the probability of a flash flood occurrence. This probability varies from 0 to 1 in understanding that the data was “non-flash flood” and “flash flood” on an S-shaped curve (sigmoid). The variable z represents flash flood causal factors, which are assumed as a linear combination. Consequently, LR requires fitting the following equation to the collected data:

$$z = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots + b_nx_n \tag{2}$$

where b_0 represents the intercept of the model, $b_i(i = 0, 1, 2, \ldots, n)$ represents the coefficient of the LR model, and $x_i(i = 0, 1, 2, \ldots, n)$ represents the flash flood conditioning variables (rainfall, water level, duration, weather, minimum temperature, and maximum temperature). The generated linear model is then a LR of the presence or absence of flash flood events (present conditions) based on the independent (pre-failure conditions) variables.

4.4 Artificial Neural Network Analysis

The ANN analysis in this study was trained using input data (flash flood factors) and ground truth labels (0 and 1, or non-flash flood and flash flood). The analysis results were then used to predict the output class (flash flood occurrences). The popularity of an ANN technique lies in its information processing characteristics, such as non-linearity, noise tolerance, and generalisation capabilities [56]. This technique was chosen for this study mainly due to the completion of the information processing through an interactive link between neurons without needing a pre-designed mathematical model [46]. As described in subsection 2.2.2, ANN is composed of three layers: input layer, hidden layer, and output layer that link the layers together. The net input to the hidden layer and output layer is given as follows:

$$y_i = \sum_{j=1}^{N} w_{ji}x_j + w_{i0} \tag{3}$$

where N represents the total number of nodes in node i’s upper layer, w_{ij} represents the weight between node i and node j, x_i represents the output value from node j, w_{i0} represents node i’s bias, and y_i represents node i’s input signal, which is then passed through a transfer function.

4.5 Validation Methods

The validity of the LR and ANN models based on these datasets were tested using statistical metrics, namely, sensitivity, specificity, and precision. The terms sensitivity and specificity refer to the number of pixels properly categorised as non-flash flood or flash flood, respectively [41, 57]. A confusion matrix approach was applied to calculate the statistical index. This approach is a binary classifier that is formulated as two-class pattern recognition [29]. To measure the predictive performance, the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) can be calculated by comparing the observed output with the model prediction output in both analyses. Therefore, based on the calculation of the TP, FP, FN, and TN values, the sensitivity, F-measure, precision, accuracy, and AUROC can be calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$F - measure = \frac{2\ \times Sensitivity\ \times Precision}{Sensitivity + Precision} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

In addition to these four indices, the receiver operating characteristic (ROC) curve can also be used to analyse both models' overall prediction ability. This technique is widely used in previous studies on flood prediction [5, 41]. As a result, when categorising instances in a dataset, the area under the ROC curve (AUROC) is frequently computed to measure a model's performance. AUROC between 0.5 and 0.6 usually suggests a weak model, whereas AUROC between 0.6 and 0.7 suggests a bad performance. AUROC between 0.7 and 0.8 suggests that the constructed model is moderately suitable for the dataset, while AUROC > 0.8 indicates that the developed model is highly suitable for the dataset [29]. AUROC values can be calculated using the following equation:

$$AUROC = \frac{(\sum TP + \sum TN)}{(P + N)} \tag{8}$$

where TP represents correctly classified flash flood pixels, TN represents correctly classified non-flash flood pixels, P represents the total number of flash flood pixels, and N represents the total number of non-flash flood pixels.

5.0 RESULTS AND DISCUSSION

The LR and ANN models have been assessed based on the six flash flood factors, as observed using the training and testing datasets listed in Table 2. The performance of these models was evaluated using statistical indices, namely, sensitivity, precision, accuracy, and F-measure. The value of the training dataset in LR model for sensitivity, precision, accuracy, and F-measure was 0.996 each. The results of the training dataset in ANN model for sensitivity, precision, accuracy, and F-measure were 0.982, 0.965, 0.982, and 0.974, respectively. Meanwhile, the value of the testing dataset in LR model for sensitivity, precision, accuracy, and F-measure was 0.997 each. The value of the testing dataset in ANN model for sensitivity, precision, and accuracy was 0.990 each, while the value for F-measure was 0.989. Based on these results, the LR model was found to be better, with a higher flash flood prediction accuracy compared to the ANN model. This result is in line with the result obtained by Bui et al. [58], who found that the single LR model was effective in delineating flash flood predictions.

