https://jurnal.unimed.ac.id/2012/index.php/geo/article/view/72252
https://doi.org/10.24114/jg.v18i1.72252 - (Vol 18, No 1 – 2026)
available at http://jurnal.unimed.ac.id/2012/index.php/geo
e-issn (2549-7057) & p-issn (2085-8167)

# Urban Flood Susceptibility Modeling Using GIS and Machine Learning in Bandar Lampung

**Alvin Pratama*, Andreas Boni Baik Simamora, Farras Ghaly**
Department of Atmospheric and Planetary Science, Faculty of Science, Institut Teknologi Sumatera, Indonesia

**ARTICLE INFO**
Article History:
Received: Januari 20, 2026
Revision: March 12, 2026
Accepted: April 08, 2026

Keywords:
Urban Flood;
Flood Susceptibility;
Geographic Information System;
Machine Learning;
Bandar Lampung

Corresponding Author
E-mail:
alvin.pratama@sap.itera.ac.id

**ABSTRACT**
Urban flooding increasingly affects rapidly urbanizing tropical cities, where terrain, rainfall, and anthropogenic surface modification interact to shape spatial flood patterns. This study develops a GIS–machine learning framework to model urban flood susceptibility in Bandar Lampung, Indonesia, using a multi-year flood inventory (2015–2024). A balanced dataset (n = 308; 1:1 flood to pseudo-absence ratio) was constructed using buffered pseudo-absence sampling with spatial separation constraints to reduce bias. Nine environmental and infrastructure-related predictors were evaluated using Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM). Model performance was assessed through five-fold stratified cross-validation, generalization gap analysis (Train AUC − CV AUC), learning curves, and a 20% hold-out test set. GB achieved the highest cross-validation performance (CV AUC = 0.8953), followed by RF (0.8782), SVM (0.8007), and LR (0.6925). However, ensemble models exhibited larger generalization gaps (RF = 0.1218; GB = 0.1047) compared to LR (0.0333), indicating stronger overfitting tendencies. Learning curves confirmed that LR maintained the most stable convergence between training and validation scores. On the independent test set (n = 61), GB achieved the highest predictive accuracy (ROC AUC = 0.9462), whereas LR showed lower discriminative performance (AUC = 0.7065) but greater validation stability. Flood susceptibility was concentrated in low-elevation areas, near major roads, and adjacent to river networks. By integrating learning curve diagnostics with cross-validation and hold-out testing, this study provides a rigorous framework for model selection in data-limited urban environments.

## INTRODUCTION

Urban flooding has become a growing environmental concern in many cities worldwide, particularly in regions experiencing rapid urban development and climate change. Changes in land use, expansion of impervious surfaces, and the intensification of extreme rainfall events have altered hydrological responses in urban landscapes, leading to more frequent and spatially complex flood occurrences (Lee et al., 2017; Zhang et al., 2024). From a geographical perspective, urban flooding reflects the interaction between physical landscape characteristics and human-induced spatial transformations, making it a critical subject for integrated spatial analysis and urban environmental studies.

In tropical regions, the impacts of urban flooding are often amplified by high rainfall intensity, short-duration storms, and dense settlement patterns. Southeast Asia has experienced a marked increase in flood-related losses over recent decades, driven by both climatic and anthropogenic factors (Nguyen et al., 2023). Indonesia, in particular, faces persistent urban flooding problems across cities of different sizes, where rapid urbanization has outpaced the development of drainage infrastructure and

the capacity for spatial planning. Flood events in Indonesian cities are not only hydrometeorological phenomena but also manifestations of spatial vulnerability shaped by topography, land cover change, and infrastructure distribution (Putri et al., 2021; Rahayu et al., 2023).

Bandar Lampung City, located in southern Sumatra, is a medium-sized urban area that exemplifies these challenges. The city is characterized by heterogeneous terrain, ranging from coastal plains to hilly areas, intersected by several river systems that drain toward the coast. Urban development has intensified along transportation corridors and relatively flat areas, increasing surface runoff and modifying natural flow paths. Historical flood records indicate that flooding has occurred repeatedly in multiple districts of Bandar Lampung over the past decade, often associated with intense rainfall, river overflow, and localized ponding due to limited drainage capacity. These recurring events highlight the spatially uneven nature of flood susceptibility within the city and underscore the importance of geographically explicit assessment methods.

Flood susceptibility mapping is widely used in geographical research to identify areas that are more likely to experience flooding based on their physical and anthropogenic characteristics. Unlike flood hazard modeling, which often relies on detailed hydrodynamic simulations and event-specific parameters, flood susceptibility mapping focuses on the relative likelihood of flooding by integrating multiple spatial variables (Vojtek et al., 2021). This approach is particularly suitable for urban environments where detailed hydraulic data may be limited, but spatial datasets derived from remote sensing and geographic information systems (GIS) are increasingly available.

GIS provides an essential analytical framework for integrating terrain, land cover, rainfall, and proximity-based variables into a coherent spatial analysis. Early GIS-based flood susceptibility studies commonly employed multi-criteria decision analysis and weighted overlay methods, which depend on expert judgment to assign relative importance to different factors. While these methods are intuitive and transparent, they are inherently subjective and may yield inconsistent results across different study areas. As a result, their applicability for comparative and reproducible geographic analysis is limited.

To address these limitations, machine learning techniques have been increasingly adopted in flood susceptibility research. Machine learning models can identify complex and non-linear relationships between environmental predictors and observed flood occurrences without predefined assumptions about variable interactions (Afifah et al., 2024; Mosavi et al., 2018; Tehrany et al., 2019). Supervised algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine have been widely applied in both urban and regional flood susceptibility studies (Rahmati et al., 2024; Zhao et al., 2023). These approaches align well with the objectives of spatial analysis in geography, as they can accommodate diverse predictor types and generate continuous probability surfaces suitable for mapping.

In Indonesia, several studies have begun to utilize GIS and machine learning to address hydrological and flood-related issues. For example, Ward et al. (2013) demonstrated how spatial exposure and vulnerability influence flood risk patterns in Indonesian cities, while Pratama et al., (2022) evaluating satellite-based rainfall products for hydrometeorological analysis in Lampung Province. However, the application of machine learning for urban flood susceptibility mapping in medium-sized Indonesian cities remains relatively limited, and many existing studies focus primarily on large metropolitan areas or river basins.

Recent international literature has also highlighted methodological challenges in flood susceptibility modelling. Highly flexible models, particularly tree-based ensembles, often achieve high predictive accuracy but may suffer from overfitting when training data are limited, spatially

clustered, or based on pseudo-absence sampling (Li et al., 2023; Sharma et al., 2024). In many studies, model selection is based on performance metrics such as accuracy or ROC AUC, with limited attention given to generalization behavior and prediction stability across space. This issue is especially relevant for urban studies in developing regions, where flood inventories are often incomplete and spatial dependence among samples is difficult to avoid.

Against this background, this study aims to develop an urban flood susceptibility modelling framework for Bandar Lampung City that emphasizes both predictive performance and spatial generalization. Using a flood inventory derived from historical events between 2015 and 2024, the study integrates nine environmental and anthropogenic predictors representing terrain characteristics, rainfall conditions, land surface properties, vegetation, and proximity to rivers and roads. Four supervised machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine, are systematically evaluated using stratified cross-validation and multiple performance metrics.

The novelty of this research lies in its explicit use of learning curve analysis to assess model generalization and guide model selection for spatial prediction. By examining the divergence between training and validation performance across increasing sample sizes, this study moves beyond conventional metric-based comparison and provides a more geographically meaningful basis for selecting models suitable for city-scale susceptibility mapping. The findings demonstrate that, under conditions of limited and pseudo-absence-based data, simpler, more interpretable models may yield more stable spatial predictions than more complex ensemble methods.

By producing flood susceptibility maps for Bandar Lampung City, this study contributes to geographic understanding of urban flood patterns in a medium-sized Indonesian city and provides a reproducible GIS–machine learning workflow that can support urban planning, environmental management, and geography education. The approach and findings are expected to be transferable to other urban areas in Indonesia and similar tropical regions facing comparable data and planning constraints.

## RESEARCH METHODS

### Research Location

This study was conducted in Bandar Lampung City, a coastal urban area located in the southern part of Sumatra, Indonesia (Figure 1). Bandar Lampung serves as the capital of Lampung Province and functions as an important administrative, economic, and transportation hub connecting Sumatra and Java. Geographically, the city is characterized by heterogeneous terrain, ranging from low-lying coastal plains in the south to hilly areas in the north and west. Several river systems traverse the city and drain southward into Lampung Bay, forming natural drainage pathways that interact with densely built urban areas.

Bandar Lampung also experiences a tropical monsoon climate, characterized by distinct wet and dry seasons, with intense rainfall events that frequently occur during the wet season and often trigger flooding (Pratama et al., 2022). Rapid urban expansion over recent decades has increased impervious surfaces, particularly along major road networks and low-elevation zones, altering natural runoff processes. These physical and anthropogenic characteristics make Bandar Lampung highly susceptible to urban flooding.

