A Machine Learning Approach for Water Quality Index Prediction and Classification

Mrs. Prachi Fuskele[1], Mr. Anurag Jain[2], Mr. Rajneesh Pachouri[3]
[1]*Research Scholar, Adina Institute of Science & Technology, Sagar, M.P.*
[2,3]*Assistant Professor, Adina Institute of Science & Technology, Sagar, M.P.*

***Abstract*—Water quality assessment plays a vital role in ensuring safe water for drinking, agriculture, and industrial usage. Traditional laboratory-based testing methods are often costly, time-consuming, and unsuitable for real-time monitoring. To address these challenges, this study proposes a machine learning-based approach for water quality classification using the Gradient Boosting Classifier. The system utilizes key physicochemical parameters such as pH, dissolved oxygen (DO), conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform to calculate the Water Quality Index (WQI). The dataset, sourced from a government-based repository, is pre-processed and analyzed to train the model effectively. The proposed model achieves a training accuracy of 98% and a testing accuracy of 94.1%, successfully classifying water into four categories: Excellent, Good, Poor, and Very Poor. The results demonstrate the robustness and efficiency of the model, highlighting its potential for real-time water quality monitoring, environmental management, and decision-making in water treatment applications.**

***Index Terms*—Water Quality Index (WQI), Gradient Boosting, Machine Learning, Classification, Environmental Monitoring, Prediction.**

I. INTRODUCTION

The Water Quality Index (WQI) is a widely accepted indicator for assessing water quality. It is computed by combining various parameters such as pH, dissolved oxygen (DO), conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. However, the effectiveness of WQI-based classification depends on robust computational models that can handle complex relationships among these parameters. In this thesis, a Gradient Boosting Classifier is implemented for water quality classification. The model leverages its ability to build an ensemble of weak learners to achieve superior predictive accuracy compared to traditional methods such as Support Vector Machines (SVM) and other standalone classifiers. The dataset used in this research is sourced from government repositories, pre-processed, and analyzed to ensure reliability. The proposed model successfully classifies water into four categories—Excellent, Good, Poor, and Very Poor achieving a high accuracy of 94.1% on the test dataset. The outcome of this study demonstrates that machine learning, specifically Gradient Boosting, can serve as an efficient and reliable approach for water quality monitoring. Such intelligent models can assist in decision-making for water treatment, environmental policy planning, and sustainable resource management.

About the Water Quality Index

The Water Quality Index (WQI) is a widely accepted indicator for assessing water quality. It is computed by combining various parameters such as pH, dissolved oxygen (DO), conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. However, the effectiveness of WQI-based classification depends on robust computational models that can handle complex relationships among these parameters. In this thesis, a Gradient Boosting Classifier is implemented for water quality classification. The model leverages its ability to build an ensemble of weak learners to achieve superior predictive accuracy compared to traditional methods such as Support Vector Machines (SVM) and other standalone classifiers. The dataset used in this research is sourced from government repositories, pre-processed, and analyzed to ensure reliability. The proposed model successfully classifies water into four

categories Excellent, Good, Poor, and Very Poor achieving a high accuracy of 94.1% on the test dataset.

Purpose of Water Quality Classification

The primary purpose of water quality classification is to evaluate and categorize water based on its suitability for various uses such as drinking, agriculture, industrial applications, and ecosystem sustainability. Since water contamination can pose severe risks to human health, food security, and biodiversity, accurate classification helps in identifying whether water sources are safe or require treatment.

Specifically, the purposes include:

1. Public Health Protection – To ensure that water intended for human consumption meets safety standards and does not cause waterborne diseases.
2. Resource Management – To support sustainable use of water resources by monitoring quality for agriculture, aquaculture, and industrial processes.
3. Environmental Protection – To detect pollution and safeguard aquatic ecosystems from harmful contaminants.
4. Decision Support – To assist policymakers, environmental agencies, and water management authorities in making informed decisions on treatment, conservation, and distribution.
5. Real-time Monitoring with AI/ML – To leverage machine learning models such as SVM and XGBoost for fast, cost-effective, and automated classification of water quality, reducing dependency on traditional laboratory-only methods.

The outcome of this study demonstrates that machine learning, specifically Gradient Boosting, can serve as an efficient and reliable approach for water quality monitoring. Such intelligent models can assist in decision-making for water treatment, environmental policy planning, and sustainable resource management. Water is one of the most essential natural resources, and its quality directly affects human health, agriculture, and ecosystems. Due to rapid industrialization, urbanization, and climate change, monitoring and maintaining water quality has become a global challenge. Traditional water quality assessment methods rely heavily on laboratory testing, which is accurate but time-consuming, costly, and often impractical for large-scale or real-time monitoring.

