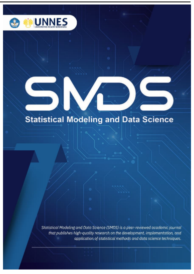
Vol. 1, No. 1 (2026) |

OPEN ACCESS

Statistical Modeling and Data Science

https://journal.unnes.ac.id/journals/smds



Comparative explainable ensemble learning for flood susceptibility mapping using Sentinel-1-derived flood inventory and multi-source geospatial data: A case study of the Semarang-Demak coastal area

M Irfan Al Hakim[1*] Tsalis Ainur Rofi[1] Taufik Dwi Ferdiansyah[1] Riski Cahya Alfiansyah[1] Abdullah Azzam Al Muhtasib[1] Nur Achmey Selgi Harwanti[1]

[1]*Department of Statistics and Data Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia*
[*]*Corresponding author, email: mirfanhakim25@students.unnes.ac.id*

Abstract

Flood susceptibility mapping is increasingly supported by remote sensing and machine learning, yet pixel-level flood inventories remain difficult to obtain in many Indonesian coastal regions where official disaster records are commonly reported at administrative scale. This study develops a comparative explainable ensemble-learning framework for flood susceptibility mapping in the Semarang-Demak coastal area by integrating a Sentinel-1-derived flood inventory with multi-source geospatial predictors from Google Earth Engine. Sentinel-1 SAR was used to derive the binary flood label from pre- and post-event change detection of the March 2024 flood, while conditioning variables represented topographic, hydrological, meteorological, spectral, land-cover, and built-environment factors. Seven ensemble models were evaluated on a balanced dataset of 1,870 samples, consisting of 935 flood and 935 non-flood observations: Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, CatBoost, and a stacking ensemble. Random Forest achieved the best overall discrimination with ROC-AUC = 0.852, accuracy = 0.791, recall = 0.932, and F1-score = 0.816 at the optimized threshold. DeLong tests and paired bootstrap resampling showed that the differences among the leading models were not statistically significant, whereas spatial block cross-validation revealed that the random-split performance estimates are optimistic for spatial generalization. SHAP interpretation, computed on training data, indicated that distance to permanent water, antecedent rainfall, NDVI, upstream drainage area, MNDWI, NDBI, river width, and distance to river were the dominant predictors, implying that flood susceptibility in the study area is governed by interacting hydrological proximity, rainfall accumulation, surface moisture, vegetation condition, and built-up intensity. The susceptibility map highlights high-risk patterns in low-lying, water-adjacent, and urbanized coastal zones. The proposed framework contributes a reproducible workflow for transforming Sentinel-1 flood detection into an explainable susceptibility model for data-limited coastal urban environments.

Keywords: *Coastal Flooding, Ensemble Learning, Flood Susceptibility, Google Earth Engine, Sentinel-1, SHAP*

ISSN xxxx-xxxx (Online)/© 2026 Authors

1. Introduction

The contribution of this study is not a new learning algorithm, but a reproducible workflow that connects three components that are rarely combined in flood susceptibility research: (1) a pixel-level flood inventory derived directly from Sentinel-1 SAR change detection, replacing the field-verified or administrative inventories used in most comparable studies; (2) a comparative evaluation of seven tree-based ensemble models with formal statistical testing of the differences among them; and (3) SHAP-based interpretation embedded in a fully cloud-based Google Earth Engine and Python pipeline. In contrast to the Jeddah reference study [1], which relied on an established flood inventory in an arid environment, the present workflow generates its own supervisory labels from SAR in a tropical coastal setting where persistent cloud cover disables optical flood mapping, and it additionally quantifies the spatial validity of its own performance estimates through spatial block cross-validation. This explicit coupling of SAR self-labelling, multi-model comparison, and explainability is the specific novelty claimed here.

Flood susceptibility mapping aims to identify locations that possess physical and environmental conditions favourable to inundation. This differs from event-based flood mapping because the target is not merely to reproduce a single observed flood extent, but to estimate the spatial tendency of flood-prone conditions from terrain, hydrological connectivity, rainfall, land cover, and urban surface characteristics. Such distinction is important for coastal urban areas, where flooding may arise from the combined effect of rainfall-runoff processes, low terrain, drainage congestion, permanent water proximity, tidal influence, and rapid land conversion [2]. Because flood susceptibility is governed by complex interactions among multiple environmental factors, conventional statistical approaches often struggle to capture nonlinear relationships [3]. Consequently, recent studies have increasingly adopted machine-learning techniques to improve predictive performance while enhancing model interpretability.

Recent studies have advanced flood susceptibility mapping through ensemble machine learning and explainable artificial intelligence. The Jeddah reference study [1], for example, combined multiple advanced tree-based learners with SHAP interpretation to model flood susceptibility in an arid urban environment. Other studies in metropolitan regions have similarly emphasized interpretable machine learning, including Random Forest [4], XGBoost [5], LightGBM [6], and SHAP-based explanations [7]. These studies show that predictive accuracy alone is insufficient; a susceptibility model must also explain which flood-conditioning factors govern spatial risk and whether the resulting map is physically plausible. However, the success of explainable machine-learning models ultimately depends on the availability of reliable spatial flood inventories for supervised training. Generating such inventories remains particularly challenging in tropical regions, where persistent cloud cover limits the effectiveness of optical satellite imagery during flood events.

In the Indonesian context, this limitation becomes even more pronounced. Although Indonesia maintains comprehensive national disaster databases through BNPB, these records are primarily event-based administrative reports rather than geospatial flood inventories [8]. Consequently, they cannot be directly used as pixel-level supervisory labels for flood susceptibility modelling. To address this limitation, Sentinel-1 SAR has become an important alternative because its microwave signals can penetrate cloud cover and detect flood-induced backscatter changes regardless of weather conditions [9]. Nevertheless, most Indonesian remote-sensing studies employ Sentinel-1 primarily for post-event flood delineation rather than transforming the resulting flood inventory into supervised labels for flood susceptibility modelling [10]. This leaves a methodological gap between flood detection and flood susceptibility inference, particularly for coastal urban areas in Indonesia where pixel-level ground truth is unavailable and multi-source geospatial predictors must substitute for field-verified inventories. The present study addresses this gap by treating the Sentinel-1-derived flood extent as a supervised label and embedding it within a comparative explainable machine-learning framework.

Addressing this methodological gap requires a study area where flood occurrence results from multiple interacting coastal and hydrological processes. The Semarang-Demak coastal area provides an appropriate natural laboratory because its low-lying coastal morphology, rapid urban expansion, drainage pressure, riverine influence, and permanent water bodies create multiple interacting flood-generating processes [11], [12]. These characteristics make the region suitable for evaluating whether Sentinel-1-derived flood inventories can effectively support explainable flood susceptibility modelling. Accordingly, this study addresses the following research questions: (1) how well do ensemble models discriminate flood and non-flood samples derived from Sentinel-1 change detection; (2) which geospatial factors contribute most to flood susceptibility; and (3) whether the spatial pattern of predicted susceptibility is consistent with the coastal-urban flood mechanism of Semarang-Demak.

The contribution of this study is not to claim a new algorithm, but to adapt an explainable ensemble-learning workflow to a data-limited Indonesian coastal setting. The novelty lies in integrating Sentinel-1-derived flood inventory, multi-source geospatial predictors, comparative ensemble modelling, and SHAP-based interpretation within a reproducible Google Earth Engine and Python workflow [13].

2. Study Area and Data

2.1. Study Area

The study area covers the Semarang-Demak coastal zone in Central Java, Indonesia. The region includes dense urban settlements, low-lying coastal plains, rivers, drainage channels, agricultural landscapes, fishponds, and permanent water features. These conditions generate a complex flood regime in which susceptibility is not controlled by rainfall alone. Instead, inundation may emerge from the interaction between antecedent precipitation, low elevation, limited drainage capacity, proximity to surface water, built-up imperviousness, and coastal hydrological influence [14].

Two additional coastal processes shape the flood regime of this area but are not represented in the present predictor set. First, the Semarang-Demak coastal plain has experienced rapid land subsidence, with measured rates of up to approximately 10–14 cm per year in the most affected districts [15], which progressively enlarges the area exposed to inundation. Second, tidal ("rob") flooding regularly inundates low-lying coastal neighbourhoods independently of rainfall, particularly when high tides coincide with reduced drainage capacity. Because both processes are time-varying and no consistent raster products are available for the event period, they were not included as predictors; Section 5.8 discusses how their absence conditions the interpretation of the model.

2.2. Data Sources and Predictor Variables

The model uses a multi-source geospatial dataset representing the major dimensions of flood-conditioning processes: topography, hydrology, rainfall, spectral surface condition, land cover, and built-up intensity. Sentinel-1 SAR was used to derive the binary flood inventory. SRTM provided elevation, slope, and aspect. MERIT Hydro represented HAND, upstream drainage area, river width, and derived distance to river. JRC Global Surface Water was used to derive permanent water and distance to permanent water. CHIRPS represented 7-day and 30-day antecedent rainfall. Sentinel-2 provided NDVI, NDWI, MNDWI, and NDBI, while ESA WorldCover supplied land-cover categories [16], [17], [18], [19], [20]. This variable design allows the model to connect flood occurrence with both static susceptibility factors and event-related rainfall conditions.

Table 1. Multi-source geospatial predictors used in the susceptibility model.

