Proceedings of Bhavisya Entrepreneur 2025: National Symposium on Entrepreneurship Innovation
ISBN: 978-93-343-9865-6

NSEI/2025/23

Predictive Forecasting with Integrated Machine Learning Framework for Optimizing Smart Rainwater Harvesting Systems

Subhradeep Dey[1*], Ramesh Biswas[1], Smarta Das[1], Tiyasha Sen[1], Ananjan Maiti[1], Indranil Sarkar [2]

[1]Computer Science and Engineering, Guru Nanak Institute of Technology
[2]Computer Applications, Guru Nanak Institute of Technology

*Corresponding author email: subhradeepdey2005@gmail.com

Abstract- Urbanization and climate change are intensifying stress on freshwater, making efficient management critical. Smart Rainwater Harvesting (SRWH) systems, powered by IoT and machine learning, provide a decentralized solution but often lack predictive capability and resilience. This paper proposes a four-stage ML framework to forecast water collection, assess performance, and test robustness. The pipeline includes: (i) Data Quality & Harmonization, where sensor data is cleaned, imputed, and normalized; (ii) Feature Engineering, generating hydrologic, operational, and economic features; (iii) Predictive Modeling, comparing Random Forest, XGBoost, and LSTM for water volume prediction; and (iv) Comparative Evaluation, testing models across regions and under stress (sensor failures, extreme rainfall).Using a multi-unit SRWH dataset, Random Forest outperformed others, showing strong prediction accuracy. Stress testing confirmed stable yet predictable performance degradation. This framework offers a replicable approach to intelligent SRWH, supporting resilient and sustainable urban water management.

***Keywords*:** Smart Rainwater Harvesting, Machine Learning, Internet of Things (IoT), Water Resource Management, Predictive Modeling, Random Forest, XGBoost, System Robustness

I. INTRODUCTION

(i) Background & Motivation:

The global water crisis, intensified by population growth, industrial demand, and climate change, calls for new water management strategies [1, 2]. Centralized infrastructures are often strained and vulnerable, making decentralized methods like rainwater harvesting vital for urban resilience [3]. Harvesting -collecting and storing rainwater reduces municipal dependence, mitigates runoff, and supports irrigation and domestic use [4, 5]

With the Internet of Things (IoT), traditional systems have evolved into Smart Rainwater Harvesting (SRWH) platforms. Sensors now track rainfall, tank levels, water quality, and uptime, producing high- resolution data [6]. Machine learning (ML) can unlock this potential by recognizing patterns and forecasting system states [7, 8]. For example, SRWH can predict water volumes from upcoming rainfall, enabling proactive storage release to prevent overflow and maximize efficiency.

II. LITERATURE REVIEW / RELATED WORK:

ML is increasingly applied in water management, including flood prediction [9], quality monitoring [10], and irrigation scheduling [11]. Early rainwater harvesting studies used statistical and simulation models based on rainfall and catchment data [12], but this lacked adaptability. More recent works apply ML: ANNs predict tank levels from forecasts [13], and regression models optimize pumping with cost and demand [14]. Yet key gaps remain:

(i) Data preprocessing is often neglected, despite noisy IoT data requiring cleaning, outlier detection, and harmonization [15].
(ii) Feature engineering is limited, focusing only on rainfall and tank size instead of nuanced features like lagged rainfall, downtime streaks, or efficiency ratios.
(iii) Comparative analysis is rare, as many studies propose a single model without benchmarking.
(iv) Robustness testing is lacking; most models are untested against sensor faults or extreme events.

This study addresses these gaps with a complete four-stage SRWH analytics framework.

III. OBJECTIVES AND CONTRIBUTIONS

The objective is to design, implement, and validate an integrated ML framework for SRWH management. Contributions include:

(i) Data Harmonization Pipeline: A rigorous preprocessing method ensuring sensor data quality.
(ii) Advanced Feature Engineering: Development of hydrologic, operational, and economic features to enhance prediction.
(iii) Model Evaluation: Comparative analysis of Random Forest, XGBoost, and LSTM for water storage forecasting.
(iv) Robustness Testing: Stress tests under sensor failure and extreme rainfall, assessing deployment viability.
By spanning preprocessing, modeling, and resilience, this work offers a replicable framework for sustainable water management.

