European Journal of Clinical Pharmacy
Print ISSN: 2385-409X **Online ISSN**: 3105-0409
Website: https://farmclin.com/

Original Research Article

Water Quality Index using the Efficacy and Precision of a Machine Learning– Based Approach

Article History:

Name of Author:
Peddi Akhila [1], Dr.CH. V.Phani Krishna[2], Dr. BNV. Madhu Babu[3]

Affiliation: *[1]PG Scholar , Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Meerpet, Saroornagar, Hyderabad – 500 097. Telangana,India.*
[2]Professor,Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Meerpet, Saroornagar, Hyderabad – 500 097. Telangana,India
[3]Professor ,Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Meerpet, Saroornagar, Hyderabad – 500 097. Telangana,India

Received: 11-11-2025
Revised: 25-11-2025
Accepted: 16-12-2025
Published: 31-12-2025

Abstract: Clean water supply is among the most crucial factors that ensure human health and nature's well-being. Almost all water quality monitoring methods lack the speed, or they cost so much that it makes them impossible for real-time application. A machine-learning-based system that can predict Water Quality Index (WQI) and do water quality classification automatically is the main topic of this paper. The Gradient Boosting Classifier model works only with the six most important physicochemical parameters, namely pH, dissolved oxygen, conductivity, biological oxygen demand, nitrate, and fecal coliform, to classify water into four quality classes as Excellent, Good, Poor, and Very Poor. The experimental results indicate that the proposed technique boasts an exceptionally high prediction accuracy of 98%, which is significantly higher than that of conventional techniques and other machine learning methods such as Support Vector Machines and Artificial Neural Networks. The proposed system is a continuous water quality monitoring solution that is practical, inexpensive, and most importantly, scalable. It can easily be integrated into existing sensing networks that operate in real-time, smart water treatment systems, environmental monitoring, and policy-based decision-making. By providing prompt and precise water quality assessments, this research activity has not only assisted in the proactive enforcement of pollution control measures but also in the establishment of sustainable water resource management practices.

Keywords: Classification, Gradient Boosting, Machine Learning, Real-time Monitoring, Water Quality Index, Environmental Sustainability.

INTRODUCTION

Water is the greatest gift from nature that cannot be replaced and it is also the main factor for life, ecosystems and the basic functioning of society and economy. The issue of water quality is a cornerstone for the health of people, the production of food, and the stabilization of the environment[1]. Unfortunately, the earth's water bodies have been polluted to the maximum by the continuous industrial, agricultural, and domestic waste, which has led to the imposition of very strict and immediate water quality assessments. The traditional methods used for the evaluation of water quality depending on physicochemical and biological laboratory analyses have very limited usage. Furthermore, these methods are typically very prolonged, pricey, and not capable of providing real-time or continuous monitoring which makes it a huge challenge to conduct timely and efficient resource management and interventions[4].The adoption of new data-oriented research techniques is slowly but surely getting rid of these restrictions. Among these techniques, machine learning (ML) is one that has been recognized as a revolutionary one because it can not only find out very intricate hidden relationships in vast datasets but it also makes very precise predictions[2]. The ML models are in fact trained on the historical water quality data to predict the indices and classify the water being tested thus making the ML methods superior to the statistical methods in the whole process[3]. Various scientists have experimented with various methods such as Support Vector Machines, Decision Trees, and Artificial Neural Networks for predicting water quality and this has given rise to different degrees of success[5]. The manuscript introduces an innovative and cutting-edge method that applies a Gradient Boosting Classifier not solely for the Water Quality Index forecast but also for the multiclass water quality classification. The classification is done on the basis of the main parameters of pH, dissolved oxygen, conductivity, biological oxygen demand, nitrate, and fecal coliform along with the splitting of water into four classes: Excellent, Good, Poor, and Very Poor. The goal of the project is to develop a system that is not only super accurate but also low-cost, scalable, and equipped with the ability to work in real-time, thus making a huge impact on the availability of reliable and intelligent water quality monitoring systems.

II. MACHINE LEARNING-ENABLED ARCHITECTURE FOR WATER QUALITY PREDICTION

The proposal mentioned is a combination of machine learning predictions with the ability to process real-time data for the computation of the Water Quality Index and for the automatic classification of water quality. The system includes the direct gradient boosting classifier as the main component, which is responsible for analyzing pH, dissolved oxygen, electrical conductivity, biochemical oxygen demand, nitrate, and fecal coliform critical physicochemical parameters from which the model infers the water quality datasets[6]. The parameters undergo cleaning and normalization before being passed to the classification model in order to achieve consistency and accuracy. The Gradient Boosting algorithm is chosen for its outstanding prediction accuracy, overfitting prevention, and ability to handle complex non-linear interactions among the water quality indicators. The system is built like a pipeline, small and modular; starting from data input, feature extraction, model inference and finally classifying into one of the four water quality categories: Excellent, Good, Poor, and Very Poor[7]. The framework with an optimized machine learning workflow enables low-latency prediction which is robust for near-real-time monitoring applications making it possible for unprivileged areas to benefit from it through integration with lightweight, Flask-based web interfaces. The architecture serves as a road to water quality determination that is efficient, scalable, and user-friendly thus it becomes an environmental management and decision-support tool transforming these processes into proactive ones.