Data source	Derived variables	Analytical role
Sentinel-1 GRD	VV/VH SAR backscatter	Pre/post event flood inventory
SRTM DEM	Elevation, slope, aspect	Topographic control
MERIT Hydro	HAND, upstream area, river width, distance to river	Hydrological connectivity
JRC Global Surface Water	Permanent water, distance to water	Water proximity and inundation context
CHIRPS Daily	7-day and 30-day rainfall	Antecedent rainfall forcing
Sentinel-2 SR	NDVI, NDWI, MNDWI, NDBI	Vegetation, water signal, built-up intensity
ESA WorldCover	Land-cover class	Surface and land-use context

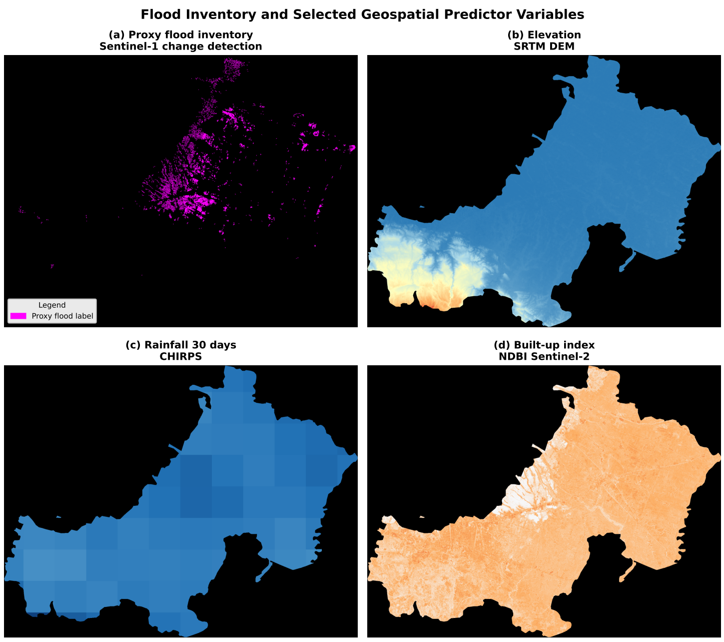


Figure 1. Sentinel-1-derived flood inventory and selected geospatial predictors. Source: processed from Google Earth Engine outputs

3. Methodology

3.1. Sentinel-1 Flood Inventory Generation

The target label was constructed from Sentinel-1 GRD VH-backscatter change detection [10] for the March 2024 flood event, using a pre-event composite (25 February – 12 March 2024) and a post-event composite (15 – 31 March 2024). Both composites are per-orbit median composites of all Interferometric Wide-swath acquisitions covering the study area within each window. Backscatter values in dB were converted to power scale, and the pre/post power ratio was computed so that a post-event backscatter drop yields a ratio greater than one.

A pixel was labelled as flood when *all* of the following conditions held: (1) the pre/post VH power ratio exceeded 1.25; (2) the post-event VH backscatter was below -18 dB, ensuring a genuinely water-like scattering regime rather than any small change; (3) the pixel was not permanent water, defined as a JRC Global Surface Water occurrence of at least 80%; (4) the SRTM slope was below $5°$, excluding steep surfaces that produce geometry-induced backscatter drops; and (5) the pixel belonged to a connected flood cluster of at least eight pixels at the 30 m sampling scale, which suppresses isolated speckle-induced false detections. Pixels satisfying these constraints were labelled as flood, while all other sampled pixels were labelled as non-flood. The combined threshold set represents a compromise between retaining plausible flood signals and limiting over-detection caused by water-like surfaces; the sensitivity of the resulting extent to the ratio threshold is quantified below.

A threshold sensitivity analysis showed that estimated flood extent declined from 56.11 km^2 at a ratio threshold of 1.15 to 44.07 km^2 at 1.50. This monotonic reduction indicates that the flood inventory is threshold-sensitive, as expected in SAR-based flood mapping. The selected threshold therefore represents a compromise between retaining plausible flood signals and limiting over-detection caused by water-like surfaces.

3.2. Sampling Strategy and Class Imbalance Handling

The raw stratified sample contained 3,935 observations, consisting of 3,000 non-flood and 935 flood samples. This imbalance emerged because the Semarang-Demak-only region contains fewer flood-labelled pixels than the surrounding-region ROI. Directly training and evaluating models on this imbalance could bias predictions toward the non-flood class and inflate accuracy while weakening flood recall.

To reduce majority-class dominance, the final modelling dataset was balanced through undersampling the non-flood class, resulting in 1,870 observations with 935 non-flood and 935 flood samples. This decision improves class comparability and makes recall, precision, and F1-score more meaningful for flood detection. However, it also reduces the diversity of non-flood samples. Therefore, the balanced dataset is interpreted as a modelling design for fair classification rather than a representation of natural flood prevalence in the landscape.

3.3. Ensemble Learning Models and Evaluation

Seven ensemble models were evaluated: Random Forest [4], Extra Trees, Gradient Boosting, XGBoost [5], LightGBM [6], CatBoost [21], and a stacking ensemble. Tree-based ensemble models were selected

Table 2. Class distribution before and after imbalance handling

Dataset	Non-flood	Flood	Total	Purpose
Raw sample	3000	935	3935	Preserves available sample distribution but is imbalanced
Balanced sample	935	935	1870	Used for final model comparison and SHAP interpretation

because flood susceptibility is expected to involve non-linear and threshold-like relationships between variables, such as abrupt increases in susceptibility near water bodies, low elevation, or high rainfall accumulation.

Table 3 reports the hyperparameter configuration of each model. The configuration follows common defaults for tabular geospatial classification with moderate depth regularization, and all models were trained with a fixed random seed (42) for reproducibility. The stacking ensemble combines Random Forest, Extra Trees, XGBoost, and LightGBM as base learners with a class-weighted logistic regression meta-learner using five-fold internal cross-validation. The optimized classification threshold for every model was determined using the Youden index from its ROC curve.

Table 3. Hyperparameter settings of the seven ensemble models (random seed = 42 for all models)

Model	Hyperparameters
Random Forest	400 trees; `min_samples_leaf` = 2; class weight = balanced
Extra Trees	400 trees; `min_samples_leaf` = 2; class weight = balanced
Gradient Boosting	250 stages; learning rate = 0.04; max depth = 3
XGBoost	350 rounds; learning rate = 0.04; max depth = 4; subsample = 0.85; column sample = 0.85; eval metric = logloss
LightGBM	350 rounds; learning rate = 0.04; 31 leaves; subsample = 0.85; column sample = 0.85; class weight = balanced
CatBoost	350 iterations; learning rate = 0.04; depth = 5; loss = Logloss; eval metric = AUC
Stacking ensemble	Base: Random Forest + Extra Trees + XGBoost + LightGBM; meta-learner: logistic regression (max iter = 2000, class weight = balanced); internal 5-fold CV

Model performance was evaluated using ROC-AUC, accuracy, precision, recall, and F1-score. In susceptibility mapping, recall is particularly important because false negatives may omit vulnerable zones. Nevertheless, precision is also relevant because excessive false positives may overstate areas requiring intervention. F1-score therefore provides a useful balance between omission and commission errors.

3.4. Data Splitting, Evaluation Protocol, and SHAP Data

The 1,870 balanced samples were divided once into a training set (75%; 1,402 samples) and an independent held-out test set (25%; 468 samples) using a stratified random split with a fixed seed. All seven models were trained exclusively on the training set, and all headline performance figures in Table 4 were computed on the held-out test set, which was not used in any stage of model fitting or tuning. Because the split is random rather than spatial, neighbouring pixels with highly similar environmental conditions can appear on both sides of the split; the spatial block cross-validation in Section 4.3 quantifies the effect of this design on generalization estimates. The SHAP analysis of the best model was computed on a random subsample of 1,000 rows drawn from the training set only, so no information from the held-out test set enters either the interpretation or the reported explanation values. The original training run was executed on Google Colab. For the additional analyses reported in this revised version (statistical model comparison, spatial block cross-validation, correlation diagnostics, and the SHAP dependence analysis), the full pipeline was re-run with the same fixed seed in two environments: a Kaggle environment (Python 3.12.13, scikit-learn 1.6.1, XGBoost 3.2.0, LightGBM 4.6.0, CatBoost 1.2.10, SHAP 0.51.0, NumPy 2.0.2, pandas 2.3.3) whose held-out test predictions reproduce six of the seven AUCs of Table 4 to six decimal places (the stacking ensemble differs by 0.002), and a local environment (Python 3.14.6, scikit-learn 1.9.0, XGBoost 3.4.1, LightGBM 4.7.0, CatBoost 1.2.10, SHAP 0.52.0, NumPy 2.5.2, pandas 3.0.5, SciPy 1.18.1). The statistical model comparison was computed on the predictions that reproduce Table 4 exactly, and all robustness diagnostics agree between the two environments within 0.006 ROC-AUC.

3.5. Statistical Comparison of Model Performance

Because the leading models differ by small ROC-AUC margins, pairwise differences against Random Forest were assessed with two complementary tests computed on the same held-out test predictions: DeLong's test for correlated ROC curves [22] and a paired bootstrap procedure with 10,000 resamples of the test set, reporting 95% confidence intervals of the AUC difference. Both tests evaluate whether the observed ranking reflects a systematic advantage or sampling variability.

3.6. Spatial Block Cross-Validation

To assess spatial generalization independently of the random split, all models were additionally evaluated with a five-fold spatial block cross-validation design. The study area was partitioned into approximately 10×10 km blocks (0.1° grid cells) defined from the geographic coordinates of each sample, and entire blocks were assigned to folds so that no block contributes samples to both training and evaluation within a fold. This design prevents the leakage of spatially autocorrelated information between neighbouring pixels across the train–test boundary and therefore provides a more conservative estimate of map generalization to unsampled areas.

3.7. SHAP Interpretation

SHAP was applied to the best-performing model to quantify the contribution of each predictor to model output. The interpretation was computed on a random subsample of 1,000 rows drawn from the training set only; the held-out test set played no role in the explanation. For binary classification, the SHAP values were interpreted for the flood class. This is methodologically important because using the raw multi-class SHAP array without selecting the positive class can produce misleading interaction-style plots. The corrected SHAP workflow therefore allows the analysis to infer not only which variables are important, but also how hydrological, rainfall, vegetation, and built-up factors shape the predicted flood susceptibility signal.

Table 4. Optimized model performance on the balanced dataset. Test-set metrics were computed on the 25% held-out split. CV AUC columns report mean ± SD over five-fold cross-validation with random stratified folds and with spatial block folds (∼10 km blocks), respectively

Model	ROC-AUC	Accuracy	Precision	Recall	F1-score	CV AUC (random)	CV AUC (spatial)
RandomForest	0.852	0.791	0.727	0.932	0.816	0.842 ± 0.019	0.593 ± 0.094
StackingEnsemble	0.850	0.793	0.747	0.885	0.810	0.841 ± 0.018	0.603 ± 0.094
CatBoost	0.849	0.786	0.743	0.876	0.804	0.835 ± 0.018	0.565 ± 0.079
XGBoost	0.842	0.778	0.720	0.910	0.804	0.825 ± 0.017	0.602 ± 0.081
LightGBM	0.836	0.782	0.734	0.885	0.802	0.824 ± 0.021	0.614 ± 0.089
ExtraTrees	0.835	0.778	0.766	0.799	0.782	0.831 ± 0.020	0.591 ± 0.082
GradientBoosting	0.831	0.767	0.736	0.833	0.782	0.816 ± 0.021	0.598 ± 0.072

4. Empirical Results

4.1. Model Performance

Random Forest produced the highest ROC-AUC (0.852) and the best F1-score (0.816) at the optimized threshold, with accuracy = 0.791, precision = 0.727, and recall = 0.932. This result indicates that Random Forest offered the strongest balance between discrimination and flood-class sensitivity in the balanced Semarang-Demak dataset.

