Machine Learning-Based Early Warning Systems for Urban Floods: A Case Study in Nilwala Basin

Abayapala A.I., Lindamulla L.M.L.K.B.
Department of Civil Engineering, Faculty of Engineering
Sri Lanka Institute of Information Technology, Malabe 10115, Sri Lanka.
abiruabayapala@gmail.com, lahiru.l@sliit.lk

ABSTRACT

This study pioneers the integration of Graph Neural Networks (GNNs) into flood forecasting systems, extending the predictive horizon from short-term forecasts to 7 days by effectively capturing spatial dependencies between rainfall stations. Focusing on the flood-prone regions of Matara and Galle districts within the Nilwala Basin, the research addresses the limitations of conventional forecasting methods by leveraging historical hydrological data, including daily rainfall records from six key stations and flow data from Pitabeddara. A hybrid machine learning framework combining Random Forest (RF) and K-Nearest Neighbors (KNN) models was developed to predict river discharge using rainfall data, overcoming challenges posed by limited water level data. The inclusion of GNNs introduces a novel approach to modeling complex spatial relationships, enabling improved accuracy in long-term flood prediction, particularly during extreme events. The proposed system demonstrates significant advancements in predictive reliability, offering a timely and accurate early warning tool to enhance disaster preparedness and risk management in the Nilwala Basin. This research underscores the transformative potential of data-driven methodologies in addressing the challenges of flood-prone regions.

KEYWORDS: *Floods, Flood Early Warning Systems, Nilwala Basin, Machine Learning*

1 INTRODUCTION

Flooding is one of the most severe natural disasters, significantly affecting human lives and economic activities. The Nilwala Basin, located in southern Sri Lanka, frequently experiences urban flooding due to its topography and seasonal rainfall patterns, especially during the Southwest monsoon. Traditional hydrological models for flood forecasting often fail to capture the complex non-linear relationships between rainfall and river discharge, leading to inaccurate predictions. Nilwala basin is located in southern region of Sri Lanka covering about 1032 km^2. This basin consists of flatlands, mountains and valleys. From the sea level the altitude varies from 0 to 925.m. Upper catchment receives about 3000 mm annual rainfall while the lower receives about 1900 mm annually. In the watershed the soil types include Red- Yellow podzolic soils, Alluvial soils and dune sands (Abeysiriwardana & Wijesekera, 2022). The Nilwala river basin experiences frequent flooding primarily during the Southwest monsoon (May-September) and the second intermonsoon periods, following the same monsoon patterns that affect other major Sri Lankan river basins including the Kelani basin (Dhanapala et al., 2022). Following Figure 1 shows the map of Nilwala river basin as per the study by Basnayake (2022).

https://doi.org/10.54389/RVXD3075

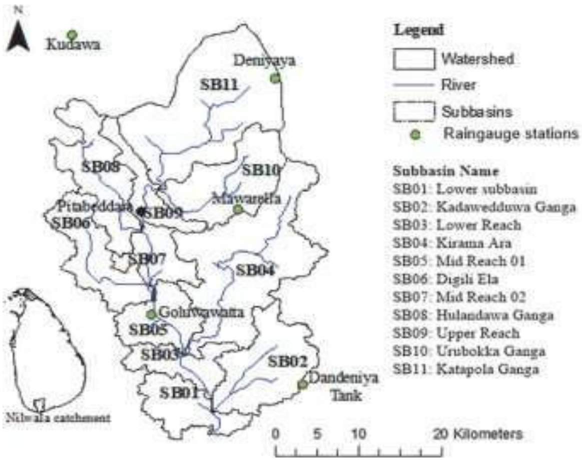

Figure 1. Location of Nilwala River, River Network, its Sub-basins, and Rain gauge Stations (Basnayake, 2022)

Flood forecasting has traditionally relied on hydrological and hydraulic models such as HEC- HMS and physically based numerical simulations (Ke et al., 2020). However, these models require extensive domain knowledge and computational resources, often lacking the flexibility to adapt to varying hydrological conditions (Samarasekara & Sooriyabandara, 2022). Following Figure 2 shows the locations of the rain gauge stations along with flood distribution in Nilwala basin (Kumara et al., 2018).

