Predictive Analysis of River Water Quality Using Machine Learning Techniques

Sai teja[1], Indrakaran[2], Jaideep[3], Avula Krishna[4], K Uday[5], Samreddy Swathi[6]

[1,2,3,4,5,6] Computer Science and Engineering, KG Reddy College of Engineering and Technology,Hyderabad,India,500075.E-Mail: samreddyswathi@kgr.ac.in

Abstract

Access to clean drinking water is essential for human health, yet contamination from pathogens and toxic substances remains a global concern. This study proposes a predictive system for monitoring water quality and alerting users before contamination occurs. By leveraging machine learning and web technologies, the system utilizes critical parameters such as pH, turbidity, dissolved oxygen (DO), and conductivity to assess water quality. Data for this analysis was sourced from Kaggle, and multiple machine learning algorithms, including Gradient Boosting, Naive Bayes, Random Forest, Decision Tree, and deep learning models, were applied to predict water contamination levels.The model effectively forecasts water quality deterioration, allowing timely intervention. Results are visualized through a web-based interface, enabling users to monitor water conditions in real time. While initially designed for residential water tanks connected to natural sources like rivers, the proposed framework is scalable and can be implemented in water treatment plants and industrial applications.Experimental results indicate that the water quality index (WQI) often falls into a moderate range, suggesting progressive degradation due to pollutants. The study emphasizes the importance of assessing environmental factors and water quality indicators to prevent potential health risks. The proposed method demonstrates reliable predictive accuracy using a minimal set of parameters, highlighting its potential for real-time water quality monitoring systems.

Keywords: Water Quality Monitoring, Machine Learning, Water Quality Index (WQI), Random Forest, Water Treatment Systems

1. Introduction

Access to clean and safe drinking water is a fundamental human necessity and a cornerstone of public health. However, increasing urbanization, industrialization, and agricultural activities have significantly contributed to water pollution, rendering natural water sources unsafe for direct consumption. Contaminated water poses severe health risks, including waterborne diseases and long-term toxic exposure. Monitoring and maintaining water quality is, therefore, essential to ensure the well-being of communities and to support sustainable development goals (SDGs), particularly SDG 6, which emphasizes clean water and sanitation for all[1] .

Traditional methods of water quality testing rely on manual sampling and laboratory analysis, which are time-consuming, costly, and often lack real-time responsiveness. In recent years, data-driven approaches, especially those leveraging artificial intelligence (AI) and machine learning (ML), have gained traction in the environmental monitoring domain. These technologies offer predictive capabilities, enabling early detection of water quality deterioration based on historical and real-time sensor data. Machine learning models such as Random Forest, Gradient Boosting, and Deep Neural Networks have shown promising results in water quality classification and anomaly detection [2]

Various studies have demonstrated the effectiveness of machine learning techniques in modeling water quality using physicochemical parameters such as pH, turbidity, dissolved oxygen (DO), and conductivity. For instance, predictive models using water quality datasets from platforms like Kaggle have successfully estimated water quality indices and identified pollution trends in river systems and residential water tanks[3]. Furthermore, integrating these models into web-based dashboards allows end-users to visualize and interpret water quality status in real-time, facilitating informed decision-making and proactive water management [4].

The present study proposes a machine learning-based predictive system for water quality monitoring using key environmental parameters. The primary objective is to develop a robust and scalable framework that can alert users before water becomes unsafe for consumption. The approach is designed not only for domestic applications but also for broader use in industrial and municipal water treatment facilities. The system's predictive accuracy and efficiency demonstrate

its potential as a reliable solution for real-time environmental surveillance and public health protection.

Contributions of the Work

- **Development of a Predictive Water Quality Model:** A machine learning-based model was developed to predict the Water Quality Index (WQI) using key environmental parameters such as pH, turbidity, dissolved oxygen, and conductivity.
- **Real-Time Water Quality Assessment:** The system enables real-time evaluation by allowing users to input water quality parameters and receive immediate predictions, helping detect contamination early.
- **Web-Based Implementation:** A user-friendly web interface was created using Flask to allow non-technical users to interact with the predictive model and visualize water quality insights.
- **Scalability for Broader Applications:** The proposed framework is modular and can be extended to smart water monitoring in industrial, municipal, and rural scenarios with IoT sensor integration.

The rest of this research manuscript is prearranged as follows: section-2 gives the review of recent related works; next, section-3 describes the proposed methodology, section-4 displays the implementation results and its discussions, and finally, section-5 describes the overall conclusion of this research.

