Aquatic Ecosystems and Environmental Frontiers, Volume 4, Issue 2, June 2026 *ISSN: 3049-186X*
DOI: https://doi.org/10.70102/AEEF/V4I2/AF260219

A Machine Learning-Based Prediction Framework for Water Quality Monitoring and Pollution Risk Assessment in Aquatic Ecosystems

[1] Wei Jian, Sino Horizon Institute for Emerging Technology Studies, Beijing, China.

[2] Liu Meilin, Sino Horizon Institute for Emerging Technology Studies, Beijing, China.

Abstract: Rivers are important for aquatic biodiversity, fisheries, drinking water supply, and agriculture, but are suffering from continuous pressure due to the discharge of untreated sewage, industrial discharge, and farm runoff. Large amounts of records are generated as a result of routine monitoring, but turning these records into timely and readily understood statements of pollution risk is not easy. This study aimed to formulate a machine learning model to predict the water quality index (WQI) based on routinely measured water quality parameters and categorize each water sample according to the level of pollution. The water records of the river water were used from more than 2800 monitoring sites in India for 11 years from 2006 to 2016. The following indicators were used: temperature, pH, conductivity, biochemical oxygen demand (BOD), nitrate, fecal coliform, and total coliform. Data screening and gap filling were performed, and a weighted WQI was then computed and further divided into four risk levels (Low, Moderate, High, and Severe). Seven prediction methods were trained and tested. The gradient boosting model was selected as the primary model for the framework because it had the highest R^2 (0.998) and the lowest mean error (0.41 index units) between the measured and predicted values of WQI, and achieved 98.3% accuracy rates in risk classification and 97.7% accuracy when tested on the subsequent non-training years. 90% of samples were in the Moderate, High, or Severe categories, with BOD and fecal coliform contributing almost 78% of the water quality variation explained. The omission of the total coliform test resulted in a minimal decrease in accuracy, but the omission of both bacterial tests resulted in a 76.1% decrease in accuracy. The framework provides a practical and cost-effective method of early warning and prioritization of monitoring effort in aquatic ecosystems, identifying hotspots.

Keywords: Water Quality Index; Pollution Risk; Machine Learning; River Monitoring; Aquatic Ecosystems; Gradient Boosting.

(Submitted: March 27, 2026; Revised: April 17, 2026; Accepted: June 02, 2026; Published: June 30, 2026)

I. Introduction

Background of Aquatic Water Quality Monitoring

Freshwater systems enhance biodiversity and supply vital goods and services to people such as drinking water, irrigation, industry and fisheries. River water quality is a result of the combined effect of natural processes and human activities in catchments (Uddin et al., 2023; Lokman et al., 2025). Long-term monitoring programs in India include parameters like temperature, pH, conductivity, biochemical oxygen demand, nitrate and bacterial counts. Rivers are still, however, vulnerable to domestic sewage, industrial effluents and agricultural runoff, all of which may lead to an increase in organic load, nutrient enrichment, dissolved minerals, and bacteria. These pressures can result in less oxygen, changing aquatic food chains, and higher ecological and human-health risks. An integrated approach to assessment is needed since the individual parameters are not sufficient to determine the overall river condition.

Water Pollution and Predictive Water Quality Assessment

Domestic sewage, industrial effluents, and agricultural runoff discharge organic matter, nutrients, dissolved salts, and microbial contaminants into rivers, causing a deterioration in water quality. All of these pressures can lead to decreased oxygen levels, changes to aquatic food webs, and increased ecological and human health risks. Several pollutants are frequently present at the same time, and individual parameters may not reflect the general condition of the water. Water quality indices were used to score the water quality by combining multiple measured water parameters; however, require full lab analysis and parameter-

specific ratings and weights (Shamsuddin et al., 2022). In addition to these, machine learning provides an alternative method that learns complex and non-linear relationships from the historical data of monitoring and subsequently is able to provide quick estimation of water quality, classification of pollution risks, and identification of those parameters that best describe the water quality of the river.

Research Problem and Motivation

Although there is increasing interest in data-driven water quality assessment, there are still some practical challenges. A lot of studies have been conducted for one river or a few stations, and thus the results are of only limited value. Prediction outcomes are frequently only presented as statistics rather than in terms of risk levels that managers can respond to. Very few research studies address the question of whether a model developed from previous years is reliable for future years or if there are expensive lab tests that could be cut back with minimal loss of accuracy. The opportunities to address these issues in a single, application-oriented framework arise with large multi-region records, such as those for Indian rivers.

Objectives of the Study

1. To build a machine learning framework to assess river water quality and identify the risks of pollution.
2. To describe river water quality on a large national dataset.
3. To determine a water quality index and provide a classification of water quality samples according to pollution risk.
4. To compare 7 machine learning and statistical prediction methods.
5. To test the reliability of the predicted water quality and possible pollution risks. To identify the most influential water quality indicators that add to water pollution risk.

Section 1 presents background information on aquatic water quality monitoring, pollution issues, the research problem, and study objectives. The conventional WQI methods, machine learning-based prediction, water quality indicators, and the research gap are discussed in Section 2. The methodology, such as data collection, parameter selection, pre-processing, prediction framework, risk classification, and evaluation measures used to develop the methodology, is presented in Section 3. Section 4 presents the results regarding water quality attributes, WQI condition, model performance, pollution risk classification, and major indicators. The findings, the implications of the study, and the limitations are discussed in Section 5, and future research directions are discussed in Section 6, which concludes the study.

Conventional Approaches for Water Quality Assessment

The surface-water evaluation is done by comparing each parameter to a regulatory standard or by using water quality index models. The typical methods used in WQI generally include parameter selection, scoring of sub-indices, and weighting and aggregation of all of them to a final WQI value (Uddin et al., 2021). While WQI methods make their results easier to interpret, the selection of parameters, weights, and class boundaries can affect the results, and generally, results will also depend on the completeness of the laboratory measurements (Uddin et al., 2021; Akhtar et al., 2021).