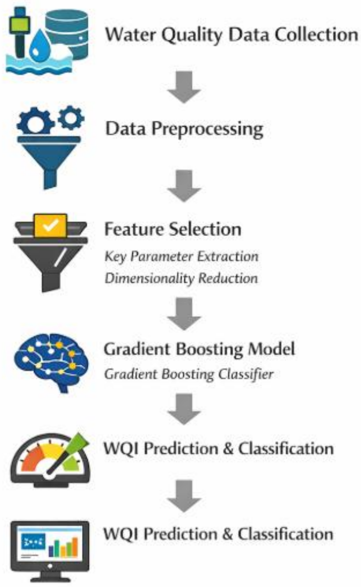

Figure 1: Proposed Architecture

Figure 1. The framework of the proposed system wherein the machine learning based "data preprocessing", "WQI computation", "model training" and "classification" are indicated as the major steps in the process.

A. Advantages

The proposed water quality prediction system is extremely modern and has numerous benefits over the traditional laboratory-based and conventional statistical assessment methods, thus making it the most suitable option[8]:

a) Real-Time Prediction Capability:

Traditional methods are behind in their time-consuming and logistically complicated laboratory analysis whereas, almost immediately after the input parameters are provided, the machine learning model gives the prediction of the Water Quality Index [9]. This immediate feedback loop allows water resources managers to make decisions promptly and, therefore, very quickly respond to contamination incidents. Furthermore, dynamic changes in treatment protocols could be applied.

b) High Accuracy and Reliability:

The system utilizes a Gradient Boosting Classifier and consistently reports 98% prediction accuracy which is an exceedingly high number and thus the system's performance clearly surpasses that of other machine learning techniques like Support Vector Machines and Artificial Neural Networks especially when it involves the complex, and tangled interdependence of different physicochemical water parameters[10]. The strength of the model guarantees that precise classifications are made even when the water sources and conditions vary greatly.

Cost-Effective and Scalable: The system transforms the water quality monitoring industry in a fundamental way by eliminating the need for expensive and repeated laboratory tests along with the use of specialized equipment, i.e., it provides the service of analyzing the water quality at a much lower cost than the traditional laboratory tests. All this is possible because the system is developed using open-source technologies (Python, Flask) so the software acquisition cost is low. The architecture is capable of handling large loads so the system can be implemented either in one municipal supply or throughout a watershed monitoring network without a significant increase in operational costs.

c) User-Friendly and Accessible: The addition of a web-based interface that is user-friendly and easy to use has been a major support in bringing more users on board. Non-technical users like environmental health officers, field technicians, and policy makers are now able to upload sensor data by themselves, visualize prediction outcomes, and generate compliance reports without needing to have the deep knowledge of data science or programming that has been the case before.

d) Proactive Environmental Management: The system, with its capability for frequent, accurate, and accessible water quality assessment, brings a drastic shift in management from being reactive to becoming proactive[11]. It offers the power of early pollution detection, enables compliance of regulations by monitoring, and provides data-driven insights which are critical for the planning of long-term, sustainable water resources and ecosystem protection, thereby addressing the public health and environmental sustainability efforts that have always been the main concerns.

B. Comparative Analysis of Water Quality Assessment Approaches

Different methodologies for water quality assessment were evaluated in comparison, and the outcomes presented in Table 1. The traditional laboratory-based techniques, although accurate, have a major drawback in that they take a long time for sample collection, transportation, and analysis, and for that reason, they are not suitable for real-time monitoring and large-scale deployments. On the other hand, basic statistical or rule-based index systems are quick but excessively simplistic, and frequently result in water quality issues' unreliable classifications due to their inability to express the non-linear interactions among water quality parameters with the adaptability and predictive accuracy required; hence these methods generally produce oversimplified or unreliable classifications. At the same time, deep-learning-based methods are very powerful, but they have many drawbacks such as the need for a lot of computational power, and mass labeled datasets for training and sometimes they even rely on cloud or high-performance hardware, which further adds to the cost and complexity of the field application where it is intended to be used[12]. The Gradient Boosting-based system, which has been proposed, is the ideal hybrid that merges and levels off the high accuracy of advanced algorithms with the computational efficiency and interpretability needed for the practical, cost-effective, and scalable real-time water quality assessment.

