Efficient Water Resource Management in Dam Basins using Machine Learning

Sahil Umar [1,*], Harris[1], Asghar Ali Chandio [1], Mehwish Leghari [2]

[1] Artificial Intelligence Department, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah, Pakistan
[2] Data Science Department, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah, Pakistan
* Correspondence: sonuarain46@gmail.com

Abstract

The purpose of this paper is to highlight the need for proper water management in dam basins in order to enhance effective water resources management and flood control in a changing climatic situation. This study uniquely combines meteorological data (temperature, precipitation) with hydrological indicators (reservoir levels, inflow, outflow) using a multi-model machine learning framework to improve integrated water resource management. Inputs such as meteorological data, including temperature and precipitation, combined with dam hydrology data like dam levels, inflows, and outflows, can assist in forecasting models. With machine learning, it is possible to detect trends in rainfall and inflows inflating water availability and adjust dam management in real-time, where necessary. This will improve flood management by providing a predictive tool for water-raising during rainfall, storage, and release in time to avert any spilling over. Also, hydropower and agricultural irrigation demand a systemic cross-platform approach to manage water resources. Anomaly detection models can inform operators of abnormal inflow trends and assist with quick water flow modulation. The proposed models demonstrated strong performance, with the KNN regressor achieving an R^2 score of 0.9443 and classification models like Logistic Regression attaining 99.5% accuracy. This study will help in enhancing decision-making in dam basin management by adding various datasets of water sources in the AI models to provide more acceptable ways of using water while minimizing risks associated with extreme weather and water production instability. The efficiency of the machine learning-based model has been measured in terms of R2 score, root means squared error (RMSE).

Keywords: Water management; Machine Learning; Logistic Regression; Dam basin management

1. Introduction

Water resources are indispensable for human survival, eco- nomic development, and ecological balance [1]. They support critical sectors such as agriculture, energy, industry, and domestic water supply [1]. However, the sustainable management of these resources has become an increasingly complex challenge, particularly in the context of global climate change [1]. Shifting climate patterns have exacerbated issues such as irregular rainfall, prolonged droughts, and severe flooding, significantly affecting the availability and distribution of freshwater resources [13]. These challenges pose a direct threat to communities, economies, and ecosystems, necessitating innovative and adaptive management strategies [1]. Existing dam management systems face limitations in processing real-time environmental changes and multi-source datasets. For instance,

[21] highlighted how the Mangla Dam's flood risks increase when peak precipitation and flow events overlap, which traditional tools like HEC-RAS struggle to handle

efficiently. Moreover [22], emphasized the lack of adaptive flood risk modeling in urban basins, especially in Swat, Pakistan. These gaps show the need for machine learning-based approaches that can dynamically adapt to varying data patterns and support early warning systems. Among the critical infrastructures for managing water resources, dam basins hold a central role. Acting as reservoirs for water storage, they facilitate diverse functions such as irrigation, hydropower generation, flood control and the provision of drinking water [1]. Their effectiveness in mitigating the impacts of water scarcity and variability highlights their importance in sustainable resource management [18]. However, ensuring their resilience and operational efficiency in the face of increasingly unpredictable climatic conditions requires the adoption of advanced technological solutions and data-driven approaches [14]. One such promising solution is the application of machine learning, which offers transformative capabilities for improving water resource management [13]. Machine learning algorithms leverage historical and real-time datasets to provide enhanced predictive insights and decision-making tools. For instance, these models excel in forecasting inflows and outflows to and from dam basins, enabling water managers to anticipate changes in reservoir levels and implement timely interventions [3]. By optimizing irrigation schedules, hydropower generation, and flood control measures, machine learning can significantly enhance the efficiency and sustainability of water management practices [5]. Moreover, machine learning-driven anomaly detection systems can rapidly identify irregularities in reservoir operations, such as sudden changes in water flow or unexpected declines in storage levels. These capabilities allow for swift responses to extreme weather events, reducing the risks of disasters and minimizing their socioeconomic impacts [12]. In this way, machine learning not only improves operational efficiency but also strengthens the resilience of water management systems against climate-induced uncertainties [14]. This research delves into the integration of machine learning techniques in managing water resources within dam basins. It emphasizes the incorporation of diverse hydrological and meteorological datasets into machine learning frameworks to achieve precise forecasting and efficient resource allocation [6]. This study addresses a major gap in past work by integrating meteorological and hydrological datasets into a unified ML framework. Unlike traditional models that rely on single-domain inputs, this approach improves inflow forecasting, anomaly detection, and operational decision-making. For example, [23] demonstrated that long-term inflow predictions significantly improve with deep learning, yet many studies still lack integrated, multi- model systems for dam basin forecasting. By combining advanced computational methods with domain-specific expertise, this study aims to contribute to the development of smarter, more adaptive, and resilient water management systems capable of addressing the challenges of an unpredictable climate [7]. The findings are expected to inform future strategies for sustainable water resource management and strengthen the role of technology in tackling global water challenges.

