DOI: 10.15849/IJASCA.220328.09

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 1, March 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Comparative Assessment of Data Mining Techniques for Flash Flood Prediction

Muhammad Hakiem Halim, Muslihah Wook, Nor Asiakin Hasbullah, Noor Afiza Mat Razali and Hasmeda Erna Che Hamid

Department of Computer Science, Faculty of Defence Science and Technology,
National Defence University of Malaysia, Kuala Lumpur, Malaysia
email: 3201298@alfateh.upnm.edu.my
Department of Computer Science, Faculty of Defence Science and Technology,
National Defence University of Malaysia, Kuala Lumpur, Malaysia
email: muslihah@upnm.edu.my (corresponding author)
Department of Computer Science, Faculty of Defence Science and Technology,
National Defence University of Malaysia, Kuala Lumpur, Malaysia
email: asiakin@upnm.edu.my
Department of Computer Science, Faculty of Defence Science and Technology,
National Defence University of Malaysia, Kuala Lumpur, Malaysia
email: noorafiza@upnm.edu.my
Centre for Research and Innovation Management, National Defence University of Malaysia, Kuala Lumpur, Malaysia
email: hasmeda@upnm.edu.my

Abstract

Data mining techniques have recently drawn considerable attention from the research community for their ability to predict flash flood phenomena. These techniques can bring large-scale flood data into real practice and have become the necessary tools for impact assessment, societal resilience, and disaster control. Although numerous studies have been conducted on data mining techniques and flash flood predictions, domain-specific flash flood prediction models based on existing data mining techniques are still lacking. Notably, this study has focused on the performance of four data mining techniques, namely, logistic regression (LR), artificial neural networks (ANN), k-nearest neighbour (kNN), and support vector machine (SVM) in a comparative assessment as prediction models. The area under the curve (AUC) was utilised to validate these models. The value of AUC was higher than 0.9 for all models. Accordingly, the outcomes outlined in this study can contribute to

Received 14 December 2021; Accepted 1 March 2022

the current literature by boosting the performance of data mining techniques for predicting flash floods through a comparison of the most recent data mining techniques.

Keywords: *Artificial neural networks (ANN), Flash flood, k-nearest neighbor (kNN), Logistic regression (LR), Support vector machine (SVM).*

1 Introduction

Over the last few decades, the incidence of high-risk hydrological events, such as flash floods has increased exponentially [1]. Flash floods have recently been deemed the world's most catastrophic natural disaster [2]. As the name implies, a flash flood occurs in a short duration, as a result of a complex combination of meteorological and hydrological extremes, such as heavy precipitation and high floods [3]. According to Cao et al. [4], the main factors contributing to flash flood occurrences are continuous heavy rainfall, topology, and geology, as well as the impact of human activities. The result is the devastating impact they have on lives, property, infrastructure, and crops.

Malaysia, a country in South East Asia, is situated near the equator and has an equatorial climate. The equatorial climate is hot and humid throughout the year, with rainfall distribution being influenced by the northeast monsoon (November to March) and the southwest monsoon (May to September). The annual rainfall in Peninsular Malaysia is 2,500 mm, while 2,300 and 3,300 mm in Sarawak and Sabah, respectively [5]. This makes Malaysia one of the wettest countries in the world [6]. However, states in the west coast, particularly Selangor, receive higher amounts of rainfall during the southwest monsoon, which often results in flash floods [7]. Flash floods in Selangor are also clearly related to rapid urban development, including the substitution of natural surfaces with roofing and concrete [8], as depicted in Fig. 1. Consequently, green and forested areas are eliminated, and the capacity of soil to absorb rain water is reduced, which would eventually damage the surrounding areas, especially to the detriment of flora and fauna [9].

Additionally, poorly maintained structures with clogged drains and inadequate drainage, as well as poor canal design and construction, have all led to the frequent occurrence of flash floods. Flash floods may occur at any time, and can cause catastrophic loss and devastation. Bari et al. [10] estimated the losses and damages per shop caused by flash floods in the commercial area of Kajang, Selangor in 2014 to be approximately RM 4,510.07. Notably, this state has suffered severe flash floods in 2002, 2008, 2011, 2016, 2019, 2020, and recently at the end of 2021. Due to the increasing recurrence of flash floods, necessary and dependable methods are required to enable managers, engineers, and authorities to better mitigate the potential factors of flash floods in Selangor.