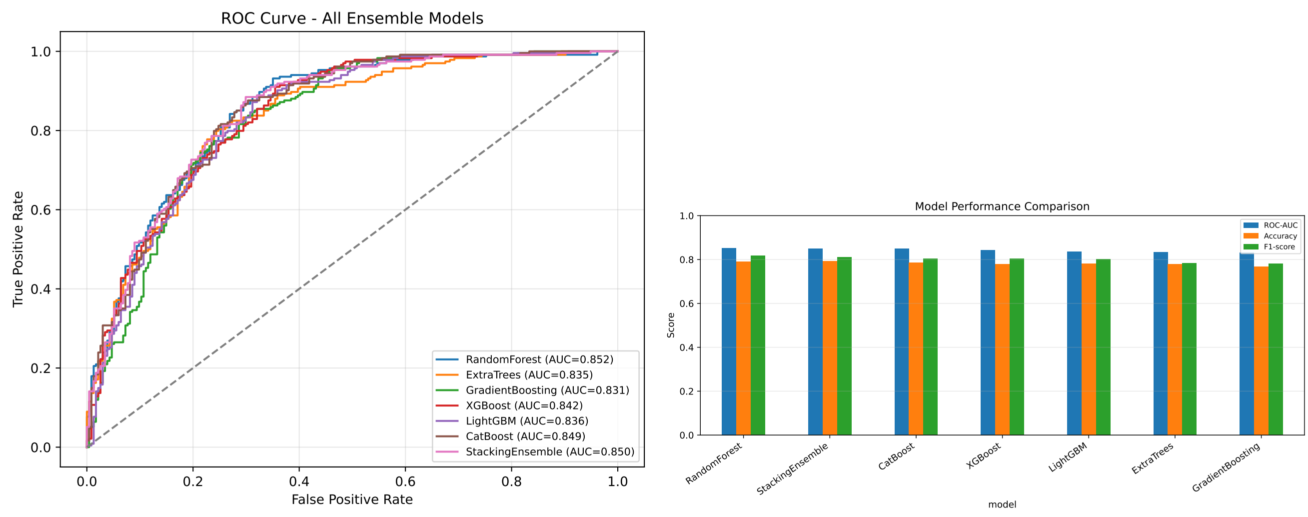
The margin between Random Forest, Stacking Ensemble, CatBoost, XGBoost, and LightGBM was not large. This suggests that the predictive signal is not unique to a single algorithm, but is consistently learnable by several tree-based ensembles. Such convergence strengthens the credibility of the geospatial predictor set: if only one algorithm performed well while the others failed, the result might indicate model-specific instability. Instead, the relatively close performance implies that the dataset contains a coherent flood-susceptibility structure.

The optimized Random Forest threshold was lower than 0.50, which is expected in a susceptibility problem where flood omission is costlier than moderate overprediction. The model prioritizes recall, meaning that more flood-prone samples are captured even at the expense of additional false positives. In risk-screening applications, this is generally preferable because areas falsely flagged as susceptible can be refined through field validation, while missed flood-prone areas may remain unmanaged.

4.2. Error Structure from Confusion Matrices

The confusion matrices show that model evaluation should not be reduced to overall accuracy. In flood susceptibility mapping, false negatives and false positives have asymmetric implications. False negatives indicate locations labelled as flood but predicted as non-flood, which may underestimate flood-prone areas. False positives indicate locations predicted as susceptible but labelled as non-flood; these may not always be purely erroneous because susceptibility refers to similarity with flood-prone conditions, not necessarily inundation during the selected event.

The high recall of Random Forest indicates that the model is conservative in capturing flood-prone samples. This behaviour is suitable for preliminary susceptibility screening, especially in coastal urban planning where underestimating potential hazard zones can be more consequential than flagging extra areas for further inspection. However, this also means that the resulting susceptibility map should be validated before being used for site-specific engineering decisions.



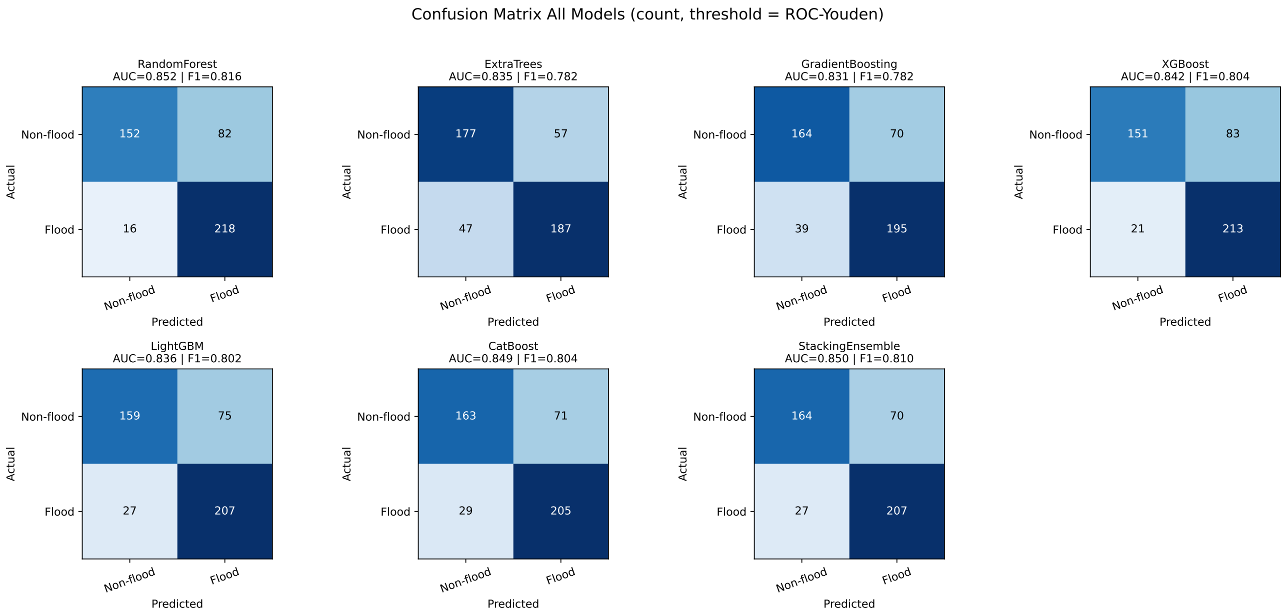
a). ROC curve **b).** Model performance summary

Figure 2. ROC curve and model performance summary. Source: model evaluation outputs

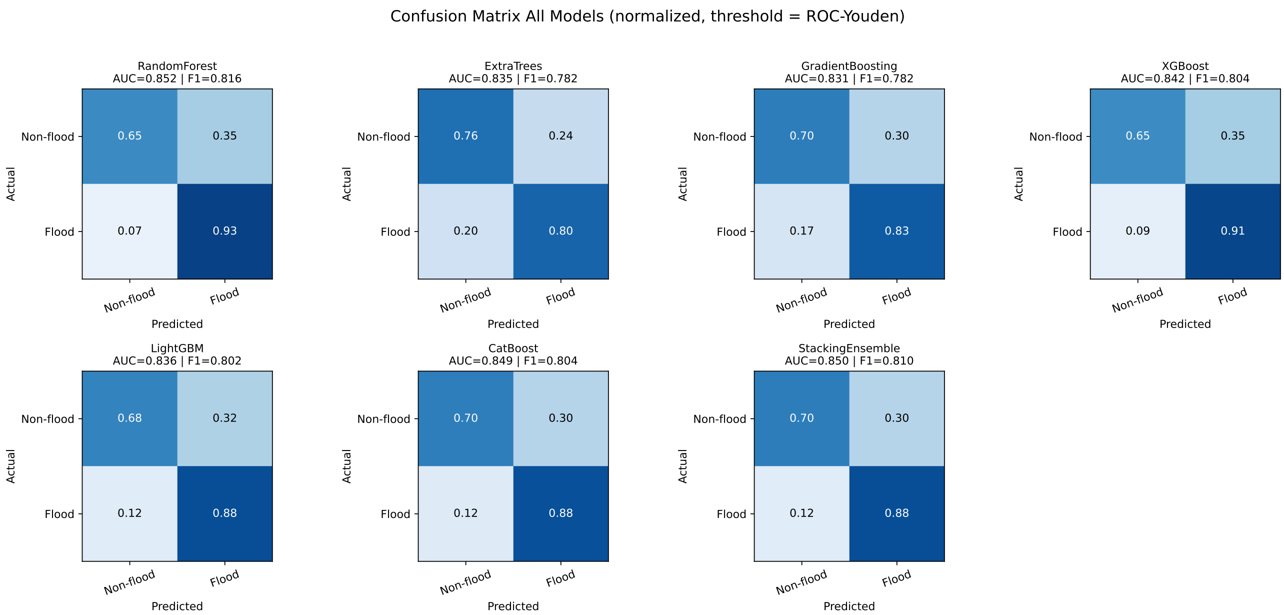
4.3. Statistical Model Comparison and Spatially Validated Performance

The DeLong tests and paired bootstrap comparisons on the held-out test predictions show that Random Forest is statistically indistinguishable from the other leading models. The differences against the stacking ensemble (ΔAUC $= 0.002$, $p = 0.652$), CatBoost (ΔAUC $= 0.003$, $p = 0.691$), XGBoost (ΔAUC $= 0.010$, $p = 0.174$), and LightGBM (ΔAUC $= 0.016$, $p = 0.072$) are all non-significant, and the 95% bootstrap confidence intervals of these differences all include zero (for example, $[-0.002, +0.033]$ against LightGBM). Only the comparisons against the two lowest-ranked models reach marginal significance: Extra Trees (ΔAUC $= 0.018$, $p = 0.034$) and Gradient Boosting (ΔAUC $= 0.022$, $p = 0.040$). Random Forest should accordingly be read as the model with the highest point estimate, statistically inseparable from the rest of the leading cluster under this configuration, with a small but detectable margin over the two weakest models.

