International Journal for Research in Climate Change and Earth Science

Published By:
Green Publications Services Private LTD.

Volume 01 Issue 01 March 2026
DOI: 10.53555/cces.v1i1.2577

ASSESSMENT OF WATER QUALITY USING PHYSICOCHEMICAL PARAMETERS AND MACHINE LEARNING TECHNIQUES

Dr Anamika Singh

Assistant Professor, Department of Geography, Vasanta College of Women, KFI, Rajghat, Varanasi, dr.anamikasghvew@gmail.com, https://orcid.org/0009-0008-8775-971X

Corresponding Author (Dr Anamika Singh)

Article History:
Article Type: Research

Received Date: 08/02/2026 **Revised Date: 07/03/1026**
Acceptance Date: 20/03/2026 **Published Date: 27/03/2026**

Abstract

Water quality degradation represents a big threat to environmental sustainability, the stability of ecosystems and public health. Water quality parameters, such as selected physicochemical parameters and machine-learning techniques, were used for assessing water quality in this study. The temperature, pH, electrical conductivity, total dissolved solids, chloride, chemical oxygen demand, and biological oxygen demand were measured. Descriptive statistics, correlation and principal component analysis were first used to analyse data distribution, relations and major pollution patterns. Then, repeated stratified cross-validation was used to compare multiple machine learning algorithms: logistic regression, decision tree, random forest, support vector machine, k-nearest neighbours, and gradient boosting. values found to be more than the selected reference limits for all observed samples. Electrical conductivity was strongly correlated with total dissolved solids (TDS), and the total variance in the data was primarily explained by dissolved solids, ionic concentration and chemical pollution as determined by principal component analysis. Gradient boosting gave the highest predictive accuracy of 0.880, F1 score of 0.837, and the ROC AUC of 0.926, followed by random forest. Total dissolved solids, electrical conductivity, and chloride were determined as the most important variables by feature-importance analysis for the estimation of the relative pollution burden.

Keywords: water quality; physicochemical parameters; machine learning; gradient boosting; random forest; pollution assessment

1. Introduction

Water is essential for human well-being, ecosystem health, food security, industrial development and socio-economic development. But fast urbanisation, intensifying agriculture, population growth, industrial effluents and climate change are increasingly polluting fresh water sources globally. The degradation of water quality is a risk to drinking water quality, aquatic biodiversity, public health and the sustainability of water-dependent communities. Global assessments highlight that water security is not just about water availability, but also about water being chemically, physically and biologically appropriate for its intended purpose (Connor, 2024). The continued degradation of freshwater systems is thus a significant environmental issue, especially in areas with limited capacity for monitoring infrastructure, treatment facilities, and regulatory oversight (du Plessis, 2023). In addition, the worldwide reporting of SDG 6.3.2 shows that many countries are still lacking in high-quality observations to undertake regular monitoring of ambient water conditions (Warner, 2021).

Good water-quality assessment is usually based on physico-chemical parameters that are indicative of the condition, composition and pollution load of water. These include temperature, pH, electrical conductivity, total dissolved solids, chloride, chemical oxygen demand and biological oxygen demand. These parameters are important factors for determining acidity or alkalinity, ionic concentration, dissolved pollutants, organic contamination, and oxygen-depleting substances. When evaluating water contamination, it is typically multidimensional; thus, their combined interpretation is more informative than evaluating a single variable. Water Quality Index approaches have therefore been developed to convert many measurements into a simplified score or class to facilitate communication, planning and regulatory decision-making. In the region, it has been demonstrated that water quality indices can be developed that represent water quality at monitoring sites in an efficient and sustainable way, provided they are well designed (Derdour et al., 2022). When many physicochemical measurements are available, the use of machine-learning-based index prediction has also been shown to be useful in terms of efficiency in the assessment (Hassan et al., 2021).

While traditional assessment practices continue to be critical, they can be time-consuming, expensive, and challenging to conduct on a network of many monitoring sites. They may also be restricted in their ability to detect nonlinear association, interactions, and complex patterns between environmental variables. A machine learning approach can learn relationships from observed data and then classify or continuously estimate them. The use of artificial neural networks (ANN) to forecast water quality has been seen in a wide variety of studies since their capacity to approximate complicated, nonlinear processes. ANN does come with its limitations, however, and the efficiency of these networks can be greatly influenced by the amount of data, their structure, and proper validation (Chen et al., 2020). Logistic regression, decision trees, random forest, support vector machine, k-nearest neighbour, gradient boosting, and ensemble models are some of the other algorithms used for water-quality prediction. Comparative studies show that none of the models is optimal in all settings and thus, model evaluation and validation are necessary overall (Hussein et al., 2023).

