Machine Learning Approaches for Water Quality Index Prediction: A Systematic Review

Group Members: Md. Rajaul Karim and Md. Jahidul Islam

GitHub Repository: https://github.com/mk19409-prog/Literature-Review-on-WQI-ML
Detailed Literature Review Sheet: https://docs.google.com/spreadsheets/d/1uNeQQFx8vY5-hyzwO9BjR6l4s6zKMueg/edit?usp=sharing&ouid=106348839836730303077&rtpof=true&sd=true

I. Introduction

Water is an essential natural resource that supports ecological balance, sustains human livelihoods, and drives agricultural, industrial, and environmental systems. Surface water is one of the numerous water sources that can be utilized to meet the freshwater needs of all societies in the world, such as rivers, lakes and reservoirs [1]. Unregulated urbanization, industrial effluents, agricultural activities, and climate-induced events have contributed to the accumulation of major pollutants in waterways. Traditional ways of assessing water quality are often based on lab analysis and manual interpretation; these methods are precise but resource-intensive, time-consuming, and ineffective for real-time or large-scale monitoring. Water Quality Index (WQI) is a quantitative method used to evaluate the overall condition of a water body based on selected physical, chemical, and biological parameters [2].

Machine learning improves water quality assessment by modeling complex and nonlinear relationships that can adapt in different geographic, seasonal, and hydrological conditions. Different ML models (such as Decision Trees, Random Forests, Support Vector Machines, Regression, and Gradient Boosting) and DL models (Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory networks, and transformers) are widely used for the prediction and forecasting of the water quality index [3]. In addition, multi-task learning frameworks allow the simultaneous prediction of multiple WQIs for different water uses by sharing common feature representations while generating task-specific outputs, improving scalability and generalization across diverse contexts [4].

Most existing WQI models are region-specific (e.g., designed for Canada or Ireland) or source-specific (e.g., rivers or wetlands) and produce a single index based on fixed parameter weightings and thresholds. Such static designs lack flexibility for multi-context environments where water bodies simultaneously support diverse uses such as drinking, agriculture, fisheries etc. These have created a critical research gap for the further review of a universal and adaptive AI-intensive WQI model that will enable mass application to a wide range of purposes with modern sustainability demands.

II. Methodology

For selection the research papers, we have followed the procedure:

keyword Used:

- Water Quality Index (WQI) Prediction/Classification and Machine Learning (ML) Algorithms
- Water Quality Index (WQI) prediction and Artificial Intelligence (AI)
- Water Quality Index (WQI) estimation and Deep Learning (DL)
- Water Quality Index (WQI) prediction and Hybrid Machine Learing Algorithms

Databases Accessed: We have accessed the following databases:

- ScienceDirect
- Scopus
- IEEE Xplore
- Web of Science
- SpringerLink
- Google Scholar
- GALILEO Search

Inclusion and Exclusion criteria We have presented some inclusion and exclusion criteria:

Inclusion Criteria

- Focused on Water Quality Index (WQI) prediction and classification using solor or hybrid machine learning.
- High-quality peer-reviewed journal (Elsevier, IEEE, springer (Minumum quartiles: Q1 and Q2)) and conference (Ranked: A and A (star)).
- Research Paper Published in Recent years (e.g, 2022, 2023, 2024, 2025 and 2026).
- Published in the English language.
- Recent/update methodological details.
- Dataset: Authentic sources, complete parameters for WQI.
- Initial selection Research papers: 35, selected based on our research (WQI): 15, finally selected: 08 journals (WQI related).

Exclusion Criteria

- Insufficient methodological details.
- Journal papers: not peer reviewed; conference papers: not list in the ranking.
- Predatory journals.
- Dataset: unauthenticated source, insufficient, and incomplete parameters.
- Not written in English.

