ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 6 No. 1 (2026) 613-621
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Assessing Flood Factors through Comparative Analysis of Machine Learning Techniques

Ong Jun Xian[1], Azme Khamis[1*]

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM1, Jalan Panchor, 84600 Muar, Johor, MALAYSIA.*

*Corresponding Author: azme@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2026.06.01.067

Article Info

Received: 4 January 2026
Accepted: 6 June 2026
Available online: 9 July 2026

Keywords

Flood, Machine Learning, K Nearest Neighbors, Random Forest, Logistic Regression, Rain, Season

Abstract

Floods are among the most frequent natural disasters, leading to loss of life, spreading of diseases, and destruction of property. Therefore, accurate and timely flood prediction is essential for effective disaster risk management. This study explores the application of Machine Learning (ML) techniques to improve flood forecasting systems in Malaysia using geological and hydrological features. The purposes of the study are to identify the factors influencing flood occurrence and their significance using Random Forest (RF), analyse hidden trends or patterns related to flood occurrence using Logistic Regression (LR) and K Nearest Neighbors, and evaluates the performance of the machine learning models proposed using evaluation metrics during the modelling process. A dataset consisting of 4217 samples was compiled for 42 study areas at 42 weather stations across Malaysia, incorporating 16 explanatory variables, including seven-day continuous rainfall characteristics, seasonal and geographical factors, and other relevant environmental attributes. Three classification models, RF, LR and KNN were trained to predict flood occurrence. The results indicate that recent rainfall (0.1444), cumulative (0.1088) and maximum daily rainfall (0.1058) are the most influential features. Among the models, LR performs better at classifying flood events with higher recall value (0.7218), while RF shows best performance in all other metrics. Overall, the best model identified is RF based on Receiver Operating Characteristic (ROC) curve (AUC=0.877).

1. Introduction

Floods are among the most frequent natural disasters affecting communities around the world [1]. Flooding is the most frequent natural disaster in Malaysia, causing pollution, interrupts social and economic activity, damages roads and infrastructures, lowers property values, increases vulnerability, and claims lives. Malaysia experiences both northeast and southwest monsoons, both bring heavy rainfall and contribute to floods, particularly in coastal areas [2]. As flood remains the most frequent natural disaster in Malaysia, causing billions of ringgits in losses each year, there is a pressing need to understand the factors that contribute to the phenomenon. This study explores the application of machine learning (ML) techniques to investigate flood related factors and improve flood forecasting systems in Malaysia, aiming to support early warning systems and disaster management strategies.

This study aims to identify the factors influencing flood occurrence and their significance using Random Forest (RF). Also, this study aims to analyse hidden trends or patterns related to flood occurrence using Logistic

Regression (LR) and K Nearest Neighbors (KNN). Finally, this study also evaluates the performance of the machine learning models proposed using evaluation metrics during the modelling process.

Previous studies such as [3] conducted a detailed analysis of monsoon season of 2022 at Pakistan's Swat River basin, revealing how unusually high rainfall, exceeding historical data by 7 to 8% triggered destructive debris flows and floods. Another research by [4] shows results indicating that over half of the studied catchments are influenced by a combination (52.1%) of recent precipitation, antecedent precipitation (excessive soil moisture), and snowmelt. [5] revealed that the features with the highest relative importance score was elevation at 0.39, followed by rainfall. In another research by [6], the results show that RF and Extreme Gradient Boosting performed best at identifying flood events, with Support Vector Machine (SVM) the lowest. Besides that, [7] conducted a study on floods prediction in flood-prone Bihar and Orissa, India using RF and Decision Tree (DT), where their results show DT outperforming RF in all metrics.

This study is divided into four sections. Section one provides the introduction, background, problem statement, objectives and significance of study while providing previous related research. Section two lists out the methodology, including research design, research data, methods and algorithms used to achieve the research objectives. Section three shows the outcomes and results based on the objectives. Last section summarizes the findings and proposes recommendations for future research.

2. Methodology

This section consists of three parts with Part one shows the data used for the study. Part two shows the algorithms used and their purpose and finally Part three explains the evaluation metrics to compare the models.

