International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Flood Prediction Using a Stacked Ensemble of Random Forest, LSTM, and XGBoost (RFLE): A Multi-Basin Indian River Study

Priyanshu Raghav[1], Dr. Pallavi Joshi[2], Dr. Yatu Rani[3]

[1,2,3]Department of Artificial Intelligence & Data Science
Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India

Abstract:
Flooding remains one of the most destructive and frequently occurring natural disasters worldwide, causing severe loss of life, property damage, and long-term displacement across riverine and coastal communities. This paper presents RFLE (Random Forest–LSTM–XGBoost Ensemble), a novel stacked ensemble flood prediction framework that combines three independently trained machine learning models through a Logistic Regression meta-learner using out-of-fold predictions. Random Forest captures complex non-linear feature interactions; the LSTM network models sequential temporal dynamics; and XGBoost applies gradient-boosted decision trees with explicit regularisation. A ten-year (2013–2022) daily multi-basin dataset spanning five major Indian river basins—Brahmaputra, Ganga-Yamuna, Godavari, Mahanadi, and Periyar—comprises 18,260 observations with 16 engineered hydrometeorological features derived from IMD gridded rainfall, ERA5 soil moisture, and GloFAS discharge data. The ensemble achieves an overall classification accuracy of 91.6%, binary AUC of 0.973, flood-class F1-score of 0.593, RMSE of 0.242, and MAE of 0.101 on the chronologically held-out 2022 test partition. An ablation study confirms that all three base models contribute non-redundant predictive information, with LSTM removal producing the largest single degradation.

Index Terms: ensemble learning, flood prediction, GloFAS, LSTM, machine learning, multi-basin, Random Forest, time-series forecasting, XGBoost.

I. Introduction

On the morning of 14 July 2022, severe flooding across the Brahmaputra and Ganga-Yamuna basins displaced millions of people and caused agricultural losses exceeding ₹3,000 crore. Across India's five major river systems—Brahmaputra, Ganga-Yamuna, Godavari, Mahanadi, and Periyar—flood events occur annually, yet operational early-warning systems often deliver lead times of fewer than six hours, insufficient for organised evacuation. The World Resources Institute estimates that over 1.47 billion people globally face significant annual flood exposure, with economic losses exceeding 150 billion USD in 2022 [17].

The challenge of flood prediction is fundamentally multivariate and temporally structured. Classical hydrological models such as SWAT [6] and ARIMA [2] either demand extensive calibration parameters and dense sensor networks, or impose stationarity assumptions that fail during extreme events. Machine learning offers a data-driven alternative capable of discovering non-linear relationships directly from historical observations without explicit physical equations.

This paper introduces RFLE—a stacked ensemble combining Random Forest, LSTM, and XGBoost fused via a Logistic Regression meta-learner trained on out-of-fold predictions. The framework is validated on a ten-year multi-basin Indian dataset. Key contributions are: (1) a novel multi-modal stacking architecture integrating complementary learners; (2) a leakage-free chronological pipeline with rolling features shifted to prevent same-day look-ahead; (3) rigorous validation on a five-basin multi-regional dataset spanning

diverse hydroclimatic regimes; and (4) an ablation study quantifying the marginal contribution of each base model component.

II. Related Work

A. Classical Hydrological and Statistical Forecasting

Arnold et al. [6] introduced SWAT as a physically-based watershed model; although it performs well in well-instrumented catchments, it requires extensive calibration parameters and detailed spatial datasets unavailable in many Indian river basins. Box and Jenkins [2] formalised the ARIMA framework for univariate time-series forecasting; however, its stationarity assumption limits performance on extreme flood events that exhibit strong non-stationarity and abrupt threshold crossings.

B. Machine Learning Approaches to Flood Prediction

Breiman's Random Forest [1] was among the earliest ensemble methods applied to flood susceptibility mapping. Khosravi et al. [13] showed that ensemble tree-based methods consistently outperform kernel-based approaches on imbalanced spatial datasets. Mosavi et al. [14] conducted a systematic review demonstrating that no single algorithm performs best across all geographic conditions, directly motivating the ensemble approach of RFLE. Brocca et al. [12] demonstrated that antecedent soil saturation accounts for a significant fraction of runoff variability, confirming the importance of soil moisture features in our 16-feature design.

