International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 3, issue no. 1, Dec 2025, 91-99
Received Date: Jul 30, 2025
Accepted Date: Sep 24, 2025
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v3i1.87014

Comparative Analysis of Traditional and Ensemble Models for Water Quality Index Prediction with Explainable AI

Amrit Kandel [1*], Rajad Shakya[2]

[1]*Department of Electronics and Computer Engineering, Thapathali, Kathmandu, Nepal, amritkandel062@gmail.com*
[2]*Department of Electronics and Computer Engineering, Thapathali, Kathmandu, Nepal, shakyarajad1@gmail.com*

Abstract

Accurate prediction of water quality is vital for effective environmental management. This study presents a comparative analysis of traditional and ensemble machine-learning models for predicting the Canadian Council of Ministers of the Environment Water Quality Index (CCME-WQI) using EPA Ireland coastal monitoring data. A standardized and leakage-proof pipeline was employed with robust scaling and multiple cross-validation across multiple random seeds to ensure stable and reproducible performance. Among all models, XGBoost achieved the best performance ($R^2 = 0.991$). Model interpretability was enabled by SHAP analysis supported by feature correlation that identified Dissolved Oxygen as the dominant factor of WQI. Overall, results illustrate the potential of ensemble learners combined with explainable AI in making accurate, interpretable, and generalizable water-quality predictions to enable data-driven environmental monitoring and decision-making.

Keywords: Water Quality Index (WQI), Machine Learning, XGBoost, SHAP, Ensemble Algorithms, Explainable AI (XAI), CCME WQI, Regression, Water Resource Management

1. Introduction

Water quality directly influences sustainable development, ecosystems, and health. Urbanization, industrial discharges, global warming, and agricultural runoff have, nonetheless, crucially deteriorated the freshwater resource worldwide (Gleick & others, 2018) Effective monitoring and management of the water resources are therefore vital to sustainability.

Water Quality Index (WQI) is a composite indicator that combines physical, chemical, and biological parameters in a single value to effectively represent total water quality status. Among different formulations, the Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI) has been most used because it is flexible and stable in handling temporal and spatial variability (WATER, 2001). Although useful, manual computation and static thresholding of traditional WQI evaluations constrain their precision and scalability. As more sensor-based and archive water quality data are coming online, more data-driven methods that can capture intricate nonlinear relationships among water parameters are needed.

Machine learning (ML) provides a viable solution to such constraints. Ensemble methods like Random Forest, XGBoost, and LightGBM have been demonstrated to deliver outstanding predictive performance and stability over standard models like Linear Regression and K-Nearest Neighbors (Chen & Guestrin, 2016; Ke, et al., 2017). Ensemble models are black boxes, however, and what they are predicting cannot be easily explained—a key consideration in environmental decision-making. To overcome this, Explainable Artificial Intelligence (XAI) models, namely SHapley Additive exPlanations (SHAP), have been found to be efficient tools to quantify the contribution of each feature towards model predictions (Lundberg & Lee, 2017).

The current study seeks to devise a comparative and interpretable prediction of WQI. To this end, we:

- Compare traditional and ensemble model performances in CCME WQI prediction using physicochemical parameters.

* *Corresponding author*

- Combine SHAP-based model interpretation to determine the impactful features affecting WQI.
- Show how joining strong preprocessing, ensemble learning, and explainability improves accuracy and transparency of water quality modeling.

2. Related Works

Recent studies prove that there is growing interest in incorporating machine learning into water quality assessment models. Han et al. (2025) utilized ensemble algorithms like Random Forest and XGBoost in predicting CCME WQI with high accuracy and in determining key features. Their study proved that ensemble methods surpass traditional methods because they can capture nonlinear relationships. Similarly, Chidiac et al. (2023) emphasized developing scalable and intelligent models which can handle noisy, high-dimensional environmental data to form the basis for today's ML-based WQI systems.

Earlier studies employed traditional ML algorithms like Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM); however, they often struggle to capture nonlinearity and interaction between features, thus reducing accuracy (Chen & Guestrin, 2016; Ke, et al., 2017). These limitations are addressed by ensemble methods like XGBoost, Random Forest, and LightGBM through feature subsampling and iterative optimization, which offer better generalization and predictive stability.