Table 2. Evaluation performance of the LR and ANN models in predicting flash flood occurrences in Selangor

Statistical Index Performance (Validation)	Dataset			
	Training		Testing	
	Logistic Regression	Artificial neural Network	Logistic Regression	Artificial neural Network
Sensitivity	0.996	0.982	0.997	0.990
Precision	0.996	0.965	0.997	0.990
Accuracy	0.996	0.982	0.997	0.990
F-measure	0.996	0.974	0.997	0.989

To identify TP and FP rates in the current study, the ROC curve was used to plot the sensitivity of the model (the percentage of existing flash flood pixels predicted correctly by the model) against 1-specificity (the percentage of predicted flash flood pixels over the total study area). The performance of the LR and ANN models were determined using the AUROC values, as shown in Fig. 4. The accuracy of the models can be explained in part by the results obtained (AUROC values).The higher the AUROC, the better the model's prediction ability [59]. These validation values were generated using the Orange software. The LR model (AUROC = 0.954) outperformed the ANN model (AUROC = 0.866) based on the results of the training dataset (Fig. 4a). In contrast, based on the results of the curve of prediction rate obtained by using the testing dataset (Fig. 4b), the AUROC value for the LR was 0.985, which outperformed the value for ANN (AUROC = 0.975), indicating that the LR was superior. It can be concluded that both the LR and ANN models were successful, with prediction rate curves greater than 0.9, indicating that these models are acceptable for this study area. Additionally, the results indicated that these models are acceptable for predicting flash floods with small differences with the real values.

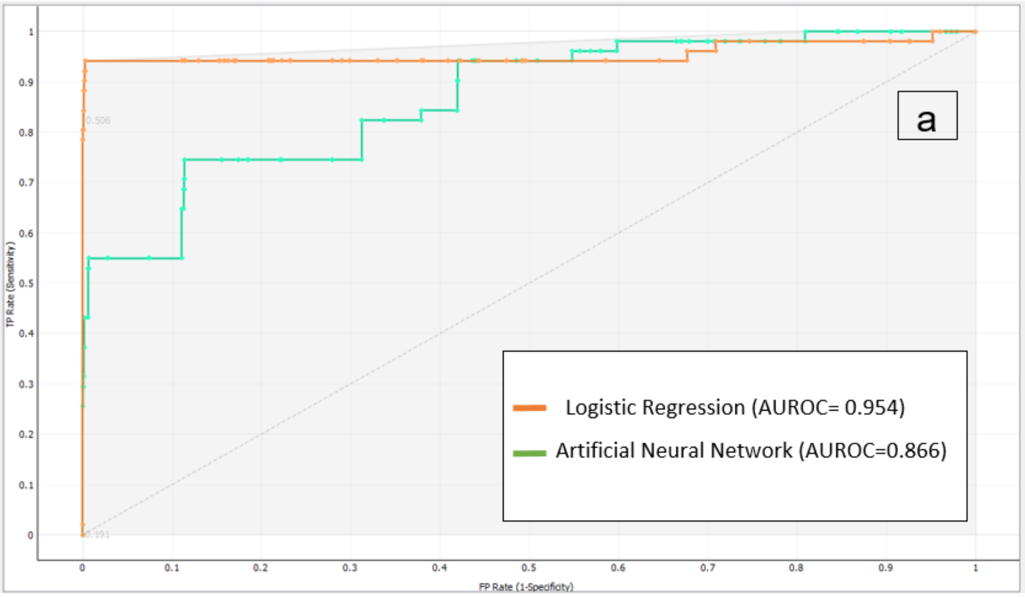


Fig. 4. Comparison of model performances using AUROC curve: (a) training dataset; and (b) testing dataset

6.0 CONCLUSION

Flash floods are the world's most destructive natural disasters that pose a serious threat to human life. Thus, flash flood prediction is critical as the first step towards minimising damages caused by floods. The prediction of flash floods in Selangor is imperative since this disaster occurs in this state every year compared to in other states in Malaysia. Hence, the results of this study may be beneficial for the management of this disaster by the authorities or policy makers to alleviate the devastating impacts of flash floods, particularly in Selangor. This study has also shown the conceptual implications based on the incorporation of six factors, namely, rainfall, water level, minimum temperature, maximum temperature, weather, and durations for predicting flash floods in Selangor. Furthermore, this study has shown that the proposed LR and ANN models are applicable for predicting flash flood occurrences in this state.

