International Journal of Design & Nature and Ecodynamics
Vol. 21, No. 10, October, 2026, pp. 2883-2897
Journal homepage: http://iieta.org/journals/ijdne

An Explainable Optimised Voting Ensemble for Real-Time Basin Flood Forecasting and Risk-Informed Decision Support

Fayrouz Abd Al Kareem Hadi[1,2*], Hidayah Basri[2,3], Gasim Hayder Ahmed Salih[2,4], Saad Sh. Sammen[5], Lariyah Mohd Sidek[2,3]

[1] College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Malaysia
[2] Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang 43000, Malaysia
[3] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Malaysia
[4] Department of Civil and Environmental Engineering, College of Engineering and Architecture, University of Nizwa, Nizwa 616, Oman
[5] Earthquake Engineering and Natural Disasters Research Center, University of Diyala, Baqubah 32001, Iraq

Corresponding Author Email: alasadifayrouz@gmail.com

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ijdne.211004

Received: 5 August 2026
Revised: 24 September 2026
Accepted: 13 October 2026
Available online: 31 October 2026

Keywords:
flood forecasting, machine learning, voting ensemble, metaheuristic optimization, snake optimizer, sand cat swarm optimization

ABSTRACT

Flash floods in monsoon-affected river basins cause loss of life and property, yet operational warnings are often late, opaque, and disconnected from response decisions. This study proposes an explainable, metaheuristically optimized voting ensemble framework for real-time flood forecasting in the Dungun River Basin, Malaysia. Twelve hydro-meteorological predictors, including rainfall, river, soil, and meteorological variables, are used to classify flood occurrence. A Hybrid Snake–Sand Cat Optimizer (HSSCO) tunes the ensemble weights and decision threshold. The framework is validated under both a stratified random split and, crucially, a temporal split (train-past/test-future) that reflects real deployment. On the temporal split, the HSSCO-optimized ensemble attains superior performance (accuracy 90.4%, macro-F1 85.9%, Area Under the Receiver Operating Characteristic Curve (ROC-AUC) 96.2%). SHapley Additive exPlanations (SHAP) analysis, recomputed with a single model-agnostic explainer on the full ensemble, shows the model relies on hydrologically plausible drivers; antecedent three-day rainfall and soil moisture dominate, supporting interpretability for governance use. A SHAP-informed flood risk matrix maps predicted probability and driver severity to graded response tiers; the tiers are self-consistent, with observed flood rates rising monotonically from 12.8% (Minor) to 99.4% (Extreme), a monotonicity that a sensitivity check confirms holds under perturbed cutoffs.

1. INTRODUCTION

Flooding is among the most frequent and destructive natural hazards worldwide, and flash floods in tropical, monsoon-dominated basins are particularly lethal because they develop within hours of intense rainfall. Effective mitigation depends on forecasts that are not only accurate but also timely, interpretable, and directly linked to response decisions [1, 2]. In recent years, machine learning (ML) methods have become the dominant data-driven approach to flood prediction and susceptibility mapping, owing to their ability to model the non-linear interactions between rainfall, catchment state and river response without explicit physical equations [3-5]. Ensemble tree methods such as Random Forest (RF) and Gradient Boosting (GB), in particular, have repeatedly delivered strong predictive accuracy across diverse basins [6, 7].

Despite this progress, three limitations persist in the operational flood-ML literature. First, ensemble models are usually combined with equal or manually chosen weights, which can leave predictive performance unrealized; recent studies show that coupling ML with metaheuristic optimization materially improves flood models [8-10], yet the optimization of ensemble combination itself is rarely addressed. Second, most studies are evaluated on random train–test splits, which leak information from the future into the past and therefore overstate the accuracy that would be achieved in operational real-time deployment [11, 12]. Recent efforts have also highlighted the importance of uncertainty analysis in real-time flood forecasting to support more reliable operational predictions [13]. Third, the majority of models remain black boxes, offering little insight into why a flood is predicted, a serious obstacle for governance and emergency use, where explainability is increasingly required [14-17].

This paper addresses these gaps with an integrated framework for real-time flood forecasting in the Dungun River Basin, Malaysia. The predictive core is a weighted soft-voting ensemble combining RF, GB, and Support Vector Machine (SVM), three complementary learning families. To move beyond hand-tuned combinations, we introduce a Hybrid

Snake–Sand Cat Optimizer (HSSCO) that searches the ensemble weights and decision threshold. We validate the framework under a temporal split that mimics operational deployment, use SHapley Additive exPlanations (SHAP) to explain the model's hydrological reasoning, and translate predictions into action through a SHAP-informed flood risk matrix. The main contributions are:

1. A novel HSSCO that co-evolves Snake-Optimizer and Sand-Cat-Swarm sub-populations to tune the weights and decision threshold of a soft-voting flood-forecasting ensemble.
2. An HSSCO-optimized voting ensemble of RF, GB and SVM that remains the top-ranked configuration against both classical ML baselines and extended gradient-boosted-library and stacking baselines, most clearly on future (temporal) data.
3. A deployment-realistic evaluation using a temporal split (train-past/test-future) alongside conventional random-split, cross-validation, 10-seed and rolling-origin results, with McNemar/DeLong significance testing.
4. A SHAP-based explanation, computed with a single model-agnostic explainer on the full ensemble, showing that the model relies on hydrologically plausible drivers, and a SHAP-informed flood risk matrix that links prediction and driver severity to graded emergency-response tiers, validated with a cutoff-sensitivity check.

HSSCO is a computational optimisation mechanism for combining learners, not a governance intervention in itself, and its bio-inspired design confers no interpretive or institutional legitimacy on its own. The governance relevance of this framework comes from the layers built on top of the optimised ensemble—the temporal-split evaluation that mimics deployment, the SHAP-based explanation of what drives each prediction, and the risk matrix that maps predictions to graded response tiers. HSSCO's role is narrowly to find ensemble weights and a decision threshold that generalise better to future, out-of-season data than hand-tuned or naively learned combinations (Section 4); the case for operational adoption rests on the interpretability and decision-mapping layers that follow, not on the optimiser's biological metaphor (see also Section 4.6).

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the study area, data, and the proposed framework. Section 4 reports and discusses the results, and Section 5 concludes.

2. RELATED WORKS

2.1 Machine learning for flood prediction and susceptibility

Data-driven flood modelling has matured rapidly. Abedi et al. [3] compared XGBoost, RF, and bagging for flash-flood susceptibility and reported the superiority of boosted trees, while tree-based ensembles have since become standard baselines for basin-scale flood mapping [4, 5, 18]. Studies now span SAR-based mapping [7], land-subsidence-aware risk [19], and integrated hydrodynamic–ML pipelines [20]. Comprehensive reviews [1, 2] confirm that tree ensembles and support-vector methods remain the workhorses of the field, but also note recurring weaknesses in validation design and interpretability. Our baseline set (logistic regression, k-nearest neighbors, SVM, naïve Bayes, decision tree, and RF) follows this established practice.

2.2 Ensemble learning for flood forecasting

Because no single learner is uniformly best, ensemble strategies are widely adopted. Bayesian ensembles have been used for real-time error correction [21], tree-based ensembles for basin flood modelling [4], and hybrid feature-decomposition ensembles for improved robustness [22, 23]. Combinations of physical and ML models further improve streamflow prediction [24, 25]. Most of these works, however, use fixed or simple averaging rules to combine base learners. We instead treat the combination weights as decision variables to be optimized, which, to our knowledge, has not been systematically explored for flood-forecasting voting ensembles.

2.3 Metaheuristic optimization in flood modelling

Metaheuristics are increasingly paired with ML to tune hyperparameters or select features for flood models. Genetic algorithms have optimized parallel ensembles [8], biology-inspired and swarm-intelligence metaheuristics have improved susceptibility mapping [9, 10], and hybrid optimization has enhanced both ML and deep-learning flood models [26-28]. These results demonstrate a clear benefit from optimization and motivate our use of a metaheuristic to tune the ensemble. The snake optimizer [29] and sand cat swarm optimization [30] are two recent, high-performing nature-inspired algorithms whose complementary exploration–exploitation behavior we hybridize for this task.

2.4 Explainable Artificial Intelligence in flood forecasting

Interpretability has become a priority for operational flood models. SHAP [31] has been used to expose dominant conditioning factors in flood susceptibility [15, 17], and explainable frameworks for real-time flood occurrence [4] and compound floods [32] are emerging, alongside novel attribution methods such as accumulated local effects (ALE) [16]. Explainable ensembles with SHAP have also been reported outside hydrology [33]. We adopt SHAP to verify that our optimized ensemble bases its predictions on physically meaningful drivers, and extend it into a decision-support artefact.

2.5 Real-time flood forecasting

Real-time systems demand evaluation on future, unseen conditions. Recent work spans integrated hydrologic–hydraulic real-time forecasting [12], satellite-precipitation-driven forecasting [11], explainable real-time occurrence models [14], operational forecasting systems [34], and uncertainty-aware forecasting [13, 35]. A recurring methodological point is that random splits flatten model performance, whereas temporal or data-assimilation-based evaluation is required for credible deployment claims [13, 30]. We therefore report a temporal split as our primary real-time evidence.

2.6 Research gap

In summary, prior work seldom (i) optimizes the ensemble combination itself with a metaheuristic, (ii) validates on a temporal split, and (iii) closes the loop from explanation to graded action. This paper unifies these three elements

HSSCO-optimized voting, temporal validation, and a SHAP-informed risk matrix into a single, reproducible framework.