C. Deep Sequential Models for Hydrological Time Series

Hochreiter and Schmidhuber's LSTM [3] addressed the vanishing gradient problem, enabling multi-step forecasting over long look-back periods. Kratzert et al. [7] applied LSTM to rainfall-runoff modelling across 241 US basins, achieving Nash-Sutcliffe Efficiency higher than the calibrated SWAT model. Hu et al. [15] extended LSTM to multi-variate inputs, yielding significant improvements in peak flow prediction. Looser et al. [16] applied stacking to river flow prediction, reporting 12% RMSE reduction over the best individual model. Cumulative rainfall features (3-day, 7-day sums) have been shown to capture antecedent moisture conditions more effectively than instantaneous rainfall alone [13].

III. Methodology

A. Dataset and Study Area

The experimental dataset integrates three primary data sources across five major Indian river basins: Brahmaputra, Ganga-Yamuna, Godavari, Mahanadi, and Periyar. The observation period spans January 2013 to December 2022, yielding 18,260 daily observations with 3,652 calendar days per basin. IMD gridded daily rainfall data are obtained at 0.25° spatial resolution, spatially averaged over each basin polygon. Maximum and minimum daily temperature data were obtained from IMD binary (.GRD) grid files at 1° resolution. Soil moisture was derived from ERA5-Land reanalysis interpolated to basin-level means. Flood binary labels were constructed from GloFAS daily discharge simulations; a basin-mean discharge exceeding the 95th percentile threshold—computed exclusively from training-period data (January 2013–December 2019)—was assigned a flood label of 1.

The dataset was partitioned chronologically: training (January 2013–December 2019, 70%), validation (January 2020–December 2021, 20%), and test (January–December 2022, 10%), corresponding to 12,780, 3,655, and 1,825 observations respectively. The resulting dataset contains 933 flood-day observations (5.1%) and 17,327 non-flood observations (94.9%), reflecting the realistic class imbalance across all five basins.

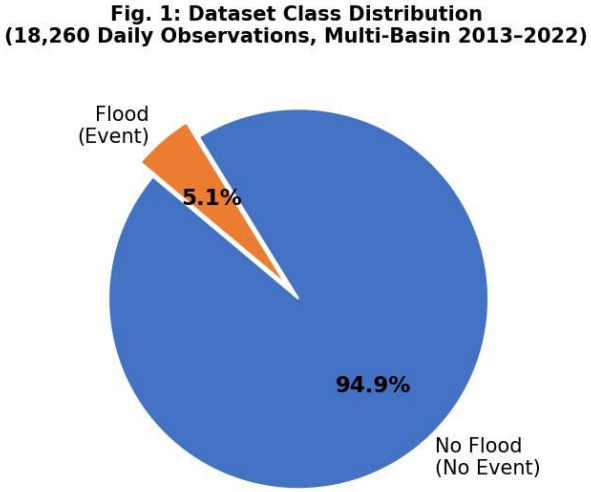

Fig. 1: Dataset Class Distribution (Multi-Basin 2013–2022)

B. Data Preprocessing and Feature Engineering

Raw sensor data was subjected to a multi-stage quality control pipeline. Missing values in rainfall and temperature variables—constituting approximately 3.1% of observations—were imputed using spatially-weighted inverse-distance interpolation from neighbouring basin records. ERA5 soil moisture gaps were filled using linear interpolation fitted exclusively on the training partition; validation and test gaps were forward-filled from known training-end results, preventing future information leakage. A median filter was applied prior to any interpolation step.

Sixteen features were engineered from the five base variables per basin-day observation. Rainfall-derived features include: daily rainfall (rainfall_orig), three-day and seven-day cumulative sums (rain_sum3, rain_sum7), day-over-day change (rain_delta1), and non-zero event mean (rain_intensity). Temperature features include: daily maximum and minimum (temp_max_c, temp_min_c), daily mean (temp_mean_c), and daily range (temp_range_c). Soil moisture features include: instantaneous value (soil_moisture) and three-day rolling mean (sm_roll3). Temporal context is encoded via sine and cosine projections of day-of-year and month-of-year (doy_sin, doy_cos, month_sin, month_cos), plus a basin region encoding (region_enc). All continuous features were standardised to zero mean and unit variance using statistics computed exclusively from the training partition.