The spatial block cross-validation results provide an important corrective perspective on the headline metrics. When entire $\sim$10 km blocks are held out, the five-fold cross-validated ROC-AUC of all models drops to the 0.56–0.61 range (Table 4, last column), compared with 0.82–0.84 under random stratified folds. This gap confirms that the random split benefits from spatial autocorrelation between neighbouring training and test pixels, and that the held-out estimates of Table 4 are optimistic with respect to generalization to geographically distinct areas. Three implications follow. First, the model ranking is stable across both validation designs, so the comparative conclusion is not an artefact of the splitting strategy. Second, the susceptibility map should be interpreted as an interpolation within the environmental and spatial domain represented by the training samples rather than a transferable predictor for other regions. Third, the persistent water-proximity signal that dominates the SHAP analysis partly reflects local spatial context, which reinforces the cautious interpretation of Section 5.2.



a). Confusion matrix counts



b). Normalized confusion matrices

Figure 3. Confusion matrices for all evaluated models. Source: model evaluation outputs. The normalized matrix shows class-wise error proportions

Table 5. Top SHAP-based predictor contributions for the flood class

Predictor	Mean |SHAP|
dist_perm_water	0.058
rain_30d	0.042
ndvi	0.036
rain_7d	0.031
upa_log	0.030
mndwi	0.026
ndbi	0.025
river_width	0.024
dist_river	0.024
ndwi	0.023
slope	0.020
elevation	0.013

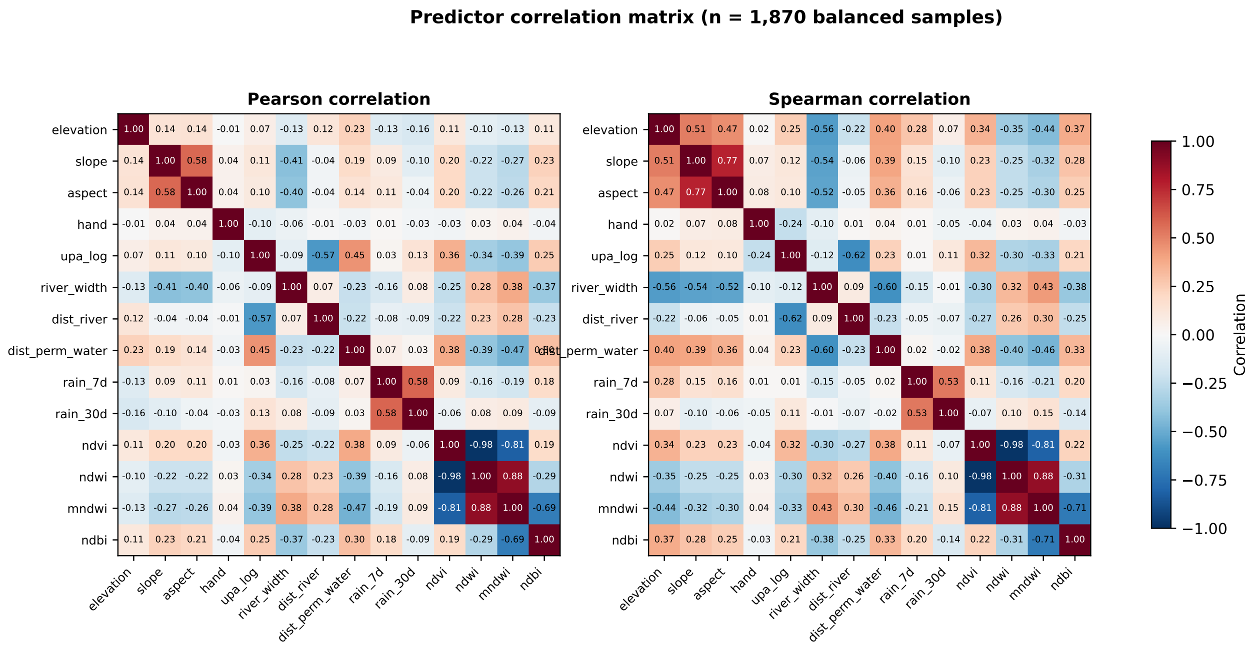


Figure 4. Pearson and Spearman correlation matrices of the continuous predictors ($n = 1{,}870$). The four Sentinel-2 spectral indices form a highly collinear block, as expected from their shared input bands

4.4. Feature Importance and SHAP Explanation

Because NDVI, NDWI, MNDWI, and NDBI are derived from overlapping Sentinel-2 bands, their pairwise Pearson correlations were inspected (Figure 4). The indices are indeed strongly collinear: NDVI and NDWI correlate at $r = -0.98$, NDWI and MNDWI at $r = 0.88$, and MNDWI and NDBI at $r = -0.69$, whereas NDBI and NDVI are only weakly correlated ($r = 0.19$). This collinearity is inconsequential for the predictive models, because tree-based ensembles are insensitive to linear dependence among predictors and the goal is discrimination rather than coefficient estimation. For the SHAP ranking, however, it implies that importance is shared among correlated indices and that individual ranks within the spectral-index block should not be over-interpreted; the aggregated signal of the index block is meaningful, but the ordering of, for example, MNDWI versus NDWI is not identifiable. The remaining non-spectral predictors are only moderately correlated (maximum $|r| = 0.58$, between 7-day and 30-day antecedent rainfall), so their SHAP ranks are more directly interpretable.

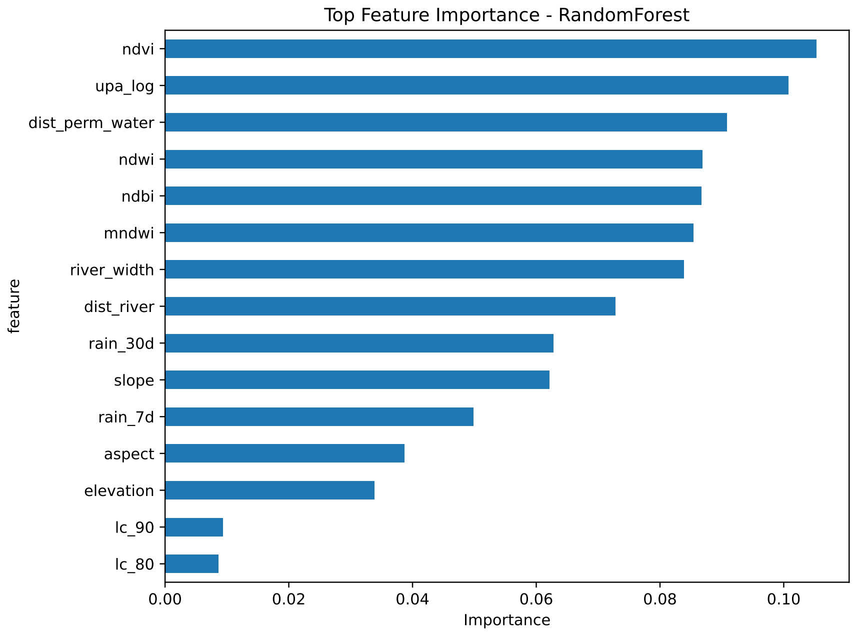
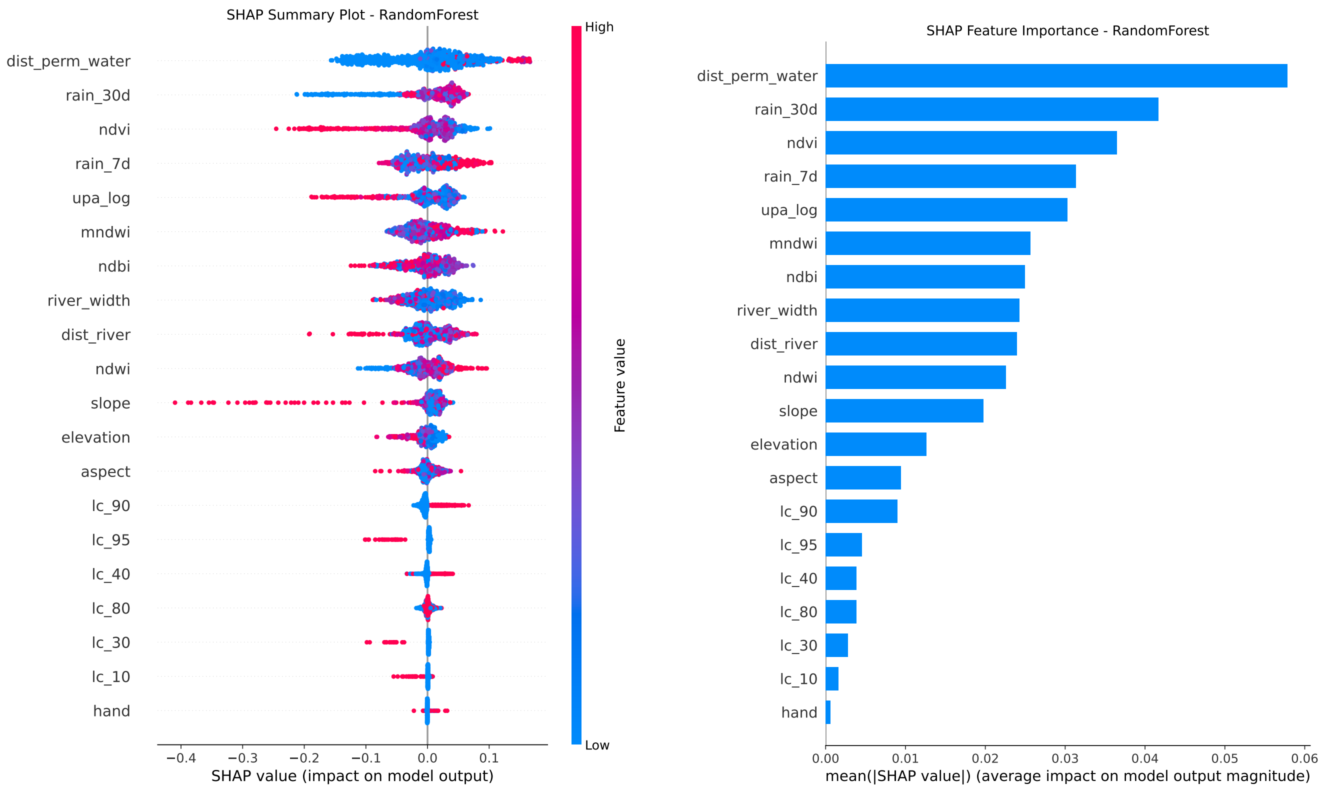


Figure 5. Random Forest feature importance. Source: Python model interpretation outputs



a). SHAP summary plot for the positive flood class

b). SHAP-based predictor contribution ranking for the flood class

Figure 6. SHAP-based interpretation of the flood classification model

The Random Forest feature importance and SHAP results jointly indicate that flood susceptibility in the Semarang-Demak coastal area is governed by interacting hydrological proximity, rainfall forcing, vegetation condition, and built-up intensity. Random Forest impurity-based importance ranked NDVI, upstream drainage area, distance to permanent water, NDWI, NDBI, MNDWI, river width, distance to river, rainfall, and slope among the key variables. The SHAP ranking refined this interpretation by identifying distance to permanent water, 30-day rainfall, NDVI, 7-day rainfall, upstream drainage area, MNDWI, NDBI, river width, distance to river, and NDWI as the strongest contributors to prediction.