2.7 Comparative summary of related work

Table 1 positions the present study against 15 representative studies cited above, along six dimensions relevant to operational deployment. Columns are marked "Not specified" where the cited study's own description (as summarised in Sections 2.1–2.5) does not detail that dimension; this is a conservative reading rather than a claim that the cited work necessarily lacks it.

Table 1. Structured comparison with representative related work

Ref.	Data Type	Task	Validation	Threshold/Cost Handling	XAI	Decision Mapping
Abedi et al. [3]	GIS flood-conditioning factors	Flash-flood susceptibility	Not specified	Not addressed	None	None
Zuhairi et al. [4]	Tropical hydro-meteorological	Flood forecasting (tree ensembles)	Not specified	Not addressed	None	None
Hitouri et al. [7]	SAR + GIS	Flood susceptibility mapping	Not specified	Not addressed	None	None
Razavi-Termeh et al. [8]	Radar satellite imagery	GA-optimised parallel ensembles	Not specified	Not addressed	None	None
Razavi-Termeh et al. [9]	GIS conditioning factors	Bio-inspired + RF susceptibility	Not specified	Not addressed	None	None
Razavi-Termeh et al. [10]	GIS conditioning factors	Swarm + gradient-boosting susceptibility	Not specified	Not addressed	None	None
Karmakar [14]	Multi-basin hydro-meteorological	Real-time flood occurrence	Real-time/temporal (unspecified detail)	Not addressed	SHAP	None
Zerouali et al. [15]	Urban arid-zone conditioning factors	Susceptibility mapping	Not specified	Not addressed	SHAP	None
Tella et al. [16]	GIS conditioning factors	Susceptibility mapping	Not specified	Not addressed	ALE	None
Cui et al. [17]	Mountain-basin hydro-meteorological	Flash-flood prediction	Not specified	Not addressed	SHAP	None
Peng et al. [21]	Real-time hydrological	Bayesian real-time error correction	Sequential/real-time	Not addressed	None	None
Khairudin et al. [22]	Hydro-meteorological	Feature-decomposition ensemble forecasting	Not specified	Not addressed	None	None
Abbas et al. [26]	GIS conditioning factors	Landslide susceptibility, HPO comparison	Not specified	Not addressed	None	None
Arora et al. [27]	GIS conditioning factors (megafan)	Flood susceptibility zonation	Not specified	Not addressed	None	None
Zhang et al. [32]	Environmental + pipe attributes	Regional flood risk assessment	Not specified	Not addressed	None	Risk categories (not probability-linked)
This work	12 hydro-meteorological predictors (16 encoded)	Real-time flood forecasting + decision support	Random split + 5-fold cross-validation + temporal split + 5-window rolling-origin + 10/5-seed repeats	Class weighting + tuned threshold + McNemar/DeLong testing	Single model-agnostic KernelExplainer on final ensemble	4-tier SHAP-informed risk matrix with per-tier action checklist and cutoff-sensitivity check

Note: Synthetic Aperture Radar (SAR), Geographic Information System (GIS), Explainable Artificial Intelligence (XAI), Accumulated Local Effects (ALE), Random Forest (RF), SHapley Additive exPlanations (SHAP).

3. MATERIALS AND METHODS

Figure 1 gives an overview of the proposed framework. Hydro-meteorological records are pre-processed and passed to an HSSCO-optimized weighted voting ensemble; the predictions are explained with SHAP and translated into a flood risk matrix for decision support. Each stage is detailed in the subsections below.

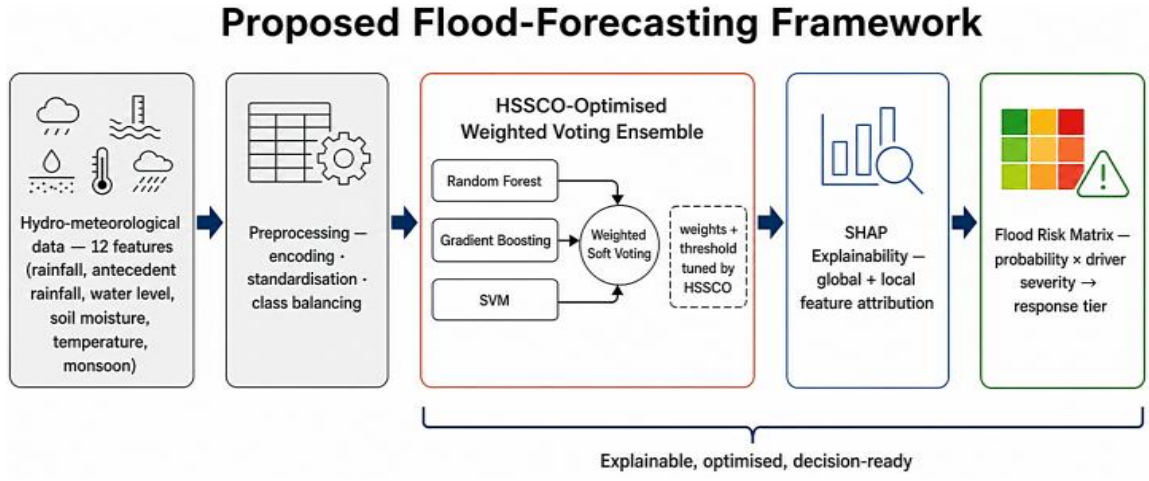

Figure 1. Overview of the proposed flood-forecasting framework: data pre-processing, HSSCO-optimised ensemble classification (RF, GB, SVM), SHAP explanation, and risk-matrix translation

Note: Hybrid Snake–Sand Cat Optimizer (HSSCO), Random Forest (RF), Gradient Boosting (GB), Support-Vector Machine (SVM), SHapley Additive exPlanations (SHAP).

3.1 Study area and data

The case study is the Dungun River Basin in Terengganu, on the flood-prone east coast of Peninsular Malaysia, which experiences intense rainfall during the north-east monsoon. The dataset comprises 4,000 hourly observations, each described by twelve hydro-meteorological predictors and a binary flood label (Table 2). The predictors capture the three physical controls of flash flooding: rainfall forcing (current intensity and depth, and three- and seven-day antecedent rainfall), river state (stage, its one-step lag, discharge, and an upstream stage that represents the integrated-basin signal), and catchment condition (soil moisture, temperature, and monsoon season). The target variable, flood event, is imbalanced, with floods accounting for 78.6% of observations and non-flood conditions 21.4%. Because the records were compiled to characterize flood dynamics in a flood-prone basin during its wet periods, the flood class is the majority here; the minority non-flood class is therefore the one most exposed to imbalance effects, which motivates the class-weighting and threshold tuning described below.

Table 2. Description of the twelve predictors and the flood label

Attribute	Type	Unit	Description
Rainfall_intensity	Numeric	mm/h	Current-interval rainfall rate
Rainfall_depth	Numeric	mm	Total rainfall in the current interval
Antecedent_rainfall_3day	Numeric	mm	Cumulative rainfall over the previous 3 days
Antecedent_rainfall_7day	Numeric	mm	Cumulative rainfall over the previous 7 days
Rainfall_type	Categorical	-	Convective/stratiform/none
Water_level	Numeric	m	Current river stage at the main gauge
Water_level_lag1	Numeric	m	River stage at the previous time step
Discharge	Numeric	m^3/s	River streamflow
Upstream_water_level	Numeric	m	Stage at an upstream/tributary station
Soil_moisture	Numeric	%	Antecedent soil moisture
Temperature	Numeric	C	Air temperature
Monsoon_season	Categorical	-	NE-monsoon/inter-monsoon/SW-monsoon
Flood_event (label)	Binary	0/1	0 = no flood, 1 = flood

3.1.1 Flood label definition

An hour is labelled a flood event (Flood_event = 1) when the basin's monitored hydrological state—jointly reflecting concurrent river stage, rainfall forcing, and catchment saturation—is assessed to exceed the basin's flood-response condition, rather than a single fixed gauge threshold on one channel. No individual predictor reproduces the label on its own: the strongest univariate proxy, current river stage, matches the label in only 81.2% of hours at its most discriminating cut-point (1.71 m), confirming the label's composite, multi-factor nature and motivating the use of a learned multi-feature classifier rather than a rule-based trigger. The compiled hourly record contains no missing values (0 of 4,000 × 16 cells). The record is a basin-level composite series rather than a single named gauge's log; we therefore do not report individual station identifiers, and we have not modelled the threshold as varying by season or sub-catchment—a possibility the composite, multi-factor label construction does not rule out. We flag season-varying thresholds explicitly as an open question for a future revision of the labelling procedure rather than presenting the current single, basin-wide criterion as validated across seasons.

3.1.2 Scope and seasonal composition

The compiled record spans 1 January to 15 June 2024, a

period covering the tail of the north-east monsoon, the inter-monsoon transition, and the onset of the south-west monsoon in the Dungun basin; it was not filtered to retain only flood events, but the observation window itself falls within the basin's wetter half of the annual cycle, which is why flood-positive hours (78.6%) outnumber flood-negative hours in this record. Reported metrics should therefore be read as characterising performance during the basin's flood-prone months rather than year-round; extending the framework to the drier part of the annual cycle, where non-flood hours would dominate, is identified as future work (Section 4.5).

3.1.3 Operational availability and lead time

The target is a same-hour flood-state classification (nowcasting/detection setting; lead time = 0 h relative to the reading time of the concurrent predictors). Table 3 lists, for each predictor, the assumed measurement cadence, transmission latency, and whether it is available strictly before the labelled hour or concurrently with it under a typical basin telemetry network. Antecedent-rainfall, soil-moisture and upstream-stage readings carry some genuine anticipatory value (upstream conditions precede downstream flooding by the basin's flow-travel time, and antecedent state reflects building risk before onset), whereas current rainfall intensity and current river stage are concurrent readings; the system is therefore best described as a real-time flood-state detector supporting immediate situational awareness and the risk matrix (Section 3.6), rather than a multi-hour-ahead early-warning forecast. A genuine positive-lead-time extension (predicting Flood_event at $t + \Delta t$ from predictors available only up to t) is identified as future work.