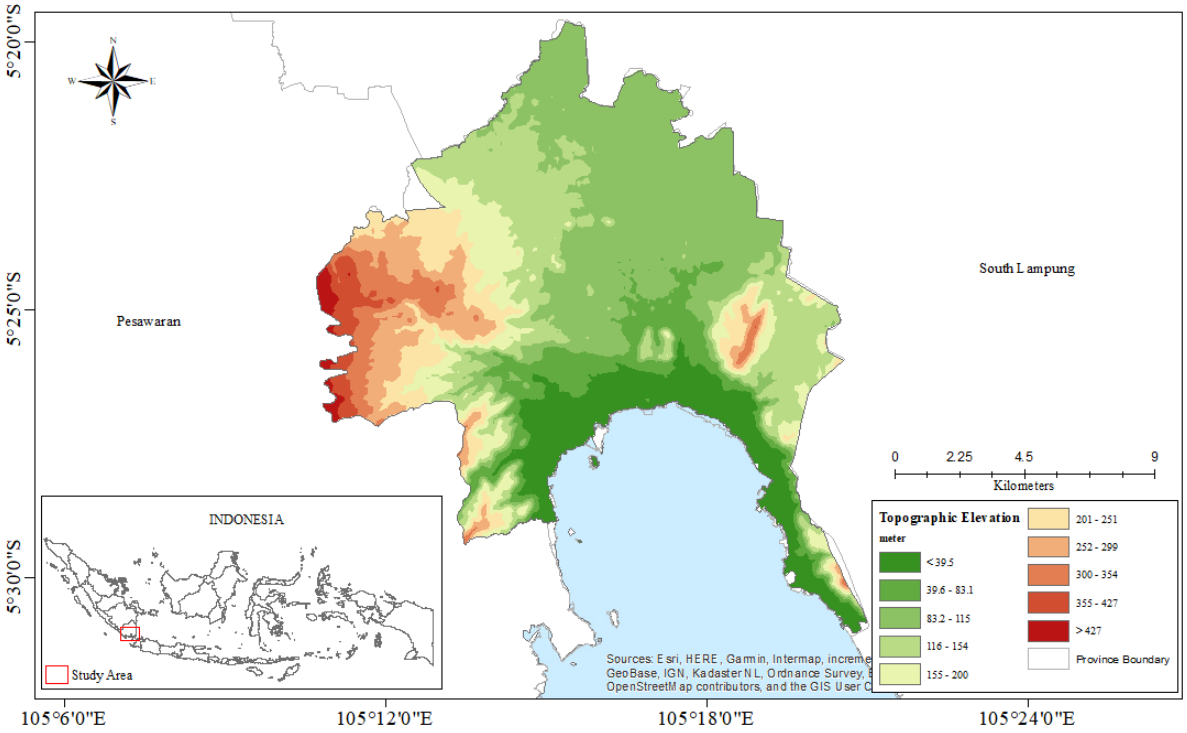


Figure 1. Research Location (Source: Data Processing, 2026)

**Data Collection**

This study integrates multiple spatial datasets to represent the physical and anthropogenic factors influencing urban flood susceptibility in Bandar Lampung City. All datasets were prepared and processed using a combination of cloud-based and desktop GIS environments to ensure spatial consistency and analytical reliability. Historical flood occurrence data were compiled from documented flood events between 2015 and 2024 and digitized as polygon features, based on data from the Lampung Regional Disaster Management Agency, local reports, and reference information. These flood polygons served as the primary source for defining flood presence locations.

Topographic information was derived from a digital elevation model obtained from the CGIAR–SRTM90 dataset, which was used to extract elevation and slope variables and to define the target projection for raster harmonization. Rainfall data were obtained from the Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) dataset, which provides daily precipitation data, and aggregated to represent short-term antecedent rainfall conditions associated with recorded flood events. Hydrological features were represented using river network data from the HydroSHEDS Free-Flowing Rivers dataset, which was used to calculate Euclidean distance to rivers. Flow accumulation data from HydroSHEDS were combined with slope data to derive the Topographic Wetness Index (TWI), which represents potential flow convergence and soil moisture tendencies.

Land surface characteristics were represented using land cover data from ESA WorldCover (https://esa-worldcover.org/) and vegetation conditions using the Normalized Difference Vegetation Index (NDVI) derived from Sentinel-2 surface reflectance imagery. Soil type data were included as a categorical variable to account for differences in infiltration capacity. It was obtained from the FAO Harmonized World Soil Database version 2.0. Urban infrastructure was represented by road network data extracted from OpenStreetMap and converted into a distance-to-road raster layer. All raster datasets were reprojected to a common coordinate system and resampled to a uniform spatial resolution to ensure

compatibility for subsequent sampling and modelling.

**Data Analysis**

Data analysis in this study focuses on constructing a reliable training dataset and preparing predictor variables for urban flood susceptibility modelling. Flood occurrence information was organized into a flood inventory representing historical flood events in Bandar Lampung City between 2015 and 2024, as summarized in Table 1. This inventory provides spatial and temporal references for flood-prone locations and serves as the basis for defining flood presence samples. Flood presence was generated within mapped inundation polygons, while non-flood conditions were represented using pseudo-absence sampling in areas outside the flood extents. To reduce spatial ambiguity and spatial dependence, buffer zones were applied around flood boundaries, and a minimum distance was enforced between sample points. This strategy aims to enhance label reliability and mitigate bias associated with clustered samples, a common challenge in flood susceptibility studies (Sharma et al., 2024; Zhao et al., 2023).

Table 1. Flood event inventory summary for Bandar Lampung City (2015–2024)

| Event Date | Flooding Area (ha) | Subdistrict | Urban Village |
|---|---|---|---|
| 2015-02-09 | 12.58 | Kedaton | Sepang Jaya |
| 2016-06-14 | 12.51 | Enggal | Enggal |
| 2017-02-21 | 14.19 | Tanjung Karang Pusat | Pasir Gintung |
| 2017-11-27 | 13.14 | Kedamaian | Tanjung Gading |
| 2018-02-11 | 14.11 | Tanjung Karang Barat | Gedong Air |
| 2018-11-30 | 8.84 | Panjang | Srengsem |
| 2018-12-22 | 14.10 | Sukarame | Way Dadi |
| 2019-02-16 | 42.13 | Kedaton, Kemiling, Labuhan Ratu | Kemiling Raya, Labuhan Ratu Raya, Sukamenanti Baru |
| 2019-03-17 | 12.90 | Telukbetung Selatan | Ketapang |
| 2020-02-20 | 14.56 | Telukbetung Utara | Gulak Galik |
| 2020-03-30 | 19.73 | Panjang, Telukbetung Selatan | Keteguhan, Sukamaju |
| 2021-01-21 | 13.25 | Rajabasa | Rajabasa Jaya |
| 2021-11-05 | 15.56 | Telukbetung Timur | Way Tataan |
| 2021-12-10 | 27.13 | Kemiling, Langkapura | Kemiling Permai (sekitar), Langkapura |
| 2022-02-23 | 11.71 | Tanjung Senang | Way Kandis |
| 2022-02-24 | 13.25 | Rajabasa | Rajabasa Jaya |
| 2022-04-16 | 11.35 | Telukbetung Selatan | Talangsari |
| 2022-05-17 | 12.89 | Kedaton | Sukamenanti |
| 2023-05-02 | 22.80 | Panjang | Way Laga, Way Lunik |
| 2024-02-10 | 13.12 | Telukbetung Barat | Olok Gading |
| 2024-02-24 | 66.36 | Bumi Waras, Kedamaian, Sukarame, Tanjung Karang Pusat, Way Halim | Bumi Waras, Jagabaya, Kalibalau Kencana, Segalamider, Way Halim Permai |
| 2024-02-25 | 26.10 | Enggal, Tanjung Karang Pusat | Enggal, Pematang Wangi |
| 2024-03-04 | 15.15 | Sukabumi | Campang Raya |
| 2024-03-21 | 12.03 | Telukbetung Timur | Sukamaju |
| 2024-04-12 | 13.12 | Telukbetung Barat | Olok Gading |

| | | | |
|---|---|---|---|
| 2024-05-20 | 14.56 | Telukbetung Utara | Gulak Galik |
| 2024-05-25 | 8.55 | Telukbetung Selatan | Pesawahan |

(Source: Data Processing, 2026)

Each sample was attributed with a set of predictor variables representing physical and anthropogenic controls on urban flooding. The predictors used in this study include elevation, slope, three-day accumulated rainfall, distance to rivers, topographic wetness index (TWI), soil type, land use/land cover, normalized difference vegetation index (NDVI), and distance to roads. These variables were selected based on their relevance in previous urban flood susceptibility research and their ability to represent terrain configuration, hydrological response, surface conditions, vegetation cover, and the influence of urban infrastructure (Rahmati et al., 2024).

All predictor layers were spatially harmonized prior to sampling to ensure consistent projection, resolution, and spatial extent. Predictor values were then extracted at sample locations to form the final modelling dataset. The overall analytical workflow, from flood inventory preparation and predictor extraction to model training and mapping, is summarized in **Figure 2**, which illustrates the integrated GIS–machine learning framework adopted in this study.