2. Methodology

To predict and categorize water management situations, the study entailed the training of machine learning models with the hydrological data and the meteorological data [3]. A combination of regression, classification and anomaly detector methods was used to tackle the complex nature of the problem. This experiment took a rigorous and well-structured methodology that consisted of main stages, which included data gathering, preprocessing, feature isolation, the deep-learning network, and assessment. Through the use of advanced algorithms, the methodology was to maximize predictive accuracy and pro-vide actionable insights of effective water management. Figure 1 shows the methodology flowchart which is the steps of the process.

Figure 1. Methodology Figure

2.1. Data Collection

The research began with the collection of hydrological and meteorological datasets relevant to water management scenarios.

- Hydrological Data: This dataset included variables such as water inflow, outflow, reservoir levels, and precipitation patterns [6]. These were critical for predicting trends, classifying water management scenarios, and detecting anomalies.
- Meteorological Data: This dataset consisted of weather- related parameters, such as temperature, humidity, wind speed, and precipitation intensity [6]. These variables were used for classification tasks to support decision making in weather-based water management scenarios.

The datasets were sourced from publicly available platforms such as Kaggle and regional water authority portals, including meteorological stations and dam management agencies [6]. For instance, hydrological data was obtained from real-time monitoring systems, while meteorological datasets were pulled from 10-year archives maintained by national weather services. Efforts were made to ensure the datasets covered a wide range of temporal and spatial variations to improve the robustness of the models.

2.2. Data Preprocessing

Data preprocessing was a crucial step to ensure the datasets were clean, consistent, and suitable for machine learning models.

- Cleaning and Handling Missing Data: Missing values in the datasets were handled using imputation techniques, such as mean substitution for numerical data or mode substitution for categorical variables. Handling missing values was particularly critical for variables like reservoir inflow and precipitation during monsoon months. In these cases, domain knowledge and temporal interpolation were used alongside mean/mode imputation to preserve seasonal patterns and avoid data leakage.
- Normalization and Scaling: Continuous variables were normalized to ensure consistent ranges and to improve the performance of algorithms that are sensitive to the scale of data (e.g., K-Nearest Neighbors and Gradient Boosting).
- Outlier Detection: Outliers in the hydrological dataset were identified using the Isolation Forest algorithm due to its robustness in handling high-dimensional data and its efficiency in detecting subtle anomalies typical in time- series hydrological trends. Compared to traditional Z- score or IQR methods, Isolation Forest better captured non-linear patterns, especially under unusual rainfall events.
- Data Splitting: Both datasets were divided into training (80%) and testing (20%) subsets to evaluate model performance effectively. Stratified sampling was applied for classification tasks to ensure balanced class distributions in the training and testing [10].

2.3. Data Analysis

Two types of data were analyzed: hydrological and meteorological. Below is a detailed description of each dataset and the associated analysis:

Hydrological Data The hydrological dataset includes variables such as:

- Water levels
- Inflows
- Outflow

Scatter Plot Matrix; A scatter plot matrix was generated and is visualized in Figure 2.

