ISSN: 2582-5526 www.thevoiceofcreativeresearch.com

The Voice of Creative Research

*Vol. 8 & Issue 3 (July 2026)*

**Peer Reviewed Article**

**DOI:** https://doi.org/10.53032/tvcr/2026.v8n3.11

# Flood Prediction in the Musi Catchment Region of Telangana Using Machine Learning and Hybrid Models

**V. Sukanya**
Assistant Professor,
Department of Computer Science and Engineering,
University College of Engineering (Autonomous),
Osmania University, Hyderabad, Telangana, India
**Email:** vuppala.sukanya@gmail.com

**K. Prathyusha**
PG Student, Department of Computer Science and Engineering,
University College of Engineering (Autonomous),
Osmania University, Hyderabad, Telangana, India
**Email:** prathyushakannegundla@gmail.com

**Abstract**

Flooding is a common natural disaster that causes severe damage and loss. The aim of flood prediction is to take preventive measures and reduce the damage caused by floods. Flood occurrence is influenced by multiple meteorological factors, rather than rainfall alone. This study used multiparameter weather data, including precipitation, humidity, temperature, surface wind, and sea level pressure, to understand the atmospheric conditions. Historical and forecast weather data for the Musi catchment region were collected from meteorological sources, such as the Coupled Model Intercomparison Project Phase 6 (CMIP6) and Aphrodite data repositories. The data from 2015 to 2049 included multiple environmental parameters. Seven model configurations were developed as four ensemble classifiers (Random Forest, XGBoost, LightGBM, and CatBoost), a two-layer Long Short-Term Memory (LSTM) network, and two hybrid architectures combining LSTM temporal encoding with XGBoost and CatBoost. Hyperparameter optimization used the Optuna framework with Tree-structured Parzen Estimation (TPE), which was applied uniformly across all models. Random Forest achieved the highest accuracy of **95.59%** with an F1-score of **85.41%**, whereas both hybrid models exceeded **94%** accuracy, substantially outperforming. The model performance was evaluated based on accuracy, precision, recall, F1-score, ROC-AUC, and per-class confusion matrices. It also supports area-wise flood risk mapping.

**Keywords:** Flood prediction, machine learning, deep learning, LSTM, hybrid models, ensemble learning, CMIP6, Musi catchment, Optuna.

## 1. Introduction

Flooding is recognized as one of the most frequent and economically devastating natural disasters worldwide, responsible for widespread loss of life and large-scale infrastructure damage. The increasing frequency and severity of flood events are caused by urbanization, deforestation, climate variability, and extreme weather conditions. This study addresses the need for robust and reliable flood forecasting systems that support early warning and informed decision-making. Traditional flood forecasting approaches primarily rely on hydrological models and rainfall threshold- based methods. This approach has limited scalability and cannot capture the complex, nonlinear relationships among multiple meteorological factors such as humidity, temperature, wind, and pressure. With advancements in artificial intelligence, machine learning (ML) and deep learning (DL) techniques have emerged as powerful data-driven alternatives capable of extracting hidden patterns from large-scale meteorological datasets.

This study presents a comprehensive flood prediction framework developed for the Musi catchment region of Telangana, India, which is historically susceptible to severe flooding owing to its geographical and climatic characteristics. The catchment is geographically bounded between latitudes 16°38′N and 17°58′N and longitudes 77°46′E and 79°48′E, encompassing a diverse range of urban, semi-urban, and rural zones, including the Greater Hyderabad Metropolitan Area. Unlike many prior studies that rely on rainfall alone, the proposed framework integrates five atmospheric variables: precipitation, relative humidity, near-surface temperature, surface wind speed, and sea-level pressure to better capture the conditions.

### *1.1 Literature Review*

Vimala et al. [1] proposed a flood prediction system employing supervised machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, Extra Trees, LightGBM, and CatBoost, to analyze historical precipitation data. Random Forest achieved approximately 89% accuracy, whereas XGBoost reached 92%; however, the study relied predominantly on precipitation-based analysis and lacked broader meteorological parameters or temporal deep-learning components.