Although predictive performance is essential, interpretability has long been a concern in environmental modeling. Rudin (2019) strongly advocated model interpretability by promoting the development and use of interpretable models, cautioning against the unquestioned reliance on black-box models. Therefore, SHAP has become a widely accepted method of interpretation to approximate global as well as local contributions of features (Lundberg & Lee, 2017). Experiments performed by Nallakaruppan et al. (2024) and Yang et al. (2024) indicated that SHAP can effectively identify influential physicochemical factors such as dissolved oxygen, pH, and hardness, creating knowledge in line with available environmental knowledge and increasing confidence among stakeholders.

Overall, then, current research shows significant accuracy and WQI prediction automation but often lacks a combined emphasis on transparency and reliability. An equilibrium towards the combination of CCME WQI architectures, ensemble ML models, and SHAP-based explainability, as employed in this study, provides a balanced pathway towards high-performance, transparent, and interpretable water quality assessments.

3. Related Theory

3.1. CCME WQI

CCME WQI integrates several water quality parameters into a single metric reflecting overall water quality in a way that makes it easy to interpret and compare across places and over time. It integrates three factors:

- **F1 (Scope):** Percentage of parameters exceeding guideline limits.
- **F2 (Frequency):** Percentage of individual tests exceeding objectives.
- **F3 (Amplitude):** Magnitude by which failed test results exceed guidelines.

The index is calculated as in (Equation 1)

$$\text{CCME WQI} = 100 - \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \quad \text{(Equation 1)}$$

Physicochemical parameters are considered with reference to environmental objectives. Each parameter contributes to F1, F2, and F3 based on frequency and degree of deviation from target range. Incorporating scope, frequency, and amplitude, CCME WQI provides a comprehensive but flexible assessment, both indicative of degree and severity of water quality issue. This makes it suitable to monitor temporal and spatial trends in complex water systems, informing environmental decisions and policy.

3.2. Robust Scaling

Robust Scaling is a data preprocessing method that scales features using statistics that are robust to outliers. As opposed to the usual scaling using the mean and standard deviation, Robust scalar uses the median and

interquartile range (IQR) and is hence more suitable for data containing outliers. It centers data using the subtraction of the median and scales using the IQR. It gives a more stable scaling for heavy-tailed or skewed distributions. The transformation function for robust scaling is given by (Equation 2).

$$x_{\text{scaled}} = \frac{x - \text{Median}(x)}{\text{IQR}(x)} \quad \text{(Equation 2)}$$

3.3. Traditional Regression Algorithms

Traditional regression algorithms make predictions based on statistical assumptions or simple rule-based systems. These models are typically interpretable and computation-wise lightweight. They work well with small to medium-sized datasets and are typically the first choice for model experimentation. Table 1 shows few of the well-known traditional algorithms are Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Decision Tree Regression.

Table 1. Traditional Regression Algorithms

Algorithm	Description
Linear Regression	Assumes a linear relationship between input and output. Optimizes by minimizing the sum of squared residuals. Simple and interpretable.
K-Nearest Neighbors (KNN)	A non-parametric method that predicts the output as the average of the k nearest training instances. Effective for non-linear data but sensitive to k and the distance metric.
Support Vector Regressor (SVR)	It fits the best line within a tolerance margin (ϵ). Effective in high-dimensional spaces and robust to outliers.
Decision Tree Regressor	A rule-based method that splits data based on feature thresholds to minimize prediction error. Interpretable but can overfit if not pruned.

3.4. Ensemble Regression Algorithms

Ensemble regression algorithms combine predictions of multiple models to improve accuracy and generalization. They are especially helpful for complex datasets where individual models can perform worse. Table 2 describes the ensemble techniques like Random Forest, XGBoost, and LightGBM, that were used in the study.

Table 2. Ensemble Regression Algorithms

Algorithm	Description
Random Forest Regressor	An ensemble of decision trees trained on bootstrapped samples with random feature selection. It averages the outputs of numerous trees to minimize variance and over fitting.
XGBoost Regressor	An optimized gradient boosting library with regularization, second-order gradient optimization, and internal missing data handling. It is well known for speed and accuracy.
LightGBM	A fast, histogram-based gradient boosting method with tree growth in the leaf-wise direction. Suitable for large datasets and has less memory usage.