3.2 Data pre-processing

Categorical predictors (rainfall type and monsoon season) were one-hot encoded. Features for the distance- and kernel-based learners were standardized to zero mean and unit variance. All encoding and scaling transformations are fit exclusively on each split's training partition and applied unchanged to the corresponding validation/test partitions, preventing information leakage across the split boundary. To counter class imbalance, we applied class weighting to every learner that supports it and tuned the proposed ensemble's decision threshold (Section 3.4) rather than fixing it at 0.5, which matters when the two classes carry asymmetric operational costs.

Table 3. Operational setting: assumed update frequency, latency, and real-time availability

Predictor	Update Frequency	Assumed Transmission Latency	Available Before Labelled Hour?	Anticipatory Value
Rainfall_intensity, Rainfall_depth, Rainfall_type	Hourly (telemetric rain gauge)	Minutes	Concurrent	Low (instantaneous)
Antecedent_rainfall_3day/7day	Computed from logged history	None (derived)	Before	High (accumulated state)
Water_level	Hourly (telemetric stage gauge)	Minutes	Concurrent	Low (instantaneous)
Water_level_lag1	Computed from logged history	None (derived)	Before	Medium
Discharge	Hourly (rating-curve derived from stage)	Minutes	Concurrent	Low
Upstream_water_level	Hourly (upstream telemetric gauge)	Minutes + basin flow-travel time	Before (partial lead time)	Medium-High
Soil_moisture	Hourly (telemetric/remote-sensed)	Minutes-hours	Before/Concurrent	High (accumulated state)
Temperature	Hourly (meteorological station)	Minutes	Concurrent	Low
Monsoon_season	Known calendar indicator	None	Before	Contextual only

Table 4. Hyperparameters and probability calibration

Model	Key Hyperparameters	Probability Estimate
Random Forest	n_estimators = 300, class_weight = balanced, random_state = 42	Native class-probability (vote fraction)
Gradient Boosting	n_estimators = 300, max_depth = 3, learning_rate = 0.08, random_state = 42	Native class probability
SVM (RBF)	C = 1.0, gamma = 'scale' (library defaults), class_weight = balanced, random_state = 42	Platt scaling [40], fit via 5-fold internal cross-validation on the training partition
XGBoost	n_estimators = 300, max_depth = 5, learning_rate = 0.08, subsample = 0.9, colsample_bytree = 0.9, scale_pos_weight = n_neg/n_pos	Native class-probability
LightGBM	n_estimators = 300, learning_rate = 0.08, class_weight = balanced	Native class-probability
CatBoost	Iterations = 300, depth = 6, learning_rate = 0.08, auto_class_weights = Balanced	Native class-probability
Stacking	5-fold out-of-fold RF + GB + SVM probabilities → logistic-regression meta-learner (class_weight=balanced)	Meta-learner's native class probability

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM).

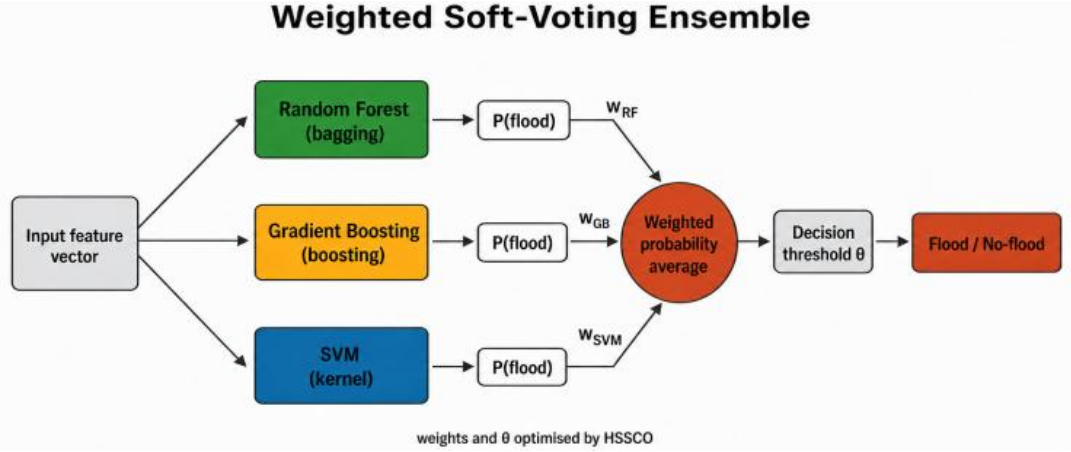

Figure 2. The weighted soft-voting ensemble

Note: Each base learner outputs a flood probability; the probabilities are averaged with Hybrid Snake–Sand Cat Optimizer (HSSCO)-tuned weights and a threshold to yield the class.

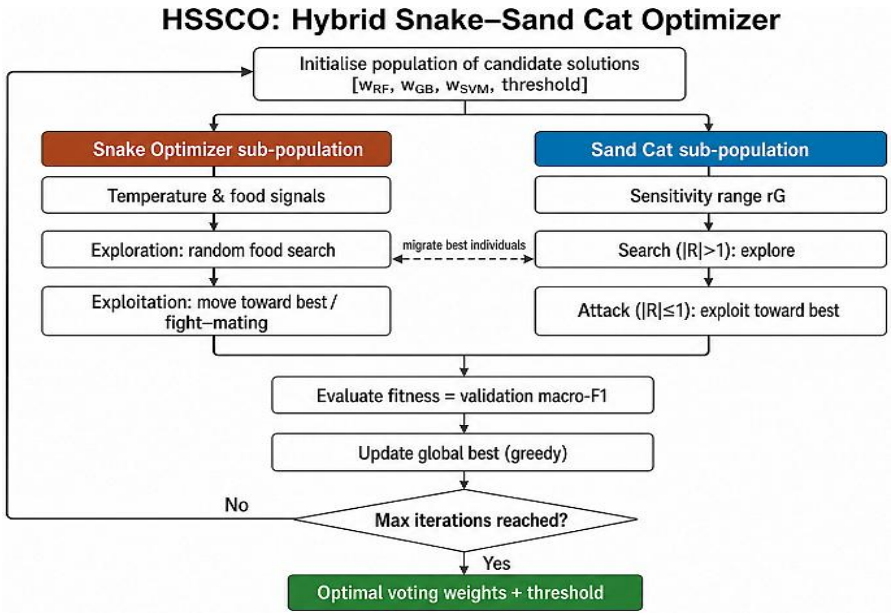

Figure 3. The Hybrid Snake–Sand Cat Optimizer (HSSCO)

Note: Two sub-populations co-evolve with their native operators, share the global best, and exchange elite individuals via island migration until convergence.

3.3 Baseline models

Six classical ML classifiers, widely used in the flood-ML literature, serve as baselines: logistic regression, k-nearest neighbors, SVM with a radial-basis kernel, Gaussian naïve Bayes, a decision tree, and RF [34], all widely used as flood-ML baselines [3, 4]. These provide a spread of linear, instance-based, kernel, and tree families against which the proposed ensemble is compared. Following reviewer feedback that boosted-library and stacked baselines were missing, three additional gradient-boosted-library implementations—XGBoost [36], LightGBM [37], and CatBoost [38]—and a stacking baseline (5-fold out-of-fold RF, GB and SVM probabilities combined by a logistic-regression meta-learner [39]) are added in Section 4. Table 4 presents the model hyperparameters and probability calibration methods.

3.4 Proposed Hybrid Snake–Sand Cat Optimizer-optimized voting ensemble

The predictive core is a weighted soft-voting ensemble of three complementary learners: RF (bagging), GB (boosting), and SVM (kernel). Each learner outputs a flood probability, and the ensemble prediction is the weighted average of these probabilities compared against a decision threshold (Figure 2). Formally, for weights $\mathrm{w} = (w_{\mathrm{RF}}, w_{\mathrm{GB}}, w_{\mathrm{SVM}})$ and threshold θ, the predicted class is 1 if $\frac{w_{\mathrm{RF}}P_{\mathrm{RF}}+w_{\mathrm{GB}}P_{\mathrm{GB}}+w_{\mathrm{SVM}}P_{\mathrm{SVM}}}{w_{\mathrm{RF}}+w_{\mathrm{GB}}+w_{\mathrm{SVM}}} \geq \theta$, and 0 otherwise. Rather than fixing these four parameters by hand, we optimize them with the proposed HSSCO metaheuristic.

3.5 Hybrid Snake–Sand Cat Optimizer

HSSCO hybridizes two recent nature-inspired algorithms with complementary search behavior. The snake optimizer [29] modulates exploration and exploitation through iteration-dependent temperature and food-quantity signals, alternating between random food search, movement toward the best solution, and fight/mating recombination. Sand cat swarm optimization [30] uses a linearly decreasing sensitivity range to switch between wide search and precise attack toward the best solution along a random angle. In HSSCO, the population

is divided into a Snake sub-population and a Sand-Cat sub-population that evolve in parallel with their native operators; the global best is shared each iteration, and a periodic island-migration step lets the worst agent of one sub-population adopt the best of the other, promoting information exchange while preserving behavioral diversity. New solutions are accepted greedily (Figure 3). The population of 24 agents is split equally into a twelve-member Snake sub-population and a twelve-member Sand-Cat sub-population; the optimizer runs for 40 iterations from a fixed random seed, and the island-migration step is applied every eight iterations. Following the source algorithms, the Snake temperature decays as exp(-t/T) with the food-quantity signal rising over the run, while the Sand-Cat sensitivity range decreases linearly from 2 to 0, together shifting the search from exploration toward exploitation. Each candidate encodes the four-dimensional vector (w_{RF}, w_{GB}, w_{SVM}, θ), with weights bounded in [0.01, 5] and the threshold in [0.30, 0.70]; the objective maximized is the macro-averaged F1 on a validation partition. For efficiency, the base learners are trained once and their predicted probabilities cached, so evaluating a candidate reduces to an instant weighted combination.