Shivappriya et al. [2] evaluated flood prediction in Kerala using K-Nearest Neighbour, Logistic Regression, Decision Tree, and Random Forest trained on rainfall datasets from 2000 to 2018. Logistic Regression achieved the highest accuracy of 88%; however, the study was limited by a small dataset and the absence of multiparameter environmental features. Rajdeep et al. [3] developed a flood monitoring system using Random Forest on multi-source meteorological, hydrological, and geospatial data, achieving 91% accuracy, but without sequential temporal learning.

Al-Qazzaz and Paşa [4] conducted a comprehensive evaluation of ensemble learning methods, including bagging, boosting, stacking, voting, and blending. Stacking-based ensemble learning achieved an overall accuracy of approximately 92%, significantly outperforming standalone approaches, although at the cost of increased computational complexity. Sarkar et al. [5] proposed a Flood Risk Index combining the Flood Hazard Index and Flood Vulnerability Index with DeepLabv3+ satellite image segmentation, achieving strong results but requiring substantial computational resources and satellite data availability. Li et al. [6] proposed an interpretable hybrid deep learning model combining Transformer and LSTM architectures for flood forecasting, demonstrating that attention-based sequential learning improves temporal pattern recognition over standalone LSTM or CNN models. Dtissibe et al. [7] compared classical machine learning and deep learning methods for flood forecasting in the far north region of Cameroon, confirming that deep learning models trained on adequate data outperform traditional approaches in regions with complex hydrological dynamics.

Mosavi et al. [8] reviewed machine learning applications in flood prediction across hydrological contexts, whereas Kratzert et al. [9] showed that LSTM networks capture rainfall-runoff dynamics more accurately supporting the temporal learning component adopted in this study.

Despite this considerable research, several gaps remain. Most current studies only use precipitation as the meteorological input and ignore the interactions between humidity, temperature, wind, and pressure. The systematic integration of hybrid gradient boosting and LSTM architectures, coupled with automated hyperparameter optimization and interactive, user-facing deployment, has not been explored for Indian river catchments. This study aims to address these gaps by proposing a multiparameter multi-model flood prediction framework optimized using Optuna and deployed as an interactive web application for real-time area-wise flood risk analysis.

### *1.2 Objectives*

This study aimed to overcome the limitations of rainfall-threshold-based flood forecasting by developing a data-driven, multi-parameter flood risk classification framework for the Musi catchment. The specific objectives of this study are as follows:

i. To develop a multiparameter flood-risk classification framework for the Musi River catchment that captures the combined influence of precipitation, humidity, temperature, surface wind speed, and sea-level pressure, rather than relying on rainfall thresholds alone.
ii. The predictive power of the models was improved using engineered accumulated rainfall, temporal, and spatial features that better represent flood-inducing conditions.
iii. To evaluate and compare ensemble machine learning models (Random Forest, XGBoost, LightGBM, and CatBoost) against hybrid CatBoost-LSTM and XGBoost-LSTM architectures, to identify the configuration that best balances structured and sequential learning for flood-risk classification.
iv. Optuna-based hyperparameter optimization was used to maximize the model accuracy and generalization and minimize overfitting.
v. To enable real-time decision-support flood risk assessment by deploying the optimized models through an interactive web application supporting area-wise, coordinate-based, and manual prediction with monthly/yearly trend visualization.

## 2. Methodology

The proposed framework follows a structured pipeline comprising data collection, preprocessing, model building, optimization, training, and evaluation, as shown in **Figure 1**.