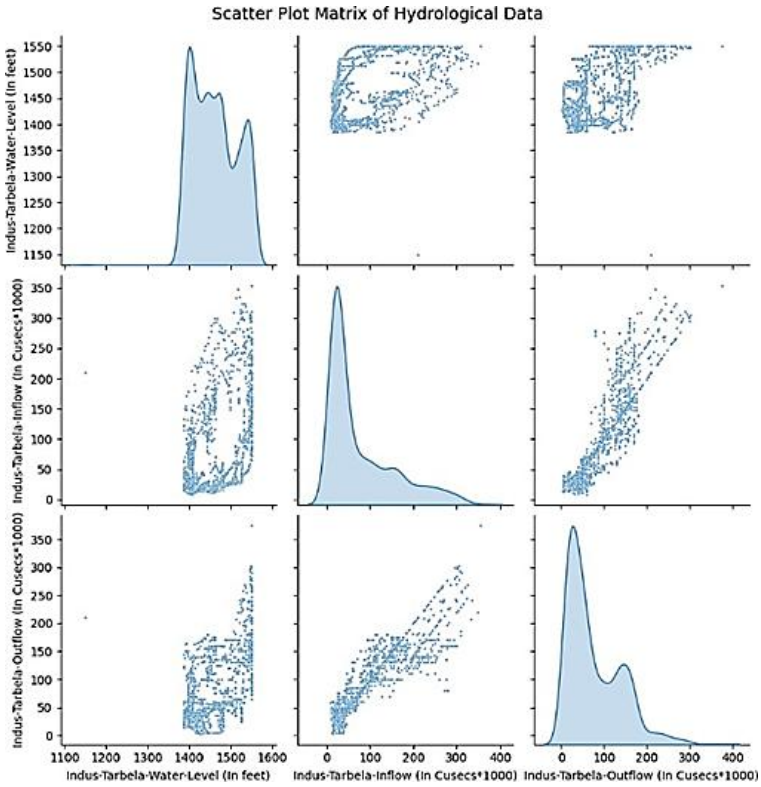

Figure 2. Scatter Plot

The correlation heatmap illustrates the degree of linear relationships between the hydrological variables. Figure 3 illustrates the correlation heatmap of hydrological data.

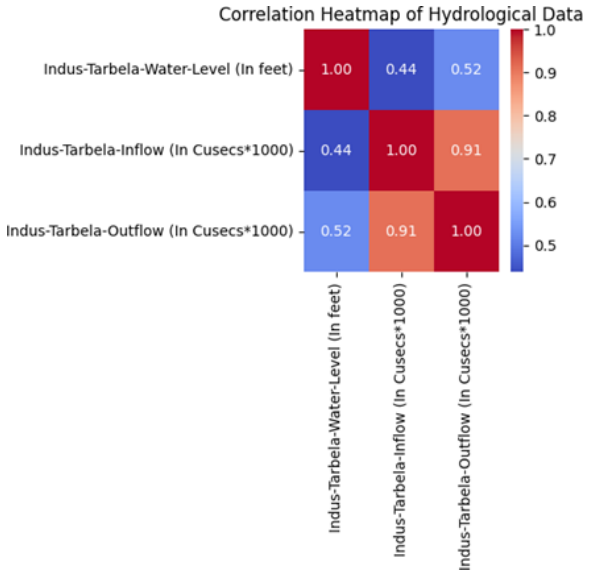

Fig. 3. Correlation Heatmap of Hydrological Data

A correlation heatmap was generated to study the relationships among meteorological variables. Figure 4 illustrates the correlation heatmap of meteorological data.