3.5. Model Explainability

Model explainability is required to know how input features affect the prediction of a machine learning model. SHAP (SHapley Additive exPlanations) is a widely used framework for model explainability. It quantifies the impact of each input feature on the model output and makes it possible to see which features contribute the most to the predictions and whether they have a positive or negative influence. This form of interpretability enhances transparency, develops trust in model outputs, and helps in the diagnosis of potential issues in the modeling process. The explainability of a model can be mathematically represented by (Equation 3).

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad \text{(Equation 3)}$$

Where $f(x)$ is the model output for input x, $\phi_0$as the base value, ϕ_i is the SHAP value for the feature i, and M is the number of features.

4. Methodology

4.1. Dataset Description

Water quality data from the Environmental Protection Agency (EPA) of Ireland covered many coastal stations like Cork Harbour and Moy Killala with 29,159 valid data and 11 physicochemical parameters such as alkalinity, ammonia, BOD, chloride, conductivity, dissolved oxygen, phosphate, pH, temperature, total hardness and true colour. Among them, 5,000 random samples of records were used for modeling based on three random seeds (42, 77, 123). This sample size was chosen to enable repeated training over multiple seeds, and 5-fold cross-validation runs while preserving reasonable computational expense. Results were compared against one another to determine consistency, and statistical tests ensured that the samples retained distributions proximally to the entire dataset. All following preprocessing, modeling, and evaluation steps were executed uniformly over these sampled datasets. Table 3 describes the features, their environmental significance, and appropriate normal ranges. Figure 1 shows the distribution of all the features of raw data.

Table 3. Description of Selected Water Quality Parameters (Abdul Hameed M Jawad, et al., 2010)

Feature	Description	Suitable Range
Alkalinity	Buffering capacity of water against pH changes	20-200 mg/L
Ammonia	Indicator of organic pollution; toxic at high levels	0.5 mg/L
BOD	Biodegradable organic matter level (oxygen demand)	< 5 mg/L
Chloride	Reflects salinity or contamination	< 250 mg/L
Conductivity	Ionic concentration in water	50-150 µS/cm
Dissolved oxygen	Critical for aquatic organisms	5 mg/L
Phosphate	Can lead to eutrophication if excessive	0.1 mg/L
pH	Acidity or alkalinity of water	6.5 – 8.5
Temperature	Affects reaction rates and DO solubility	25ºC
Total Hardness	Calcium and magnesium content	60-180 mg/L
True colour	Color due to dissolved organic/inorganic matter	15 TCU

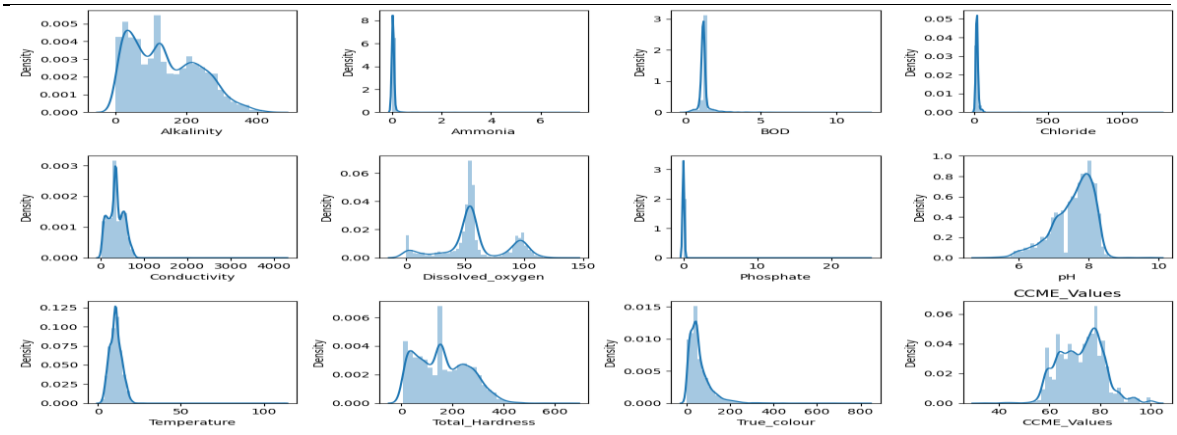

Figure 1. Distribution of selected features in the sample dataset.