Figure 2. Matara District flood inundation region with water and rain gauge stations (Kumara et al., 2018)

Machine Learning in Flood Forecasting: Machine learning methods such as Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and Support Vector Regression (SVR) have been successfully applied in hydrological modeling. These models outperform traditional statistical approaches by learning complex data patterns directly from historical records (Mosavi et al., 2018).

Hybrid and Advanced ML Models: The use of hybrid models combining multiple machine learning techniques have shown promising results in flood forecasting. For example, combining Random Forest with LSTM improves predictive accuracy by capturing both spatial and temporal patterns (Ganjirad & Delavar, 2023). Graph Neural Networks (GNN) is gaining popularity in hydrological forecasting due to their ability to model spatial dependencies. In the context of flood prediction, GNNs can capture the relationships between rainfall events across different parts of a river basin, improving the accuracy of runoff predictions (Zhao et al., 2024). A study by Shi et al. (2023) focused on computing water levels as a flood forecasting method using Graph Transformer Network (FloodGTN) for river systemsThe system is designed to incorporate external factors such as precipitation, tidal patterns, and the operational configurations of hydraulic infrastructure (eg: dam discharges, gate positions, and pump activity) along the river (Shi et al., 2023). The above study reported 70% accuracy improvement over HEC-RAS, however no raw MAE or root mean square error (RMSE) were reported.

This study aims to develop a Machine Learning (ML)-based early warning system for urban floods in the Nilwala Basin, leveraging historical hydrological data. To achieve this, the research focuses on the following objectives:

- To identify the required variables to understand the flood patterns in the Nilwala Basin.

- To build a relationship between rainfall in several stations and the flow variation at a specific station.
- To use combined machine learning techniques for accurate predictions.
- To train the machine learning models for one-week early predictions in flow.
- To enhance the accuracy of flow prediction in the machine learning models.

By employing a combination of Random Forest (RF), K-Nearest Neighbors (KNN), and Graph Neural Networks (GNN), this research seeks to provide a clear path for accurate short and long-term flood predictions, enhancing disaster preparedness and risk management.

2 METHODOLOGY

This study adopts a data-driven approach to develop an effective early warning system for flood prediction in the Nilwala Basin using a combination of machine learning techniques: Random Forest (RF), K-Nearest Neighbors (KNN), and Graph Neural Network (GNN). The methodology is divided into the following key components.

- Data collection and pre-processing
- Model development – several approaches
- Feature Engineering
- Model Evaluation and Validation
- Long-term predictions and enhancements

2.1 Data Collection and Preprocessing

Data Collection: The primary data used in this study consists of historical rainfall and river flow records from January 1, 2012, to December 31, 2022. Rainfall data were collected from six meteorological stations within the Nilwala Basin: Deniyaya, Kotapola, Athuela, Deegala, Goluwawatta, and Kekenadura. These stations were selected based on their geographical distribution and the availability of reliable data. The river flow data were obtained from the Pitabeddara station, located downstream, which is a critical point for monitoring flood risk in the basin.

The collected data required significant preprocessing due to issues such as missing values, inconsistent time intervals, and data quality. The daily rainfall data, , were merged with the hourly river flow data by converting the flow data to daily maximum values to match the rainfall dataset's time scale. Missing rainfall and flow records were identified and were adjusted using the data from neighboring stations to improve model accuracy (Subramanya, 2013). This method compares the provided data with the normal value for precipitation at a particular time period (e.g., hourly. daily, weekly, monthly). Due to the heavy variation of the normal precipitation following

$$P_x = \frac{N_x}{M}\left(\frac{P_1}{N_1} + \frac{P_1}{N_1} + \dots + \frac{P_m}{N_m}\right) \qquad 01$$

Table 1 shows the data availability of the stations.

Table 1. Visual representation of the percentages of missing data in the selected stations

Station	Data Feature	Available Amount	% Available
Deniyaya	Rainfall	3196/4018	79.54%
Kotapola	Rainfall	2861/4018	71.20%
Athuela	Rainfall	2494/4018	62.07%
Deegala	Rainfall	2464/4018	61.32%
Goluwawatta	Rainfall	3376/4018	84.02%
Kekenadura	Rainfall	3987/4018	99.23%
Pitabeddara	Flow	4018/4018	100%

To ensure consistency across different input variables, MinMax Scaling was applied. This method normalized the data to the range 0 to 1, reducing the influence of variables with larger magnitudes and improving model convergence during training.

