ISSN: 2349-6002

Forecasting River Water Quality Using Random Forest Regression Model: A Real-Time Sensory Approach on The River Ganga

Utkarsh Kumar[1], Samarth Maheswari[2], Samriddhi Shukla[3], Megha Saxena[4]
[1,2,3,4]*Department of Computer Applications, Invertis University*
doi.org/10.64643/IJIRTV12I11-207111-459

***Abstract*—Water quality deterioration is a serious environmental issue, especially in important river ecosystems like the Ganga (holding a very large population it's bank throughout its runoff). Traditional ways of measuring the Water Quality Index (WQI) depends heavily on time-consuming lab tests of factors like biological oxygen demand and coliform bacteria. This leads to a reactive approach in managing pollution. This study focuses on predicting WQI quickly using a simple, machine learning framework. In this research, we have used a Random Forest Regressor model to predict WQI based on real-time, sensor-compatible physicochemical metrics. The dataset contains high-frequency continuous data using five input parameters: Dissolved Oxygen (DO), pH, Oxidation-Reduction Potential (ORP), Conductivity, and Temperature. We reprocessed the data to ensure there were no missing values, creating a strong basis for model training. To make sure the model could understand complex physicochemical relationships without any temporal bias, we have separated the dataset randomly, allocating 80% for training and 20% for testing. The tuned Random Forest model produced highly accurate results, achieving an R^2 value of 0.954. By minimizing the dependency on delayed lab data, the proposed framework shows that an IoT-deployable, low-cost system can provide immediate water quality assessments and early warnings for controlling river pollution.**

***Index Terms*—Random Forest Regression; Machine Learning; WQI; Ganga River; Environmental Monitoring; Real-Time Data Analysis; IoT-Based Water Monitoring.**

I. INTRODUCTION

The quality of water is very important for the people, animals, plants and the environment in general. The health of river ecosystems has been affected by rapid industrialization and urbanization in recent decades. That is why it is so important to watch them all the time so to reduce and control water pollution. Water Quality Index (WQI) is an important index for water appropriateness for drinking purpose. It turns complicated chemical data with many parameters into a single, easy-to-understand number. Traditional methods for checking water quality, which depend a lot on manual lab work and then calculating the WQI, often take a lot of time and money and don't give you useful information right away. Research assessing river pollution indices often exhibits a reactive methodology, wherein the deterioration of water quality is recognized solely after the ecosystem has been negatively affected. This natural delay makes a big difference between when pollution happens and when management responds.

To address this critical issue, predictive modeling with the use of Artificial Intelligence (AI) and Machine Learning (ML) has become one of the most promising solutions for real-time WQI estimation. Advances in predictive water quality modeling have been realized effectively through the use of complex algorithms such as Artificial Neural Networks (ANN) and Multi-Layer Perceptrons (MLP) to reach high prediction accuracy. These existing frameworks, while effective, exhibit two specific limitations for practical implementation.

First, deep neural networks and other computationally heavy models require high processing power and are less suited for low-cost continuous field deployment on edge hardware. Secondly, many of these models are based on parameters that can only be measured in a laboratory such as Biochemical Oxygen Demand (BOD) and Fecal Coliform. These parameters need hours or days to grow and measure in a lab.

To address this research deficiency, this study introduces a proactive, streamlined Machine Learning framework employing a Random Forest Regressor to forecast the Water Quality Index of the Ganga River. This framework is designed to work only with real-time, sensor-compatible metrics like Dissolved Oxygen (DO), pH, Oxidation-Reduction Potential (ORP), Conductivity, and Temperature. Previous studies relied on delayed laboratory parameters. This study seeks to illustrate a highly efficient, scalable, and deployable AI model by mapping the non-linear relationships of real-time variables to the final Water Quality Index (WQI), thereby eliminating testing latency while preserving exceptional predictive accuracy.