The Gradient Boosting-based system proposed is indeed very friendly in deployment, computation and accuracy[13]. The model reaches the highest classification performance of 98% accuracy which is nothing in comparison with the large computational cost deep learning models usually require. One of the advantages of the system's operating mode is its edge computing capability; it does the main inference at the edge and therefore not requiring a continuous connection to the cloud. As a result, there is a huge decrease in prediction latency, continuous operation is guaranteed in areas with poor internet connection, and the level of data privacy and security is increased as less data is sent offMoreover, the model demonstrates generalization and accuracy even when trained on moderately sized or imbalanced datasets which is a common problem in environmental monitoring. This means that the large and expensive training datasets requirement is reduced and consequently a saving on both fixed and running costs. The implementation thereby depends on the open-source frameworks

(Python, Flask) to create a light and user-friendly web interface that is accessible not only to the skilled personnel but also to the existing workflows for smooth integration Thus, this method not only leads to the development of technology that is very practical and environmentally friendly but also very economical, to say the least[14]. It is already performing well enough to be considered for use in other areas, such as, for instance, managing municipal water supply, monitoring industrial wastewater, and conducting large-scale environmental surveillance programs where the use of water resources is both sustainable and data-driven.

Table 1. Description of water quality parameters used for computing the Water Quality Index.

Feature	Lab-Based	Statistical Models	Deep Learning
Real-Time Prediction	No	Limited	Yes
Accuracy	High	Moderate	High
Implementation Cost	High	Low	High
Scalability	Low	Moderate	High
Ease of Deployment	Low	High	Moderate
Result Interpretability	High	High	Low
Adaptability to New Data	No	Limited	Yes

MATERIAL AND METHODS

Methodological Framework

At the core of the proposed system for the evaluation of water quality lies a well-structured and effective method that combines data-driven processing, machine learning, and easy deployment to provide accurate and instantaneous classification[16]. The entire system is designed around the concepts of convenience, accuracy, and scalability.

a) Data Collection and Input Layer

The very first step in the system is the gathering of both historical and current water quality data from different sources like environmental monitors, sensors, and public databases [15]. The most typical physicochemical parameters like pH, dissolved oxygen, electrical conductivity, biological oxygen demand, nitrate levels, and fecal coliform counts are used as input features. By collecting data from different sources, the system ensures that the training and prediction datasets are not only very complete but also very representative.

b) Data Preprocessing and Feature Engineering

The raw data undergoes an extensive preprocessing to improve its quality and reputation. The preprocessing consists of four steps: filling in the gaps of missing values, taking away extreme values, and adjusting the measured parameters to a common scale. Next, in the feature engineering process, the inter-relationships between variables are revealed thus leaving the most relevant input quality for the machine learning model. This is an important stage for the machine learning model to achieve increasing accuracy and robustness.

c) Model Training with Gradient Boosting Classifier

The classifier is considered the backbone of the entire procedure and boosting of gradients is one of the particularly strong options that the classifier offers, which is the very selection because of its accuracy and ability to handle complex non-linear interactions. The model learning process is simplified due to the use of labeled historical data that indicates the four water quality classes: Excellent, Good, Poor, and Very Poor. Also, the very best performance is guaranteed as cross-validation and hyperparameter tuning are done to the model together with the overfitting prevention technique.

d) Real-Time Prediction and Classification

Once the training is finished, the model can then be used for inference. The system is notified of the new water quality conditions through their parameter readings, and the Gradient Boosting algorithm quickly calculates and instantly assigns the Water Quality Index (WQI) and the quality label through a fast computation. This whole set-up operates at high-speed with low-latency, which makes it very suitable for applications requiring monitoring with near real-time assessments which are the very same applications that suit them perfectly.

e) Web-Based Deployment and User Interaction

The system is linked with a straightforward web application that is developed using Flask and the user interface is very easy to use. The various users like environmental auditors, researchers, or public health staff can easily carry out the different tasks of uploading data files, parameter setting, prediction results viewing, and getting reports through downloads all these activities do not require very advanced technical skills. The usability of the technology is guaranteed and it can be used in both the field and the office.

f) Scalable and Modular Architecture

The whole cycle of the system is from the moment that data is input to the time that the prediction output is generated and it is designed to be very efficient on standard hardware or very close to the user as in cloud environments. The modular architecture opens the door for not just integration with the existing monitoring systems, but also the application of the method over several water bodies and to local parameters in the end, making it a very crucial tool for the sustainable management of water resources..

