International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

Deep Multi Target Classification Framework for Water Quality Monitoring and Pollution Severity Level Assessment

Mugala Poojitha, T. Srilekha*

Vaagdevi Engineering College, Warangal, 506005, Telangana, India.

*Correspondence: T. Srilekha

Abstract

Water quality evaluation is essential for protecting ecosystems and safeguarding human health. Conventional monitoring techniques primarily rely on manual sample collection followed by laboratory-based chemical and biological testing. Although these methods provide reliable results, they require considerable time, labor, and resources, making them unsuitable for continuous or real-time monitoring and delaying critical environmental decisions. This study introduces an intelligent and automated framework for water quality analysis and pollution prediction by leveraging advanced Machine Learning (ML) algorithms. The proposed solution is developed as a Flask-based web application that integrates Classification and Regression Tree (CART) techniques for both pollution category identification and Water Quality Index (WQI) estimation. The implemented models include Linear Logistic Regression (LLR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Auto-Interpretable TaoTree (AITT). Among the classification models, AITT delivers the highest prediction accuracy for pollution level identification, while its regression variant achieves the best R^2 score for WQI estimation, demonstrating strong predictive capability. The application accepts 16 water quality parameters as user input and instantly generates pollution classifications along with WQI predictions. In addition, the platform provides Exploratory Data Analysis (EDA) visualizations, comparative model evaluation, and model retraining functionality, enabling environmental professionals to analyze trends and update predictive models using newly available datasets. By integrating CART-based machine learning workflows with the Flask framework, the system streamlines data preprocessing, model development, prediction, and visualization, minimizing manual intervention while improving accuracy and consistency. Built using scikit-learn, imodels, and pandas, the proposed framework offers a scalable, interpretable, and dependable approach to water quality assessment, supporting effective environmental monitoring and informed decision-making.

Keywords: water quality assessment, water quality index (WQI), pollution monitoring, environmental safety, public health, automated monitoring system, real-time analysis, flask web application, data preprocessing.

1. INTRODUCTION

Water is an essential resource for sustaining human life, and maintaining its quality is vital for improving overall health and living standards. The suitability of water for human consumption depends on whether it satisfies established water quality standards. Water that meets these standards is classified as potable and is safe for drinking, whereas non-potable water is intended for purposes other than direct human consumption. Before being used, non-potable water generally undergoes extensive treatment processes; however, it may still remain unsuitable for drinking [1]. According to the World Health Organization (WHO), in 2020, over 74% of the global population, representing nearly 5.8 billion people, had access to safely managed drinking water services where water was adequately treated to satisfy minimum safety and quality requirements, as illustrated in Fig. 1. Despite this progress, more than 2 billion people worldwide continue to rely on contaminated or unsafe drinking water sources [2]. Furthermore, reports indicate that approximately 1.8 billion individuals depend on non-potable water for their daily needs [3]. This lack of access to safe water has severe health consequences, particularly for children. The World Health Organization reported in 2017 that diarrhoeal diseases are responsible for the deaths of nearly 525,000 children under five years of age every year [4].

In earlier years, obtaining clean freshwater from surface and groundwater sources was comparatively less challenging than it is today. Rapid industrialization, population growth, and economic development have significantly increased the demand for water, leading to a rise in water pollution. In addition, inadequate wastewater management practices and the improper utilization of water resources have further contributed to the deterioration of water quality [5]. Consequently, continuous monitoring of water quality has become essential for identifying pollution levels, protecting freshwater resources, and supporting the implementation of effective environmental management strategies. However, accurately predicting water quality remains a complex task due to the large number of physical, chemical, and biological factors involved. As a result, researchers across the world have devoted significant efforts to developing reliable methods for water quality assessment because of its direct impact on human health and environmental sustainability.

