International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Urban Water Assessment using Auto Interpretable Learning for Pollution Classification and Quality Index Modeling

M. Amareswar[1]*, Salwar Sai Vardhan[1], Pala Deepak[1], Bhusarapu Dolasura Veera Venkata Ganesh[1]

[1]Department of Computer Science and Engineering (DS), Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: M. Amareswar

Abstract

Water quality assessment plays a crucial role in ensuring environmental safety and public health. Traditional approaches depend on manual sampling and laboratory-based chemical and biological analyses. These methods are labor-intensive, time-consuming, and often fail to provide real-time insights, limiting their applicability for timely decision-making. This research proposes an intelligent, automated framework for water quality monitoring and pollution prediction using modern Machine Learning (ML) techniques. The system is implemented as a Flask-based web application integrated with Classification and Regression Tree (CART) models for both pollution level classification and Water Quality Index (WQI) prediction. The models utilized include Linear Logistic Regression (LLR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Auto-Interpretable TaoTree (AITT). For pollution classification, the AITT model achieves the highest accuracy, whereas for WQI regression, the same model, adapted for regression, produces the best R^2 score, demonstrating robust predictive performance. Users provide 16 water sample parameters through the interface, enabling instant predictions for pollution levels and WQI values. The system also features Exploratory Data Analysis (EDA) visualizations, model performance comparison, and retraining capabilities, empowering environmental engineers to monitor trends and adapt models to new datasets. By combining CART-based ML pipelines with Flask, the system automates preprocessing, model training, prediction, and visualization, significantly reducing manual effort and enhancing reliability. Implemented using scikit-learn, imodels, and pandas, this framework delivers a scalable, interpretable, and accurate solution for water quality assessment, providing actionable insights for environmental management and 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

Human life, in general, depends on the availability of water, and the water quality is one of the important factors affecting the practical improvement of daily life. A specific standard of water quality needs to be reached for the water to be considered potable water, which is safe for humans to drink. In contrast, non-potable water is the water that is used for everything except human use. Many large-scale procedures are carried out on non-potable water before use, although it remains unfit for direct human consumption [1]. According to the World Health Organization, in 2020 more than 74% of the world's population (5.8 billion people) use safely managed water where the water is treated well to reach the minimum limit of safety and quality standards as shown in Fig. 1. However, more than 2 billion people in the World have access only to polluted or undrinkable water [2]. For example, according to it was reported that approximately 1.8 billion people worldwide use non-potable water sources [3]. As a result, it affects the lives of people especially children, resulting in their death. According to a 2017 report

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

from the World Health Organization, about 525,000 children under the age of five die from diarrhoea every year [4].

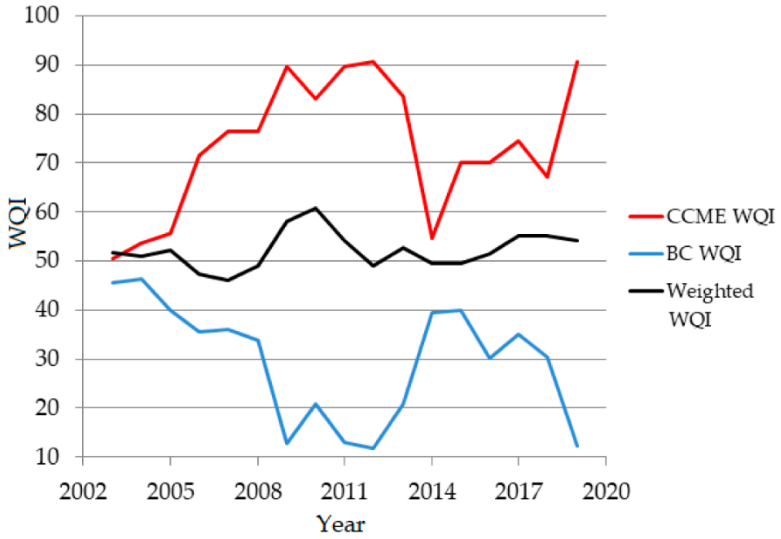

Fig. 1: Water quality index per very year.