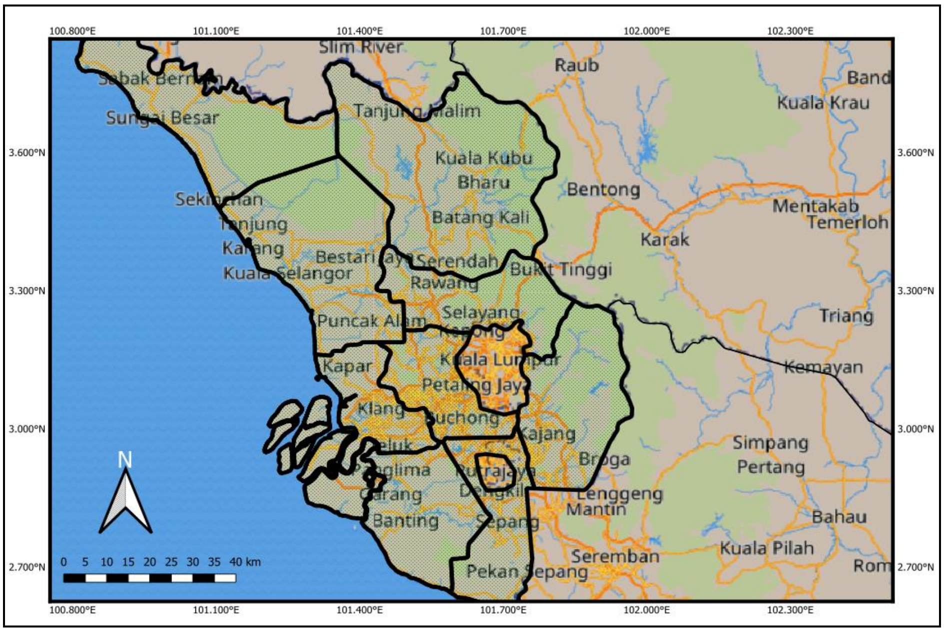

Fig. 1. Development activities in Selangor

The emergence of new data mining techniques has prompted many researchers to increase research activities in order to obtain accurate flash flood prediction values. In general, data mining techniques are grounded in machine learning techniques that can intelligently generate rules and patterns from large amounts of data. Data mining techniques were used by recent researchers to predict natural disasters, such as flood [11], flash flood [12], landslide [13], hurricane [14], earthquake [15], and tsunami [16]. Makhtar et al. [17] asserted that data mining techniques that are utilised in the field of natural disaster can aid in disaster control measures. Thus, data mining techniques are seen as a promising approach for gaining rapid access to diverse hazard assessments [18]. Therefore, this situation has become the motivation for the present study to assess the performance of data mining techniques for flash flood prediction, especially in the state of Selangor.

2 Related Work

Flash flood prediction is a challenging task due to its practical value in popular science and meteorology. Recently, several major efforts have been done to solve flash flood prediction problems using data mining techniques, with successful outcomes. Panahi et al. [19] used deep learning 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. Shirzadi et al. [20] 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 risks in flash flood-prone areas around the world. In another catchment, Janizadeh et al. [21] 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 were appropriate for mapping flash flood vulnerability in different places, thereby, able to protect people from catastrophic flooding.

Logistic regression, classification and regression trees, artificial neural network, random forest, support vector machine, and decision tree, along with a statistical method have been used for making flash flood predictions in Romania by Costache, Hong, et al. [22]. They found that hybrid models were able to achieve high performance, with prediction accuracies of more than 85%. In contrast, Costache [23] performed a comparative assessment of flash flood potential indexes, namely, frequency ratio and weights of evidence, with logistic regression and support vector machine. The results revealed that the highest accuracy can be attributed to support vector machine with weights of evidence (80.1%), followed by support vector machine with frequency ratio (79.7%), logistic regression with weights of evidence (77.2%), and logistic regression with frequency ratio (76.6%). The k-nearest neighbour, along with extreme gradient boosting, was used for flash flood prediction mapping in Egypt [24]. Although the results revealed that the extreme gradient boosting (90.2%) performed better than k-nearest neighbour (80.7%), this was only true for the data available at that time. In contrast, other studies have revealed that k-nearest neighbour often show a high accuracy value in predicting floods [5], [25], [26].

In the Malaysian context, Razali et al. [5] used Bayesian networks, decision trees, k-nearest neighbours, and support vector machines for making flood risk prediction in Kuala Krai, Kelantan. Their results showed that these techniques could produce high accuracy values of up to 99%. The artificial neural network was used by Raja Mohamad and Wan Ishak [27] to develop a prediction model for the reservoir flood stage in Timah Tasoh, Perlis. The results revealed that the predicted model has achieved more than 90% accuracy. Recently, Shaaban et al. [28] performed a comparative performance of three data mining techniques, namely, decision tree, naive Bayes, and support vector machine for making flood predictions in Kemaman, Terengganu. Their findings indicated that the performance of the decision tree was better than the other two techniques.

The present 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 [19]–[21], Romania [22], [23], and Egypt [24] for predicting flash flood occurrences. Although recently, various data mining techniques have been applied to predict floods in Malaysia [5], [27], [28], these studies were focused on the type of monsoon floods, not flash floods. To the best of our knowledge, studies that applied data mining techniques for predicting flash floods in Malaysia were conducted more than 10 years ago (see [29] and [30]). Hence, a novel finding of the current flash flood occurrence must be revealed.

The current trend of making flash flood prediction includes utilising deep learning neural networks, Bayesian belief network, decision tree, logistic regression, multilayer perceptron, quadratic discriminant analysis, classification and regression trees, artificial neural network, random forest, support vector machine, k-nearest neighbours, and naïve Bayes. Accordingly, the preferred data mining techniques have been logistic regression [21]–[23], artificial neural network [22], [27], k-nearest neighbours [5], [24]–[26], and support vector machine [5], [23], [28].

To improve the results obtained by Kia et al. [29] and Wardah et al. [30], the present study aimed to predict potential occurrences of flash floods using a comparative assessment of the results of the following data mining techniques: logistic regression (LR); artificial neural networks (ANN); k-nearest neighbours (kNN); and support vector machine (SVM). These techniques were chosen because they are new in the field of flash flood prediction in Selangor. Moreover, the performance of each technique for predicting flash floods was almost perfect, with reliable efficiency of close to 1 [5], [21]–[28]. Evaluating the performance of prediction models is a critical step in the assessment of data mining techniques. Thus, this study employed the area under the curve (AUC), since it is the most imperative indicator to validate the results and to test the performance of each model [22]. Additionally, the AUC has been a standard technique in most geo-hazard modelling studies [31].