The dominance of distance to permanent water is physically plausible. In a coastal lowland, locations near permanent water bodies, rivers, ponds, drainage channels, and wet surfaces often share hydrological connectivity with flood pathways. The importance of antecedent rainfall reflects the role of accumulated precipitation in increasing runoff potential, saturating surfaces, and loading drainage systems. The presence of both 7-day and 30-day rainfall among the influential predictors suggests that the model captures both short-term event forcing and antecedent wetness conditions.

Vegetation and water indices provide additional surface-state information. NDVI separates vegetated or agricultural areas from built-up and bare surfaces, and its SHAP contribution is directionally consistent with susceptibility being higher in low-vegetation, moisture-laden, or disturbed zones. NDWI and MNDWI capture open-water and moisture signals, with higher index values producing positive SHAP contributions that push predictions toward the flood class. NDBI reflects built-up intensity; its positive mean |SHAP| implies that denser urban surfaces, where imperviousness limits infiltration, are associated with higher model-predicted susceptibility.

Hydrological variables such as upstream drainage area, river width, and distance to river reinforce the water-proximity argument at the channel scale. Locations with larger upstream contributing areas or narrower separation from river channels consistently produce positive SHAP values, consistent with the physical expectation that drainage convergence amplifies inundation potential. These variables should be interpreted as derived predictors from global hydrological datasets rather than field-surveyed drainage infrastructure; their contribution nevertheless supports the expectation that susceptibility is higher where terrain and drainage structures concentrate water.

Notably, `elevation` and `slope` rank last in SHAP contribution, with mean |SHAP| values of 0.013 and 0.020 respectively. In mountainous or hilly study areas, these variables typically dominate flood susceptibility models because terrain controls runoff pathways directly. Their low rank here is an informative finding for a coastal lowland: when topographic relief is minimal and confined within a narrow elevation range, the discriminating signal shifts from terrain structure to water proximity and rainfall accumulation. The SHAP rankings therefore reflect the physical character of the Semarang-Demak setting rather than a deficiency of the predictor set.

4.5. Spatial Susceptibility Pattern

The final susceptibility map indicates that higher susceptibility classes are concentrated in zones that combine water proximity, lowland morphology, wet surface signals, and built-up pressure. This pattern is consistent with the coastal-urban flood mechanism of Semarang-Demak. Areas close to permanent water or drainage features can receive concentrated runoff or experience limited drainage capacity, while built-up zones may increase impervious surfaces and reduce infiltration.

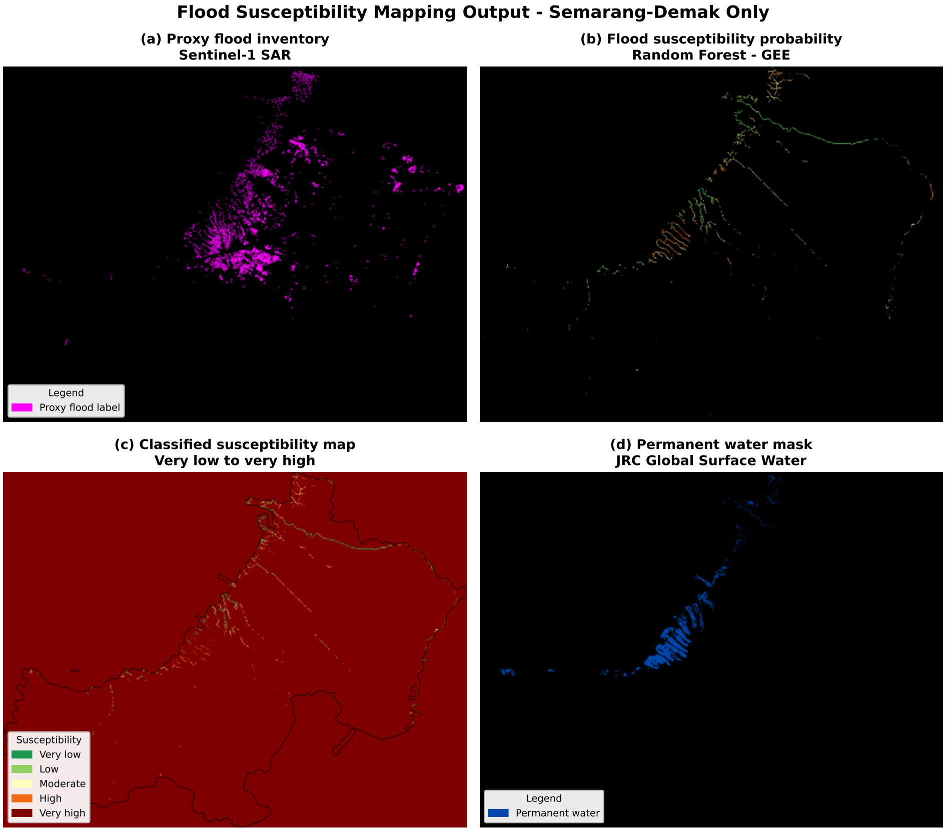


Figure 7. Final flood susceptibility mapping output for the Semarang-Demak coastal area. Source: Random Forest raster implementation in Google Earth Engine

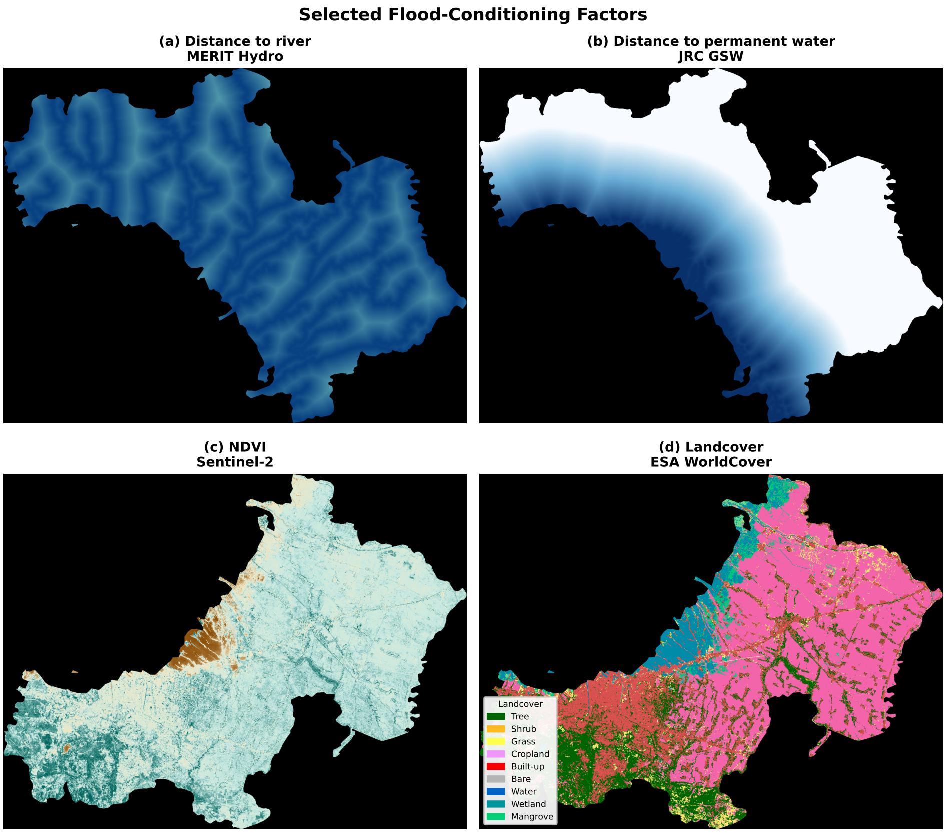


Figure 8. Selected flood-conditioning factors used to support spatial interpretation. Source: processed geospatial predictors from Google Earth Engine

The map should not be read as a real-time forecast. It represents spatial similarity to previously detected flooded pixels under the selected predictor conditions. Therefore, a high-susceptibility pixel does not necessarily mean that flooding occurred at that exact location on the selected date. Instead, it indicates that the location has geospatial characteristics resembling flood-labelled samples. This distinction is central to susceptibility mapping and prevents overinterpretation of the probability values as event-specific inundation certainty.

The classified map is useful for communicating broad susceptibility gradients, but the continuous probability map retains more analytical information. The five-class map can support planning communication, while the probability surface is more appropriate for technical prioritization because it preserves subtle spatial differences.

5. Discussion

5.1. Comparative Performance of the Ensemble Models

The comparative results show that all seven ensemble models were able to discriminate between flood and non-flood samples derived from the Sentinel-1-based flood inventory. Random Forest achieved the highest ROC-AUC (0.852) and F1-score (0.816), followed by the Stacking Ensemble (ROC-AUC = 0.850; F1-score = 0.810) and CatBoost (ROC-AUC = 0.849; F1-score = 0.804). XGBoost, LightGBM, ExtraTrees, and Gradient Boosting produced slightly lower ROC-AUC values of 0.842, 0.836, 0.835, and 0.831, respectively. The substantial overlap among the ROC curves further indicates that the predictive behaviour of the evaluated models was broadly similar.

Although Random Forest obtained the highest values for the principal discrimination metrics, the differences among the leading models were relatively small. Therefore, the results should not be interpreted as evidence that Random Forest is universally superior to other ensemble approaches. Instead, its performance indicates a marginal advantage under the present predictor configuration, sampling design, and Sentinel-1-derived flood labels. The statistical comparisons in Section 4.3 formalize this reading: Random Forest is statistically inseparable from the stacking ensemble, CatBoost, XGBoost, and LightGBM, and holds only a small margin over the two lowest-ranked models, so the model ranking should be treated as descriptive rather than decisive.

The normalized confusion matrices provide additional insight into this difference. Random Forest correctly classified approximately 93% of the flood-labelled test samples, resulting in the highest flood recall among the evaluated models, although this was accompanied by a non-flood recall of approximately 65%. Consequently, its practical advantage is more appropriately interpreted as a relatively stronger ability to identify flood-labelled samples rather than as substantial overall superiority. This distinction is important because the choice of the best-performing model may depend on whether the application prioritizes sensitivity to flood-labelled locations or balanced discrimination between flood and non-flood classes.