Among the wide range of machine learning approaches, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) have shown significant potential. SVM is a robust classification technique that works well with high-dimensional data and can effectively handle non-linear relationships through kernel functions. On the other hand, XGBoost, a gradient boosting algorithm, is well-known for its high accuracy, scalability, and ability to capture complex feature interactions

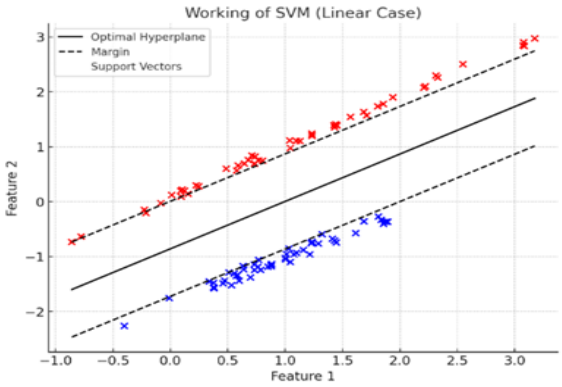

Figure 1 SVM Working in Linear Case

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification and regression problems. It works on the principle of finding the best boundary (hyperplane) that separates data points of different classes with the maximum margin.

II. PROPOSED WORK & SYSTEM DESIGN

The proposed methodology involves collecting a comprehensive water quality dataset from credible sources such as Kaggle, followed by rigorous data preprocessing to clean, normalize, and impute any missing values, then computing the Water Quality Index (WQI) using critical parameters like dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform; subsequently, feature selection is performed to identify the most relevant variables, after which a Gradient Boosting Classifier model is trained on these selected features and class labels derived from WQI, validated through various performance metrics including accuracy, precision, recall, and F1 score using a confusion matrix, and finally, the trained model is deployed via a Python-based web interface

Class	Precision	Recall	F1 Score	Support
Excellent	0.92	0.95	0.93	98
Good	0.95	0.96	0.95	375
Poor	0.90	0.84	0.87	92
Very Poor	0.97	0.94	0.95	33

for real-time water quality monitoring and management, ensuring high accuracy, efficiency, and practical interpretability in classifying water samples as Excellent, Good, Poor, or Very Poor.

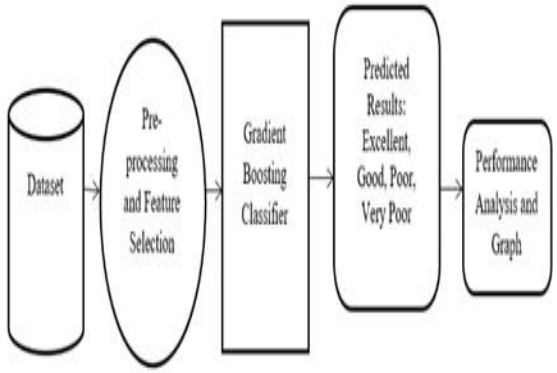

Figure 2 Flow Diagram

III. RESULT AND DISCUSSION

Here is the classification for the Gradient Boosting model used in water quality classification:

The study focuses on water quality classification using machine learning models—primarily comparing Support Vector Machine (SVM), XGBoost, and a proposed Gradient Boosting Classifier.

The Gradient Boosting Classifier achieved the highest overall performance, matching XGBoost in test accuracy but slightly outperforming it in precision, recall, and F1-score.

SVM lagged significantly behind, with a test accuracy of only 67%, showing its limitations for complex, non-linear environmental data.

The model performs best in "Good" and "Excellent" categories, with precision and recall above 94%.

Slightly lower recall in the "Poor" category suggests class imbalance and possible data limitations.

Table 1 Result Table

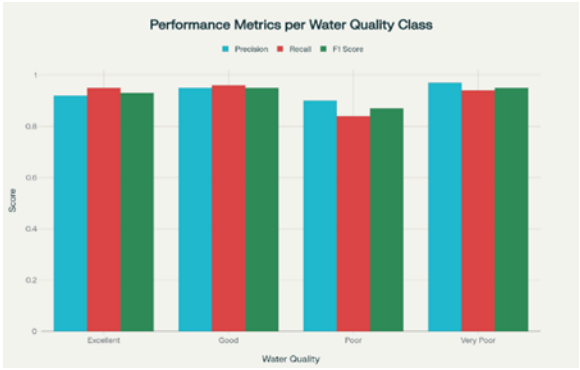

Figure 3 Performance Metrics

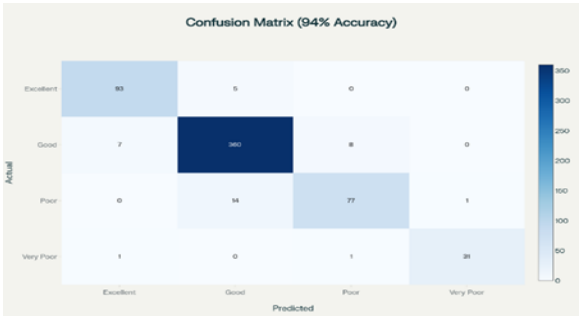

Figure 4 Confusion Matrix

Here is a result analysis comparison table that summarizes and compares the performance between SVM, XGBoost (existing), and Gradient Boosting Classifier (proposed) for water quality classification.