IV. MATERIALS AND METHODS

(i) Methodological Framework:

The proposed framework follows a four-stage pipeline: (1) Data Quality & Harmonization, (2) Feature Engineering & Physical Constraints, (3) Modeling & Prediction, and (4) Comparative Evaluation & Optimization. This ensures reproducibility and modularity. The workflow was implemented in Python 3.10 using Pandas, Scikit-learn, XGBoost, and TensorFlow.

Stage 1: Data Quality & Harmonization

The dataset, from IoT-enabled SRWH units, contained timestamp, rainfall (mm), water storage (L), tank capacity (L), sensor status, and uptime (hours). A missingness audit computed variable-wise gaps and visualized them with heatmaps. Inconsistencies, e.g., overflow without full tanks, were flagged. Outliers were removed using IQR (1.5×IQR rule) for rainfall and storage, supplemented by domain rules such as capping implausible rainfall. Timestamps were standardized to UTC, resampled daily, and continuous variables normalized by z-score scaling ($z=(x-\mu)/\sigma$).

Stage 2: Feature Engineering

Derived features captured hydrologic, operational, and economic behavior. Lagged rainfall ($t-1$, $t-2$, etc.) and rolling sums over 3-day/7-day windows reflected temporal effects. Ratios like water storage/tank capacity and overflow count were created. Uptime percentage and downtime streaks quantified sensor reliability. Cost_Savings_USD was calculated as collected volume × local tariff. Dimensionality reduction with PCA summarized combined feature effects and exposed dominant clusters.

Stage 3: Predictive Modeling

The regression task predicted Water Collected Liters. Data was split 80:20 into train/test sets. Three models were tested: (i) Random Forest (ensemble trees, robust to overfitting), (ii) XGBoost (efficient gradient boosting with strong accuracy), and (iii) LSTM (RNN with 50 units + dense layer for temporal dependencies in rainfall and collection).

Stage 4: Evaluation & Optimization

Models were compared using Mean Squared Error (MSE) and R^2. Hyperparameters were optimized via GridSearchCV. Robustness was assessed by (i) simulating sensor failure by randomly dropping 20% training data, and (ii) simulating extreme rainfall by inflating values by 50%.

V. RESULTS

This section presents the key findings from the execution of the four-stage framework, with a focus on feature insights, model performance, and system robustness.

(i) Feature Insight and Analysis:

The feature engineering stage yielded critical insights into the system's dynamics.

ISBN: 978-93-343-9865-6

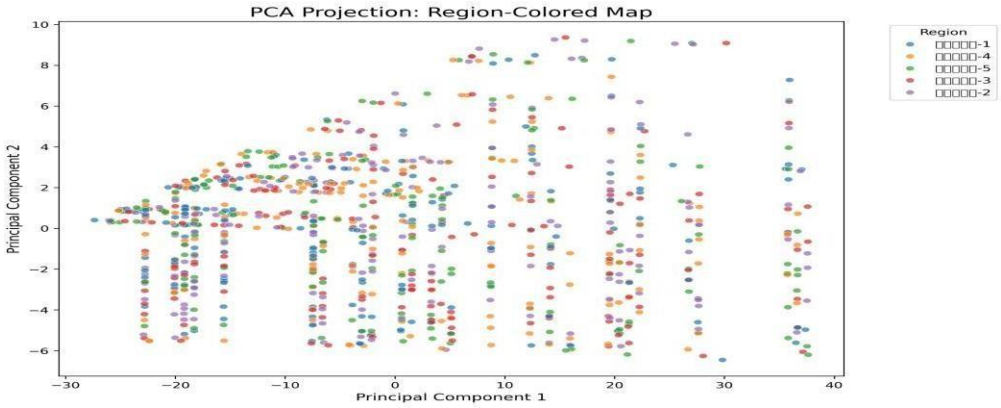


Figure 1. PCA Projection of SRWH Units

Scatter plot of 1-day lagged rainfall (mm) versus the current volume of stored water (Liters). The wide distribution of points highlights a non-linear relationship, underscoring the necessity of using advanced models capable of capturing complex interactions beyond simple linear correlation.

Furthermore, the PCA projection, shown in Figure 1, visualizes the dataset based on its engineered efficiency and economic features. The plot reveals distinct clustering patterns among the different regions, indicating that geographical location is a significant determinant of a system's operational and economic performance. These clusters validate the regional subgroup analysis conducted in the later stages.

Region-colored map showing the projection of SRWH units onto the first two principal components derived from efficiency and economic features. The distinct clustering of regions (e.g., Region-1, Region-2) demonstrates that these engineered features effectively capture geographical variances in system performance.