3 Methodology

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

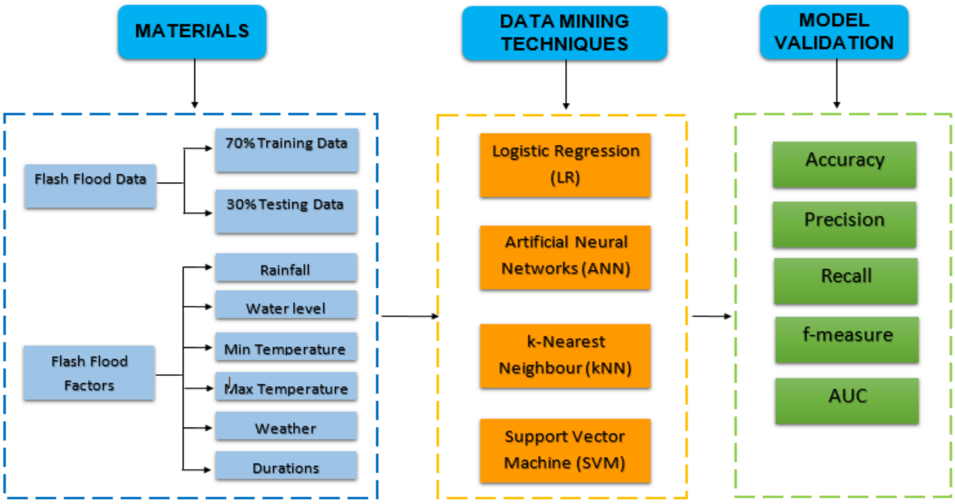

Fig. 2. Research methodology of this study

3.1 Materials

3.1.1 Study Area

This study was conducted in Selangor, which is a state in the western part of Peninsular Malaysia. The following Fig. 3 shows that Selangor shares its northern border with the state of Perak, its southern border with the state of Negeri Sembilan, and its eastern border with the state of Pahang, while its west border faces the Straits of Malacca. The study area covered 32 different locations in Selangor, which were located between the latitudes of 2° 40' 24.6" N and 3° 48' 27.0" N, and longitudes of 101° 31' 56.6" E and 101° 21' 70.0" E.

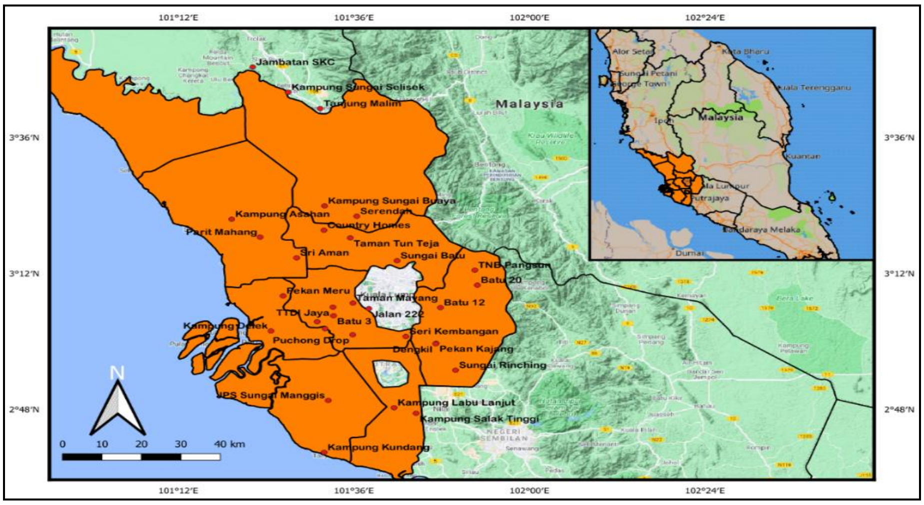

Fig. 3. A map of Selangor

3.1.2 Flash Flood Data

The data set for this study was collected from the websites of the Department of Irrigation and Drainage, Selangor (DID) and the Malaysian Meteorological Department (MMD). Water level and rainfall data were collected from the DID, while weather, and minimum and maximum temperature readings were collected from the MMD. Data for the dates were gathered during this study. Data related to different locations in Selangor were also collected, such as the area, district, main basin, and the sub-river basin. A total of 9,665 datasets were collected between June 2020 and March 2021 from 32 different locations. These datasets were then divided into a training dataset (70%, 6,765 data) and a testing dataset (30%, 2,900 data). The training dataset was used to build the data mining models, while the testing dataset was used to validate the models. The split percentage for training (70%) and testing (30%) was chosen because these percentages have been used as a standard measure in most data mining techniques and flash flood studies [11], [12], [32]. Table 1 tabulates a sample of the dataset collected in Selangor on June 3, 2020.

Table 1. A sample of dataset in Selangor

Area	Weather	*Rainfall	*Water level	Min Temp.	Max Temp.	Flash Flood
Kg. Asahan	Sunny	80.00	7.80	26.00	33.00	Yes
Sri Aman	Sunny	5.00	4.85	25.00	33.00	No
Parit Mahang	Sunny	2.00	2.56	25.00	33.00	No
Kg. Delek	Sunny	0.00	-0.59	26.00	33.00	No
Pekan Meru	Sunny	0.00	2.92	26.00	33.00	No
Taman Sri Muda	Thunder	0.00	2.33	26.00	33.00	No
Tugu Keris	Sunny	0.00	2.88	26.00	33.00	No
TTDI Jaya	Thunder	0.00	3.43	26.00	33.00	No
Batu 3	Thunder	0.00	2.48	26.00	33.00	No
Taman Mayang	Thunder	7.00	14.51	26.00	33.00	No
Puchong Drop	Thunder	0.00	5.16	26.00	33.00	No
Jalan 222	Thunder	75.00	17.03	26.00	33.00	Yes
Seri Kembangan	Thunder	0.00	35.34	26.00	33.00	No
Taman Tun Teja	Rainy	1.00	33.16	25.00	33.00	No
Sungai Batu	Sunny	17.00	49.28	25.00	33.00	No
Country Homes	Rainy	36.00	16.02	25.00	33.00	No
Serendah	Thunder	10.00	34.77	25.00	33.00	No
Jambatan SKC	Sunny	25.00	17.24	25.00	33.00	No
Tanjung Malim	Sunny	80.00	36.67	25.00	33.00	Yes
Kg. Sungai Selisek	Sunny	0.00	24.47	25.00	33.00	No
Kg. Sungai Buaya	Thunder	33.00	14.36	25.00	33.00	No
TNB Pangsun	Thunder	0.00	132.60	25.00	33.00	No