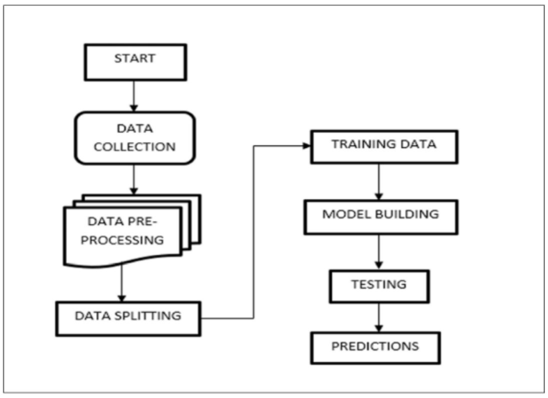


**Figure 1.** Flowchart of the proposed flood prediction pipeline

### *2.1 Dataset*

The study area was the Musi catchment in Telangana (16°38′N–17°58′N, 77°46′E–79°48′E), which was selected for its recurrent flood exposure and mixed urban–rural land use, including the Hyderabad metropolitan area. Meteorological data were obtained from CMIP6, the Earth System Grid Federation (ESGF), and the Aphrodite gridded precipitation product, covering 2015–2049; the 2015–2025 subset was used for training and validation, whereas 2026–2049 projections were reserved for future flood-risk estimation. Five atmospheric variables were used: precipitation, relative humidity, near-surface temperature, surface wind speed, and sea-level pressure, together with spatial (latitude, longitude, zone) and temporal (date, month, season, year) attributes.

Raw records were resampled to daily resolution and temporally aligned across sources, and coarse-resolution CMIP6 fields were statistically downscaled to the catchment scale. Missing values were imputed by linear interpolation for gaps under three days and by climatological substitution for longer gaps. Duplicates were removed, rolling-mean smoothing was applied to suppress noise, and all features were normalized prior to training. Flood indicators were derived from 3-day, 5-day, and 7-day rolling precipitation sums, alongside seasonal and locational covariates. The prediction target was formulated as a three-class flood-risk variable based on accumulated precipitation: Low (< 20 mm), Medium (20-40 mm), and High (> 40 mm), providing a practically interpretable early warning classification. For the LSTM-based models, the processed records were reshaped into fixed-length sliding-window sequences.