The relatively close performance of the ensemble models also provides an important basis for interpreting the subsequent predictor analysis. The convergence of ROC-AUC values suggests that the principal spatial signal contained in the multi-source predictor dataset can be learned by different tree-based ensemble algorithms. Accordingly, the SHAP analysis is not interpreted as evidence that the environmental relationships are unique to Random Forest, but rather as an examination of the environmental information that contributes to the predictions of the selected model. Nevertheless, model performance and predictor importance remain conditional on the present dataset and should not be generalized directly to other flood environments without independent validation.

5.2. Dominant Role of Permanent-Water Proximity and Antecedent Rainfall

The SHAP analysis identifies distance to permanent water (`dist_perm_water`) as the most influential predictor, with the largest mean absolute SHAP value (approximately 0.058). Derived from JRC Global Surface Water, this variable represents the spatial relationship between locations and persistent surface-water features. Its high contribution indicates that the spatial configuration of persistent water contains substantial information for distinguishing flood-labelled from non-flood samples. This interpretation is physically plausible in the Semarang-Demak coastal environment, where rivers, drainage features, ponds, and other persistent water surfaces form interconnected components of a low-lying hydrological landscape. Previous flood susceptibility studies have likewise identified distance or spatial relationships to water bodies as important predictors of flood-prone conditions [23], [24], [25].

The direction of the SHAP relationship requires particular caution. Higher values of `dist_perm_water` are generally associated with positive SHAP values, whereas lower values tend to occur toward the negative side of the SHAP axis. Therefore, the result does not support a simple interpretation that locations closer to permanent water are necessarily more susceptible to flooding.

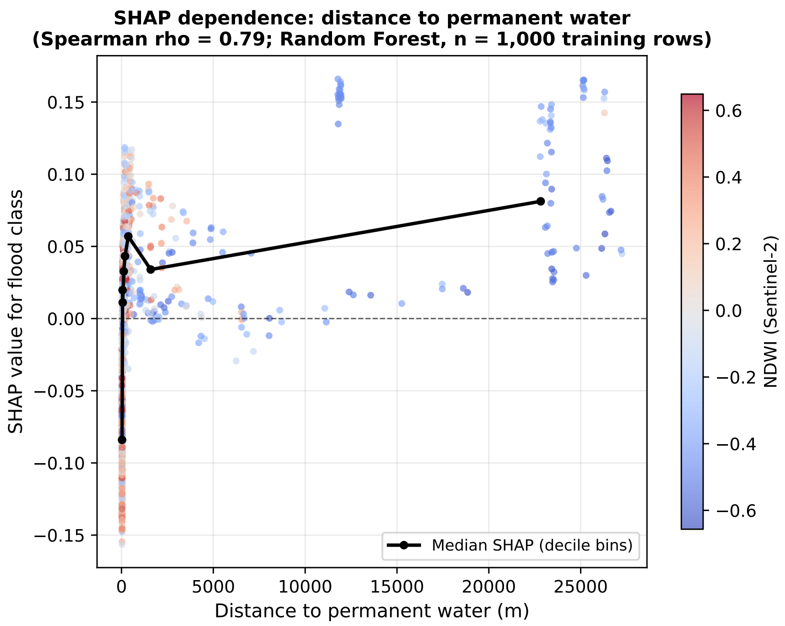


Figure 9. SHAP dependence plot for distance to permanent water (Random Forest, 1,000 training rows). Points are individual samples coloured by NDWI; the black line and band show the decile-binned median and interquartile range of the SHAP value. The relationship is strongly non-monotonic: samples within the first distance bin (median $\approx$ 30 m) receive clearly negative SHAP values, while all farther bins shift toward positive contributions

The dependence plot in Figure 9 makes this pattern explicit. The binned median SHAP value is negative (approximately -0.08) for the closest decile at roughly 30 m from permanent water, rises steeply across the next bins to approximately $+0.01$ to $+0.06$ at distances of 60 m to 370 m, and remains positive but more dispersed at larger distances, with an overall Spearman correlation of $\rho = 0.79$ between the variable and its SHAP value. The direction is therefore the opposite of the naive distance-decay expectation, and the most plausible explanation lies in how the training labels were constructed. Permanent water is explicitly excluded from the flood label, and the sampling grid starts at the edge of the permanent-water mask; the pixels nearest to permanent water are therefore frequently narrow shoreline bands that remain unlabelled even when occasionally inundated, so the model learns to treat the immediate shoreline fringe as relatively less flood-like than the low-lying ground slightly farther away. In addition, the negative-SHAP closest bin coincides with higher NDWI values (Figure 9, colour scale), indicating persistent surface moisture rather than event inundation. The high SHAP importance of `dist_perm_water` hence indicates that the spatial arrangement of persistent water features provides contextual information for the classification. This apparently non-monotonic or counter-intuitive relationship may reflect the fact that persistent water features are spatially associated with broader drainage, pond, coastal, and land-use configurations rather than functioning solely as direct sources of inundation. It may also partly reflect the spatial structure of the Sentinel-1-derived flood inventory itself. Thus, the predictor should be interpreted as a spatial contextual variable rather than as direct evidence of channel overtopping or a causal effect of water proximity.

Antecedent rainfall was the second major source of predictive information. The 30-day rainfall variable (`rain_30d`) ranked second, with a mean absolute SHAP value of approximately 0.042, whereas 7-day rainfall (`rain_7d`) ranked fourth at approximately 0.031. Higher values of both variables generally occur toward positive SHAP values, indicating that greater accumulated rainfall tends to be associated with predictions of the flood class. This relationship is physically plausible because accumulated precipitation can contribute to antecedent wetness and hydrological loading, potentially reducing the capacity of the landscape to accommodate additional rainfall and increasing the likelihood of surface water accumulation [9], [26]. Importantly, this interpretation represents a plausible physical mechanism rather than a directly measured process, because the present model does not explicitly measure groundwater level, soil moisture, or infiltration capacity.

The stronger contribution of 30-day rainfall compared with 7-day rainfall is particularly noteworthy. The result suggests that precipitation accumulated over a longer antecedent period contains more discriminatory information than rainfall over the shorter temporal window considered here. This may indicate that pre-event hydrological conditions are relevant to the spatial distribution of the flood-labelled samples, rather than the observed flood pattern being associated solely with immediate rainfall forcing. Similar emphasis on antecedent precipitation or cumulative rainfall has been reported in flood susceptibility and flood modelling studies [9].

Nevertheless, the present result should not be interpreted as evidence that long-term rainfall accumulation is universally more important than short-duration rainfall. In Semarang-Demak, where flooding may arise through interacting rainfall, drainage, riverine, and coastal processes, the relative contribution of rainfall accumulation is likely to depend on the specific flood event and its antecedent conditions.

5.3. Contribution of Surface Characteristics from Sentinel-2

NDVI was the third most influential predictor, with a mean absolute SHAP value of approximately 0.036. Higher NDVI values are predominantly associated with negative SHAP contributions, whereas lower values more frequently occur toward positive SHAP values. This indicates that locations characterized by lower vegetation-index values tend to be associated with higher predicted flood susceptibility within the present model. A plausible physical interpretation is that vegetation-related surface conditions can differentiate areas with greater vegetation cover from more intensively developed, sparsely vegetated, or water-influenced surfaces. Vegetated surfaces can enhance infiltration and influence runoff processes, whereas impervious or sparsely vegetated surfaces may permit greater generation of surface runoff [27].

The observed negative contribution of high NDVI is broadly consistent with previous flood susceptibility studies reporting that vegetation-related indices can contribute to the discrimination of flood-prone areas [24]. However, NDVI should not be interpreted as a direct causal determinant of flooding. The index is a remotely sensed proxy for vegetation-related surface characteristics and can also vary with land-cover type, seasonal vegetation condition, moisture, and other environmental factors. Consequently, its SHAP contribution is better interpreted as evidence that vegetation-related surface conditions contain useful spatial information for distinguishing the Sentinel-1-derived flood labels. In Semarang-Demak, this information is particularly relevant because agricultural and vegetated surfaces coexist with expanding built-up areas and water-influenced coastal landscapes.

The water-sensitive indices derived from Sentinel-2 also contributed meaningfully to the model. MNDWI and NDWI obtained mean absolute SHAP values of approximately 0.026 and 0.023, respectively, and higher values generally occurred toward positive SHAP contributions. This behaviour is consistent with the role of these indices in representing water-related and moisture-sensitive surface characteristics. Higher index values are generally associated with stronger indications of surface-water presence, while MNDWI can improve the discrimination of water features from built-up and vegetated surfaces [28]. This spectral sensitivity may therefore provide additional information for distinguishing flood-labelled from non-flood locations. Nevertheless, the indices should not be interpreted as direct measurements of flood occurrence because their values can also reflect persistent surface water and other water-influenced land-cover conditions.

The contribution of these Sentinel-2 indices is particularly relevant because the flood inventory was derived from Sentinel-1 SAR. Their predictive contribution therefore provides complementary optical surface information associated with the spatial distribution of the SAR-derived flood labels. This does not constitute independent validation of the Sentinel-1 inventory, because the Sentinel-2 variables are predictors within the same supervised modelling framework. Rather, the result indicates that the flood-labelled spatial pattern contains information that is also reflected in optical characteristics of the surface.

NDBI produced a mean absolute SHAP value of approximately 0.025, indicating a meaningful contribution to the model predictions. However, the SHAP distribution does not show a clear monotonic relationship between NDBI values and SHAP contributions, as both relatively low and high NDBI values occur across positive and negative SHAP values. This pattern suggests that the contribution of built-up intensity is not uniform across the study area and may be context-dependent. Because NDBI provides information related to built-up and impervious surface characteristics, variations in NDBI may reflect differences in surface conditions associated with infiltration and runoff processes [29], [30]. In the Semarang-Demak setting, this relationship is particularly relevant because developed areas coexist with low-lying coastal terrain and interconnected drainage systems. Nevertheless, NDBI should be interpreted as a proxy for built-up surface characteristics rather than as direct evidence that urbanization alone caused the observed inundation.

5.4. Hydrological Connectivity and Upstream Drainage Conditions

The hydrological variables derived from MERIT Hydro provided another important component of the predictive signal. Upstream drainage area (`upa_log`) ranked fifth, with a mean absolute SHAP value of approximately 0.030, while river width and distance to river each had mean absolute SHAP values of approximately 0.024. These variables represent different dimensions of hydrological structure: upstream drainage area describes the contributing area of the drainage network, whereas river width and distance to river represent the spatial relationship between locations and river channels.

Importantly, the SHAP directions were not uniform across these variables. Higher `upa_log` values were predominantly associated with negative SHAP values, while lower values tended to be associated with positive contributions. Similarly, higher river-width values were more frequently distributed toward negative SHAP values, whereas larger distance-to-river values tended to occur more frequently on the positive side of the SHAP axis. These patterns indicate that the model did not learn a simple relationship in which greater hydrological connectivity uniformly corresponds to greater flood susceptibility. Instead, the contribution of hydrological variables appears to depend on their spatial configuration and interaction with other environmental conditions.