Batu 12	Sunny	0.00	40.93	25.00	33.00	No
Kg. Pasir	Sunny	0.00	47.99	25.00	33.00	No
Pekan Kajang	Thunder	0.00	22.33	25.00	33.00	No
Sungai Rinching	Thunder	0.00	20.42	25.00	33.00	No
Batu 20	Thunder	0.00	88.27	25.00	33.00	No
JPS Sungai Manggis	Rainy	0.00	0.87	25.00	33.00	No
Kg. Kundang	Thunder	0.00	1.50	25.00	33.00	No
Dengkil	Thunder	0.00	3.43	25.00	33.00	No
Kg. Labu Lanjut	Rainy	0.00	3.01	25.00	33.00	No
Kg. Salak Tinggi	Thunder	0.00	6.92	25.00	33.00	No

*Remarks: Rainfall (light: 1–10 mm; moderate: 11–30 mm; heavy: 30–60 mm; very heavy: > 60 mm); Water level (normal: < 5 m; alert: 5–6 m; warning: 6–7 m; danger: > 7 m)

3.1.3 Flash Flood Factors

The influencing flash flood factors were selected mainly based on specific study areas associated with the literature review. Prior studies have found that flash floods can primarily be determined based on four fundamental factors, namely, precipitation, topography [33], geology, and human activities [4]. Based on the selection criteria (e.g., objectivity, representativeness, and availability) and the mechanism of flash flood formation, six factors were preliminarily identified, namely, rainfall, water level, minimum and maximum temperature, weather, and durations. Table 2 lists the description of each flash flood factor employed in this study.

Table 2. 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

3.2 Data Mining Techniques

3.2.1 Logistic Regression

The probability of flash flood occurrences was constructed using the LR model. This technique was chosen because it can incorporate all 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 model 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 [34]. This model was also found to be effective for predicting the presence or absence of features based on the values of predictor variables. 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. 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) [18], [35], which is expressed in Eq. (1):

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

where *p* represents the probability of a flash flood occurrence. This probability varied from 0 to 1 in understanding that the data was "no flash flood" and "flash flood" on an S-shaped curve (sigmoid). The variable *z* represents flash flood causal factors, which were assumed as a linear combination in this study. Consequently, the LR model required Eq. (2) to be fitted 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 flash flood factors (rainfall, water level, duration, weather, and minimum and maximum temperatures). The generated linear model can then become the LR model for the presence or absence of flash flood events (present conditions) based on the independent (pre-failure conditions) variables.

3.2.2 Artificial Neural Networks

The ANN analysis in this study was trained using input data (flash flood factors) and ground truth labels (0 and 1, or no 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 [36]. 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 [33]. ANN is composed of three layers, namely, input layer, hidden layer, and an output layer that links these layers together. The net input to the hidden layer and output layer is given by Eq. (3), 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 the upper layer of node i, w_{ij} represents the weight between node i and node j, x_i represents the output value from node j, while w_{i0} represents the bias in node i, and it also represents the input signal of node i, which is then passed through a transfer function [33].

3.2.3 k-Nearest Neighbours

The kNN technique uses the k most similar neighbours to calculate the prediction of flash flood occurrences. The number of similar observations that produces the best prediction, or k, will be determined. If this value is too high, the kNN model will overgeneralise; if the value is too small, it will lead to a large variation in the prediction [25]. The selection of k was performed by evaluating different values of k within a range and selecting the value that produced the "best" prediction. To assess the different values of k, the sum of squared error (SSE) evaluation criteria [24], as shown in Eq. (4), can be used:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

where SSE values that are less than one show that the predictions are accurate. After establishing a value for k during the training phase, the model can be utilised to make flash flood predictions.

3.2.4 Support Vector Machine

The SVM model can derive intrinsic rules from an enormous number of complex input and output variables [37]. A training dataset of known sample data, $T = \{x_1, x_2, \dots, x_n, y\}$, was considered with x_i as the ith input (rainfall, water level, minimum temperature, maximum temperature, weather, durations), $(x_i \in R_n)$, y as the output, and i = 1, 2, …, n. Then, to achieve the maximum interval, these data were separated into two categories using an n-dimensional hyperplane. Thus, the calculation steps for the algorithm are as given in Eq. (5) and Eq. (6), as follows:

$$\frac{1}{2} \|w\|^2 \tag{5}$$

$$y_i\big((w \,.\, x_i) + b\big) \geq 1 \tag{6}$$

where ||w|| is the coefficient vector that defines the orientation of the hyperplane normal, b is the offset of the hyperplane from the origin, and (·) denotes the scalar product operation. Once the optimal hyperplane has been determined, the following optimisation problem can be solved using Lagrangian multipliers, as shown in Eq. (7):

$$L = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i\big(y_i((w \,.\, x_i) + b) - 1\big) \tag{7}$$

where λ_i is the Lagrangian multiplier [13]. Standard approaches can be used to solve Eq. (7) by utilising dual minimisation, with respect to w and b. A separating hyperplane can be defined, as shown in Eq. (8), in the case of linear separable data:

$$y_i\big((w \,.\, x_i) + b\big) \geq 1 - \xi_i \tag{8}$$

and Eq. (8) becomes Eq. (9):

$$L = \frac{1}{2}\|w\|^2 - \frac{1}{vn}\sum_{i=1}^{n} \xi_i \tag{9}$$

where $v \in \{0,1\}$ is introduced as misclassification. The kernel function must be chosen carefully in SVM modelling. The linear kernel function (LN), polynomial kernel function (PL), radial basis function (RBF), and sigmoid kernel function (SIG) are the most commonly used kernel types for SVM analysis [38]. In this study, the RBF, $K(x_i, x_j)$, was selected to perform the SVM analysis, as shown in Eq. (10):

$$K\,(x_i, x) = e^{(-\gamma\|x_i - x\|^2)}, \gamma > 0 \tag{10}$$

where γ denotes the kernel function's parameter. Next, $\gamma = \frac{1}{2}\sigma^2$, where σ is an adjustable parameter that governs the kernel's performance, is sometimes used to parameterise kernel functions [39]. RBF can produce efficient interpolation, but it may have some flaws when it comes to longer-range extrapolation [40].