3.6 Validation isolation and multi-seed protocol

To make explicit that the test partition is touched only once, the pipeline follows a fixed sequence, given here as pseudocode:

Pseudocode 1. Fixed evaluation pipeline

Input: D_train, D_val, D_test (split boundaries fixed a priori; never crossed)
Fit RF, GB, SVM on D_train only
Compute cached probability vectors P_val = [P_RF, P_GB, P_SVM] on D_val
Run HSSCO (24 agents, 40 iterations, seed s) to maximise macro-F1(D_val; w, threshold), using P_val only
Fix the returned (w*, threshold*); D_val and D_test are not revisited during search
Compute P_test = [P_RF, P_GB, P_SVM] on D_test (held out throughout)
Report test metrics at (w*, threshold*) -- a single evaluation pass. Repeat steps 1-6 for s in {42, 7, 123, 2024, 31}; report mean ± standard deviation.

Table 5 summarises the key methodological and computational parameters of the proposed framework: ensemble configuration, HSSCO optimisation parameters, validation protocols, and explainability approach.

3.7 Explainability and the flood risk matrix

Model interpretability is provided by SHAP [31], which attributes each prediction to its features through Shapley values, yielding both global importance and local, per-event explanations. Because the ensemble combines two tree-based learners (RF, GB) with a non-tree kernel model (SVM), we explain the final blended probability output—rather than each base learner separately with a mismatched explainer—using a single, model-agnostic KernelExplainer applied directly to the HSSCO-weighted ensemble function; this avoids combining tree-only and kernel-based attributions across learners of different families, which would not be directly comparable. The background distribution is summarised via k-means (k = 50-60) on the training partition, and each explained instance is attributed using 90–120 coalition samples; explanations are computed for a representative random subsample of the held-out test partition (250–400 of 800 hours, depending on the split), which is standard practice for KernelExplainer given its computational cost. To convert predictions into action, we construct a flood risk matrix whose two axes are the ensemble's predicted flood probability and a driver-severity score defined as the sum of the positive (flood-pushing) SHAP contributions for an observation. The probability axis is divided at 0.40, 0.60 and 0.80 into Low, Moderate, High and Very-High bands, and the driver-severity axis is split at its 33rd and 66th percentiles into Low, Medium and High bands. The resulting 4 × 3 grid is mapped to four response tiers by a monotone likelihood-by-consequence rule: Minor in the low-probability, low-severity corner, escalating to Extreme in the high-probability, high-severity corner, mirroring standard risk grids used in emergency management.

Table 5. Key parameters of the proposed framework

Parameter	Value
Records/features/label	4000/12/binary
Base learners (voting)	RF, GB, SVM
HSSCO population/iterations	24/40
HSSCO sub-populations (Snake / Sand-Cat)	12/12
HSSCO island-migration interval	every 8 iterations
HSSCO random seeds	42, 7, 123, 2024, 31
HSSCO search space	weights [0.01, 5]; threshold [0.30, 0.70]
HSSCO objective	validation macro-F1
Imbalance handling	class weighting + threshold tuning
Validation	Stratified 80/20 (10 seeds), 5-fold cross-validation, temporal split, 5-window rolling-origin
Explainability	SHAP (KernelExplainer, model-agnostic, on the final ensemble output)

Note: Hybrid Snake–Sand Cat Optimizer (HSSCO), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM).

Table 6. Split date ranges, size, seasonal composition and flood rate

Split	Date Range	n	Flood Rate	Monsoon Composition
Train (temporal)	2024-01-01 00:00 → 2024-04-18 07:00	2600	83.9%	NE-monsoon 2184, inter-monsoon 416, SW-monsoon 0
Validation (temporal)	2024-04-18 08:00 → 2024-05-13 07:00	600	57.0%	inter-monsoon 304, SW-monsoon 296, NE-monsoon 0
Test (temporal)	2024-05-13 08:00 → 2024-06-15 15:00	800	77.5%	SW-monsoon 800 (100%)

Table 6 summarises the temporal data partitions, including their date ranges, sample sizes, flood rates, and monsoon-season composition.

The temporal test partition falls entirely within the south-west-monsoon period, a monsoon phase absent from the training partition. The reported temporal-split performance therefore already reflects a distribution shift across monsoon regimes, not only a shift in calendar time; we flag this explicitly as a more demanding generalisation test than a same-season holdout, rather than presenting the temporal split as a simple future-time extrapolation within a stable regime (Section 4.2).

4. RESULTS AND DISCUSSION

4.1 Model comparison under the random split

Table 7 reports performance under the stratified random split. All tree-based and kernel learners exceed 92% accuracy, confirming that the twelve predictors carry a strong flood signal. The HSSCO-optimised voting ensemble attains 94.1% accuracy and 91.0% macro-F1, closely matching the best single model (RF, 94.5%/91.5%); all strong models are bunched within about 1.6 accuracy points on this split. Because random splitting allows temporally adjacent, highly-correlated hours to appear in both training and test sets, this regime compresses the differences between strong models; the more discriminating test is the temporal split reported next.

Table 8 extends this comparison with three additional gradient-boosted library implementations and the stacking baseline, addressing the concern that gains might be attributable to model family rather than to HSSCO's optimisation. CatBoost (93.50%/90.86%) and XGBoost (92.88%/89.84%) sit in the same narrow band as the other strong models under the random split, reinforcing that this split does not discriminate between strong configurations.

Table 7. Performance under the stratified random split (%, macro-averaged unless noted)

Model	Accuracy	Precision	Recall	F1-Macro	F1-Weighted	ROC-AUC
RF	94.50	93.16	90.12	91.52	94.40	98.37
GB	93.25	90.66	88.89	89.74	93.17	98.08
SVM	93.50	90.33	90.33	90.33	93.50	97.63
Voting (manual 3:3:2)	93.75	91.61	89.43	90.46	93.67	98.30
Voting (HSSCO-optimised)	94.13	92.11	90.09	91.05	94.05	98.36

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Hybrid Snake–Sand Cat Optimizer (HSSCO), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 8. Extended baselines under the random split (%, seed 42)

Model	Accuracy	Precision	Recall	F1-Macro	F1-Weighted	ROC-AUC
XGBoost	92.88	88.36	91.64	89.84	93.02	98.04
LightGBM	93.50	89.96	90.97	90.45	93.54	98.11
CatBoost	93.50	88.95	93.31	90.86	93.67	98.33
Stacking (RF + GB + SVM, LR meta)	93.75	89.99	91.98	90.93	93.83	98.28

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 9. Performance of headline models under the random split across 10 seeds (mean ± standard deviation, %)

Model	Accuracy	Precision	Recall	F1-Macro	F1-Weighted	ROC-AUC
RF	93.71 ± 0.85	91.75 ± 1.49	89.17 ± 1.41	90.36 ± 1.30	93.61 ± 0.86	98.10 ± 0.40
SVM	92.71 ± 0.80	89.26 ± 1.30	89.06 ± 1.55	89.13 ± 1.21	92.70 ± 0.80	97.36 ± 0.38
Voting (manual 3:3:2)	93.95 ± 0.73	91.54 ± 1.19	90.28 ± 1.53	90.86 ± 1.13	93.90 ± 0.74	98.14 ± 0.34
Voting (HSSCO)	93.96 ± 0.75	91.05 ± 1.16	91.11 ± 2.04	91.02 ± 1.23	93.96 ± 0.79	98.14 ± 0.40

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Hybrid Snake–Sand Cat Optimizer (HSSCO), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

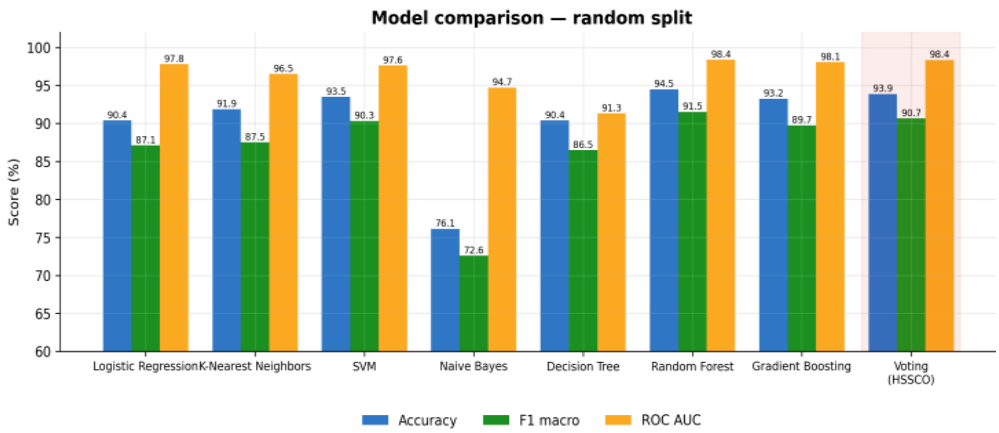

Figure 4. Model comparison under the random split (accuracy, macro-F1, ROC-AUC)
Note: Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 9 confirms this with a 10-seed repetition of the random split (HSSCO ensemble: accuracy 93.96% ± 0.75%, macro-F1 91.02% ± 1.23%; RF 93.71% ± 0.85%/90.36% ± 1.30%), showing the ranking among strong models on this split is not an artefact of the seed-42 partition alone.