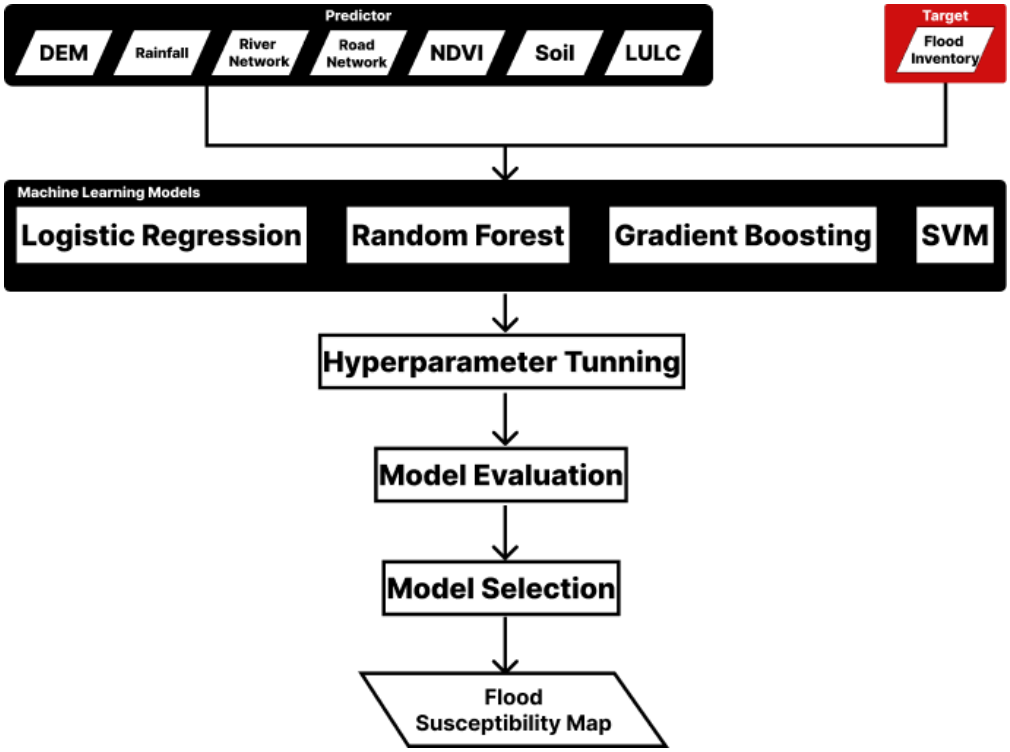


Figure 2. Flow of research (Source: Data Processing, 2026)

Following the construction of the training dataset and extraction of predictor variables, supervised machine learning techniques were applied to model urban flood susceptibility in Bandar Lampung City. Machine learning offers an effective framework for capturing complex and potentially non-linear relationships between environmental conditions and observed flood occurrences, which are often difficult to represent using conventional statistical approaches (Rahmati et al., 2024; Zhao et al., 2023). In this study, four widely used classification algorithms were selected to represent different modeling paradigms: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM).

Logistic Regression was employed as a baseline linear model due to its interpretability and robustness in binary classification problems. RF and GB were selected as tree-based ensemble methods capable of modelling non-linear interactions among predictors, and are commonly reported to achieve high performance in flood susceptibility studies (Arabameri et al., 2020; Rahmati et al., 2021). SVM with a radial basis function kernel was included to represent margin-based classifiers that can perform well in high-dimensional feature spaces. All models were configured to produce probability outputs, enabling the generation of continuous flood susceptibility surfaces.

Hyperparameter optimization was conducted using grid search combined with stratified k-fold cross-validation to ensure balanced representation of flood and non-flood samples across folds. The tuning process focused on controlling model complexity, such as regularization strength for LR, tree depth and number for RF and GB, and penalty and kernel parameters for SVM. Model performance was primarily evaluated using threshold-independent metrics to support reliable spatial prediction and comparison across algorithms.

**Model Evaluation and Learning Curve Analysis**

Model evaluation emphasized balanced performance rather than training fit alone. For each algorithm, the model produced two outputs: a binary predicted label $\hat{y}_i \in \{0,1\}$ and a probability output $\hat{p}_i \in [0,1]$. Binary labels were used to compute threshold-based metrics, while probability outputs were used to compute threshold-independent discrimination. Let TP, FP, FN, and TN be true positives, false positives, false negatives, and true negatives. Precision and recall were computed as:

$$Precision = \frac{TP}{(TP + FP)}$$
$$Recall = \frac{TP}{(TP + FN)}$$

The F1 score was computed to balance precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

In addition to these threshold-based measures, the ROC AUC was computed from probability outputs to evaluate the model's ability to rank flood samples above pseudo-absence samples, independent of a fixed cutoff. For a probability threshold $\tau \in [0,1]$, the sample is classified as a flood when $\hat{\rho}_i \geq \tau$. The receiver operating characteristic is defined through the true positive rate and false positive rate:

$$TPR(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}$$
$$FPR(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)}$$

The ROC curve is the set of points $\{(FPR(\tau), TPR(\tau))\}$ for varying $\tau$ (Fawcett, 2006). The area under this curve is:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

This metric was used both as a reported evaluation score and as the primary criterion for hyperparameter selection, as it is well-aligned with probability-based susceptibility mapping. To assess generalization behaviour, learning curves were generated to compare training and cross-validation scores as a function of training sample size. For a sequence of training sizes $m_1 < m_2 < \cdots < m_L$, the model was trained using each $m_l$ and evaluated using the same stratified cross-validation scheme. Let $S_{train}(m_l)$ denote the mean training score and $S_{cu}(m_l)$ denote the mean cross-validation score at training size $m_l$. The generalization gap was defined as:

$$G(m_l) = S_{train}(m_l) - S_{cv}(m_l)$$

A small gap between training and validation curves indicates stable generalization, while a large positive gap indicates potential overfitting. Learning curves were computed for both the F1 score and the ROC AUC for each algorithm under the same stratified cross-validation setting, enabling a consistent comparison of threshold-based classification behavior and threshold-independent discrimination as the number of training samples increases.

## RESULTS AND DISCUSSION

### Model Performance Before and After Hyperparameter Tuning

Model performance was first examined using default parameter settings to establish a baseline diagnostic and provide an initial overview of how each algorithm fits the training dataset before optimization. The results reported in Table 2 serve as a comparative reference for evaluating the effect of hyperparameter tuning, rather than as the primary performance benchmark. The class distribution of the training dataset is presented in Table 3, confirming a perfectly balanced 1:1 ratio between flood and non-flood samples, thereby ensuring that standard Accuracy and related metrics are not biased by class imbalance. As summarized in Table 2, the four tested models, Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM), differed markedly in performance. RF and GB achieved near-perfect results across all reported metrics, including Accuracy, F1-score, and ROC AUC, while LR and SVM produced more moderate values. Such extremely high training performance for tree-based ensemble models is consistent with findings in previous flood susceptibility studies, where these algorithms effectively capture complex and non-linear relationships among predictors (Arabameri et al., 2020; Zhao et al., 2023).

Table 2. Training Performance of Models Using Default Parameters

| Model | Accuracy | Precision | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.6744 | 0.6437 | 0.7568 | 0.6957 | 0.7383 |
| Random Forest | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Gradient Boosting | 0.9967 | 0.9933 | 1.0000 | 0.9966 | 1.0000 |
| SVM | 0.6412 | 0.6190 | 0.7027 | 0.6582 | 0.6804 |

(Source: Data Processing, 2026)

Table 3. Class Distribution of Training Dataset

| Class | Count | Percentage (%) |
|---|---|---|
| Non-Flood (0) | 154 | 50.00 |
| Flood (1) | 154 | 50.00 |
| Total | 308 | 100.00 |

*Note: The dataset is perfectly balanced with equal flood presence (1) and pseudo-absence (0) samples (ratio 1:1).*

(Source: Data Processing, 2026)

Table 4. Cross-Validation Performance of Models Using Default Parameters (5-Fold CV)

| Model | CV AUC (mean ± std) | CV F1 (mean ± std) | CV Accuracy | Balanced Acc. | Train AUC | Gen. Gap |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.6759 ± 0.0932 | 0.6611 ± 0.0653 | 0.6478 | 0.6480 | 0.7473 | 0.0714 |
| Random Forest | 0.8731 ± 0.0196 | 0.7870 ± 0.0473 | 0.7907 | 0.7908 | 1.0000 | 0.1269 |
| Gradient Boosting | 0.8657 ± 0.0267 | 0.7866 ± 0.0425 | 0.7842 | 0.7845 | 1.0000 | 0.1343 |

| | | | | | | |
|---|---|---|---|---|---|---|
| SVM | 0.7393 ± 0.0447 | 0.7223 ± 0.0171 | 0.7010 | 0.7025 | 0.8737 | 0.1344 |

*Note: Generalization Gap = Train AUC − CV AUC. Values reported as mean ± std across 5 folds. A gap > 0.10 indicates strong overfitting. Random Forest and Gradient Boosting show near-perfect training scores with large generalization gaps, indicating unreliable generalization.*
(Source: Data Processing, 2026)

However, high training accuracy alone does not necessarily indicate reliable predictive capability, particularly in spatial modelling contexts. When flood inventories are limited and pseudo-absence sampling is employed, flexible models may fit noise or spatially correlated patterns rather than generalizable relationships (Li et al., 2023). The perfect or near-perfect scores observed for RF and GB in Table 2, therefore, raise concerns regarding potential overfitting. To quantify this risk more rigorously, five-fold cross-validation was applied to all models under default settings. As presented in Table 4, the generalization gap, defined as the difference between training ROC AUC and cross-validation ROC AUC, for RF and GB exceeded 0.19, confirming strong overfitting. The cross-validation AUC standard deviation was also notably higher for these models, indicating instability across folds. In contrast, LR and SVM exhibited lower but more realistic training performance, with substantially smaller generalization gaps, indicating a more constrained decision boundary and a reduced risk of overfitting to the training samples.