Nevertheless, this study is the first to investigate flash flood prediction in Selangor based on six factors using the LR and ANN models. Although both techniques were shown to achieve better performance with high accuracy values, their integration with other data mining techniques, such as SVM, decision tree, naïve Bayes, association rules, and cluster analysis should be considered for future works.

7.0 ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Higher Education Malaysia (MOHE), and the Centre for Research and Innovation Management, National Defence University of Malaysia for supporting this research (RACER/1/2019/ICT04/UPNM/1).

List of Reference

[1] Prama, M., Omran, A., Schröder, D., & Abouelmagd, A. (2020). Vulnerability assessment of flash floods in Wadi Dahab Basin, Egypt. Environmental Earth Sciences, 79, 1-17.

[2] Estrada, M. A. R., Koutronas, E., Tahir, M., & Mansor, N. (2017). Hydrological hazard assessment: THE 2014–15 Malaysia floods. International journal of disaster risk reduction, 24, 264-270.

[3] Ali, A. (2018). Flood inundation modeling and hazard mapping under uncertainty in the Sungai Johor Basin, Malaysia. CRC press.

[4] Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536.

[5] Pham, B. T., Avand, M., Janizadeh, S., Phong, T. V., Al-Ansari, N., Ho, L. S., ... & Prakash, I. (2020). GIS based hybrid computational approaches for flash flood susceptibility assessment. Water, 12(3), 683.

[6] Saravi, S., Kalawsky, R., Joannou, D., Rivas Casado, M., Fu, G., & Meng, F. (2019). Use of artificial intelligence to improve resilience and preparedness against adverse flood events. Water, 11(5), 973.

[7] A. Rashid, R. A., Nohuddin, P. N., & Zainol, Z. (2017). Association rule mining using time series data for Malaysia climate variability prediction. In Advances in Visual Informatics: 5th International Visual Informatics Conference, IVIC 2017, Bangi, Malaysia, November 28–30, 2017, Proceedings 5 (pp. 120-130). Springer International Publishing.

[8] Razali, N., Ismail, S., & Mustapha, A. (2020). Machine learning approach for flood risks prediction. IAES International Journal of Artificial Intelligence, 9(1), 73.

[9] Vafakhah, M., Mohammad Hasani Loor, S., Pourghasemi, H., & Katebikord, A. (2020). Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. Arabian Journal of Geosciences, 13, 1-16.

[10] Chang LiChiu, C. L., Chang FiJohn, C. F., Yang ShunNien, Y. S., Kao IFeng, K. I., Ku YingYu, K. Y., Kuo ChunLing, K. C., & Ir. Mohd, Z. M. A. (2019). Building an intelligent hydroinformatics integration platform for regional flood inundation warning systems.

[11] Widiasari, I. R., & Nugroho, L. E. (2017, November). Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining. In 2017 International Conference on Innovative and Creative Information Technology (ICITech) (pp. 1-5). IEEE.

[12] Costache, R. (2019). Flash-flood Potential Index mapping using weights of evidence, decision Trees models and their novel hybrid integration. Stochastic Environmental Research and Risk Assessment, 33(7), 1375-1402.

[13] Janizadeh, S., Avand, M., Jaafari, A., Phong, T. V., Bayat, M., Ahmadisharaf, E., ... & Lee, S. (2019). Prediction success of machine learning methods for flash flood susceptibility mapping in the Tafresh watershed, Iran. Sustainability, 11(19), 5426.

[14] ong, H., Panahi, M., Shirzadi, A., Ma, T., Liu, J., Zhu, A. X., ... & Kazakis, N. (2018). Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. Science of the total Environment, 621, 1124-1141.

[15] Kanani-Sadat, Y., Arabsheibani, R., Karimipour, F., & Nasseri, M. (2019). A new approach to flood susceptibility assessment in data-scarce and ungauged regions based on GIS-based hybrid multi criteria decision-making method. Journal of hydrology, 572, 17-31.

[16] Jangyodsuk, P., Seo, D. J., Elmasri, R., & Gao, J. (2015, November). Flood prediction and mining influential spatial features on future flood with causal discovery. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 1462-1469). IEEE.

[17] Ruslan, F. A., Zakaria, N. K., & Adnan, R. (2013, August). Flood modelling using artificial neural

network. In 2013 IEEE 4th Control and System Graduate Research Colloquium (pp. 116-120). IEEE.

[18] Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena, 125, 91-101.

