Enhancing precision flood mapping: Pahang's vulnerability unveiled

Tahmina A. K*[1], Siventhiran S B[2 ¶] , Maheswaran S [3 &,] Saravana Selvan[4 &], Sreeramanan S[5] , Low J An [6 &], Leela A[7] , Prahankumar R[8 &], Lokeshmaran A[9] , Boratne AV[10] , Abdullah, M. T[11 ¶]

Abstract

Flooding in Malaysia is considered one of the most impactful natural disasters. Annually, Pahang experiences substantial destruction due to floods. The aim of this research is to address the urgent issue of flood susceptibility in Pahang, Malaysia. To achieve this, a combination of Geographic Information System (GIS) and Ensemble Machine Learning (EML) will be utilized. By considering nine factors from a geospatial database that contribute to flooding, the areas prone to floods will be mapped. The mapping process will be carried out using the ArcGIS environment, and a model called Random Forest (RF)-embedding will be developed using the Ensemble Machine Learning (EML) technique. To determine the most influential factors in flooding, Feature Selection (FS) will be employed. The accuracy of the flood susceptibility models will be assessed by analyzing the Area Under the Curve (AUC). Flood susceptibility mapping is a complex procedure with uncertainties, but our research can contribute to flood management in vulnerable regions by improving flood models and providing spatial outcomes to help decision-makers implement risk reduction strategies.

KEYWAODS : Flood susceptibility; vulnerability; Geographic Information System; Ensemble Machine Learning; Pahang.

Introduction

WHO defines a disaster as a significant event that exceeds the capacity of local resources, necessitating external assistance[1] . Floods are a frequently occurring natural disaster that can lead to widespread illness and loss of life on a global scale. The extent of damage to communities is influenced by factors such as geography, population density, and infrastructure [2]. The occurrence of hydrometeorological disasters such as floods, heavy rainfall, and tropical storms has been linked to climate change[3]. In Malaysia, particularly in its eastern region, the annual monsoon floods pose significant dangers to individuals residing near riverbanks [4,5] .

Malaysia is characterized by a tropical climate, featuring high temperatures, humidity, and abundant rainfall. The country experiences distinct monsoon seasons that impact both the peninsular and insular regions. The Northeast Monsoon, occurring from November to March, brings heavy rainfall and rough seas, often leading to flooding in the eastern part of the peninsula. On the other hand, the Southwest Monsoon dominates from May to September, primarily affecting the southwestern coastal areas of Sabah. The First and Second Inter-monsoonal Periods fall in between these two main monsoon seasons. Malaysia receives significant annual rainfall, with Maxwell's Hill recording the highest amount at 5,000 mm [6–10] .

Pahang, the third largest state in Malaysia, is located in the basin of the Pahang River, extending from the east coast to Endau. Positioned in the East Coast region of Malaysia, it spans an area of 35,965 km2. Pahang showcases a tropical landscape and encounters an equatorial climate marked by elevated humidity all year round. The coastal area of Pahang in Peninsular Malaysia is significantly affected by strong

northerly winds and intense rainfall during the early and mid-monsoon period, resulting in severe floods from November to January. In contrast to other regions in Peninsular Malaysia, the east-coast section of Pahang receives higher annual precipitation levels with more pronounced variations. Pahang has witnessed significant flood occurrences that inflicted substantial damage on the state, commencing with the catastrophic flood event in 1926, which was documented as the most severe flood incident in history. The year 2014 witnessed one of the most devastating flood events in Pahang and the entire East Coast of Peninsular Malaysia. The major flood event in Pahang in 2014 caused extensive damage, particularly to the local economy and alterations in the hydrological system. Notably, Pahang encountered major floods in 1971 and 2021, resulting in widespread destruction and influencing both the local economy and hydrological system [6,11,12] .

The utilization of GIS-based machine learning algorithms to construct models for forecasting natural calamities in particular areas has witnessed a surge. Ensemble Machine Learning (EML) amalgamates numerous classifiers to enhance accuracy, resulting in more precise predictions compared to the utilization of a single classifier. The incorporation of ensemble models strives to diminish prediction errors and enhance overall accuracy. [13,14] .