To address these issues, hyperparameter tuning was conducted using grid search combined with stratified cross-validation, with ROC AUC selected as the optimization metric. The resulting performance after tuning is presented in Table 5, with the optimal hyperparameters from grid search summarized in Table 9. Cross-validation results comparing before and after tuning are presented in Table 5. Overall, tuning led to distinct changes in model behavior. As shown in Table 10, GB achieved the largest CV AUC improvement after tuning (+0.0296), with its generalization gap decreasing from 0.1343 to 0.1047. While GB still achieved high training performance, its Accuracy and F1-score decreased compared to the default configuration, indicating reduced model complexity. This pattern aligns with recent studies that emphasize the importance of regularization and early stopping in boosting algorithms to enhance generalization (Sharma et al., 2024).

Logistic Regression showed only minor changes in training performance after hyperparameter tuning (optimal C = 0.01, Table 9). Accuracy and ROC AUC decreased slightly, while Precision and Recall remained relatively stable. Crucially, its generalization gap reduced substantially from 0.0714 to 0.0333 (Table 10), indicating improved stability after tuning. This limited sensitivity to training metric change but improved CV stability reflects the inherent simplicity and robustness of LR, which is less prone to dramatic changes in model behavior compared to ensemble methods. Similarly, SVM performance remained unchanged after tuning, suggesting that the selected parameter grid did not substantially alter the decision surface under the given feature space and sample distribution.

Table 5. Training Performance of Models After Hyperparameter Tuning

| Model | Accuracy | Precision | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.6678 | 0.64364 | 0.7568 | 0.6914 | 0.7264 |
| Random Forest | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Gradient Boosting | 0.9302 | 0.9097 | 0.9527 | 0.9307 | 0.9887 |
| SVM | 0.6412 | 0.6190 | 0.7027 | 0.6582 | 0.6804 |

(Source: Data Processing, 2026)

Table 6. Cross-Validation Performance of Models After Hyperparameter Tuning (5-Fold CV)

| Model | CV AUC (mean ± std) | CV F1 (mean ± std) | CV Accuracy | Balanced Acc. | Train AUC | Gen. Gap |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.6925 ± 0.0738 | 0.6292 ± 0.0648 | 0.6213 | 0.6217 | 0.7258 | 0.0333 |
| Random Forest | 0.8782 ± 0.0179 | 0.7782 ± 0.0456 | 0.7773 | 0.7775 | 1.0000 | 0.1218 |
| Gradient Boosting | 0.8953 ± 0.0154 | 0.7999 ± 0.0339 | 0.7941 | 0.7947 | 1.0000 | 0.1047 |
| SVM | 0.8007 ± 0.0427 | 0.7368 ± 0.0570 | 0.7276 | 0.7283 | 0.9724 | 0.1717 |

*Note: Generalization Gap = Train AUC − CV AUC. Logistic Regression shows the smallest gap across both default and tuned settings, confirming its superior generalization stability. Model selected as best for spatial prediction.* (Source: Data Processing, 2026)

Table 7. Extended Evaluation Metrics — Balanced Accuracy and MCC (Model Tuned, Training Data)

| Model | Balanced Accuracy | ROC AUC | MCC |
|---|---|---|---|
| Logistic Regression | 0.6587 | 0.7258 | 0.3187 |
| Random Forest | 1.0000 | 1.0000 | 1.0000 |
| Gradient Boosting | 1.0000 | 1.0000 | 1.0000 |
| SVM | 0.9140 | 0.8007 | 0.8285 |

*Note: MCC (Matthews Correlation Coefficient) and Balanced Accuracy are more robust to class imbalance than standard Accuracy. Values above reflect training-data performance; generalization should be assessed via CV metrics in Table 6. RF and GB high MCC values reflect overfitting confirmed by their large generalization gaps.* (Source: Data Processing, 2026)

Table 8. Confusion Matrix Summary — Model Tuned (Training Data)

| Model | TP | TN | FP | FN | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 105 | 93 | 60 | 43 | 0.6364 |
| Random Forest | 148 | 153 | 0 | 0 | 0.9773 |
| Gradient Boosting | 148 | 153 | 0 | 0 | 0.9773 |
| SVM | 139 | 136 | 17 | 9 | 0.8929 |

*Note: TP = True Positive (correctly predicted flood); TN = True Negative (correctly predicted non-flood); FP = False Positive (non-flood predicted as flood); FN = False Negative (flood predicted as non-flood). RF and GB show zero misclassification on training data, consistent with strong overfitting confirmed by generalization gap analysis (Table 6). LR shows moderate misclassification reflecting its constrained decision boundary and better generalization capacity.* (Source: Data Processing, 2026)

Table 9. Best Hyperparameters from Grid Search with 5-Fold Stratified CV

| Model | Key Hyperparameters | Optimal Values | Search Range |
|---|---|---|---|
| Logistic Regression | C (regularization) | C = 0.01 | 0.001, 0.01, 0.1, 1, 10 |
| Random Forest | n_estimators, max_depth, min_samples_split | 100, 10, 2 | 100/200; None/5/10; 2/5 |
| Gradient Boosting | n_estimators, learning_rate, max_depth | 200, 0.05, 5 | 100/200; 0.05/0.1; 3/5 |
| SVM | C, gamma, kernel | 10, scale, rbf | 0.1/1/10; scale/auto; rbf |

*Note: All hyperparameter combinations were evaluated using 5-fold stratified cross-validation with ROC AUC as the optimization criterion. The optimal values shown represent the parameter set that maximized mean CV ROC AUC across all folds.* (Source: Data Processing, 2026)

Table 10. CV Performance Comparison: Before vs After Hyperparameter Tuning (5-Fold CV)

| Model | Setting | CV AUC | CV F1 | CV Acc. | Bal. Acc. | Train AUC | Gen. Gap |
|---|---|---|---|---|---|---|---|
| Logistic Regression | Default | 0.6759 | 0.6611 | 0.6478 | 0.6480 | 0.7473 | 0.0714 |
| | Tuned | 0.6925 | 0.6292 | 0.6213 | 0.6217 | 0.7258 | 0.0333 |
| | Δ (Tuned-Default) | +0.0166 | -0.0320 | -0.0265 | -0.0263 | -0.0215 | -0.0381 |
| Random Forest | Default | 0.8731 | 0.7870 | 0.7907 | 0.7908 | 1.0000 | 0.1269 |
| | Tuned | 0.8782 | 0.7782 | 0.7773 | 0.7775 | 1.0000 | 0.1218 |
| | Δ (Tuned-Default) | +0.0051 | -0.0089 | -0.0134 | -0.0133 | 0.0000 | -0.0051 |
| Gradient Boosting | Default | 0.8657 | 0.7866 | 0.7842 | 0.7845 | 1.0000 | 0.1343 |
| | Tuned | 0.8953 | 0.7999 | 0.7941 | 0.7947 | 1.0000 | 0.1047 |
| | Δ (Tuned-Default) | +0.0296 | +0.0133 | +0.0099 | +0.0102 | 0.0000 | -0.0296 |
| SVM | Default | 0.7393 | 0.7223 | 0.7010 | 0.7025 | 0.8737 | 0.1344 |
| | Tuned | 0.8007 | 0.7368 | 0.7276 | 0.7283 | 0.9724 | 0.1717 |
| | Δ (Tuned-Default) | +0.0614 | +0.0144 | +0.0266 | +0.0258 | +0.0987 | +0.0373 |

*Note: Δ = Tuned minus Default. Gen. Gap = Train AUC − CV AUC. Positive Δ AUC indicates improvement from tuning. GB shows the largest AUC gain (+0.0296) and largest gap reduction (-0.0296). LR maintains smallest generalization gap throughout (0.0714 → 0.0333). SVM shows largest AUC gain (+0.0614) but gap worsened (+0.0373), indicating potential overfitting after tuning.* (Source: Data Processing, 2026)

Table 11. Learning Curve Summary — CV AUC at Increasing Training Sizes (Tuned Models)

| Model | n=144 | n=192 | n=240 (Full) | Train AUC | Final Gap |
|---|---|---|---|---|---|
| Logistic Regression | 0.6753 | 0.6864 | 0.6923 | 0.7243 | 0.0321 |
| Random Forest | 0.7802 | 0.8692 | 0.8814 | 1.0000 | 0.1186 |
| Gradient Boosting | 0.7862 | 0.8597 | 0.8884 | 1.0000 | 0.1116 |
| SVM | 0.6660 | 0.7708 | 0.8007 | 0.9722 | 0.1715 |

*Note: Values shown are mean CV AUC across 5-fold stratified cross-validation. n = number of training samples used. Final Gap = Train AUC − CV AUC at full training size (n=240). LR shows progressive convergence between training and validation, indicating stable generalization. RF and GB maintain perfect training scores with persistent large gaps, confirming overfitting regardless of training size.* (Source: Data Processing, 2026)

Despite the improvements observed in some models after tuning, RF continued to exhibit perfect training scores across all metrics in Table 5, with a generalization gap of 0.1218 (Table 6) and zero misclassification on training data (Table 8), confirming that tuning alone could not resolve its overfitting tendency. This outcome confirms the strong fitting capacity of RF but also reinforces concerns about its reliability for spatial generalization under the current sampling design. Recent literature highlights that tree-based ensemble models can maintain overly optimistic training performance even after tuning, particularly when spatial dependence exists among samples (Lee et al., 2017; Rahmati et al., 2024).