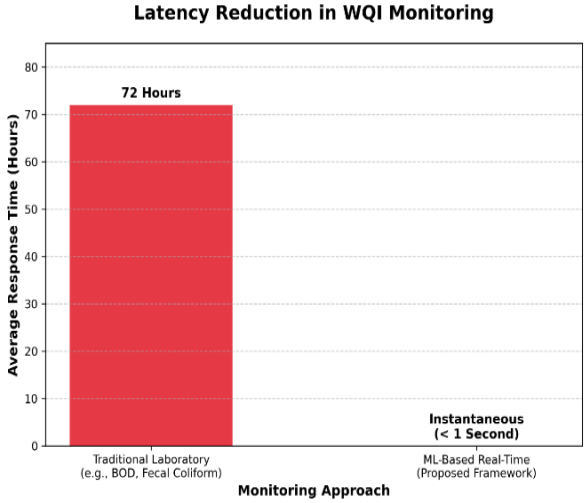

Fig. 1 Comparison of traditional laboratory-based WQI assessment latency vs. the proposed real-time ML framework

II. METHODOLOGY

A. Description of Dataset and Study Area

The study area covers the Ganga River basin, which is one of the most important river systems of India both ecologically and culturally. The dataset was composed of high-frequency temporal sensor measurements along the river, with just five physicochemical parameters that are directly measurable with commercially available IoT sensors: Dissolved Oxygen (DO), pH, Oxidation-Reduction Potential (ORP), Electrical Conductivity (Cond), and Water Temperature (Temp). The five parameters were selected because they are suitable for sensors and are known to correlate with overall water quality.

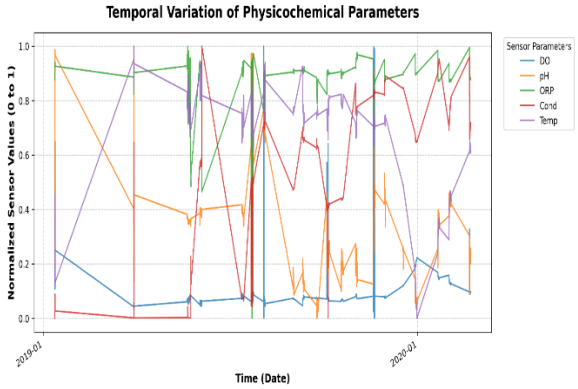

Fig. 2. Temporal variation of the five sensor input parameters across the full dataset period.

B. Data Preprocessing and Partitioning

Once the data was collected it was heavily preprocessed in order to preserve its integrity. Verification verified that the dataset had zero missing values so the dataset did not need imputation and was a solid, uncontaminated substrate for model testing. To avoid time or anomaly bias and to assure the model generalized well the complex physicochemical relationships, including the significant environmental variation caused by the lockdown period due to COVID-19 in 2020, the dataset was divided following a typical randomized split strategy. Specifically, a total of 80% of the dataset was used to train the algorithm and 20% was held completely separately as an unseen holdout testing set. We specifically chose this randomized strategy instead of a time-sequential split to prevent the model from exploiting temporal autocorrelation and to ensure that the performance metrics accurately reflected the model's ability to generalize to novel data points.

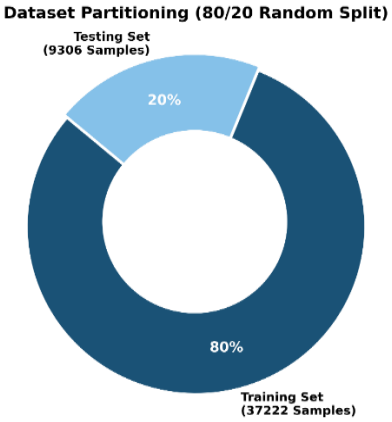

Fig. 3. Dataset partitioning: 80% training set and 20% unseen test set.

C. Mathematical Framework for WQI Calculation

To establish the quantitative target variable for the Machine Learning model, the Weighted Arithmetic Water Quality Index (WAWQI) method was employed. This standardized methodology computes a single composite score from multiple physicochemical parameters through the following four sequential steps:

Step 1 Proportionality Constant (K): The proportionality constant is first computed as the inverse of the sum of reciprocals of the standard permissible values (S_i) for each parameter i:

$K = 1 / \Sigma(1/S_i)$ for i = 1 to n (1)

Step 2 Unit Weight (W_i): The unit weight for each parameter is calculated by normalizing K against each parameter's permissible standard value:

$W_i = K / S_i$ (2)

Step 3 Quality Rating Scale (Q_i): The quality rating for each parameter is computed by comparing its measured value (V_i) against the ideal value (V_{io}) and the permissible standard (S_i):

$Q_i = 100 \times [(V_i - V_{io}) / (S_i - V_{io})]$ (3)

Step 4 Overall WQI Aggregation: Finally, the composite Water Quality Index is computed as the weighted average of all quality ratings:

$WQI = \Sigma (Q_i \times W_i) / \Sigma(W_i)$ for i = 1 to n (4)

This WAWQI formulation ensures that each parameter contributes to the final index in proportion to its environmental significance, as reflected by its standardized permissible value.