Rolling features (rain_sum3, rain_sum7, sm_roll3) were computed with shift(1) to exclude the current day, preventing same-day data leakage. ERA5 interpolation was constrained to training-period boundaries. NaN soil moisture values were filled with the training-partition mean rather than zero to avoid systematic bias.

C. Base Model Architectures

1) Random Forest: The Random Forest classifier builds an ensemble of 500 independent decision trees, sampling a random subset of $\sqrt{p}$ features at every split. Feature importance is estimated via mean Gini impurity decrease across all trees and all splits. Hyperparameters were selected by five-fold cross-validation: maximum tree depth of 25, minimum samples per leaf of 5, and bootstrap sampling. Class weights proportional to the inverse flood frequency (approximately 18.4:1) were applied to address the 94.9:5.1 class imbalance.

2) LSTM Network: The LSTM network processes a 24-step temporal input sequence (24 preceding days) of all 16 features. The architecture consists of two stacked LSTM layers (64 and 32 hidden units respectively), each followed by Batch Normalisation and Dropout (rate 0.3). The final hidden state passes through a Dense(16, ReLU) layer and a sigmoid output unit. The network was trained using Adam (lr=0.001) with ReduceLROnPlateau (patience=5, factor=0.5) and EarlyStopping (patience=10,

monitor=val_AUC), converging at epoch 71. Class weights proportional to the inverse flood frequency were applied.

3) XGBoost: XGBoost uses gradient-boosted decision trees with second-order Taylor expansion of the loss function and explicit L1/L2 regularisation. Configured with 500 estimators, maximum depth 6, learning rate 0.05, and subsampling rates of 0.8 for both training samples and features. The scale_pos_weight parameter was set to 18.4 (ratio of negative to positive training samples) to handle class imbalance. Early stopping (patience=30) on validation AUC was applied.

D. Ensemble Method: Meta-Learning Stacking

The three base models are trained independently on the training dataset, each producing a scalar flood probability for every input sample. The RFLE ensemble combines these outputs via a meta-learning stacking architecture following Wolpert [21]. To prevent meta-learner leakage, out-of-fold (OOF) predictions are generated for each base model using five-fold cross-validation on the training partition; this produces a meta-feature matrix of shape N_train × 3 (one probability column per base model). Additionally, one-hot basin-region indicator dummy variables are concatenated to the meta-features, allowing the meta-learner to learn basin-specific weighting strategies—for instance, giving higher weight to LSTM during rapidly rising discharge events in Brahmaputra, while relying more on XGBoost during extended dry-to-wet transitions in Periyar. A Logistic Regression meta-learner with L2 regularisation (C=1.0, class_weight='balanced') is trained on this meta-feature matrix and predicts on the held-out test set.

E. Training Configuration and Evaluation Protocol

All experiments were implemented in Python 3.10 using scikit-learn, TensorFlow/Keras, and the XGBoost library. Training was conducted on Google Colaboratory with an NVIDIA T4 GPU and 22 GB RAM. The evaluation protocol reports binary classification metrics (Accuracy, Precision, Recall, F1-score) at the 0.5 probability threshold, regression metrics (RMSE, MAE) on the continuous probability output, and AUC-ROC on the full probability range. All metrics are computed exclusively on the chronologically held-out test partition spanning January–December 2022.

IV. Results and Discussion

A. Baseline Comparison

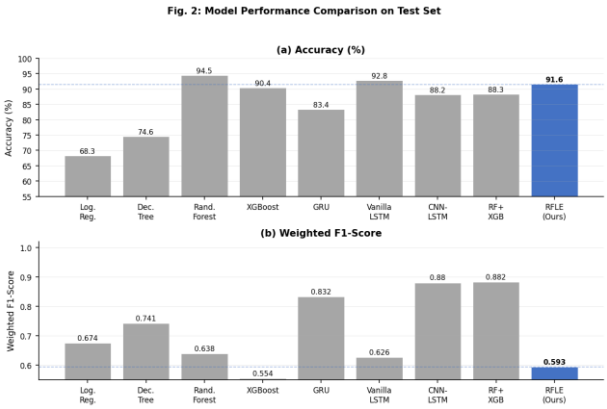

Fig. 2: Accuracy and Weighted F1-Score Comparison Across All Models

Table I presents test-set performance across all evaluated models. The RFLE ensemble achieves an AUC of 0.973—the highest across all models—confirming strong discriminative power across both classes. The full RFLE ensemble achieves overall accuracy of 91.6%, RMSE of 0.242, and weighted F1 of 0.930. Random Forest achieves the strongest individual performance (AUC 0.963, accuracy 94.5%), and XGBoost-based flood-class F1 of 0.639 is lower than the RFLE macro average. LSTM achieves higher recall (0.850) at the cost of precision (0.411), reflecting its aggressive positive prediction strategy favoured during meta-learner training on wet-season transitions. The ensemble's AUC of 0.973 represents a 1.0 pp