2.1 Data

Table 1 describes the dataset used in this study which is based on various data on features related to flood occurrence. The data used to build the dataset is retrieved from several sources [8][9][10][11]. The data collected from different sites are combined into one dataset. The dataset consists of 17 variables including the responding variable. Among the 16 predictor variables, 6 are geographical features on elevation, location and tree cover, while the remaining 10 variables are seasonal and weather features including a 7-day continuous rainfall and the cumulative and maximum daily rainfall. The season feature will be classified into 4 categories, namely Northwest Monsoon, 1st Inter-monsoon, Southwest Monsoon, and 2nd inter-monsoon as defined by Meteorological Department of Malaysia. The responding variable, flood occurrence, is a binary variable noting '1' for flood and '0' for no flood.

Table 1 Data Description

Variables	Meaning	Data Type
location	Location of weather station/study	Categorical
avr_elevation	Average elevation	Numerical
min_elevation	Minimum elevation	Numerical
max_elevation	Maximum elevation	Numerical
forest_cover	Natural Forest cover percentage	Numerical
tree_cover(cd>30%)	Tree cover percentage with canopy density >30%	Numerical
season	Season	Categorical
day1	Rainfall on 7 days from flood	Numerical
day2	Rainfall on 6 days from flood	Numerical
day3	Rainfall on 5 days from flood	Numerical
day4	Rainfall on 4 days from flood	Numerical
day5	Rainfall on 3 days from flood	Numerical
day6	Rainfall on 2 days from flood	Numerical
day7	Rainfall on 1 day from flood	Numerical
cuml_7rain	Cumulative rainfall of all 7 days	Numerical
max_rainfall	Maximum daily rainfall	Numerical
y	Flood Occurence	Categorical

2.2 Data Preprocessing

The initial dataset requires several pre-processing steps to ensure suitability for the machine learning models. Categorical variables 'y', 'season', and 'location' are transformed into numerical formats suitable for algorithms. The data was split into training and testing sets, with training data consisting of 80% of the total data and testing set consisting of the remaining 20%. To overcome imbalance in data, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training dataset. SMOTE is a statistical technique for increasing the number of cases in the dataset in a balanced way, by generating new instances from existing minority cases as input, without changing the number of majority cases [12]. Balancing the class distribution is to prevent the proposed models from being biased towards the majority non-flood class. The numerical features in the training data were standardized for LR and KNN models. Standardization was performed on the quantitative variables to ensure that all predictors contributed equally to the regression model as distance and gradient based models are most affected by range of features [13].

2.3 Algorithms

2.3.1 Random Forest

Random Forest (RF) is a robust machine learning algorithm that builds multiple decision trees using random subsets of the dataset and features during training [14]. This randomness introduces diversity among the trees, helping to reduce overfitting and enhance prediction accuracy.

When performing RF classification, the decision on each tree nodes is based on the Gini Index with Equation (1) showing the formula [15].

$$Gini = 1 - \sum_{i=1}^{c} \left(p_i\right)^2 \tag{1}$$

where p_i represents the relative frequency of the class and c represents the number of classes.

During the training process, the algorithm calculates how much each feature contributes to reducing the Gini impurity across all the decision trees in the forest.

2.3.2 K Nearest Neighbors

K-Nearest Neighbors (KNN) algorithm is a supervised, non-parametric learning method that classifies or predicts the grouping of a data point based on its proximity to neighboring points [16]. It works by identifying the "k" nearest data points in the training set to a new, unlabelled data point and then using these nearest neighbors to make a prediction. KNN uses distance metrics to identify nearest neighbor for classification tasks. The most used distance metric is the Euclidean distance, with Equation (2) showing the formula [17].

$$d\left(x, X_i\right) = \sqrt{\sum_{j=1}^{n} \left(x_j - X_{ij}\right)^2} \tag{2}$$

where $x = (x_1, x_2, \ldots x_n)$ is the query point, $Xi = (X_{i1}, X_{i2}, \ldots X_{in})$ is the i-th point in the dataset, n is the dimension, and $d(x, X_i)$ is the distance between points x and X_i .

To visually interpret how KNN operates within the feature space, a 3D visualization was generated focusing on the three most important features identified in the overall study.