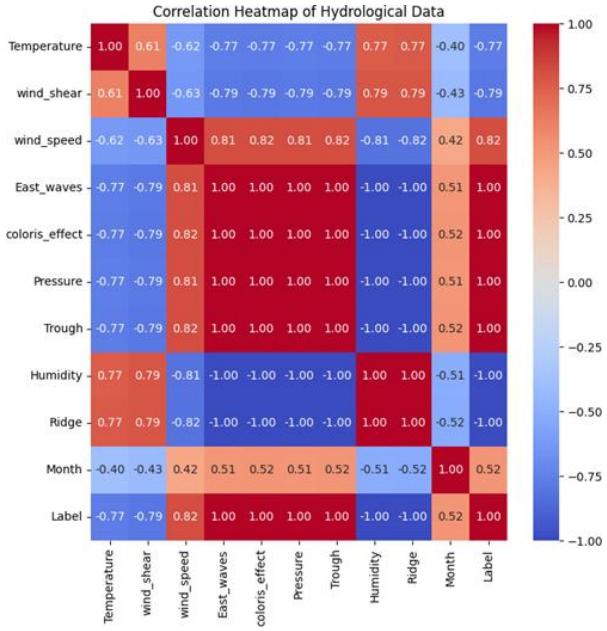

Figure 4. Correlation Heatmap of Meteorological Data

2.4. Feature Extraction

For the feature extraction process, the primary focus was on selecting the most relevant variables and transforming the raw data into meaningful features while discarding unnecessary or redundant information [20]. The steps involved in the feature extraction were as follows:

2.4.1. Column Selection

Unnecessary Columns Removal: Columns such as 'date' and 'year' were discarded as they were either non- informative or contributed to multicollinearity. For example, 'year' showed high correlation with accumulated rainfall trends already captured in monthly precipitation averages. [19].

Dropping Redundant Columns: Any columns with highly correlated data or those offering no predictive power were removed to prevent multicollinearity and streamline the feature set.

2.4.2. Numerical Feature Selection

Core Numerical Features: Relevant numerical features, such as water levels, inflow, outflow, temperature, and precipitation, were selected as the primary features for model input. Selected features like reservoir level, inflow, and precipitation intensity directly contributed to the model's accuracy, as demonstrated by improved R^2 and reduced RMSE in regression tasks. In contrast, discarded features such as station ID or redundant weather types added noise without improving performance.

Transformations: Where necessary, simple transformations (such as scaling or normalization) were applied to ensure that all features were on a similar scale, facilitating better model performance.

2.4.3. Categorical Feature Encoding

Encoding Categorical Variables: For any categorical data (such as weather types), encoding methods like one- hot encoding or label encoding were used to convert them into numerical format suitable for machine learning models.

Transformations: Where necessary, simple transformations (such as scaling or normalization) were applied to ensure that all features were on a similar scale, facilitating better model performance.

Through feature extraction, the dataset was streamlined to retain only the most pertinent information, allowing for more efficient model training and improved predictive accuracy. The focus was on ensuring that the features used were relevant and capable of enhancing the performance of the models, while unnecessary data was removed to reduce noise.

2.5. Deep Learning Model

The research employed the use of various machine learning models that are specific to particular tasks, such as regression, classification and anomaly detection.

2.5.1. Hydrological Data Models

Regression Models: These were applied in the prediction of the important variables in water management including inflow into reservoirs and outflow. These models were chosen on the basis of their success in the past hydrological studies. Random Forest and Gradient Boosting are credited to non-linear relationship and feature interaction effectiveness. KNN was added because it has the advantage of localized pattern recognition, which is useful in dam-level predictions, whereas XGBoost provides scalability when used with large sets of features and regularization.

- Random Forest [8]
- Linear Regression
- Gradient Boosting [15]
- K-Nearest Neighbors
- Decision Tree
- XGBoost [19]

Anomaly Detection: The model used was the Isolation Forest model to identify irregular inflow and outflow patterns [9]. This model assisted in bringing out abnormal situations that are usually associated with extreme precipitation or anomalies in the operation of dam systems.