Hence, the overarching goal of this research is to develop a comprehensive flood susceptibility mapping framework for the Pahang State, Malaysia using a Geographic Information System (GIS) and Ensemble Machine Learning (EML) approach. This framework aims to enhance disaster preparedness, inform land use planning, and mitigate the impacts of recurrent floods on communities and infrastructure in the region.

Conceptual framework

This conceptual framework integrates principles from disaster management, resilience, geographic information systems (GIS), and machine learning to analyze flood susceptibility and identify vulnerable areas.

Analytical Tools

GIS is essential for spatial data analysis and decision-making, particularly in flood susceptibility mapping. It integrates geospatial data to examine spatial relationships and visualize vulnerable areas. Machine Learning, specifically ensemble methods like Random Forest, provide advanced techniques for analyzing complex datasets and improving the accuracy of flood susceptibility predictions.

Steps:

1. *Develop an Integrated GIS-Based Framework*. The objective is to establish a robust GIS-based framework for flood susceptibility mapping in the Pahang State. This involves compiling and integrating geospatial datasets related to topography, hydrology, land use, and climate variables to create a comprehensive database for analysis.

2. *Apply Ensemble Machine Learning Algorithms*. The objective is to apply ensemble machine learning algorithms, such as Random Forest (RF) and Gradient Boosting Machines (GBM), to the integrated dataset to develop predictive models of flood susceptibility. This objective includes feature selection, model training, validation, and evaluation to ensure the accuracy and reliability of the susceptibility maps.

3. *Generate Actionable Insights for Decision-Making*. The objective is to generate actionable insights from the flood susceptibility maps to support informed decision-making and disaster management strategies. This involves identifying vulnerable areas, assessing the factors contributing to flood risk, and recommending targeted interventions and mitigation measures to reduce the impacts of floods on communities, infrastructure,

and the environment.

By achieving these objectives, the research aims to contribute to the development of evidence-based policies and practices for flood risk reduction and resilience building in the Pahang State, ultimately enhancing the region's capacity to adapt to and mitigate the impacts of recurrent floods.

Expected Result

The main goal of this research is to understand the susceptibility to floods in a specific area and develop targeted approaches to reduce their impact. The Pahang region in Malaysia has experienced frequent floods, causing significant challenges to the community, infrastructure, and economy. Despite efforts to mitigate flood risks, there is a need for accurate and comprehensive flood vulnerability mapping. Conventional techniques often lack precision and fail to consider environmental factors and the dynamic nature of climate change and land use patterns. This research aims to address these limitations and inform decision-making and disaster management strategies. The research is expected to forecast that the combination of these flood influencing factors will collectively serve as the main determinants of flood susceptibility in Pahang. It is anticipated that the Ensemble Machine Learning (EML) method, particularly the Random Forest-embedding model, will effectively capture the intricate relationships among these factors to generate precise flood susceptibility forecasts. Moreover, the study will posit that the identified susceptibility categories, spanning from very low to very high, will offer a dependable depiction of the diverse levels of flood vulnerability in different areas of Pahang. The validation procedure, which involves Area Under the Curve (AUC) analysis, is expected to provide a robust evaluation of the model's predictive accuracy, thereby bolstering confidence in the resultant flood susceptibility map for Pahang..

Materials and methods

Study Area

Pahang in Peninsular Malaysia has been chosen as the research site due to its annual monsoon floods, which harm the local population.

Study design and Data Collection Tool

Flood influencing factors.

According to the data available for Pahang and a comprehensive literature search, a total of nine factors have been identified as potential indicators of heightened flood susceptibility in the context of modelling studies. These factors encompass elevation, slope, curvature, flow direction, flow accumulation, distance from river, rainfall, land-use, and geology. Together, these parameters effectively capture the topographical and hydrometeorological conditions that contribute to the overall vulnerability of the region to flooding events[15,16] .

Digital Elevation Models (DEMs) have demonstrated their indispensable role in ensuring the precision of hydrodynamic models [17] . The Earth data platform provided access to the 30 m resolution Shuttle Radar Topography Mission (SRTM) DEM Version 3, from which the digital elevation data will be obtained [18] .