Machine Learning-Based Water Quality Prediction

The increasing use of machine learning, previous studies have predicted water quality indices and classified the pollution conditions. Studies in Algeria, Vietnam, and Pakistan demonstrated that water quality index values can be accurately estimated by using a variety of methods, including decision trees, support vector machines, random forests, and boosting models, and water quality classes can be assigned (Kouadri et al., 2021; Khoi et al., 2022; Ahmed et al., 2021). More recent reviews also suggest uses of machine learning in surface water, groundwater, drinking water and wastewater, as well as marine environments, and point to issues with data quality, model uncertainty and generalisability (Zhu et al., 2022;

Yan et al., 2024). Another classification analysis also shows that feature selection and ensemble learning have been shown to increase the reliability of the prediction (Nasir et al., 2022; Abuzir & Abuzir, 2022). However, many of the studies conducted so far are focused on individual rivers, only one station, or on region-specific data, so these studies have limited operational applicability across more rivers (Kouadri et al., 2021; Ahmed et al., 2021; Yan et al., 2024).

Water Quality Indicators and Pollution Risk Assessment

The choice of water quality parameters is important to water quality prediction, since each is a measure of a different source of pollution or ecological pressure. Parameters like oxygen demand, bacterial counts, nutrients, pH, and conductivity, which are related to organic, microbial, nutrient, and salinity pollution, have been used in previous studies. Machine learning has been demonstrated to correctly classify water quality within a river basin and/or coastal systems with a smaller subset of informative parameters, indicating that the correct choice of indicators can help with an accurate assessment with fewer monitoring efforts (Malek et al., 2022).

Research Gap

There are some lacunae in the literature. First, uncertainty in index models and their predictions is still not well addressed, and recent work has advocated for models that estimate and communicate this uncertainty. Secondly, the majority of prediction studies have been restricted to a single river or area, and there is little evidence of large multi-region records. Third, model performance is often not evaluated during other time periods that were not included in the model training, but in order to be used in operation, such testing should also be done. Only a small number of studies convert the predicted index values into meaningful levels of pollution risk and investigate what expensive laboratory tests might be dropped without compromising the reliability of the index value. These gaps were addressed in the present study.

II. Methodology

Research Design

The study was conducted in two stages, using a quantitative, data-driven design. The first stage involved the establishment of a data foundation, which included the collection of monitoring records, screening and gap-filling of data, the selection of parameters, and the calculation of a water quality index with an associated risk level. The second stage developed and evaluated the prediction framework, covering training and comparison of prediction approaches, classification of risk levels, and ranking of pollution indicators. Results from both stages were used to inform management decision-making.

Study Area and Data Collection

The study area included rivers in India, which has a wide range of climates, land use, and population density. The data consisted of records from the national monitoring network for river water quality that were publicly released from 2006 to 2016, and were open-mixed into a collection of 8,885 records. Records labeled as groundwater (1,909) were omitted from the analysis, and only surface aquatic ecosystems were considered, resulting in 6,976 river water records. They were collected from over 2,800 monitoring locations across 28 states and union territories, most of which are from Maharashtra, Odisha, Karnataka, Andhra Pradesh, Kerala, Uttar Pradesh, and Tamil Nadu. The number of records per year ranged from 368 in 2006 to 1,264 in 2011.

Selection of Water Quality Parameters

Seven parameters were chosen as the ones most frequently measured over the network and were also representative of the main types of river pollution shown in Table 1. Oxygen solubility and biological activity are affected by temperature. pH is a measure of acidic and alkaline conditions that impacts aquatic organisms. Conductivity is a measure of the dissolved salts caused by industry and agriculture. BOD is

used to indicate organic pollution and oxygen depletion potential. Nutrient enrichment is expressed in the form of nitrate. The presence of fecal coliforms and total coliforms is a sign of sewage contamination and a health hazard. Each parameter was given a weight based on its ecological and health significance, with particular emphasis on organic and fecal pollution.

Table 1: Water quality parameters, their relevance and assigned weights.

Parameter	Unit	Pollution aspect represented	Weight
Temperature	°C	Thermal Condition, Oxygen Solubility	0.06
pH	–	Acidity or Alkalinity	0.16
Conductivity	µmhos/cm	Dissolved Salts, Effluent Input	0.10
BOD	mg/L	Organic Load, Oxygen Depletion	0.24
Nitrate	mg/L	Nutrient Enrichment	0.10
Faecal coliform	MPN/100 mL	Sewage Contamination	0.22
Total coliform	MPN/100 mL	General Bacterial Contamination	0.12

Data Preparation and Analysis

There were four steps in the data preparation process. All values were first checked for physical plausibility; pH values exceeding the range of possibilities and temperatures exceeding 45 °C were considered recording errors and coded as missing values. Secondly, records that were missing more than 2 of the 7 parameters were excluded, which would only affect 5 records and leave 6,971 samples for analysis. Third, the remaining 63 missing values were replaced by the median value for the same state, thus maintaining regional conditions. Fourth, the variability of bacterial counts, conductivity, BOD, and nitrate was several orders of magnitude, and thus these values were reduced to a compressed scale in order not to be overwhelmed by a few extreme samples in the modeling process. Descriptive statistics, comparison with common water quality limits, and rank correlations were used to describe the data.

Each of these was then converted to a water quality index. All parameters were scaled into a range of 0 to 100, with 100 being excellent and 0 being very poor condition, according to rating bands determined by commonly used water quality criteria for designated uses in India. The ratings were aggregated with the weights shown in Table 1 to arrive at a rating of 0-100 points.

Development of the Water Quality Prediction Framework

The framework was broken down into four layers as illustrated in figure 1. The seven parameters measured are passed to the input layer. The preparation layer screens, fills, and rescales the data and divides it into a training portion (70%, 4,879 samples) and an independent testing portion (30%, 2,092 samples). There are six comparison models and the core model in the prediction layer. The output layer outputs the predicted index as well as the risk level associated with the index and the indicators that affect the index most importantly.

The seven prediction approaches ranged from simple ones, such as multiple linear regression, which assumes that there is a linear relationship between the variables, to more flexible ones, such as k-nearest neighbors, which predicts based on the most similar past samples; support vector regression; an artificial neural network; a single decision tree; random forest, which predicts by averaging many decision trees; and gradient boosting, which constructs a sequence of small decision trees, where each one corrects the errors of the previous trees. For each approach, the training part was trained and tested using five-fold cross-checking, which involved dividing the training data into five subsets and using each one of them once for testing. It was decided to use gradient boosting as the core of the framework, as it performed well. The core model was also trained using only data from the years 2006 through 2013 (5,033 samples) and evaluated on data from 2014 through 2016 (1,938 samples), to assess the reliability over time.