Water-quality modelling has progressed from simple prediction to recently accepting new challenges. The sensors, connected devices, and machine-learning models can be integrated into an IoT system to enable continuous monitoring and quick detection of any abnormal conditions (Bhardwaj et al., 2022). AI-powered machine-learning platforms can simplify the process of selecting and tuning a model, and EAI can explain why certain variables are more important than others in making predictions (Madni et al., 2023). SVMs have been useful in predicting indices, especially when used in conjunction with sensitivity analysis and proper parameter selection (Mamat et al., 2023). Moreover, the reliability of the model can be enhanced by optimizing the hyperparameters, like the grid search, which can be done to find optimal values of the hyperparameters instead of using default values (Shams et al., 2024). Preliminary optimization of modelling frameworks has also been suggested to enhance the prediction of water-quality characteristics of importance (Suwadi et al., 2022).

Another public-health dimension to water-quality prediction is that of public health. Accurate classification can facilitate the identification of safer water and aid early response to prevent widespread exposure of water by contamination. In this regard, machine-learning models have been used for public-health and drinking-water quality decision support (Özsezer & Mermer, 2024). Class-balancing techniques and explainable AI have also been employed in potability prediction studies to tackle the issue of unequal outcome distribution and enhance interpretability (Patel et al., 2022). Various frameworks have been developed for the recognition of safe drinking water and estimation of index categories through multiple learning algorithms (Torky et al., 2023). However, small sample size, data leakage, missing observations, class imbalance and outliers can compromise model performance. It is therefore important to preprocess the data transparently, validate it repeatedly and interpret it carefully.

Technical innovation and institutional cooperation are needed to develop sustainable water management. To turn water-quality evidence into action, there is a need for water-quality evidence to be shared through partnerships developed at the international level, for a shared approach to monitoring, and for a coordinated approach to governance (Koncagül & Connor, 2023). In this context, an integration of physicochemical assessment and machine learning can facilitate quicker interpretation, identification of influential parameters and prioritisation of monitoring efforts.

Objectives

1. Evaluation of water quality with some physicochemical parameters.
2. To benchmark the accuracy of several machine-learning algorithms.
3. To determine the most significant physicochemical parameters affecting the water-quality classification.

2. Materials and Methods

2.1 Research Design

A quantitative analytical design was used in this study to determine the water quality based on the physicochemical parameters and machine learning techniques. The data used for the analysis was secondary: "Water Quality" (Kadiwal, 2021).

2.2 Dataset and Variables

The data consisted of 35 water samples collected, with the parameters measured being temperature, pH, electrical conductivity, total dissolved solids, chloride, chemical oxygen demand, and biological oxygen demand. The physicochemical parameters were used as the predictor variables, and the overall water-quality category was derived as the target variable.

2.3 Data Preprocessing

Duplicated, missing and outliers were checked in the data set. Median imputation was used if necessary, when there were missing values for numbers. Environmental outliers were only added if they were plausible, and scale-sensitive variables were standardized before model development.

Predictors were not included that were derived from the status variables were not included to prevent data leakage.

2.4 Exploratory Data Analysis

The distributions and relationships of the parameters were analysed by descriptive statistics, histograms, boxplots and correlation analysis. Correlations were either Pearson or Spearman, depending on the distribution of the data, and a heat map was created to detect any potential multicollinearity.

2.5 Machine Learning Models

Water quality classification was done using logistic regression, decision tree, random forest, support vector machine, k-nearest neighbours and gradient boosting models. The data set was split in 80% training and 20% testing.

2.6 Model Validation and Evaluation

Due to the small number of samples, stratified cross-validation and grid search optimisation methods were employed. Confusion matrices, F1-Score, Recall, Precision, Accuracy, and ROC-AUC were used to evaluate the model's performance. The model that had the most consistent validation results was chosen as the best-performing model.