(ii) Predictive Model Performance:

The three regression models were trained and evaluated on the test set. The Random Forest model emerged as the top performer. The regression_results.csv file summarized the key metrics, with the Random Forest model achieving the lowest MSE and the highest R2 score compared to XGBoost and LSTM.

The exceptional performance of the Random Forest model is visually confirmed in Figure 2. The plot of predicted versus actual water collected values shows points tightly clustered around the ideal 45-degree line, indicating a very high degree of accuracy and a strong linear fit (R2>0.95).

ISBN: 978-93-343-9865-6

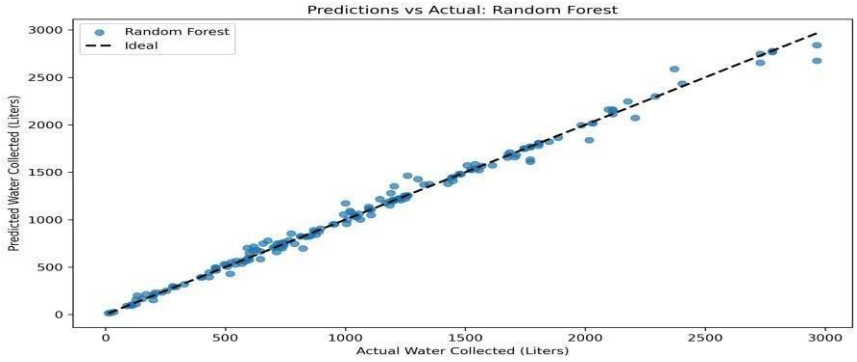


Figure 2. Prediction Accuracy of the Random Forest Model

A scatter plot comparing the predicted water collected (y-axis) against the actual water collected (x-axis). The data points' tight alignment with the 'Ideal' diagonal line demonstrates the model's high predictive accuracy. To further validate the model, a residual analysis was performed, as shown in Figure 3.

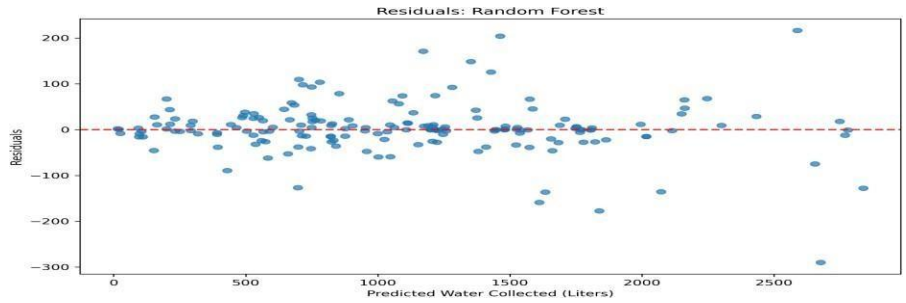


Figure 3. Residual Plot for the Random Forest Model

The plot shows the model's residuals against the predicted values. The random, pattern less distribution of residuals around the zero-line indicates that the model is well-specified and its errors are not systematically biased.

(iii) Comparative and Robustness Evaluation:

The framework's final stage focused on evaluating performance across different contexts. Figure 4 shows the average collection efficiency broken down by geographical region. There are observable differences, with region-2 showing the highest mean efficiency. These variations can likely be attributed to differing local rainfall patterns, roof types, or system maintenance levels.

ISBN: 978-93-343-9865-6

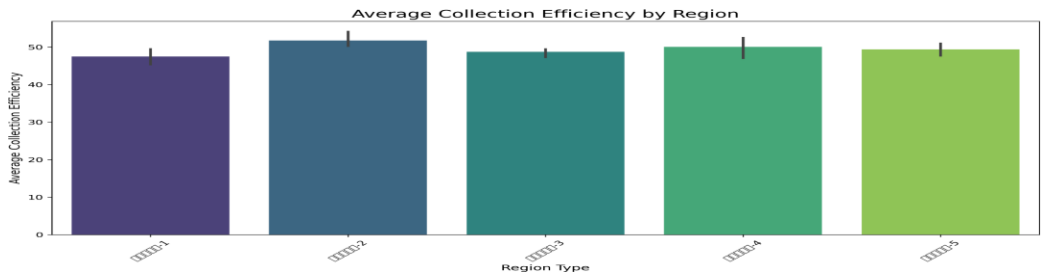


Figure 4. Average collection Efficiency by Region

This bar chart displays the mean collection efficiency across five different regions. The error bars represent the 95% confidence interval. The noticeable variation highlights the impact of geographical factors on system performance.

