A Hybrid Approach to Water Quality Classification Using SVM and Xgboost Method

Akash B. Koli, Beldar Faijan Shaikh Akil, Lohar Bhavesh Kantilal, Pawar Darshan Madhukar, Patil Rushikesh Sanjay

Department of Computer Engineering, D.N. Patel College of Engineering, Shahada, Dist. Nandurbar, Maharashtra, India

DOI: https://doi.org/10.51244/IJRSI.2025.12060080

Received: 22 June 2025; Accepted: 26 June 2025; Published: 09 July 2025

ABSTRACT

This project focuses on the classification of waterquality using machine learning methods—Support Vector Machine (SVM) and XGBoost. The system uses various chemical indicators like pH, dissolved oxygen, turbidity, and conductivity to predict the water quality status. The dataset is preprocessed and important features are extracted before being passed into the models. After evaluating multiple models, XGBoost showed higher accuracy and robustness compared to SVM. The system aims to help environmental authorities monitor and improve water resources more effectively.

Keywords: Water Quality, Machine Learning, SVM, XGBoost, Classification

INTRODUCTION

Water is one of the most critical natural resources that plays a vital role in supporting life on earth. It is used for a wide range of purposes, including drinking, irrigation, industrial uses, and aquatic life maintenance. However, the quality of water is often compromised due to various pollutants, which can negatively impact human health and the ecosystem. Hence, monitoring and managing water quality is of utmost importance.

Traditionally, water quality assessment is performed through expensive laboratory tests, which are not practical for real-time monitoring. Moreover, conventional methods lack accuracy and require a considerable amount of time and effort to process data. Therefore, there is a need for an efficient and cost-effective approach to monitor water quality in real-time.

In recent years, machine learning techniques have emerged as a promising solution for various environmental applications, including water quality monitoring. In this project, we propose a novel approach that utilizes the advantages of machine learning techniques to predict water quality index and water quality class. The proposed method aims to provide an accurate and efficient solution for realtime water quality monitoring and management.

This project focuses on developing a model that can predict water quality class based on various water quality parameters, including pH, dissolved oxygen, temperature, and electrical conductivity. The proposed approach uses Gradient Boosting Classifier to predict water quality as Excellent, Good, Poor, and Very Poor. The accuracy and effectiveness of the proposed approach are demonstrated through a comprehensive evaluation and analysis of the model's performance.

LITERATURE SURVEY

Sharma, K., Mishra, A., & Verma, R. (2020) : This study investigates the use of various machine learning techniques to classify drinking water quality. The authors experimented with models like Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN) using a dataset containing multiple physicochemical features. The results revealed that SVM outperformed other models with higher accuracy and better handling of nonlinear relationships. The study highlights the importance of preprocessing steps like feature scaling and dimensionality reduction to enhance performance. [1]

Gupta, M. & Patil, S. (2021): In this research, ensemble learning techniques—Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost)—were compared for water quality classification. The study emphasized the efficiency of ensemble models in handling noisy and incomplete environmental data. Among the tested models, XGBoost delivered the highest accuracy and generalization due to its regularization capabilities. The authors concluded that ensemble-based models are well-suited for real-world environmental classification tasks. [2]

Aditya Kadiwal (Kaggle Dataset Publisher, 2022): This dataset-centric study provides a practical platform for building water quality prediction systems. The dataset includes 10 essential features such as hardness, solids, sulphates, and chloramines. Several researchers have used this dataset to test different classifiers including Logistic Regression, SVM, and Random Forest. The importance of preprocessing—null value treatment, correlation filtering, and feature scaling—is highlighted as a critical step for improving classification performance. The dataset has since become a benchmark in water quality ML research. [3]

Singh, R., Kumar, A., & Das, P. (2019): This review paper explores the evolution of AI-driven approaches in the domain of water quality assessment. It contrasts traditional chemical-based lab tests with modern predictive techniques using Artificial Intelligence (AI) and Machine Learning (ML). Techniques such as Artificial Neural Networks (ANN), SVM, and hybrid models are analyzed. The study finds that AI models significantly reduce operational time and enable real-time monitoring, particularly useful in smart city and IoT-integrated systems. [4]