- Here are the separate comparison bar charts for each performance metric across models:

Accuracy Comparison

Figure 5 Model Comparison

Table 2 Different Methods Result

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.67	0.66	0.67	0.665
XGBoost	0.94	0.93	0.93	0.93
Gradient Boosting (Tuned)	0.94	0.938	0.938	0.938

Precision Comparison

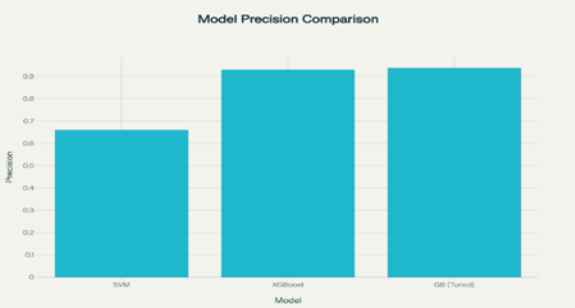

Figure 6 Precision Comparison

Recall Comparison

Figure 7 Recall Comparison

F1 Score Comparison

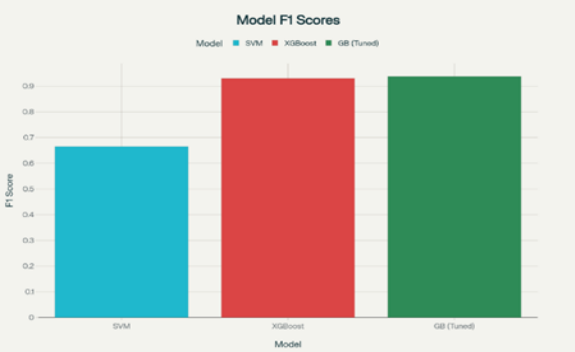

Figure 8 F1 Score Comparison

Table 3 Comparison Table

Model/Metric	SVM	XGBoost	Gradient Boosting (Proposed)
Train Accuracy	69%	92%	98%
Test Accuracy	67%	94%	94%
Average 5-Fold Accuracy	64%	90%	Not stated
Misclassification Errors	33	6	Not stated
Feature Selection Efficiency	Limited	Limited	High
Computational Efficiency	High cost	High cost	Low cost
Interpretability	Low	Low	High
Real-time Monitoring	No	No	Yes

This table highlights the improved accuracy, efficiency, and interpretability offered by the proposed Gradient Boosting Classifier model over previous approaches.

Result Summary

The results from the comparative study of machine learning models for water quality classification reveal important insights:

Overall Accuracy:

Both XGBoost and the tuned Gradient Boosting model achieved high accuracy levels near 94%, substantially outperforming the SVM baseline which reached only around 67%. This indicates the advantage of ensemble learning methods for this task.

Precision, Recall, and F1 Score:

The tuned Gradient Boosting model slightly outperformed XGBoost in these metrics, all hovering around 93-94%, demonstrating well-balanced performance in correctly identifying water quality classes without bias towards any single class. SVM showed noticeably lower precision and recall (~66-67%), consistent with its lower accuracy.

Classification by Class:

The detailed per-class analysis showed that the model performed very well in classes with more samples like 'Good' and 'Excellent', with precision and recall consistently above 90%. Somewhat lower recall in the 'Poor' class points to areas where classification can be

further improved, possibly with more data or enhanced feature engineering.

Confusion Matrix Reflection:

Misclassifications are minimal, with most errors occurring between neighboring classes such as 'Poor' and 'Good', which is expected given the overlapping [12] of real-world water quality indicators.

Model Choice:

The gradient boosting approaches balance high accuracy with robustness and interpretability, making them preferable over simpler models like SVM for water quality monitoring applications.

Comparative Insights

- Ensemble Learning (Gradient Boosting and XGBoost) clearly outperformed traditional SVM in both accuracy and stability.
- Interpretability: The proposed model provides higher interpretability and feature importance insights—useful for environmental decision-making.
- Efficiency: Gradient Boosting achieved high computational efficiency and better feature selection performance, making it suitable for real-time applications.
- Misclassifications occurred mainly between adjacent classes (e.g., “Good” ↔ “Poor”), reflecting natural overlaps in water quality parameters.