[19] Khosravi, K., Nohani, E., Maroufinia, E., & Pourghasemi, H. R. (2016). A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. Natural hazards, 83, 947-987.

[20] 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. Journal of hydrology, 512, 332-343.

[21] Al-Azzam, O., Sarsar, D., Seifu, K., & Mekni, M. (2014). Flood prediction and risk assessment using advanced geo-visualization and data mining techniques: a case study in the Red-Lake valley. Journal of Theoretical and Applied Information Technology, 87(3), 18-27.

[22] 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. Journal of hydrology, 504, 69-79.

[23] Panigrahi, B. K., Das, S., Nath, T. K., & Senapati, M. R. (2018). An application of data mining techniques for flood forecasting: application in rivers Daya and Bhargavi, India. Journal of The Institution of Engineers (India): Series B, 99, 331-342.

[24] Asmawi, M. Z., Mustofa, N. S., Abd Aziz, A. R., & Rahim, A. A. (2020). The Rainwater Harvesting (RWH) as A Flash Flood Mitigation Measure in Ken Rimba Shah Alam, Selangor. Journal of Social Science and Humanities, 3(1), 14-22.

[25] Bari, M. A., Alam, L., Alam, M. M., Rahman, L. F., & Pereira, J. J. (2021). Estimation of losses and damages caused by flash floods in the commercial area of Kajang, Selangor, Malaysia. Arabian Journal of Geosciences, 14, 1-9.

[26] Bhuiyan, T. R., Hasan, M. I., Reza, E. A. C., & Pereira, J. J. (2018). Direct impact of flash floods in Kuala Lumpur City: Secondary data-based analysis. ASM Science Journal, 11(3), 145-157.

[27] Makhtar, M., Harun, N. A., Aziz, A. A., Zakaria, Z. A., Abdullah, F. S., & Jusoh, J. A. (2017). An association rule mining approach in predicting flood areas. In Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016 Proceedings Second (pp. 437-446). Springer International Publishing.

[28] Dikshit, A., Pradhan, B., & Alamri, A. M. (2021). Pathways and challenges of the application of artificial intelligence to geohazards modelling. Gondwana Research, 100, 290-301.

[29] 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.

[30] Ibrahim, K. S. M. H., Huang, Y. F., Ahmed, A. N., Koo, C. H., & El-Shafie, A. (2022). A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. Alexandria Engineering Journal, 61(1), 279-303.

[31] Suparta, W., Rahman, R., & Singh, M. S. J. (2014, June). Monitoring the variability of precipitable water vapor over the Klang Valley, Malaysia during flash flood. In IOP Conference Series: Earth and Environmental Science (Vol. 20, No. 1, p. 012057). IOP Publishing.

[32] Tang, K. H. D. (2019). Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations. Science of the Total Environment, 650, 1858-1871.

[33] Khalid, M. S. B., & Shafiai, S. B. (2015). Flood disaster management in Malaysia: An evaluation of the effectiveness flood delivery system. International Journal of Social Science and Humanity, 5(4), 398.

[34] Buslima, F. S., Omar, R. C., Jamaluddin, T. A., & Taha, H. (2018). Flood and flash flood geo-hazards in Malaysia. Int. J. Eng. Technol, 7(4), 760-764.

[35] Samsuri, N. O. R. A. S. H. I. K. I. N., Abu Bakar, R., & Unjah, T. A. N. O. T. (2018). Flash flood impact in Kuala Lumpur–Approach review and way forward. International Journal of the Malay World and Civilisation, 6(1), 69-76.

[36] Mohtar, W. H. M. W., Abdullah, J., Maulud, K. N. A., & Muhammad, N. S. (2020). Urban flash flood index based on historical rainfall events. Sustainable Cities and Society, 56, 102088.

[37] Hua, A. K. (2018). Applied GIS in environmental sensitivity development based slope failure. International Journal of Research, 5(16), 1286-1289.

[38] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[39] Wook, M., Ismail, S., Yusop, N. M. M., Ahmad, S. R., & Ahmad, A. (2019). Identifying priority antecedents of educational data mining acceptance using importance-performance matrix analysis.

Education and Information Technologies, 24(2), 1741-1752.

[40] Panahi, M., Jaafari, A., Shirzadi, A., Shahabi, H., Rahmati, O., Omidvar, E., ... & Bui, D. T. (2021). Deep learning neural networks for spatially explicit prediction of flash flood probability. Geoscience Frontiers, 12(3), 101076.