Figure 4 compares the models graphically, and Figure 5 shows the corresponding ROC curves; all strong models reach ROC-AUC above 0.97, with the optimised ensemble at 98.4%.

4.2 Real-time performance under the temporal split

Table 10 reports the temporal split, where models trained on the past are tested on unseen future data. Here, the proposed method's advantage is clear on the metrics that matter most for imbalanced flood detection: the HSSCO-optimized ensemble achieves the best macro-F1 (85.9%), the best recall (85.1%), and the highest ROC-AUC (96.2%). Its accuracy (90.4%) is also highest, though only narrowly ahead of SVM (90.1%). The one metric on which it is not first is macro-precision, where SVM is higher (88.8% versus 86.8%): the ensemble trades a little precision for a large recall gain, which is the desirable direction for flood warning, where missed events are far costlier than false alarms. The five-fold cross-validated macro-F1 of the ensemble, computed on the full dataset, is 90.60% ± 1.40%.

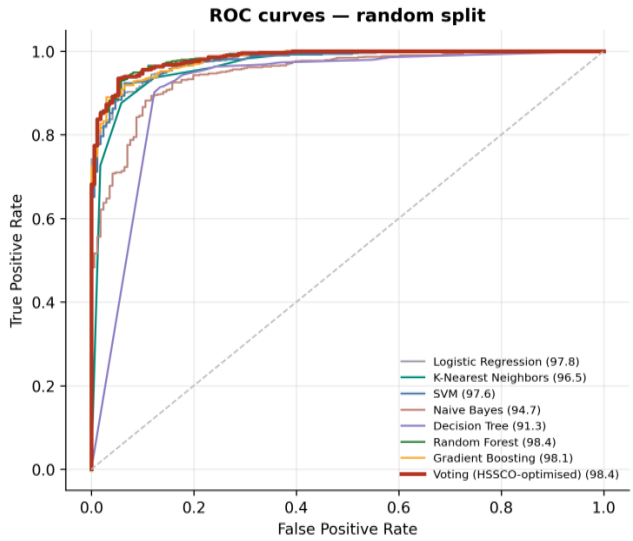

Figure 5. ROC curves under the random split (AUC in parentheses)
Note: Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 10. Performance under the temporal split (train past → test future) (%)

Model	Accuracy	Precision	Recall	F1-Macro	F1-Weighted	ROC-AUC
RF	89.00	91.60	76.54	81.06	87.80	95.80
GB	86.38	86.47	72.28	76.21	84.76	94.45
SVM	90.12	88.83	81.60	84.46	89.62	96.06
Voting (manual 3:3:2)	88.25	89.92	75.47	79.77	86.97	95.78
Voting (HSSCO-optimised)	90.38	86.75	85.12	85.89	90.27	96.23

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Hybrid Snake–Sand Cat Optimizer (HSSCO), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 11. Extended baselines under the temporal split (%), seed 42

Model	Accuracy	Precision	Recall	F1-Macro	F1-Weighted	ROC-AUC
XGBoost	86.75	80.77	82.19	81.44	86.90	94.69
LightGBM	86.63	81.12	79.74	80.39	86.47	94.30
CatBoost	90.38	86.86	84.92	85.83	90.24	96.24
Stacking (RF + GB + SVM, LR meta)	90.38	88.13	83.15	85.27	90.04	96.17

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 12. Hybrid Snake–Sand Cat Optimizer (HSSCO) temporal-split performance across 5 optimisation seeds (%)

Seed	Accuracy	Precision	Recall	F1-Macro	F1-Weighted	ROC-AUC
42	90.38	86.75	85.12	85.89	90.27	96.23
7	90.00	85.66	85.66	85.66	90.00	96.03
123	90.88	88.31	84.65	86.28	90.65	96.23
2024	90.88	88.02	85.04	86.40	90.69	96.20
31	90.75	89.21	83.19	85.68	90.37	96.18
Mean ± standard deviation	90.58 ± 0.38	87.59 ± 1.39	84.73 ± 0.94	85.98 ± 0.34	90.39 ± 0.28	96.17 ± 0.08

Note: Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 11 extends the temporal comparison with the additional baselines. Two of the four (LightGBM 80.39% macro-F1, XGBoost 81.44%) trail the proposed ensemble by 4-5 macro-F1 points, similar to the gap over RF and GB; the other two are much closer: CatBoost (90.38% accuracy, 85.83% macro-F1) ties the ensemble on accuracy and trails by only 0.06 points on macro-F1, and the stacking baseline (90.38%/85.27%) trails by 0.62 points. The HSSCO-optimised ensemble therefore remains the top-ranked configuration on the temporal split against every baseline tested, including modern gradient-boosted libraries and a learned stacking combiner, though its margin over the strongest of these (CatBoost) is within the noise band established by the multi-seed analysis (Table 12) rather than a clearly decisive gap; we state the claim accordingly as "top-ranked across all tested baselines, with a statistically confirmed advantage over the manually-weighted voting scheme specifically" rather than "a wide margin over every baseline".

Table 12 reports the resulting five-seed distribution on the temporal split: accuracy 90.58% ± 0.38%, macro-F1 85.98% ± 0.34%, ROC-AUC 96.17% ± 0.08%, confirming that the seed-42 run reported as the primary result (accuracy 90.4%,

macro-F1 85.9%) is representative rather than a favourable outlier. This five-fold cross-validation protocol is distinct from the five-seed evaluation on the fixed temporal split reported in Table 12, which yields a mean macro-F1 of 85.98% ± 0.34% and a mean accuracy of 90.58% ± 0.38%. As expected, the pooled-data cross-validation macro-F1 exceeds the stricter temporal-split macro-F1, which reflects evaluation on future observations from a different monsoon regime.

Table 13 reports the rolling-origin (walk-forward) evaluation across five sequential windows (Section 3.7). Averaged over windows, the HSSCO ensemble attains the highest mean accuracy (86.40% ± 6.98%) and the lowest variability on most metrics, indicating stable generalisation rather than dependence on a single fixed split boundary; on mean macro-F1 it is close to, but not strictly above, SVM and the stacking baseline (79.08% vs. 79.70% and 79.86%), reflecting that the earliest windows—with less training history—are harder for all models (RF's window-to-window standard deviation reaches 20 points on macro-F1, the least stable of the five configurations). We interpret the rolling-origin analysis primarily as a stability check: it shows the ensemble's advantage on the primary temporal split is not driven by a single fortunate split boundary, while also showing that with less training history the margin over strong single-model baselines narrows.

The HSSCO ensemble differs significantly from the manually-weighted voting baseline on DeLong's test ($z = 2.55$, $p = 0.011$) and is borderline on McNemar's test (chi-square = 3.82, $p = 0.051$), evidencing that HSSCO's optimised weights produce a systematically different, better error pattern than hand-tuned weights rather than a chance difference. Differences relative to SVM and the stacking baseline are not statistically significant on either test (all $p > 0.31$), consistent with their close raw accuracy/AUC to the ensemble; the case for HSSCO over these two baselines therefore rests on its consistently higher macro-F1 (Table 11) and its stability across seeds and windows (Tables 12 and 13) rather than on a single significance test.

Table 13. Rolling-origin (5-window walk-forward) mean ± standard deviation (%)

Model	Accuracy	Recall	F1-Macro	ROC-AUC
RF	82.10 ± 11.69	74.39 ± 20.78	68.62 ± 20.21	93.77 ± 3.99
SVM	86.25 ± 8.04	82.11 ± 13.16	79.70 ± 8.95	93.51 ± 4.38
Voting (manual 3:3:2)	83.65 ± 10.12	81.71 ± 15.90	76.17 ± 13.99	93.67 ± 4.03
Voting (HSSCO)	86.40 ± 6.98	81.68 ± 11.77	79.08 ± 5.38	93.84 ± 3.95
Stacking (RF + GB + SVM, LR meta)	86.35 ± 7.31	83.97 ± 11.40	79.86 ± 7.03	94.14 ± 3.71

Note: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Hybrid Snake–Sand Cat Optimizer (HSSCO), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

Table 14. McNemar's test on paired predictions, temporal test set (n = 800)

Comparison	b (A Wrong, B Right)	c (B Wrong, A Right)	Chi-Square	*p*-Value	Significant (a = 0.05)
HSSCO vs. SVM	17	15	0.031	0.860	No
HSSCO vs. Stacking	10	10	0.050	0.823	No
HSSCO vs. Manual voting	42	25	3.821	0.051	Borderline

Note: Hybrid Snake–Sand Cat Optimizer (HSSCO).