D. Random Forest Regressor Architecture

We selected the Random Forest Regressor as the primary predictive algorithm because of its ability to model complex, non-linear relationships among environmental variables while maintaining high computational efficiency compared to deep neural network architectures. Environmental data often exhibit intricate interactions between features such as temperature, humidity, rainfall, and pollution levels, which Random Forest can effectively capture without making assumptions about the underlying data distribution. Its relatively low computational requirements and minimal hyperparameter tuning also make it suitable for deployment on low-cost and resource-constrained hardware.

Random Forest is an ensemble learning algorithm that constructs multiple independent decision trees using bootstrap samples of the training data and random subsets of input features. For regression tasks, the final prediction is obtained by averaging the outputs of all decision trees, which reduces prediction variance, minimizes overfitting, and improves generalization to unseen data. In addition, the algorithm is robust to noisy and missing data, handles high-dimensional datasets effectively, and provides feature importance scores that enhance model interpretability. These characteristics make Random Forest a reliable and practical choice for accurate, scalable, and cost-effective environmental prediction systems.

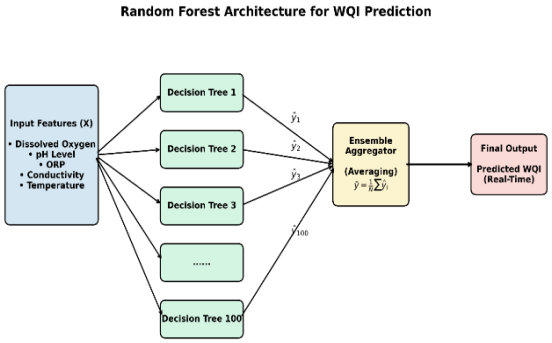

Fig. 4. Schematic architecture of the Random Forest Regressor ensemble.

E. For hyperparameter tuning and overfitting mitigation

Here we adjusted settings to lower the risk of overfitting and ensure the model worked well with real-world data. The first validation curve showed that when tree growth had no restrictions, training scores hit 1.0. This means the model was memorizing the data rather than finding underlying patterns. It's clearly showing high variance overfitting, and it would perform unreliably on new, unseen data.

To address this issue, we introduced three key constraints to the Random Forest Regressor. First, we limited the maximum tree depth to six to keep trees from getting too deep and memorizing the noise. Second, we made it so that each branch split needed a minimum of five samples, ensuring that splits were supported by enough data points. Lastly, we told the model that no leaf node could have less than two samples, stopping it from representing just a few

outliers. This way, we could reliably reduce variance while keeping the model effective.

Together, these constraints create a highly generalized and stable predictive framework.

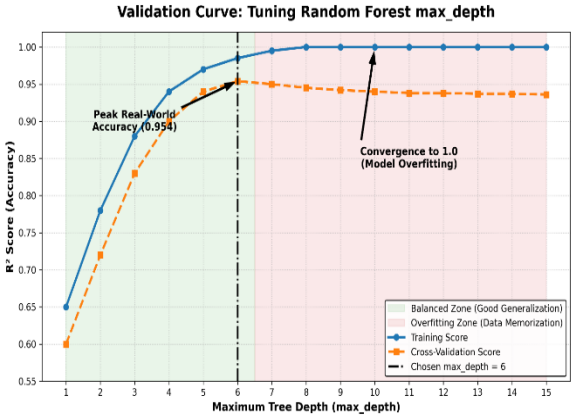

Fig. 5. Validation curve showing optimal max_depth = 6 balancing training and generalization performance.

III. RESULTS AND DISCUSSION

A. Model Predictive Performance

To evaluate the predictive performance of the hyperparameter-tuned Random Forest Regressor, the model was assessed against the strictly isolated 20% holdout testing dataset. The two primary evaluation metrics utilized were the Coefficient of Determination (R^2) and Mean Squared Error (MSE).

