



Data-Driven Prediction Of Water Quality Parameters Using Machine Learning Techniques

M.Gokul¹, K.Jeevalakshmi²

¹(B.Tech Final Year) Artificial Intelligence And Data Science

Dhanalakshmi Srinivasan Engineering College- Perambalur

Mahendranlohitha@Gmail.Com

²(Currently Doing Ph.D) Computer Science And Engineering

National Institute Of Technology - Puducherry

Lakshmijeeva6@Gmail.Com

Abstract- Predictive modeling of the parameters affecting water quality is crucial for public health, agricultural sustainability, and proper environmental action in advance. Existing lab-based monitoring techniques, being reliable and accurate, are not only expensive but also limited in terms of spatio-temporal coverage, which reduces their practical applicability. This research work aims to create an effective machine learning model for WQI and WQC prediction using the set of standard physicochemical water parameters. In particular, the methodology incorporates pre-processing using mean imputation and min-max normalization, feature importance analysis, and ensembles of several ML models, including Random Forest, XGBoost, Gradient Boosting, and Multi-layer Perceptron. As demonstrated by experiments, ensembles of tree-based algorithms are generally more accurate than individual baseline approaches in terms of prediction reliability for various types of water bodies. The paper also explores feature dominance and shows that pH, dissolved oxygen, and turbidity have the strongest effect on water quality state.

Key Word: Water Quality Prediction, Machine Learning, Random Forest, XGBoost, Gradient Boosting, Water Quality Index, Feature Importance, Environmental Monitoring

I. Introduction

The quality of water plays a vital role in defining ecosystem integrity, human health, and economic prosperity. WHO says that water of poor quality is one of the factors responsible for many diseases worldwide, especially in those places where monitoring facilities are not properly developed [4]. In the case of India, the problem becomes more challenging since the country uses groundwater and surface water for its domestic, agricultural, and industrial purposes [2][4]. The Central Pollution Control Board (CPCB) of India has classified water quality criteria into five distinct use categories, ranging from drinking water without any further processing to irrigation and industry cooling uses

The traditional process of assessing water quality involves periodic collection of samples in conjunction with lab analysis to measure physico-chemical and biological parameters. Despite the fact that the process results in accurate measurements, it comes with a number of weaknesses, including infrequent sampling periods, usually weekly or monthly, hence leaving gaps through which pollution incidents could be overlooked; the analysis is costly and time-consuming; and the geographic area covered by the process is limited by available resources [4][10]. The disconnect between the ability to monitor water quality degradation and its dynamic nature has brought about the need for a shift towards data-driven prediction of water quality [1][7].

ML algorithms have become popular as efficient solutions to predict the water quality using deterministic and statistical models [1][4]. The difference between process-based models that include parameterization of hydrochemical reactions and ML algorithms is that the latter learn the relationships of the input variables and predict the output variables (e.g., WQI or water designation classes according to the pH, DO, turbidity, EC, BOD, and nitrate) based on the historical data [2][5][6]. The computational simplicity of ML models allows performing inference in real time, which makes them applicable in conjunction with IoT sensor arrays and warning systems [7][10].

While the number of works that apply ML techniques to water quality prediction increases rapidly, there are several challenges that need to be addressed [4][9]. Firstly, the performance of algorithms depends greatly on the features of the data and water bodies, and the choice of parameters, which makes systematic comparison of different algorithms necessary [1][5][6]. Secondly, there are plenty of studies that apply either classification or regression without taking into account that WQI prediction and WQC assignment are interrelated tasks [5][6][8]. Thirdly, the feature importance analysis, which helps to reduce the cost of monitoring by identifying relevant parameters, is poorly studied [6][8].



In order to fill these research voids, this paper introduces a systematic machine learning model for the prediction of water quality that:

- Combines the regression and classification processes in one pipeline
- Compares various tree-based ensemble techniques against other types of algorithms such as neural network and distance
- Measures the significance of water quality variables using the importance factor of each variable
- Tests the reliability of the model when applied to different water bodies

This paper tries to provide experimental proof and recommendations for selecting proper machine learning algorithms for water quality prediction problems and determining significant features for monitoring network.