Pollution Risk Classification

The values of the indices were divided into 4 pollution risk levels in Table 2. Low: Water quality is good, and adverse effects on aquatic life are minimal. The Moderate level is an indication of pollution that is noticeable and needs regular attention. The High level represents significant pollution that may affect aquatic communities and/or limit water use. The Severe level indicates highly polluted water which must be dealt with immediately. The index values were predicted by each approach and were then mapped to risk categories based on the same boundaries, so the results can be directly compared to the risk levels based on measured values.

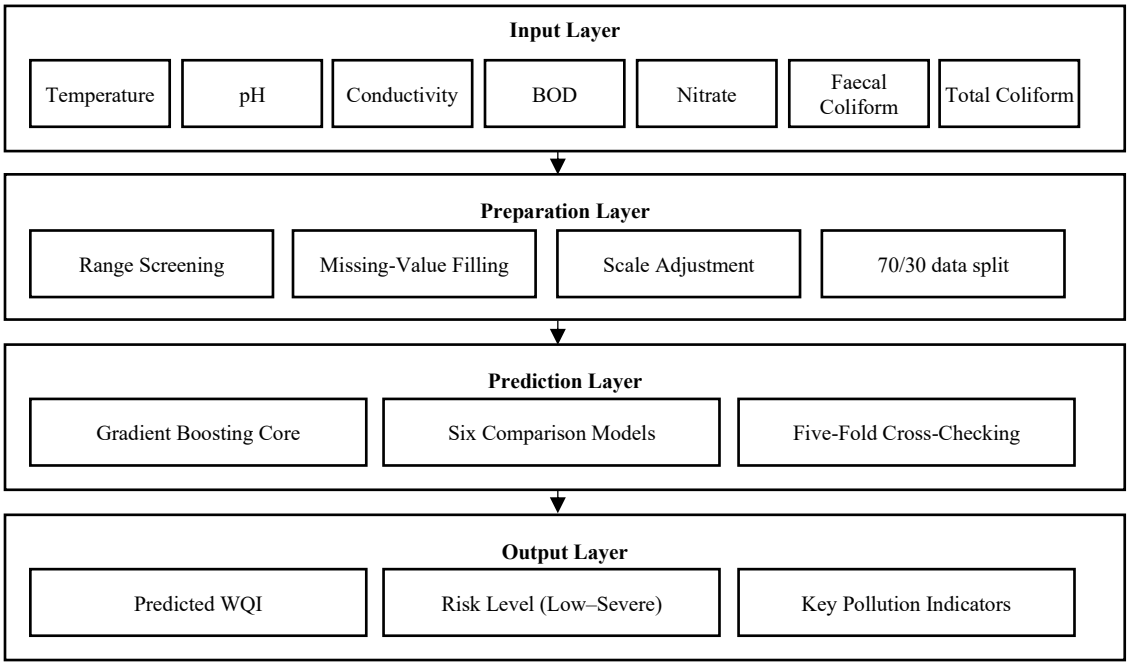


Figure 1: Layered architecture of the proposed water quality prediction framework

Performance Evaluation Measures

The accuracy of the prediction was tested in three different ways. The coefficient of determination (R^2) will indicate the percentage of the variation in the observed index that is explained by the predictions, and a value of 1 will indicate that the predictions are in perfect agreement. The root mean square error (RMSE) and mean absolute error (MAE) display the average size of the prediction errors in index units (smaller is better). The overall accuracy, the precision, the recall, the balanced F1 score, and the kappa statistic (the agreement beyond what would be expected by chance) were used to assess risk classification. The importance of each indicator was assessed by permuting each one separately and measuring the drop in prediction accuracy; the higher the drop, the more important the indicator.

Table 2: Pollution Risk Levels and their Interpretation

Risk level	WQI range	Interpretation for Aquatic Ecosystems
Low	80–100	Good Condition; Suitable for Most Aquatic Life and Uses
Moderate	60–79.9	Noticeable Pollution; Regular Monitoring Needed
High	40–59.9	Substantial Pollution; Ecological Stress Likely
Severe	Below 40	Heavily Polluted; Urgent Intervention Required

III. Results

Characteristics of Water Quality Parameters

The water quality parameters are classified according to their characteristic features. The measured parameters have been summarised in Table 3. There was not much variation in temperature and pH. The median values were 25.8 °C and 7.7, respectively. Conductivity, BOD, nitrate and bacterial counts, on the other hand, were very skewed, with mean values much greater than median values due to a limited number of very polluted samples. The median counts of fecal and total coliforms were 179 and 863 MPN/100 mL, respectively, with maximum counts (290 million MPN/100 mL and 1.3 billion MPN/100 mL) showing a high level of sewage contamination at some stations. The widespread distribution of these parameters is shown in figure 2.

Table 3: Descriptive Statistics of River Water Quality Parameters (N = 6,971)

Parameter	Mean	Median	Std. deviation	Minimum	Maximum
Temperature (°C)	24.37	25.80	4.82	2.00	39.00
pH	7.60	7.70	0.54	2.60	13.70
Conductivity (µmhos/cm)	807.42	296.00	2,785.90	0.20	47,156.00
BOD (mg/L)	4.30	2.00	12.39	0.00	534.50
Nitrate (mg/L)	1.20	0.49	3.65	0.00	209.60
Fecal Coliform (MPN/100 mL)	174,514.83	179.00	5,860,402.01	0.00	294,579,250
Total Coliform (MPN/100 mL)	602,633.13	863.00	19,426,097.68	0.00	1,339,515,250