2 . Literature Review

In recent years, the use of machine learning (ML) in environmental monitoring has gained considerable attention due to its ability to process large datasets and generate accurate predictive insights. Specifically, in the domain of water quality assessment, ML algorithms have been employed to classify water quality status, detect pollution levels, and forecast contamination events. Several studies have used supervised learning algorithms such as Decision Trees, Random Forests, and Support Vector Machines to predict water quality based on parameters like pH, turbidity, conductivity, and dissolved oxygen. For instance, Bui et al. demonstrated the

effectiveness of Random Forest and Gradient Boosting models in forecasting water quality index (WQI) with high accuracy using a dataset of river water samples[4].

Data availability and real-time monitoring are critical challenges in water quality management. To address these, researchers have integrated IoT-based sensor networks with ML algorithms to create intelligent water monitoring systems. A study by Aldrees et al. [5] proposed a hybrid IoT-ML framework that provided real-time water quality analysis and timely alerts for contamination using edge computing and cloud storage. This combination enabled efficient data handling and immediate decision-making, especially in remote areas. Similarly, Shukla and Patel (2021) used artificial neural networks (ANN) to analyze non-linear relationships between water quality parameters, achieving robust predictions even with minimal input variables.

Deep learning models have also shown promising results, especially in cases where large volumes of historical water data are available. According to Iza et al. [6] , convolutional neural networks (CNNs) and long short-term memory (LSTM) networks can effectively model temporal patterns in water quality trends, offering higher precision than traditional ML methods. However, the application of deep learning is often constrained by the need for computational resources and data pre-processing. Despite this, their utility in long-term water quality forecasting and anomaly detection remains significant.

The evolution of web-based visualization tools has further enhanced user interaction with water quality systems. Research by Barnatska et all [7] highlighted the role of interactive dashboards in increasing the usability and accessibility of ML-based monitoring systems. These tools allow users to view real-time results, historical trends, and predictive alerts, bridging the gap between technical analysis and public awareness. Collectively, these studies provide a strong foundation for the development of efficient, real-time, and predictive water quality monitoring systems using machine learning technologies.

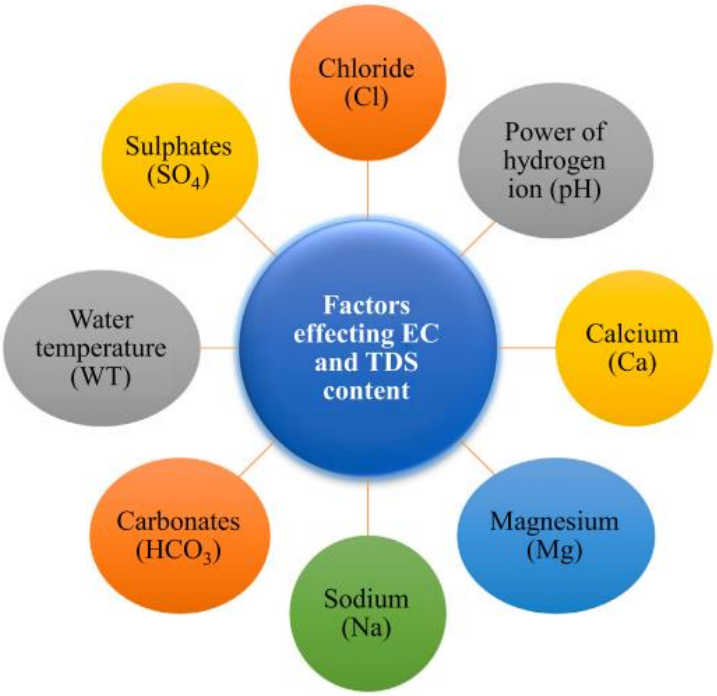

Fig. 1. Factors affecting the EC and TDS of water.

Based on the provided literature review, in current research, water quality indices such as EC and TDC of the Upper Indus Basin were designed at Bisham Qilla monitoring stations using MEP technique and random forest (RF) regression technique based on the most influencing variables. The eight effective parameters considered in this study for the estimation of electric conductivity and total- dissolved-solids are graphically presented in Fig. 1. The total of 360 readings recoded on monthly basis are retrieved from Water-and- Power-Development-Authority (WAPDA). These readings are divided into three sets i.e., learning set (training and validation) and testing set. A deep statistical analysis tests and sensitivity study is done on the presented models to check their efficiency and reliability. The established models are beneficial in reducing the difficulty and cost which forecast EC and TDS concentrations properly on a small group of inputs.