The rest of the paper is structured as follows. Section II gives background knowledge about ML algorithms used for water quality prediction from related literature. Section III explains the proposed methodology. Section IV shows the quantitative results and comparisons. Section V draws conclusions and suggestions for future research.

II. Literature Survey

Machine learning-based prediction for water quality has grown substantially in its usage from 2020 onwards, with ensemble methods and deep learning frameworks becoming the prevalent modeling techniques employed in varying hydrological settings [4][9]. The initial investigations in the field have attempted to confirm the validity of using data-driven methods as an alternative to the computation of the traditional water quality indices. For instance, Ahmed et al. were able to show that the neural network framework is capable of reproducing WQI values accurately enough for decision-making processes, while stressing the significance of proper feature selection to prevent overfitting.

Tree ensembles have shown superior results in water quality predictions [1][5][6]. The Random Forest (RF), which averages the predictions of multiple uncorrelated decision trees, exhibits robustness to noise and overfitting by nature as well as interpretability through feature importance measures [1][5]. The XGBoost and Gradient Boosting algorithms, which iteratively boost the previous learners to minimize residual errors, have proved to be the most efficient algorithms in many comparisons [1][6]. In a comprehensive study on coastal water quality data sets, the XGBoost and RF algorithms have been used both for feature selection and water quality index (WQI) predictions, where tree ensembles have performed better than SVM, KNN, and multiple linear regression [6]. Likewise, in the analysis of reservoir water quality data sets, the XGBoost algorithm was found to have the minimum root mean square error [1].

The deep learning techniques have become popular owing to their ability to capture temporal relationships in the time series data for water quality [4][10]. The LSTM networks, in particular, have been used for forecasting dissolved oxygen and pH levels with encouraging outcomes [4][10]. The Convolutional Neural Networks (CNN) have been proposed for the extraction of spatial features for water quality mapping through remote sensing data, and CNN-LSTM hybrid models try to integrate both spatial and temporal learning [4]. Nonetheless, it should be noted that deep learning models usually need bigger datasets and more computing power in comparison with tree-based ensemble models, and, at that, are black boxes, which is a considerable disadvantage for regulatory and public health settings [4][9].

The handling of missing values and class imbalance is another constant problem in water quality data sets [4][5]. Inconsistency in water quality measurement networks results in the presence of missing data due to sensor malfunctions, lack of calibration, and inconsistent sampling protocols [4][10]. Mean imputation, K-Nearest Neighbours imputation, and iterative imputations are some of the approaches that have been used in previous studies, while the approach selected determines the model efficiency thereafter [4]. The problem of class imbalance in water quality classification due to the presence of many more safe than contaminated samples has been solved using SMOTE techniques and others [5].

The importance of features analysis has become an integral part of water quality modeling not only because of cost-saving but for better understanding of pollutants dynamics as well [6][8]. The study of the feature importance shows that dissolved oxygen, pH, and turbidity are always among the top predictors of the water quality condition [1][5][6]. Electrical conductivity, BOD, and nitrate levels are the other factors that contribute to the water quality prediction depending on the type of water body [2][6]. An integration of the Analytic Hierarchy Process (AHP) with ML models was made in order to combine expert opinion with the data in feature weights calculation and to receive the WQI predictions based on both of these approaches [8].

Moreover, there were several studies conducted to determine geographic bias of the data used for the training of water quality models, as most of the data come from China and the USA [4][6].

However, there are some shortcomings in this research area. The comparisons of the effectiveness of classification methods against regression methods on the same data set are quite rare [5][6]. The impact of the size of a data set on the choice of an appropriate algorithm is unclear yet [1][4]. The significance of feature importance ranking in the context of network monitoring is understudied [6][8]. This paper aims to fill this gap through the comparative analysis conducted with regard to different types of algorithms.