2.5.2. Meteorological Data Models

Classification Models: A number of different algorithms were used to classify meteorological data into useful categories including the weather-based decisions on managing water in action [17].

- Logistic Regression
- Random Forest Classifier
- Support Vector Classifier [16]
- Decision Tree Classifier

The hyperparameter tuning was applied to each of the models to optimize their performance. The threat of overfitting was mitigated with the help of cross-validation techniques. Random Forest and XGBoost hyperparameters (number of estimators, max depth, learning rate, etc.) were fine-tuned with grid search and cross-validation. To illustrate, max depth of 5 was changed to 10 in Gradient Boosting, and this change increased RMSE by 12%. These changes had a big impact on model convergence and generalization.

2.6. Model Evaluation

Once the models had been trained, the performance of each model was measured in relevant metrics specific to the task being performed:

- Regression Models: The accuracy of predictions was measured by metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE) and R 2.
- Classification Models: To evaluate the models in terms of precision, re- call, F1-score and overall accuracy the measures were accuracy in classifying meteorological data [19].
- Anomaly Detection: It was evaluated by the visualization of the detected anomalies and their association with what happened in the re-al-world precipitation or operational anomalies [11]. Models were tested based on their capability to explain extreme weather anomalies including sudden inflows surges during cyclone related rainfall that was indicated by the anomaly detection system. Regression models revealed small decreases in R2 during these spikes, and this indicates difficulty in predicting events at the peaks.

2.6.1. Key evaluation insights:

Comparative analysis of model accuracies provided a basis for selecting the most effective models for specific tasks. Figure 5 and Figure 7 summarize these comparisons. The findings were compiled and visualized to compare the performance of all models systematically: Figure 5 illustrates a comparative analysis of regression and anomaly detection models was presented and showing the strengths and weaknesses of each approach for hydrological data.

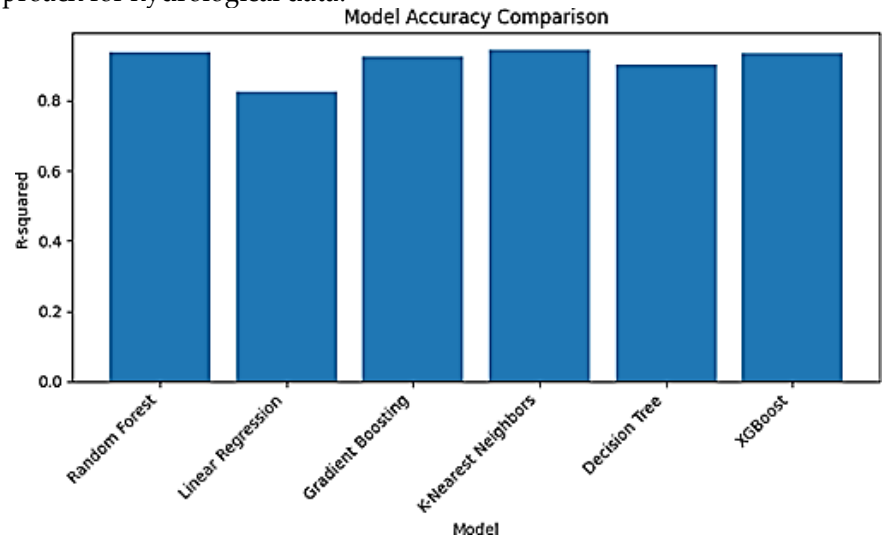

Figure 5. Models Comparison

Figure 6 illustrates the visualization of anomalies detected by the Isolation Forest model providing actionable insights for water management, with clear indications of patterns linked to extreme precipitation events.