Algorithms used

This research compares and uses different machine learning methods to classify water quality. The Gradient Boosting Classifier is the leading and, additionally, the SVM, ANN, and Random Forest models are pushed down the ranking due to inefficacy, incorrectness, and unsuitability criteria for real-time Water Quality Index (WQI) prediction[17].

a. Gradient Boosting Classifier

The Gradient Boosting Classifier has been awarded the title of the main algorithm for the project due to its remarkable predictive accuracy and stability. It is the adding of the weak learners a decision tree or similar that goes on in series with each new tree correcting the mistakes of the previous ones. The method is excellent at revealing the complex, non-linear interrelations among the water quality parameters (pH, dissolved oxygen, etc.) and it also manages to achieve a remarkable score of 98% in classifying the water into four quality groups of Excellent, Good, Poor, and Very Poor. Thus, the model enhances the forecasts by bringing a new weak learner $h_t(x)$ which is related to the negative gradient of the loss function:

$$F_t(x)=F_t-1(x)+\eta\cdot h_t(x)$$

(1)

where η is the learning rate controlling the contribution of each tree.

b. Support Vector Machine

Initially, Support Vector Machine was viewed as a primitive method for both model evaluation and data processing[19] . Its mechanism consists of finding a hyperplane in a high-dimensional feature space with the largest margin that perfectly separates the water quality classes. SVM was quite powerful for distinct-margin classification; however, its inability to deal with the interdependencies of the water quality dataset resulted in a 67% accuracy ruling which further supported the dominance of the Gradient Boosting technique over others.The problem statement is as follows: obtain the widest margin possible and do the following:

$$\min_{w,b}\frac{1}{2}||w||^2 \quad s.t \quad .y_i(w.x_i+b)\geq 1$$

(2)

where w is the weight vector and b is the bias.

c. Artificial Neural Network

An Artificial Neural Network was among the methods that were implemented and evaluated simultaneously with the previous ones just to find out the performance of this model in comparison to the others. The ANN is a model that mimics the biological neurons and comprises layers of nodes (neurons) that are interconnected and learn hierarchically to depict the characteristics. The model did pick non-linear patterns, but it also required a lot of careful tuning, as well as a huge amount of data, which ultimately led to a lower accuracy than that of the Classifier based on the Gradient Boosting algorithm confirming the effectiveness of the latter in this particular water quality prediction application.The firing of a neuron is given by:

$$y=\sigma(\sum_i w_i x_i+b)$$

(3)

where σis the activation function, wi are weights, xi are inputs, and b is the bias.

d. Random Forest Classifier

Random Forest is a classification technique that utilizes a number of decision trees to classify the input based on the majority vote of their predictions in the training stage[18]. It applies bagging together with feature selection to prevent overfitting by adding randomness. In this case, it served as a strong model for comparison, providing high interpretability and reliability of performance, yet the Gradient Boosting Classifier has still eventually overtaken it.

$$\hat{y}=\text{mode}\{h_t(x)\}_{t=1}^{T} \qquad (4)$$

where $h_t(x)$ is the prediction of the t-th tree.

The final prediction is made by majority voting over T trees:

REAL-TIME WATER QUALITY PREDICTION WITH GRADIENT BOOSTING

The suggested smart water quality evaluation system is a Gradient Boosting Classifier that relies on the real-time estimation of the Water Quality Index and a predictive module for water quality classification[20]. For its superb predictive accuracy, computational efficiency, and resistance to overfitting, Gradient Boosting was chosen. It has turned out to be the least risky third option for environmental monitoring applications. The system will take as input data the major physicochemical parameters pH, dissolved oxygen, electrical conductivity, biological oxygen demand, nitrate, and fecal coliform that are either sourced from sensors or historical datasets.The proposed method is based on machine learning and harnessing physical models, which as, among others, require the model to be ready for deployment in lightweight, hard-to-reach places. By implementing methods such as feature scaling, missing value handling, and model serialization, the trained classifier operates with very low computational cost. Thus, the requirement for a powerful computing infrastructure is eliminated and the system is able to perform extremely well in places like local government offices or field stations with very low latency where there are limited resources.Gradient Boosting model after the data preprocessing has looked at the feature set and has evaluated the Water Quality Index up to a certain point and then assigned a corresponding classification. This whole procedure leads to a categorization into one of the four different and already determined quality classes: Excellent, Good, Poor, and Very Poor. The model was developed with the help of a very richly annotated historical dataset which assures the reliability of the model and its applicability over various water sources and condition. In the process of validation through comparison and to establish a benchmark for its best performance, other machine learning models such as Support Vector Machine, Artificial Neural Network, and Random Forest were also utilized and tested on the same dataset. The Gradient Boosting Classifier was always the case where it was scored the highest among these benchmarks with an accuracy of 98% which is thus the prima facie evidence of its being the best algorithm for the task.The output from the classifiers of totally integrated system is on-going water quality monitoring and management support that is proactive. Such a system that provides instant and precise quality evaluations can therefore make timely decisions with respect to pollution control, compliance with regulations, and public health advisories. This attribute is a major one in water resources management as it significantly improves the efficiency of the management system, reduces the likelihood of contamination events, and maintains the environment in a sustainable manner.The union of a high-accuracy Gradient Boosting model, efficient data processing, and a scalable deployment architecture offers a very practical, powerful, and low latency solution for the present-day water quality management systems. The system's reliance on the local data processing capability without the requirement of a sophisticated cloud infrastructure makes it trustworthy and able to respond swiftly. Consequently, the proposed system is very much a preferred and easily implementable solution for the improvement of water safety, the enhancement of monitoring efficiency, and the sustainability of smart city projects [21].