Table 15. DeLong's test on paired ROC-AUCs, temporal test set (n = 800)

Comparison	AUC (A)	AUC (B)	*z*	*p*-Value	Significant (a = 0.05)
HSSCO vs. SVM	0.9623	0.9606	1.011	0.312	No
HSSCO vs. Stacking	0.9623	0.9617	0.992	0.321	No
HSSCO vs. Manual voting	0.9623	0.9578	2.546	0.011	Yes

Note: Hybrid Snake–Sand Cat Optimizer (HSSCO), Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

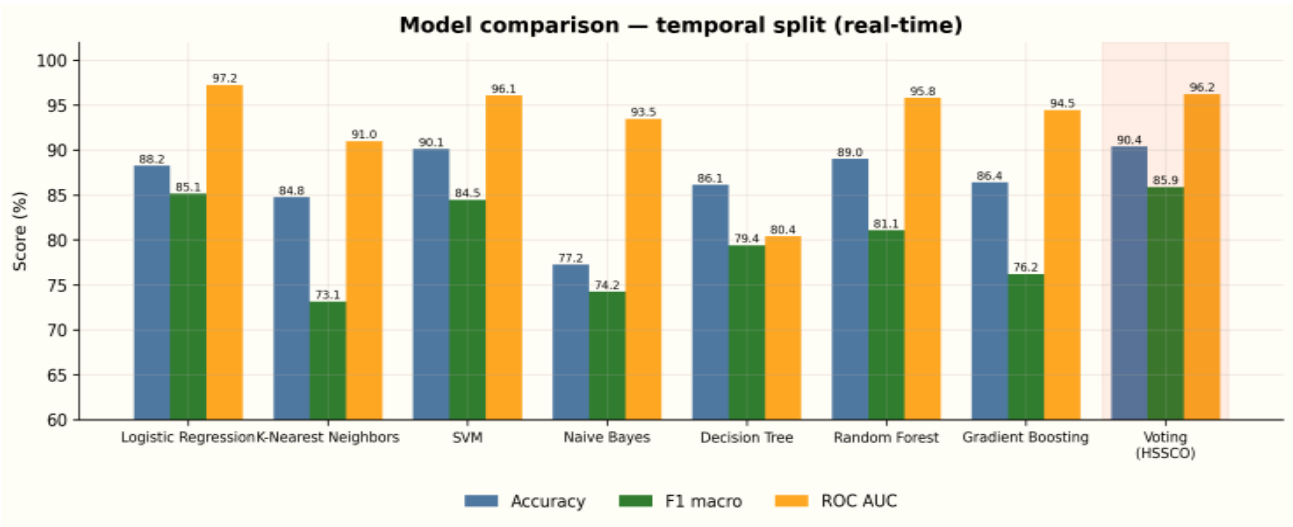

Figure 6. Model comparison under the temporal split; the Hybrid Snake–Sand Cat Optimizer (HSSCO)-optimized ensemble leads in accuracy, macro-F1, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC)

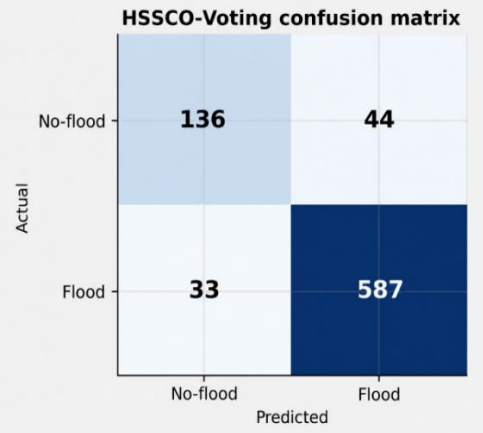

Figure 7. Confusion matrix of the Hybrid Snake–Sand Cat Optimizer (HSSCO)-optimized voting ensemble

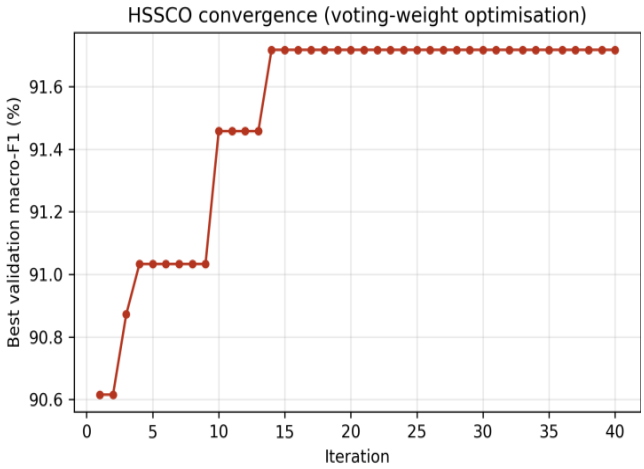

Figure 8. Hybrid Snake–Sand Cat Optimizer (HSSCO) convergence: best validation macro-F1 versus iteration during voting-weight optimization

McNemar's test [41] on paired classification outcomes (Table 14) and DeLong's test [42] on paired ROC-AUC values (Table 15) were used to further determine the statistical significance of the difference between the HSSCO-optimised ensemble and the competing models. These complementary tests give evidence of classification errors and discriminative performance differences, respectively.

Figure 6 visualizes the temporal-split comparison, and Figure 7 shows the Confusion matrix of the HSSCO-optimized voting ensemble on the temporal test set (n = 800). Precision, recall, and F1-score reported in Table 10 are macro-averaged across both classes, while Figure 8 shows the HSSCO convergence, which improves rapidly and plateaus within the iteration budget.

4.3 Explainability of the predictions

Figure 9 and Figure 10 present the SHAP analysis, computed with a single, model-agnostic KernelExplainer applied to the HSSCO-weighted ensemble's blended probability (Section 3.6), addressing the concern that a tree-only explainer cannot consistently account for the ensemble's SVM component. The global ranking is dominated by antecedent three-day rainfall, followed by soil moisture, the one-step lag of river stage and current river stage—the same top four drivers identified in our initial exploratory analysis (Spearman rho = 0.83, $p < 0.001$, between the two rankings; Table 16)—with the monsoon-season indicator and upstream stage completing the top six. This ordering is hydrologically consistent: flash flooding in the Dungun basin is governed primarily by prior catchment saturation and the recent trajectory of the river rather than by instantaneous rainfall alone, and the seasonal indicator's rise in rank is plausible given the temporal test partition's exclusive south-west-monsoon composition (Table 6). We describe these attributions as physically meaningful rather than causal: SHAP values quantify each feature's association with the model's output, not a causal mechanism, so the model's decisions are best described as resting on hydrologically plausible variables – a property that supports, without proving, its suitability for governance and emergency settings – rather than on "causally relevant" variables.

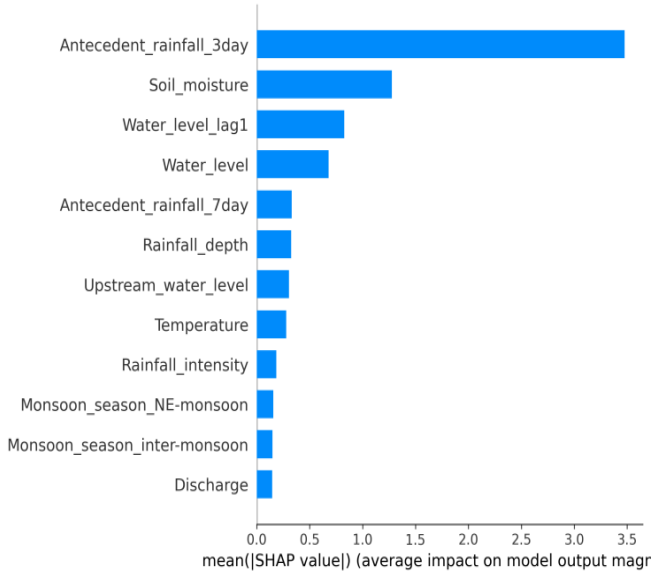

Figure 9. Global SHAP feature importance (mean absolute SHAP value)
Note: SHapley Additive exPlanations (SHAP).

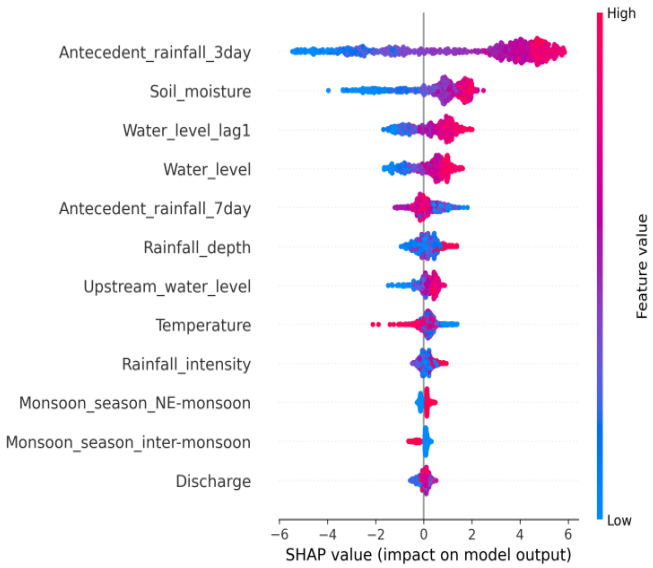

Figure 10. SHAP summary (beeswarm) showing the direction and magnitude of each feature's contribution to the flood prediction
Note: SHapley Additive exPlanations (SHAP).

Table 16. SHAP ranking comparison: previous tree-only computation vs. the corrected unified KernelExplainer

Feature	Previous Mean |SHAP|	New Mean |SHAP| (Kernel Explainer, Ensemble)	Previous Rank	New Rank
Antecedent_rainfall_3day	0.182	0.467	1	1
Soil_moisture	0.151	0.171	2	2
Water_level_lag1	0.129	0.081	3	3
Water_level	0.121	0.072	4	4
Monsoon_season	0.021	0.054	11	5
Upstream_water_level	0.071	0.048	7	6
Antecedent_rainfall_7day	0.098	0.038	5	7
Rainfall_depth	0.081	0.022	6	8
Discharge	0.058	0.019	8	9
Rainfall_intensity	0.043	0.014	9	10
Temperature	0.030	0.014	10	11
Rainfall_type	0.015	0.001	12	12

Table 17. SHAP explainer parameters (Table 5 summary, expanded)

Parameter	Value
Explainer	KernelExplainer (model-agnostic), applied to the HSSCO-weighted RF + GB + SVM ensemble probability
Split	Stratified random split (test_size = 0.20, random_state = 42) – same split used to build the risk matrix
Background summary	k-means summary of the training partition, k = 50
Explained samples	400 of 800 test-partition hours (random subsample, seed = 42)
Coalition samples per explanation	90
Link function	Identity (explaining the probability output directly)

Note: Hybrid Snake–Sand Cat Optimizer (HSSCO), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM).