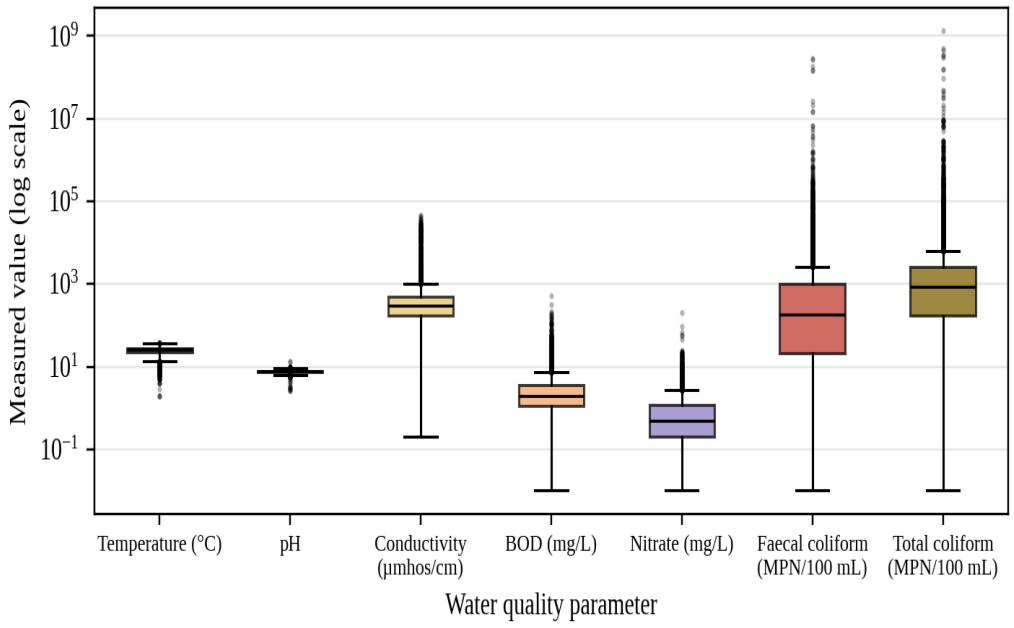


Figure 2: Distribution of Water Quality Parameters Across River Monitoring Samples.

BOD was found to be higher than the common water quality limits in 31.1% of the samples; total coliform was higher than 5,000 MPN/100 mL in 18.2% of the samples, fecal coliform higher than 2,500MPN/100 mL in 15.3% of the samples, and conductivity was higher than 750 µmhos/cm in 13.6% of the samples shown in figure 3. In comparison, pH values were found outside the pH 6.5 to 8.5 range in just 3.1% of samples and nitrate concentrations above 10 mg/L in just 1.2% of samples. Therefore, the most common pollution problems were found to be organic and fecal. Rank correlations revealed that the index

was most highly correlated with BOD (−0.77), followed by total coliform (−0.60), fecal coliform (−0.59), and conductivity (−0.51). Fecal coliform and total coliform were closely correlated (0.84), and BOD was moderately correlated with conductivity (0.45), indicating that organic and salt pollution frequently co-occur.

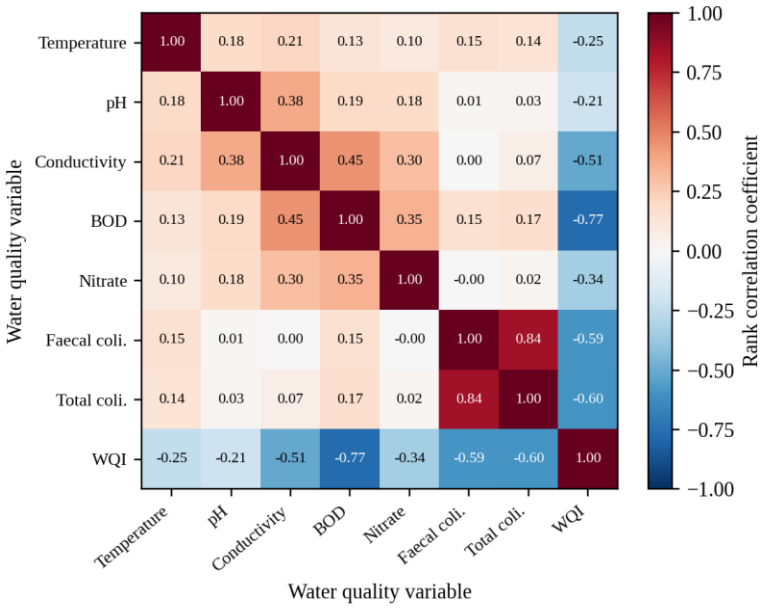


Figure 3: Rank Correlation Matrix of Water Quality Parameters and the Water Quality Index

Water Quality Status of the Study Area

The average was 73.4 (s.d. 13.5) and the typical river sample was in the Moderate risk level. Annual mean values figure (4a) declined from 77.2 in 2006 to between 70.6 and 73.2 during 2007 to 2014, then rose to about 80.6 and 80.9 in 2015 and 2016. The proportion of Low-risk samples also rose, to approximately 60% in the last two years figure (4b). This late enhancement is a part of this improvement, which may be due to changes in the stations sampled and thus should be interpreted with caution.

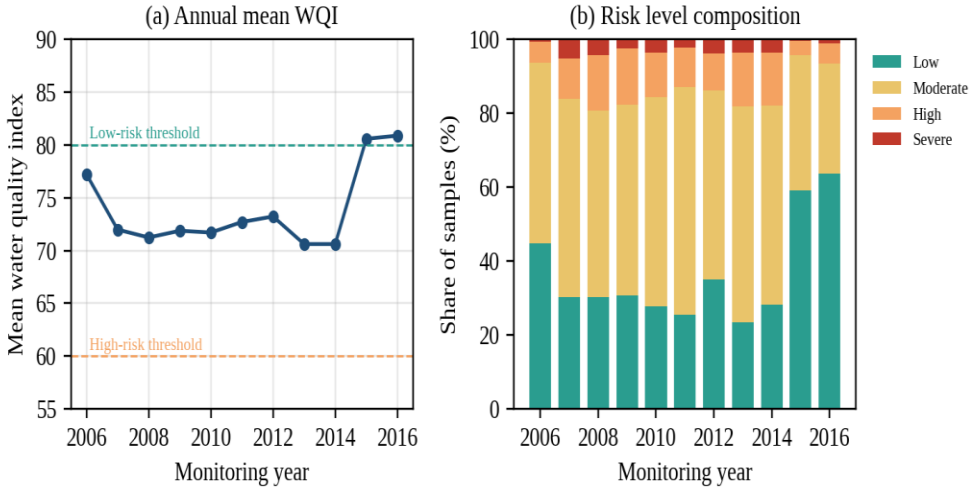


Figure 4: Temporal Pattern of (a) Annual Mean Water Quality Index and (b) Risk Level Composition, 2006–2016