TABLE I
RECENT STATE-OF-THE-ART

Reference	Problem addressed	Methods	Findings and contributions	Limitations/Gaps	Dataset
(Bataineh et al., 2026) [5], [Q1, IEEE]	Comparison to traditional methods, a hybrid model incorporated XGBoost boosting with SHAP-based feature-informed neural network initialization can increase the accuracy of Water Quality Index prediction.	Models: ANN, XGBoost, and Hybrid ANN + XGBoost model (proposed model) Baseline models for comparison: Random Forest, Support Vector Regression, Standard ANN, Standalone XGBoost	When compared to random initialization, SHAP-based weight initialization enhanced neural network convergence and performance. Accuracy: 86.9%, F1-score: 84.9%, ROC-AUC: 89.4%	The dataset employed binary classification, although multiple classification is necessary in real life. There is no specific reference of potability measures.	Public Water Potability dataset from Kaggle:
(Nishat et al., 2025) [6], [Q1, Springer]	Dhaka rivers are severely contaminated by sewage, industrial discharge, and urban runoff; precise WQI forecasting is required for environmental management.	Machine Learning Models (14): ANN, RFR, Decision Tree Regression, Linear Regression, Ridge Regression, SGD Regressor, XGBoost and etc.	The best result was obtained by ANN (RMSE=2.34, MAE=1.24, NSE=0.97, R2=0.97), followed by Random Forest; ML greatly increases the accuracy of WQI forecasting.	Reliance on historical data, lack of climatic elements, regional dataset limitations, and restricted physicochemical parameters. Very small dataset	262 samples (2001–2023) from the Bangladesh Water Development Board dataset; rivers: Buriganga, Turag, Balu, and Tongi Khal; parameters: DO, pH, EC, TDS, Fe, and Cl
(Han et al. 2025) [3] [Q1, Elsevier]	How to use machine learning to precisely estimate and categorize the Water Quality Index (WQI) to enhance environmental sustainability and urban waste management.	LSTM (Deep learning), Random Forest, Decision Tree, Support Vector Machine.	LSTM performed better than all other models; RF was second best; DT and SVM fared worse.	Single-municipality dataset; restricted geographic generalization; potential bias in the dataset; computational difficulty; problems with scalability	Telangana Post-Monsoon Groundwater Quality Data, a Kaggle dataset; parameters: pH, EC, TDS, and SAR; WQI divided into nine classifications (Good–Poor)
(Mohseni et al. 2024) [7] [Q1, Elsevier]	Groundwater quality must be accurately predicted using AI models because traditional water quality monitoring is expensive and ineffective.	ANN, Multiple Linear Regression, Support Vector Machine, Random Forest, XG-Boost and Stacking Ensemble Model;	All ML models obtained high prediction accuracy; the ensemble model demonstrated the best prediction accuracy, with XG-Boost outperforming the single models.	Lack of meteorological or environmental factors; small dataset size; restricted geographic coverage; computational resource requirements	54 groundwater samples were taken from 54 wards in Ujjain City, India; the sources were hand pumps, bore wells, and dug wells; the parameters were pH, turbidity, EC, TDS, alkalinity, hardness, chloride, and fluoride.
(Satish et al. 2024) [8] [Q1, Elsevier]	How can the shortcomings of traditional neural networks be overcome by a hybrid machine learning framework to increase the accuracy of Water Quality Index (WQI) prediction?	Advanced Deep Learning model (hybrid/optimized architecture) Traditional ML models: ANN, SVM, RF, and etc.,	shown stability in WQI prediction for complicated nonlinear connections.	Lack of real-time deployment, computational complexity,	Physicochemical parameters including pH, DO, BOD, COD, turbidity, temperature, and more are included in this multi-parameter water quality dataset.
(Uddin et al. 2023) [9] [Q1, Elsevier]	Different classification techniques used by the current Water Quality Index (WQI) models lead to uncertainty and inconsistency in the classification of water quality.	Total 7 WQI Models Used And Machine Learning Classifiers Used: Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor. XGBoost	WQM and RMS models demonstrated more reliability, whereas XGBoost performed better than any other model. SVM and NB fared rather poorly, whereas KNN demonstrated 100% accuracy but overfitting danger.	Study restricted to a single coastal area (Cork Harbour); potential overfitting in the KNN model; reliance on the authors' earlier improved WQI approach;	catchments.ie (EPA, Ireland)
(Hridoy et al. 2025) [10] [Q1, Elsevier]	Protecting aquatic ecosystems and controlling the risks of aquatic diseases depend on accurate water quality classification and exact WQI prediction. Existing machine learning studies lack interpretability and integrated classification + regression frameworks.	Machine Learning Classification Models + Regression Models (WQI Prediction)	LightGBM and XGBoost are the best classification models. XGBoost Regressor is the best regression model. In both classification and regression tasks, ensemble gradient boosting models perform better than conventional machine learning models.	The dataset's spatiotemporal metadata is limited. Relatively modest dataset size (170 samples initially) Absence of deep learning models	Source: Mendeley Data Repository; 4300 processed records for machine learning modeling; 170 original water samples; 14 physicochemical and biological characteristics