The process of obtaining fresh water from ground and surface water in the past was easier than now. This is due to the increased dependence of human life on the availability of water. A rise in the issues of water pollution is also a result of the industrial and economic growth that humanity has reached, as well as a lack of knowledge about the right use of wastewater and adequate water use [5]. It is important to monitor water quality to find out the degree of water pollution, ensure access to clean water resources and apply effective guidelines for the protection of Water Resources. Predicting water quality is a difficult task. Many researchers have made a great effort in determining water quality because of its importance to human life.

Therefore, it is an urgent necessity for humans to provide safe drinking water as it maintains the health of the kidneys, and the intestines nourish the muscles and help to maintain body fluid. To ensure the potability of water, it is important to devise new technologies and methods. The water quality index is a method used for measuring water quality which reflects the impact of different water standards on its quality. The calculation of WQI is necessary to determine the usability of water and to know the water specifications It converts complex analyses of water properties and huge amounts of data into easy information that can be understood and used by specialists and non-specialists. The water quality measurement index has been evaluated by many international studies as the basis for measuring various water indicators.

2. LITERATURE SURVEY

Xu, et al. [6] Proposed an automated water quality assessment framework where we formalise a predictive model using ML to infer the water quality and level of pollution using collected water and sediments samples. Firstly, due to the sparsity of sample collection locations, the amount of sediment samples of water is limited, and the dataset is incomplete. Therefore, after an extensive investigation on various data imputation methods' performance in water and sediment datasets with different missing data rates, we chose the best imputation method to process the missing data. Afterwards, the water sediment sample will be tagged as one of four levels of pollution based on some guidelines and then the ML model will use a specific technique named classification to find the relationship between the

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

data and the result. After that, the result of prediction can be compared to the real result so that it can be checked whether the model is good and whether the prediction is accurate.

Liu, et al. [7] Focused on a water quality prediction model which requires high-quality data. In the process of construction and operation of smart water quality monitoring systems based on Internet of Things (IoT), more big data are produced at a high speed, which has made water quality data complicated. Taking advantage of the good performance of long short-term memory (LSTM) deep neural networks in time-series prediction, a drinking-water quality model was designed and established to predict water quality big data with the help of the advanced deep learning.

Hangan, et al. [8] Presented an overview of the latest research related to information and communication technology systems for water resource monitoring, control and management. The main objective of our review is to show how emerging technologies offer support for smart administration of water infrastructures. The paper covered research results related to smart cities, smart water monitoring, big data, data analysis and decision support. Our evaluation reveals that there are many possible solutions generated through combinations of advanced methods. Emerging technologies open new possibilities for including new functionalities such as social involvement in water resource management. This review offered support for researchers in water monitoring and management to identify useful models and technologies for designing better solutions.

Lingling Zhu, et al. [9] Used deep learning methods such as Convolutional Neural Networks (CNN) and Long-Short Term Memory to classify water quality. Next, it identifies the air quality in Urban Development. The convolutional LSTMs use convolutional layers and the recurrent connections found in LSTMs. This allows the model to capture spatial dependencies in the input data in addition to the temporal dependencies captured by the recurrent connections. We also use thorough exploratory analysis to investigate the various beach habitats and the kinds of trash discovered in multiple places. Lowering water pollution and raising air quality are both strategies that can be employed to ensure sustainable urban development. The performance metrics such as accuracy, recall, precision, and F1-score are evaluated and classify the water pollution efficiently. In the water pollution dataset, the algorithms of RNN 65%, DBN 78%, LSTM 82%, and the proposed work of Conv.LSTM 92%. Similarly, for the air pollution dataset, the algorithms of RNN 60%, DBN 75%, LSTM 80%, and the proposed work of Conv.LSTM 91%.

