International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

IoT-Based Indian River Water Quality Prediction Model

Abha Patak[1], Omkar Hulawale[2], Ajinkya Lahane[3], Tushar Aswar[4], Abhishek Ganore[5]

[1]Professor, Department of Computer Engineering Dr. D. Y. Patil College of Engineering and Innovation Varale, Pune, India
[2,3,4,5]Student, Department of Computer Engineering Dr. D. Y. Patil College of Engineering and Innovation Varale, Pune, India

Abstract
This study aims to create an IoT based model that can monitor and predict the quality of river water in India. It will use real time data, machine learning and deep learning technique. The system uses sensors for pH, temperature, dissolved oxygen, and turbidity, connected to a microcontroller for data collection. The data collected are sent to a cloud database for storage and processing. A machine learning model, based on the best fit algorithm, will be trained to predict the Water Quality Index (WQI) and determine whether the water is safe or polluted. This model will offer an effective, low-cost, and scalable way of monitoring the environment. The proposed system will accurately predict water quality, allowing for continued, low-cost monitoring and early detection of warnings. Integration of IoT with machine learning provides a reliable scalable solution for water monitoring. This combination will help agencies monitor and respond quickly to pollution warnings and supports a sustainable water quality prediction model.

Keywords: Internet of Things, Water Quality, Prediction Model, Machine Learning, Sensors, River Pollution.

1. INTRODUCTION

Water plays a fundamental role in sustaining life, supporting agriculture, and maintaining ecological balance. In India, river systems such as the Ganga, Yamuna, Narmada, Tapi, and Godavari are crucial sources for domestic, industrial, and agricultural use. However, increasing urbanization, industrial discharge, and agricultural runoff have caused significant degradation of river water quality, posing serious threats to public health and aquatic ecosystems. Regular monitoring and prediction of water quality have therefore become essential for effective environmental management and sustainable water use. Conventional water quality assessment methods rely on manual sampling and laboratory-based testing, which, although accurate, are time-consuming and incapable of providing real-time data [1]. These limitations hinder timely detection of pollution events and reduce the ability to make proactive management decisions. To overcome these challenges, researchers have increasingly turned toward data-driven and automated prediction approaches using machine learning (ML) and deep learning (DL) models. Recent studies have shown that machine learning techniques such as Support Vector Machines (SVM), Random Forests, and ensemble methods can effectively predict Water Quality Index (WQI) values with

higher accuracy compared to traditional statistical models [1], [4], [7], [8]. These methods can process large and complex datasets, enabling better classification of water quality levels and detection of key influencing parameters. Deep learning models, particularly Long Short-Term Memory (LSTM) networks and hybrid neural architectures, have also demonstrated superior performance in modeling temporal dependencies and non-linear variations in water quality data [2]–[4]. Such models are capable of learning seasonal and spatial patterns, making them more reliable for forecasting long-term water trends. Parallel developments in the Internet of Things (IoT) have revolutionized environmental monitoring by enabling real-time data collection from multiple water quality sensors. IoT-based systems now allow continuous measurement of key parameters such as pH, turbidity, dissolved oxygen (DO), and total dissolved solids (TDS), which can be transmitted wirelessly for analysis [10]. Cloud-integrated IoT frameworks have further improved data storage, visualization, and remote accessibility, providing a scalable infrastructure for water monitoring and analytics [9]. However, most existing IoT implementations focus primarily on monitoring rather than integrating predictive intelligence, limiting their ability to forecast water quality dynamics and support preventive measures. Several studies have explored hybrid or optimized models to enhance prediction accuracy. For example, grid search optimization techniques have been applied to improve ML model performance [1], while hybrid ML and temporal models have demonstrated potential for more accurate forecasting [3]. Nonetheless, many of these approaches remain region-specific and lack generalization across diverse water systems such as those found in India, where hydrological and anthropogenic factors vary widely [6], [8], [9]. Given these limitations, the integration of IoT-based realtime data acquisition with advanced ML and DL models offers a promising direction for future research. The proposed IoTbased Indian River Water Quality Prediction System aims to address existing challenges by combining sensor-based data collection, cloud storage, and hybrid predictive analytics using models like LSTM and XGBoost. This integrated framework is expected to deliver accurate, continuous, and scalable water quality forecasting that can support governmental and environmental agencies in achieving sustainable water resource management.