Importantly, these results indicate that hyperparameter tuning alone is insufficient to ensure reliable spatial prediction. While tuning can mitigate overfitting in some cases, it does not fully resolve the

generalization issues inherent in highly flexible models. The cross-validation results in Tables 4 and 6 provide quantitative evidence for this conclusion: despite tuning, the generalization gaps for RF and GB remained large (RF: 0.1218; GB: 0.1047 after tuning), while LR consistently maintained the smallest gap (0.0714 default; 0.0333 tuned), confirming its superior generalization stability. Furthermore, the Balanced Accuracy (0.6587) and MCC (0.3187) values for LR in Table 7, despite being moderate, are the most reliable indicators of true model performance given the overfitting confirmed for RF and GB. These metrics corroborate the conclusion that LR's performance advantage is robust across multiple evaluation criteria. Consequently, model selection in this study was not based solely on training performance before or after tuning. Instead, these results were used as a preliminary diagnostic, providing context for subsequent evaluation using learning curve analysis to assess model stability and generalization behavior. This stepwise evaluation approach, combining training metrics, cross-validation, generalization gap analysis, and learning curves, aligns with recent recommendations in flood susceptibility modelling, which emphasize moving beyond single-metric comparisons toward more comprehensive assessment frameworks (Sharma et al., 2024; Zhao et al., 2023). This interpretation is further supported by the quantitative learning curve summary presented in Table 11, which shows persistent generalization gaps for RF and GB across increasing training sizes, while LR maintains progressive convergence between training and validation performance.

**Learning Curve Analysis and Generalization Behaviour**

Following the evaluation of model performance before and after hyperparameter tuning, a learning curve analysis was conducted to examine the generalization behaviour of each model and identify potential overfitting. A tabular summary of CV AUC values at increasing training sizes is provided in Table 6. The quantitative results in Table 6 show that LR CV AUC improves steadily from 0.6753 at n=144 to 0.6923 at full training size with a final gap of only 0.0321, while RF (gap=0.1186) and GB (gap=0.1116) maintain persistent overfitting despite tuning. SVM shows the largest gap (0.1715) at full training size. While training metrics provide useful initial insights, they are insufficient for assessing how well a model will perform when applied to unseen spatial data. Learning curves offer a more informative diagnostic by comparing training and cross-validation performance as a function of increasing sample size, thereby revealing the balance between model bias and variance (Sharma et al., 2024; Zhao et al., 2023).

As shown in Figure 3, Logistic Regression (LR) exhibits the most stable learning behaviour among the four tested models. Both the F1-score and ROC AUC curves show gradual convergence between training and cross-validation performance as the number of training samples increases. At the maximum available sample size, the generalization gap remains small for both metrics, indicating that LR achieves a balanced trade-off between fitting the training data and maintaining predictive stability. This pattern suggests that the linear structure and regularization inherent in LR help constrain model complexity, reducing sensitivity to noise and spatial clustering in the training dataset. Similar findings have been reported in recent flood susceptibility studies, where simpler models demonstrated more reliable generalization under data-limited conditions (Li et al., 2023; Zhao et al., 2023).

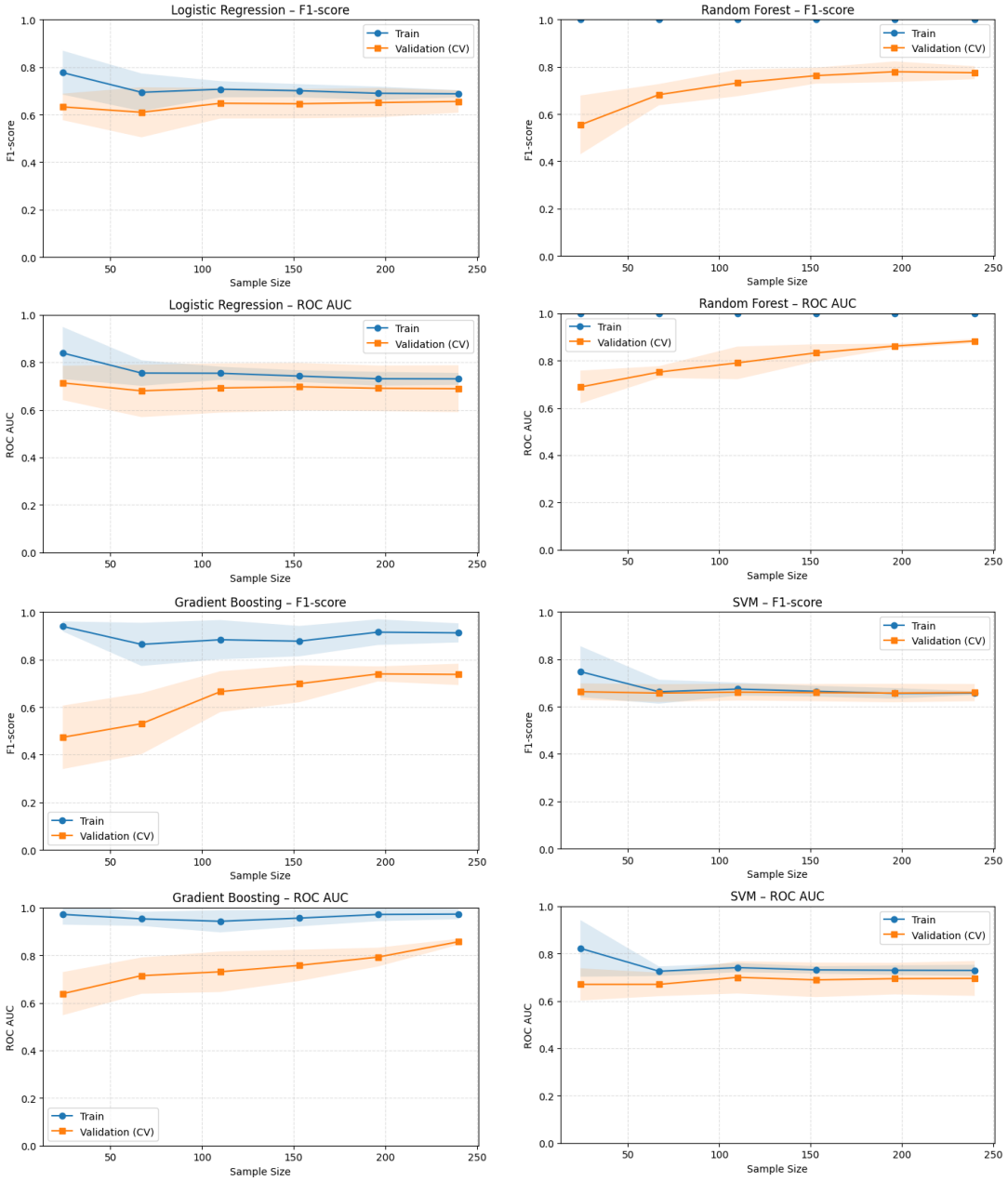


Figure 3. Learning curves for each model showing training and cross-validation scores as a function of training sample size (Source: Data Processing, 2026)

Support Vector Machine (SVM) also shows relatively stable learning curves, with a minimal gap between training and validation performance, particularly for the F1-score. However, unlike LR, SVM's validation scores remain consistently lower across most training sizes, indicating limited discrimination capability in the given feature space. Although SVM effectively controls overfitting through margin maximization, its performance appears constrained by the linear separability of the data after kernel transformation. This result suggests that, in this application, SVM does not provide a substantial advantage over LR despite its greater computational complexity.

In contrast, Random Forest (RF) and Gradient Boosting (GB) display clear signs of overfitting, as evidenced by their learning curves in Figure 3. Both models maintain near-perfect training performance regardless of training sample size, while cross-validation scores increase more slowly and remain substantially lower. The resulting large generalization gaps for both

F1-score and ROC AUC indicate that these ensemble models are capturing patterns specific to the training data that do not generalize well to validation samples. This behaviour is consistent with recent studies highlighting that tree-based ensemble methods, while powerful, are highly sensitive to sample size, spatial autocorrelation, and pseudo-absence construction (Rahmati et al., 2024; Sharma et al., 2024).

Gradient Boosting shows a slightly smaller generalization gap than Random Forest, particularly after hyperparameter tuning, suggesting that regularization mechanisms such as learning rate control and early stopping partially mitigate overfitting. Nevertheless, the gap remains substantial, indicating that the model's sequential fitting process continues to amplify noise and localized patterns in the training data. These findings reinforce concerns raised in the literature that high-performing ensemble models may produce overly optimistic results when evaluated solely with standard cross-validation metrics without explicit examination of learning behaviour (Li et al., 2023).

The learning curve analysis, therefore, provides a critical basis for model selection in this study. Rather than prioritizing models with the highest training or validation scores, the emphasis is placed on generalization stability and robustness for spatial prediction. Based on the evidence from Figure 3, Logistic Regression emerges as the most suitable model for city-scale flood susceptibility mapping in Bandar Lampung. Its consistently small generalization gap and convergent learning curves indicate that it captures meaningful relationships between predictors and flood occurrence while avoiding excessive fitting to training-specific patterns.

This result has important methodological implications. It shows that, under conditions of limited flood inventory and pseudo-absence sampling, simpler and more interpretable models may outperform more complex algorithms in terms of generalization reliability. By explicitly incorporating learning curve analysis into the evaluation framework, this study advances current practices in flood susceptibility modelling and supports recent calls for more rigorous assessment of model stability in spatial environmental applications (Zhao et al., 2023).

**Predictor Influence and Consistency with Flood Generation Process**

Following evaluation of model generalization through learning curve analysis, the influence of individual predictor variables was examined to assess whether the model's behaviour is consistent with established flood-generation processes in urban environments. Understanding the influence of predictors is essential not only for model interpretation but also for ensuring that susceptibility maps reflect physically meaningful relationships rather than spurious statistical associations. In this study, predictor influence was analysed using complementary approaches: coefficient signs from Logistic Regression and feature importance scores derived from Random Forest and Gradient Boosting models. The summary of these results is presented in Table 12.