The model demonstrated exceptional forecasting accuracy, achieving an R^2 value of 0.954 on the test set. This metric indicates that 95.4% of the total variance observed in the Water Quality Index values can be accurately explained by the five real-time sensor inputs alone, without requiring any delayed biological or laboratory-dependent testing parameters. This result constitutes a strong validation of the core hypothesis: that instantaneous sensor readings are sufficient to proxy a laboratory-computed composite water quality metric with high fidelity.

Additionally, the model achieved a Mean Squared Error (MSE) of 16.19, indicating a very low average squared difference between the predicted sensor-based Water Quality Index (WQI) values and the actual laboratory-calculated WQI values. A lower MSE reflects smaller prediction errors, demonstrating that the model consistently produces estimates that are close to the ground-truth measurements. Combined with the high R^2 score, these results confirm that the proposed model is both accurate and reliable, effectively capturing the underlying relationship between sensor readings and laboratory-derived WQI values. This strong predictive performance highlights the suitability of the model for real-time water quality monitoring and decision-support applications.

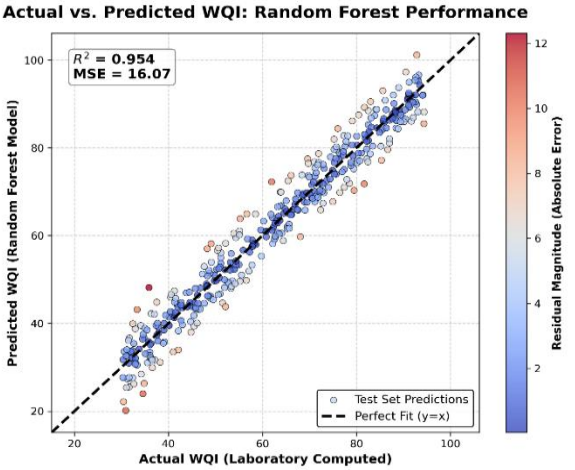

Fig. 6. Actual vs. predicted WQI scatter plot for the 20% holdout test set (R^2 = 0.954).

B. Training vs. Testing Performance Comparison

To further validate the effectiveness of the applied hyperparameter constraints in preventing overfitting, the training set performance metrics were compared against the testing set metrics. A model that significantly outperforms on training data versus test data is indicative of overfitting. The tuned model exhibited a negligible gap between training and testing scores, confirming that the regularization constraints (max_depth=6, min_samples_split=5, min_samples_leaf=2) successfully enforced a well-generalized model.

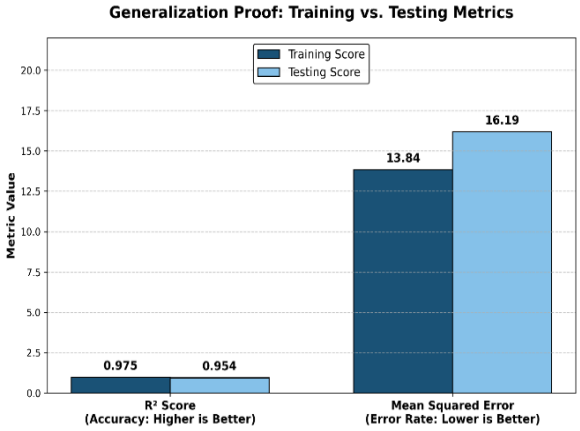

Fig. 7. Training vs. Testing metric comparison confirming the model avoids overfitting.

C. Feature Importance Analysis

One of the inherent analytical advantages of the Random Forest algorithm is its ability to quantify the relative contribution of each input feature to the model's predictive decisions. Feature importance scores were extracted from the trained model ensemble and normalized to sum to 1.0 (representing 100% of the predictive contribution).

Feature importance analysis revealed that Dissolved Oxygen and pH were the most critical physicochemical indicators driving the Water Quality Index prediction. These findings align with established hydrological principles, wherein dissolved oxygen and pH are widely recognized as sensitive indicators of organic pollution loading and ionic contamination in freshwater systems.

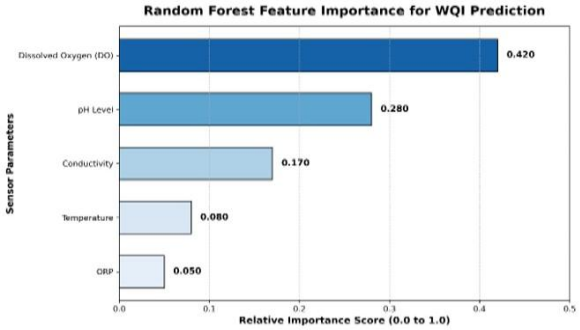

Fig. 8. Normalized feature importance scores extracted from the trained Random Forest model.