### *2.2 Model Architecture*

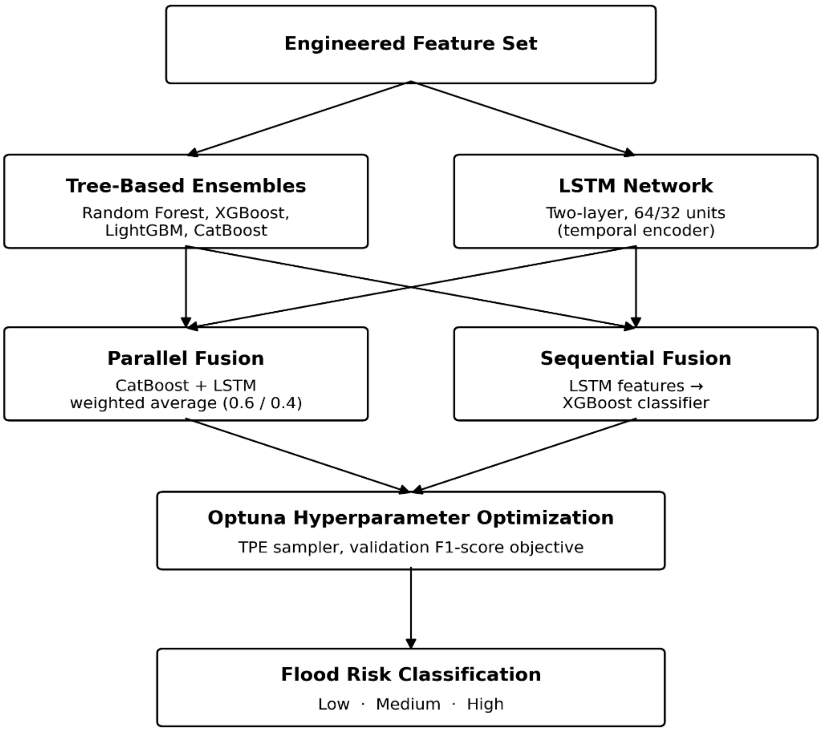


**Figure 2.** Architecture of models used for Flood risk Classification

As Shown in **Figure 2**, Four gradient-boosting and bagging ensembles were implemented: Random Forest, XGBoost, LightGBM, and CatBoost were selected for their established robustness on structured tabular data. A two-layer LSTM network (64 and 32 units, 0.2 dropout) was implemented separately to model the temporal dependencies in the atmospheric time series. Two hybrid configurations were constructed to combine both learning modes: a parallel CatBoost-LSTM ensemble fusing output probabilities via weighted averaging (0.6 CatBoost, 0.4 LSTM) and a sequential XGBoost-LSTM configuration in which LSTM-derived temporal embeddings served as input features to a downstream XGBoost classifier. Seven model configurations were developed and benchmarked using a common evaluation protocol.

### *2.3 Training and Evaluation*

All models were trained on an 80:20 stratified train-test split and tuned using Optuna's Tree-structured Parzen Estimator, with the validation F1-score as the optimization objective to counteract class imbalance across the three risk categories. The search space covered the estimator count, tree depth, learning rate, and regularization coefficients for the ensemble models, and unit count, dropout rate, batch size, sequence length, and epoch count for the LSTM and hybrid models. The model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices computed on the held-out test set. The final tuned models were deployed through a Streamlit-based web application, providing area-wise, coordinate-based, and manually parameterized flood-risk queries, alongside monthly and yearly probability-trend visualization.

## 3. Results

All models were trained on 80% of the data and evaluated on the remaining 20% of the test split. Performance was measured using accuracy, precision, recall, and F1-score, with confusion matrices used to visualize the per-class classification behavior across the Low, Medium, and High flood-risk categories.

### *3.1 Model Performance Comparison*

**Table 1** summarizes the precision, recall, F1-score, and accuracy achieved by each model on the held-out test dataset. The tuned Random Forest model achieved the highest overall performance, with an accuracy of 95.59% and an F1-score of 85.41%, confirming that well-optimized ensemble learning delivers the most robust flood-risk predictions for the Musi catchment dataset. Among the hybrid models, XGBoost + LSTM achieved 94.53% accuracy with an F1-score of 81.38%, closely followed by CatBoost + LSTM with 94.39% accuracy and an F1-score of 81.06%, both consistently outperforming.

**Table 1**. Performance metrics of all implemented models

| Algorithm | Precision (%) | Recall (%) | F1-Score (%) | Accuracy (%) |
|---|---|---|---|---|
| Random Forest | 82.89 | 88.09 | 85.41 | 95.59 |
| XGBoost | 81.31 | 81.82 | 81.56 | 94.57 |
| LightGBM | 79.82 | 81.82 | 80.81 | 94.30 |
| CatBoost | 73.86 | 77.50 | 77.50 | 93.06 |
| LSTM | 60.00 | 27.27 | 37.50 | 86.67 |
| CatBoost + LSTM (Hybrid) | 80.31 | 81.82 | 81.06 | 94.39 |
| XGBoost + LSTM (Hybrid) | 81.25 | 81.50 | 81.38 | 94.53 |

### *3.2 Confusion Matrix Analysis*

This confusion matrix shows the count of correct and incorrect predictions across the Low, Medium, and High flood-risk categories.