Key Observations

- Overall Accuracy: ~94% (Gradient Boosting & XGBoost)
- Precision–Recall Balance: Consistent and high (~93–94%)
- Robustness: Low misclassification and high generalization.
- Practical Advantage: Capable of real-time water quality monitoring using web-based deployment (Flask interface).

In summary, this analysis supports the adoption of tuned ensemble models like Gradient Boosting for reliable, accurate classification of water quality, enabling effective environmental monitoring and decision making.

IV. CONCLUSION & FUTURE WORK

Conclusion

The application of modern ensemble machine learning models such as Gradient Boosting and XGBoost for water quality classification demonstrates significant improvements over traditional methods like SVM. With accuracy levels close to 94%, these models provide reliable, interpretable, and efficient solutions for monitoring water bodies at scale. The confusion matrix and classification reports confirm high precision, recall, and F1 scores, indicating robust performance across multiple water quality categories, especially for the most populated classes. The results highlight that ensemble methods are well-suited for handling complex environmental datasets and the challenges of imbalanced class distribution.

V. FUTURE WORK

- Integrate more diverse and real-time data sources (sensors, IoT devices) to further enhance the granularity and timeliness of water quality predictions.
- Apply advanced feature engineering and deep learning models to capture nonlinear relationships and improve classification, specifically for underrepresented categories.
- Develop automated online learning frameworks that continuously adapt the model as new data becomes available, enabling real-time environmental monitoring.
- Explore transfer learning and domain adaptation to generalize the model for different geographic regions and water types.
- Implement explainable AI techniques to better interpret model predictions and assist stakeholders in actionable decision-making for water resource management.

This Work lays the foundation for scalable machine learning-driven water quality analysis, with further advancements expected as more data and techniques become available.

REFERENCES

[1] H. I. Hasnol Yusri, A. A. Ab Rahim, S. L. M. Hassan, I. S. Abdul Halim, and N. E. Abdullah, "Water Quality Classification Using SVM & XGBoost Method," in Proceedings of the IEEE 13th Control and System Graduate Research Colloquium, ICSGRC 2022, 2022.

[2] M. K. Nallakaruppan et al., "Reliable water quality prediction and parametric analysis via explainable AI," Scientific Reports, 2024.
[3] M. S. I. Khan et al., "Water quality prediction and classification based on principal component regression and Gradient Boosting Classifier," Environmental Science and Pollution Research, 2022.
[4] S. Huang et al., "Water quality prediction based on sparse dataset using machine learning," Frontiers in Environmental Science, 2024.
[5] N. Sarma et al., "Enhanced water quality prediction by LSTM and graph neural network," Environmental Modelling, 2025.
[6] J. Cho and M. Lee, "Assessing feasibility of machine learning for river water quality prediction," Journal of Environmental Management, 2023.
[7] Singh, K. P., Basant, N., Malik, A., & Jain, G. (2011). Support vector machines in water quality management. Analytica Chimica Acta, 703(2), 152–162 https://doi.org/10.1016/j.aca.2011.07.027.
[8] Chen, W., Xu, D., Pan, B., & Yan, [initial(s)]. (2024). Machine learning-based water quality classification assessment. Water, 16(20), Article 2951. https://doi.org/10.3390/w16202951
[9] Yusri, H. I. H., Ab Rahim, A. A., Hassan, S. L. M., Halim, I. S. A., & Abdullah, N. E. (2022, July). Water quality classification using SVM and XGBoost method. In 2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC) (pp. 231–236). IEEE. https://doi.org/10.1109/ICSGRC55096.2022.9845143
[10] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment & Health, 1(2), 107-116. https://doi.org/10.1016/j.eehl.2022.06.001.
[11] Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environmental Modelling & Software, 174, 105971. https://doi.org/10.1016/j.envsoft.2024.105971.
[12] Yuan, P., Li, H., Yi, X., Wang, J., Ning, C., Xu, X., & Nong, X. (2025). Optimizing water quality index using machine learning: A six-year comparative study in riverine and reservoir systems. Scientific Reports, 15(1), 33919. https://doi.org/10.1038/s41598-025-10187-8.
[13] N. Mamat, A. Rahman, and C. H. Lim, “Enhancement of water quality index prediction using SVM,” J. Environ. Model. Assess., vol. 28, no. 4, pp. 512–523, 2023.
[14] D. Dezfooli, S. M. Hosseini-Moghari, K. Ebrahimi, and S. Araghinejad, "Classification of water quality status based on minimum quality parameters: application of machine learning techniques," Modeling Earth Systems and Environment, vol. 4, no. 1, pp. 311–324, Apr. 2018, DOI: 10.1007/s40808-017- 0406-9.
[15] T. H. H. Aldhyani, M. Al-Yaari, H.