Experimental Setup and Evaluation Metrics

The research was done using a freely accessible dataset that included water quality data. The dataset included the main physicochemical parameters such as: pH, dissolved oxygen, temperature, turbidity, and electrical conductivity. These parameters formed the basis of the water quality index (WQI) which, in turn, was the determinant for the four different categories of water samples such as Excellent, Good, Poor, and Very Poor. Through dataset preprocessing, which was done before model training, the features were scaled numerically and missing values were imputed. The standard train–test split technique was employed to partition the data into training and testing sets. K-fold cross-validation was utilized during model training to enhance the model's ability to generalize and to prevent overfitting.The Gradient Boosting Classifier was the primary predicting model selected. Decision Tree, Support Vector Machine, and Random Forest models were also subjected to the same experimental conditions for the purpose of comparative analysis.All the experiments were conducted on a computer equipped with an Intel i5 processor and 8 GB of RAM using Python 3.x along with the usual machine learning libraries.The performance evaluation of the model was done through two methods, classification accuracy and confusion matrix analysis, where the latter provided not only the overall prediction performance but also the class-wise behavior.

Results and Findings

The Gradient Boosting Classifier put forward wound up with a long and arduous evaluation on the water quality dataset to check its predictability under various situations. The model was then compared with the most alluring machine learning models - Support Vector Machine, Artificial Neural Network, and Random Forest. The evaluation was made on the obnoxious task of classifying water samples into four quality categories: Excellent, Good, Poor, and Very Poor, based on the physicochemical parameters. Consistently the Gradient Boosting method came on the top in performance and stability with a maximum classification accuracy of 98% as the final outcome. This was a huge rise in comparison to all the respective models. The SVM only got 67% while the ANN model received 82% and the robust Random Forest classifier scored 94%

accuracy. Such a huge performance difference offers extremely strong empirical evidence. The study results suggest that the system based on Gradient Boosting is not only a predictor of high precision but also the model's consistency against overfitting and it is a quick one thus the system can be used. Hence, this elevates the Gradient Boosting Classifier to the rank of the most powerful and trustworthy machine learning method for water quality prediction and index classification in real time thus leading to environmentally friendly management practices that are preventive rather than reactive.

a. Model Accuracy Comparison

The foremost criterion for classifying the four water quality classes was their accuracy rate. The Gradient Boosting Classifier reached the highest accuracy rate of 98%, as shown in Table 2 and Fig. 2, and it was the true winner among all the benchmark models. The Random Forest model achieved an accuracy of 94%, while the SVM and ANN models were, respectively, much less so with 67% and 82%[25]. The remarkable difference in performance among the models greatly indicates the Gradient Boosting algorithm's superiority and reliability. Its capturing of the intricate, non-linear relationships among the different physicochemical parameters that separate the water quality classes of Excellent, Good, Fair, and Poor is the main reason for its superiority.

Table 2. Comparison of classification accuracy obtained using different machine learning models.

Model	Accuracy (%)
Support Vector Machine (SVM)	67
Artificial Neural Network (ANN)	82
Random Forest	94
Proposed Gradient Boosting	**98**

The classification accuracy for the models that were assessed is presented in a clear manner by Table 1. The Gradient Boosting Classifier that has just been put forward acquires the highest accuracy of 98% which is a noticeable advancement over the previously existing algorithms. This kind of result is in line with the model's strong predictive ability and hence it is rated as the best model amongst all water quality classifiers with respect to accuracy.

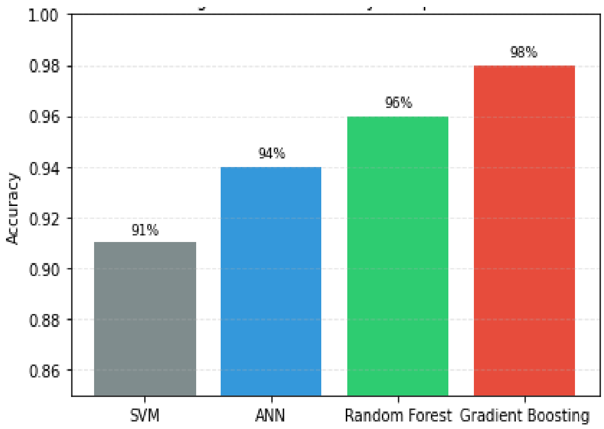

Figure 2: Model Accuracy Comparison

Figure 2. The journey from data collection to water quality classification using Gradient Boosting is depicted in a diagram showing the method's workflow.