Clear regional differences were observed in figure 5. West Bengal (58.4) and Uttar Pradesh (58.5) had the lowest mean index values, in the High-risk range, indicating high pressure on the rivers in the densely populated Ganga plain. The relatively poor conditions were also observed in Punjab (66.5) and Maharashtra (69.8). The most favorable condition was observed in the states of Himachal Pradesh (88.9) and Madhya Pradesh (80.4), with the lowest mean value of 75.0 in the State of Tamil Nadu.

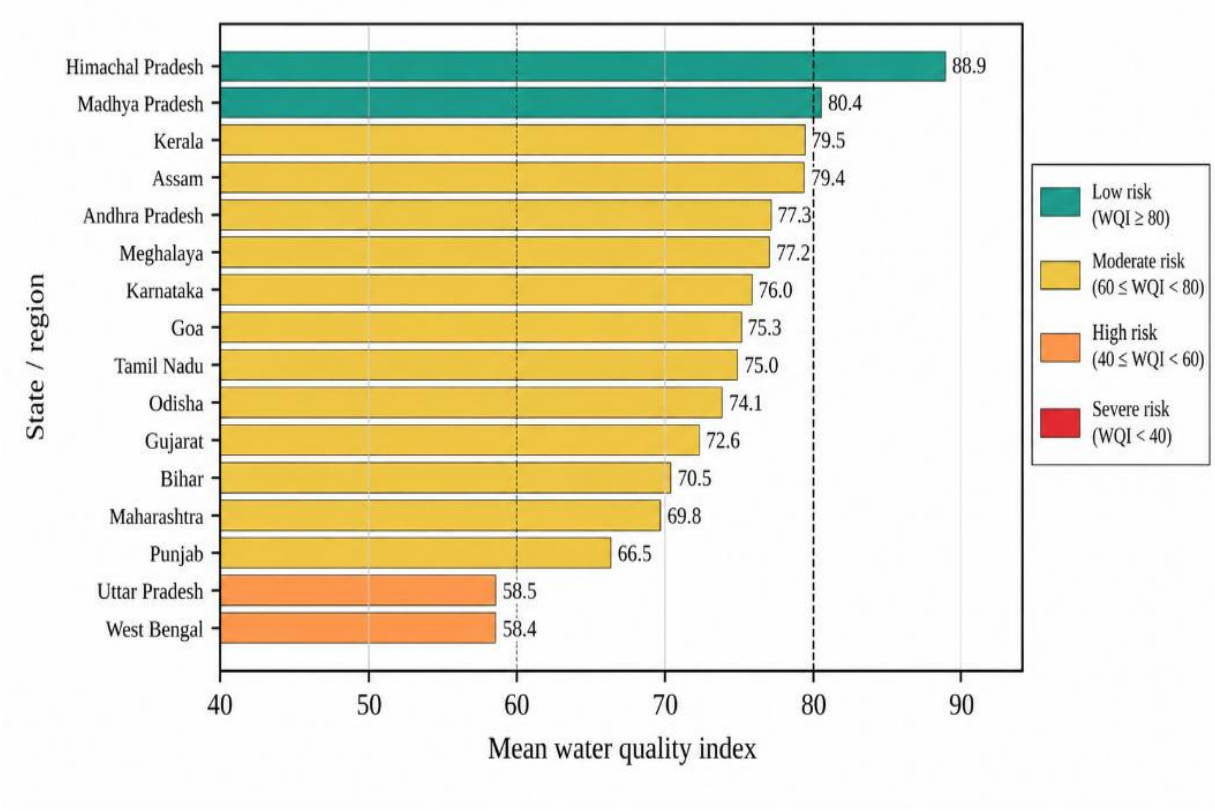


Figure 5: Mean Water Quality Index by State for States with at Least 100 River Samples

Prediction Performance of the Proposed Framework

The results from the gradient boosting part of the framework showed a good fit for the observed index. On the independent test data, it had an R^2 of 0.998, RMSE of 0.65 index units, and MAE of 0.41 index units, and five-fold cross-checking resulted in a mean R^2 of 0.997 with very little variance. The deviation between predicted and observed values was small and close to the line of perfect agreement over the entire index range figure 6(a), with very few prediction errors being larger than ±2 index points Figure 6(b). The model's performance did not deteriorate when only trained with the data from 2006–2013 and tested with 2014–2016 data, with R^2 of 0.996 and RMSE of 0.83.

Comparison of Prediction Approaches

The seven approaches are compared with each other in table 4 and figure 7(a) and figure 7(b). The differences in the approaches to explaining the variation in the index occurred. Multiple linear regression had the lowest R^2 (0.892) and RMSE (4.48), which is not applicable to the stepwise changes of the parameters and the non-linear relationship between the parameters and water quality. The results were superior when using K-nearest neighbours, the neural network or support vector regression and ranged from R^2 0.952 to 0.973. The best performance was obtained using tree methods: single decision tree obtained 0.975, random forest 0.987 and gradient boosting 0.998. This ranking was the same in both test data and cross-checking and was not due to chance.

Table 4: Performance of the Prediction Approaches on Independent Test Data.