J. K. Pandya, et al. [10] Presented a novel approach to predicting water potability by developing an advanced ensemble model and an interactive visualization dashboard. A comprehensive dataset of water quality parameters was collected and pre-processed to ensure data integrity. An ensemble model combining DTs, RF, Gradient Boosting Machines (GBM), XGBoost and Neural Networks was constructed, leveraging the strengths of each algorithm to enhance predictive accuracy. The model achieved an accuracy of 96.7%, precision of 96.7%, recall of 100%, and F1-score of 98.4%, outperforming existing models in the literature.

Yituo Zhang, et al. [11] Proposed an integrated EMD-LSTM model that combines the data preprocessing module cantered on empirical mode decomposition (EMD) and the long short-term memory (LSTM) neural network prediction module to improve the accuracy of the modelling-based detection methods. In the integrated EMD-LSTM model, EMD allows retaining outliers and utilizing data on non-aligned moments, which contributes to capturing data patterns, while powerful nonlinear mapping and learning ability of LSTM neural network enables the time series prediction of water quality.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Chen, et al. [12] Conducted extensive investigation and analysis on ANN-based water quality prediction from three aspects, namely feedforward, recurrent, and hybrid architectures. Based on 151 papers published from 2008 to 2019, 23 types of water quality variables were highlighted. The variables were primarily collected by the sensor, followed by specialist experimental equipment, such as a UV-visible photometer. Five different output strategies, namely Univariate-Input-Itself-Output, Univariate-Input-Other-Output, Multivariate-Input-Other(multi)-output, Multivariate-Input-Itself-Other-Output, and Multivariate-Input-Itself-Other (multi)-Output, are summarized. From results of the review, it can be concluded that the ANN models can deal with different modelling problems in rivers, lakes, reservoirs, wastewater treatment plants, groundwater, ponds, and streams. The results of many of the review articles are useful to researchers in prediction and similar fields.

Mohammad Ehteram, et al. [13] Developed a new hybrid model for predicting the WQI. The study uses a combination of a convolutional neural network, clockwork recurrent neural network, and M5 Tree to predict a WQI. The M5T model lacks advanced operators for extracting meaningful data from water quality parameters, so the new model enhances its ability to analyse intricate patterns. The general linear model analysis of variance is an improved version of the ANOVA. Our study uses the GLM-ANOVA to determine significant inputs. As all input variables had $p < 0.050$, they were defined as significant variables. Results showed that NH-NL and PH had the highest and lowest impact, respectively. Our study used the CNN-CRNN-M5T, CNN-CRNN, CRNN-M5T, CNN-M5T, CNN, CRNN and M5T models to predict the WQI of a large basin in Malaysia.

Guohao Zhang, et al. [14] Constructed an integrated deep learning framework for predicting water pollutant concentrations, incorporating several key modules including data preprocessing, frequency decomposition, feature enhancement, sample augmentation, and decoder regression prediction. In the established model, an improved wavelet transform algorithm is first employed to address the issue of original data being unable to effectively distinguish detailed features, thereby accurately extracting the periodicity and volatility characteristics of the data. Secondly, an encoder module based on the Informer architecture enhances various frequency domain features and further improves the quality of features and their correlation with labels through distillation techniques.

Huang, et al. [15] Proposed a hybrid deep learning and ML framework that classifies Water Quality Index (WQI) categories using Convolutional Neural Networks (CNN), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Multi-Layer Perceptron (MLP). CNN processes time-series groundwater data as 2D matrices, leveraging dilated convolutions and regularization techniques to extract multiscale temporal patterns. The model was trained and validated on post-monsoon groundwater datasets from Telangana, India, employing fourfold cross-validation. Among the algorithms tested, CNN achieved the best performance with RMSE: 0.0654 and R^2: 0.9981, reducing prediction error by 18–48% over KNN, NB, and MLP. These results highlight CNN's superior ability to learn from spatial–temporal dynamics while maintaining computational efficiency. Unlike previous studies focused solely on regression models or singular algorithms, this study combined multiple classifiers into a unified prediction pipeline, enhancing adaptability across heterogeneous datasets.