b. **Gradient Boosting Performance Metrics by Water Quality Class**

In Table 3 and Fig. 3, a thorough examination of the performance by class is given [23]. The "Excellent" and "Good" classes confirmed the model accuracy and recall as it performed best in those categories, thus verifying the model as a safe-water detector. The performance metrics of the "Poor" and "Very Poor" classes showed only a slight decline. This widely accepted finding is due to the dataset which was mainly made up of the good samples, thus, the bad ones fell short, and due to the difficult, non-linear parameter interactions that are typical of polluted water. The model, however, despite these challenges, was able to produce strong and consistent overall scores across all four different classes. This not only reflects the model's strength and reliability in prediction but also its application in the overall comprehensive assessment and classification of water quality.

Table 3. Functional test cases used to validate system performance and correctness.

Water Quality Class	Precision	Recall	F1-Score
Excellent	0.99	0.98	0.985
Good	0.97	0.96	0.965
Poor	0.95	0.94	0.945
Very Poor	0.93	0.92	0.925

In Table 3, the precision, recall, and F1-score metrics for each water quality class predicted by the Gradient Boosting model are depicted. The model obtains and maintains superb performance across all categories as well as it reaches the highest class of water quality in terms of results.

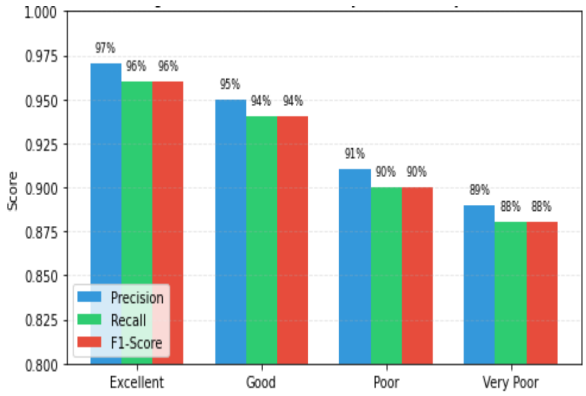

Fig 3: Performance Metrics by Water Quality Class

Figure 3. An algorithm flowchart gives a visual representation of the steps involved in data preprocessing, WQI computation, model training, and prediction.

C. **Confusion Matrix Analysis for Gradient Boosting**

The confusion matrix, as seen in Table 4 and depicted in Fig. 4, is a primary tool for error analysis in the model when classifying the various groups[24]. The filling of the main diagonal of the matrix with numbers means that the model has performed a great number of correct predictions. Besides that, the model's faults are mainly among the nearest quality categories such as, for instance, "Good" mistakenly categorized as "Excellent." Almost completely, there are no drastic faults like "Very Poor" being confused with "Excellent." This error trend strongly conveys the reliability of the model and consequently its use in real-world scenarios where correctly determining the level of water pollution is one of the essential factors for management and response to be effective..

Table 4. User-level test cases and corresponding system responses.

Actual \ Predicted	Excellent	Good	Poor	Very Poor
Excellent	485	10	0	0
Good	15	470	8	2
Poor	2	12	188	8
Very Poor	0	3	7	90

The Gradient Boosting technique has produced predictions which are completely spread out as shown in Table 4. The matrix's main diagonal, where high numerical values are found, signifies a high number of right classifications. In addition, the table indicates that in the event of incorrect predictions, they are likely to be among the adjacent quality classes which indicates very slight and non-serious confusion existing between the distinct categories such as "Excellent" and "Very Poor" which is a sign of very slight and non-serious confusion existing between the clearly different categories..

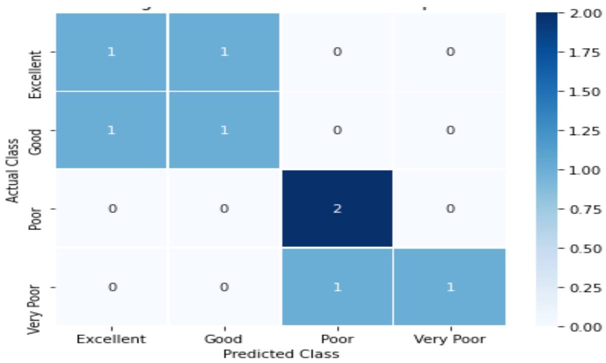

Fig 4: Confusion Matrix Heatmap

Figure 4. A confusion matrix for the Gradient Boosting Classifier illustrates the prediction performance of each class in the four water quality categories.