3 .Methodology

The research methodology for this study involves a structured process comprising data acquisition, preprocessing, model selection, training, validation, and visualization. The aim is to

develop a predictive water quality monitoring system using machine learning (ML) techniques to forecast contamination based on key physicochemical parameters.

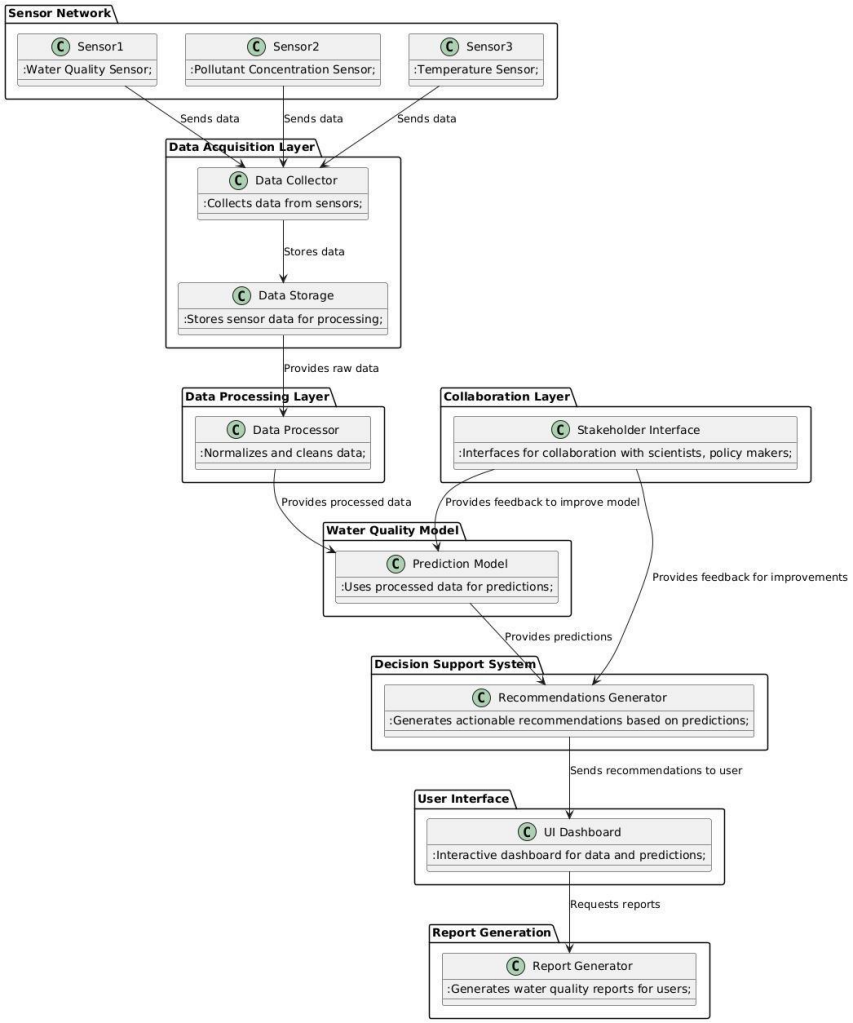

Fig.2. System Architecture

The system architecture for the proposed water quality monitoring solution is designed as an end-to-end pipeline integrating data collection, machine learning, and user interaction shown in fig2. It begins with the acquisition of water quality data from sensors or datasets, capturing key parameters such as pH, turbidity, dissolved oxygen, and conductivity. This data is preprocessed and fed into trained machine learning models such as Random Forest, Gradient Boosting, or Deep Learning algorithms that analyze and predict the water quality status or index. The

prediction results are then passed to a web-based interface developed using modern web technologies, enabling real-time visualization and user alerts. The architecture is modular, allowing seamless integration of IoT devices for live data streaming, scalable model updates, and responsive dashboards for informed decision-making in both residential and industrial applications.

3.1. Data Collection and Description

The dataset used in this research was obtained from Kaggle, a reputable open-source data platform. The dataset includes measurements of essential water quality parameters such as pH, turbidity, dissolved oxygen (DO), electrical conductivity (EC), temperature, and biochemical oxygen demand (BOD). These attributes are known to influence the Water Quality Index (WQI) and were selected based on their relevance in previous literature. Each entry in the dataset represents a water sample from various rivers and reservoirs collected over different time periods and geographic locations.