3. PROPOSED SYSTEM

The proposed system presents an advanced and intelligent water-quality assessment framework that utilizes ML techniques to accurately predict pollution levels and estimate WQI. Unlike conventional approaches that rely on fixed thresholds and manual evaluation, this system analyzes complex patterns within historical water-quality datasets to deliver more accurate and adaptive predictions. The architecture incorporates key components such as data preprocessing, feature scaling, label encoding,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

model training, evaluation, and real-time prediction, all integrated within a Flask-based web application as shown in Fig. 2. It employs multiple ML models including RF, GB, LLR, DT, along with advanced interpretable models like AITT and ET, ensuring high performance, reliability, and transparency in decision-making.

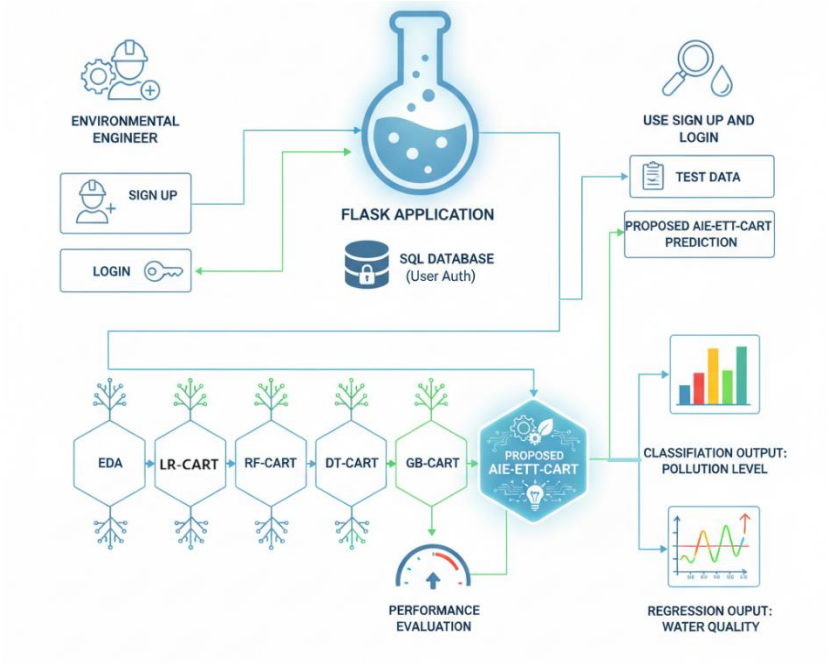

Fig 2: Proposed system architecture

These models work together to classify pollution levels into Low, Medium, or High and to regress the WQI value based on physicochemical parameters such as pH, temperature, DO, turbidity, COD, BOD, nitrates, and heavy metal concentrations. Additionally, the system provides a secure, role-based user interface where environmental engineers can train models, compare performance metrics, and conduct EDA through interactive dashboards. Normal users can access a simplified prediction module to input water sample parameters and obtain immediate pollution and WQI results. All computations, model loading, and prediction processes occur in real time, enabling a highly responsive and automated environmental monitoring solution. Overall, the proposed system enhances decision-making accuracy, reduces manual effort, and provides a scalable, data-driven framework for water quality monitoring and environmental protection.

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.