2.7 Feature Importance

To determine the most significant parameters for water-quality classification, the physicochemical parameters were ranked using two methods, feature importance and permutation importance.

2. 8 Ethical Considerations

There were no human participants, personal information or identifiable biological data in the study; the study was based on secondary data with physicochemical water measurements. So, there was no need for formal ethical approval and informed consent.

3. Results

3.1 Descriptive Characteristics of Physicochemical Parameters

The descriptive statistics of the Water Quality dataset are presented in Table 1. The number of samples available for testing with the instruments was 35 for temperature, pH, electrical conductivity, chemical oxygen demand, biological oxygen demand and total dissolved solids and 22 for chloride. The temperature was comparatively less variable, ranging from 26.50 to 37.00 °C, with the mean being 31.75 ± 2.55 °C. The mean pH was 8.10 ± 1.78, which indicates that the pH of some samples was alkaline.

Significant differences were found in the parameters related to pollution. Mean EC was 3.76 ± 1.57 mS/cm, while mean TDS was 2551.80 ± 1848.96 mg/L. COD and BOD recorded high mean concentrations of 1997.46 ± 235.46 mg/L and 681.17 ± 207.91 mg/L, respectively. The relatively broad TDS range of 286–8416 mg/L suggested high variation in the amounts of dissolved pollutants in the samples. The normalised distribution of the measured parameters is presented in Figure 1.

Table 1. Descriptive statistics of physicochemical water-quality parameters

Parameter	N	Mean	Standard deviation	Minimum	Median	Maximum
Temperature (°C)	35	31.75	2.55	26.50	32.00	37.00
pH	35	8.10	1.78	5.49	8.13	11.71
EC (mS/cm)	35	3.76	1.57	0.87	4.12	6.74
Chloride (mg/L)	22	443.59	118.12	225.00	456.50	667.00
COD (mg/L)	35	1997.46	235.46	1147.00	2013.00	2351.00
BOD (mg/L)	35	681.17	207.91	278.00	708.00	1034.00
TDS (mg/L)	35	2551.80	1848.96	286.00	2460.00	8416.00

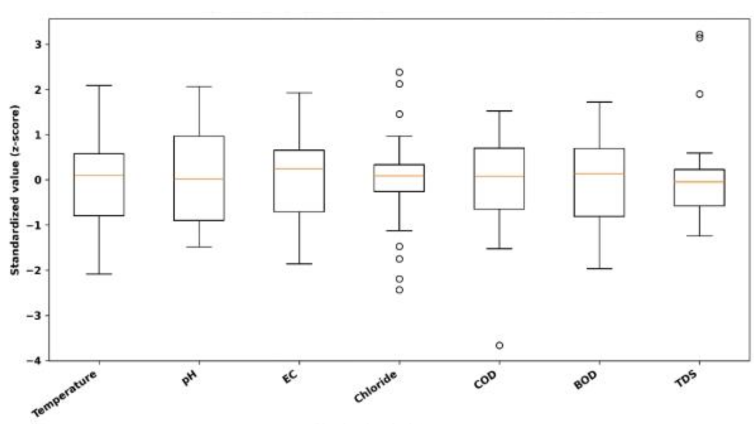

Figure 1. Distribution of standardised physicochemical water-quality parameters

3.2 Water-Quality Limit Exceedances

Significant degradation in water quality was evident for the evaluated samples in the compliance assessment. The 14 samples (40.0%) had pH values not within the reference range of 6–9 as reported in Table 2. Twenty-two samples recorded values of EC > 3 mS/cm, and 20 samples had TDS values > 2000 mg/L. Of 22 samples that had chloride measurements, 20 exceeded the 250 mg/L limit.

The highest level of pollution was noted for COD and BOD. The COD concentration exceeded the limit of 250mg/L for all samples, while the BOD exceeded the limit of 30mg/L for all samples. If that is the case, a simple acceptable-versus-unacceptable target would have included just one class. A relative pollution burden classification was thus created for exploratory machine learning analysis. A result of this yielded 21 samples with lower burden and 14 samples with higher burden.