2.3.3 Logistic Regression

Logistic regression (LR) is a supervised ML algorithm that classifies output in the form of discrete or categorical variables [18]. It estimates the probability that a given input belongs to a certain class using the sigmoid (logistic) function shown in Equation (3) [19].

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + exp(-z)} \tag{3}$$

where σ(x) is the sigmoid function, exp is the exponential function, and e is the Euler's number. LR solves classification task by learning from a training set, a vector of weights w_i and a bias term b. Each weight w_i is a real number and is associated with the input features xi. The resulting single number z expresses the weighted sum of the evidence for the class as shown in Equation (4)

$$z = w_1 x_1 + w_2 x_2 + \ldots + w_i x_i + b \tag{4}$$

Applying the sigmoid function, this gives a value 0 or 1, which is interpreted as probability shown in Equation (5).

$$P(y=1|x)=\hat{y}=\frac{1}{1+e^{w^T x+b}} \tag{5}$$

Classification rule is shown in Equation (6).

$$\hat{y}=\begin{cases}1, P(y=1|x)\geq c\\ 0, otherwise\end{cases} \tag{6}$$

where c is the decision boundary. universally set initially at 0.5.

A key advantage of the Logistic Regression model is the interpretability of its coefficients. Each coefficient represents the expected change in the mean of the transformed response given that the predictor changes by 1 unit on the coded scale. The coefficients corresponding specifically to the daily rainfall variables (day 1 to day 7) were visualized using a bar chart.

2.3.4 Evaluation Metrics

Evaluation metric plays a critical role in achieving the optimal classifier during the classification training [20]. The evaluation metrics used in this study are accuracy, precision, recall, specificity and F1-score. The calculation of the metrics are shown in Equations (7), (8), (9), (10) and (11) respectively [21].

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

$$Precision=\frac{TP}{TP+FP} \tag{8}$$

$$Recall/Sensitivity=\frac{TP}{TP+FN} \tag{9}$$

$$Specificity=\frac{TN}{TN+FP} \tag{10}$$

$$F1\ \text{Score}=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{11}$$

Here, TP represents True Positives, FP represents False Positives, TN represents True Negatives and FN represents False Negatives. Besides the five metrics, a Receiver Operating Characteristic (ROC) Curve is plotted. ROC is a visual representation of the trade-offs between the true positive rate (TPR) and false positive rate (FPR) at various thresholds [22]. Area Under the Curve (AUC) value gives a single scalar value ranging from 0 to 1 that shows the performance of the model. The closer the AUC value is to 1, the better the model.

3. Results

This part shows the results obtained from the study. Part one shows the feature importance evaluation result by RF. Part two and three explains the rainfall patterns, coefficients and feature space. Part three summarizes the comparative analysis of the three models based on evaluation metrics.

3.1 Feature Importance Evaluation

Table 2 shows the feature importance output of the RF model. Day 7 rain (rainfall on or one day before flood occurrence) and maximum daily rainfall dominate the model, with importance 0.1444 and 0.1088 each. Maximum daily rainfall is the third most important feature with importance 0.1058. Day 6 rain (rainfall on a day or two days prior to rainfall) comes close at fourth place with 0.0778. Season comes fifth at 0.062632. The rest of the seasonal and weather features in descending order are day 3 (0.0576), day 5 (0.0562), day 4 (0.0436), day 1 (0.0389), with day 2 (0.0368) coming last. Days 1 and 2 (rainfall 6-7 days prior of flood) came out second last and last in weather features and 13th and 15th overall. This shows that most floods are influenced by recent rainfall.

Among geographical features, tree cover percentage of location with canopy density over 30% is the highest and 7th highest overall with 0.0568. Elevation features like average (0.0510), maximum (0.0463) and minimum elevation (0.0429) show moderate but non-negligible influence on the model. The remaining features, natural forest cover percentage (0.0376) and location of study (0.0328) came third last and last among feature importance, showing insignificant influence.

On the whole, based on the analysis, rainfall variables are more influential than geographical variables in terms of flood prediction model building as they provide direct information on the most influential flood factor which is rain.