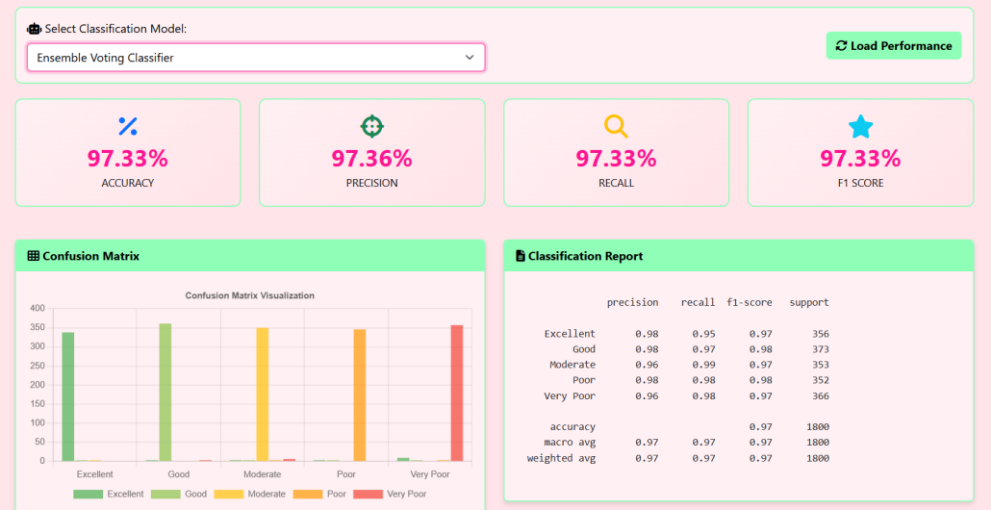

Fig. 3: AITT classifier performance visualization

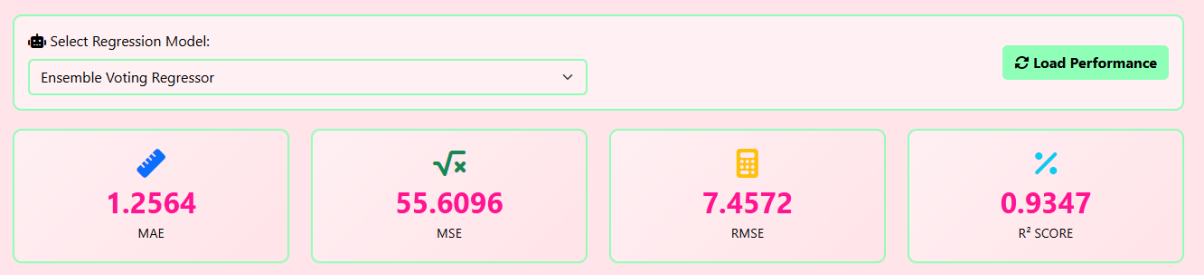

Fig 4: AITT regressor model performance visualization.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

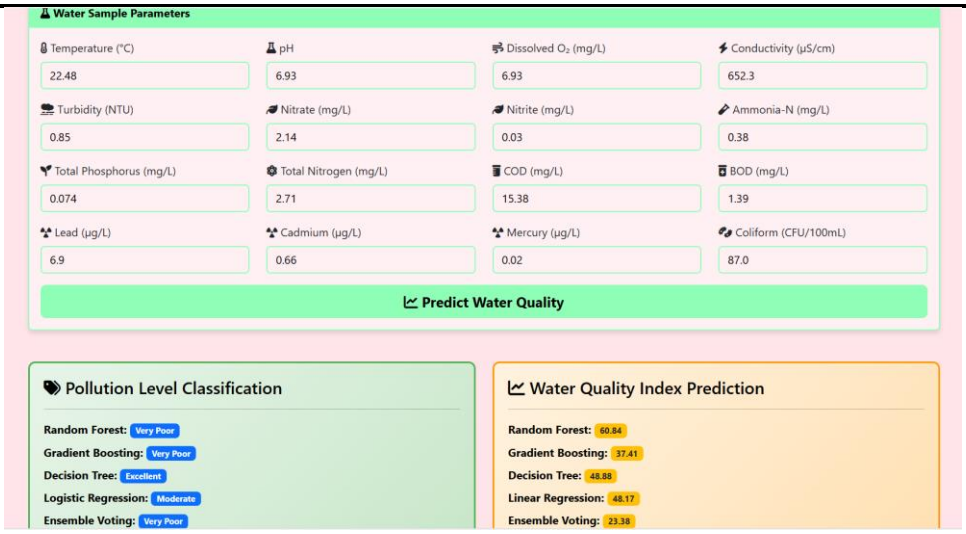

Fig 5: Prediction Interface for Water Quality Assessment

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.