Table 2. Compliance and pollution-burden characteristics of the samples

Indicator	**Reference criterion**	**Within limit, n (%)**	**Above/outside limit, n (%)**	**Missing**
pH	6–9	21 (60.0)	14 (40.0)	0
EC	≤3 mS/cm	13 (37.1)	22 (62.9)	0
TDS	≤2000 mg/L	15 (42.9)	20 (57.1)	0
Chloride	≤250 mg/L	2 (9.1)	20 (90.9)	13
COD	≤250 mg/L	0 (0.0)	35 (100.0)	0
BOD	≤30 mg/L	0 (0.0)	35 (100.0)	0

3.3 Relationships Among Physicochemical Parameters

Both weak and strong correlation was observed between the physicochemical parameters by Spearman correlation analysis. The complete correlation structure is presented in Figure 2, and the main associations are presented in Table 3. EC showed the highest correlation with TDS, which was 0.837. This has shown that the higher the electrical conductivity of the water sample, the higher the concentration of the dissolved solids.

A moderate correlation was observed between TDS and COD (rs = 0.401), which indicated that higher concentrations of dissolved solids was related to higher chemical oxygen demand. The positive relationship of COD was less with pH (rs = 0.266) and BOD (rs = 0.227). This indicates that BOD had a weak and negative correlation with temperature (rs = −0.274). Chloride showed weak correlations with the other variables, suggesting that its variation was relatively independent of the other variables in the observations available.

Table 3. Selected Spearman correlations among physicochemical parameters

Parameter pair	Spearman coefficient	Strength and direction
EC–TDS	0.837	Strong positive
COD–TDS	0.401	Moderate positive
Temperature–BOD	−0.274	Weak negative
pH–COD	0.266	Weak positive
EC–COD	0.250	Weak positive
COD–BOD	0.227	Weak positive
Temperature–Chloride	0.224	Weak positive
pH–TDS	0.161	Weak positive
EC–BOD	−0.123	Weak negative
Chloride–TDS	−0.043	Negligible negative

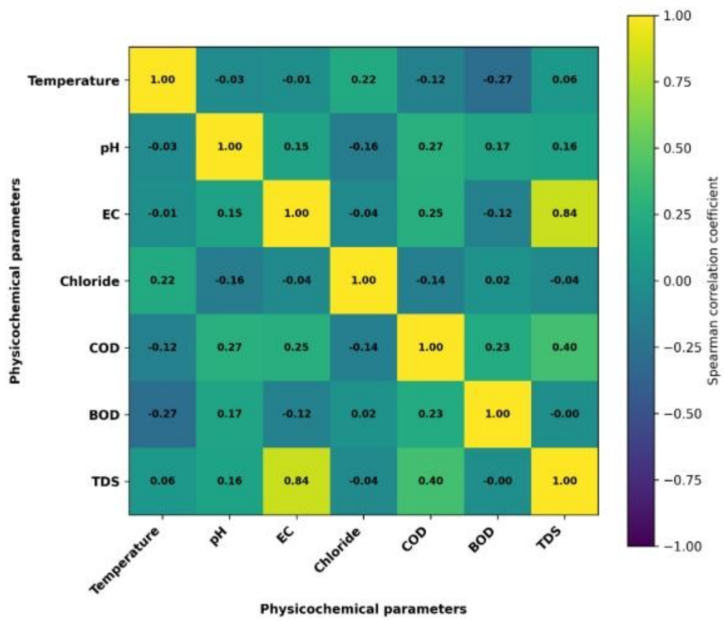

Figure 2. Spearman correlation heatmap of physicochemical water-quality parameters

3.4 Principal Component Structure

The physicochemical parameters were median imputed and standardised before performing the PCA. The first 3 components accounted for 64.22% of the total variance as shown in Table 4. PC1 accounted for 28.46% of the variance and was mainly related to the dissolved-solid and chemical-pollution dimension, with TDS, EC and COD being the main components. PC2 accounted for 19.53% of the total variance and was principally correlated with BOD and temperature, while PC3 accounted for 16.23% of the total variance and was largely correlated with chloride and temperature.

The total variance was 89.50% accounted for by the first five components. Thus, from these findings, it was concluded that water-quality variability was multidimensional and one physicochemical parameter was not sufficient to fully describe these variations.