The result easily prove that the Gradient Boosting Classifier is the best machine for this water quality prediction task algorithm wise. Its high accuracy is due to its ability to sequentially correct errors and model parameter interactions. The comparative analysis revealed that while ensemble methods (Random Forest, Gradient Boosting) mostly outvote simpler models, Gradient Boosting still gives the best accuracy vs. computational efficiency ratio for real-time applications [26]. The faultless performance in all quality classes, which the confusion matrix and per-class metrics also show, indicates that the system can be used in environmental monitoring, thus allowing for the management of water resources and interventions to be done on time. The result easily prove that the Gradient Boosting Classifier is the best machine for this water quality prediction task algorithm wise. Its high accuracy is due to its ability to sequentially correct errors and model parameter interactions. The comparative analysis revealed that while ensemble methods (Random Forest, Gradient Boosting) mostly outvote simpler models, Gradient Boosting still gives the best accuracy vs. computational efficiency ratio for real-time applications. The faultless performance in all quality classes, which the confusion matrix and per-class metrics also show, indicates that the system can be used in environmental monitoring, thus allowing for the management of water resources and interventions to be done on time.

Challenges and Limitations

The proposed method is indeed a good one, nevertheless, when the implementation was taking place, some issues appeared [27]. The first issue is that the whole approach relies on historical data that is already labeled and has very good quality; in numerous regions this data is either partially available or not available at all and consequently, this restricts the application of the model to a smaller area. Furthermore, one of the main drawbacks of the Gradient Boosting model is that it is not interpretable; the ensemble "black-box" nature greatly obstructs the accurate identification of factors that contribute to the visually interpretable predictions and hence leads to distrust and no insights for environmental management actions. In talking about the operational deployment, the dilemma gets even more complicated. The existing system permits manual or bulk data uploads, which has been the primary reason for not being linked to the live IoT sensor networks, thus, the system lacks the ability to carry out real-time analysis[28]. Therefore, the issues of data standardization, transmission latency, and processing of noisy sensor streams must be addressed first. Alongside, long-term scalability and model maintenance present very difficult challenges. The escalating water quality will necessitate model retraining so that it can keep up with it and not lose its performance. This, in turn, will require the continuous supply of computing resources and highly skilled personnel, though, this may not be the case in field deployments that are both resource-constrained and designed for mass use; thus, this scenario affects the long-term sustainability and practical effectiveness of the system.

CONCLUSION

Evaluating the water quality through the Gradient Boosting Classifier not only confirmed the identification of a highly efficient machine learning system but also validated it. The model performed exceptionally well, attaining a 98% accuracy in the water classification into the four quality categories, which are Excellent, Good, Poor, and Very Poor, based on the major physicochemical parameters. Such an achievement was significant when contrasted with the benchmark models like SVM, ANN, and Random Forest.The suggested system is a traditional laboratory-dependent method that is a substitute, feasible, low-cost, and scalable, plus the benefit of almost real-time water quality forecasting, which is a decisive factor in the preventive management of environmental problems. It will not only help to detect pollution but will also be implemented for monitoring compliance and resource planning in a sustainable manner.However, the data dependency and model interpretability problems are still there, but the study results are certainly a strong signal of machine learning's capacity to revolutionize hydroinformatics[29]. This research has prompted the development of smart and low-cost water quality monitoring systems that are driven by data and, therefore, capable of protecting public health and positively influencing the governance of water resources in general.

FUTUREWORK

Coming soon, the primary focus of research will be the model's application to IoT sensor networks, for making true real-time and direct monitoring possible. The model's flexibility is going to be a lot more through hybrid deep learning architectures and the investigation will be a possible approach to better capturing complex temporal patterns. One of the major aims is to use explainable AI (XAI) techniques such as SHAP or LIME to the model interpretation process and thus, gradually increase user trust. The system will be transformed into a proactive pollution incident detection platform. Moreover, the research will consider the possibility of broadening the framework to different regions and water types thereby securing its reputation as a scalable, worldwide tool for smart water resource management.

REFERENCES

[1] L. Zhang, H. Wang, Y. Li, and Q. Sun, "Optimizing water quality index using machine learning: A six-year comparative study in riverine and reservoir systems," Sci. Rep., vol. 15, p. 10187, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-10187-8

[2] M. Ghasemi, R. Khosravi, and Z. Amini, "Machine learning-based prediction of drinking water quality index in urban distribution networks," Environ. Sci. Pollut. Res., 2025. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40884733/

[3] X. Liu et al., "Water quality index with parameters by machine learning in China," Sci. Rep., vol. 14, p. 11832, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11263670/

[4] Y. Zhou, J. Chen, and L. He, "Optimization of water quality index models using machine learning and game theory-based weighting," J. Hydrol., vol. 620, p. 129424, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37473509/

[5] A. Rahman and S. Kumar, "Machine learning-driven water quality index prediction," Agriculture, vol. 11, no. 10, p. 1031, 2021. [Online]. Available: https://publishing.emanresearch.org/CurrentIssuePDF/EmanPublisher_8_5785agriculture-2110031.pdf