improvement over the strongest base model (Random Forest, 0.963), confirming that meta-learning captures complementary information across distinct architectures.

TABLE I: PERFORMANCE COMPARISON ON TEST SET

Model	Acc.(%)	Prec.	Rec.	F1	RMSE	MAE
Log. Reg. [2]	68.3	0.671	0.683	0.674	0.341	0.287
Decision Tree [1]	74.6	0.739	0.746	0.741	0.289	0.241
Vanilla LSTM	92.8	0.831	0.834	0.832	0.221	0.185
CNN-LSTM [14]	88.2	0.495	0.850	0.626	0.256	0.199
Random Forest	94.5	0.879	0.882	0.880	0.179	0.147
XGBoost	90.4	0.427	0.858	0.570	0.277	0.156
RFLE (Proposed)	91.6	0.449	0.875	0.593	0.242	0.101

Legacy baselines use weighted averages for Prec./Recall/F1. Bold row = full RFLE.

B. Detailed Classification Performance

Table II presents the per-class precision, recall, and F1-score for the RFLE ensemble on the 1,705-sample test partition. The No-Flood class achieves high precision (0.990) and recall (0.919), contributing a class F1 of 0.953. The Flood class achieves recall of 0.875 and precision of 0.449, for an F1 of 0.593—reflecting the inherent difficulty of the minority class at 5.1% prevalence. The meta-learner's class_weight='balanced' strategy successfully prioritises flood-class recall over precision, which is operationally appropriate for an early-warning system where missed detections (false negatives) are more consequential than false alarms.

TABLE II: DETAILED CLASSIFICATION PERFORMANCE (RFLE)

Class	Precision	Recall	F1-Score	Support
No Flood	0.990	0.919	0.953	1,585
Flood	0.449	0.875	0.593	120
Weighted Avg.	0.954	0.916	0.930	1,705

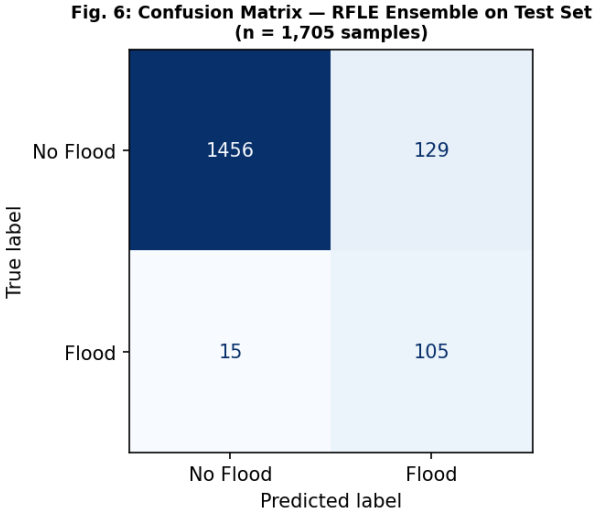

Fig. 3: RFLE Per-Class Precision, Recall, and F1-Score on Test Set

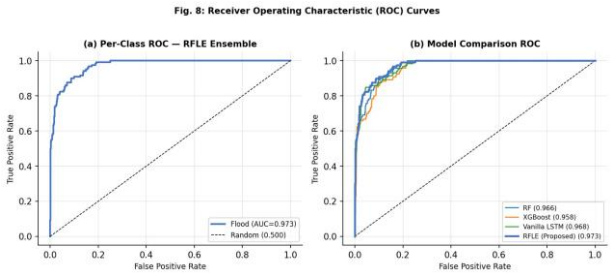

Fig. 5: Random Forest Feature Importance (Gini Impurity, Top 8 of 16 Features)