Table 4. Principal component variance and major loadings

Component	Variance explained (%)	Cumulative variance (%)	Major positive loadings	Major negative loadings
PC1	28.46	28.46	TDS (0.632), EC (0.503), COD (0.434)	—
PC2	19.53	47.99	BOD (0.604), COD (0.325)	Temperature (−0.524), EC (−0.435)
PC3	16.23	64.22	Chloride (0.716), Temperature (0.487), BOD (0.299)	EC (−0.287)

3.5 Machine-Learning Model Performance

Repeated stratified five-fold cross-validation (5FCV) was used for comparing six machine learning algorithms. Table 5 and Figure 3 present the average level of classification performance. Gradient boosting was the best overall performance with an accuracy of 0.880, F1 score of 0.837 and ROC-AUC of 0.926. Random forest gave the second-highest ROC-AUC value of 0.917 and an accuracy of 0.811, and an F1-score of 0.780.

The decision tree also showed a relatively good performance of accuracy of 0.806 and recall of 0.860. Logistic regression and support vector machine yielded moderate results, with the k-nearest neighbours giving the least performance, especially in terms of the recall and the F1 score. The relatively large differences among the validation runs were due to the small number of runs and suggested that the performance predictions should be used with caution.

Table 5. Repeated cross-validation performance of machine-learning models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic regression	0.640	0.533	0.587	0.530	0.737
Decision tree	0.806	0.718	0.860	0.770	0.821
Random forest	0.811	0.752	0.860	0.780	0.917
Support vector machine	0.709	0.593	0.713	0.635	0.760
K-nearest neighbours	0.594	0.320	0.213	0.241	0.627
Gradient boosting	**0.880**	**0.863**	0.840	**0.837**	**0.926**

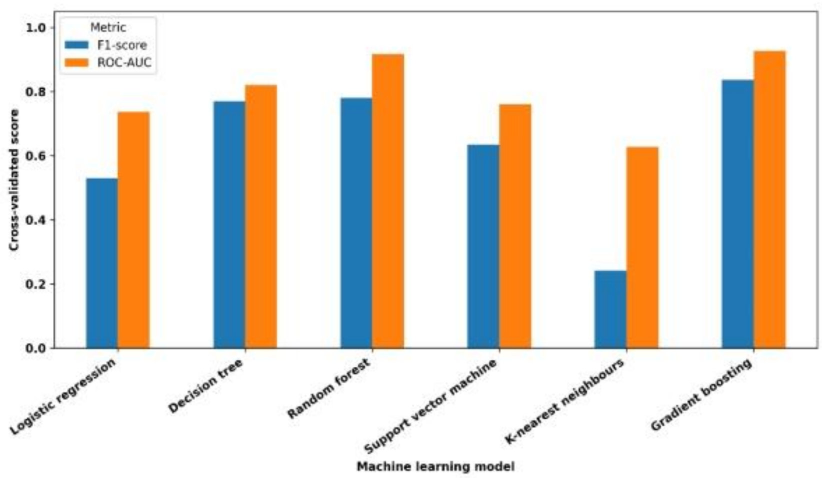

Figure 3. Cross-validated performance of the evaluated machine-learning models

3.6 Importance of Physicochemical Predictors

The variables that most strongly contributed to the identification of classes of pollution burden were identified via permutation importance from the random forest model. TDS had the highest mean permutation importance (0.201) as observed in Table 6 and Figure 4. The second most important parameter was EC (value 0.125), followed by chloride (value 0.077).

The smallest contributions were given by temperature, pH and BOD, while the lowest was for COD. The small contribution of COD and BOD was partly attributed to the uniformly high concentrations of these parameters in all the samples; both were greater than the reference concentrations and hence did not discriminate among the relative pollution burden groups.

Table 6. Random forest permutation importance of physicochemical predictors

Rank	Parameter	Mean importance	Standard deviation
1	TDS	0.201	0.065
2	EC	0.125	0.053
3	Chloride	0.077	0.029
4	Temperature	0.026	0.022
5	pH	0.025	0.024
6	BOD	0.022	0.024
7	COD	0.008	0.017