The presence of flooding is largely impacted by the slope of the land, as steeper slopes can accelerate the flow of water over the surface, hindering its ability to seep into the ground [19] . The shape of a surface, as determined by its curvature, indicates whether it is convex, concave, or flat, indicating changes in slope inclination. Concave surfaces tend to collect flood water, increasing the likelihood of flooding [20] . The direction of

flow plays a crucial role in determining the path that surface water will take and the potential for flooding [21]

An increase in flow accumulation coincides with an increase in vulnerability to flooding [19] . In this research, the distance from rivers was estimated using the Euclidean distance tool in ArcGIS software, which utilized a raster layer depicting the river network. The ArcGIS platform will be used to generate maps for the elevation, slope, curvature, flow direction, flow accumulation, and distance from river, which will be subsequently categorized into sub-classes using the natural break classification method.

Flooding occurs when there is a sudden increase in water levels in rivers, lakes, and reservoirs due to intense rainfall, often resulting in inadequate drainage [22] . We will be using data from 10 precipitation stations in Pahang, including Cameron Highlands, Bentong, Bera, Kuantan, Lipis, Maran, Pekan, Raub, Rompin, and Temerloh, to create a rainfall distribution map for the research area. We will employ the Inverse Distance Weighted (IDW) approach, utilizing a 10-year dataset from 2012 to 2021, to construct the map [23] . This method ensured that the rainfall patterns in the area being studied were accurately depicted.

The properties of drainage systems are significantly affected by changes in land use and land cover (LULC) in the upstream watersheds. These modifications directly impact the occurrence of surface overflow and the land surface's capacity to absorb water, ultimately playing a role in the frequency and intensity of flooding events [24] . The global geological and LULC data will be obtained from the worldwide geological maps database provided by the USGS and the Global data [25] . The LULC map will be created using the ArcGIS platform, delineating seven distinct categories: water bodies, trees, flooded vegetation, crops, built area, bare terrain, and rangeland. In the case of

the Geology map of Pahang, and will be segmented into nine primary soil features, based on the USGS-USA soil taxonomy [25,27]

Random forest (RF) Embedding classifier.

The random forest technique demonstrates strong predictive accuracy and is adept at managing large datasets for regression and classification purposes. By training numerous decision trees concurrently through bootstrapping, aggregation, and bagging methods, the RF method consistently outperforms alternative techniques in accuracy and prevents overfitting. Moreover, the training process for the RF-embedding model is quicker, leading to superior classification accuracy [28].

Feature Selection

Feature selection is crucial for improving model efficiency, eliminating unnecessary data, preventing overfitting, and enhancing generalization on test data. In this study, an embedded feature selection method using a shuffling algorithm was used to create random probes based on the original variables. These probes were combined with the variables to train a Random Forest regression model, which determined the significance of each variable (Z-score). Variables with a Z-score higher than the maximum Z-score among the random probes were considered important [29] . In this context, the DML algorithm uses the embedded Mean Decrease Accuracy (MDA) measure. It typically splits based on "gini" for Gini impurity and "entropy" for information gain, mathematically defined as p(xi) for each possible value i of random variable x and c for the number of classes in the dataset (Eq 1,2) [30–32].

$$Entropy: \ H(x) = -\sum_{i=1}^{n} p(x_i)\ log_2\ p(x_i) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

$$Gini(E) = 1 - \sum_{i=1}^{C} pi^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

The RF learning model, using multiple decision trees, is more accurate than a single decision tree. It combines random feature selection and bagging for classification and regression. In this study, a popular machine learning FS method ranked flood influencing factors. This algorithm is widely endorsed by researchers for its strong predictive performance, high accuracy, and ease of interpretation. It iteratively generates rankings by shuffling features and identifying consistently important ones [29,33] .

Ethical declaration

Our research study has been granted ethical approval by both the Research Management Centre (RMC) Committee [Application Ref No: AIMST/RMC/AUHAEC/FRGS/25022022/01, Date: 25/02/2022] and the AIMST University Human Ethics Committee (AUHEC) [Application Ref No: AUHEC/FOM/22/09/2023/, Date: 22/09/2023].