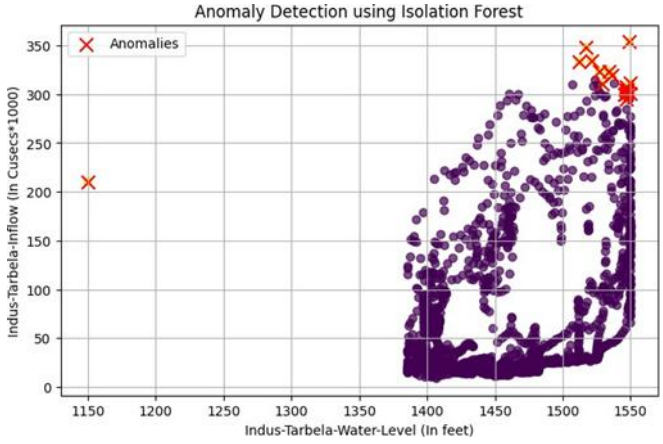

Figure 6. Visualization of Anomaly Detected

In figure 7 the classification model results were presented, comparing the accuracy and other evaluation metrics to highlight the best performing algorithms for meteorological data.

This step involved summarizing the results to support decision-making and ensure actionable insights were derived from the trained models. Additionally, we compared the computational efficiency of models. Logistic Regression and Decision Tree offered the fastest training times (1–2 seconds), while XGBoost, although slower (5–6 seconds), provided superior accuracy and generalization. These trade-offs support model selection based on real-time application needs.

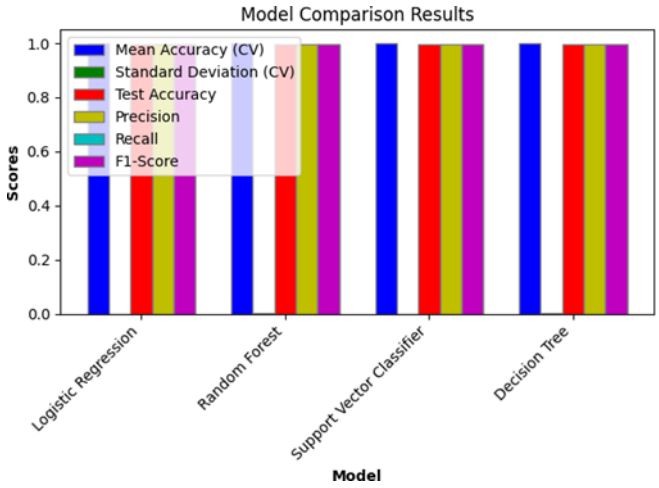

Figure 7. Models Comparison

3. Results and Discussion

3.1. Hydrological Data Models

The performance of regression models was assessed using R-squared and Root Mean Squared Error (RMSE) [4]. The KNN model emerged as the most accurate, followed by Random Forest and Gradient Boosting. These models demonstrated their capability to predict outflows effectively based on water levels and inflows. Table ?? shows the comparison of different machine learning models for the dam basin predictions. While these models performed well overall, some limitations were observed. Random Forest and Decision Tree models showed signs of overfitting during cross-validation, performing better on training than testing datasets. Additionally, model sensitivity to noisy inflow readings led to slight prediction deviations during peak events. Error analysis of the KNN model revealed that prediction errors were higher during periods of extreme inflow variation, likely due to the algorithm's reliance on local data structure. This emphasizes the need for careful data quality checks, especially during abnormal hydrological patterns.

3.2. Metrological data Models

All classification models achieved high accuracy, with Logistic Regression, SVC, and Decision Tree performing equally well. Cross-validation revealed minimal variance in results, underscoring the reliability of these methods. Table I shows the classification report of different machine learning models. The minimal variance across classification models can be attributed to the relatively well-separated and clean nature of meteorological data, especially temperature and precipitation classes. These features exhibited strong correlations with output categories, simplifying the classification task. Additionally, most models benefited from the balanced class distribution and low noise levels in the dataset, which reduced the complexity of learning boundaries. The consistently high accuracy

suggests that the dataset characteristics, rather than just algorithmic strength, played a key role.