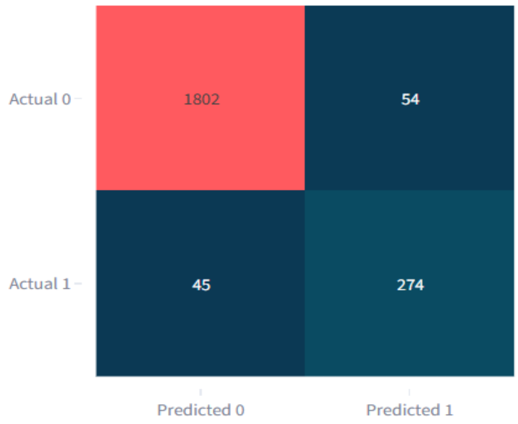


**Figure 3.** Random Forest Confusion Matrix

**Figure 3** presents the confusion matrix of the tuned Random Forest model on the test set. The model correctly classified 1,802 of 1,856 negative (non-flood/low-risk) instances and 274 of 319 positive (flood-risk) instances, reflecting balanced performance across classes and effective reduction of overfitting through the aggregation of multiple decision trees trained on different subsets of the multi-parameter weather data.

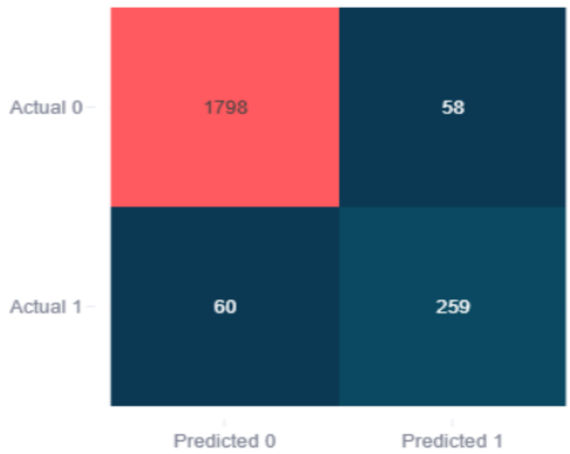


**Figure 4.** XGBoost Confusion Matrix

**Figure 4** presents the confusion matrix of XGBoost, which performs efficiently on the large-scale multiparameter weather dataset through optimized boosting techniques.

### *3.3 Comparison with Existing Studies*

**Table 2** compares the proposed framework with representative studies from the literature. The proposed framework is the only one among the compared studies to integrate six meteorological and spatial parameters with future CMIP6 climate projections, and it achieved the highest reported accuracy of 95.59%.

**Table 2.** Comparison of the proposed framework against related studies

| Study | Best Model | Best Accuracy | Multi-Parameter | Limitation |
|---|---|---|---|---|
| Vimala et al. [1] | XGBoost | 92% | No | Precipitation-only; no temporal model |
| Shivappriya et al. [2] | Logistic Reg. | 88% | No | Small dataset; no deep learning |
| Rajdeep et al. [3] | Random Forest | 91% | Partial | No temporal modelling; no hybrid |
| Al-Qazzaz & Paşa [4] | Stacking Ensemble | 92% | Partial | High complexity; no CMIP6 projections |
| Sarkar et al. [5] | Gradient Boost | N/A ($R^2$) | Yes | No temporal DL; satellite dependency |
| Proposed Framework | Random Forest (Tuned) | 95.59% | Yes (6 params) | Region-specific; future work: IoT integration |

## 4. Conclusions

This study provides an integrated and intelligent approach for flood prediction by employing advanced techniques with multiparameter meteorological data. The main aim of this work is to develop a system that predicts the flood risk based on the environmental conditions and provides easy to use features like prediction for an area, manual input analysis and visual representation of trends. The results show that this goal was successfully achieved using a well-structured and modular approach.

Another key role of this work is the inclusion of multi-parameter data such as precipitation, humidity, temperature, surface wind and sea level pressure. The system is more suitable for real-world applications in disaster management because it significantly improves the reliability of predictions. This system is different from traditional methods, which mostly rely on rainfall. This provides a broader view of the atmospheric conditions that contribute to the occurrence of a flood event.

Streamlit provide real-time flood risk prediction for pre-defined locations and manually entered meteorological parameters. Visualization features, such as monthly and yearly trend graphs, help users better understand seasonal patterns and potential flood risks.

In conclusion, this study successfully demonstrated flood prediction in the Musi catchment region. The system not only allows accurate and reliable predictions but also provides practical tools for analysis and decision-making, demonstrating the potential of data-driven methods in disaster management and providing a scalable and extensible solution to real-world flood prediction systems.