Table 2 Feature Importance Output

Feature	Meaning	Importance
day7	Rainfall on 1 day before flood	0.1444
cuml_7rain	Cumulative rainfall of all 7 days	0.1088
max_rainfall	Maximum daily rainfall	0.1058
day6	Rainfall on 2 days before flood	0.0780
season	Season	0.0626
day3	Rainfall on 4 days before flood	0.0576
tree_cover(cd>30%)	Tree cover percentage with canopy density >30%	0.0568
day5	Rainfall on 3 days before flood	0.0562
avr_elevation	Average elevation	0.0510
max_elevation	Maximum elevation	0.0463
day4	Rainfall on 4 days before flood	0.0436
min_elevation	Minimum elevation	0.0429
day1	Rainfall on 7 days before flood	0.0389
forest_cover	Natural forest cover percentage	0.0376
day2	Rainfall on 6 days before flood	0.0368
location	Location of weather station/study	0.0328

3.1.1 Rainfall Patterns

Fig. 1 shows the LR coefficients on all rainfall variables. Day 7 has the strongest positive effect (2.1 ± 0.1). This shows that the model's prediction relies heavily on this variable. This suggests that the flood is highly responsive to immediate precipitation, meaning buildup matters, but the final-day rainfall is dominant. Day 6 has a strong negative effect (–1.2 ± 0.1), this may be due to multicollinearity, where day 6 rainfall is statistically correlated with Day 7 rainfall, causing LR to assign opposite signs to balance the shared effect. This negative coefficient does not mean rainfall reduces flood risk. Rainfall on day 1 and 2 have very small positive values, showing slight, almost negligible effect, mostly noise. Day 4 has a significant negative value (-0.9 ± 0.1), which is a show of slightly opposite effect. Day 5 has a significant positive coefficient (0.9 ± 0.1), showing modest contribution to the model.

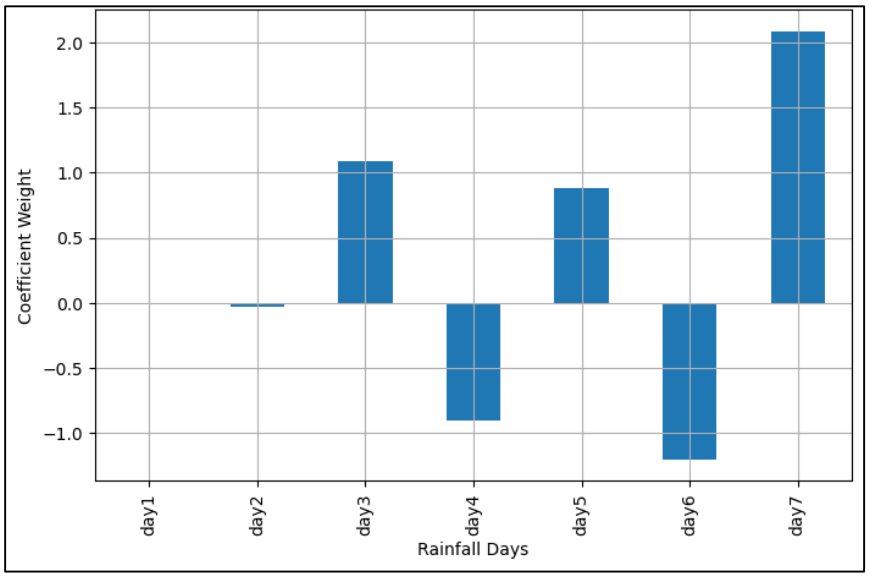


Fig. 1 Logistic Regression coefficients on rainfall variables

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 613-621

Overall, LR coefficients show that rainfall far from the occurrence of floods contributes weakly to the model. Flood risk is controlled mostly by rainy days prior to flood.

3.2 Visualization of Feature Space

Fig. 2 shows a KNN 3D plot visualizing a flood classification model based on the three most important features based on RF feature importance evaluation. The three features are maximum daily rainfall, cumulative 7-day rainfall, and day 7 rainfall. In the plot, each data point is classified in red (1/flood) or blue (0/no flood). The plot illustrates how data points naturally cluster in this 3D space.