III. LITERATURE REVIEW

Recent research on Water Quality Index (WQI) prediction shows a strong transition from conventional weighted arithmetic models toward machine learning (ML), deep learning (DL), and hybrid ensemble approaches. Across the reviewed papers, the central problem is the limitation of traditional WQI frameworks, including rigid parameter weighting, inconsistent classification schemes, region-specific formulations, and poor generalization across environmental contexts. An exclusive summary of the literature review with most crucial seven papers are presented in Table I. A detailed literature review can be found from the access link to the Google Sheet:

Google Sheet

Most studies relied on region-specific datasets, including:

- Kaggle Water Potability dataset (3276 samples)
- Telangana groundwater dataset (Kaggle)
- Coastal water data from catchments.ie (EPA, Ireland)

Input features typically included physicochemical water quality parameters such as pH, DO, BOD, COD, EC, TDS, turbidity, sulfate, hardness, salinity, and sometimes climate

indicators. However, except catchment dataset, many datasets were small and geographically limited [9].

Most studies applied supervised ML algorithms such as Random Forest (RF), Support Vector Machine (SVM), XGBoost, KNN, Decision Tree, and Artificial Neural Networks (ANN). Results comparison is summarized in the *Key findings and contributions* column of Table I. More recent works incorporated:

- Hybrid models (e.g., ANN + XGBoost with SHAP-based initialization – Bataineh et al., 2026) [5]
- Stacking ensemble models (Mohseni et al., 2024) [7]
- Advanced deep learning architectures (LSTM, optimized DL frameworks) (Han et al., 2025; Satish et al., 2024) [8]

Deep learning models, especially LSTM, consistently outperformed traditional ML methods when temporal patterns were involved. Ensemble and hybrid frameworks generally achieved the highest predictive accuracy and robustness, as summarized in Table I.

However, several important limitations remain. Many studies rely on small, region-specific datasets, restricting broader applicability and scalability. Most models operate under single-task settings, focusing only on regression or classification rather than integrating both. Environmental, climatic, and land-use factors are often insufficiently incorporated. Hybrid and ensemble models can be computationally demanding, and few studies validate their real-time deployment. Most importantly, there is still no universal, adaptive WQI framework capable of functioning across diverse water-use contexts.

Overall, the literature demonstrates clear progress in ML/DL-based WQI prediction, with ensemble and deep learning models showing superior performance. However, current approaches remain fragmented, localized, and task-specific. There is a strong need for a scalable, multitask, and adaptive AI-driven WQI architecture that integrates diverse datasets, supports multi-class classification and regression simultaneously, and generalizes across environmental contexts.

IV. Synthesis and Conclusion

Water Quality Index (WQI) models have developed to condense complex environmental information into a single representative score. Typically, WQI models adhere to four fundamental steps executed in order to generate the WQI, which include: (a) choosing water quality parameters, (b) generating the sub-indices for those parameters, (c) assigning weights to the parameters, and ultimately, (d) using an aggregation function to compute the final index [1].