2. OBJECTIVES

- To design and implement an IoT-based system for real-time monitoring of river water quality.
- To apply hybrid Machine Learning and Deep Learning models (XG Boost and LSTM) for accurate prediction of the Water Quality Index (WQI) and key physicochemical parameters.
- To improve environmental decision-making through real-time data visualization and predictive analytics on cloud platforms. Figures and Tables

3. LITERATURE REVIEW:

Water quality prediction has become a vital research area in environmental monitoring, especially with the growing availability of large-scale data and the integration of Artificial Intelligence (AI) and Internet of Things (IoT) technologies. Researchers have increasingly focused on developing intelligent systems that can both monitor and forecast water quality parameters such as pH, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), and Total Dissolved Solids (TDS).

The following section reviews key contributions in this field and highlights their methodologies, advantages, and limitations.

- Mahmoud Y. Shams, Ahmed M. Elshewey, El-Sayed M. El-Kenawy, Abdelhameed Ibrahim, Fatma M. Talaat, and Zahraa Tarek [1] introduced a machine learning-based water quality prediction model

optimized using a grid search approach. Their work demonstrated that hyperparameter tuning significantly enhances the performance of algorithms such as Random Forest and Gradient Boosting when applied to environmental datasets. The study established a baseline for data-driven prediction systems, emphasizing the importance of balancing model complexity with interpretability.

- K. Chen et al. [2] performed a comparative study of multiple ML algorithms on large-scale surface water datasets. They identified that ensemble models—especially Random Forest and Gradient Boosting—outperformed traditional regressionbased methods in both accuracy and feature importance analysis. The study also emphasized identifying the most influential water parameters to improve model explainability and environmental decision-making.
- Fernandez del Castillo, M. Verduzco Garibay, D. D´ıazVazquez, C. Yebra-Montes, L. E. Brown, A. Johnson, A. ´ Garcia-Gonzalez, and M. S. Gradilla-Hernandez [3] proposed ´ a hybrid machine learning framework combined with temporal analysis to improve river water prediction. Their model successfully captured time-dependent patterns, illustrating how temporal context enhances the precision of long-term water quality forecasts.
- K. Kalaivanan and J. Vellingiri [4] examined various machine learning models for water quality prediction and highlighted the crucial role of preprocessing, feature scaling, and normalization. Their research revealed that proper handling of missing data and feature selection directly impacts the predictive accuracy and reliability of ML models used in aquatic environments.
- Rahmi Fadhilah, Heri Kuswanto, and Dedy Dwi Prastyo [5] addressed class imbalance issues in water quality datasets by implementing resampling techniques in classification models. Their results indicated that combining ensemble learning with resampling improved prediction robustness and minimized bias in skewed datasets—a frequent challenge in environmental monitoring.
- Sau, Chakraborty, and Majumder [6] applied Polynomial Neural Networks (PNN) to estimate the Surface Water Quality Index (WQI) for peri-urban rivers. Their study demonstrated the capacity of neural networks to model non-linear interactions among pollutants, offering a flexible approach for modeling complex environmental systems.
- P. M. Pujar and H. H. Kenchanavvar [7] implemented a Support Vector Machine (SVM) model for river water quality analysis and prediction. Their research concluded that SVMs provide robust classification even with smaller datasets, making them suitable for localized water quality assessment where sensor data are limited.
- M. H. Nishat, M. H. R. B. Khan, T. Ahmed, S. N. Hossain, A. Ahsan, M. M. El-Sergany, Md. Shafiquzzaman, M. A. Imteaz, and M. T. Alresheedi [8] conducted a comparative analysis of machine learning models for Dhaka's river systems. They demonstrated that ensemble algorithms like Random Forest and XGBoost achieved the highest accuracy in predicting the Water Quality Index (WQI), proving their suitability for dynamic river ecosystems.
- K. P. Rasheed Abdul Haq and V. P. Harigovindan [9] proposed a hybrid deep learning framework integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models for smart aquaculture. Their work validated the effectiveness of hybrid deep models in processing real-time multivariate sensor data, offering insights applicable to river water quality prediction.
- Finally, H. M. Forhad, Md. Ripaj Uddin, R. S. Chakrovorty, A. M. Ruhul, H. M. Faruk, Sarker Kamruzzaman, Nahid Sharmin, AHM Shofiul Islam Molla Jamal, Md. Mezba-Ul Haque, and AKM M. Morshed [10] developed an IoT-based real-time water quality monitoring system for water treatment plants. Their framework collected data from multiple sensors (pH, turbidity, DO, and

temperature) and transmitted it to a cloud platform for visualization, proving the feasibility of integrating IoT with environmental analytics.