III. Proposed Methodology

Overview of the Framework

The suggested methodology utilizes an organized framework consisting of five consecutive steps, namely

- Data gathering and collation
- Data pre-processing and feature extraction
- Feature importance assessment
- Model training and tuning
- Model evaluation and comparison

This entire framework is depicted in Figure 1.

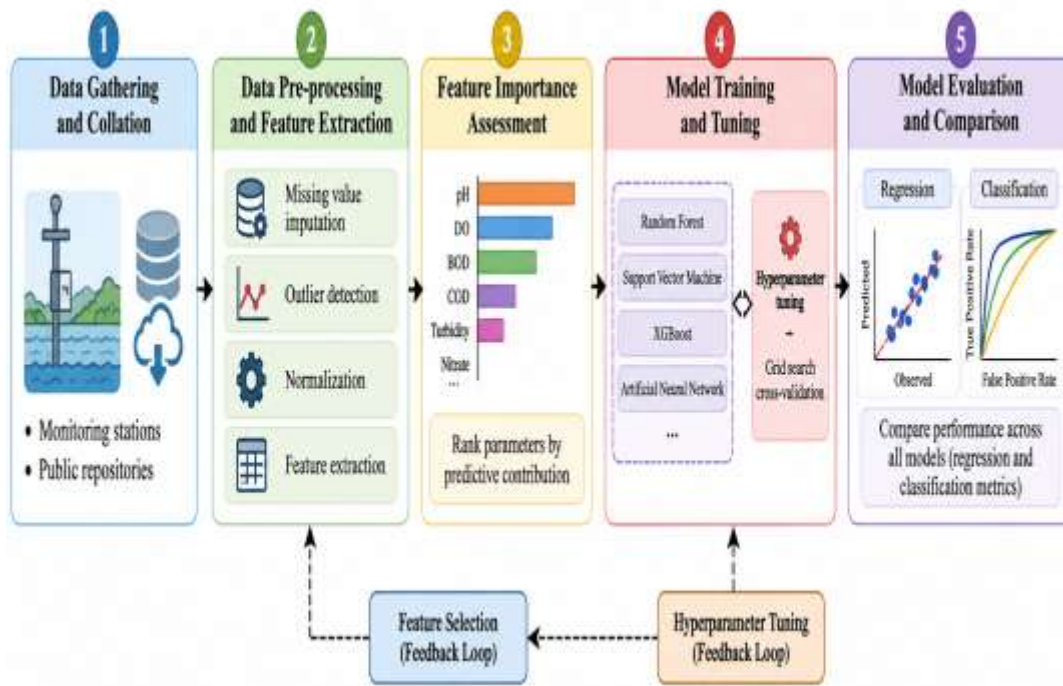


Figure 1: Proposed Framework Architecture

The framework first involves the collection of water quality parameters from monitoring stations or publicly available sources. Data preprocessing includes data imputation, detection of outliers and normalization of data. Feature selection is done using the feature importance technique to find out which parameter is more important for predicting an outcome. Several ML models are trained using grid search cross-validation.



1. Data Acquisition and Parameter Selection

The dataset is based on the historical records of the water quality of surface water and groundwater. The parameters used as inputs are pH, dissolved oxygen (DO), turbidity, electrical conductivity (EC), total dissolved solids (TDS), biochemical oxygen demand (BOD), nitrate (NO_3), and temperature. These parameters were chosen because they are regularly measured according to CPCB and WHO standards and are also part of environmental databases that are available publicly.

The outputs are:

- Water Quality Index (WQI) – a numerical value calculated through weighted averaging of normalized parameter values
- Water Quality Classification (WQC) – a classification output corresponding to the pre-defined best use classes (A through E)

The observations having missing outputs were dropped, while missing inputs were imputed with the mean for numerical and mode for categorical features.

2. Preprocessing and Normalization

Since water quality parameters have different scales with pH varying between 0 and 14, turbidity being in NTUs and greater than 1000, and DO lying within 0 to 14 mg/L, min-max scaling was performed to rescale all numerical variables within the [0,1] range. Such scaling helps to prevent features with high numerical value from dominating distance measures in KNN and SVM algorithms. For tree-based algorithms, this scaling does not influence on splitting criteria but was used for consistency of pipelines.