Machine learning is data-driven approach that can capture complex, nonlinear relationships among water quality parameters and predict accurate index without relying on fixed mathematical assumptions. Table I illustrated an exclusive summary on existing studies. Different machine learning and deep learning models are incorporated to predict water quality indices, refereed to column *Methods* in Table I have prediction accuracy ranged from 84% to 99% in different dataset for a single water quality index. However, these models are mostly used for estimating single purpose water quality index. On the other hand, most existing research overlook interpretablity, where the explainable artificial intelligence (XAI) helps not only to visualize the parameters responsible for a prediction, but also assist verify which parameters suppose to contribute for an accurate prediction. In real life environment, classification of water quality index play a vital role. In the previous work most of the study solely focus on either binary classification or merging different classes, emphasizing less on real classification of the water quality index models.

In machine learning based evaluation, dataset play a vital role. In order to calculate WQIs, parameters selection is crucial step as the different indexing models have different set of input parameters. In the existing exploration we found a large collection of water quality datasets. But, only few datasets have the required parameters to calculate different water quality indices. Amon all the datasets, we found Ireland Catchments dataset have the all required parameters with a vast collection data items. It provides comprehensive information on Ireland's rivers, lakes, groundwater, estuaries, and coastal waters, assessed under the EU Water Framework Directive (WFD). This standardized data collection ensures reliability and supports transferability to other regions with minimal recalibration.

Our project will focus on addressing a crucial limitation of the need of an universal water quality index, incorporating machine learning, where most current WQI models are developed for specific regions, such as Canada or Ireland, or tailored to particular water sources like rivers or wetlands. These models typically generate a single index using predefined parameter weights and fixed threshold values. As a result, they lack the flexibility required for multi-use environments where the same water body may serve drinking, irrigation, fisheries, and other purposes simultaneously. This limitation highlights an important research gap: the need for a more universal and adaptive AI-driven WQI framework capable of supporting large-scale application across diverse environmental contexts while aligning with modern sustainability requirements.

References

[1] M. G. Uddin, S. Nash, and A. I. Olbert, "A review of water quality index models and their use for assessing surface water quality," *Ecological indicators*, vol. 122, p. 107218, 2021.

[2] M. M. Syeed, M. S. Hossain, M. R. Karim, M. F. Uddin, M. Hasan, and R. H. Khan, "Surface water quality profiling using the water quality index, pollution index and statistical methods: A critical review," *Environmental and Sustainability Indicators*, vol. 18, p. 100247, 2023.

[3] Z. Han, S. Zhang, L. He *et al.*, "Predicting and investigating water quality index by robust machine learning methods," *Journal of Environmental Management*, vol. 381, p. 125156, 2025.

[4] M. G. Uddin, S. Nash, M. T. M. Diganta, A. Rahman, and A. I. Olbert, "Robust machine learning algorithms for predicting coastal water quality index," *Journal of Environmental Management*, vol. 321, p. 115923, 2022.

[5] A. Al Bataineh, B. Vamsi, and S. A. Smith, "A hybrid machine learning framework for water quality index prediction using feature-based neural network initialization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[6] M. H. Nishat, M. H. R. B. Khan, T. Ahmed, S. N. Hossain, A. Ahsan, 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. 1, p. 31, 2025.

[7] U. Mohseni, C. B. Pande, S. C. Pal, and F. Alshehri, "Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model," *Chemosphere*, vol. 352, p. 141393, 2024.

[8] N. Satish, J. Anmala, M. R. Varma, and K. Rajitha, "Performance of machine learning, artificial neural network (ann), and stacked ensemble models in predicting water quality index (wqi) from surface water quality parameters, climatic and land use data," *Process Safety and Environmental Protection*, vol. 192, pp. 177–195, 2024.

[9] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "Performance analysis of the water quality index model for predicting water state using machine learning techniques," *Process Safety and Environmental Protection*, vol. 169, pp. 808–828, 2023.

[10] M. A. A. M. Hridoy, A. I. Shawkat, C. Bordin, M. R. Acharjee, A. Masood, A. O. Baki, and M. A. Al Mamun, "Advanced machine learning models for accurate water quality classification and wqi prediction: Implications for aquatic disease risk management," *Science of the Total Environment*, vol. 1008, p. 180965, 2025.