Approach	R^2	RMSE	MAE	Cross-Check R^2	Risk Accuracy (%)
Multiple Linear Regression	0.892	4.48	3.24	0.894	86.2
K-nearest Neighbours	0.952	2.99	2.24	0.950	90.5
Support Vector Regression	0.973	2.26	1.76	0.971	92.3
Artificial Neural Network	0.968	2.44	1.91	0.967	91.5
Decision Tree	0.975	2.13	1.25	0.975	93.6
Random Forest	0.987	1.57	0.93	0.986	95.8
Gradient Boosting (Proposed)	0.998	0.65	0.41	0.997	98.3

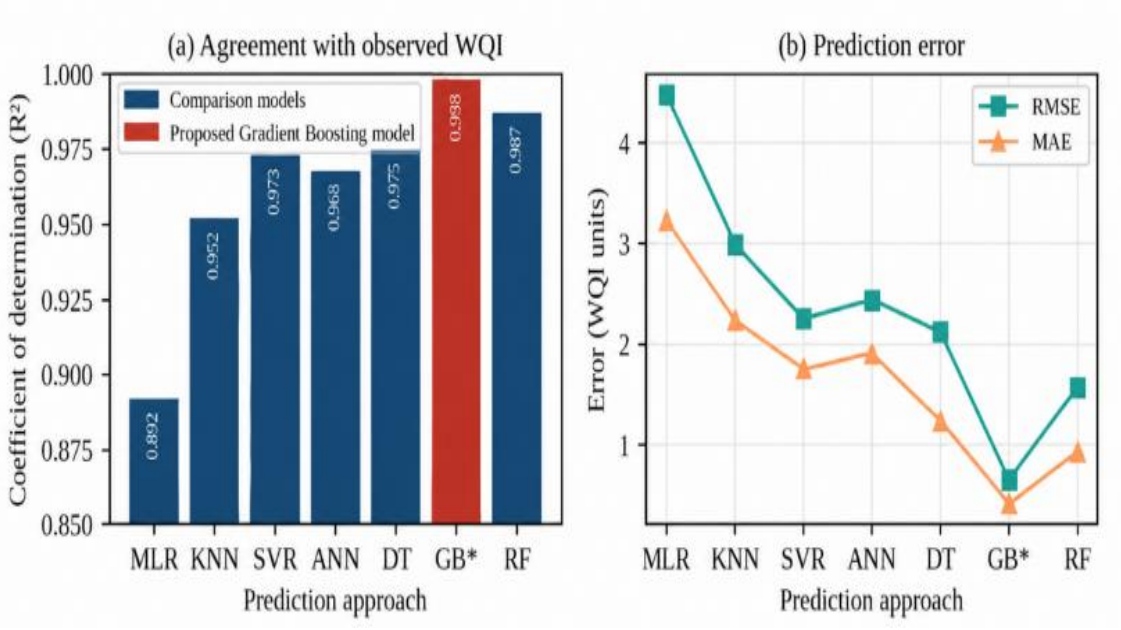


Figure 6: Comparison of Prediction Approaches: (a) Coefficient of Determination and (b) Prediction Error. GB* Denotes the Proposed Gradient Boosting Core

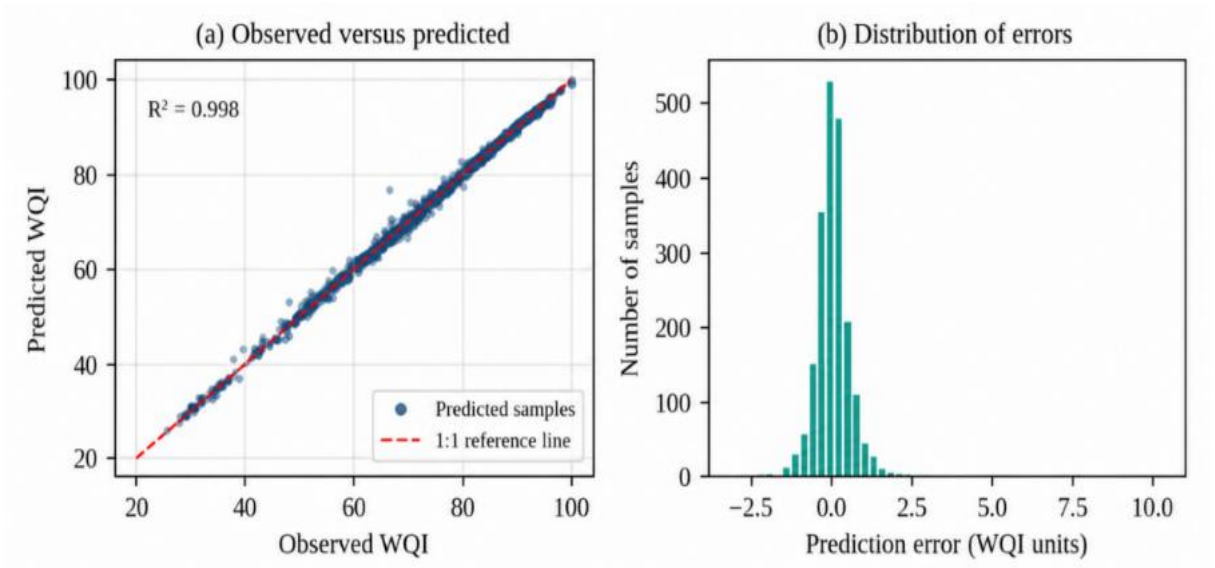


Figure 7: Prediction Performance of the Proposed Framework: (a) Observed Versus Predicted Index and (b) Distribution Of Prediction Errors

Classification of Pollution Risk Levels

The 6,971 river samples, 34.5% were classified as Low risk, 51.4% as Moderate, 11.3% as High and 2.9% as Severe figure (8b). So, almost two thirds of samples were moderately polluted, and approximately 1 in 7 samples were High or Severe pollution. The framework correctly classified 98.3% of test samples as the index values were converted to risk levels, and had macro-averaged precision of 98.7%, recall of 97.7%, F1 of 98.2% and kappa value of 0.971, which shows almost complete agreement. The recognition rates were: 97.9% for the Low samples, 98.8% for the Moderate samples, 97.4% for the High samples and 96.7% for the Severe samples. There were no misclassifications to the High or Severe levels and no Low-risk sample misclassified as High or Severe figure (8a). Risk accuracy on the later year test was also high at 97.7%.