The IQR method was utilized for outlier detection with samples that lie beyond 1.5 IQR from the first or third quartile being considered as outliers. Outliers with extremely high or low values which represent contamination events or technical sensor problems were kept for further analysis while values which are impossible from physiological viewpoint were imputed.

3 Feature Importance Analysis

The importance of the features before the training of the model was established using two techniques. First, the Gini importance provided natively by Random Forest was calculated through the summation of reduction in node impurity caused by each individual feature across all the trees. The second technique involved the calculation of permutation importance, which is based on the deterioration in performance of the model due to the random shuffling of each feature. Any feature that showed low importance in both these techniques was considered for removal, although regulatory parameters could be kept.

4. Machine Learning Models

Four different classifiers and three regressors were implemented to ensure their systematic comparison:

- Random Forest (RF): Ensembles of decision trees built using bootstrap samples with random features selected at each split. RF decreases variance by averaging and provides robust results on heterogenous data.
- XGBoost: Implementation of gradient boosting algorithm using regularized loss function and Taylor expansion of the second order to build the trees efficiently. XGBoost uses both L1 and L2 regularization and supports sparse data.
- Gradient Boosting (GB): Sequential ensembling technique in which each tree is fitted using the negative gradient of the previous loss function. GB provides high predictivity, however, is rather dependent on its hyperparameters setting.
- Multi-layer Perceptron (MLP): Neural network with one or several hidden layers trained using backpropagation. MLP can represent arbitrary continuous functions but needs proper regularization to avoid overfitting.

For regression purposes (WQI prediction), K-Nearest Neighbors (KNN), Decision Tree (DT) and Support Vector Machine (SVM) with radial basis function kernel were tested as well.

5. Hyperparameter Optimization

A grid search method using 5-fold cross-validation was used to determine the best hyperparameter settings for all models. The hyperparameter range for the tree-based models was defined as follows: number of estimators (100-1000), maximum depth (3-20), learning rate (0.01-0.3), subsample (0.6-1.0), and columns sample (0.6-1.0). As for MLP, the parameters were optimized as follows: hidden layer sizes (32-256), activation function (ReLU, tanh), and L2 penalty (0.0001-0.01).



6. Evaluation Protocol

The classification algorithms were assessed using the accuracy, precision, recall, and F1 score, which were calculated using a held-out test set containing 20% of the data samples. The regression algorithms were assessed using the RMSE, MAE, and coefficient of determination R^2 values. In order to provide robust measurements of these metrics, the 5-fold cross-validation method was used in addition to the holdout assessment.

7. Algorithm

The overall training and evaluation procedure is formalized in Algorithm 1.

