International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Water Quality Classification Using SVM And XGBoost Method

Beldar Faijan Shaikh Akil[1], Lohar Bhavesh Kantilal[2], Pawar Darshan Madhukar [3], Patil Rushikesh Sanjay[4]

***Mr. Aakash . B. Kholi.* (*Asst. Prof*)**

Department of Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada

Faijan Shaikh Akil Beldar, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada, Lohar Bhavesh Kantilal, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada, Pawar Darshan Madhukar, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada, Patil Rushikesh Sanjay, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada.

Abstract – This project focuses on the classification of water quality 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.

***Key Words*:** Water Quality, Machine Learning, SVM, XGBoost, Classification

1.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 real-time 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.

2. LITERETURE SURVEY

A. Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status

Authors : A. Danades, D. Pratama, D. Anggraini, and D. Anggriani

Summary : Water is classified into four status of water quality, which good condition, lightly polluted, medium polluted and heavyly polluted. The classification status of water quality is very important to know the proper use and handling. Accuracy in classification of the quality status is very important, so that both of the classification algorithm K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are used. The classification of status of water quality based on the parameters. This study discusses the comparison algorithm KNN and SVM in classification of water quality status, a comparison conducted to determine the value that algorithm has the highest accuracy of the

determination water Quality Status Classification, testing KNN and SVM algorithm using 10-fold Cross Validation. Based on the result of the test, the highest average value of accuracy is SVM because the accuracy value is higher, it is 92.40% at linear kernel. The average value of KNN accuracy is only 71.28% at K=7.

B. **Support vector machines in water quality management**

Authors : K. P. Singh, N. Basant, and S. Gupta

Summary : Support vector classification (SVC) and regression (SVR) models were constructed and applied to the surface water quality data to optimize the monitoring program. The data set comprised of 1500 water samples representing 10 different sites monitored for 15 years. The objectives of the study were to classify the sampling sites (spatial) and months (temporal) to group the similar ones in terms of water quality with a view to reduce their number; and to develop a suitable SVR model for predicting the biochemical oxygen demand (BOD) of water using a set of variables. The spatial and temporal SVC models rendered grouping of 10 monitoring sites and 12 sampling months into the clusters of 3 each with misclassification rates of 12.39% and 17.61% in training, 17.70% and 26.38% in validation, and 14.86% and 31.41% in test sets, respectively. The SVR model predicted water BOD values in training, validation, and test sets with reasonably high correlation (0.952, 0.909, and 0.907) with the measured values, and low root mean squared errors of 1.53, 1.44, and 1.32, respectively. The values of the performance criteria parameters suggested for the adequacy of the constructed models and their good predictive capabilities. The SVC model achieved a data reduction of 92.5% for redesigning the future monitoring program and the SVR model provided a tool for the prediction of the water BOD using set of a few measurable variables. The performance of the nonlinear models (SVM, KDA, KPLS) was comparable and these performed relatively better than the corresponding linear methods (DA, PLS) of classification and regression modeling.

C. **Efficient optimization of support vector machine learning parameters for unbalanced datasets**

Authors : T. Eitrich and B. Lang

Summary : Support vector machines are powerful kernel methods for classification and regression tasks. If trained optimally, they produce excellent separating hyperplanes. The quality of the training, however, depends not only on the given training data but also on additional learning parameters, which are difficult to adjust, in particular for unbalanced datasets. Traditionally, grid search techniques have been used for determining suitable values for these parameters. In this paper, we propose an automated approach to adjusting the learning parameters using a derivative-free numerical optimizer. To make the optimization process more efficient, a new sensitive quality measure is introduced. Numerical tests with a well-known dataset show that our approach can produce support vector machines that are very well tuned to their classification tasks.

D. **Designing and accomplishing a multiple water quality monitoring system based on SVM**

Authors : Z. Pang and K. Jia

Summary : Along with the fast development of economy, the rational allocation of water resources has become one of the prerequisites which ensure the sustainable development of a country. Building a water quality surveillance and evaluation mechanism serving for managing regional water environment and addressing sudden contamination accident is of great importance. Based on the monitoring data on water quality, a multiple water quality monitoring system based on SVM is designed and accomplished. It established a corresponding water quality evaluation model by using the Gauss Radial Basis Function, and through offline studies of samples to guarantee the effectiveness and timeliness of this system. Besides the corresponding water quality classifying categories and instance interface are also figured out in the paper. This system has been successfully applied in the Central Line Project of South-to-North Water Diversion project, and according to the results the system is stable, effective and safety.

E. **XGBoost: A scalable tree boosting system**

Authors : T. Chen and C. Guestrin

Summary : Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

3.OBJECTIVES

- **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.

4.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.

Once the best-performing model is identified, it is prepared for deployment to ensure usability in real-world scenarios. A simple user interface or API can be developed using platforms like Flask or Streamlit, allowing users to input water sample parameters and receive instant classification results. This interface enhances accessibility and enables field experts or authorities to use the model for live predictions.

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.

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 Gradient Boosting Classifier, 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 Gradient Boosting Classifier outperformed the existing system even after the parameters had been tuned. In conclusion, this project highlights the significance of water quality and the need for an efficient and economical solution to monitor and manage it. The proposed approach, utilizing the advantages of machine learning techniques, provides an accurate and effective solution for predicting the water quality index and water quality class. The approach achieves a high Train Accuracy of 98% and Test Accuracy of 94%, indicating its potential for real-time

monitoring and management of water quality. The model developed in this study can predict water quality as Excellent, Good, Poor, and Very Poor, enabling various applications such as water treatment, environmental monitoring, and aquatic life management. Overall, this project demonstrates the potential of machine learning techniques in the field of water quality monitoring and management, and it can be further improved and expanded to meet the increasing demand for efficient and reliable water quality management systems.

REFRENCES

I. A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status," in Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016, Feb. 2017, pp. 137–141. DOI: 10.1109/FIT.2016.7857553.

II. K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," Analytica Chimica Acta, vol. 703, no. 2, pp. 152–162, Oct. 2011, DOI: 10.1016/j.aca.2011.07.027.

III. T. Eitrich and B. Lang, "Efficient optimization of support vector machine learning parameters for unbalanced datasets," Journal of Computational and Applied Mathematics, vol. 196, no. 2, pp. 425–436, Nov. 2006, DOI: 10.1016/j.cam.2005.09.009.

IV. Z. Pang and K. Jia, "Designing and accomplishing a multiple water quality monitoring system based on SVM," in Proceedings - 2013 9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2013, 2013, pp. 121–124. DOI: 10.1109/IIHMSP. 2013.39.

V. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, vol. 13-17-August-2016, pp. 785–794. DOI: 10.1145/2939672.2939785.