### *4.1 Future Research Directions*

The results of this study provide several opportunities for future investigations. The proposed flood prediction system has great potential for achieving accurate and intelligent flood risk assessments. However, there are many possibilities for further improvement and extension. Future work can be conducted to improve the prediction accuracy, adaptability, scalability, and real-time applicability to make the system more robust and suitable for large-scale deployment.

- Integration with Advanced Sensors: The integration of additional data sources, such as river water level sensors, soil moisture sensors, and satellite imagery with sensor fusion techniques, can provide a better understanding of flood conditions.
- Advanced Deep Learning Architectures: Examining attention mechanisms, transformer architectures for time-series data, and spatio-temporal graph neural networks may enhance the system's capability to model complex environmental interactions.

- Online Implementation: Building a fully online system capable of processing live data and providing immediate predictions is essential for early warning applications.
- Robustness to Environmental Changes: Enhancing the robustness of the system to various environmental conditions, including different terrains and changing climate patterns, would increase its applicability. Continuous learning methods can assist the system in staying effective over time.
- Regional Models: Adaptive regional models could be built rather than a single model. These models would be tailored to the specific geography and climate of the location and would increase the accuracy of predictions on a smaller scale.

**Author Contributions**
All authors contributed equally.

**Funding**
This study received no external funding.

**Data Availability Statement**
The meteorological datasets analyzed in this study were derived from publicly available repositories: the Coupled Model Intercomparison Project Phase 6 (CMIP6), Earth System Grid Federation (ESGF), and Aphrodite precipitation dataset. The Processed datasets and trained model artifacts are available from the corresponding author upon reasonable request.

**Conflicts of Interest**
The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this study.

**Acknowledgement**
The authors gratefully acknowledge the Department of Computer Science and Engineering, University College of Engineering (Autonomous), Osmania University, Hyderabad, for providing the infrastructure and support necessary to conduct this research.

## References

[1] Vimala, S., et al. (2024). Flood prediction using supervised machine learning algorithms. In Proceedings of the International Conference on Advanced Computing (pp. 1–8) https://doi.org/10.1109/ICOSEC61587.2024.10722348
[2] Shivappriya, S. N., et al. (2025). Predicting flood incidences in Kerala using machine learning models. Journal of Environmental Informatics, 45(1), 1–14.https://doi.org/10.1109/ICAECA63854.2025.11012196
[3] Rajdeep, B., et al. (2025). Flood monitoring using machine learning. In Proceedings of the IEEE International Conference on Environmental Data Science (pp. 1–6).https://doi.org/10.1109/ICKECS65700.2025.11035740
[4] Al-Qazzaz, Z. A., & Paşa, Y. (2025). Evaluation of ensemble learning methods for flood prediction. Water Resources Research, 61(3), 1–19.https://doi.org/10.1109/SIU66497.2025.11112464

[5] Sarkar, S., et al. (2025). Machine learning-driven framework for flood risk index calculation. Remote Sensing of Environment, 300, 113892.https://doi.org/10.1109/IEMENTECH65115.2025.10959467

[6] Li, X., Zhang, Y., & Wang, H. (2024). An interpretable hybrid deep learning model for flood forecasting based on Transformer and LSTM. Journal of Hydrology, 627, 1–14. https://doi.org/10.1016/j.ejrh.2024.101873

[7] Dtissibe, F. Y., et al. (2024). A comparative study of machine learning and deep learning methods for flood forecasting in the Far-North region of Cameroon. Journal of Hydrology: Regional Studies, 48, 1–15. https://doi.org/10.1016/j.sciaf.2023.e02053

[8] Mosavi, A., Ozturk, P., & Chau, K.-W. (2018). Flood prediction using machine learning models: Literature review. *Water*, 10(11), 1536. https://doi.org/10.3390/w10111536

[9] Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018