3.3 Model validation

The performance of the LR, ANN, kNN, and SVM models were evaluated using the most commonly used statistical metrics, namely, accuracy, precision, recall, and f-measure [41], [42]. In this study, accuracy referred to the overall accuracy of these models, precision referred to the probability that the model will predict flash flood occurrence, recall referred to the probability that the model can detect flash flood occurrence from the total number of occurrences, and f-measure represented the harmonic mean of precision and recall. The formulas for calculating these metrics are given in Eqs. (11)–(14):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

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

$$f - measure = \frac{2 \times recall \times precision}{recall + precision} \tag{14}$$

TP, TN, FP, and FN denoted true positive, true negative, false positive, and false negative, respectively. This study has also used the 10-fold cross-validation, as it is the standard for predicting the error rate of a data mining technique when a single, or fixed number of data is given [42]. The receiver operating characteristic (ROC) is one of the most imperative evaluation measures for determining the capability of data mining models. The area under the curve (AUC) is a ROC performance metrics. Hence, this study utilised the ROC to analyse the overall capability of the flash flood prediction models, while the metric values of AUC were used to validate the model performance. Generally, an AUC of 0.5–0.6 indicates a weak performance, and an AUC of 0.6–0.7 implies a poor performance. A classifier with an AUC of 0.7 to 0.8 demonstrates a modest level of performance and an AUC value of greater than 0.8 shows that the developed model is well-suited for the given dataset [32]. The AUC value can be calculated using the following Eq. (15):

$$AUC = \frac{(\sum TP + \sum TN)}{P+N} \tag{15}$$

where TP and TN are the numbers of pixels that are correctly classified, P is the total number of pixels with flash flood phenomena, and N is the total number of pixels without flash flood phenomena.

4 Results, Analysis and Discussions

4.1 Results and Analysis

As previously stated, this study utilised 9,665 datasets that have been separated into training and testing datasets at 70% and 30% of the total dataset, respectively. The first set was used to construct models, while the second set was used to validate models [32]. Then, the LR, ANN, kNN, and SVM models were assessed based on the training and the testing datasets. The performance of these models was evaluated using four statistical metrics, namely, accuracy, precision, recall, and f-measure. The performance results are as shown in Table 3. Based on this table, the performance of the training dataset using the kNN model exhibited the highest values of 0.999 for all statistical metrics (accuracy, precision, recall, and f-measure). Meanwhile, the performances of the training dataset using the other models were quite similar with the performance using kNN, with LR = 0.998, ANN = 0.997, and SVM = 0.998 for all metrics values of accuracy, precision, recall, and f-measure. With the testing dataset, the kNN model has also produced the best accuracy (0.997), precision (0.997), recall (0.997), and f-measure (0.997), in comparison with the metrics values obtained using the other three models.

Table 3. Evaluation performance of the LR, ANN, kNN, and SVM models in predicting flash flood occurrences in Selangor

Models	Sample	TP	TN	FP	FN	Accuracy	Precision	Recall	f-measure
LR	Training	106	6641	4	14	0.998	0.998	0.998	0.998
	Testing	42	2848	1	9	0.997	0.996	0.997	0.996
ANN	Training	99	6641	4	21	0.997	0.997	0.997	0.997
	Testing	38	2849	0	13	0.996	0.996	0.996	0.995
kNN	Training	115	6641	4	5	0.999	0.999	0.999	0.999
	Testing	43	2848	1	8	0.997	0.997	0.997	0.997
SVM	Training	110	6641	4	10	0.998	0.998	0.998	0.998
	Testing	40	2848	1	11	0.996	0.996	0.996	0.996

The results of the ROC curve and AUC for the flash flood prediction models are illustrated in Fig. 4 (training dataset) and Fig. 5 (testing dataset). The ROC curve analysis using the training dataset showed that the kNN model received the highest value of AUC (0.987), followed by SVM (0.986), ANN (0.983), and LR (0.981). Meanwhile, the ROC curve analysis using the testing dataset showed that the SVM and kNN models received the highest values of AUC at 0.971 and 0.961, respectively, followed by ANN (0.959), and LR (0.946).