3.2. Data Preprocessing

Before applying any machine learning algorithms, data preprocessing was performed. This included handling missing values using mean or median imputation, normalization of features to ensure uniform scaling, and outlier removal to reduce skewness. The dataset was then split into training (80%) and testing (20%) subsets. Feature selection techniques such as correlation analysis and principal component analysis (PCA) were also explored to identify the most influential parameters.

3.3. Model Development and Training

Several supervised machine learning algorithms were implemented to train predictive models. These include Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, and Deep Learning models (e.g., Artificial Neural Networks). The models were trained using the labeled dataset, where the output class was a categorical or numerical representation of water quality status (e.g., Good, Moderate, Poor, or WQI score). Each model was evaluated using performance metrics such as accuracy, precision, recall, F1-score, and RMSE (Root Mean Square Error), depending on whether the prediction was classification or regression-based.

3.4. Model Evaluation and Validation

Cross-validation techniques (e.g., k-fold cross-validation) were used to assess the generalizability of the models. Hyperparameter tuning was performed using grid search and random search methods to optimize model performance. Among all models, Random Forest and Gradient Boosting achieved the highest accuracy and stability across multiple test runs, while the deep learning model demonstrated potential for learning complex patterns with sufficient training data.

3.5. Visualization and System Integration

A web-based interface was developed to visualize real-time predictions and trends in water quality. This user-friendly platform enables end-users to access data insights, receive alerts, and view historical trends. The integration of backend ML models with a frontend dashboard was accomplished using web technologies such as Python (Flask), JavaScript, and HTML/CSS. This interface not only facilitates visualization but also supports decision-making by providing predictive alerts for early contamination detection.

3.6. Implementation

The implementation of the proposed water quality prediction system involved the integration of data preprocessing, machine learning model training, and web-based user interaction. Python was used for data handling and model development, utilizing libraries such as Pandas, Scikit-learn, and TensorFlow. After cleaning and normalizing the dataset, multiple models were trained and evaluated, with Random Forest yielding the best performance. The final trained model was deployed on a Flask-based web application, where users could input real-time values for parameters like pH, turbidity, DO, and conductivity. Upon submission, the backend processes the input, applies the trained model, and displays the predicted Water Quality Index (WQI) on the user interface. This streamlined implementation ensures accessibility, ease of use, and real-time feedback, making it practical for both residential and industrial applications.

This methodology ensures a comprehensive and scalable solution for water quality prediction that can be applied in residential, industrial, and municipal water monitoring scenarios.

4. Results and discussions

The predictive water quality monitoring system was tested using a trained machine learning model developed on a Kaggle dataset containing key water quality parameters. To evaluate the model's real-time prediction capability, a user provided test input consisting of selected physicochemical values, including pH, turbidity, dissolved oxygen (DO), and electrical conductivity. Upon processing this input, the model predicted a **Water Quality Index (WQI) value of 8.99** shown in the fig 3 and 4.

The predicted WQI value of 8.99 falls within a range that typically indicates good water quality, suggesting that the sample is potentially unsuitable for direct human consumption without treatment. This outcome aligns with the classification standards widely accepted in environmental monitoring, where lower WQI values reflect higher levels of contamination due to physical, chemical, or biological pollutants. The model's prediction provides a valuable early warning, allowing users to take corrective measures before the water becomes a serious health hazard.