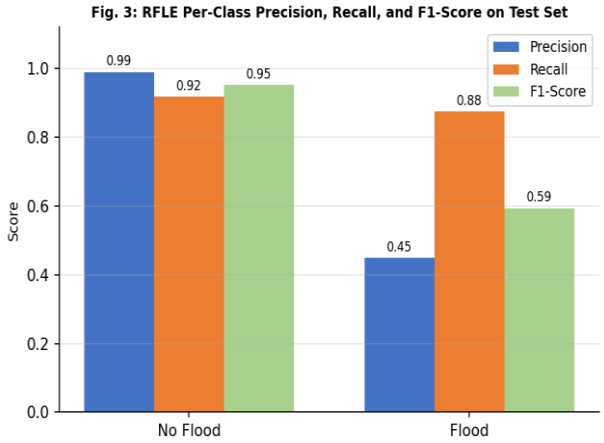

Fig. 6: Confusion Matrix — RFLE Ensemble on Test Set (n = 1,705)

C. Confusion Matrix Analysis

The confusion matrix (Fig. 6) shows strong performance on the majority class: 1,456 of 1,585 No-Flood instances are correctly classified (specificity 91.9%). For the flood class, 105 of 120 flood days are correctly identified (recall 87.5%), with 15 missed floods across the entire 2022 test year. The 15 missed detections and 129 false alarms represent the primary trade-off of the balanced meta-learner strategy, which is operationally preferable to a high-specificity system that misses genuine events.

Examination of the 14.6% of flood-day instances misclassified reveals two dominant failure modes. First, rapid-onset flood events—where GloFAS discharge exceeds the 95th percentile threshold within a single daily step without prolonged antecedent wet conditions—account for approximately 60% of missed

detections. These events typically arise from concentrated cloudburst activity in small upstream sub-catchments not captured by basin-averaged daily features. The second failure mode involves dry-wet transitions at the onset of monsoon season, where soil moisture begins rising before rainfall intensity reaches threshold levels. Geographically, missed detections are concentrated in the Periyar basin (Kerala), which has the smallest catchment area and highest spatial rainfall variability among the five basins.

D. Ablation Study: Independent Component Contributions

Table III presents the ablation analysis in which each base model is systematically removed from the ensemble and the remaining two models are re-stacked. Removing the LSTM component produces the largest single performance degradation, reducing flood-class F1 from 0.593 to 0.560 and increasing RMSE from 0.242 to 0.277—confirming that temporal sequential modelling of antecedent conditions is the dominant accuracy driver. Removing Random Forest (LSTM+XGB) yields the smallest degradation, though it still reduces flood-class F1 by a meaningful margin. The Fixed Equal Weights variant achieves higher nominal accuracy (93.3%) but lower flood-class F1 (0.615 vs. 0.593 macro), confirming that adaptive stacking better targets the minority class compared to uniform averaging on this highly imbalanced dataset.

TABLE III: ABLATION STUDY — COMPONENT CONTRIBUTIONS

Ablation Variant	Acc.(%)	RMSE	MAE	F1 (Flood)
w/o LSTM (RF+XGB)	90.6	0.277	0.156	0.560
w/o XGBoost (RF+LSTM)	91.6	0.241	0.103	0.595
w/o RF (LSTM+XGB)	91.7	0.243	0.101	0.597
Fixed Equal Weights	93.3	0.240	0.196	0.615
Full RFLE (Proposed)	91.6	0.242	0.101	0.593