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

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 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.

Table 2: Comparative analysis of regression models for WQI prediction

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 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

The proposed Water Quality Prediction System effectively integrates ML techniques with a Flask-based web interface, delivering real-time, reliable, and interpretable predictions for water pollution levels and WQI. The system employs systematic preprocessing, trains multiple models, and evaluates performance using standard metrics, ensuring predictions are accurate and scientifically meaningful. The AITT CART model demonstrates superior capability by capturing complex patterns in the dataset, providing high accuracy and robust performance while maintaining interpretability, a key requirement in environmental applications. The system streamlines the traditionally complex process of water quality assessment by automating tasks that otherwise require laboratory analysis and extensive manual effort. Environmental engineers can perform EDA, train and compare multiple models, and deploy predictions efficiently using the integrated backend, while users benefit from a simple web interface that delivers consistent, transparent, and actionable water quality insights.

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] Xu, X.; Lai, T.; Jahan, S.; Farid, F.; Bello, A. A ML Predictive Model to Detect Water Quality and Pollution. Future Internet 2022, 14, 324. https://doi.org/10.3390/fi14110324

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[7] Liu, P.; Wang, J.; Sangaiah, A.K.; Xie, Y.; Yin, X. Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment. Sustainability 2019, 11, 2058. https://doi.org/10.3390/su11072058

[8] Hangan, A.; Chiru, C.-G.; Arsene, D.; Czako, Z.; Lisman, D.F.; Mocanu, M.; Pahontu, B.; Predescu, A.; Sebestyen, G. Advanced Techniques for Monitoring and Management of Urban Water Infrastructures—An Overview. Water 2022, 14, 2174. https://doi.org/10.3390/w14142174

[9] Lingling Zhu, Zuhra Junaida Binti Mohamad Husny, Noor Aimran Samsudin, HaiPeng Xu, Chongyong Han, Deep learning method for minimizing water pollution and air pollution in urban environment, Urban Climate, Volume 49, 2023, 101486, ISSN 2212-0955, https://doi.org/10.1016/j.uclim.2023.101486.

[10] 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," in IEEE Access, vol. 13, pp. 92406-92428, 2025, doi: 10.1109/ACCESS.2025.3573589

[11] Yituo Zhang, Chaolin Li, Yiqi Jiang, Lu Sun, Ruobin Zhao, Kefen Yan, Wenhui Wang, Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model, Journal of Cleaner Production, Volume 354, 2022, 131724, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2022.131724

[12] Chen, Y.; Song, L.; Liu, Y.; Yang, L.; Li, D. A Review of the Artificial Neural Network Models for Water Quality Prediction. Appl. Sci. 2020, 10, 5776. https://doi.org/10.3390/app10175776

[13] Mohammad Ehteram, Ali Najah Ahmed, Mohsen Sherif, Ahmed El-Shafie, An advanced deep learning model for predicting water quality index, Ecological Indicators, Volume 160, 2024, 111806, ISSN 1470-160X, https://doi.org/10.1016/j.ecolind.2024.111806

[14] Guohao Zhang, Cailing Wang, Hongwei Wang, 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, Volume 197, 2025, 107058, ISSN 0957-5820, https://doi.org/10.1016/j.psep.2025.107058

[15] Huang, T., Chau, K.Y., Zhan, S. et al. Hybrid deep learning and ML framework for high-precision water quality prediction in urban systems. Environ Dev Sustain (2025). https://doi.org/10.1007/s10668-025-06447-2