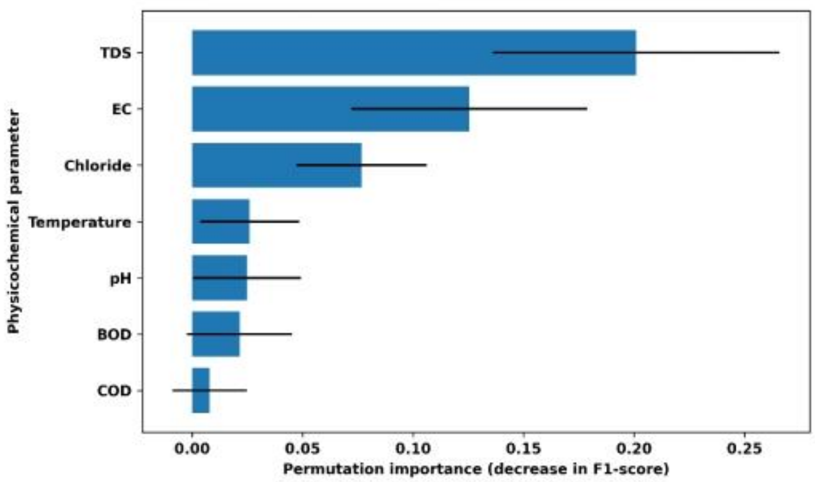

Figure 4. Permutation importance of physicochemical parameters in the random forest model

3.7 Overall Results

The overall results showed significant water physical and chemical deterioration in the water samples examined. All the samples had COD and BOD levels that were above their reference values, and EC, TDS and chloride levels had frequent exceedance. The two parameters (EC and TDS) had high correlation and were the main dimension of dissolved pollutants in PCA. Gradient boosting (GB) showed the best classification results of the tested machine-learning techniques; TDS, EC and chloride were found to be the most important indicators of the relative pollution burden. The sample size was small, and the target variable was derived, which makes the results of the machine learning interpretation exploratory and not externally generalisable.

4. Discussion

The present work explores water quality assessment by employing a physicochemical parameter and various machine-learning models. Results indicated that the samples were degraded, especially in the chemical oxygen demand, biological oxygen demand, total dissolved solids, electrical conductivity, and chloride. All the samples showed a high level of organic and chemical pollution as the levels of COD and BOD were above the reference values. The high frequency of exceeding TDS, EC, and chloride indicated high concentration of dissolved ions and anthropogenic contamination. Studies have also noted that physicochemical parameters can offer complementary information and should be interpreted rather than as individual parameters when characterising overall water quality (Chatterjee et al., 2024).

One of the statistically significant relationships was the positive correlation between EC and TDS. This was to be expected as the more ions in the solution, the more conductive it will be. The moderate correlation between TDS and COD shows that samples with higher dissolved solids tended to have higher chemical oxygen demand. The patterns reinforce the use of an integrated water-quality assessment instead of using individual water quality parameters. Weighted water-quality indices have also been applied to integrate several aspects of water-quality and enhance groundwater pollution pattern interpretation (Das & Das, 2024). Another benefit of optimized models of the WQI is that the parameter weighting can enhance the representation of water-quality conditions and minimize the redundancy among correlated parameters (Ding et al., 2023).

The principal component analysis indicated that TDS, EC and COD were the major polluting parameters present in the dominant pollution dimension. The first three components accounted for 64.22% of the total variance, thus indicating the multidimensionality of water quality. The first component mainly explained the dissolved solids and chemical pollution, while the second and third components captured the organic load, the temperature variation, and chloride related variation. Dimensionality reduction is used to group pollutants and simplify the model. The use of random-forest-based customisation of water-quality indices has also revealed that data-driven feature structures can provide improved local relevance over conventional water-quality indices that are fixed (Lee et al., 2023).

Out of the evaluated algorithms, the highest overall performance was obtained by gradient boosting with an accuracy score of 0.880, F1 score of 0.837 and ROC-AUC score of 0.926. Random forest was a strong performer with the ROC-AUC of 0.917. The results indicate that ensemble tree-based methods were more appropriate for the nonlinear and interactive relationships between the physicochemical data. Similarly, ensemble techniques have been shown to perform well when predicting WQIs in previous studies (Sidek et al., 2024). When the environmental relationships are complicated, combined learners often outperform each other and conventional machine learning and neural networks (Satish et al., 2024). A strong coastal WQI modelling has also shown that the use of ensemble algorithms is beneficial in terms of the stability of the predictions (Uddin et al., 2022).