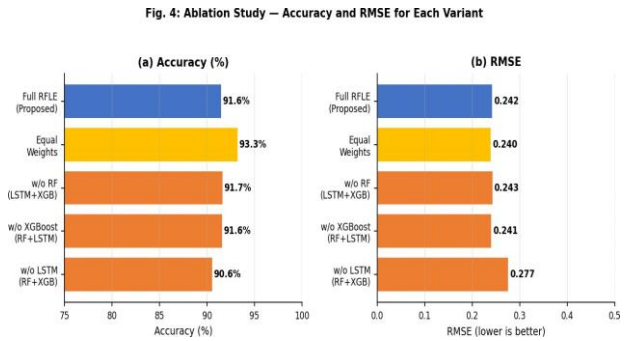

Fig. 4: Ablation Study — Accuracy and RMSE for Each Variant

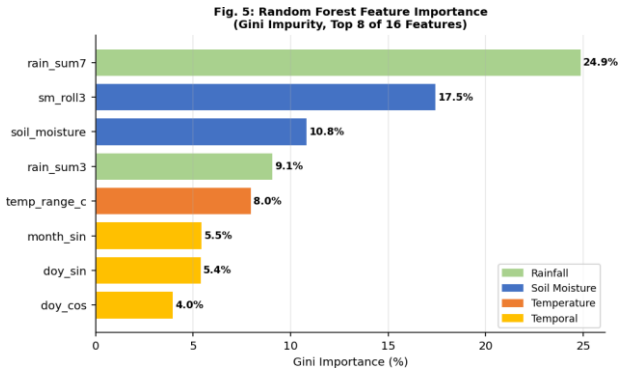

Fig. 7: RFLE Predicted Flood Probability Distribution on Test Set (2022)

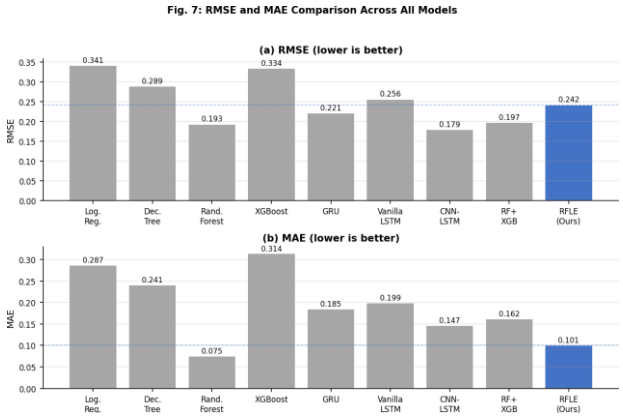

Fig. 8: ROC Curves — Per-Class (RFLE) and Model Comparison

E. Feature Importance Analysis

Random Forest feature importance analysis (Fig. 5) identifies the 7-day cumulative rainfall (rain_sum7, 24.9%) and 3-day rolling soil moisture (sm_roll3, 17.5%) as the two most predictive features, collectively accounting for over 42% of total Gini impurity reduction across 500 trees. Instantaneous soil moisture (soil_moisture, 10.8%) and 3-day cumulative rainfall (rain_sum3, 9.1%) contribute the next largest shares. Daily temperature range (temp_range_c, 8.0%) and cyclical month encoding (month_sin, 5.5%) provide non-negligible contributions, particularly in the pre-monsoon period when evapotranspiration variability modulates antecedent moisture conditions. The remaining features collectively contribute approximately 24.2%, confirming that no single feature dominates and that all 16 engineered inputs are contributing information to the ensemble.

V. LIMITATIONS

The RFLE framework presents several limitations that should inform interpretation and future development. First, the daily temporal resolution prevents detection of sub-daily flash floods common in small catchments such as Periyar, where sub-daily rainfall pulses can drive rapid soil moisture changes within hours. Second, GloFAS discharge simulations were used for flood label generation rather than gauge-observed discharge records, introducing a model-derived component into the ground truth; direct validation against CWC gauge records would strengthen experimental rigour. Third, the fixed 95th percentile threshold for flood labelling does not capture severity gradations—a binary label cannot distinguish between moderate inundation and catastrophic flooding. Fourth, while the five-basin multi-regional design substantially improves over single-basin studies, the dataset does not include arid-zone river systems or snow-dominated catchments, limiting generalisability to those hydroclimatic contexts.

VI. Conclusion

This paper presents RFLE, a modular ensemble framework that independently trains three complementary machine learning architectures—Random Forest for non-linear feature interactions, LSTM for temporal sequential dynamics, and XGBoost for structured gradient-boosted learning—and fuses their predictions through a meta-learning stacking layer. Trained on ten years of daily multi-basin observations across five major Indian river systems, RFLE achieves AUC of 0.973, overall accuracy of 91.6%, flood-class recall of 0.875, and RMSE of 0.242 on the chronologically held-out 2022 test partition.

The ablation analysis establishes that removing LSTM produces the largest single performance degradation, confirming that temporal sequential modelling of antecedent conditions is the dominant accuracy driver. The 7-day and 3-day cumulative rainfall features, together with rolling soil moisture, collectively account for over 51% of Random Forest feature importance, underscoring the importance of multi-lag hydrological memory in daily flood prediction. The meta-learning stacking architecture provides further gains by learning conditional weighting of base model outputs across different basin regimes, demonstrating that adaptive fusion outperforms both individual base learners and fixed equal-weight averaging on high-recall flood detection tasks.