4.2. Data Preprocessing

Preprocessing was done with scikit-learn pipelines to prevent data leakage. Numerical features were scaled robustly using the interquartile range for outlier treatment. The data was divided between training and test sets in an 80:20 proportion and all the transformations were fitted to the training folds only during cross-validation time.

4.3. Model Selection and Training

Traditional regression models (Linear Regression, K-Nearest Neighbors, Support Vector Regressor, Decision Tree Regressor) and ensemble models (Random Forest, LightGBM, XGBoost) were explored. Ensemble models were selected because they can model nonlinear interaction and provide stable predictions.

GridSearchCV was used to find the best hyper parameters for all the models on training data. The tuned hyperparameters values are tabulated as shown in Table 4. Models' performance was evaluated through 5-fold KFold cross-validation repeated across three different random seeds (42, 77, 123) for reproducibility purposes. An additional site holdout test was done, excluding whole monitoring sites from training and reserving them for test use to account for spatial generalization. All steps were performed within scikit-learn pipelines to maintain preprocessing consistency and prevent leakage.

Table 4. Tuned hyperemeters for all algorithms

Algorithm	Hyperparameters
Linear Regression	fit_intercept = 'False'
K-Nearest Neighbors (KNN)	leaf_size = 10, n_neighbors = 7, weights = 'distance', p = 1, algorithm = 'auto'
Support Vector Regressor (SVR)	C = 1, epsilon = 0.1, gamma = 'auto', kernel = 'rbf'
Decision Tree Regressor	Max_depth = 10, min_sample_split = 5, min_sample_leaf = 1
Random Forest Regressor	N_estimators = 200, max_depth = 12
LightGBM Regressor	N_estimators = 500, max_depth = 6, learning_rate = 0.05, subsample = 0.8, num_leaves = 31, min_child_samples = 5, min_split_gain = 0.0
XGBoost Regressor	N_estimators = 200, max_depth = 3, learning_rate = 0.2 , subsample = 1.0, colsample_bytree= 0.7

4.4. Evaluation Metrics

Model performance was evaluated in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). Results along with the standard deviation across the 5 folds were provided to report variability and model stability. Site holdout test results were also reported to demonstrate generalization to previously unseen sites. They are briefly discussed in Table 5.

Table 5. Evaluation Metrics for Regression Models

Metric	Formula	Description
Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Calculates the square root of the mean of the squared differences between actual and predicted values. More harshly penalizes larger errors.
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Calculates mean of the absolute difference between actual and predicted values. More interpretable but less sensitive to large errors.

Coefficient of Determination (R^2 Error)	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}$	Indicates how well the model explains the variance of the target variable. A score nearer to 1 implies better model fit.

4.5. Model Explainability using SHAP

To describe the results of the top-performing model (XGBoost), SHapley Additive exPlanations (SHAP) was employed. SHAP values provide the contribution of each input feature to the prediction of the model, making global (whole feature importance) and local (individual prediction) explainability possible. Feature contributions were assessed across the sampled data and cross-validation folds to give consistent insights. This approach facilitated transparency, identified key drivers of water quality predictions.

5. Results and Discussion

5.1. Model Performance Evaluation

The performance of all the models was evaluated on three regular regression measures: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). The comparative performance of all the models is reported in Table 6.