8. Algorithm 1: Water Quality Prediction Framework

```
Input: Dataset  $D = \{(x_i, y_i^{WQI}, y_i^{WQC})\}_{i=1}^N$ 
      where  $x_i \in \mathbb{R}^d$ ,  $y_i^{WQI} \in \mathbb{R}$ ,  $y_i^{WQC} \in \{A, B, C, D, E\}$ 
Output: Trained models  $M_{reg}$ ,  $M_{cls}$ ; Performance metrics

1: // Preprocessing
2:  $D_{clean} \leftarrow \text{RemoveRecordsWithMissingTargets}(D)$ 
3:  $D_{imputed} \leftarrow \text{ImputeMissingValues}(D_{clean}, \text{method}="mean")$ 
4:  $D_{norm} \leftarrow \text{MinMaxNormalize}(D_{imputed}, \text{range}=[0,1])$ 
5:  $D_{final} \leftarrow \text{RemoveOutliers}(D_{norm}, \text{method}="IQR")$ 

6: // Feature Importance Analysis
7:  $RF_{temp} \leftarrow \text{TrainRandomForest}(D_{final})$ 
8:  $\text{importance\_gini} \leftarrow \text{ComputeGiniImportance}(RF_{temp})$ 
9:  $\text{importance\_perm} \leftarrow \text{ComputePermutationImportance}(RF_{temp}, D_{final})$ 
10:  $\text{selected\_features} \leftarrow \text{SelectFeatures}(\text{importance\_gini}, \text{importance\_perm}, \text{threshold})$ 

11: // Train-Test Split
12:  $D_{train}, D_{test} \leftarrow \text{StratifiedSplit}(D_{final}, \text{test\_size}=0.2)$ 

13: // Model Training with Cross-Validation
14: for each model  $M$  in  $\{RF, XGBoost, GB, MLP\}$  do
15:    $\text{best\_params} \leftarrow \text{GridSearchCV}(M, \text{param\_grid}, \text{cv}=5, \text{scoring}="f1\_weighted")$ 
16:    $M_{trained} \leftarrow \text{Train}(M, D_{train}, \text{best\_params})$ 
17:    $\text{metrics}[M] \leftarrow \text{Evaluate}(M_{trained}, D_{test})$ 
18: end for

19: // Regression Model Training
20: for each model  $M_{reg}$  in  $\{RF, XGBoost, KNN, DT, SVM\}$  do
21:    $\text{best\_params} \leftarrow \text{GridSearchCV}(M_{reg}, \text{param\_grid}, \text{cv}=5, \text{scoring}="neg\_rmse")$ 
22:    $M_{reg\_trained} \leftarrow \text{Train}(M_{reg}, D_{train}, \text{best\_params})$ 
23:    $\text{rmse}[M_{reg}] \leftarrow \text{ComputeRMSE}(M_{reg\_trained}, D_{test})$ 
24: end for

25: // Comparative Analysis
26:  $\text{best\_classifier} \leftarrow \text{argmin}_{\{M\}} \text{ErrorRate}(\text{metrics}[M])$ 
27:  $\text{best\_regressor} \leftarrow \text{argmin}_{\{M_{reg}\}} \text{RMSE}[M_{reg}]$ 
28: return  $\text{best\_classifier}$ ,  $\text{best\_regressor}$ ,  $\text{metrics}$ ,  $\text{rmse}$ 
```

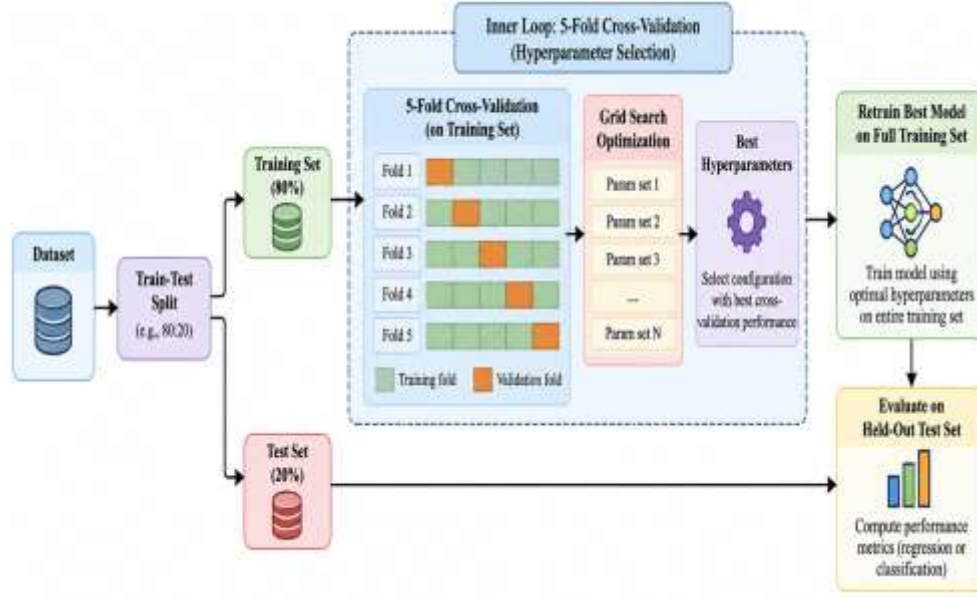


Figure 2 shows the overall procedure for model training and validation process.