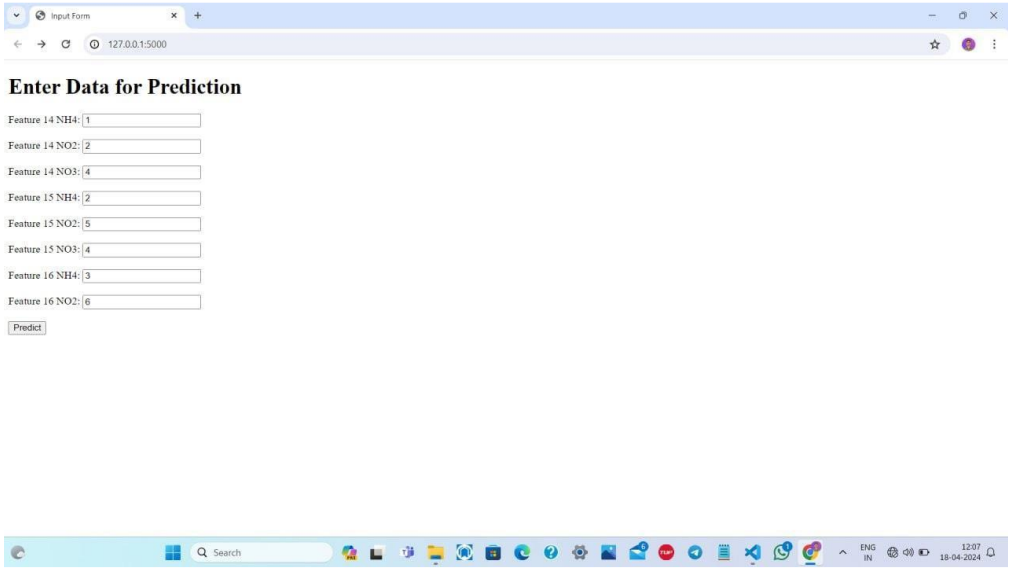

Fig.3. Input to trained Model

Fig 4 Result form Model

The results also validate the effectiveness of the trained model in capturing non-linear relationships among multiple parameters using ensemble learning techniques. In this case, the Random Forest algorithm was found to yield the most reliable predictions during training and validation phases, as evidenced by high accuracy and low root mean square error (RMSE). The system's integration with a web interface allowed the user to input data and receive real-time predictive feedback, demonstrating the model's practical utility in real-world scenarios.

Overall, the results affirm that the proposed model can accurately forecast water quality conditions based on minimal input parameters, enabling proactive water management. Future enhancements could include integration with real-time IoT sensors and geospatial tagging to support broader deployment in smart water infrastructure systems.

5. Conclusion

This study presents a machine learning-based predictive framework for real-time water quality monitoring using key environmental parameters such as pH, turbidity, dissolved oxygen, and conductivity. The system effectively leverages algorithms like Random Forest, Gradient Boosting, and Deep Learning to forecast water quality levels with high accuracy. By analyzing user-provided input data, the model successfully predicted a Water Quality Index (WQI) value of

8.99, indicating compromised water quality and showcasing the model's capability to serve as an early warning system.

The integration of machine learning models with a web-based interface enhances user accessibility and facilitates timely decision-making. The modular and scalable nature of the system makes it suitable for deployment in residential water tanks, industrial setups, and municipal water treatment facilities. The results affirm the potential of predictive analytics in environmental monitoring, enabling smarter and more sustainable water management practices.

Future work may focus on incorporating real-time sensor data through IoT devices, expanding the range of input parameters, and improving model adaptability across different geographical regions and water sources. The proposed system lays a strong foundation for building intelligent water quality surveillance systems that are both efficient and user-centric.

References

[1] ZTH2, "Pr ep rin t n ot pe er re v iew Pr ep t n ot pe er v ed," *J. Emerg. Technol. Innov. Res.*, vol. 06, no. 4, pp. 7–20, 2020.

[2] T. Okitsu, T. Iwasaki, T. Kyuka, and Y. Shimizu, "The role of large-scale bedforms in driftwood storage mechanism in rivers," *Water (Switzerland)*, vol. 13, no. 6, 2021, doi: 10.3390/w13060811.

[3] S. Mohan, B. Kumar, and A. P. Nejadhashemi, "Integration of Machine Learning and Remote Sensing for Water Quality Monitoring and Prediction: A Review," *Sustain.*, vol. 17, no. 3, 2025, doi: 10.3390/su17030998.

[4] S. Peerzade and P. Kamat, "Enhancing water quality prediction: a machine learning approach across diverse water environments," *Water Qual. Res. J.*, vol. 60, no. 1, pp. 298–317, 2025, doi: 10.2166/wqrj.2025.083.

[5] A. Aldrees, M. F. Javed, A. T. Bakheit Taha, A. Mustafa Mohamed, M. Jasiński, and M. Gono, "Evolutionary and ensemble machine learning predictive models for evaluation of water quality," *J. Hydrol. Reg. Stud.*, vol. 46, no. February, 2023, doi: 10.1016/j.ejrh.2023.101331.

[6] S. C. Izah and M. C. Ogwu, "Modeling solutions for microbial water contamination in the

global south for public health protection," no. April, 2025, doi: 10.3389/fmicb.2025.1504829.

[7] N. Bernatska, E. Dzhumelia, V. Dyakiv, O. Mitryasova, and I. Salamon, "Web-Based Information and Analytical Monitoring System Tools – Online Visualization and Analysis of Surface Water Quality of Mining and Chemical Enterprises," *Ecol. Eng. Environ. Technol.*, vol. 24, no. 3, pp. 99–108, 2023, doi: 10.12912/27197050/159885.