Table 6. Performance comparison of regression models

Model	Seed = 42			Seed = 77			Seed = 123		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Linear Regressor	0.3280 ± 0.0083	0.4387 ± 0.0179	0.8076 ± 0.0120	0.3334 ± 0.0128	0.4483 ± 0.0191	0.7987 ± 0.0129	0.3269 ± 0.0053	0.4411 ± 0.0214	0.8024 ± 0.0201
KNN Regressor	0.1155 ± 0.0049	0.1944 ± 0.0157	0.9168 ± 0.0124	0.1211 ± 0.0032	0.1957 ± 0.0133	0.9168 ± 0.0109	0.1150 ± 0.0040	0.1909 ± 0.0175	0.9172 ± 0.0156
Support Vector Regressor	0.1218 ± 0.0031	0.1809 ± 0.0070	0.9279 ± 0.0074	0.1302 ± 0.0031	0.1963 ± 0.0133	0.9164 ± 0.0105	0.1196 ± 0.0050	0.1739 ± 0.0085	0.9317 ± 0.0072
Decision Tree Regressor	0.0723 ± 0.0052	0.1440 ± 0.0184	0.9541 ± 0.0107	0.0687 ± 0.0054	0.1338 ± 0.0199	0.9601 ± 0.0135	0.0673 ± 0.0026	0.1396 ± 0.0065	0.9558 ± 0.0052
Random Forest Regressor	0.0452 ± 0.0027	0.1074 ± 0.0155	0.9726 ± 0.0080	0.0437 ± 0.0010	0.1058 ± 0.0159	0.9741 ± 0.0074	0.0459 ± 0.0026	0.1103 ± 0.0171	0.9725 ± 0.0075
LightGBM	0.0329 ± 0.0025	0.0772 ± 0.0107	0.9863 ± 0.0041	0.0283 ± 0.0023	0.0812 ± 0.0143	0.9849 ± 0.0051	0.0297 ± 0.0044	0.0771 ± 0.0183	0.9856 ± 0.0067
XGBoost Regressor	**0.0339 ± 0.0029**	**0.0629 ± 0.0028**	**0.9913 ± 0.0008**	**0.0343 ± 0.0019**	**0.0630 ± 0.0009**	**0.9912 ± 0.0027**	**0.0322 ± 0.0018**	**0.0630 ± 0.0037**	**0.9910 ± 0.0011**

Bold values denote the best performance.

Amongst all the models, XGBoost Regressor had the best performance with the smallest RMSE and largest R^2 score. This indicates the strong ability of the model in identifying non-linear correlations in environmental data. Random Forest and LightGBM also had a good performance, indicating the power of ensemble learning. Conversely, regular models like Linear Regression and KNN demonstrated relatively weaker accuracy, indicating their inability to adapt to identify sophisticated interactions

Figure 2 visualizes the predicted vs. actual WQI values for all algorithms to further assess the model predictions. A closer alignment of points along the ideal fit line indicates better model performance.

5.2. Model Explainability using SHAP

The SHAP (SHapley Additive exPlanations) framework was employed to explain the predictions made by the best-performing model (XGBoost). Figure 3 shows the pairwise correlation between the physiochemical parameters and the WQI value. Figure 4 shows the SHAP summary plot, which identifies the most influential features across the dataset.

The SHAP summary plot provides a visual representation of how much each input feature contributes to the predicted Water Quality Index (WQI). Features are listed by importance on the y-axis, and the x-axis contains SHAP values representing each feature's contribution to the prediction, either positive or negative. Each point is a sample from the data, color indicating the actual value of the feature (red for high, blue for low).

Correlation heatmap and SHAP summary plot together highlight contributing factors towards the Water Quality Index (WQI). It was observed that Dissolved Oxygen had the strongest positive correlation with WQI, followed by negative correlations of True Colour, Total Hardness, and Conductivity, as they are negatively affecting water quality when present in excess. SHAP summary plot further confirms these findings, with Dissolved Oxygen showing the highest positive impact on model output as the most significant feature, and True Colour and Total Hardness having negative SHAP values reducing the predicted WQI. Correlation and SHAP consistency reinforce model interpretability and demonstrate that the learned feature relations align with environmental understanding.