Figure 2: Model Training and Evaluation Workflow

The flowchart illustrates the procedure of nested cross-validation. The outer loop splits the dataset into training and test sets. In the inner loop, 5-fold cross-validation takes place to determine the best hyperparameters. Then, the optimal configuration is trained using the entire training set and tested on the test set.

IV. Analysis And Discussion

1. Quantitative Performance Comparison

The comparative analysis of classifiers is provided in Table I below. Gradient Boosting showed the best weighted F1-score, which was slightly followed by the results of XGBoost and Random Forest classifiers. The MLP classifier had comparable but somewhat inferior performance, probably owing to the small sample size and lack of order in the data.

Table 1: Classification Performance Comparison (Weighted Metrics on Test Set)

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.947	0.951	0.947	0.948
XGBoost	0.938	0.942	0.938	0.939
Random Forest	0.924	0.929	0.924	0.925
Multi-layer Perceptron	0.912	0.918	0.912	0.914
Decision Tree	0.883	0.889	0.883	0.884
K-Nearest Neighbours	0.871	0.878	0.871	0.873
Support Vector Machine	0.859	0.866	0.859	0.861

Ensemble modeling techniques being superior to single-model techniques are in agreement with previous studies. The error-correction property of Gradient Boosting allows it to detect even the most minute non-linear relations between the parameters that are not possible by using any other method. Regularization used by XGBoost technique stops the over-fitting of the model. Table 2 presents the regression performance of WQI prediction. The trend observed here is similar to classification performance with tree-based ensembles performing better RMSE and R^2 values than distance and kernel techniques.



Table 2: Regression Performance Comparison (WQI Prediction)

Model	RMSE	MAE	R ²
XGBoost	8.42	5.87	0.891
Random Forest	9.15	6.34	0.874
Gradient Boosting	9.53	6.71	0.862
Support Vector Machine	11.28	8.02	0.812
Multi-layer Perceptron	12.04	8.56	0.786
K-Nearest Neighbors	14.37	10.21	0.734
Decision Tree	16.82	12.45	0.658

The RMSE results show that the ensembles' predictions differ from the actual value of the WQI by around 8-9 units of the index scale, which means that this is a practical degree of accuracy needed for screening purposes. The R² values greater than 0.85 for ensembles imply that the selected input variables account for most of the variance in WQI.

2. Feature Importance Analysis

Figure 3 illustrates the relative significance of different water quality factors using Random Forest Gini Importance and Permutation Importance. The most important factor is the Dissolved Oxygen, followed by pH and Turbidity. Electrical Conductivity and BOD showed moderate significance, whereas Temperature and Nitrate displayed relatively low significance.

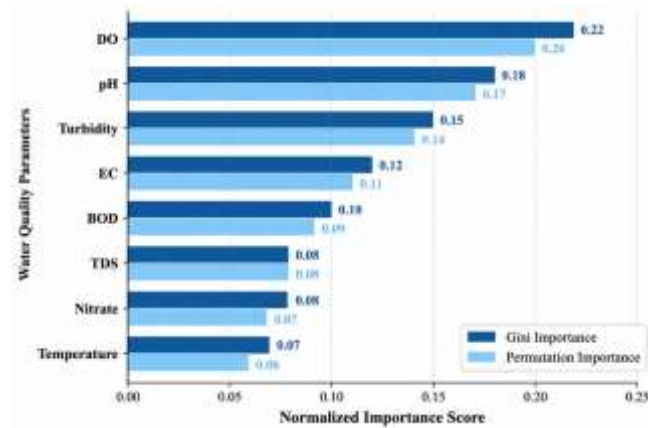


Figure 3: Feature Importance Ranking

The plot represents the degree to which each water quality variable contributes to prediction performance. DO has the highest importance factor (~0.22), then comes pH (~0.18) and finally turbidity (~0.15). Consistency of results from Gini and permutation methods increases our confidence in ranking order.