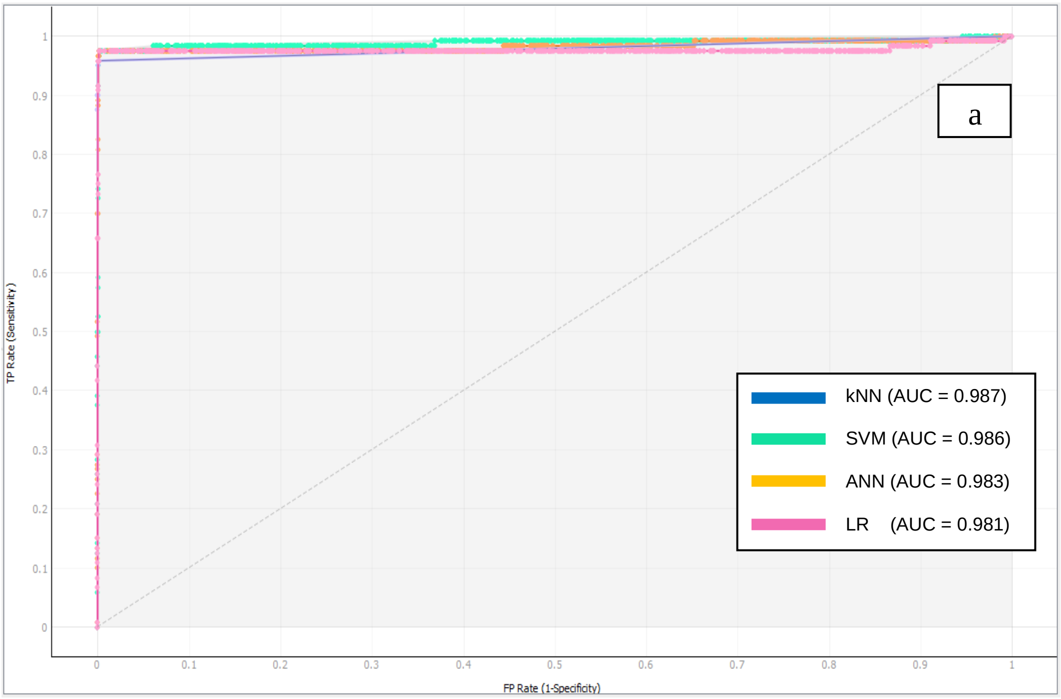

Fig. 4. ROC curve and AUC of the LR, ANN, kNN, and SVM models using the training dataset

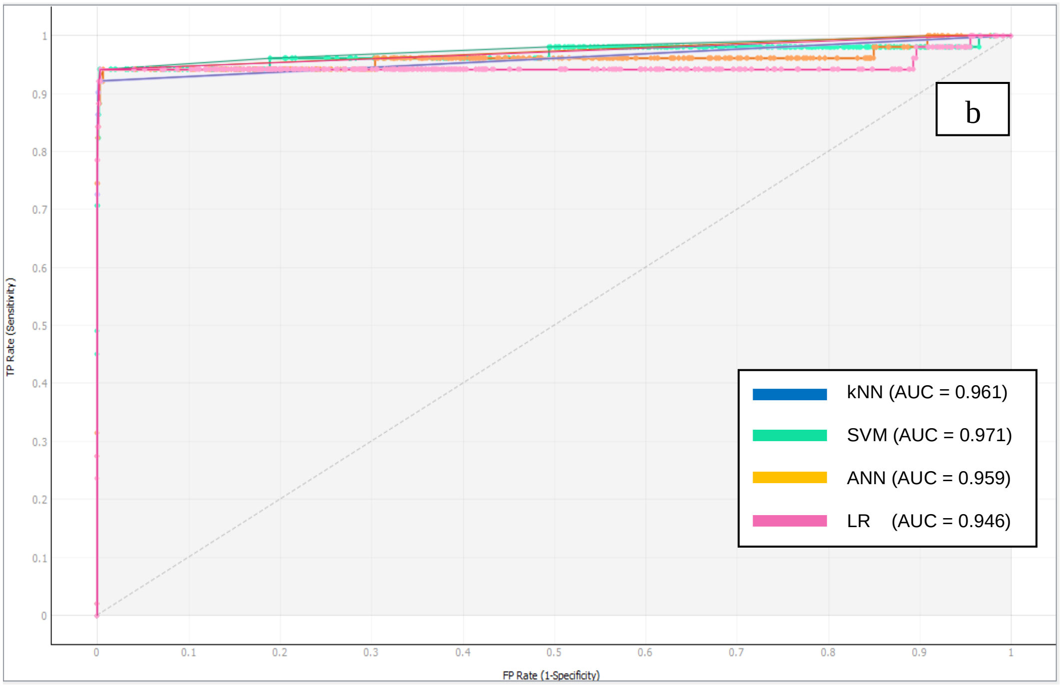

Fig. 5. ROC curve and AUC of the LR, ANN, kNN, and SVM models using the testing dataset

4.2 Discussion

Flash floods are impacted by a multitude of factors, and their occurrence can never be predicted entirely. As a result, it is vital to choose appropriate assessment techniques, strengthen the prediction model, and increase the accuracy of assessment outcomes. Numerous ways of measuring the occurrence of flash floods have been developed by researchers worldwide, and each of these models has distinct advantages and disadvantages. For example, the absence of appropriate screening processes for flood factors and the subsequent model construction are both relatively complex and require professional knowledge. Cao et al. [4] stated that the utilised model should be straightforward and highly comprehensible. Hence, four commonly used data mining models, with six appropriate factors, were used for predicting flash flood occurrences in Selangor, Malaysia. The data mining models in this study were LR, ANN, kNN, and SVM, while the factors involved were rainfall, water level, minimum and maximum temperatures, weather, and duration.

As previously mentioned, Table 3 shows the performance comparison results between LR, ANN, kNN, and SVM models. These results showed that kNN has the most accurate values for predicting flash floods, with 0.999 and 0.997 accuracy for the training and testing datasets, respectively. Meanwhile, kNN is one of the simplest techniques used mostly for classification and regressions. It is a predictive model that does not require complicated mathematical equations and can be described as a technique for a non-parametric, supervised learning and pattern classifier [24]. Subsequently, the AUC values were greater than 0.9 for all LR, ANN, kNN and SVM models. These results were corroborated by the findings by Bui et al. [32] and Janizadeh et al. [21], who had also obtained AUC values of more than 0.9 when predicting flash flood occurrences. Overall, the present study showed that the kNN was the best flash flood prediction model compared to the other models (LR, ANN, and SVM).

5 Conclusion

Flash flood prediction modelling is a critical task that needs to be undertaken in the study area of the Selangor state. This is because this state has repeatedly experienced flash floods in recent years. The present study has shown that the proposed LR, ANN, kNN, and SVM models were applicable for predicting flash flood occurrences in this state. All models have shown great performances based on the validation results, with accuracy of more than 90%. Hence, the results of this study may be beneficial for the local government agencies and decision-makers in relation to this disaster. Specifically, the authorities, or policy makers could utilise this knowledge to alleviate the devastating impacts of flash floods before they occur, particularly in Selangor. In the future, other advanced data mining techniques will be examined, such as deep learning, boosted regression tree, and random forest for making flash flood prediction and their performance will be compared with the four techniques in this study. Additionally, the effect of flash flood factors on the performance of these techniques will also be evaluated.