2.2 Model Development

The study implemented three machine learning models, each tailored to capture different aspects of the rainfall-flow relationship in the Nilwala Basin.

Random Forest (RF) Model

The Random Forest algorithm was chosen for its robustness in handling non-linear relationships and its ability to process large datasets effectively. RF is an ensemble learning method that builds multiple decision trees during training and averages their predictions, thereby reducing overfitting and increasing accuracy. Lagged rainfall data (1-day, 2-day, and 3-day lags) were included as input features to capture the delayed effect of rainfall on river flow. Additionally, rolling averages of rainfall were used to represent cumulative rainfall impacts.

These features become input predictors for each sample in the training dataset. The RF algorithm built an ensemble of decision trees using samples from the training data and selecting random features subjects (statistical metrics from different stations and times). The model objective was to build a nonlinear relationship between the statistical features and target flow values by averaging predictions.

The dataset was split into training (80%) and testing (20%) subsets. The model was trained on the training set, and its performance was evaluated on the test set using metrics such as Root Mean Square Error (RMSE) and R-squared (R^2) values.

K-Nearest Neighbors (KNN) Model

The KNN model was trained by constructing feature vectors from the same set of statistical rainfall and flow characteristics. The model was employed to capture localized patterns in the data, leveraging the similarity between recent rainfall and historical events. KNN is a non-parametric algorithm that identifies the k most similar data points (neighbors) and uses their outcomes to predict the target variable.

The input features for the KNN model included daily rainfall data from all six stations and the maximum daily flow value at Pitabeddara. A dynamic approach was used to determine the optimal value of k (number of neighbors) based on model performance during cross-validation.

KNN does not require an explicit training phase like other models; instead, it relies on the stored dataset to make predictions. The training phase involved storing these feature vectors along with observed flow values. The optimal number of neighbors k was determined via cross-validation by minimizing prediction error metrics on the validation set.

The KNN model's performance was assessed using RMSE to quantify prediction errors. Additionally, a confusion matrix was constructed to evaluate the model's ability to classify high-flow events, critical for flood warnings.

Long Short-Term Memory (LSTM) Model

The Long Short-Term Memory (LSTM) model used in this study follows a structured process to improve flood prediction in the Nilwala Basin. Initially, the model's development began with rigorous data collection and preprocessing, where historical daily rainfall data from six stations and hourly river discharge data (converted to daily maxima) were compiled.

Once gathered, the data underwent normalization to balance the input scales and reduce training biases. Subsequently, time-series sequences were created to capture temporal dependencies using a look-back window, enabling the model to learn from past rainfall patterns to influence future discharge. The LSTM architecture, featuring input, hidden memory cells, and output layers, was designed to effectively capture both short and long-term dependencies, which are crucial for understanding flood dynamics. The model was then trained using a training dataset and evaluated on separate validation data, utilizing optimization techniques and loss functions such as Root Mean Square Error (RMSE) to fine-tune performance. It was trained using mini-batch gradient descent, feeding batches of sequences. Model output was analyzed by comparing actual vs. predicted flows, with confusion matrices providing insight into classification accuracy.

Graph Neural Network (GNN) Model

The GNN model was implemented to address the complex spatial dependencies between the rainfall stations and the river flow at Pitabeddara. GNNs are well-suited for this task as they can effectively model relationships between nodes (stations) in a graph structure, capturing both spatial and temporal patterns. A graph was constructed where each node represented a rainfall station, and the edges represented the influence of rainfall on downstream flow. The edge weights were determined based on the correlation between rainfall at each station and the flow at Pitabeddara.

The GNN model consisted of three layers - An input layer to receive rainfall data, a hidden layer with Graph Convolutional Networks (GCN) for feature extraction, and an output layer that provided the predicted weekly maximum flow. The hidden layer applied convolutional operations across the graph to aggregate information from neighboring nodes, enhancing the model's ability to learn complex spatial relationships. The GNN was trained end-to-end using supervised learning, minimizing mean squared error between predicted and observed flows. The training process involved iterative message passing among nodes to aggregate spatial information. In addition to spatial features, the GNN model incorporated a temporal aggregation layer that processed a sequence of past rainfall data (up to 7 days) to capture long-term dependencies and delayed hydrological responses.