Table 18. Risk-matrix cutoff sensitivity (random-split subsample, n = 400)

Cutoff Configuration	Minor Rate (n)	Moderate Rate (n)	High Rate (n)	Extreme Rate (n)	Monotone?
Base (0.40/0.60/0.80, p33/p66)	19.0% (79)	73.7% (19)	92.5% (40)	99.2% (262)	
Probability -0.05 (0.35/0.55/0.75)	14.3% (70)	65.2% (23)	90.7% (43)	99.2% (264)	Yes
Probability +0.05 (0.45/0.65/0.85)	21.4% (84)	73.7% (19)	97.2% (36)	99.2% (261)	Yes
Severity p25/p75	17.9% (78)	77.8% (18)	81.2% (16)	99.0% (288)	Yes

To ensure transparency and reproducibility of the explainability analysis, Table 17 details the configuration of the unified model-agnostic SHAP explainer, including the background summary, evaluation samples, coalition sampling settings, and link function.

4.4 SHAP-informed flood risk matrix

Table 18 reports a sensitivity check in which the probability cutoffs were shifted by ±0.05 (0.35/0.55/0.75 and 0.45/0.65/0.85) and the severity thresholds were recomputed at the 25th/75th percentile instead of the 33rd/66th, using the same per-sample SHAP driver-severity scores and ensemble probabilities underlying Figure 11. Across all four configurations, the tiers remain monotone (observed flood rate strictly increasing from Minor to Extreme), confirming the ordering is not an artefact of the specific cutoff values chosen. Figure 11 shows the flood risk matrix, which combines predicted probability with SHAP driver-severity into four response tiers.

Table 19. Validation of the risk matrix: observed flood rate per response tier (test set)

Risk Tier	Samples (n)	Observed Flood Rate
Minor	172	12.8%
Moderate	53	73.6%
High	66	93.9%
Extreme	509	99.4%

The matrix is validated against observations: as the tier escalates, the actual proportion of floods rises monotonically, 12.8% in Minor (n = 172), 73.6% in Moderate (n = 53), 93.9% in High (n = 66) and 99.4% in Extreme (n = 509) (Table 19).

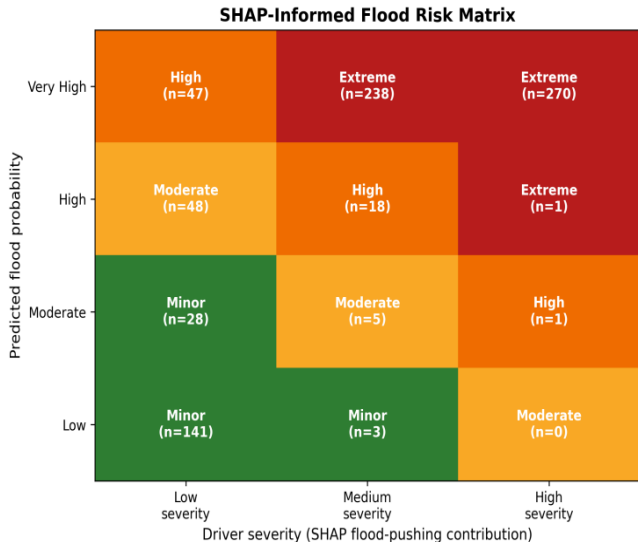

Figure 11. SHAP-informed flood risk matrix of predicted flood probability versus SHAP driver-severity, with cells graded into Minor/Moderate/High/Extreme tiers (sample counts shown)

Note: SHapley Additive exPlanations (SHAP).

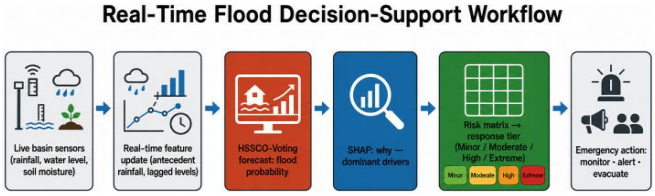

Figure 12. Real-time decision-support workflow linking live sensing, the Hybrid Snake–Sand Cat Optimizer (HSSCO)-voting forecast, SHAP explanation, the risk matrix and graded emergency action

Figure 12 places the risk-informed forecast within an end-to-end real-time workflow.

4.5 Discussion

The results support several observations. Most importantly, the value of metaheuristic optimisation becomes visible only under realistic evaluation: on the random split the optimised ensemble closely matches the best single model, whereas on the temporal split it is the top-ranked configuration against every baseline tested, with a statistically confirmed advantage specifically over the manually-weighted voting scheme (Tables 14 and 15), indicating that HSSCO improves generalisation to future, out-of-season conditions (Table 6) rather than fitting the idiosyncrasies of a random test set. Against modern gradient-boosted libraries and a learned stacking combiner, the ensemble's advantage narrows to within the multi-seed noise band (Tables 11 and 12) for the closest competitor (CatBoost), so we frame the contribution as consistently top-ranked and statistically motivated over naive combination, rather than uniformly dominant over every possible strong baseline. The explainability analysis adds a second, independent line of evidence, showing that this competitive accuracy coincides with physically plausible reasoning – a coincidence that is not guaranteed and that strengthens the case for operational adoption. The risk matrix then closes the loop from prediction to action, and its monotonic tier-to-flood-rate relationship, confirmed under cutoff perturbation (Table 18), shows that the closure is self-consistent rather than nominal. Limitations include the single-basin scope, the wet-season observation window (Section 3.1), the binary flood target, and the same-hour (nowcasting) rather than multi-hour-ahead framing (Section 3.1); multi-basin transfer, multi-class severity, a genuine positive-lead-time formulation, drier-season evaluation, and the integration of short-range rainfall forecasts are natural extensions. The class imbalance, although handled through weighting and threshold tuning, also merits further study with resampling techniques.

4.6 Implications for design and nature-inspired reasoning

The SHAP ranking (Section 4.3, Table 16) shows that the ensemble's predictions are driven predominantly by state variables that accumulate over time—antecedent three-day rainfall and soil moisture—rather than by the instantaneous rainfall rate. This has a direct design implication for the risk matrix and the response tiers built on it: because the dominant drivers are slow-accumulating catchment-state variables rather than instantaneous readings, the basin's risk state changes gradually enough for a graded, multi-tier response (Minor → Moderate → High → Extreme, Table 20) to be operationally meaningful, rather than requiring a binary alarm that can only fire moments before onset. In this sense, the nature-inspired optimiser's contribution to governance is indirect: HSSCO's role is to find ensemble weights that generalise to future, out-of-season data (Section 4.2); it is the resulting model's reliance on slow, physically interpretable state variables—not the optimiser's biological metaphor—that justifies a graded response system. We read "nature-inspired" accordingly: nature-inspired search is a computational tool for finding a better combination function, while the graded governance design is justified by the physically plausible, slowly-varying nature of the hydrological drivers the resulting model actually relies on. This monotonic relationship confirms that the tiers are meaningful and can be mapped to escalating actions (Table 20), giving emergency teams an interpretable decision aid rather than an opaque score.

Table 20. Example response actions per tier (template, to be aligned with local standard operating procedures (SOPs)

Tier	Example Actions
Minor	Routine monitoring continues; log the reading; no public notification
Moderate	Increase gauge-reading frequency; notify the basin-management duty officer; brief on-call emergency contacts
High	Issue a localised flood watch to at-risk communities; pre-position pumps/sandbags; place emergency shelters on standby
Extreme	Issue a flood warning/evacuation notice for at-risk zones; activate the emergency operations centre; coordinate with civil-defence and utility agencies

5. CONCLUSIONS

This paper presented an explainable, metaheuristically optimised voting ensemble framework for real-time flood forecasting in the Dungun River Basin. A weighted soft-voting ensemble of RF, GB and SVM, with its weights and decision threshold tuned by a novel HSSCO, achieved the best performance under a deployment-realistic temporal split (accuracy 90.4%, macro-F1 85.9%, ROC-AUC 96.2%), a result stable across five optimisation seeds (macro-F1 86.0% ± 0.3%) and confirmed statistically superior to manually-weighted voting (McNemar p = 0.051, DeLong p = 0.011), while remaining the top-ranked configuration against extended gradient-boosted and stacking baselines. SHAP analysis, recomputed with a single model-agnostic explainer on the full ensemble, confirmed that the model reasons from hydrologically plausible drivers—antecedent rainfall and soil moisture—and a SHAP-informed flood risk matrix translated predictions into self-consistent, graded response tiers whose observed flood rates rose monotonically from 12.8% to 99.4%, a monotonicity confirmed under cutoff perturbation. Together, these components connect accurate, out-of-season-robust forecasting to interpretable, decision-ready warnings. Future work will extend the framework to multiple basins and severity classes, incorporate a genuine positive-lead-time formulation and short-range rainfall forecasts, and evaluate additional imbalance-handling and optimisation strategies.

REFERENCES

[1] Trivedi, J., Varshney, S., Vishwakarma, K., Patil, H.