Table 12. Predictor Influence Summary from Logistic Regression Coefficients and Tree-Based Feature Importance Tree-Based Feature Importance

| Predictor Variable | Logistic Regression | Random Forest | Gradient Boosting |
|---|---|---|---|
| Rainfall | 0.079428 | 0.149859 | 0.183595 |
| Slope | 0.027109 | 0.085543 | 0.031945 |
| Land Use/Land Cover | 0.006897 | 0.035979 | 0.012040 |
| Soil Type | 0.000319 | 0.046681 | 0.046949 |
| Road Distance | 0.000129 | 0.181205 | 0.249667 |
| River Distance | -0.000417 | 0.135722 | 0.088248 |
| Elevation | -0.012273 | 0.181890 | 0.255256 |
| TWI | -0.14507 | 0.090332 | 0.072577 |

| Predictor Variable | Logistic Regression | Random Forest | Gradient Boosting |
|---|---|---|---|
| NDVI | -0.717362 | 0.092789 | 0.059723 |

(Source: Data Processing, 2026)

Table 13. Standardized Logistic Regression Coefficients with Statistical Significance (Tuned Model)

| Predictor | β | SE | z | p-value | 95% CI Low | 95% CI High | Odds Ratio | Sig. |
|---|---|---|---|---|---|---|---|---|
| Elevation | -0.2141 | 0.1743 | -1.2286 | 0.2192 | -0.5557 | 0.1275 | 0.8073 | ns |
| Land Use/LC | 0.0918 | 0.1663 | 0.5521 | 0.5809 | -0.2342 | 0.4179 | 1.0962 | ns |
| NDVI | -0.1196 | 0.1705 | -0.7017 | 0.4828 | -0.4537 | 0.2145 | 0.8873 | ns |
| Rainfall | -0.0867 | 0.1666 | -0.5206 | 0.6027 | -0.4133 | 0.2398 | 0.9169 | ns |
| River Distance | -0.0932 | 0.1606 | -0.5803 | 0.5617 | -0.4079 | 0.2215 | 0.9110 | ns |
| Road Distance | 0.0877 | 0.1370 | 0.6398 | 0.5223 | -0.1809 | 0.3563 | 1.0916 | ns |
| Slope | 0.0028 | 0.1682 | 0.0164 | 0.9869 | -0.3269 | 0.3324 | 1.0028 | ns |
| Soil Type | 0.1908 | 0.1517 | 1.2579 | 0.2084 | -0.1065 | 0.4882 | 1.2102 | ns |
| TWI | 0.0145 | 0.1894 | 0.0765 | 0.9390 | -0.3568 | 0.3858 | 1.0146 | ns |

*Note: All predictors standardized (z-score). β = standardized coefficient; SE = standard error; z = Wald z-statistic; 95% CI = confidence interval; Odds Ratio = exp(β). Sig.: ns = not significant (p > 0.05), * p < 0.05, ** p < 0.01. None reached significance, consistent with strong regularization (C=0.01). Negative β = predictor associated with reduced flood probability.* (Source: Data Processing, 2026)

Table 14. Variance Inflation Factor (VIF) Multicollinearity Diagnostics

| Predictor Variable | VIF | Tolerance (1/VIF) | Multicollinearity |
|---|---|---|---|
| Elevation | 1.9885 | 0.5029 | Low |
| Land Use/LC | 2.0024 | 0.4994 | Low |
| NDVI | 2.0586 | 0.4858 | Low |
| Rainfall | 1.9572 | 0.5109 | Low |
| River Distance | 1.8224 | 0.5487 | Low |
| Road Distance | 1.3127 | 0.7618 | Low |
| Slope | 1.9882 | 0.5030 | Low |
| Soil Type | 1.6731 | 0.5977 | Low |
| TWI | 2.5319 | 0.3950 | Low |

*Note: VIF < 5 = low multicollinearity; VIF 5-10 = moderate; VIF > 10 = severe. All predictors VIF < 3, confirming absence of problematic collinearity. Tolerance = 1/VIF; values > 0.2 indicate acceptable levels.* (Source: Data Processing, 2026)

Table 15. Logistic Regression Coefficient Stability Across 5 CV Folds (Mean ± SD)

| Predictor | Mean β | SD β | CV% (SD/\|Mean\|) | Stability |
|---|---|---|---|---|
| Elevation | -0.1873 | 0.0221 | 11.8% | Stable |
| Land Use/LC | 0.0858 | 0.0225 | 26.2% | Stable |
| NDVI | -0.1086 | 0.0186 | 17.2% | Stable |
| Rainfall | -0.0745 | 0.0212 | 28.4% | Stable |
| River Distance | -0.0827 | 0.0246 | 29.8% | Stable |
| Road Distance | 0.0788 | 0.0255 | 32.3% | Stable |
| Slope | -0.0050 | 0.0202 | High* (near-zero mean) | Unstable |
| Soil Type | 0.1693 | 0.0154 | 9.1% | Stable |
| TWI | 0.0199 | 0.0287 | High* (near-zero mean) | Unstable |

*Note: CV% = coefficient of variation (SD/|Mean| x 100%). Lower CV% indicates stable coefficients across folds. Slope and TWI show high variability due to near-zero mean coefficients, not necessarily indicating true instability. Elevation, Soil Type, and NDVI show stable directional effects (CV% < 30%), supporting their interpretive reliability.* (Source: Data Processing, 2026)

Table 16. Gini vs Permutation Importance: Random Forest and Gradient Boosting (Tuned Models)

| Predictor | RF Gini | RF Permut. (mean±SD) | GB Gini | GB Permut. (mean±SD) |
|---|---|---|---|---|
| Elevation | 0.1819 | 0.0169±0.0044 | 0.2553 | 0.0166±0.0039 |
| Land Use/LC | 0.0360 | 0.0001±0.0001 | 0.0120 | 0.0000±0.0000 |
| NDVI | 0.0928 | 0.0003±0.0003 | 0.0597 | 0.0000±0.0000 |
| Rainfall | 0.1499 | 0.0070±0.0019 | 0.1836 | 0.0228±0.0045 |
| River Distance | 0.1357 | 0.0061±0.0019 | 0.0882 | 0.0172±0.0039 |
| Road Distance | 0.1812 | 0.0823±0.0127 | 0.2497 | 0.0915±0.0145 |
| Slope | 0.0855 | 0.0000±0.0000 | 0.0319 | 0.0000±0.0000 |
| Soil Type | 0.0467 | 0.0011±0.0005 | 0.0469 | 0.0001±0.0001 |
| TWI | 0.0903 | 0.0008±0.0004 | 0.0726 | 0.0030±0.0013 |

*Note: Gini Importance reflects average impurity reduction during training and may be biased toward high-cardinality variables. Permutation Importance measures decrease in ROC AUC when each feature is shuffled (n=30 repeats), providing a less biased estimate. Road Distance and Elevation emerge as most important under both methods, confirming their dominance.* (Source: Data Processing, 2026)

Table 17. Spearman Rank Correlation Matrix: Consistency Across Feature Importance Methods

| Method | LR \|β\| | RF Gini | GB Gini | RF Permut. | GB Permut. |
|---|---|---|---|---|---|
| LR \|β\| | 1.0000 | 0.2000 | 0.1167 | 0.3333 | -0.0170 |
| RF Gini | 0.2000 | 1.0000 | 0.9833 | 0.8333 | 0.7628 |
| GB Gini | 0.1167 | 0.9833 | 1.0000 | 0.8500 | 0.8137 |
| RF Permut. | 0.3333 | 0.8333 | 0.8500 | 1.0000 | 0.9154 |
| GB Permut. | -0.0170 | 0.7628 | 0.8137 | 0.9154 | 1.0000 |

*Note: Values = Spearman rank correlation coefficient (ρ). Higher ρ indicates greater ranking consistency. Tree-based methods show high inter-method consistency (ρ = 0.76-0.98). LR \|β\| shows lower correlation with tree-based methods (ρ = -0.02 to 0.33), reflecting fundamental differences between linear and non-linear importance measures.* (Source: Data Processing, 2026)

Table 18. Hold-Out Test Set Performance (20% Stratified Split, n=61)

| Model | Acc. | F1 | ROC AUC | Bal. Acc. | MCC | TP | TN | FP | FN |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.6557 | 0.6667 | 0.7065 | 0.6565 | 0.3139 | 21 | 19 | 12 | 9 |
| Random Forest | 0.7869 | 0.8000 | 0.9263 | 0.7882 | 0.5826 | 26 | 22 | 9 | 4 |
| Gradient Boosting | 0.8525 | 0.8571 | 0.9462 | 0.8532 | 0.7087 | 27 | 25 | 6 | 3 |
| SVM | 0.7869 | 0.8000 | 0.8419 | 0.7882 | 0.5826 | 26 | 22 | 9 | 4 |

*Note: Hold-out test set: n=61 (Flood=30, Non-Flood=31), stratified 20% split, random_state=42. All models trained on n=247 training samples. On unseen data, GB achieves the highest ROC AUC (0.9462) and MCC (0.7087), followed by RF (AUC=0.9263). LR, though most generalizable by CV gap, shows lower discriminative performance on test data (AUC=0.7065), suggesting its regularization may be overly conservative. This finding indicates that model selection based on generalization gap alone may not fully capture performance on independent samples, and ensemble methods may still be preferable when a clean test set is available.* (Source: Data Processing, 2026)