Rahman, A., Zhang, L., & Ahmed, N. (2022): This paper focuses on the use of the XGBoost algorithm in environmental dataset classification, including both air and water quality datasets. XGBoost's performance is evaluated in terms of its scalability, speed, and ability to handle missing values. The authors demonstrate that XGBoost not only achieves superior accuracy but also provides interpretability through feature importance analysis. This makes it suitable for environmental policy-making and monitoring systems where explainability is essential. [5]

Bharti, A., & Mishra, S. (2021): This study compares the performance of Logistic Regression, Decision Tree, and Random Forest classifiers in predicting water potability. Using a real-world dataset, the authors observed that Random Forest achieved the highest accuracy due to its ability to handle large feature sets and reduce variance. The paper also discusses the role of synthetic data balancing techniques like SMOTE to address class imbalance issues. The study suggests that combining preprocessing with ensemble learning enhances overall performance in potability prediction tasks. [6]

OBJECTIVE

To develop a machine learning-based system capable of accurately classifying water quality by analyzing key physicochemical parameters such as pH, turbidity, dissolved oxygen, and conductivity.

To compare the performance of multiple classification algorithms Particularly Support Vector Machine (SVM) and XGBoost, in order to determine the most effective model for water quality prediction.

To extract and preprocess relevant water quality features by handling missing values, normalizing data, and selecting the most influential parameters contributing to water classification.

To implement a user-friendly interface or dashboard for inputting test data and displaying classification results, allowing users or authorities to quickly assess the water quality status.

To improve prediction accuracy and reduce false classifications, ensuring that the chosen machine learning model is robust, efficient, and practical for real-world water monitoring systems.

To deploy the final trained model in a way that supports real-time or batch evaluation of water samples, ensuring high reliability, scalability, and integration into broader environmental monitoring frameworks.

METHODOLOGY

The methodology for developing the water quality classification system is structured into several key stages, starting with data preprocessing. Initially, a dataset containing water quality parameters and their corresponding quality labels (such as Safe, Moderate, or Polluted) is collected from authentic sources like Kaggle or government water boards. The dataset is carefully examined to handle missing or inconsistent values, and irrelevant or duplicate records are removed. Each feature is then normalized to ensure that the machine learning algorithms can effectively learn from the data without bias caused by differing units or scales.

Next, feature extraction and selection are performed to identify the most influential parameters affecting water quality classification. Parameters such as pH, turbidity, total dissolved solids (TDS), conductivity, and dissolved oxygen (DO) are analyzed for their impact on water classification. Correlation analysis is conducted to reduce multicollinearity among features. This ensures that only the most relevant and independent features are retained, improving model efficiency and performance.

Following this, two machine learning algorithms—Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost)—are trained and tested using the preprocessed dataset. The dataset is split into training and testing sets to evaluate each model's performance objectively. The models are assessed using key performance metrics such as accuracy, precision, recall, and F1-score. Hyperparameter tuning is also performed to optimize each model, ensuring the best possible classification performance.

Before final deployment, the complete system undergoes rigorous validation to ensure reliability, speed, and scalability. The methodology ensures that the final water quality classification system is not only accurate but also robust, user-friendly, and suitable for real-time environmental monitoring applications. It supports proactive decision-making by providing timely and trustworthy insights into water safety levels.

This section discusses the classification algorithms of SVM and XGBoost, whose performances are evaluated in the water quality classification. The methodology includes the project workflow, method description, and evaluation performance criteria, illustrated in Figure.