2. LITERATURE SURVEY

Huang, et al. [6] Developed a hybrid framework that integrates deep learning and machine learning techniques for classifying Water Quality Index (WQI) categories. The proposed approach employed Convolutional Neural Networks (CNN), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Multi-Layer Perceptron (MLP). CNN transformed time-series groundwater information into two-dimensional matrices and utilized dilated convolution along with regularization methods to capture multiscale temporal features. The framework was evaluated using post-monsoon groundwater data collected from Telangana, India, through fourfold cross-validation. Experimental findings showed that CNN achieved the best performance with an RMSE of 0.0654 and an R^2 value of 0.9981, lowering prediction errors by 18–48% compared with KNN, NB, and MLP. The study demonstrated that CNN effectively captures spatial and temporal relationships while remaining computationally efficient, and its integrated prediction pipeline improved adaptability across diverse datasets. J. K. Pandya, et al. [7] Introduced an advanced framework for predicting water potability by combining an ensemble learning model with an interactive visualization dashboard. Initially, water quality parameters were collected and carefully preprocessed to maintain data consistency and reliability. The proposed ensemble incorporated Decision Trees (DT), Random Forest (RF), Gradient Boosting Machines (GBM), XGBoost, and Neural Networks, utilizing the strengths of each algorithm to improve prediction performance. The model produced an accuracy of 96.7%, precision of 96.7%, recall of 100%, and an F1-score of 98.4%, exceeding the performance of previously reported approaches. Chen, et al. [8] Performed a comprehensive review of Artificial Neural Network (ANN)-based techniques for water quality prediction by examining feedforward, recurrent, and hybrid neural network architectures. The analysis covered 151 research articles published between 2008 and 2019 and identified 23 commonly used water quality variables obtained mainly through sensors and specialized instruments such as UV-visible photometers. The study also summarized five different output modelling strategies and concluded that ANN techniques are highly effective for modelling water quality across rivers, lakes, reservoirs, groundwater, ponds, streams, and wastewater treatment facilities, providing valuable insights for future investigations.

Xu, et al. [9] Designed an automated framework for evaluating water quality and pollution levels using machine learning with water and sediment samples. Since the available sediment dataset contained missing values due to limited sampling locations, multiple data imputation methods were examined to determine the most effective approach for handling incomplete data. After preprocessing, the samples were categorized into four pollution classes according to standard guidelines. A classification algorithm was then trained to establish the relationship between water quality attributes and pollution categories. Finally, predicted outputs were compared with actual observations to assess the overall accuracy and reliability of the proposed framework. Guohao Zhang, et al. [10] Proposed an integrated deep learning model for predicting water pollutant concentrations by combining data preprocessing, frequency decomposition, feature enhancement, sample augmentation, and decoder-based regression modules. An enhanced wavelet transform algorithm was first utilized to identify periodic and fluctuating characteristics within the original data. An Informer-based encoder was subsequently applied to strengthen frequency-domain feature representations and improve their association with target variables through feature distillation, resulting in more precise pollutant concentration predictions. Mohammad Ehteram, et al. [11] Presented a hybrid Water Quality Index (WQI) prediction model by integrating Convolutional Neural Networks (CNN), Clockwork Recurrent Neural Networks (CRNN), and the M5 Tree algorithm. Since the standalone M5 Tree lacked sufficient capability for extracting complex information from water quality parameters, the hybrid approach improved feature representation and predictive accuracy. The General Linear Model Analysis of Variance (GLM-ANOVA) was employed to determine significant input variables, with all variables satisfying $p < 0.050$. Performance evaluation demonstrated the effectiveness of CNN-CRNN-M5T and other hybrid combinations, including CNN-CRNN, CRNN-M5T, CNN-M5T, CNN, CRNN, and M5T, for WQI prediction in a large river basin in Malaysia.

Lingling Zhu, et al. [12] Utilized deep learning techniques, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), for water quality classification and urban air quality analysis. The proposed Conv-LSTM architecture combined convolutional operations with recurrent memory units to capture both spatial and temporal dependencies within environmental datasets. In addition, detailed exploratory analysis was conducted to examine beach environments and waste distribution patterns. Experimental results indicated that Conv-LSTM achieved 92% accuracy for water pollution classification, outperforming RNN (65%), DBN (78%), and LSTM (82%). Likewise, for air pollution prediction, Conv-LSTM attained 91% accuracy, surpassing RNN (60%), DBN (75%), and LSTM (80%). Hangan, et al. [13] Reviewed recent advancements in information and communication technologies for monitoring, controlling, and managing water resources. The survey discussed emerging technologies associated with smart cities, intelligent water monitoring, big data analytics, and decision support systems. The