[41] Shirzadi, A., Asadi, S., Shahabi, H., Ronoud, S., Clague, J. J., Khosravi, K., ... & Bui, D. T. (2020). A novel ensemble learning based on Bayesian Belief Network coupled with an extreme learning machine for flash flood susceptibility mapping. Engineering Applications of Artificial Intelligence, 96, 103971.

[42] Wardah, T., Bakar, S. A., Bardossy, A., & Maznorizan, M. (2008). Use of geostationary meteorological satellite images in convective rain estimation for flash-flood forecasting. Journal of Hydrology, 356(3-4), 283-298.

[43] 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. Environmental earth sciences, 67, 251-264.

[44] Tehrany, M. S., Kumar, L., Jebur, M. N., & Shabani, F. (2019). Evaluating the application of the statistical index method in flood susceptibility mapping and its comparison with frequency ratio and logistic regression methods. Geomatics, Natural Hazards and Risk.

[45] Nandi, A., Mandal, A., Wilson, M., & Smith, D. (2016). Flood hazard mapping in Jamaica using principal component analysis and logistic regression. Environmental Earth Sciences, 75, 1-16.

[46] Zhao, G., Pang, B., Xu, Z., Peng, D., & Xu, L. (2019). Assessment of urban flood susceptibility using semi-supervised machine learning model. Science of the Total Environment, 659, 940-949.

[47] Viteri López, A. S., & Morales Rodriguez, C. A. (2020). Flash flood forecasting in São Paulo using a binary logistic regression model. Atmosphere, 11(5), 473.

[48] Kourgialas, N. N., & Karatzas, G. P. (2017). A national scale flood hazard mapping methodology: The case of Greece–Protection and adaptation policy approaches. Science of the Total Environment, 601, 441-452.

[49] Nikoo, M., Ramezani, F., Hadzima-Nyarko, M., Nyarko, E. K., & Nikoo, M. (2016). Flood-routing modeling with neural network optimized by social-based algorithm. Natural hazards, 82, 1-24.

[50] Cao, Y., Jia, H., Xiong, J., Cheng, W., Li, K., Pang, Q., & Yong, Z. (2020). Flash flood susceptibility assessment based on geodetector, certainty factor, and logistic regression analyses in Fujian Province, China. ISPRS International Journal of Geo-Information, 9(12), 748.

[51] Bui, D. T., Hoang, N. D., Martínez-Álvarez, F., Ngo, P. T. T., Hoa, P. V., Pham, T. D., ... & 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. Science of The Total Environment, 701, 134413.

[52] Costache, R., Ngo, P. T. T., & Bui, D. T. (2020). Novel ensembles of deep learning neural network and statistical learning for flash-flood susceptibility mapping. Water, 12(6), 1549.

[53] Rasyid, A. R., Bhandary, N. P., & Yatabe, R. (2016). Performance of frequency ratio and logistic regression model in creating GIS based landslides susceptibility map at Lompobattang Mountain, Indonesia. Geoenvironmental Disasters, 3, 1-16.

[54] Pradhan, B., & Lee, S. (2010). Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. Environmental Earth Sciences, 60, 1037-1054.

[55] Xu, C., Xu, X., Dai, F., Wu, Z., He, H., Shi, F., ... & Xu, S. (2013). Application of an incomplete landslide inventory, logistic regression model and its validation for landslide susceptibility mapping related to the May 12, 2008 Wenchuan earthquake of China. Natural hazards, 68, 883-900.

[56] Chiang, Y. M., Chang, F. J., Jou, B. J. D., & Lin, P. F. (2007). Dynamic ANN for precipitation estimation and forecasting from radar observations. Journal of Hydrology, 334(1-2), 250-261.

[57] Khosravi, K., Shahabi, H., Pham, B. T., Adamowski, J., Shirzadi, A., Pradhan, B., ... & Prakash, I. (2019). A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods. Journal of Hydrology, 573, 311-323.

[58] Bui, D. T., Panahi, M., Shahabi, H., Singh, V. P., Shirzadi, A., Chapi, K., ... & Ahmad, B. B. (2018). Novel hybrid evolutionary algorithms for spatial prediction of floods. Scientific reports, 8(1), 15364.

[59] Islam, A. R. M. T., Talukdar, S., Mahato, S., Kundu, S., Eibek, K. U., Pham, Q. B., ... & Linh, N. T. T. (2021). Flood susceptibility modelling using advanced ensemble machine learning models. Geoscience Frontiers, 12(3), 101075.