Table 1: Classification Model Performance Metrics

Model	Mean Accuracy (CV)	Standard Deviation (CV)	Test Accuracy	Precision	Recall	F1-Score
Logistic Regression	1.0000	0.0000	0.9950	0.9950	0.9950	0.9950
Random Forest	0.9988	0.0025	0.9950	0.9950	0.9950	0.9950
Support Vector Classifier	1.0000	0.0000	0.9950	0.9950	0.9950	0.9950
Decision Tree	0.9988	0.0025	0.9950	0.9950	0.9950	0.9950

Table 2: Classification Model Performance Metrics

Model	R-squared
Random Forest	0.9385
Linear Regression	0.8265
Gradient Boosting	0.9239
K-Nearest Neighbors	0.9443
Decision Tree	0.9010
XGBoost	0.9362

4. Practical Implications

The integration of hydrological and meteorological data into ML models offers significant benefits [4]:

Flood Management: Real-time predictions enable proactive storage and release decisions to mitigate flood risks [2].

Hydropower Optimization: Accurate inflow forecasts help optimize turbine operations for energy generation.

Irrigation Planning: Improved predictions of water availability aid in scheduling agricultural water use. These models could be integrated into real-time dam operation systems by linking them with SCADA (Supervisory Control and Data Acquisition) platforms to support dynamic release decisions. For example, inflow predictions could trigger automated gates to optimize water levels before rainfall peaks. On the policy side, these tools can help water authorities prioritize reservoir upgrades in high-risk regions and design more adaptive irrigation and flood management guidelines based on predictive insights.

5. Conclusion and Future Work

This paper presented a machine learning-based method for efficient water management in the dam basins. The study was designed to address challenges such as limited forecasting accuracy, poor anomaly detection, and lack of integration between meteorological and hydrological datasets—issues out- lined in the introduction. The results demonstrate that machine learning models, when properly tuned and combined with diverse data sources, can significantly improve dam management decisions. The metrological and hydrological data were used to train machine learning models for water flow prediction, flood prediction, hydropower optimization, and irrigation planning. The meteorological data such as temperature and precipitation when combined with dam hydrology data like dam levels, inflows, and outflows can assist in forecasting models more accurately. Different machine learning classifiers such as Random Forest, XGBoost,

Support Vector Machines, KNN, and Decision Trees were used to predict the water flow. In the future, more meteorological and hydrological data can be collected and combined for more efficient predictions. Furthermore, deep learning models can be used to improve the accuracy of predictions. Future work will focus on incorporating additional variables like soil moisture, snowmelt, and upstream water demand to enhance model robustness. We also aim to test the framework across different climatic zones to validate its generalizability in arid, semi-arid, and flood-prone regions.