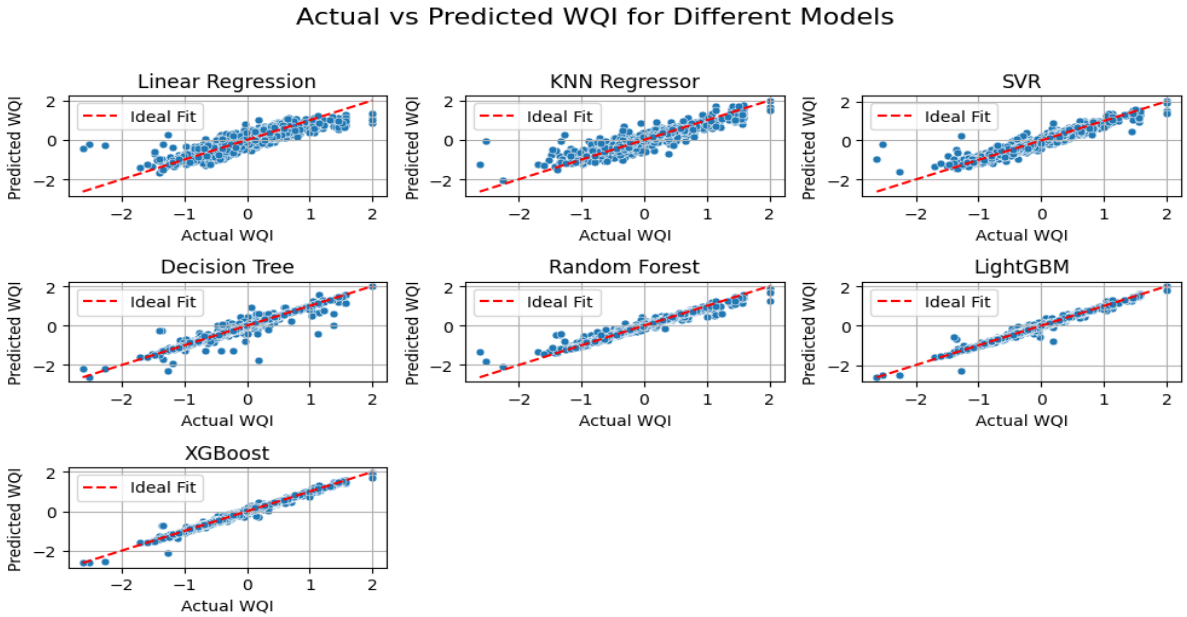

Figure 2. Predicted vs Actual WQI values for different regression models.

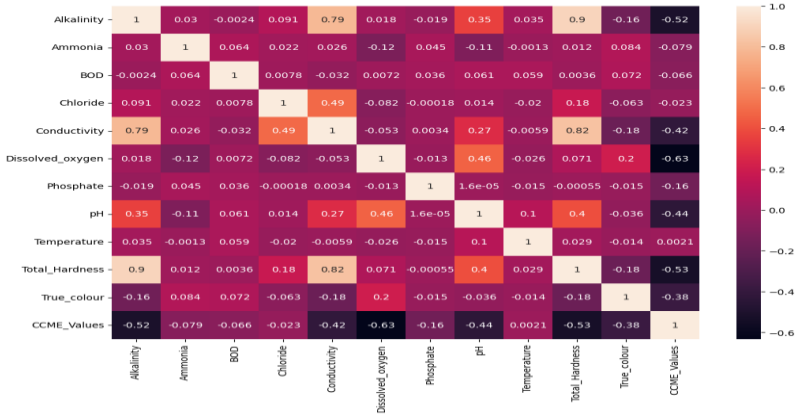

Figure 3. Correlation heatmap of the features

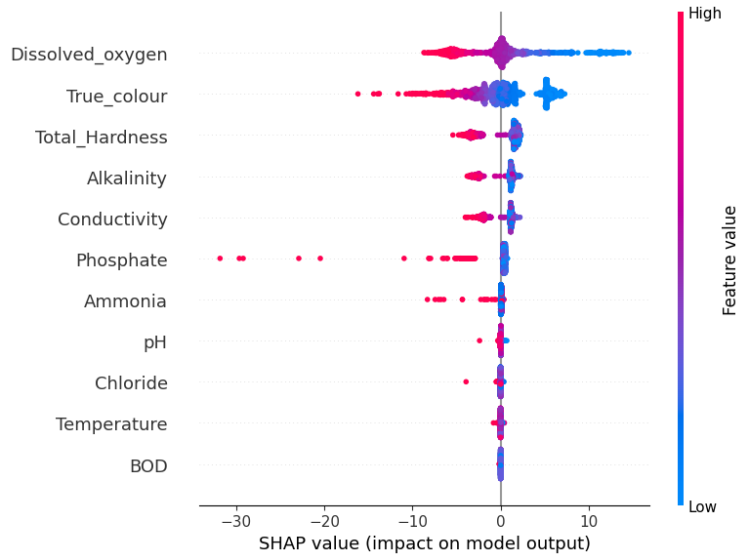

Figure 4. SHAP summary plot for XGBoost: global feature importance.