From the reviewed literature, it is evident that substantial progress has been achieved in ML, DL, and IoT-based water quality systems. However, most existing studies focus on either prediction without real-time IoT integration or real time monitoring without predictive analytics. Furthermore, there is limited research specifically tailored to Indian river ecosystems, which are influenced by unique climatic and anthropogenic factors. This gap underscores the need for a hybrid IoT–Cloud–ML framework that integrates both temporal prediction (LSTM) and high-performance learning (XGBoost) to enable accurate, real-time, and scalable river water quality forecasting for Indian conditions.

4. RESEARCH GAP

A review of existing literature shows substantial progress in the areas of statistical modeling, machine learning, deep learning, and IoT-assisted water quality monitoring. However, several critical gaps still remain unaddressed.

- **Limitations of Conventional Regression-Based Models:** Early studies applying Multiple Linear Regression (MLR) and similar statistical techniques [1], [5] performed adequately on small datasets but failed to model the non-linear, multivariate, and seasonally dynamic behavior of river water systems. These models also demonstrate high sensitivity to missing or noisy data, making them unsuitable for large-scale or real-time deployment.
- **Inadequate Temporal Modeling in Classical ML Approaches:** Machine learning algorithms such as Random Forest, SVM, and other ensemble-based WQI predictors [5], [7], [8] improved accuracy over regression methods, but they primarily function as static classifiers or regressors. Their inability to learn temporal dependencies limits their effectiveness for long-term water quality forecasting.
- **Partial Adoption of Deep Learning without Generalization:** Deep learning models such as LSTM and hybrid neural architectures have been successfully used for sequential water quality prediction [2], [4]; however, most implementations are region-specific, parameter restricted, or dataset-dependent. As a result, they lack scalability and do not generalize well to diverse ecosystems such as Indian rivers, where seasonal and anthropogenic variations are more complex.
- **IoT Systems Lacking Predictive Intelligence:** IoT enabled monitoring frameworks [10] have enabled continuous sensing and wireless transmission of key parameters such as pH, DO, TDS, and temperature. Yet, the majority of these systems provide only real-time alerts and dashboards without incorporating ML/DL based forecasting models capable of predicting future water quality conditions.
- **Absence of Unified IoT–Cloud–Hybrid ML Architecture:** Although individual ML, DL, and IoT-based approaches have been explored, very few studies combine real-time sensing, cloud analytics, and hybrid prediction models (e.g., LSTM + XG Boost) into a single end-to end framework. Existing hybrid models [3], [4], [7] are mostly implemented offline and are not connected to scalable IoT-based data pipelines.
- **Limited Research on Indian River Ecosystems:** Most published work either evaluates foreign datasets or small-scale experimental water bodies. There is a notable shortage of predictive and IoT-integrated research specifically designed for Indian river basins, which are affected by unique climatic, industrial, and agricultural pollution dynamics [6], [8], [9].

Gap Conclusion: Current research provides either real time monitoring without prediction or prediction without real time IoT integration. No existing work offers a scalable, IoT enabled, cloud-connected, hybrid

LSTM XG Boost forecasting system tailored for Indian river water quality

5. DATASET DESCRIPTION

The dataset used in this project is from Kaggle named "Indian Water Quality Data (2021-2023)" sourced from the Central Pollution Control Board (CPCB) of India, the national organization responsible for the prevention and control of water and air pollution.

This dataset contains water quality data collected from various monitoring stations across 17 different states in India between the year 2021 and 2023. It covers different types of water bodies, primarily Rivers and Drains.