The decision tree performed relatively well in terms of recall, compared to logistic regression, support vector machine and k-nearest neighbours. This variation proves that algorithm selection should be done on an empirical basis, not on assumption. The tuned models built using machine learning algorithms can vary significantly according to the pre-processing, tuning parameters and validation procedure (Shams et al., 2024). Comparative WQI studies have also revealed that model ranking depends on the set of data, the location and the intended objective as defined by the model targets (Hussein et al., 2023). SVMs are capable of reasonable performance, but the accuracy will be affected by kernel selection, scaling and predictor selection (Mamat et al., 2023). Lake-water studies have also shown the structure and size of the available data is linked to the predictive performance (Prasad et al., 2021).

TDS, EC and chloride were the most important predictors of relative pollution burden for the permutation importance. This discovery is consistent with the analysis results for water-quality classifications (Chatterjee et al., 2024) that indicate that a small number of discriminative physicochemical variables can account for a large portion of the variation in water-quality classifications. Another key aspect of EAI solutions is to focus on understanding the contribution of individual predictors to the model decision-making, beyond just reporting accuracy (Madni et al., 2023). In contrast, COD and BOD had lower importance as both parameters were measured above the limits in all samples and contributed little to the separation of the relative pollution groups.

A word of caution about these findings. A small sample size makes the estimates unstable, and "overfits" them. Machine learning techniques can provide better predictions in sparse data sets, but there is still a lot of uncertainty when observations are limited (Huang et al., 2024). The feature importance and classification performance might have been affected by the missing chloride values. A study on WQI estimation has revealed that if not properly addressed, missingness may affect the WQI values, thereby compromising the model's reliability (Sierra-Porta, 2024). Furthermore, it is crucial to consider uncertainty-aware modelling as traditional performance measures can hide prediction variability (Uddin et al., 2023).

In conclusion, the study shows that ensemble machine learning and physicochemical assessment can reveal the dominant pollution patterns and help with exploratory classification. But more data, external validation, seasonal data and other parameters are required for models to support operational use.

5. Conclusion

This study showed that the use of machine learning for water-quality assessment, together with a physicochemical assessment, is helpful. The results indicated a significant difference in the measured parameters, especially for chemical oxygen demand, biological oxygen demand, total dissolved solids, electrical conductivity and chloride. The COD and BOD values were above the selected reference limit for all samples, which suggests there is a significant amount of organic and chemical pollution. Additional analysis in the form of correlation analysis was also performed for electrical conductivity and total dissolved solids, which revealed a strong positive relationship between the two, and principal component analysis further confirmed that the main sources of variation were dissolved solids, ionic concentration and chemical pollution. Out of the evaluated machine-learning models, gradient boosting performed best, followed closely by random forest. The results indicate that ensemble tree-based algorithms are successful in modelling the nonlinear relationships between the physicochemical variables. The feature importance analysis showed that TDS, EC and chloride are the most important features for predicting the relative pollution burden, while COD and BOD were not so influential for classification, as their concentration was almost constant in all samples. Although these encouraging results have been obtained, the limited number of samples, missing chloride observations, and the classification of the target limit the generalisability of the results. Thus, the results of the machine learning should be viewed as exploratory.

References

1. Bhardwaj, A., Dagar, V., Khan, M. O., Aggarwal, A., Alvarado, R., Kumar, M., ... & Proshad, R. (2022). Smart IoT and machine learning-based framework for water quality assessment and device component monitoring. Environmental Science and Pollution Research, 29(30), 46018-46036.
2. Chatterjee, T., Gogoi, U. R., Samanta, A., Chatterjee, A., Singh, M. K., & Pasupuleti, S. (2024). Identifying the most discriminative parameter for water quality prediction using machine learning algorithms. Water, 16(3), 481.
3. Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A review of the artificial neural network models for water quality prediction. Applied Sciences, 10(17), 5776.
4. Connor, R. (2024). The United Nations World Water Development Report 2024: water for prosperity and peace; executive summary.
5. Das, C. R., & Das, S. (2024). Coastal groundwater quality prediction using objective-weighted WQI and machine learning approach. Environmental Science and Pollution Research, 31(13), 19439-19457.
6. Derdour, A., Jodar-Abellan, A., Pardo, M. Á., Ghoneim, S. S., & Hussein, E. E. (2022). Designing efficient and sustainable predictions of water quality indexes at the regional scale using machine learning algorithms. Water, 14(18), 2801.
7. Ding, F., Zhang, W., Cao, S., Hao, S., Chen, L., Xie, X., ... & Jiang, M. (2023). Optimization of water quality index models using machine learning approaches. Water research, 243, 120337.
8. du Plessis, A. (2023). Water resources from a global perspective. In South Africa's Water predicament: Freshwater's unceasing decline (pp. 1-25). Cham: Springer International Publishing.
9. Hassan, M. M., Hassan, M. M., Akter, L., Rahman, M. M., Zaman, S., Hasib, K. M., ... & Mollick, S. (2021). Efficient prediction of water quality index (WQI) using machine learning algorithms. Human-Centric Intelligent Systems, 1(3), 86-97.