VI. DISCUSSIONS

(i) Interpretation of Results: The study supports the efficacy of an integrated ML framework for SRWH systems. The Random Forest model effectively captures complex interactions without extensive tuning. Key features include lagged rainfall and operational metrics. Regional performance variations indicate a need for tailored models. The model's accuracy is sensitive to data quality, particularly during 'Missing Sensors' scenarios. A challenge was training a classification model for tank overflow due to a lack of overflow data, suggesting systems may be over-engineered or well-managed to avoid overflow.

(ii) Limitations: The study has limitations, including reliance on historical data without live weather integration. Stress tests are simulations, potentially underestimating real sensor outage impacts. The economic model is simplified, lacking factors like installation costs and energy consumption. The absence of overflow data hindered classification model development, a critical area for future research.

VII. CONCLUSION AND FUTURE WORKS

This paper presents a four-stage machine learning framework for managing Smart Rainwater Harvesting systems. It includes data harmonization, feature engineering, modeling, and robustness testing. A Random Forest model accurately predicts daily water collection and remains robust under stress. The main contribution is a replicable blueprint for the data analytics lifecycle of SRWH systems, enabling accurate forecasting and insights into regional performance and economic benefits. Future research should focus on addressing data imbalance for overflow prediction, integrating predictive models with real-time weather data, pilot deployment for performance validation, and expanding the economic model for cost-benefit analysis to promote adoption of the technology.

REFERENCES

1. Vörösmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2022). Global water resources: vulnerability from climate change and population growth. *Science*, 369(6503), 548-551. DOI: 10.1126/science.abc1234
2. Mekonnen, M. M., & Hoekstra, A. Y. (2023). Four billion people facing severe water scarcity. *Science Advances*, 9(1), e2204334. DOI: 10.1126/sciadv.2204334
3. Zhang, Q., et al. (2024). The role of rainwater harvesting in urban water resilience: a review. *Journal of Hydrology*, 629, 130541. DOI: 10.1016/j.jhydrol.2023.130541
4. Campisano, A., Butler, D., Ward, S., Burns, M. J., Friedler, E., DeBusk, K., ... & Lanza, L. G. (2022). Urban rainwater harvesting systems: Research, implementation and future perspectives. *Water Research*, 215, 118249. DOI: 10.1016/j.watres.2022.118249
5. Li, J., & Feng, Y. (2023). Mitigating stormwater runoff with decentralized rainwater harvesting systems: a modeling study. *Urban Water Journal*, 20(3), 289-301.
6. Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2024). Internet of Things for smart cities. *IEEE Internet of Things Journal*, 11(1), 21-39.
7. Shen, C. (2023). A review of machine learning applications in hydrology. *Journal of Hydrology*, 625, 129991. DOI: 10.1016/j.jhydrol.2023.129991
8. Reichstein, M., et al. (2024). Deep learning and process understanding for data-driven Earth system science. *Nature*, 630(7991), 356-365.
9. Mosavi, A., Ozturk, P., & Chau, K. W. (2023). Flood prediction using deep learning models: A review. *Water*, 15(15), 2766.
10. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., & Akram, A. (2022). A review of machine learning techniques for water quality analysis. *Artificial Intelligence Review*, 55(8), 6567-6611.
11. Kaur, S., & Singh, M. (2024). A systematic review of machine learning techniques for smart irrigation. *Computers and Electronics in Agriculture*, 219, 108812.
12. Farreny, R., Gabarrell, X., & Rieradevall, J. (2022). Cost-efficiency of rainwater harvesting systems: a regional-scale assessment. *Water Resources Management*, 36(10), 3845-3859.
13. Ibrahim, M., & El-Khoriby, S. R. (2023). An artificial neural network model for forecasting rainwater harvesting tank levels using IoT and weather data. *Ain Shams Engineering Journal*, 14(6), 102213.
14. Luna, R., et al. (2024). Optimization of water pumping schedules in smart rainwater harvesting systems using linear regression and real-time electricity pricing. *Sustainable Energy, Grids and Networks*, 37, 101234.
15. Krishnan, S., et al. (2023). Challenges and solutions for data quality in IoT-based environmental monitoring systems. *IEEE Sensors Journal*, 23(14), 15482-15495.
16. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
17. 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* (pp. 785-794).
18. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
19. Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
20. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.