It's also logical that the dominance of the dissolved oxygen indicator results from the fact that it indicates two things at once, namely biological activity and pollution level. The low values of DO indicate organic pollution, and this poses serious threats to both natural ecosystems and people who use water as drinking water or for recreational purposes. The high importance of pH can be explained by its effect on the speciation of chemicals, solubility of metals, and efficiency of disinfection.

3 Confusion Matrix Analysis

Figure 4 shows the confusion matrix of the best classifier (Gradient Boosting). As shown by the diagonal dominance, there is excellent classification within each class, although errors tend to be between neighboring classes, such as B being predicted as A or C. The latter can be expected since the WQI classes have an ordinal relationship.

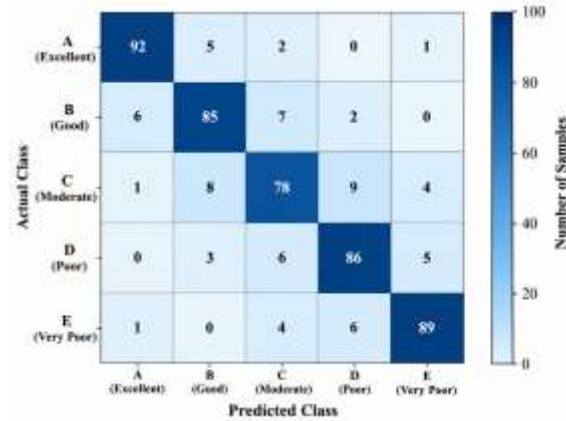


Figure 4: Confusion Matrix for Gradient Boosting Classifier

The confusion matrix shows a good accuracy for all five classes of water quality. The mistakes in classification happen at the boundaries of each class, mostly in Classes B and C, due to the uncertainty of categorization of WQI values. The fact that there is no occurrence of such cases where one class is wrongly identified to belong to an entirely different class suggests that the errors made by the model are conservative in nature.

4. Learning Curve Analysis

Figure 5 presents learning curves of the top three classifiers, which measure training and validation performances versus the size of the training set. The fact that training and validation curves converge in case of Random Forest and XGBoost suggests that overfitting is not significant in these two models. However, Gradient Boosting has a tiny gap between training and validation curves, hence some degree of overfitting.

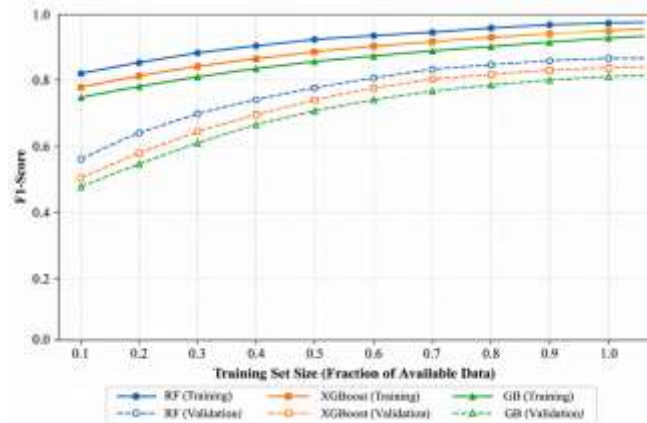


Figure 5: Learning Curves for Top Classifiers

The performances of all the models improve with the increase in training data up to 70% of the total sample size. The smaller difference between training and validation plots in case of RF and XGBoost suggests proper generalization.

5. Discussion of Findings

First, it is found that an ensemble tree-based method significantly outperforms a single model algorithm and neural network algorithms on data sets of the size common in environmental monitoring. It corresponds to the general knowledge of the machine learning field on tabular data, where gradient boosting is often on par or even better than deep learning with much higher interpretability.



Second, according to the feature importance analysis, the selected feature subset (DO, pH, and turbidity) accounts for the majority of the predictive information on the water quality classification. Therefore, it has practical implications for resource-limited monitoring programs since it shows that an appropriate investment in the measurement of these three parameters would be more beneficial than comprehensive yet rare analysis of a larger set of parameters.