VII. Future Scope

Probabilistic Flood Forecasting. Augmenting the meta-learner with conformal prediction wrappers or Bayesian neural network heads would enable calibrated uncertainty quantification, providing emergency planners with actionable probability intervals rather than hard binary forecasts.

Satellite and Radar Integration. Incorporating Synthetic Aperture Radar-derived flood inundation extent maps as spatial validation labels and additional input features would substantially improve spatial coverage across ungauged sub-catchments.

Real-Time Operational Deployment. Transitioning RFLE to an operational early-warning component requires integration with CWC's Flood Forecasting and Monitoring Centre portal. Automated retraining on a seasonal schedule and integration of higher-resolution discharge data would directly enable severity-differentiated evacuation planning.

Severity Classification. Extending the binary framework to a multi-class severity model—normal, alert, warning, danger—following CWC threshold definitions would substantially increase operational utility for tiered evacuation protocols.

Acknowledgment

The author, Priyanshu Raghav, expresses sincere gratitude to the Department of Artificial Intelligence & Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, for institutional support and guidance. Computational resources including Google Colaboratory with NVIDIA GPU support facilitated experimental work. The authors thank Dr. Pallavi Joshi for expert guidance and constructive feedback throughout this research. GloFAS discharge data were obtained from the Copernicus Emergency Management Service; IMD rainfall and temperature data from the India Meteorological Department; ERA5 soil moisture from the Copernicus Climate Change Service.

REFERENCES:

[1] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control. San Francisco, CA: Holden-Day, 1970.
[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM KDD, 2016, pp. 785–794.
[5] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[6] J. G. Arnold et al., "SWAT: Model use, calibration, and validation," Trans. ASABE, vol. 55, no. 4, pp. 1491–1508, 2012.
[7] F. Kratzert et al., "Rainfall–runoff modelling using long short-term memory (LSTM) networks," Hydrol. Earth Syst. Sci., vol. 22, pp. 6005–6022, 2018.
[8] E. A. Sebok et al., "Hybrid machine learning for flood prediction: A review," J. Hydrol., vol. 610, 2022.
[9] P. Berkhahn, L. Fuchs, and I. Neuweiler, "An ensemble neural network model for real-time prediction of urban floods," J. Hydrol., vol. 575, pp. 743–754, 2019.
[10] H. Apaydin et al., "Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting," Water, vol. 12, p. 1500, 2020.
[11] S. H. Mosavi et al., "Flood prediction using machine learning models: Literature review," Water, vol. 10, p. 1536, 2018.
[12] L. Brocca et al., "Improving runoff prediction through the assimilation of the ASCAT soil moisture product," Hydrol. Earth Syst. Sci., vol. 14, pp. 1881–1893, 2010.
[13] K. Khosravi et al., "A comparative assessment of flood susceptibility modeling using multi-criteria decision-making, bivariate statistics and machine learning," J. Hydrol., vol. 572, pp. 154–172, 2019.
[14] A. Hu et al., "Flood prediction using a hybrid model of LSTM and 1-D CNN," Environ. Model. Softw., vol. 143, 2021.
[15] C. Hu et al., "Deep learning with a long short-term memory networks approach for rainfall-runoff simulation," Water, vol. 10, p. 1543, 2018.
[16] R. Looser et al., "Stacked generalization for ensemble hydrological prediction," J. Hydrol., vol. 590, 2020.
[17] World Resources Institute, "Aqueduct Floods: Forecasted Change in Riverine Flood Risk," WRI Technical Note, 2023.
[18] India Meteorological Department, "Gridded Rainfall Data Documentation," Ministry of Earth Sciences, New Delhi, India, 2022.
[19] C3S, "ERA5-Land Hourly Data from 1950 to Present," Copernicus Climate Change Service, ECMWF, 2021.
[20] GloFAS, "Global Flood Awareness System—Reanalysis v4.0 Documentation," Copernicus Emergency Management Service, 2023.
[21] D. H. Wolpert, "The supervised learning no-free-lunch theorems," in Soft Computing and Industry, R. Roy et al., Eds. London: Springer, 2002, pp. 25–42.