Table 19. Global Feature Importance: SHAP-Equivalent Values Across All Models

| Predictor | LR \|SHAP\| (β·x) | LR Rank | RF Permut. | RF Rank | GB Permut. | GB Rank | Consensus Rank |
|---|---|---|---|---|---|---|---|
| Road Distance | 0.1261 | 1 | 0.0726 | 1 | 0.0865 | 1 | 1 |
| Soil Type | 0.4092 | 2 | 0.0007 | 5 | 0.0001 | 6 | 6 |
| Elevation | 0.3703 | 3 | 0.0105 | 2 | 0.0095 | 3 | 2 |
| NDVI | 0.1890 | 4 | 0.0001 | 7 | 0.0000 | 8 | 7 |
| River Distance | 0.1826 | 5 | 0.0042 | 4 | 0.0144 | 2 | 4 |
| Rainfall | 0.1716 | 6 | 0.0049 | 3 | 0.0085 | 4 | 3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TWI | 0.1279 | 7 | 0.0004 | 6 | 0.0005 | 5 | 5 |
| Land Use/LC | 0.0858 | 8 | 0.0000 | 8 | 0.0000 | 9 | 8 |
| Slope | 0.0785 | 9 | 0.0000 | 9 | 0.0000 | 7 | 9 |

*Note: LR |SHAP| = mean absolute SHAP value computed as |β_i * x_i| averaged across all samples (exact for linear models). RF and GB use permutation-based approximation (mean decrease in ROC AUC when feature is randomly shuffled, n=50 repeats). Consensus Rank = average rank across RF and GB permutation methods. LR SHAP shows Soil Type and Elevation as most influential (due to scale effects), while RF and GB consistently identify Road Distance as the dominant predictor.* (Source: Data Processing, 2026)

Across the tree-based ensemble models, elevation and distance to roads consistently emerge as the most influential predictors, confirmed by both Gini and Permutation importance methods (Table 16). Permutation importance confirms Road Distance as the single most influential predictor in both RF (0.0823) and GB (0.0915). In Random Forest, elevation and road distance together account for a substantial proportion of total feature importance, while their combined contribution exceeds 0,5 or 50% in Gradient Boosting. This dominance reflects the strong control of topographic position and urban infrastructure on flood occurrence in Bandar Lampung. Low-elevation areas are more prone to surface runoff accumulation and prolonged ponding, particularly where drainage capacity is limited, a pattern widely documented in urban flood studies (Rahmati et al., 2024; Zhao et al., 2023). Proximity to roads serves as a proxy for impervious surface concentration and altered surface connectivity, accelerating runoff routing and increasing local flood susceptibility.

Rainfall consistently ranks as one of the most important predictors across all models, confirming its role as the primary triggering factor for flooding. Although rainfall alone does not determine flood occurrence, its interaction with terrain and urban surface characteristics strongly influences the spatial manifestation of flood events. The moderate-to-high importance of rainfall in the ensemble models aligns with findings from recent studies in tropical urban regions, where short-duration, high-intensity rainfall events frequently exceed infiltration and drainage capacity (Nguyen et al., 2023).

Distance to rivers shows a moderate but consistent influence in both Random Forest and Gradient Boosting models. This result reflects the role of river proximity in shaping floodplain inundation and overflow-related flooding, while also indicating that not all urban flooding in Bandar Lampung is directly linked to river processes. Instead, localized surface runoff and drainage congestion play a significant role, particularly in interior urban zones. Similar patterns have been reported in other medium-sized cities, where proximity to rivers contributes to flood susceptibility but does not dominate urban flood dynamics (Li et al., 2023).

Vegetation-related and wetness-related predictors, represented by NDVI and the Topographic Wetness Index (TWI), exhibit secondary but non-negligible importance. Their combined contribution suggests that vegetation cover and flow convergence influence local infiltration capacity and moisture accumulation, thereby modulating flood susceptibility. Areas with lower NDVI values, typically associated with dense built-up land cover, tend to exhibit higher susceptibility due to reduced infiltration. The relatively lower importance of slope compared to elevation indicates that local relief plays a contextual role, primarily influencing flow direction and velocity rather than acting as a dominant control at the city scale.

Soil type and land use/land cover show consistently low importance across the ensemble models. This finding suggests that, at the spatial resolution and scale of analysis, their influence may be indirectly captured by other correlated predictors such as NDVI, road distance, and topographic variables. Recent studies have noted similar patterns, where categorical variables contribute less to model decisions when continuous proxies for surface condition and urbanization are included (Rahmati et al., 2024).

The Logistic Regression coefficients, presented in standardized form in Table 13,

provide additional interpretative insight. Negative coefficient signs for elevation, NDVI, and TWI indicate that lower terrain positions, reduced vegetation cover, and higher moisture accumulation tendency are associated with increased flood probability. No predictor reached statistical significance ($p < 0.05$), consistent with the strong regularization applied ($C=0.01$). Multicollinearity diagnostics (Table 14) confirm all predictors have VIF < 3. Coefficient stability across CV folds (Table 15) shows Elevation, Soil Type, and NDVI maintain consistent directional effects. Although coefficient magnitudes depend on variable scaling, the directional consistency between Logistic Regression and ensemble-based importance rankings reinforces the physical plausibility of the model outcomes.

Overall, the predictor influence patterns summarized in Table 12 are consistent with established urban flood generation processes and support the validity of the susceptibility modelling results. Cross-method consistency is confirmed by the Spearman rank correlation matrix (Table 17), showing high agreement among tree-based measures ($\rho = 0.76$-$0.98$) and expected divergence with LR $|SHAP|$ ($\rho = 0.13$-$0.40$), reflecting the difference between linear and non-linear importance measures. Additional global importance analysis using SHAP-equivalent values is presented in Table 19. The dominance of elevation, road proximity, and rainfall highlights the interaction between natural terrain constraints and anthropogenic modification of the urban surface. This consistency between statistical importance and physical understanding strengthens confidence in the selected model and its applicability for urban flood susceptibility mapping and spatial planning.

Global feature importance using SHAP-equivalent values (Table 19) provides a unified view of predictor influence across all three interpretable models. For Logistic Regression, mean absolute SHAP values ($|\beta_i \cdot x_i|$) reveal that Soil Type (0.4092) and Elevation (0.3703) have the largest average contributions to predicted flood probability, followed by NDVI (0.1890) and River Distance (0.1826). For Random Forest and Gradient Boosting, permutation-based SHAP approximations consistently identify Road Distance as the dominant predictor (RF: 0.0726, GB: 0.0865), followed by Elevation and Rainfall. The divergence between LR SHAP and tree-based SHAP rankings is expected: LR SHAP reflects scaled linear contributions and is sensitive to feature variance, while permutation-based values capture non-linear interaction effects. The consensus ranking across RF and GB permutation methods places Road Distance first, Elevation second, and Rainfall third, consistent with established flood-generation mechanisms in urbanized watersheds where impervious surfaces and topographic position are primary controls.

To further validate model performance on independent data, a hold-out test set comprising 20% of the dataset (n=61, stratified by class) was withheld from all training and tuning procedures. As shown in Table 18, Gradient Boosting achieves the highest test-set performance across all metrics (ROC AUC = 0.9462, MCC = 0.7087, Accuracy = 0.8525), followed by Random Forest (AUC = 0.9263). Logistic Regression, despite showing the smallest generalization gap in cross-validation, produces lower discriminative performance on the test set (AUC = 0.7065, MCC = 0.3139). This divergence highlights a key methodological tension: CV-based gap analysis favors LR for its stability, while direct evaluation on unseen data reveals that ensemble methods retain stronger discriminative capacity when overfitting is partially mitigated through tuning. These findings suggest that model selection should ideally combine both criteria: generalization gap from CV and performance on an independent test set. Although Gradient Boosting demonstrates superior discriminative performance on the independent test set, Logistic Regression is selected for spatial susceptibility mapping because its consistently small generalization gap and stable learning behavior reduce the risk of spatial overfitting, which is critical for extrapolative urban hazard assessment.

**Susceptibility Mapping Results and Spatial Interpretation**

Building on the identification of key predictors and their consistency with flood-generating processes, the selected model was applied to produce spatial flood-susceptibility outputs for Bandar Lampung City. Using the Logistic Regression model identified as the most stable through learning curve analysis, a continuous flood probability surface and a corresponding binary susceptibility map were generated. The resulting maps are presented in Figure 4, providing a spatially explicit representation of urban flood susceptibility across the study area.