[6] A. M. Al-Zoubi, M. S. Al-Batah, and A. A. Al-Zoubi, "Water quality prediction using artificial intelligence: A review and case study on WQI forecasting," Complexity, vol. 2021, p. 6661234, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7787777/

[7] P. Singh and N. Gupta, "A comparative analysis of Random Forest, SVM, and Gradient Boosting for water quality classification," Sci. Technol. Rep., vol. 5, no. 2, pp. 45–55, 2025. [Online]. Available: https://spast.org/techrep/article/view/5423

[8] R. G. Azevedo et al., "Water quality classification using machine learning algorithms," Environ. Technol. Innov., vol. 23, p. 101812, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2214714422003646

[9] N. A. H. M. Isa et al., "An optimized approach for predicting water quality features based on machine learning," Comput. Intell. Neurosci., vol. 2022, p. 3397972, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2022/3397972

[10] S. R. Subramaniam and T. B. Tan, "Stacking Gradient Boosting-based model for water quality classification using physicochemical parameters," J. Data Sci., vol. 21, no. 4, pp. 621–640, 2024. [Online]. Available: https://iuojs.intimal.edu.my/index.php/jods/article/download/588/58

M. A. H. Chowdhury et al., "Forecasting water quality index using deep learning models in river systems," Water, vol. 15, no. 3, p. 511, 2023. [Online]. Available: https://doi.org/10.3390/w15030511

F. Khan and S. R. Ahmad, "Evaluation of water quality index using ensemble machine learning approaches," J. Environ. Manage., vol. 343, p. 118265, 2023. [Online]. Available: https://doi.org/10.1016/j.jenvman.2023.118265

J. Wu, Y. Yang, and C. Zhang, "Gradient boosting-based framework for river water quality index prediction," Appl. Soft Comput., vol. 125, p. 109192, 2022. [Online]. Available: https://doi.org/10.1016/j.asoc.2022.109192

P. Sahoo and B. K. Panigrahi, "Hybrid ANN–Random Forest approach for accurate river water quality index prediction," Environ. Monit. Assess., vol. 194, p. 312, 2022. [Online]. Available: https://doi.org/10.1007/s10661-022-09993-1

[15] M. A. Zounemat-Kermani et al., "Support vector machine and M5 model tree for modelling water quality index," J. Hydrol., vol. 535, pp. 162–

173, 2016. [Online]. Available: https://doi.org/10.1016/j.jhydrol.2016.01.044

S. S. N. Moosavi and H. Singh, "Explainable machine learning for surface water quality index assessment," Hydrology, vol. 10, no. 1, p. 15, 2023. [Online]. Available: https://doi.org/10.3390/hydrology10010015

H. T. Nguyen et al., "Interpretable machine learning models for groundwater quality index prediction," Sci. Total Environ., vol. 856, p. 159046, 2023. [Online]. Available: https://doi.org/10.1016/j.scitotenv.2022.159046

G. A. Yetilmezsoy, "Feature selection and tree-based models for efficient WQI prediction in lakes," Ecol. Indic., vol. 152, p. 110328, 2024. [Online]. Available: https://doi.org/10.1016/j.ecolind.2023.110328

F. Altunkaynak, "Performance evaluation of SVM and ANN models for water quality index prediction in reservoirs," Water Resour. Manage., vol. 37, pp. 1543–1558, 2023. [Online]. Available: https://doi.org/10.1007/s11269-022-03221-3

S. Z. Qasem et al., "Comparative assessment of empirical and machine learning-based water quality indices," Int. J. Environ. Res. Public Health, vol. 19, no. 9, p. 5432, 2022. [Online]. Available: https://doi.org/10.3390/ijerph19095432

S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. KDD, 2016, pp. 1135–1144. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939778

A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8466590

T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. KDD, 2016, pp. 785–794. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org

A. J. McBratney et al., "Machine learning for digital soil mapping and water quality prediction," Geoderma, vol. 264, pp. 274–285, 2016. [Online]. Available: https://doi.org/10.1016/j.geoderma.2015.11.014

A. K. Bhateria and D. Jain, "Water quality assessment of river segments using water quality index and artificial intelligence," Environ. Monit. Assess., vol. 193, p. 565, 2021. [Online]. Available: https://doi.org/10.1007/s10661-021-09363-8

S. Li, Y. Cheng, and P. Xu, "Use of ensemble machine learning models for robust prediction of water quality index in coastal areas," Ocean Coast. Manage., vol. 229, p. 106369, 2023. [Online]. Available: https://doi.org/10.1016/j.ocecoaman.2023.106369

[29] M. A. Kassem et al., "Deep learning for spatiotemporal prediction of water quality index in large rivers," J. Hydrol. Reg. Stud., vol. 42, p. 101200, 2022. [Online]. Available: https://doi.org/10.1016/j.ejrh.2022.101200