This finding is consistent with the understanding that flood susceptibility is not determined solely by distance from a river. Previous studies have emphasized the roles of drainage configuration, contributing area, channel characteristics, and the interaction between surface conditions and hydrological pathways in determining inundation patterns [1], [23], [24], [26]. In a coastal environment, water may be redistributed through interconnected rivers, drainage channels, ponds, and low-lying surfaces, meaning that the effect of a hydrological variable can vary spatially. Therefore, the combined contribution of `upa_log`, river width, and distance to river is more informative than interpreting any one of these variables in isolation.

The relatively high contribution of `upa_log` compared with elevation and slope further suggests that hydrological structure provides predictive information beyond simple terrain descriptors. One possible explanation is that the drainage network captures aspects of water connectivity that are not fully represented by elevation or slope alone. However, this interpretation remains inferential because the present predictor set does not explicitly represent hydraulic infrastructure such as embankments, flood gates, pumping systems, or drainage capacity. The observed SHAP patterns should therefore be interpreted as evidence of spatial association with hydrological structure rather than direct evidence of a particular hydraulic mechanism.

5.5. Topography and the Characteristics of a Low-Lying Coastal Environment

Slope and elevation were less influential than permanent-water proximity, antecedent rainfall, NDVI, and the principal hydrological variables, with mean absolute SHAP values of approximately 0.020 and 0.013, respectively. Aspect had an even smaller contribution. These results do not indicate that topography is physically unimportant for flooding. Rather, they suggest that, within the environmental and spatial conditions represented by the present Semarang-Demak dataset, simple terrain descriptors provide less discriminatory information than several variables describing water-related conditions and surface characteristics.

The SHAP directions are nevertheless physically plausible: higher slope and elevation values generally tend to contribute negatively, whereas lower values are more frequently associated with positive contributions. Lower and flatter locations can provide greater potential for water accumulation, and this relationship has been widely recognized in flood susceptibility modelling [24], [26]. However, the relatively modest SHAP magnitudes indicate that elevation and slope alone do not adequately distinguish the flood-labelled locations in the present case.

One possible explanation is the low-relief character of the Semarang-Demak coastal zone. When the study area is dominated by relatively low-lying terrain, the spatial variability of elevation and slope may provide less discriminatory information than variables that capture water proximity, rainfall accumulation, and hydrological connectivity. In this context, a lower SHAP importance for elevation does not imply that elevation is irrelevant to flood processes. Instead, it indicates that its incremental predictive contribution is comparatively limited within this particular study area and predictor configuration. This distinction is important when comparing the present results with studies conducted in areas characterized by larger topographic gradients, where elevation and slope may have greater discriminatory power [23], [24].

5.6. Land-Cover Variables and Categorical Predictor Interpretation

The ESA WorldCover land-cover variables contributed less strongly than the principal continuous environmental predictors. Among the land-cover variables, `lc_90` showed the largest mean absolute SHAP contribution, followed by `lc_95`, `lc_40`, `lc_80`, `lc_30`, and `lc_10`, while HAND produced the smallest contribution in the SHAP ranking. These results indicate that the model relied more strongly on continuous measures of water proximity, rainfall, surface condition, and hydrological structure than on the individual land-cover categories.

The categorical nature of the WorldCover variables requires particular caution in interpreting the SHAP results. Unlike rainfall, elevation, or distance, the numerical class identifiers do not represent an ordered physical gradient. Therefore, a positive or negative SHAP value associated with a land-cover category should not be interpreted as evidence that increasing the numerical class value increases or decreases flood susceptibility. Instead, the land-cover variables should be interpreted in terms of how much information each encoded category contributes to model discrimination. This approach avoids assigning an artificial ordinal meaning to categorical land-cover classes.

The relatively smaller contribution of the land-cover variables also does not imply that land cover is irrelevant to flooding. Rather, the result suggests that much of the spatial differentiation captured by the model can be represented more directly through continuous environmental gradients such as water proximity, rainfall accumulation, vegetation characteristics, built-up intensity, and hydrological structure. In Semarang-Demak, where different land-cover classes occur within a strongly interconnected coastal environment, these continuous variables may capture spatial contrasts that are not fully represented by categorical land-cover labels alone.

5.7. Spatial Interpretation of the Predicted Susceptibility

The final susceptibility map generated using Random Forest provides a spatial representation of the environmental relationships learned from the multi-source predictor dataset. When interpreted together with the Sentinel-1-derived flood inventory and the permanent-water information, the map should be understood as a generalization of the environmental characteristics associated with flood-labelled pixels to the broader study area. It is therefore not a direct reproduction of the observed Sentinel-1 flood extent.

The spatial pattern is consistent with the SHAP interpretation in that the model integrates information from permanent-water proximity, antecedent rainfall, hydrological structure, vegetation-related surface conditions, and built-up intensity. The resulting susceptibility surface therefore represents the combined contribution of multiple environmental dimensions rather than the influence of a single conditioning factor. This is particularly useful for a coastal region such as Semarang-Demak, where flood-related processes can vary spatially across urban, agricultural, drainage, riverine, and coastal environments.

Nevertheless, the susceptibility map should be interpreted as a relative representation of environmental similarity to the flood-labelled samples rather than as a deterministic flood forecast. A high susceptibility value indicates that the environmental conditions at a location resemble those associated with the Sentinel-1-derived flood labels. It does not mean that the location must be inundated during every rainfall event, nor does the predicted value represent a directly measured flood probability, depth, or duration. This distinction is particularly important because the supervisory target is a remotely sensed proxy for inundation rather than a field-verified measurement of hydrodynamic flood characteristics.

5.8. Implications of Excluding Land Subsidence and Tidal Dynamics

The present predictor set deliberately represents rainfall-runoff and hydrological-proximity mechanisms, and it does not include land subsidence or tidal stage. This omission is a practical necessity rather than a judgement that these processes are unimportant. No consistent, event-period raster products exist for either process at the 30 m modelling scale: land-subsidence estimates for Semarang-Demak derive from intermittent PS-InSAR campaigns [15], and tidal conditions vary hourly and would require in-situ gauge records or tide models that cannot currently be sampled as per-pixel features inside Google Earth Engine. Because the susceptibility target is a single-event inventory, static or slowly varying proxies for these processes would also be difficult to interpret unambiguously.

Their absence nevertheless conditions what the model can and cannot express. Land subsidence and tidal flooding in Semarang-Demak operate largely through the same pathways the model does capture: they raise low-lying ground toward the water table, expand the footprint of permanent and semi-permanent water, and intensify drainage congestion. Part of their influence is therefore absorbed, in a time-averaged sense, by the variables the model does observe, notably elevation, HAND, distance to permanent water, and the coastal land-cover classes. What the model cannot express is the temporal dynamics: it cannot distinguish a location that floods because of heavy antecedent rainfall from one flooded primarily by a spring tide, nor can it project forward as subsidence progressively enlarges the exposed area. Consequently, the susceptibility map should be read as a rainfall-runoff-and-proximity assessment of present-day conditions, not as a total coastal-hazard map. Areas subject to rapid subsidence or frequent rob flooding may be underrepresented if their surface characteristics resemble those of non-subsiding inland areas, and the map should be intersected with subsidence-rate and tidal-inundation layers before being used for coastal-zone planning or adaptation decisions.

5.9. Implications and Limitations of the Sentinel-1-Based Supervised Framework

The results demonstrate the potential of combining a Sentinel-1-derived flood inventory with multi-source geospatial predictors for supervised flood susceptibility modelling in a data-limited coastal setting. The framework provides a way to transform an event-derived flood inventory into a spatial learning target and relate the resulting flood-labelled pattern to environmental variables that extend beyond the observed inundation footprint. The relatively similar performance of the evaluated ensemble models further suggests that the underlying spatial signal is sufficiently consistent to be learned by multiple tree-based approaches.

At the same time, the findings should not be interpreted as establishing a universally optimal model or a universal hierarchy of flood-conditioning factors. Random Forest performed marginally better than the other models for the present dataset, but the difference in ROC-AUC among the leading models was less than 0.01. Likewise, the SHAP ranking represents the relative contribution of predictors within this particular modelling framework and study area. These rankings may change under different flood events, predictor configurations, sampling strategies, or environmental settings.

Several limitations should also be considered. First, the Sentinel-1-derived flood inventory is a remotely sensed proxy and may contain uncertainties associated with SAR backscatter behaviour, surface scattering, permanent-water masking, and the selected change-detection threshold. Such uncertainties can propagate into the supervised learning stage. Consequently, the susceptibility relationships learned by the models may partly reflect characteristics of the particular flood event used to construct the supervisory labels rather than stable relationships applicable to all flood events.

Second, because the observations are spatially distributed, the random train-test split benefits from spatial autocorrelation between neighbouring samples. The spatial block cross-validation performed in this study quantifies this effect explicitly: when entire $\sim$10 km blocks are held out, cross-validated ROC-AUC falls from 0.82–0.84 to 0.56–0.61 across all models (Section 4.3). The headline metrics in Table 4 should therefore be read as within-domain discrimination estimates, while the spatially validated figures represent a more conservative bound on generalization to geographically distinct locations. Future work should extend this design to multi-event inventories, where blocks from different flood events can serve as genuinely independent evaluation units.

Third, several important coastal flood processes are not explicitly represented in the current predictor set. Tidal conditions, land subsidence, hydraulic infrastructure, and other dynamic coastal processes can influence inundation patterns in Semarang-Demak but are not directly captured by the present framework. Previous research has highlighted the importance of land subsidence in the coastal Semarang-Demak environment [15]. Incorporating such dynamic variables in future multi-event studies could provide a more complete representation of the physical processes underlying coastal flooding.

Future research should therefore combine multi-event Sentinel-1 flood inventories with spatially independent validation across events and additional coastal predictors, including tidal conditions and land-subsidence information. Such extensions would allow researchers to test whether the predictor relationships identified in the present study remain stable across different flood events and environmental conditions. They could also help distinguish relationships that are specific to the spatial structure of a particular flood inventory from those that represent more transferable flood-susceptibility characteristics. In this way, the Sentinel-1-based supervised framework can serve as a basis for progressively more generalizable flood susceptibility assessment in Indonesian coastal environments.