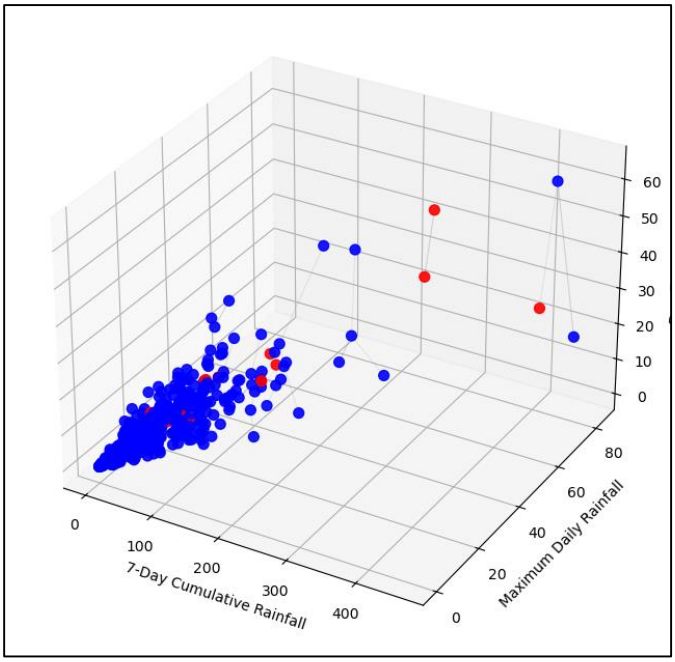
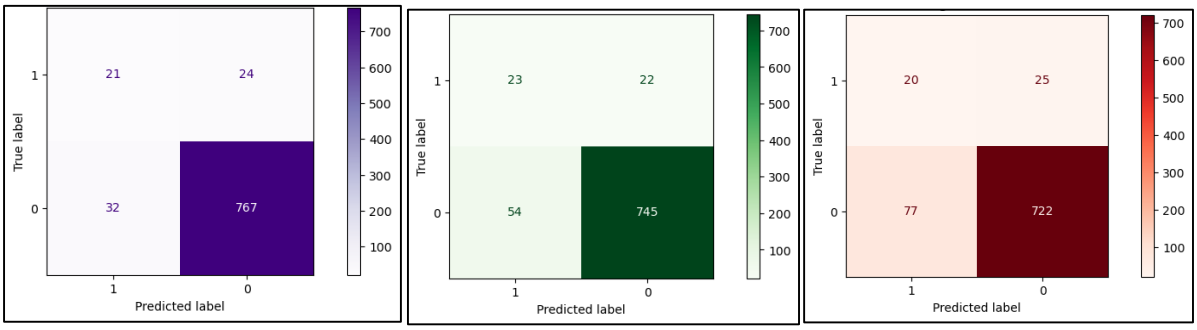


Fig. 2 K Nearest Neighbors Intepretation 3D Model

Fig. 2 clearly demonstrates that flood risk is directly proportional to higher values across all three rainfall metrics. The no flood (blue) region primarily occupies the space corresponding to low-to-moderate values on all axes, while the flood (red) region clusters in the area of high rainfall values. This indicates a strong separability between the two classes. Flood events are most strongly associated with high values on these axes, suggesting that a significant volume of recent rainfall (day 7) combined with a high peak intensity (maximum daily rainfall), especially when the soil is already saturated (high cumulative 7-day rainfall), is the primary driver of flood occurrence. The figure visually shows the imbalance in data and shows that while most events cluster in low-rainfall areas, flood events typically occupy distinct regions characterized by higher cumulative rainfall and maximum daily intensity. The grey lines in the visualization connect each data point to its single nearest neighbor. This figure is key to interpreting the model's mechanism. A new data point falling into the high-rainfall red region would be classified as a flood event, proving that high rainfall accumulation is a direct predictor of flooding.

3.3 Confusion Matrices

Fig. 3 shows the confusion matrices of the three models applied in this study. From the matrices, LR has the highest TP value (23) while RF has the highest TN value (767). RF performed better at lower FP value (32) while LR with the lower FN value (22). KNN ranks the lowest among three models based on the 4 metrics.