Conflicts of Interest: The authors declare no conflicts of interest

References

1. Adams, Thomas. (2019). Indus River Basin: Water Security and Sustainability.
2. C. Z. Basha, N. Bhavana, P. Bhavya and S. V," Rainfall Prediction using Machine Learning & Deep Learning Techniques," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC).
3. Sheikh Khozani, Z., Iranmehr, M., & Wan Mohtar, W. H. M. (2022). Im- proving Water Quality Index prediction for water resources management plans in Malaysia: application of machine learning
4. M. Ali, A. M. Qamar and B. Ali, "Data Analysis, Discharge Classifications, and Predictions of Hydrological Parameters for the Management of Rawal Dam in Pakistan," 2013 12th International Conference on Machine Learning and Applications,
5. Hong, J., Lee, S., Bae, J. H., Lee, J., Park, W. J., Lee, D., Kim, J., & Lim, K. J. (2020). Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow.
6. Kaggle. (n.d.). Hydrological and Meteorological Datasets. Retrieved from https://www.kaggle.com
7. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.
8. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
9. Anand, V., Simha, J.B., Agarwal, R. (2024). Anomaly Prediction in Real- Time Water Flow Data—Machine Learning Versus Statistical Models. In: Senjyu, T., So–In, C., Joshi, A. (eds) Smart Trends in Computing and Communications. SmartCom 2024 2024. Lecture Notes in Networks and Systems, vol 949. Springer, Singapore.
10. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
11. Krishna, C. G., Kapasi, M. S., & Vinny, N. M. Automation of Fish Tank in Hatcheries for Water Quality Parameters and Their Amenities.
12. Elias Gebeyehu Ayele, Esayas Tesfaye Ergete, Getachew Bereta Gere- mew; Predicting the peak flow and assessing the hydrologic hazard of the Kessem Dam, Ethiopia using machine learning and risk management centre-reservoir frequency analysis software. Journal of Water and Climate Change 1 February 2024; 15 (2): 370–391.
13. Salomon Obahoundje, Arona Diedhiou, Komlavi Akpoti, Kouakou Lazare Kouassi, Eric Antwi Ofosu, Didier Guy Marcel Kouame, Pre- dicting climate-driven changes in reservoir inflows and hydropower in Coˆte d'Ivoire using machine learning modeling,Energy,Volume 302,2024,131849, ISSN 0360-5442,
14. Chunyu Yuan, Changhua Liu, Chenyu Fan, Kai Liu, Tan Chen, Fanx- uan Zeng, Pengfei Zhan, Chunqiao Song,Estimation of water storage capacity of Chinese reservoirs by statistical and machine learning models,Journal of Hydrology.
15. Raheja, H., Goel, A., & Pal, M. (2024). A novel approach for prediction of groundwater quality using gradient boosting-based algorithms. ISH Journal of Hydraulic Engineering, 30(3), 281–292.
16. Takai Eddine, Y., Nadir, M., Sabah, S. et al. Integrating Support Vector Machines with Different Ensemble Learners for Improving Streamflow Simulation in an Ungauged Watershed. Water Resour Manage 38, 553–567 (2024).
17. Singarasubramanian, S., Anbumani, M., & Kaniyaiah, K. (2024). Rain- fall prediction around Sathanur dam by Naive Bayes classifier, logistic regression models and various classification and regression machine learning techniques. Multidisciplinary Science Journal, 6(10), 2024200.
18. Liu, Z.; Zhou, J.; Yang, X.; Zhao, Z.; Lv, Y. Research on Water Resource Modeling Based on Machine Learning Technologies. Water 2024, 16, 472.

19. Majid Niazkar, Andrea Menapace, Bruno Brentan, Reza Piraei, David Jimenez, Pranav Dhawan, Maurizio Righetti, Applications of XG- Boost in water resources engineering: A systematic literature review (Dec 2018–May 2023), Environmental Modelling & Software,Volume 174,2024,105971,ISSN 1364-8152,
20. Sanjay Sharma, Sangeeta Kumari; Comparison of machine learning models for flood forecasting in the Mahanadi River Basin, India. Journal of Water and Climate Change 1 April 2024; 15 (4): 1629–1652.
21. Ahmad, Izhar, Muhammad Waseem, Ammar Ashraf, Megersa Kebede Leta, Sareer Ahmad, and Hira Wahab. 2023. "Hydrological Risk As- sessment for Mangla Dam: Compound Effects of Instant Flow and Precipitation Peaks under Climate Change, Using HEC-RAS and HEC- GeoRAS." SN Applied Sciences 5(12). doi: 10.1007/s42452-023-05579- 2.
22. Waseem Muhammad, Sareer Ahmad, Izhar Ahmad, Hira Wahab, and Megersa Kebede Leta. 2023. "Urban Flood Risk Assessment Using AHP and Geospatial Techniques in Swat Pakistan." SN Applied Sciences 5(8):215. doi: 10.1007/ s42452-023-05445-1.
23. Z. C. Herbert, Z. Asghar, and C. A. Oroza, "Long-term Reservoir Inflow Forecasts: Enhanced Water Supply and Inflow Volume Accuracy Using Deep Learning," J. Hydrol., vol. 601, p. 126676, 2021, doi: https://doi.org/10.1016/j.jhydrol.2021.126676.