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

study demonstrated that combining multiple advanced technologies enables more efficient management of water infrastructure while introducing additional capabilities such as community participation in water resource management. The review serves as a valuable reference for researchers developing next-generation water monitoring systems. Yituo Zhang, et al. [14] Introduced an integrated Empirical Mode Decomposition–Long Short-Term Memory (EMD-LSTM) model to enhance water quality prediction accuracy. The proposed framework combined an EMD-based preprocessing stage with an LSTM prediction model. EMD preserved abnormal observations while effectively utilizing non-aligned data samples to reveal hidden data characteristics, whereas the LSTM network exploited its nonlinear learning capability to accurately predict time-series water quality data, leading to improved forecasting performance. Liu, et al. [15] Developed a water quality prediction model using high-quality datasets generated by Internet of Things (IoT)-enabled smart monitoring systems. As continuous sensing produced large volumes of complex water quality data, the study adopted Long Short-Term Memory (LSTM) deep neural networks to effectively capture temporal dependencies within the data. The proposed model successfully predicted drinking water quality from continuously generated time-series information, demonstrating the capability of deep learning techniques for intelligent and real-time water quality monitoring.

3. PROPOSED SYSTEM

This Research Water Quality Prediction System is designed as an integrated framework that combines data preprocessing, machine learning model development, and a Flask-based web application to provide accurate prediction of water pollution levels and the Water Quality Index (WQI). Initially, the water quality dataset is processed to remove inconsistencies, define input features, and prepare the data for model training. As illustrated in Fig. 2, multiple classification and regression models are trained and evaluated to identify the best-performing predictive model. The trained models and preprocessing components are then stored and deployed through the Flask backend for real-time prediction. The application also includes authentication, Exploratory Data Analysis (EDA), model comparison, and prediction modules that enable environmental engineers to analyze data and users to obtain instant pollution level and WQI predictions. This integrated architecture automates the complete water quality assessment workflow, reducing manual effort while improving prediction accuracy, interpretability, and decision-making efficiency.

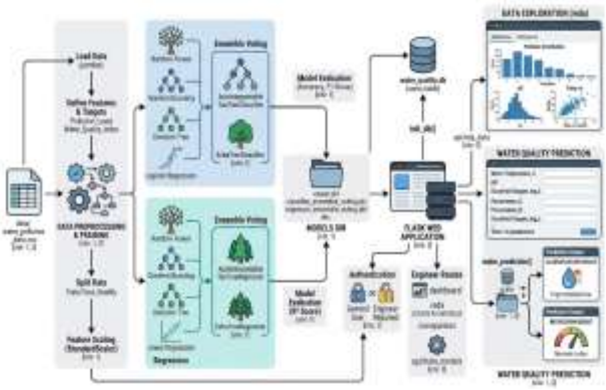

Fig 2: Proposed system architecture for urban water assessment

Data Collection and Preprocessing

The system begins by loading the water quality dataset using the Pandas library. The required input features and target variables, namely Pollution Level and Water Quality Index (WQI), are identified and processed. Data preprocessing operations such as cleaning, handling inconsistencies, train-test splitting, and feature scaling using StandardScaler are performed to prepare high-quality input data for model training.

Model Training for Classification and Regression

The preprocessed dataset is used to train multiple machine learning models for both classification and regression tasks. Classification models predict pollution categories, while regression models estimate the Water Quality Index. Ensemble learning techniques are employed to improve prediction performance by combining the strengths of several individual models.

Model Evaluation and Model Storage

Each trained model is evaluated using standard performance metrics to determine its effectiveness. Classification models are assessed using Accuracy and F1-Score, whereas regression models are evaluated using the R^2 Score. The best-performing models, along with the feature scaler, are saved in the models directory for deployment and future prediction.

Flask-Based Web Application Integration

The trained models are integrated into a Flask web application that manages authentication, backend processing, and communication between users and predictive models. The application provides dedicated routes for dashboard visualization, EDA analysis, model comparison, retraining, and prediction. This architecture enables efficient interaction between the machine learning models and the web interface.

Real-Time Prediction and Visualization

Users enter sixteen water quality parameters through the web interface to obtain instant predictions. The system applies the stored preprocessing pipeline, loads the trained models, and generates both pollution level classification and Water Quality Index predictions. The prediction results, along with EDA visualizations and performance comparisons, are displayed through an interactive interface, enabling effective water quality monitoring and informed environmental decision-making.

4. RESULTS ANALYSIS

The results section presents the key findings of the study in a clear and organized manner. It highlights the main outcomes obtained from the data analysis, showing patterns, trends, or relationships relevant to the research objectives. The section may include tables, graphs, or charts to support the findings and make them easier to understand. It focuses on factual information without interpretation, allowing readers to see what was discovered. The results are usually structured logically, often following the order of the research questions or hypotheses. This section provides a concise summary of what the study revealed.