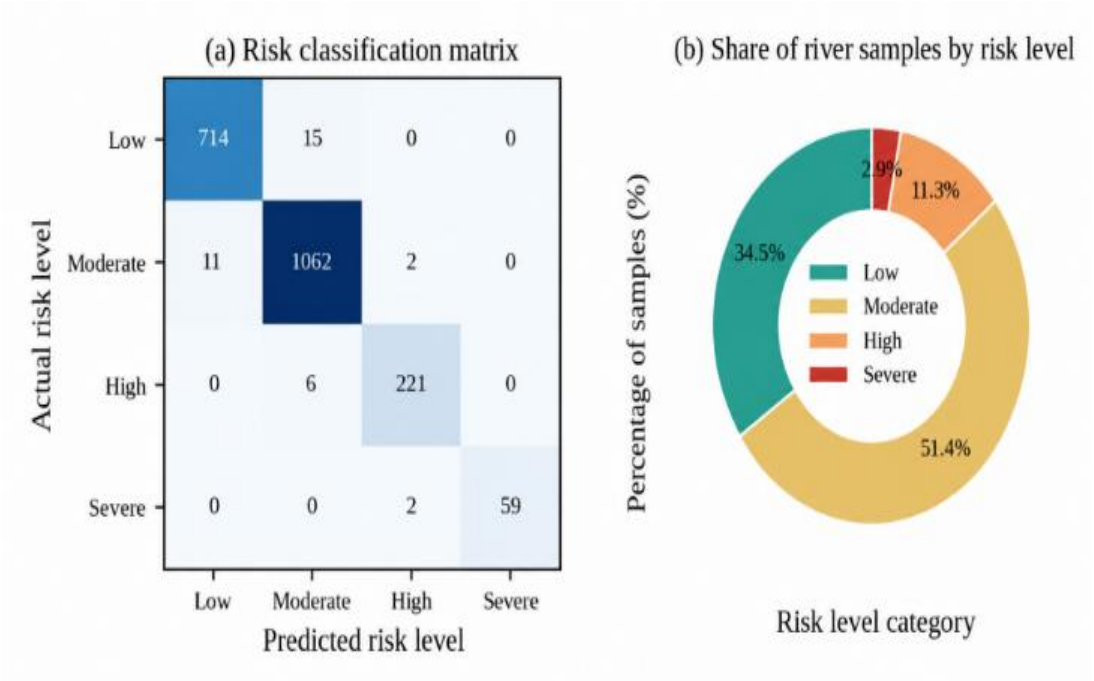


Figure 8: Pollution Risk Assessment: (a) Actual Versus Predicted Risk Levels on Test Data and (b) Share of River Samples in Each Risk Level

Identification of Major Pollution Indicators

Indicator importance figure (9a) revealed that BOD had the greatest influence with 47.2% of the predictive influence, followed by faecal coliform (30.5%). Moderate influence was observed for total coliform (8.7%) and conductivity (8.3%) whereas pH (3.9%), nitrate (1.1%) and temperature (0.4%) had little influence. The combined effects of organic and faecal pollution thus accounted for nearly 78% of the variation in river water quality predicted.

Reduced-indicator scenarios figure (9b) demonstrated what monitoring might look like if it was reduced. The removal of total coliform, which essentially was the same as faecal coliform, resulted in a slight decrease in the risk accuracy from 98.3% to 96.5%, while the R^2 value was still at 0.989. However, its removal decreased R^2 to 0.745 and risk accuracy to 76.1%, proving that at least one bacterial test is necessary to be able to assess the risk reliably.

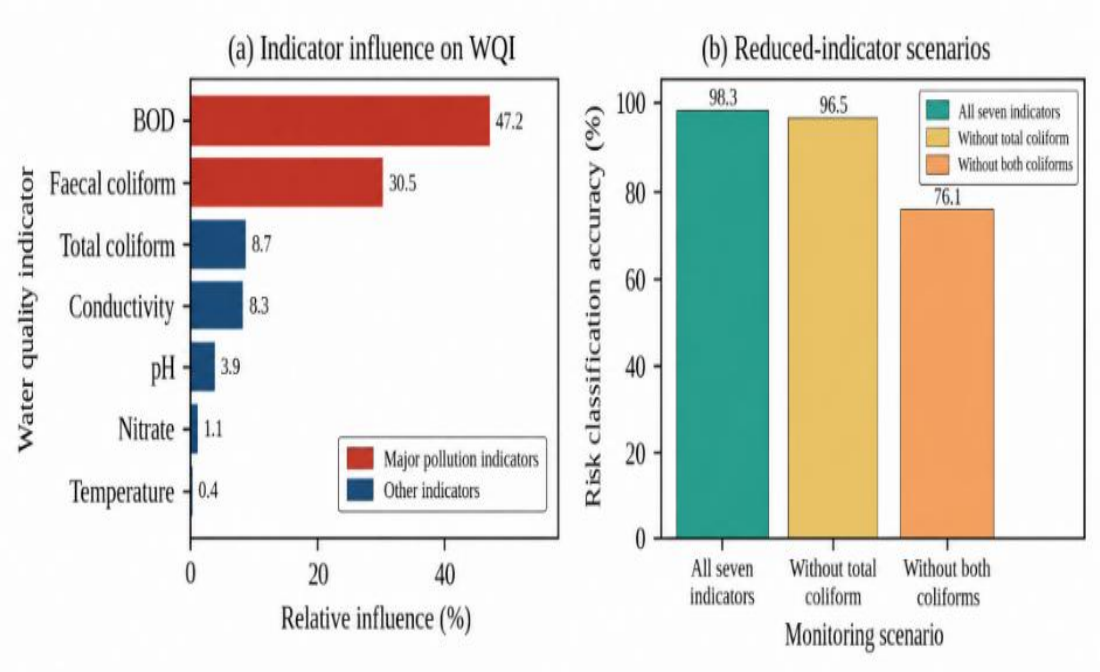


Figure 9: (a) Relative Influence of Water Quality Indicators and (b) Risk Classification Accuracy Under Reduced-Indicator Monitoring Scenarios

IV. Discussion

Interpretation of the Main Findings

The findings confirm that the routinely observed parameters include sufficient information to provide highly reliable estimation of overall water quality and pollution risk of the river. The near-perfect functioning of the gradient boosting core is conditioned by its ability to reflect the stepwise nature of water quality dynamics, which emerges when every single parameter reaches critical thresholds. High accuracy of predictions on the data of later years indicates that the dependencies established based on historical information were stable and could be used operationally. Importantly, the fact that all the errors in the classification occur between adjacent risk levels has practical significance as well – it proves that the framework never produces misleading false alarms or fails to identify seriously polluted sections of the river.