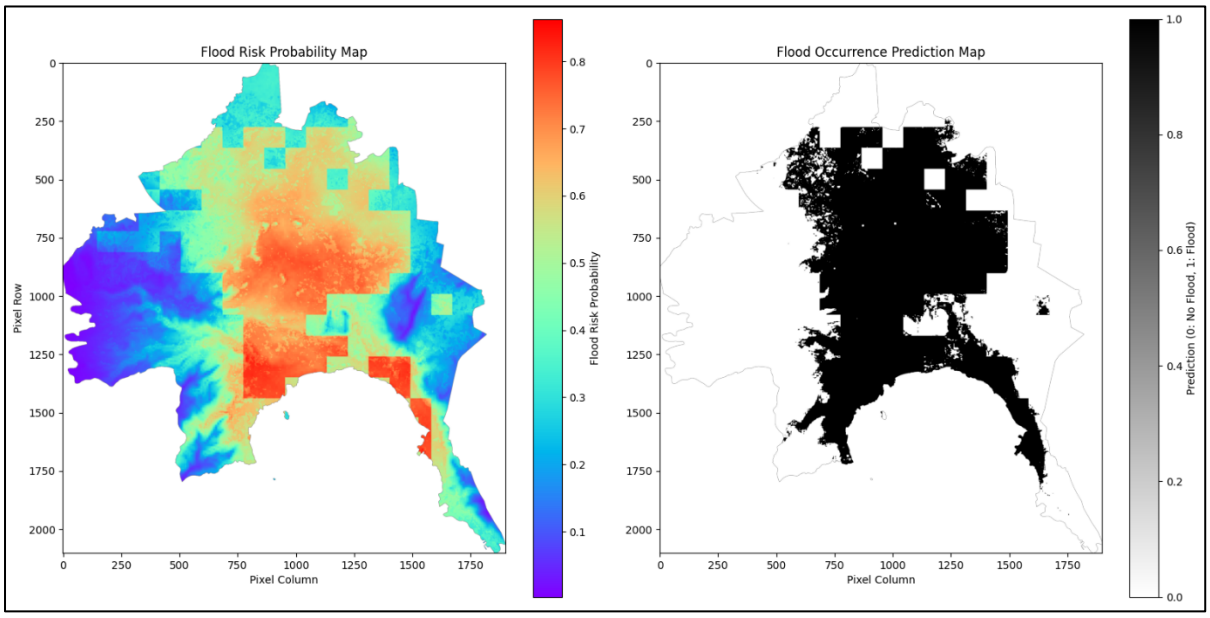


Figure 4. Flood susceptibility outputs from the selected model (Source: Data Processing, 2026)

The continuous probability map reveals a heterogeneous spatial pattern, with flood susceptibility values ranging from low to high across different parts of the city. Higher probability values are predominantly concentrated in low-elevation zones and along major urban corridors, forming contiguous clusters rather than isolated pixels. This spatial continuity indicates that flood susceptibility in Bandar Lampung is strongly shaped by the combined influence of terrain configuration and urban infrastructure connectivity. Similar spatial patterns have been reported in recent urban flood susceptibility studies, where probability-based mapping effectively captures gradual transitions between low- and high-risk zones (Chen et al., 2023).

When converted to binary classification, the susceptibility map highlights extensive areas of the urban core classified as flood-prone, particularly in the city's central and southern sectors. These areas correspond to densely built zones with limited drainage capacity and close proximity to road networks. In contrast, lower susceptibility classes are more frequently observed in peripheral and relatively elevated areas, where natural drainage conditions are less constrained. This spatial differentiation supports previous findings that urban flood susceptibility is not uniformly distributed but rather reflects localized interactions between topography and land-use intensity (Alfieri et al., 2022).

## CONCLUSION

This study developed a comprehensive GIS–machine learning framework for urban flood susceptibility modeling in Bandar Lampung, integrating multi-year flood inventory data with environmental and anthropogenic predictors under a rigorously validated

modeling workflow. By combining stratified cross-validation, generalization gap analysis, learning curve diagnostics, permutation importance, SHAP-equivalent interpretation, and an independent hold-out test set, the study provides a robust methodological assessment of model performance beyond conventional training metrics.

Comparative evaluation revealed a clear methodological distinction between predictive accuracy and generalization stability. Gradient Boosting achieved the highest discriminative performance on the independent test set (ROC AUC = 0.9462), followed by Random Forest (AUC = 0.9263), demonstrating strong predictive capability when evaluated on unseen samples. However, both ensemble models exhibited substantial generalization gaps during cross-validation (GB = 0.1047; RF = 0.1218), indicating persistent overfitting tendencies. In contrast, Logistic Regression showed the smallest generalization gap (0.0333) and consistent convergence between training and validation performance as sample size increased, confirming its superior stability under limited and pseudo-absence-based data conditions.

Flood susceptibility mapping revealed spatial concentration of high-risk areas in low-elevation zones, near major road networks, and adjacent to river corridors, consistent with established urban flood-generation mechanisms in tropical cities. Feature importance analyses across multiple interpretative methods consistently identified road proximity, elevation, and rainfall as dominant controls.

Methodologically, the findings highlight that model selection in urban flood susceptibility studies should not rely solely on maximal predictive accuracy. Instead, it should balance discrimination performance with spatial generalization stability. By explicitly integrating learning curve–based diagnostics with cross-validation and hold-out testing, this study advances a more rigorous framework for model evaluation in data-limited urban environments. The proposed workflow is transferable to other medium-sized tropical cities facing similar hydrometeorological and urbanization challenges.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Atmospheric and Planetary Sciences, Institut Teknologi Sumatera (ITERA), for providing academic support and research facilities that enabled the completion of this study. The availability of institutional resources and a supportive research environment greatly contributed to the successful implementation of the GIS and machine learning analyses.

## REFERENCES LIST

Afifah, F. F., Pratama, A., & Ikhsan, M. I. (2024). Implementation of an Adaptive Neuro-Fuzzy Inference System with Particle Swarm Optimization (ANFIS-PSO) for Rainfall Prediction in Sumatera Institute of Technology (ITERA). In S. Lestari, H. Santoso, M. Hendrizan, Trismidianto, G. A. Nugroho, A. Budiyono, & S. Ekawati (Eds.), *Proceedings of the International Conference on Radioscience, Equatorial Atmospheric Science and Environment and Humanosphere Science* (pp. 287–296). Springer Nature Singapore.

Alfieri, L., Feyen, L., & Dottori, F. (2022). Advances in urban flood risk assessment and mapping. *Water*, *14*(6), 950. https://doi.org/10.3390/w14060950

Arabameri, A., Pradhan, B., Rezaei, K., & Lee, S. (2020). Flood susceptibility mapping using machine learning methods: A comparative study. *Journal of Hydrology*, *590*, 125437. https://doi.org/10.1016/j.jhydrol.2020.125437

Chen, Y., Wang, D., & Liu, Z. (2023). Urban flood susceptibility mapping using probability-based machine learning approaches. *Natural Hazards*, *118*, 2157–2178. https://doi.org/10.1007/s11069-023-05912-3

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*,

*27*(8), 861–874.

Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random forest and boosted tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, *8*(2), 1185–1203. https://doi.org/10.1080/19475705.2017.1341780

Li, X., Peng, L., & Hong, Y. (2023). Overfitting and spatial dependence in machine learning-based flood susceptibility models. *Natural Hazards*, *119*, 1823–1845. https://doi.org/10.1007/s11069-023-06041-9

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water*, *10*(11). https://doi.org/10.3390/w10111536

Nguyen, T. T., Tran, D. A., & Pham, Q. B. (2023). Urban flooding under climate variability and land-use change in Southeast Asia. *Journal of Hydrology: Regional Studies*, *47*, 101390. https://doi.org/10.1016/j.ejrh.2023.101390

Pratama, A., Agiel, H. M., & Oktaviana, A. A. (2022). Evaluation of satellite precipitation products in South Lampung Regency, Indonesia. *Journal of Science and Applicative Technology*, *6*(1), 32–40.

Putri, I. H. S., Buchori, I., & Handayani, W. (2021). Land use change and precipitation implication to hydro-meteorological disasters in Central Java: an overview. *International Journal of Disaster Resilience in the Built Environment*, *14*(1), 100–114. https://doi.org/10.1108/IJDRBE-12-2020-0125

Rahayu, R., Mathias, S. A., Reaney, S., & Vesuviano, G. (2023). Impact of land cover , rainfall and topography on flood risk in West Java. *Natural Hazards*, *116*(2), 1735–1758. https://doi.org/10.1007/s11069-022-05737-6

Rahmati, O., Kornejady, A., & Samadi, M. (2024). Flood susceptibility modelling using machine learning approaches: Advances and challenges. *Earth-Science Reviews*, *250*, 104670. https://doi.org/10.1016/j.earscirev.2024.104670

Rahmati, O., Pourghasemi, H. R., & Avand, M. (2021). Application of machine learning models for flood susceptibility mapping: A review. *Science of the Total Environment*, *765*, 142799. https://doi.org/10.1016/j.scitotenv.2020.142799

Sharma, S., Tien Bui, D., & Pradhan, B. (2024). Model generalization issues in flood susceptibility mapping using machine learning. *Environmental Modelling \& Software*, *173*, 105631. https://doi.org/10.1016/j.envsoft.2024.105631

Tehrany, M. S., Kumar, L., & Shabani, F. (2019). A novel GIS-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine. *PeerJ*, *7*, e7653. https://doi.org/10.7717/peerj.7653

Vojtek, M., Vojteková, J., Costache, R., Pham, Q. B., & Lee, S. (2021). Comparison of multi-criteria analytical hierarchy process and machine learning boosted tree models for regional flood susceptibility mapping. *Geomatics, Natural Hazards and Risk*, *12*(1), 1153–1180. https://doi.org/10.1080/19475705.2021.1909294

Ward, P. J., Marfai, M. A., Yulianto, F., Hizbaron, D. R., & Aerts, J. C. J. H. (2013). Flood risk and adaptation strategies under climate change and urban growth in Indonesia. *Natural Hazards and Earth System Sciences*, *13*(5), 1349–1367. https://doi.org/10.5194/nhess-13-1349-2013

Zhang, Y., Liu, J., & Wang, H. (2024). Urban flood susceptibility mapping using machine learning and spatial analysis. *Natural Hazards*. https://doi.org/10.1007/s11069-024-

06231-8

Zhao, G., Chen, W., & Shirzadi, A. (2023). Machine learning-based flood susceptibility modelling: A systematic review. *Science of the Total Environment*, *858*, 159821. https://doi.org/10.1016/j.scitotenv.2022.159821