ACKNOWLEDGEMENTS
This study was supported by the Ministry of Higher Education Malaysia (MOHE) under the RACER/1/2019/ICT04/UPNM/1 grant. The authors fully acknowledge the National Defence University of Malaysia (NDUM) and MOHE for the approved fund, which made this research viable and effective.

References

[1] Costache, R. (2019). Flash-Flood Potential assessment in the upper and middle sector of Prahova river catchment (Romania). A comparative approach between four hybrid models. *Science of the Total Environment*, 659, 1115-1134.

[2] Islam, M. M., Ujiie, K., Noguchi, R., & Ahamed, T. (2022). Flash flood-induced vulnerability and need assessment of wetlands using remote sensing, GIS, and econometric models. *Remote Sensing Applications: Society and Environment, 25*(August 2021), 100692.

[3] Malek, N. S. A., Zayid, S., Ahmad, Z., Ya'Acob, S., & Bakar, N. A. A. (2020). Data understanding for flash flood prediction in urban areas. *Journal of Environmental Treatment Techniques*, *8*(2), 770-778.

[4] 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), 1-22.

[5] Razali, N., Ismail, S., & Mustapha, A. (2020). Machine learning approach for flood risks prediction. *IAES International Journal of Artificial Intelligence, 9*(1), 73-80.

[6] Khalid, M. S., & Shafiai, S. (2015). Flood disaster management in Malaysia: An evaluation of the effectiveness flood delivery system. *International Journal of Social Science and Humanity, 5*(4), 398-402.

[7] Samsuri, N., Bakar, R. A., & Unjah, 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.

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

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

[10] 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*(3), 1-9.

[11] El-Haddad, B. A., Youssef, A. M., Pourghasemi, H. R., Pradhan, B., El-Shater, A. H., & El-Khashab, M. H. (2021). Flood susceptibility prediction using four machine learning techniques and comparison of their performance at Wadi Qena Basin, Egypt. *Natural Hazards, 105*(1), 83-114.

[12] Abu El-Magd, S. A. (2022). Random forest and naïve Bayes approaches as tools for flash flood hazard susceptibility prediction, South Ras El-Zait, Gulf of Suez Coast, Egypt. *Arabian Journal of Geosciences, 15*(3), 1-12.

[13] Nhu, V. H., Zandi, D., Shahabi, H., Chapi, K., Shirzadi, A., Al-Ansari, N., Singh, S. K., Dou, J., & Nguyen, H. (2020). Comparison of support vector machine, bayesian logistic regression, and alternating decision tree algorithms for shallow landslide susceptibility mapping along a mountainous road in the west of Iran. *Applied Sciences, 10*(15), 1-27.

[14] D'Amico, D. F., Quiring, S. M., Maderia, C. M., & McRoberts, D. B. (2019). Improving the hurricane outage prediction model by including tree species. *Climate Risk Management*, *25*(October 2018), 100193.

[15] Yousefzadeh, M., Hosseini, S. A., & Farnaghi, M. (2021). Spatiotemporally explicit earthquake prediction using deep neural network. *Soil Dynamics and Earthquake Engineering, 144*(August 2020), 106663.

[16] Yao, Y., Yang, X., Lai, S. H., & Chin, R. J. (2021). Predicting tsunami-like solitary wave run-up over fringing reefs using the multi-layer perceptron neural network. *Natural Hazards, 107*(1), 601-616.

[17] 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 *Advances in Intelligent Systems and Computing* (pp. 437-446). Springer International Publishing.

[18] Li, X., Yan, D., Wang, K., Weng, B., Qin, T., & Liu, S. (2019). Flood risk assessment of global watersheds based on multiple machine learning models. *Water, 11*(8), 1-18.

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

[20] Shirzadi, A., Asadi, S., Shahabi, H., Ronoud, S., Clague, J. J., Khosravi, K., Pham, B. T., Ahmad, B. Bin, & 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.

[21] Janizadeh, S., Avand, M., Jaafari, A., Van Phong, T., Bayat, M., Ahmadisharaf, E., Prakash, I., Pham, B. T., & Lee, S. (2019). Prediction success of machine learning methods for flash flood susceptibility mapping in the Tafresh watershed, Iran. *Sustainability, 11*(19), 1-19.

[22] Costache, R., Hong, H., & Pham, Q. B. (2020). Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. *Science of the Total Environment, 711*, 134514.

[23] Costache, R. (2019). Flash-Flood Potential assessment in the upper and middle sector of Prahova river catchment (Romania). A comparative approach between four hybrid models. *Science of the Total Environment, 659*, 1115-1134.

[24] El-Magd, S. A. A., Pradhan, B., & Alamri, A. (2021). Machine learning algorithm for flash flood prediction mapping in Wadi El-Laqeita and surroundings, Central Eastern Desert, Egypt. *Arabian Journal of Geosciences, 14*(4), 1-14.

[25] Costache, R., Pham, Q. B., Sharifi, E., Linh, N. T. T., Abba, S. I., Vojtek, M., Vojteková, J., Nhi, P. T. T., & Khoi, D. N. (2020). Flash-flood susceptibility assessment using multi-criteria decision making and machine learning supported by remote sensing and GIS techniques. *Remote Sensing*, *12*(106), 1-26.

[26] Snehil, & Goel, R. (2020). Flood damage analysis using machine learning techniques. In *International Conference on Smart Sustainable Intelligent Computing and Applications. ICITETM2020* (pp. 78-85). Science Direct.