High accuracy must be interpreted in the right context, however. Since the index is constructed using the same set of parameters, the framework primarily focuses on the task of reproducing and accelerating the index-based approach to water quality assessment.

Implications for Aquatic Ecosystem Monitoring

The framework can be used to facilitate aquatic ecosystem monitoring in a number of ways. To be able to give advance warning of risk levels if organic or bacterial pollution increases, the risks are quickly estimated so that managers have time to investigate the source of the pollution and prevent ecological damage. State-level and station-level summaries can help to detect pollution hotspots and help target the location of additional monitoring stations or biological surveys. Findings that total coliform is only adding additional information over faecal coliform suggest that monitoring effort could be focused on additional parameters that are not currently monitored, such as dissolved oxygen, heavy metals or other biological indicators, that would provide a more complete indicator of ecosystem health.

Practical Implications for Pollution Management

The findings, for pollution control, clearly indicate the need for attention to sewage control. Wastewater collection and treatment in High and Severe risk areas, especially in the densely populated plains, is likely to achieve the greatest improvement in river condition. The framework can also be used to monitor the impact of such interventions over time, by comparing the risk levels that were predicted before and after interventions. It can be easily embedded in current monitoring processes, current monitoring dashboards or sensor-based systems in a minimum additional investment, as it makes use of only a small number of common parameters, and its output can be expressed easily to local authorities and to the public.

Study Limitations

There are a number of caveats to be pointed out. The data set did not include dissolved oxygen, which is an important measure of the health of aquatic ecosystems. The weights and bands in the index are derived from criteria widely accepted and used in the industry, but are subject to expert judgement, and alternative combinations of criteria could result in differing levels of risk. The number of samples and stations differed among years and states, and this could impact interpretation of temporal and regional patterns. This data was last updated in 2016, and the situation may vary from this time. Finally, the framework restores an index using the same inputs; future work will see how well it can predict future conditions based on previous measurements, and how well it can include land use, rainfall and flow of the rivers.

V. Conclusion

The study machine learning-based prediction framework has been developed which uses 6,971 data samples taken from water quality monitoring points in river ecosystems across India between 2006 and 2016 and with more than two and half thousand points of water quality monitoring. Seven frequently measured parameters were screened, weighted, combined into a WQI and classified into four pollution risk levels. Seven prediction methods were explored and gradient boosting was used as the basis of the framework. The framework had an R2 of 0.998 for the water quality index and a correct risk class for 98.3% of the test samples with no errors of the water quality index in later years that did not form part of the training process. Almost two-thirds of the river samples were of moderate to high pollution with the highest pollution levels recorded in the highly populated Ganga plain. BOD and faecal coliform were found to be the most prevalent pollution indicators, a clear indication of the pollution originating from sewage. There is the potential to eliminate total coliform without a significant loss of accuracy, but at least one indicator bacteria should be kept. The environmental value of the framework is that it transforms the everyday monitoring data into timely and consistent, transparent statements of pollution risk to facilitate early warning, identification of hotspot areas and prioritisation of wastewater management. It is easy to use because of its minimal data requirements and simple outputs. Limitations are that there is no dissolved oxygen, estimation of weights is done by experts and spatial and temporal sampling is not uniform. Improved future research should include the monitoring of dissolved oxygen, nutrients, metals, and biological indicators, and should be expanded to include water quality forecasts for future conditions and tested with more recent and continuous monitoring data from sensors.

References

[1] Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. *Ecological indicators*, *122*, 107218. https://doi.org/10.1016/j.ecolind.2020.107218

[2] Akhtar, N., Ishak, M. I. S., Ahmad, M. I., Umar, K., Md Yusuff, M. S., Anees, M. T., ... & Ali Almanasir, Y. K. (2021). Modification of the water quality index (WQI) process for simple calculation using the multi-criteria decision-making (MCDM) method: a review. *Water*, *13*(7), 905. https://doi.org/10.3390/w13070905

[3] Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science*, *11*(12), 190. https://doi.org/10.1007/s13201-021-01528-9

[4] Khoi, D. N., Quan, N. T., Linh, D. Q., Nhi, P. T. T., & Thuy, N. T. D. (2022). Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. *Water*, *14*(10), 1552. https://doi.org/10.3390/w14101552

[5] Ahmed, M., Mumtaz, R., & Hassan Zaidi, S. M. (2021). Analysis of water quality indices and machine learning techniques for rating water pollution: a case study of Rawal Dam, Pakistan. *Water Supply*, *21*(6), 3225-3250. https://doi.org/10.2166/ws.2021.082

[6] Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, *169*, 808-828. https://doi.org/10.1016/j.psep.2022.11.073

[7] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., ... & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, *1*(2), 107-116. https://doi.org/10.1016/j.eehl.2022.06.001

[8] Yan, X., Zhang, T., Du, W., Meng, Q., Xu, X., & Zhao, X. (2024). A comprehensive review of machine learning for water quality prediction over the past five years. *Journal of Marine Science and Engineering*, *12*(1), 159. https://doi.org/10.3390/jmse12010159

[9] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, 102920. https://doi.org/10.1016/j.jwpe.2022.102920

[10] Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. *Water Quality Research Journal*, *57*(3), 152-164. https://doi.org/10.2166/wqrj.2022.004

[11] Lokman, A., Ismail, W. Z. W., & Aziz, N. A. A. (2025). A review of water quality forecasting and classification using machine learning models and statistical analysis. *Water*, *17*(15), 2243. https://doi.org/10.3390/w17152243

[12] Malek, N. H. A., Wan Yaacob, W. F., Md Nasir, S. A., & Shaadan, N. (2022). Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. *Water*, *14*(7), 1067. https://doi.org/10.3390/w14071067

[13] Shamsuddin, I. I. S., Othman, Z., & Sani, N. S. (2022). Water quality index classification based on machine learning: A case from the Langat River Basin model. *Water*, *14*(19), 2939. https://doi.org/10.3390/w14192939