Fig 3 shows, the system displays the performance results of the AITT, which combines multiple models to achieve significantly higher accuracy and reliability in predicting water pollution levels. The top section presents impressive performance metrics, including accuracy, precision, recall, and F1-score each exceeding 97%, indicating outstanding predictive capability. The Confusion Matrix visualization demonstrates the classifier’s strong ability to correctly classify all five pollution categories with minimal misclassification. Additionally, the Classification Report provides detailed precision, recall, and F1-score values for each class, along with the total support count. This figure highlights the superiority of the AITT Classifier compared to individual models and confirms its role as the most robust and accurate classification model in the system.

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

Fig. 3: AITT classifier performance visualization

Fig 4: AITT regressor model performance visualization.

Fig 4 the AITT regressor achieves near-perfect alignment between predicted and actual WQI values. Residuals are minimal and evenly distributed, reflecting accurate capture of both low and high WQI ranges. With an R^2 of 0.9347, low MAE of 1.2564, and RMSE of 7.4572, the model demonstrates strong predictive capability across the entire dataset. This visualization confirms AITT as the most reliable model for WQI prediction, effectively representing complex patterns in water quality data.

Fig 5: Prediction Interface for Water Quality Assessment

Fig 5 shows, the prediction interface enables users to input various water sample parameters and generate real-time predictions for pollution level and water quality index. The upper section provides input fields for sixteen key physical, chemical, and biological parameters such as temperature, pH, dissolved oxygen, turbidity, nutrients, heavy metals, and coliform count. After entering the values, users can click the “Predict Water Quality” button to obtain results. The system displays outputs from multiple ML models, including RF, GB, DT, Logistic Regression, and AITT. The left panel shows pollution level classification results, while the right panel presents numerical WQI predictions for each model. This interface provides an interactive, user-friendly environment for evaluating water quality based on real-time input data.

Model	Accuracy	Precision	Recall	F1-Score
RF model	79.39%	79.53%	79.39%	79.37%
GB model	53.17%	53.04%	53.17%	53.03%
DT model	20.61%	12.41%	20.61%	14.98%
Logistic Regression	24.11%	23.95%	24.11%	23.30%
AITT model	**97.33%**	**97.36%**	**97.33%**	**97.33%**

Table 1: Comparative analysis of classification models for water pollution level prediction

Table 1 presents the comparative performance of various classification models for predicting water pollution levels. Among traditional models, the RF achieves the highest accuracy of 79.39%, while GB, DT, and Logistic Regression exhibit significantly lower performance. The AITT model outperforms all other models with an accuracy of 97.33%, along with consistently high precision, recall, and F1-score. These results demonstrate the superior capability of the AITT model in capturing complex patterns in water quality data, providing highly reliable predictions for environmental monitoring.

Model	**MAE**	**MSE**	**RMSE**	**R^2 Score**
RF model	9.1117	477.5896	21.8538	0.4391
GB model	23.4803	745.7748	27.3089	0.1241
DT model	25.2127	848.3128	29.1258	0.0037
Linear Regression	25.2216	849.9767	29.1544	0.0018
AITT model	**1.2564**	**55.6096**	**7.4572**	**0.9347**

Table 2: Comparative analysis of regression models for WQI prediction

Table 2 compares the performance of different regression models for WQI prediction. Among conventional models, RF shows moderate performance with an R^2 of 0.4391, while GB, DT, and LR exhibit low predictive capability. The AITT model demonstrates superior accuracy with an MAE of 1.2564, RMSE of 7.4572, and an R^2 of 0.9347. These results highlight AITT's effectiveness in capturing complex patterns in water quality data, making it the most reliable model for WQI prediction.

5. CONCLUSION

This research successfully combines Machine Learning (ML) algorithms with a Flask-based web application to provide accurate, real-time, and easily interpretable predictions of water pollution levels and the Water Quality Index (WQI). The framework follows a structured workflow that includes data preprocessing, training of multiple predictive models, and performance evaluation using standard assessment metrics to ensure reliable and meaningful prediction outcomes. Among the implemented models, the AITT CART model achieves the best overall performance by effectively identifying complex relationships within the dataset while maintaining high prediction accuracy and model interpretability, which are essential for environmental monitoring applications. By automating the water quality assessment process, the system significantly reduces the dependence on time-consuming laboratory testing and manual analysis. Furthermore, the integrated platform enables environmental engineers to perform Exploratory Data Analysis (EDA), train and compare different predictive models, and generate accurate predictions through a unified backend, while the user-friendly web interface

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

provides fast, transparent, and actionable insights for effective water quality management and decision-making.