[27] Raja Mohamad, R. N. M., & Wan Ishak, W. H. (2019). Forecasting the flood stage of a reservoir based on the changes in upstream rainfall pattern. *Journal of Technology and Operations Management*, *14*(2), 46-52.

[28] Shaaban, N. N., Hassan, N., Mustapha, A., & Mostafa, S. A. (2021). Comparative Performance of Supervised Learning Algorithms for Flood Prediction in Kemaman, Terengganu. *Journal of Computer Science, 17*(5), 451-458.

[29] Wardah, T., Abu Bakar, S. H., 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.

[30] 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*(1), 251-264.

[31] Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., & Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of Environmental Management, 217*, 1-11.

[32] Bui, D. T., Hoang, N. D., Martínez-Álvarez, F., Ngo, P. T. T., Hoa, P. V., Pham, T. D., Samui, P., & Costache, R. (2020). A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area. *Science of the Total Environment, 701*, 134413.

[33] 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.

[34] 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), 1-16.

[35] Tehrany, M. S., Shabani, F., Neamah Jebur, M., Hong, H., Chen, W., & Xie, X. (2017). GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques. *Geomatics, Natural Hazards and Risk, 8*(2), 1538-1561.

[36] Dtissibe, F. Y., Ari, A. A. A., Titouna, C., Thiare, O., & Gueroui, A. M. (2020). Flood forecasting based on an artificial neural network scheme. *Natural Hazards, 104*(2), 1211-1237.

[37] Xiong, J., Li, J., Cheng, W., Wang, N., & Guo, L. (2019). A GIS-based support vector machine model for flash flood vulnerability assessment and mapping in China. *ISPRS International Journal of Geo-Information, 8*(7), 1-23.

[38] He, Y., Ma, D., Xiong, J., Cheng, W., Jia, H., Wang, N., Guo, L., Duan, Y., Liu, J., & Yang, G. (2021). Flash flood vulnerability assessment of roads in China based on support vector machine. Geocarto International, 0(0), 1-25.

[39] Ma, M., Liu, C., Zhao, G., Xie, H., Jia, P., Wang, D., Wang, H., & Hong, Y. (2019). Flash flood risk analysis based on machine learning techniques in the Yunnan Province, China. *Remote Sensing, 11*(170), 1-16.

[40] Tehrany, M. S., Jones, S., & Shabani, F. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena, 175*(December 2018), 174-192.

[41] Kantardzic, M. (2020). *Data Mining: Concepts, Models, Methods, and Algorithms* (3rd Ed.). John Wiley & Sons.

[42] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th Ed.). Morgan Kauffman.

Notes on contributors

Muhammad Hakiem Halim received a Bachelor of Computer Science (Computer Security) in 2019 from Malaysia's National Defence University (NDUM). He then spent a year as a research assistant at the NDUM. He is currently pursuing a Master of Computer Science degree at NDUM. His research interests include data mining techniques, computer security, and data analysis across a broad range of domains, most notably flash flood phenomena.

Muslihah Wook received her PhD in Information Science from Universiti Kebangsaan Malaysia in 2017, her Master of Computer Science from Universiti Putra Malaysia in 2004, and her Bachelor of Information Technology (Hons) from Universiti Utara Malaysia in 2001. Her research interests include big data analytics and data mining applications in various domains, particularly in education, security and defence. She is currently working as a Senior Lecturer in the Department of Computer Science, Faculty of Defence Science and Technology, National Defence University of Malaysia (NDUM). She has become a member of the International Association of Computer Science and Information Technology (IACSIT) and the Institute of Research Engineers and Doctors (IRED) in 2011 and 2013, respectively. She has been appointed as a technical reviewer of Education and Information Technologies and Journal of Big Data—Springer journals indexed by WoS and Scopus (Q1), and other outstanding journals as well.

Nor Asiakin Hasbullah received her PhD in information technology and quantitative sciences from MARA University of Technology Malaysia in 2017, and she completed a three-month attachment at Glasgow Caledonian University in Glasgow in 2012 for privacy research.She holds a Master of Science in Information Technology from MARA University of Technology Malaysia in 2006, and a

Bachelor of Information Technology (hons) from Universiti Kebangsaan in 2003. Currently, she is a senior lecturer in the Department of Computer Science at the National Defence University of Malaysia (NDUM). Her research interests are in the field of privacy, data protection, information security and ethics in ICT. She is a member of the Malaysian Board of Technologists and Informatics Intelligence Special Interest Group, NDUM. She has published and presented most of her research findings and articles at various international conferences and international journals.

Noor Afiza Mat Razali holds a Bachelor's Degree in Computer and Information Engineering, Master of Science in Computer Science and PhD of Science in Computer Science from Japanese Universities. Afiza is an Associate Professor at the Faculty of Defence Science and Technology at the National Defence University of Malaysia (NDUM). Afiza's research and expertise are in the areas of cybersecurity, disaster management systems, big data analytics, human-computer interaction, artificial intelligence, robotics, and blockchain technology. Afiza is also a professional techno-gist, a recognition given by the Malaysian Board of Technologists and appointed as the expert reviewer for the Malaysian Board of Technologists and the Ministry of Science, Technology and Innovation (MOSTI) Grand Challenge.

Hasmeda Erna Che Hamid received her Master of Information Technology from the International Islamic University of Malaysia (IIUM) in 2014 and her Bachelor of Computer Science (Hons) from the University of Technology Malaysia (UTM) in 2009. Prior to joining the National Defence University of Malaysia (NDUM), she was a Webmaster at Digital Media Broadcasting Services Sdn Bhd, an IT Officer at the Malaysian Rubber Board (LGM), an Assoc. Software Engineer at XYBASE MSC Sdn Bhd, and an Analyst Programmer at Systematic Conglomerate Sdn Bhd. Currently, Hasmeda is a doctoral candidate at the NDUM and has conducted research on disaster management using machine learning techniques. Her research interests are artificial intelligence, data management, and software engineering.