The dataset includes parameters crucial for assessing water quality:

- **Physical Parameters**: Temperature
- **Chemical Parameters:** pH, Conductivity, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Nitrate-N
- **Biological Parameters:** Fecal Coliform, Total Coliform Each row represents a specific monitoring location and provides the minimum and maximum recorded values for these parameters over a given year.

Columns Description

- **STN code:** Station Code, a unique identifier for the monitoring location • Monitoring Location: Descriptive name for the location where the sample was taken • Year: The year the data was recorded (2021, 2022, 2023)
- **Type Water Body:** Type of water body such as 'River' or 'Drain'
- **State Name:** Indian state where the monitoring station is located
- **Temperature (C) – Min/Max:** Minimum and Maximum temperature recorded in degrees Celsius
- **Dissolved – Min/Max:** Minimum and Maximum pH values
- **Conductivity (µmho/cm) – Min/Max:** Minimum and Maximum electrical conductivity
- **BOD (mg/L) – Min/Max:** Minimum and Maximum Biochemical Oxygen Demand
- **Fecal Coliform (MPN/100ml) – Min/Max:** Minimum and Maximum levels of Fecal Coliform bacteria
- **Total Coliform (MPN/100ml) – Min/Max:** Minimum and Maximum levels of Total Coliform bacteria
- **Fecal – Min/Max:** Additional fecal bacteria measurements when available

6. WORKING

As shown in Figure 1, the proposed system integrates IoT, Cloud Computing, and Machine Learning to predict water quality efficiently.

The process begins with **Data Collection**. IoT sensors placed at different water sources automatically collect data, which is continuously sent to the system. If IoT devices are unavailable, data can be manually entered through a web portal, providing flexibility.

The collected data is sent to the **Cloud** for storage and easy access. Cloud storage allows for handling large volumes of data and accessing it from anywhere. Next, data goes through Data Preprocessing, which ensures clean, consistent, and normalized data ready for analysis. Missing or incorrect values are handled, redundant data is removed, and all readings are standardized.

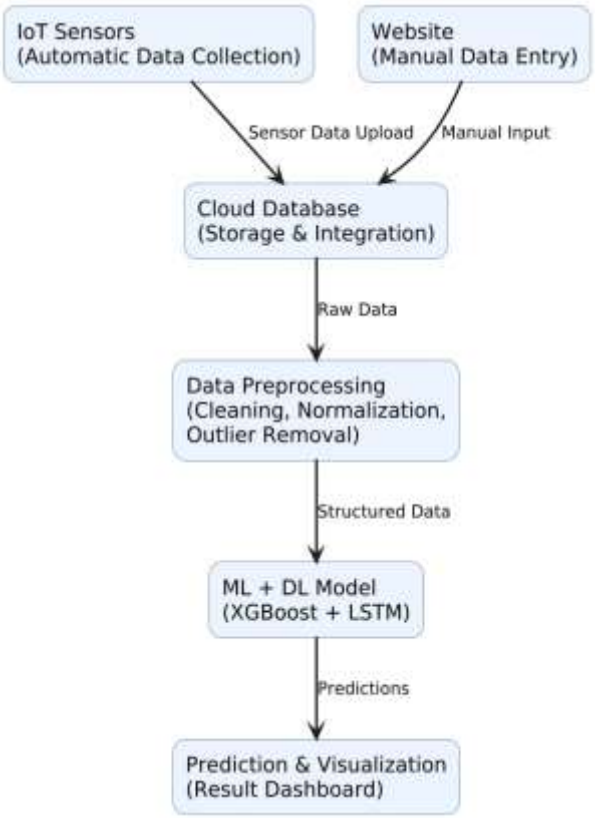

Figure 1. System Workflow

The preprocessed data is then passed to the Machine Learning module, which consists of two main models:

- **XGBoost (eXtreme Gradient Boosting):** Fast and effective for structured data, identifies relationships between water quality features, improving accuracy.
- **LSTM (Long Short-Term Memory):** A deep learning model capable of understanding time-based patterns, analyzing historical data to predict future outcomes.

Together, these models form a hybrid learning system that enhances prediction performance. Finally, the results are displayed on a dashboard or visual interface, enabling users to interpret predictions without advanced technical knowledge.