REFERENCES

[1]. Varila M., "What Is Potable Water? Your Guide to Understanding Types of Water", viralrang, 2020. [Online]. Available: https://viralrang.com/what-is-potable-water-your-guide-to-understanding-types-of-water/#

[2]. UNECE, "miyah alshrob," who, (2022). [Online]. Available: https://www.who.int/ar/news-room/fact-sheets/detail/drinking-water

[3]. Fluence news team, "What Is Potable Water?", fluencecorp, 2019. [Online]. Available: https://tinyurl.com/2qj936u9.

[4]. World Health Organization, "Preventing diarrhoea through better water, sanitation and hygiene: exposures and impacts in low- and middle-income countries," World Health Organization (Report), Villars-sous-Yens, Switzerland, 2019.

[5]. World Health Organization, "Diarrhoeal disease," who, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease.

[6]. Huang, T., Chau, K.Y., Zhan, S. et al., "Hybrid deep learning and ML framework for high-precision water quality prediction in urban systems," Environmental Development and Sustainability, 2025. [Online]. Available: https://doi.org/10.1007/s10668-025-06447-2

[7]. J. K. Pandya, S. S. Khandelwal, R. K. Tipu and K. S. Pandya, "Advancing Water Quality Management: An Integrated Approach Using Ensemble ML and Real-Time Interactive Visualization," IEEE Access, vol. 13, pp. 92406–92428, 2025. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3573589

[8]. Chen, Y., Song, L., Liu, Y., Yang, L. and Li, D., "A Review of the Artificial Neural Network Models for Water Quality Prediction," Applied Sciences, vol. 10, no. 17, p. 5776, 2020. [Online]. Available: https://doi.org/10.3390/app10175776

[9]. Xu, X., Lai, T., Jahan, S., Farid, F. and Bello, A., "A ML Predictive Model to Detect Water Quality and Pollution," Future Internet, vol. 14, no. 11, p. 324, 2022. [Online]. Available: https://doi.org/10.3390/fi14110324

[10]. Guohao Zhang, Cailing Wang, Hongwei Wang and Yu Tao, "Advanced deep learning model for predicting water pollutants using spectral data and augmentation techniques: A case study of the Middle and Lower Yangtze River, China," Process Safety and Environmental Protection, vol. 197, p. 107058, 2025. [Online]. Available: https://doi.org/10.1016/j.psep.2025.107058

[11]. Mohammad Ehteram, Ali Najah Ahmed, Mohsen Sherif and Ahmed El-Shafie, "An advanced deep learning model for predicting water quality index," Ecological Indicators, vol. 160, p. 111806, 2024. [Online]. Available: https://doi.org/10.1016/j.ecolind.2024.111806

[12]. Lingling Zhu, Zuhra Junaida Binti Mohamad Husny, Noor Aimran Samsudin, HaiPeng Xu and Chongyong Han, "Deep learning method for minimizing water pollution and air pollution in urban environment," Urban Climate, vol. 49, p. 101486, 2023. [Online]. Available: https://doi.org/10.1016/j.uclim.2023.101486

Received: 30-05-2026 | Accepted: 01-07-2026 | Published: 08-07-2026

[13]. Hangan, A., Chiru, C.-G., Arsene, D., Czako, Z., Lisman, D.F., Mocanu, M., Pahontu, B., Predescu, A. and Sebestyen, G., "Advanced Techniques for Monitoring and Management of Urban Water Infrastructures—An Overview," Water, vol. 14, no. 14, p. 2174, 2022. [Online]. Available: https://doi.org/10.3390/w14142174

[14]. Yituo Zhang, Chaolin Li, Yiqi Jiang, Lu Sun, Ruobin Zhao, Kefen Yan and Wenhui Wang, "Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model," Journal of Cleaner Production, vol. 354, p. 131724, 2022. [Online]. Available: https://doi.org/10.1016/j.jclepro.2022.131724

[15]. Liu, P., Wang, J., Sangaiah, A.K., Xie, Y. and Yin, X., "Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment," Sustainability, vol. 11, no. 7, p. 2058, 2019. [Online]. Available: https://doi.org/10.3390/su11072058