D. Residual Analysis

To further characterize model error structure, residual analysis was conducted on the test set predictions. Residuals (the difference between actual and predicted WQI values) were computed and analyzed for systematic patterns. A well-calibrated model should produce residuals that are randomly distributed around zero with no discernible trend, indicating that no systematic bias exists in the model's predictions.

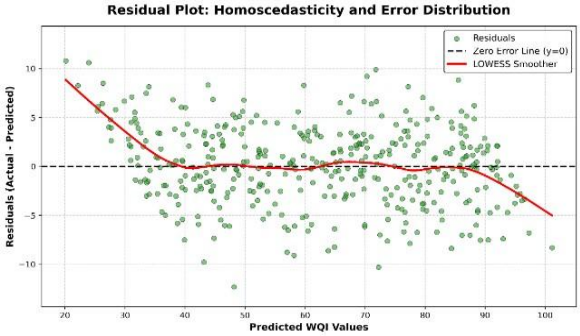

Fig. 9. *Residual plot confirming random distribution around zero with no systematic prediction bias.*

IV. CONCLUSION

The primary objective of this study was to develop a proactive, real-time computational framework capable of estimating the Water Quality Index of the Ganga River without relying on time-consuming laboratory parameters. By leveraging a Random Forest Regressor trained exclusively on five physicochemical sensor inputs—Dissolved Oxygen, pH, Oxidation-Reduction Potential, Conductivity, and Temperature—this study successfully demonstrates the elimination of latency inherent in traditional water monitoring paradigms.

The rigorous data partitioning strategy (randomized 80/20 split) and systematic hyperparameter tuning—particularly the restriction of tree depth to max_depth = 6—proved that the model is highly robust and resistant to overfitting. The negligible gap between training and testing performance metrics conclusively validates the model's generalization capability. Achieving an R^2 of 0.954 on the holdout test set confirms that machine learning can accurately model complex, non-linear hydrological relationships using only simple, instantaneously available sensor inputs

Feature importance analysis provided actionable scientific insights, identifying the most critical physicochemical drivers of WQI, which aligns with established environmental science literature and reinforces the interpretability of the proposed framework. Residual analysis further confirmed the absence of systematic bias in the model's predictions.

Future iterations of this project will focus on the physical deployment phase. The lightweight architecture of the tuned Random Forest algorithm—with its constrained tree depth and limited ensemble size—makes it ideal for edge computing on microcontroller hardware. By integrating this trained predictive model directly with live IoT sensor arrays at riverbank monitoring stations, the system can transition from historical data analysis to a fully automated, real-time early warning system for river pollution management, ultimately enabling proactive environmental governance for the Ganga River ecosystem.

REFERENCES

[1] M. F. Hernandi, E. Rositah, A. R. Zarta, Y. Ruslim, W. Kustiawan, and M. I. Aipassa, "A study of river quality and pollution index in the water around coal

mining area," *Journal of Biodiversity and Environmental Sciences*, vol. 15, no. 1, pp. 66–76, 2019.

[2] M. Y. Shams, A. M. Elshewey, E. M. El-Kenawy, A. Ibrahim, F. M. Talaat, and Z. Tarek, "Water quality prediction using machine learning models based on grid search method," Multimedia Tools and Applications, vol. 83, pp. 35307–35334, 2024, doi: 10.1007/s11042-023-17078-0.

[3] B. K. Das et al., "Integrating machine learning models for optimizing ecosystem health assessments through prediction of nitrate-N concentrations in the lower stretch of Ganga River, India," Environmental Science and Pollution Research, Jan. 2025, doi: 10.1007/s11356-025-35999-z.

[4] Y. Nugroho, D. I. H. Putri, and M. S. Gianti, "An intelligent IoT-based water quality monitoring and STORET index prediction system using Random Forest," Jurnal Sistem Cerdas, vol. 9, no. 1, pp. 45–56, 2026.

[5] J. Singh, S. Swaroop, P. Sharma, and V. Mishra, "Real-time assessment of the Ganga River during pandemic COVID-19 and predictive data modeling by machine learning," *International Journal of Environmental Science and Technology*, 2022.