(a) (b) (c)

Fig. 3 Confusion Matrices for (a) Random Forest, (b) Logistic Regression; and (c) K Nearest Neighbors Model Performance Evaluation

Table 3 Model Performance Metrics

Algorithm	Accuracy	Precision	Recall	Specificity	F1-Score
KNN	0.8791	0.9260	0.6740	0.9036	0.8992
RF	**0.9336**	**0.9391**	0.7133	**0.9599**	**0.9362**
LR	0.9100	0.9355	**0.7218**	0.9324	0.9208

Table 3 shows the outputs of evaluation metrics for RF, LR and KNN where RF outperform other models in accuracy (0.9336), precision (0.9391), specificity (0.9599), and F1-score (0.9362). LR shows best performance in recall metric (0.7218), showing that the model slightly edges out RF in predicting real flood events, with RF more accurate in classifying non-flood events. KNN is the least performer of the three models in all metrics. Since both RF and LR have similar metrics and both show selected advantages over each other, the final decision is based on the AUC value from the ROC curve shown in Fig. 4.

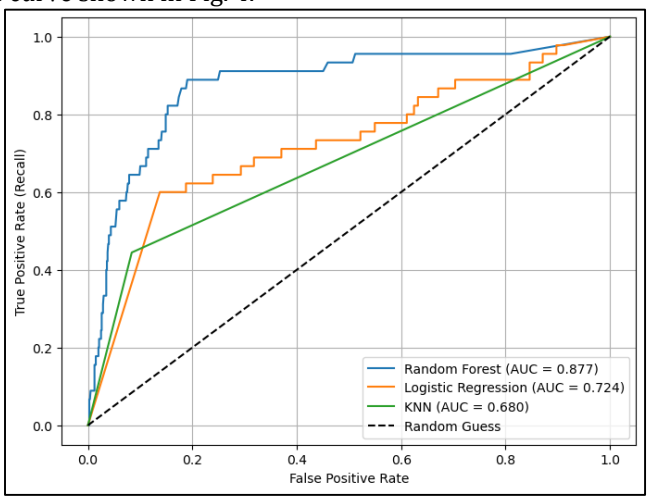


Fig. 4 ROC curves and their AUC scores

RF completely outperforms the other models with AUC=0.877 followed by LR (0.724) and KNN (0.680). KNN is the least-performing model with AUC very close to 0.5 which is the score for a random guess. This suggests the KNN model has a very limited ability to distinguish between flood and non-flood cases, no matter the classification threshold. The visual gap between the blue (RF) and orange (LR) lines is large, corresponding to the difference in their F1-score.

4. Conclusion

Overall, the most influential features in flood prediction are cumulative, most recent and maximum daily rainfall. LR coefficients show that recent rainfall contributes most to the prediction ability of the model. The better model for predicting flood occurrence based on geographical, seasonal and weather features is the RF classifier. Although LR has a higher recall, the value merely edges out RF, while showing relative weakness in other metrics.

Future work in this study should focus on incorporating more predictive features and adopting more advanced algorithms. The study revealed that short-term rainfall accumulation and sudden spikes increase in

rainfall intensity tend to cause floods rather than long term rainfalls. Thus, the initial important recommendation is the increase of storm-related dynamics and detailed geographical information on study location. Suggested features are rainfall intensity index, surface of study area covered in water like rivers or ponds, and slope. Other statistical characteristics like standard deviation and skewness of rainfall can be considered because they represent irregularity and spike-like rainfalls, which are usually accompanied by floods.

This study revealed that simple linear models and distance-based classifiers are less effective. In this regard, models like XGBoost and LightGBM ought to be tried since they are exceptionally good at dealing with non-linearities and interactions between features. It is also possible that SVM and NN will provide better performance, particularly with better features. Although KNN will always be a good tool of investigating similarities in rainfalls, it is not the best concluding predictive model considering the overlapping flood and non-flood characteristics.