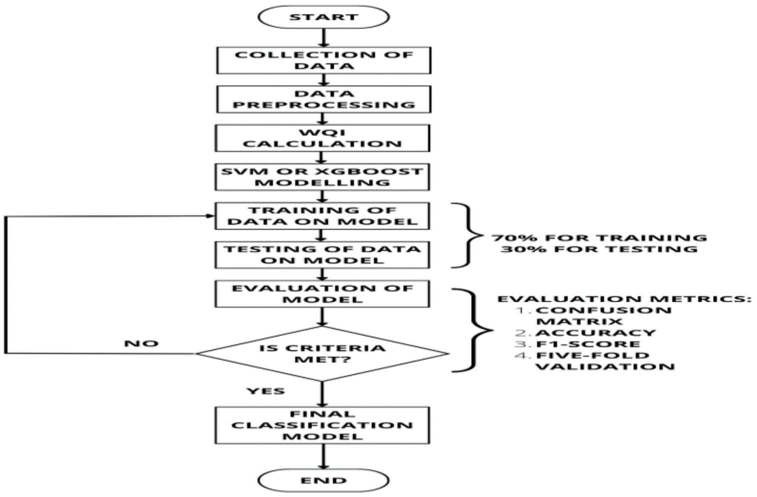

Fig. Flowchart of Methodology

RESULTS

In terms of using the K-fold validation method, fivefold were used in determining the accuracy of the models. The tuned SVM model was used as the final model, resulting in a 65% accuracy. In contrast, XGBoost had a mean accuracy score of 90% for the five-fold validation. Even using a different dataset for each fold, XGBoost proved superior to the SVM method, illustrated in Table.

Table: Number of Folds with Accuracy Score

Number of Fold	**SVM Accuracy**	**XGBoost**
1	0.66	0.92
2	0.68	0.88
3	0.61	0.90
4	0.58	0.90
5	0.67	0.94
Mean	0.64	0.94

CONCLUSION

Water quality is important in determining whether the water source is qualified for consumption. WQI is essential to classify whether the water is safe for consumption. Rather than requiring expensive and complex analysis to test the water quality, this research uses two machine learning algorithms, SVM and XGBoost, to predict water quality using readily available water quality parameters. The parameters employed for the classification algorithm are dissolved oxygen. pH. conductivity, biological oxygen demand, nitrate, fecal coliform. and total coliform. The outcome showed that XGBoost outperformed the SVM algorithm even after the parameters had been tuned. The results showed that SVM resulted in more misclassification of data than XGBoost. Despite the achievement of this research, improvements, such as applying more parameters to the model or using even more advanced deep learning algorithms, can be applied in future research to improve water quality classification.

REFERENCES

1. Sharma, K., Mishra, A., & Verma, R. (2020). Water Quality Prediction Using Machine Learning Techniques.
2. Gupta, M., & Patil, S. (2021). Comparative Study of Ensemble Methods for Water Quality Classification.
3. Kadiwal, A. (2022): Machine Learning-Based Water Potability Prediction System. (Kaggle Dataset Publisher)
4. Singh, R., Kumar, A., & Das, P. (2019). A Review on Water Quality Assessment Using Artificial Intelligence.
5. Rahman, A., Zhang, L., & Ahmed, N. (2022). Application of XGBoost for Environmental Dataset Classification.
6. Bharti, A., & Mishra, S. (2021). Prediction of Water Potability Using Classification Algorithms.
7. Deshmukh, R., & Jain, M. (2020). Potable Water Quality Prediction Using Decision Tree and SVM. International Journal of Scientific Engineering and Research, 8(5), 123–129.
8. Kumar, N., & Chauhan, S. (2021). Forecasting Water Contamination Using Machine Learning Models. Journal of Environmental Engineering and Studies, 7(2), 88–95.
9. Fatima, S., Ahmed, R., & Sheikh, I. (2022). Smart Water Monitoring Using IoT and Machine Learning. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 11(6), 54–60.
10. Reddy, V., & Kumar, P. (2020). Groundwater Quality Prediction Using SVM and ANN: A Case Study in India. Water Resources and Environmental Engineering, 6(3), 199–205.