2.3 Feature Engineering

Feature engineering was critical to improve model performance. The following techniques were applied.

- Lagged rainfall data were included to account for the time it takes for rainfall to translate into increased river flow.
- Three-day rolling averages of rainfall were used to smooth out short-term fluctuations and emphasize the impact of sustained rainfall events.
- For time-series models like GNN, sliding window approach was used to create sequences of input data, allowing the model to learn from the temporal order of rainfall events.

2.4 Model Evaluation and Validation

The performance of each model was rigorously evaluated using both regression and classification metrics.

- Root Mean Square Error (RMSE) was used as the primary metric to quantify the prediction error, with lower values indicating better model performance.
- R-squared (R^2) provided a measure of the variance in the observed data explained by the model, indicating the goodness of fit.
- A confusion matrix was generated for each model's prediction accuracy in flow values.
- The Random Forest model served as a strong baseline, capturing non-linear relationships effectively but showing limitations in handling temporal dependencies. The KNN model performed well at localized flow predictions but struggled with sudden changes in flow rates. The GNN model outperformed both, demonstrating superior predictive capability by leveraging spatial and temporal dependencies, particularly during prolonged rainfall events.

2.5 Long-Term Prediction and Enhancements

To extend the predictive horizon to one week, the GNN model was tailored for long-term forecasting. The model provided maximum flow in 7 days' time as the output, which is critical for flood management and early warning systems.

3 RESULTS

The proposed framework demonstrated a marked improvement in flood prediction accuracy, with the GNN model outperforming conventional approaches such as Random Forests, achieving a 50% reduction in RMSE and a fivefold increase in R^2 compared to traditional models. Additionally, high- flow accuracy improved significantly, with GNN achieving 85%, representing a 15% enhancement over KNN and a 21% improvement over Random Forest, underscoring its superior performance in predicting extreme events. A time-series plot of predicted vs actual MaxFlow showed visibly closer peak alignment after the inclusion of elevation. Table 2 compares the models' performance using key metrics (RMSE and R^2), and their ability to classify high-flow events.

Table 2. The performance comparison of the four models

Metric	Random Forest	KNN	LSTM	GNN
RMSE (Testing)	61.44	45.21	34.89	30.57
R-squared (R^2)	0.14	0.31	0.59	0.68
High-Flow Accuracy	70%	78%	82%	85%

The Random Forest model captured non-linear relationships between rainfall and river flow but struggled with temporal dependencies, leading to reduced accuracy during peak flow events. In contrast, KNN performed better on short-term forecasts, yet its reliance on historical patterns made it less reliable during unusual weather conditions. LSTM excelled at capturing temporal patterns, adeptly handling the delayed impact of rainfall on river flow, but it was limited in modeling spatial dependencies between rainfall stations. The Graph Neural Network (GNN) outperformed all other models by effectively leveraging both spatial and temporal data. Its graph-based architecture captured the complex interactions between rainfall events and downstream flow, making it the most accurate and reliable model for flood forecasting in the study. The comparison of the predictions and actual values shown in Figure 3. Although the model did not capture the exact values similar to the actual values, it is noteworthy given the loss of data in the inputs, the model was able to capture the pattern of the variation. Furthermore, this can be integrated to predict water levels and predicting flood or non-flood scenario using a water level threshold.

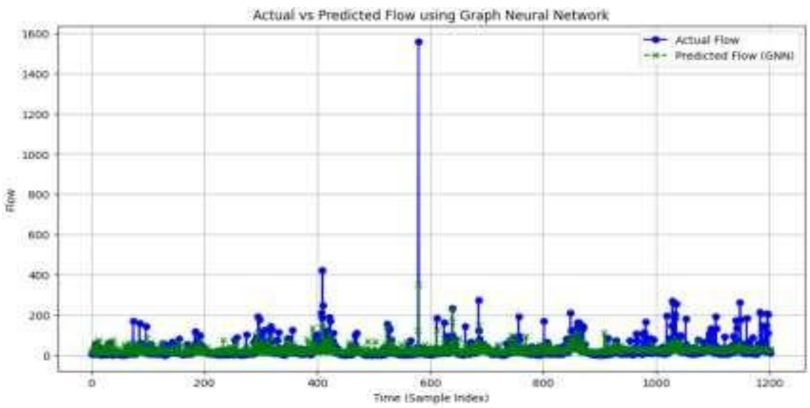

Figure 3. Actual vs. Predicted Flow at Pitabeddara Station using Graph Neural Network