6. Conclusion

The findings demonstrate that flood susceptibility in the Semarang-Demak coastal area is primarily associated with the spatial and environmental characteristics of water-related conditions, antecedent rainfall, vegetation, drainage structure, and developed surfaces. SHAP analysis identified distance to permanent water as the most influential predictor, followed by antecedent rainfall, NDVI, upstream drainage area, MNDWI, NDBI, river width, distance to river, and NDWI. These results indicate that the spatial pattern of flood-labelled locations is not governed by a single conditioning factor, but rather reflects the combined information contained in water proximity, rainfall accumulation, hydrological connectivity, surface characteristics, and built-up conditions. Importantly, the SHAP results represent predictor contributions to the model predictions and should not be interpreted as direct evidence of causal relationships.

Among the evaluated ensemble models, Random Forest achieved the strongest overall performance, with a ROC-AUC of 0.852 and an F1-score of 0.816. However, DeLong tests and paired bootstrap comparisons showed that Random Forest is statistically inseparable from the other leading models (all $p \geq 0.07$), with only marginal separation from the two lowest-ranked models, so it should be regarded as the best point estimate under the present configuration rather than a statistically distinguishable winner. Five-fold spatial block cross-validation additionally showed that random-split estimates are optimistic for spatial generalization, with cross-validated ROC-AUC falling to the 0.56–0.61 range when $\sim$10 km blocks are held out. The model's relatively high flood recall further indicates its ability to identify flood-labelled samples, although this was accompanied by a lower recall for the non-flood class.

The resulting susceptibility map shows that higher susceptibility is concentrated in coastal zones characterized by combinations of water-related conditions, lowland morphology, wet surface signals, and built-up pressure. The map therefore provides a spatial screening representation of environmental similarity to the Sentinel-1-derived flood-labelled samples rather than a deterministic prediction of inundation during future rainfall events. Its primary value is to support the identification and prioritization of areas requiring further flood-risk assessment and management.

Overall, this study demonstrates the potential of transforming Sentinel-1-derived flood detection into a supervised and explainable flood susceptibility framework for data-limited coastal environments. The integration of multi-source geospatial predictors, comparative ensemble learning, and SHAP interpretation provides both predictive and spatially interpretable information for the Semarang-Demak coastal area. Nevertheless, the findings remain conditional on the selected flood event, Sentinel-1-derived supervisory labels, predictor configuration, and validation strategy. Future research should incorporate multi-event flood inventories, spatially independent validation, and additional dynamic coastal predictors such as tidal conditions, land subsidence, and hydraulic infrastructure to evaluate the transferability and generalizability of the identified susceptibility relationships.

Data Availability

The balanced dataset (1,870 rows), the raw sample, the exact train-test split assignment (seed 42), and all machine-readable outputs of the revised analyses, including test predictions, DeLong/bootstrap tables, cross-validation summaries, fold assignment, correlation matrices, and SHAP summaries, are publicly available at Kaggle dataset.

Acknowledgement

The authors thank the anonymous reviewers for their constructive comments and suggestions. Gratitude is also extended to all parties who supported the data collection and geospatial processing stages of this study.

References

[1] B. Zerouali, A. H. Almaliki, and C. A. G. Santos, "Flood susceptibility mapping in arid urban areas using SHAP-enhanced stacked ensemble learning: A case study of Jeddah," *Journal of Environmental Management*, 2025, ISSN: 0301-4797. DOI: 10.1016/j.jenvman.2025.127128.

[2] A. Kaur, H. S. Lakra, S. Ali, R. Shaw, and A. Ahirwar, "Microscale coastal flood risk assessment methods amid climate change," *Marine Development*, vol. 4, no. 1, p. 13, Jul. 2026, ISSN: 3004-832X. DOI: 10.1007/s44312-026-00083-8.

[3] W. F. M. Tique, M. Sapena, M. Weigand, S. Groth, C. Geiß, and H. Taubenböck, "A comparative assessment of data-driven flood susceptibility mapping in Nigeria," *Natural Hazards*, vol. 122, no. 4, pp. 1–29, Feb. 2026, ISSN: 1573-0840. DOI: 10.1007/s11069-025-07868-y.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 0885-6125. DOI: 10.1023/A:1010933404324.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[6] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017, pp. 3149–3157.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[8] Pusdatinkom BNPB, "Regulation of the national disaster management authority no. 7 of 2023 on guidelines for implementing disaster event and impact data standards," [Peraturan BNPB No 7 Tahun 2023 tentang Petunjuk Pelaksanaan Standar Data Kejadian dan Dampak Bencana], Pusat Data, Informasi dan Komunikasi Kebencanaan (Pusdatinkom BNPB), Regulation 7, 2023.

[9] J. Zhao, M. Li, Y. Li, P. Matgen, and M. Chini, "Urban flood mapping using satellite synthetic aperture radar data: A review of characteristics, approaches, and datasets," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 1, pp. 237–268, Mar. 2025, ISSN: 2168-6831. DOI: 10.1109/MGRS.2024.3496075.

[10] UN-SPIDER, *Recommended practice: Flood mapping using Sentinel-1 data in Google Earth Engine*, United Nations Platform for Space-based Information for Disaster Management and Emergency Response, 2024.

[11] Husnayaen et al., "Physical assessment of coastal vulnerability under enhanced land subsidence in Semarang, Indonesia, using multi-sensor satellite data," *Advances in Space Research*, vol. 61, pp. 2159–2179, 8 Apr. 2018, ISSN: 0273-1177. DOI: 10.1016/J.ASR.2018.01.026.

[12] S. L. Sakina, N. S. Ningsih, and A. Prayogi, "Investigating wave characteristics in the Java Sea and potential influencing factors: A case study of coastal flooding along the northern coast of Semarang City, Indonesia (May 23, 2022)," *Anthropocene Coasts*, vol. 8, pp. 44–, 1 Dec. 2025, ISSN: 2561-4150. DOI: 10.1007/S44218-025-00122-9.

[13] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, Dec. 2017, ISSN: 0034-4257. DOI: 10.1016/j.rse.2017.06.031.

[14] M. A. Marfai and L. King, "Potential vulnerability implications of coastal inundation due to sea level rise for the coastal zone of Semarang City, Indonesia," *Environmental Geology*, vol. 54, no. 6, pp. 1235–1245, Sep. 2008, ISSN: 0943-0105. DOI: 10.1007/s00254-007-0906-4.

[15] H. Z. Abidin, H. Andreas, I. Gumilar, Y. Fukuda, Y. E. Pohan, and T. Deguchi, "Land subsidence in coastal city of Semarang (Indonesia): Characteristics, impacts and causes," *Geomatics, Natural Hazards and Risk*, vol. 6, no. 5–7, pp. 401–427, Sep. 2012, ISSN: 1947-5705. DOI: 10.1080/19475705.2012.692336.

[16] T. G. Farr et al., "The Shuttle Radar Topography Mission," *Reviews of Geophysics*, vol. 45, no. 2, RG2004, Jun. 2007, ISSN: 8755-1209. DOI: 10.1029/2005RG000183.

[17] D. Yamazaki, D. Ikeshima, J. Sosa, P. D. Bates, G. H. Allen, and T. M. Pavelsky, "MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset," *Water Resources Research*, vol. 55, no. 6, pp. 5053–5073, Jun. 2019, ISSN: 0043-1397. DOI: 10.1029/2019WR024873.

[18] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418–422, Dec. 2016, ISSN: 0028-0836. DOI: 10.1038/nature20584.

[19] C. Funk et al., "The climate hazards infrared precipitation with stations — a new environmental record for monitoring extremes," *Scientific Data*, vol. 2, no. 1, p. 150 066, Dec. 2015, ISSN: 2052-4463. DOI: 10.1038/sdata.2015.66.

[20] D. Zanaga et al., *ESA WorldCover 10 m 2020 v100*, European Space Agency (ESA). Dataset released under CC BY 4.0 licence, 2021. DOI: 10.5281/zenodo.5571936.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[22] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988, ISSN: 0006-341X. DOI: 10.2307/2531595.

[23] R. Costache, S. C. Pal, C. B. Pande, A. R. M. T. Islam, F. Alshehri, and H. G. Abdo, "Flood mapping based on novel ensemble modeling involving the deep learning, Harris Hawk optimization algorithm and stacking based machine learning," *Applied Water Science*, vol. 14, pp. 78–, 4 Mar. 2024, ISSN: 2190-5495. DOI: 10.1007/S13201-024-02131-4.

[24] M. Rahimi et al., "Integrating geospatial intelligence and machine learning for flood susceptibility mapping," *Scientific Reports*, vol. 16, 1 Feb. 2026, ISSN: 2045-2322. DOI: 10.1038/S41598-026-41014-3.

[25] N. Gülcan and Ö. Ekmekcioğlu, "Unveiling the performance of pre-processing approaches in machine learning based flood susceptibility mapping," *Natural Hazards*, vol. 122, pp. 199–, 5 Feb. 2026, ISSN: 1573-0840. DOI: 10.1007/S11069-026-08034-8.

[26] E. Alcântara, Y. C. Guimarães, C. F. P. Baião, and J. R. Mantovani, "Interpretable machine learning for flood susceptibility mapping in the metropolitan region of São Paulo, Southeast Brazil," *Discover Geoscience*, vol. 3, 1 Dec. 2025, ISSN: 2948-1589. DOI: 10.1007/S44288-025-00362-9.

[27] O. Aniramu, O. Iyanda, and O. Orimoogunje, "Remote sensing and GIS-based modelling of land use dynamics and urban flood risk in Lagos Megacity for future flood mitigation," *Scientific Reports*, vol. 16, 1 Feb. 2026, ISSN: 2045-2322. DOI: 10.1038/S41598-026-38544-1.

[28] R. L. Bhatt, S. Prabha, M. Aithsham, and P. Dogra, "Change detection analysis of surface water bodies using pixel-based model in Jammu District, India," *Discover Geoscience*, vol. 4, 1 Jan. 2026, ISSN: 2948-1589. DOI: 10.1007/S44288-025-00381-6.

[29] T. A. Kebede, B. T. Hailu, and K. V. Suryabhagavan, "Evaluation of spectral built-up indices for impervious surface extraction using Sentinel-2A MSI imageries: A case of Addis Ababa City, Ethiopia," *Environmental Challenges*, vol. 8, Aug. 2022, ISSN: 2667-0100. DOI: 10.1016/J.ENVC.2022.100568.

[30] M. V. Shukuru and E. W. Machumu, "Analysis of the impervious surface expansion on urban runoff," *African Journal of Land Policy and Geospatial Sciences*, vol. 7, pp. 1068–1085, 4 Oct. 2024, ISSN: 2657-2664. DOI: 10.22004/AG.ECON.346794.