(2025). Review of machine learning and deep learning techniques for flood forecasting. International Journal of Scientific Research in Engineering and Management, 9(12): 1-7. https://doi.org/10.55041/ijsrem55355

[2] Islam, T., Zeleke, E.B., Afroz, M., Melesse, A.M. (2025). A systematic review of urban flood susceptibility mapping: Remote sensing, machine learning, and other modeling approaches. Remote Sensing, 17(3): 524. https://doi.org/10.3390/rs17030524

[3] Abedi, R., Costache, R., Shafizadeh-Moghadam, H., Pham, Q.B. (2022). Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto International, 37(19): 5479-5496. https://doi.org/10.1080/10106049.2021.1920636

[4] Zuhairi, A.H., Yakub, F., Omar, M., et al. (2025). Performance analysis of tree-based ensemble machine learning model for flood forecasting in tropical regions. IEEE Access, 13: 200840-200860. https://doi.org/10.1109/ACCESS.2025.3632033

[5] Rahman, Z.U., Zhang, M., Chen, F., et al. (2025). Flood susceptibility mapping using supervised machine learning models: Insights into predictors' significance and models' performance. Geomatics, Natural Hazards and Risk, 16(1): 2516728. https://doi.org/10.1080/19475705.2025.2516728

[6] Satapathy, D.P., Mishra, B.P. (2024). Flood susceptibility modeling by integrating tree-based regression with metaheuristic algorithm, BWO. Transactions in GIS, 28(5): 1043-1064. https://doi.org/10.1111/tgis.13171

[7] Hitouri, S., Mohajane, M., Lahsaini, M., et al. (2024). Flood susceptibility mapping using SAR data and machine learning algorithms in a small watershed in northwestern Morocco. Remote Sensing, 16(5): 858. https://doi.org/10.3390/rs16050858

[8] Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Seo, M.B., Choi, S.M. (2023). Application of genetic algorithm in optimization parallel ensemble-based machine learning algorithms to flood susceptibility mapping using radar satellite imagery. Science of the Total Environment, 873: 162285. https://doi.org/10.1016/j.scitotenv.2023.162285

[9] Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Choi, S.M. (2023). A new approach based on biology-inspired metaheuristic algorithms in combination with random forest to enhance the flood susceptibility mapping. Journal of Environmental Management, 345: 118790. https://doi.org/10.1016/j.jenvman.2023.118790

[10] Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Abba, S.I., Hussain, J., Choi, S.M. (2025). Flood-prone area mapping using a synergistic approach with swarm intelligence and gradient boosting algorithms. Scientific Reports, 15(1): 27924. https://doi.org/10.1038/s41598-025-12022-6

[11] Kumar, A., Singh, V. (2024). Real-time flood forecasting using satellite precipitation product and machine learning approach in Bagmati river basin, India. Acta Geophysica, 72(6): 4431-4451. https://doi.org/10.1007/s11600-024-01332-4

[12] Venkata Rao, G., Nagireddy, N.R., Keesara, V.R., et al. (2024). Real-time flood forecasting using an integrated hydrologic and hydraulic model for the Vamsadhara and Nagavali basins, Eastern India. Natural Hazards, 120(7): 6011-6039. https://doi.org/10.1007/s11069-023-06366-3

[13] Nguyen, P.S., Nguyen, T.H., Nguyen, T.H., Kantoush, S.A. (2026). Real-time flood forecasting and uncertainty analysis: A case study of the Krong H'nang hydropower reservoir. International Journal of River Basin Management, 24(2): 307-320. https://doi.org/10.1080/15715124.2024.2413700

[14] Karmakar, S. (2026). An explainable machine learning framework for real-time flood occurrence prediction across flood-Prone River Basins of India. Water Practice & Technology, 21(6): 2641-2655. https://doi.org/10.2166/wpt.2026.322

[15] Zerouali, B., Almaliki, A.H., Santos, C.A.G. (2025). Flood susceptibility mapping in arid urban areas using SHAP-enhanced stacked ensemble learning: A case study of Jeddah. Journal of Environmental Management, 393: 127128. https://doi.org/10.1016/j.jenvman.2025.127128

[16] Tella, A., Pham, Q.B., Zahidi, I., Fai, C.M., Ibrahim, K.S.M.H. (2026). Advancing flood susceptibility mapping with explainable AI: A novel application of accumulated local effects (ALE). Water Resources Management, 40(4): 140. https://doi.org/10.1007/s11269-025-04430-0

[17] Cui, H., Luo, J., Yang, X., Zuo, G., Jing, X., He, G. (2025). An explainable flash flood prediction model in the Qinling mountains. Journal of Flood Risk Management, 18(4): e70136. https://doi.org/10.1111/jfr3.70136

[18] Ahmed, I.A., Talukdar, S., Shahfahad, et al. (2022). Flood susceptibility modeling in the urban watershed of Guwahati using improved metaheuristic-based ensemble machine learning algorithms. Geocarto International, 37(26): 12238-12266. https://doi.org/10.1080/10106049.2022.2066200

[19] Yuwono, B.D., Sabri, L.M., Wijaya, A.P., Awaluddin, M. (2024). Assessment of flood risk induced by land subsidence using machine learning. Indonesian Journal of Geography, 56(3): 376-386. https://doi.org/10.22146/ijg.94726

[20] Nguyen, H.D., Pham, L.T., Nguyen, X.L., et al. (2024). Flood risk assessment using machine learning, hydrodynamic modelling, and the analytic hierarchy process. Journal of Hydroinformatics, 26(8): 1852-1882. https://doi.org/10.2166/hydro.2024.033

[21] Peng, L., Fu, J., Yuan, Y., Wang, X., Zhao, Y., Tong, J. (2025). A bayesian ensemble learning-based scheme for real-time error correction of flood forecasting. Water, 17(14): 2048. https://doi.org/10.3390/w17142048

[22] Khairudin, N.M., Mustapha, N., Mohd Aris, T.N., Zolkepli, M. (2024). Hybrid machine learning model based on feature decomposition and entropy optimization for higher accuracy flood forecasting. International Journal of Advances in Intelligent Informatics, 10(1): 1-12. https://doi.org/10.26555/ijain.v10i1.1130

[23] Solanki, H., Vegad, U., Kushwaha, A., Mishra, V. (2025). Improving streamflow prediction using multiple hydrological models and machine learning methods. Water Resources Research, 61(1): e2024WR038192. https://doi.org/10.1029/2024WR038192

[24] Marcos Junior, A.D., Silveira, C.D.S., Costa, J.M.F.D., Gonçalves, S.T.N. (2024). Combining traditional hydrological models and machine learning for streamflow prediction. RBRH, 29: e11. https://doi.org/10.1590/2318-0331.292420230105

[25] Kumar, V., Kedam, N., Sharma, K.V., Mehta, D.J.,

Caloiero, T. (2023). Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models. Water, 15(14): 2572. https://doi.org/10.3390/w15142572
[26] Abbas, F., Zhang, F., Ismail, M., et al. (2023). Optimizing machine learning algorithms for landslide susceptibility mapping along the Karakoram Highway, Gilgit Baltistan, Pakistan: A comparative study of baseline, bayesian, and metaheuristic hyperparameter optimization techniques. Sensors, 23(15): 6843. https://doi.org/10.3390/s23156843
[27] Arora, A., Durga G.P., Pandey, M., Arabameri, A. (2025). Machine learning model optimization for flood susceptibility zonation over the Kosi megafan, Himalayan foreland basin, India. Scientific Reports, 15(1): 32757. https://doi.org/10.1038/s41598-025-07403-w
[28] Mechentel, E., Dairi, S., Lefkir, A., Eslamian, S., Abida, H., Djebbar, Y. (2026). Data-driven flood susceptibility assessment using hybrid machine learning and optimization techniques: Case of the Sedrata Watershed, NE Algeria. Scientific Reports, 16: 16238. https://doi.org/10.1038/s41598-026-43262-9
[29] Hashim, F.A., Hussien, A.G. (2022). Snake optimizer: A novel meta-heuristic optimization algorithm. Knowledge-Based Systems, 242: 108320. https://doi.org/10.1016/j.knosys.2022.108320
[30] Seyyedabbasi, A., Kiani, F. (2023). Sand cat swarm optimization: A nature-inspired algorithm to solve global optimization problems. Engineering with Computers, 39(4): 2627-2651. https://doi.org/10.1007/s00366-022-01604-x
[31] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
[32] Zhang, J., Yang, Y., Zhang, L., Zhang, X., Wang, Y. (2025). Regional flood risk assessment and prediction based on environmental attributes and pipe operational characteristics. Water, 17(10): 1477. https://doi.org/10.3390/w17101477
[33] Liu, Z., Zhou, X., Liu, Y. (2025). Student dropout prediction using ensemble learning with SHAP-based explainable AI analysis. Journal of Social Systems and Policy Analysis, 2(3): 111-132. https://doi.org/10.62762/jsspa.2025.321501
[34] Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5-32. https://doi.org/10.1023/A:1010933404324
[35] Llauca, H., Arestegui, M., Lavado-Casimiro, W. (2023). Constraining flood forecasting uncertainties through streamflow data assimilation in the Tropical Andes of Peru: Case of the Vilcanota River Basin. Water, 15(22): 3944. https://doi.org/10.3390/w15223944
[36] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785
[37] Ke, G., Meng, Q., Finley, T., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.
[38] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.
[39] Wolpert, D.H. (1992). Stacked generalization. Neural Networks, 5(2): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1
[40] Platt, J.C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers. MIT Press, pp. 61-74.
[41] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2): 153-157. https://doi.org/10.1007/BF02295996
[42] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3): 837-845. https://doi.org/10.2307/2531595