10. Huang, S., Xia, J., Wang, Y., Lei, J., & Wang, G. (2024). Water quality prediction based on sparse dataset using enhanced machine learning. Environmental Science and Ecotechnology, 20, 100402.
11. Hussein, E. E., Jat Baloch, M. Y., Nigar, A., Abualkhair, H. F., Aldawood, F. K., & Tageldin, E. (2023). Machine learning algorithms for predicting the water quality index. Water, 15(20), 3540.
12. Kadiwal, A. (2021). Water quality: Drinking water potability [Data set]. Kaggle. https://www.kaggle.com/datasets/adityakadiwal/water-potability
13. Koncagül, E., & Connor, R. (2023). The United Nations World Water Development Report 2023: partnerships and cooperation for water; facts, figures and action examples.
14. Lee, H., Park, S., Hang, V., Nguyen, M., & Shin, H. S. (2023). Proposal for a new customization process for a data-based water quality index using a random forest approach. Environmental Pollution, 323, 121222.
15. Madni, H. A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., ... & Ashraf, I. (2023). Water-quality prediction based on H2O AutoML and explainable AI techniques. Water, 15(3), 475.
16. Mamat, N., Mohd Razali, S. F., & Hamzah, F. B. (2023). Enhancement of water quality index prediction using support vector machine with sensitivity analysis. Frontiers in Environmental Science, 10, 1061835.
17. Özsezer, G., & Mermer, G. (2024). Prediction of drinking water quality with machine learning models: A public health nursing approach. Public health nursing, 41(1), 175-191.
18. Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., ... & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Computational Intelligence and Neuroscience, 2022(1), 9283293.
19. Prasad, D. V. V., Venkataramana, L. Y., Kumar, P. S., Prasannamedha, G., Soumya, K., & Poornema, A. J. (2021). Prediction on water quality of a lake in Chennai, India using machine learning algorithms. Desalination and Water Treatment, 218, 44-51.
20. Satish, N., Anmala, J., Varma, M. R., & Rajitha, K. (2024). 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, 192, 177-195.

21. Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications, 83(12), 35307-35334.
22. Sidek, L. M., Mohiyaden, H. A., Marufuzzaman, M., Noh, N. M., Heddam, S., Ehteram, M., ... & Sammen, S. S. (2024). Developing an ensembled machine learning model for predicting water quality index in Johor River Basin. Environmental Sciences Europe, 36(1), 67.
23. Sierra-Porta, D. (2024). Assessing the impact of missing data on water quality index estimation: a machine learning approach. Discover Water, 4(1), 11.
24. Suwadi, N. A., Derbali, M., Sani, N. S., Lam, M. C., Arshad, H., Khan, I., & Kim, K. I. (2022). An optimized approach for predicting water quality features based on machine learning. Wireless Communications and Mobile Computing, 2022(1), 3397972.
25. Torky, M., Bakhiet, A., Bakrey, M., Ismail, A. A., & Seddawy, A. I. E. (2023). Recognizing safe drinking water and predicting water quality index using machine learning framework. International Journal of Advanced Computer Science and Applications, 14(1).
26. Uddin, M. G., Nash, S., Diganta, M. T. M., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. Journal of Environmental Management, 321, 115923.
27. Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. Water Research, 229, 119422.
28. Warner, S. (2021). Progress on Ambient Water Quality: Tracking SDG 6 Series-Global Indicator 6.3. 2 Updates and Acceleration Needs.