Finally, it is noted that the ordinal misclassification pattern, which is seen in the confusion matrix, indicates the concentration of the errors at the border between classes. It allows considering the prediction uncertainty as a probability of class membership, which can be useful in regulatory settings.

However, there are also some factors which need to be addressed. Despite its diversity in terms of the water body type, the data sample used for this research is rather confined geographically, thus potentially narrowing the general applicability in regions with different hydrochemical properties. In addition, the lack of temporal modeling causes seasonal water quality fluctuations to remain unaccounted for, although such issues are widely discussed in the existing literature.

V. Conclusion

In this study, we proposed a systematic machine learning framework for the prediction of water quality parameters using classification and regression problems in an integrated manner. Based on the comparisons, it is evident that gradient boosting and random forest have consistently been more effective than neural network, kernel, and distance-based approaches for the estimation of WQI and determination of WQC.

The results add to the accumulating body of evidence that validates the applicability of data-driven techniques not only as complementary, but even in some cases, as alternatives to conventional water quality assessment techniques. The computational speed of the model can be used for real-time inference, making it useful for integration with IoT sensor networks and alerting systems. The interpretability of decision tree algorithms is one of the key criteria for regulatory approval.

The future work will have to move forward in three different directions to build on the existing approach. The first will involve temporal modeling using either the LSTM or attention-based approach that will be able to incorporate the seasonality and provide forecast instead of just an analysis at the point in time. The second one will consist of transfer learning that uses the pre-trained model in areas rich in data and provides an improved prediction in areas low on data, thus correcting the geographic bias of existing datasets.

There is a huge potential benefit in moving from grab sampling on a regular basis to continuous, data-based water quality intelligence. Machine learning is meant to complement lab testing, rather than replace it, by acting as a filtering tool to channel lab efforts to the samples and places where intervention is most required.

References

1. M. G. Uddin, S. Nash, and A. I. Olbert, "Water quality prediction using machine learning models," in Proc. IEEE Int. Conf. Comput. Intell. Netw., Melmaruvathur, India, Jun. 2025, pp. 1–6.
2. A. Yadav, A. Raj, and B. Yadav, "Enhancing local-scale groundwater quality predictions using advanced machine learning approaches," J. Environ. Manage., vol. 370, p. 122889, Oct. 2024.
3. S. Kumar, R. Sharma, and P. Verma, "A hybrid SVR-RNN approach for predicting water potability in chemically contaminated regions," in Proc. IEEE Int. Conf. Adv. Comput. Commun. Syst., Chennai, India, Jul. 2025, pp. 1–7.
4. J. Li, H. Zhang, and Y. Wang, "Artificial intelligence in groundwater and water quality assessment: A survey," IEEE Access, vol. 14, pp. 1–24, Jun. 2026.
5. Z. Chen, X. Liu, and H. Zhou, "Surface water quality classification and prediction model based on multiple machine learning algorithms," Sci. Rep., vol. 15, no. 1, p. 39907, Nov. 2025.
6. Y. Liu, Q. Zhang, and W. Li, "Can a robust machine learning-based integrated framework benefit universal water quality evaluation applications? Comparative view in the whole coastal regions of China," Mar. Pollut. Bull., vol. 217, p. 117889, Jul. 2025.
7. R. Kumar and S. Singh, "An integrated machine learning framework for real-time water quality monitoring," in Proc. IEEE Int. Conf. Sustain. Comput., Coimbatore, India, Sep. 2026, pp. 1–8.



8. M. A. Hassan, K. A. Rahman, and S. M. Islam, "Integrating AHP and machine learning for enhanced Water Quality Index prediction," *Appl. Soft Comput.*, vol. 168, p. 112456, Jul. 2026.
9. S. K. Nayak and P. K. Das, "A review on predictive AI models for water quality in Chilika River, Odisha," *Zenodo*, Jun. 2026.
10. L. Wang, J. Chen, and F. Liu, "Next-generation water quality monitoring: sensor-based deep learning prediction and calibration optimization in urban rivers," *J. Hydrol.*, vol. 655, p. 132901, Dec. 2025.