Acknowledgement

The author would like to thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Ong Jun Xian, Azme Khamis;* ***data collection:*** *Ong Jun Xian;* ***analysis and interpretation of results:*** *Ong Jun Xian, Azme Khamis;* ***draft manuscript preparation:*** *Ong Jun Xian, Azme Khamis. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Our World in Data. (2024). Number of natural disaster events. https://ourworldindata.org/grapher/number-of-natural-disaster-events

[2] Department of Irrigation and Drainage Malaysia. (2017). Flood management. https://www.water.gov.my/index.php/pages/view/419?mid=244

[3] Bazai, N. A., Alam, M., Cui, P., Hao, W., Khan, A. P., Waseem, M., Shunyu, Y., Ramzan, M., Wanhong, L., & Ahmed, T. (2025). Dynamics and impacts of monsoon-induced geological hazards: a 2022 flood study along the Swat River in Pakistan. *Natural Hazards and Earth System Sciences*, 25(3), 1071– 1093. https://doi.org/10.5194/nhess-25-1071-2025

[4] Motta, M., De Castro Neto, M., & Sarmento, P. (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. *International Journal of Disaster Risk Reduction*, 56, 102154. https://doi.org/10.1016/j.ijdrr.2021.102154

[5] Khaltri, S., Kokane, P., Kumar, V., & Pawar, S. (2022). Prediction of waterlogged zones under heavy rainfall conditions using machine learning and GIS tools: a case study of Mumbai. GeoJournal, 88(S1), 277–291. https://doi.org/10.1007/s10708-022-10731-3

[6] Mdegela, L., Municio, E., De Bock, Y., Luhanga, E., Leo, J., & Mannens, E. (2023). Extreme rainfall event classification using machine learning for Kikuletwa River floods. Water, 15(6), 1021. https://doi.org/10.3390/w15061021

[7] Kunverji, K., Shah, K., & Shah, N. (2021). A flood prediction system developed using various machine learning algorithms. *In Proceedings of the 4th International Conference on Advances in Science & Technology* (ICAST2021).

[8] Government of Malaysia. (2023). Kluang, Johor - Weather and climate | Data.gov.my. data.gov.my. https://data.gov.my/dashboard/weather-and-climate/kluang

[9] Malaysian Meteorological Department. (2025). *Weather phenomena*. https://www.met.gov.my/en/pendidikan/fenomena-cuaca/

[10] Topographic Map. (2025). Kluang. Topographic Map. https://en-gb.topographicmap.com/map-74mlt6/Kluang/?center=1.67545%2C103.86249&zoom=7

[11] Global Forest Watch. (2025). Forest Monitoring, Land Use & Deforestation Trends | Global Forest Watch. https://www.globalforestwatch.org/

[12] Microsoft. (2025). *SMOTE component*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote?view=azureml-api-2

[13] Vinay, S. (2021). Standardization in machine learning. *LinkedIn. Recuperado de https://www. linkedin. com/pulse/standardization-machine-learning-sachin-vinay*.

[14] Ren, L., & Cao, S. (2022). Application of feature selection based on elastic network and random forest in the evaluation of sports effects. *Journal of Electrical and Computer Engineering*, 2022, 1–9. https://doi.org/10.1155/2022/2794104
[15] Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A., & Alzahrani, A. K. (2022). Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *Big Data and Cognitive Computing*, 6(1), 13. https://doi.org/10.3390/bdcc6010013
[16] Ukey, N., Yang, Z., Li, B., Zhang, G., Hu, Y., & Zhang, W. (2023). Survey on Exact kNN Queries over High-Dimensional Data Space. *Sensors*, 23(2), 629. https://doi.org/10.3390/s23020629
[17] GeeksforGeeks. (2025). KNeArest Neighbor(KNN) algorithm. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/
[18] Lee, F. F. (2025). *What is logistic regression?* IBM. https://www.ibm.com/think/topics/logistic-regression
[19] Jurafsky, D., & Martin, J. H. (2025). Speech and language processing (Draft of January 12, 2025). https://web.stanford.edu/~jurafsky/slp3/
[20] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.
[21] Vujovic, Ž. Đ. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6). https://doi.org/10.14569/ijacsa.2021.0120670
[22] Chugh, V. (2024). *AUC and the ROC curve in machine learning*. DataCamp. https://www.datacamp.com/tutorial/auc