4 CONCLUSION

This study presents a novel machine learning-based framework for flow prediction in the Nilwala Basin, integrating Random Forest, KNN, and GNN models. The GNN model, in particular, showed strong performance in capturing spatial and temporal patterns, making it a valuable tool for early warning systems. The results indicate that the proposed methodology significantly improves prediction accuracy compared to traditional hydrological models.

The following are the key findings of this project.

- Currently, the model predicts only flow values, but for it to be more useful in flood risk management, it needs to be updated to predict floods using a proper threshold. Defining and incorporating flood thresholds based on historical flood data or established flood levels will allow the model to classify high risk events more accurately.
- Integration of Graph Neural Network can provide a proper forecasting network, by treating time-series data as a graph.
- The combination of RF, KNN, and GNN models provides a comprehensive approach to handling the complexities of flood prediction.

- GNN’s ability to model spatial relationships enhances the prediction of extreme flood events, offering a practical solution for disaster preparedness.
- Data quality and completeness remain critical to model performance, highlighting the need for improved data-collection infrastructure.
- The integration of elevation data helps address topographical variability in the basin, enabling more precise predictions of flow and flood hydrographs.

REFERENCES

Abeysiriwardana, H. D., & Wijesekera, N. T. S. (2022). Application of GIS and Logistic Regression for Flood Susceptibility Mapping in Nilwala River Basin, Sri Lanka. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, *55*(2), 1. https://doi.org/10.4038/engineer.v55i2.7503

Basnayake, B. M. L. A. (2022). Prioritization of Sub-basins of Nilwala River Basin for Flood Exposure using Multi Criteria Decision Analysis. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, *55*(1). https://doi.org/10.4038/engineer.v55i1.7483

Dhanapala, L., Gunarathna, M. H. J. P., Kumari, M. K. N., Ranagalage, M., Sakai, K., & Meegastenna, T. J. (2022). Towards Coupling of 1D and 2D Models for Flood Simulation—A Case Study of Nilwala River Basin, Sri Lanka. *Hydrology*, *9*(2), 17. https://doi.org/10.3390/hydrology9020017

Ganjirad, M., & Delavar, M. R. (2023). FLOOD RISK MAPPING USING RANDOM FOREST AND SUPPORT VECTOR MACHINE. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *X-4/W1-2022*, 201–208. https://doi.org/10.5194/isprs-annals-X- 4-W1-2022-201-2023

Ke, Q., Tian, X., Bricker, J., Tian, Z., Guan, G., Cai, H., Huang, X., Yang, H., & Liu, J. (2020). Urban pluvial flooding prediction by machine learning approaches – a case study of Shenzhen city, China. *Advances in Water Resources*, *145*, 103719. https://doi.org/10.1016/j.advwatres.2020.103719

Kumara, M. T., Madushanka, W. D. K., & Nandaseela, S. de S. (2018). *Spatial Distribution of Floods in Mathara District: With Special Focus on 2003 and 2017 Flood Events*. https://ir.kdu.ac.lk/handle/345/2460

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water*, *10*(11), 1536. https://doi.org/10.3390/w10111536

Shi, J., Stebliankin, V., Wang, Z., Wang, S., & Narasimhan, G. (2023). *Graph Transformer Network for Flood Forecasting with Heterogeneous Covariates* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2310.07631

Subramanya, K. (2013). *Engineering Hydrology* (Fourth Edition). https://www.goodreads.com/book/show/24967969-engineering-hydrology

Zhao, X., Wang, H., Bai, M., Xu, Y., Dong, S., Rao, H., & Ming, W. (2024). A Comprehensive Review of Methods for Hydrological Forecasting Based on Deep Learning. *Water*, *16*(10), 1407. https://doi.org/10.3390/w16101407