References

[1] UNDRR. Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction. Source United Nations Office for Disaster Risk Reduction United Nations General Assembly 2017. https://www.undrr.org/publication/report-open-ended-intergovernmental-expert-working-group-indicators-and-terminology (accessed April 5, 2024).

[2] Du W, Fitzgerald GJ, Clark M, Hou XY. Health impacts of floods. Prehosp Disaster Med 2010;25:265–72. https://doi.org/10.1017/S1049023X00008141.

[3] Sun F, Lai X, Shen J, Nie L, Gao X. Initial allocation of flood drainage rights based on a PSR model and entropy-based matter-element theory in the Sunan Canal, China. PLoS One 2020;15:e0233570. https://doi.org/10.1371/journal.pone.0233570.

[4] Khan MMA, Shaari N, Nahar A, Baten MdA, Nazaruddin DA. Flood impact assessment in Kota Bharu, Malaysia: a statistical analysis 2014.

[5] Nurul Ashikin A, Nor Diana MI, Siwar C, Alam MM, Yasar M. Community preparation and vulnerability indices for floods in Pahang State of Malaysia. Land (Basel) 2021;10:1–23. https://doi.org/10.3390/land10020198.

[6] DARUL MAKMUR. Portal Rasmi Kerajaan Negeri Pahang 2024. https://www.pahang.gov.my/ (accessed April 5, 2024).

[7] ASM - Academy of Sciences Malaysia. Assessment on the Sustainability of the Tasik Chini Basin and Tasik Chini Biosphere Reserve - Official Portal Academy of Sciences Malaysia 2023. https://www.akademisains.gov.my/asm-publication/assessment-on-the-sustainability-of-the-tasik-chini-basin-and-tasik-chini-biosphere-reserve/ (accessed April 22, 2024).

[8] Saimi FM, Hamzah FM, Toriman ME, Jaafar O, Tajudin H. Trend and Linearity Analysis of Meteorological Parameters in Peninsular Malaysia. Sustainability 2020;12:9533. https://doi.org/10.3390/su12229533.

[9] Wong C, Liew J, Yusop Z, Ismail T, Venneker R, Uhlenbrook S. Rainfall Characteristics and Regionalization in Peninsular Malaysia Based on a High Resolution Gridded Data Set. Water (Basel) 2016;8:500. https://doi.org/10.3390/w8110500.

[10] Britannica. Pahang | History, Nature & Industry 2024. https://www.britannica.com/place/Pahang (accessed April 5, 2024).

[11] Muhammad NS, Abdullah J, Julien PY. Characteristics of Rainfall in Peninsular Malaysia. J Phys Conf Ser 2020;1529:052014. https://doi.org/10.1088/1742-6596/1529/5/052014.

[12] Kamarudin MKA, Toriman ME, Abd Wahab N, Abu Samah MA, Abdul Maulud KN, Mohamad Hamzah F, et al. Hydrological and climate impacts on river characteristics of pahang river basin, Malaysia. Heliyon 2023;9. https://doi.org/10.1016/j.heliyon.2023.e21573.

[13] Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble Learning for Disease Prediction: A Review. Healthcare (Switzerland) 2023;11. https://doi.org/10.3390/healthcare11121808.

[14] Shirzadi A, Soliamani K, Habibnejhad M, Kavian A, Chapi K, Shahabi H, et al. Novel GIS based machine learning algorithms for shallow landslide susceptibility mapping. Sensors (Switzerland) 2018;18. https://doi.org/10.3390/s18113777.

[15] Khoirunisa N, Ku C-Y, Liu C-Y, Esteban D, López-Gutiérrez J-S, Negro V, et al. A GIS-Based Artificial Neural Network Model for Flood Susceptibility Assessment. International Journal of Environmental Research and Public Health 2021, Vol 18, Page 1072 2021;18:1072. https://doi.org/10.3390/IJERPH18031072.

[16] Nguyen VN, Yariyan P, Amiri M, Tran AD, Pham TD, Do MP, et al. A New Modeling Approach for Spatial Prediction of Flash Flood with Biogeography Optimized CHAID Tree Ensemble and Remote Sensing Data. Remote Sensing 2020, Vol 12, Page 1373 2020;12:1373. https://doi.org/10.3390/RS12091373.