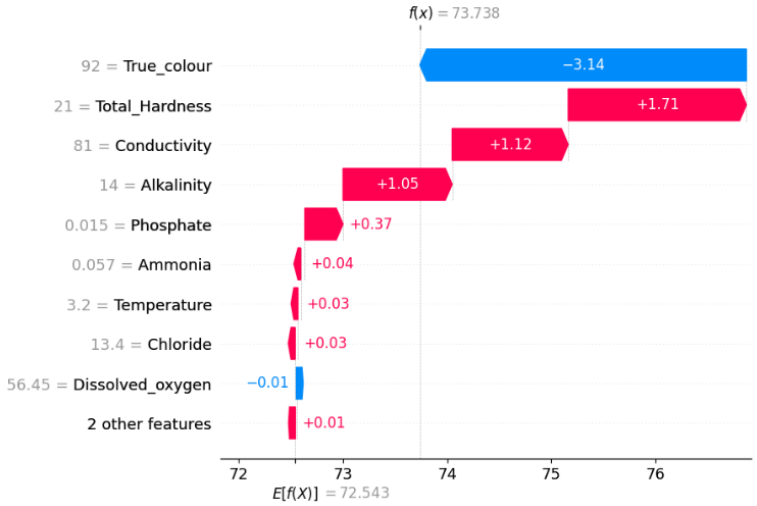

Figure 5. SHAP force plot illustrating feature contribution for a single instance.

In addition, local interpretability was explored using SHAP force plots for individual predictions. These visualizations showed the contribution of features for specific output values, making the model trustworthy and transparent.

Figure 5 indicates the contribution of each feature to the model's prediction of a specific instance. The base value of the model is 72.543, and the final prediction for this row is 73.738. True Colour had the highest negative impact, reducing the prediction by 3.14 units. Total Hardness, Conductivity, and Alkalinity features drove the prediction higher, with Total Hardness contributing the most positively (+1.71). This plot illustrates how the values of the single features in a particular data point affected the model output.

6. Conclusion

The present study compared traditional and ensemble machine-learning models in predicting CCME-WQI using a leakage-proof standardized pipeline. The models were robust using several random seeds and five-fold cross-validation, with small standard deviations for performance metrics.

XGBoost and LightGBM were the most stable and accurate of all the models. SHAP-based explainability identified dissolved oxygen, true colour, total hardness and alkalinity as the most influential features, and triangulation with correlation and partial dependence plots indicated consistent feature behavior. Overall, the proposed framework demonstrates that ensemble learners, when combined with explainable AI, provide an accurate, stable, and interpretable solution for water-quality assessment.

7. Limitations and Future Enhancements

Current study was limited to EPA Ireland coastal data, and this may restrict model generalization. Multi-region and multi-country datasets can be extended in the future, in addition to the inclusion of additional spatial and temporal variables. Other more interpretable baselines such as GLM and GAM can also be explored for comparison of interpretability. Spatial cross-validation and hybrid ensemble approaches will also be considered for added robustness and facilitation of real-time water-quality monitoring.

References

Abdul Hameed M Jawad, A., Haider S, A. & Bahram K, M., 2010. Application of water quality index for assessment of Dokan lake ecosystem, Kurdistan region, Iraq. *Journal of Water Resource and protection,* Volume 2010.

Chen, T. & Guestrin, C., 2016. *Xgboost: A scalable tree boosting system.* s.l., s.n., pp. 785--794.

Chidiac, S. et al., 2023. A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives. *Reviews in Environmental Science and Bio/Technology,* 22(2), pp. 349--395.

Gleick, P. H. & others, 2018. *The world's water volume 8: The biennial report on freshwater resources.* s.l.:Springer.

Han, Z., Zhang, S., He, L. & others, 2025. Predicting and investigating water quality index by robust machine learning methods. *Journal of Environmental Management,* Volume 381, p. 125156.

Ke, G. et al., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems,* Volume 30.

Lundberg, S. M. & Lee, S.-I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems,* Volume 30.

Nallakaruppan, M. K. et al., 2024. Reliable water quality prediction and parametric analysis using explainable AI models. *Scientific Reports,* 14(1), p. 7520.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence,* 1(5), pp. 206--215.

WATER, C., 2001. Canadian Water Quality Guidelines for the Protection of Aquatic Life. *User's Manual. Canadian Council of Ministers of the Environment. USA.*

Yang, S. et al., 2024. Estimating the water quality index based on interpretable machine learning models. *Water Science \& Technology,* 89(5), pp. 1340--1356.