7. CONCLUSION

This project presents an intelligent, efficient, and scalable framework for river water quality prediction by integrating Machine Learning (ML), Deep Learning (DL), Cloud Computing, and Internet of Things (IoT) technologies. The proposed system effectively combines real-time data acquisition from IoT sensors with advanced analytical models such as XGBoost and LSTM, enabling accurate forecasting of critical water quality parameters.

Unlike conventional manual testing or static ML systems, this hybrid IoT–Cloud–ML approach ensures continuous monitoring, dynamic prediction, and remote accessibility. The integration of cloud storage facilitates secure and scalable data handling, while the ML models improve decision-making accuracy by learning temporal and non-linear relationships among environmental variables. As a result, the system not only identifies current water quality but also anticipates future degradation trends, which is vital for pollution prevention and sustainable water management.

The implementation of this model is particularly relevant to Indian river ecosystems, where rapid industrialization and population growth have intensified pollution challenges. By leveraging real-time IoT data and predictive intelligence, this system can assist government agencies, environmental boards, and researchers in developing proactive strategies for river conservation and pollution control.

Overall, this project demonstrates that the fusion of IoT and ML technologies offers a cost-effective, automated, and reliable solution for water quality forecasting. In the future, this framework can be expanded by integrating more advanced deep learning architectures, satellite-based sensing data, and geospatial analytics to enhance prediction precision and adaptability across diverse environmental contexts.

8. REFERANCE

1. M. Y. Shams, A. M. Elshewey, E.-S. 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- 16737-4.
2. K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, and H. Ren, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," Water Research, vol. 171, p. 115454, 2020, doi: 10.1016/j.watres.2019.115454.
3. A. Fernandez del Castillo, M. Verduzco Garibay, D. D ´ ´ıaz-Vazquez, C. ´ Yebra-Montes, L. E. Brown, A. Johnson, A. Garcia-Gonzalez, and M. S. Gradilla-Hernandez, "Improving river water quality prediction with ´ hybrid machine learning and temporal analysis," Ecological Informatics, vol. 82, p. 102655, 2024, doi: 10.1016/j.ecoinf.2024.102655.
4. K. Kalaivanan and J. Vellingiri, "Survival study on different water quality prediction methods using machine learning," Nature Environment and Pollution Technology, vol. 21, no. 3, pp. 1259–1267, 2022, doi: 10.46488/NEPT.2022.v21i03.82.
5. R. Fadhilah, H. Kuswanto, and D. D. Prastyo, "Performance evaluation of classification methods utilizing resampling techniques for water quality prediction on imbalanced data," Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 26091–26099, 2025, doi: 10.48084/etasr.11832.
6. A. Sau, T. Chakraborty, and M. Majumder, "Prediction of surface water quality index of a peri-urban river by polynomial neural networks," Grenze International Journal of Engineering and Technology, vol. 10, no. 2, p. 466, 2024.
7. P. M. Pujar and H. H. Kenchanavvar, "Water quality analysis and prediction of river water using support vector machine model," Grenze International Journal of Engineering and Technology, vol. 7, no. 1, p. 121, 2021.
8. M. H. Nishat, M. H. R. B. Khan, T. Ahmed, S. N. Hossain, A. Ahsan, M. M. El-Sergany, M. Shafiquzzaman, M. A. Imteaz, and M. T. Alresheedi, "Comparative analysis of machine learning models for predicting water quality index in Dhaka's rivers of Bangladesh," Environmental Sciences Europe, vol. 37, no. 31, 2025, doi: 10.1186/s12302-025-01078-w.
9. K. P. Rasheed Abdul Haq and V. P. Harigovindan, "Water quality prediction for smart aquaculture using hybrid deep learning models," IEEE Access, vol. 10, pp. 60078–60098, 2022, doi: 10.1109/ACCESS.2022.3180482.

10. H. M. Forhad, M. R. Uddin, R. S. Chakrovorty, A. M. Ruhul, H. M. Faruk, S. Kamruzzaman, N. Sharmin, A. H. M. S. I. M. Jamal, M. M. Haque, and A. K. M. M. Morshed, "IoT based real-time water quality monitoring system in water treatment plants (WTPs)," Heliyon, vol. 10, p. e40746, 2024, doi: 10.1016/j.heliyon.2024.e40746.