[17] Xu K, Fang J, Fang Y, Sun Q, Wu C, Liu M. The Importance of Digital Elevation Model Selection in Flood Simulation and a Proposed Method to Reduce DEM Errors: A Case Study in Shanghai. International Journal of Disaster Risk Science 2021;12:890–902. https://doi.org/10.1007/s13753-021-00377-z.

[18] Earthdata. Earthdata Search Search. NASA USAGov 2023. https://search.earthdata.nasa.gov/search (accessed July 6, 2023).

[19] Chaulagain D, Ram Rimal P, Ngando SN, Nsafon BEK, Suh D, Huh JS. Flood susceptibility mapping of Kathmandu metropolitan city using GIS-based multi-criteria decision analysis. Ecol Indic 2023;154:110653. https://doi.org/10.1016/j.ecolind.2023.110653.

[20] Ramesh V, Iqbal SS. Urban flood susceptibility zonation mapping using evidential belief function, frequency ratio and fuzzy gamma operator models in GIS: a case study of Greater Mumbai, Maharashtra, India. Geocarto Int 2022;37:581–606. https://doi.org/10.1080/10106049.2020.1730448.

[21] Towfiqul Islam ARM, Talukdar S, Mahato S, Kundu S, Eibek KU, Pham QB, et al. Flood susceptibility modelling using advanced ensemble machine learning models. Geoscience Frontiers 2021;12:101075. https://doi.org/10.1016/j.gsf.2020.09.006.

[22] Liu X, Zhou P, Lin Y, Sun S, Zhang H, Xu W, et al. Influencing Factors and Risk Assessment of Precipitation-Induced Flooding in Zhengzhou, China, Based on Random Forest and XGBoost Algorithms. International Journal of Environmental Research and Public Health 2022, Vol 19, Page 16544 2022;19:16544. https://doi.org/10.3390/IJERPH192416544.

[23] Paul Stackhouse. NASA POWER | DAVe. Esri, TomTom, Garmin, FAO, NOAA, USGS, EPA, USFWS 2024. https://power.larc.nasa.gov/data-access-viewer/ (accessed June 6, 2024).

[24] Sugianto S, Deli A, Miswar E, Rusdi M, Irham M. The Effect of Land Use and Land Cover Changes on Flood Occurrence in Teunom Watershed, Aceh Jaya. Land 2022, Vol 11, Page 1271 2022;11:1271. https://doi.org/10.3390/LAND11081271.

[25] USGS. EarthExplorer. Science for a Changing World 2024. https://earthexplorer.usgs.gov/ (accessed April 20, 2024).

[26] USGS. U.S. Geological Survey. The Water Cycle 2022. https://www.usgs.gov/special-topics/water-science-school/science/water-cycle (accessed April 13, 2024).

[27] Steinshouer DW, Qiang J, McCabe PJ, Ryder RT. Maps showing geology, oil and gas fields, and geologic provinces of the Asia Pacific region. 1999. https://doi.org/10.3133/ofr97470f.

[28] Netzer M, Baumgartner C, Baumgarten D. Predicting prediction: A systematic workflow to analyze factors affecting the classification performance in genomic biomarker discovery. PLoS One 2022;17. https://doi.org/10.1371/journal.pone.0276607.

[29] Chen Y, Ma L, Yu D, Zhang H, Feng K, Wang X, et al. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. Ecol Indic 2022;135:108545. https://doi.org/10.1016/j.ecolind.2022.108545.

[30] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70–9. https://doi.org/10.1016/j.neucom.2017.11.077.

[31] Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics 2022;2:927312. https://doi.org/10.3389/fbinf.2022.927312.

[32] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2021;2:1–21. https://doi.org/10.1007/s42979-021-00592-x.

[33] Masrur Ahmed AA, Deo RC, Feng Q, Ghahramani A, Raj N, Yin Z, et al. Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. J Hydrol (Amst) 2021;599:126350. https://doi.org/10.1016/j.jhydrol.2021.126350.