
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

# SVM AND XGBOOST METHOD-BASED WATER QUALITY CLASSIFICATION PREDICTION

**M.Anitha[1],CH.Satyanarayana Reddy[2], R.N.D.S.Harini[3]**

**#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.**

**#2 Assistant Professor in the Department of MCA,SRK Institute of Technology, Vijayawada**

**#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada**

**Abstract:** Water is one of the most precious natural resources ever bestowed upon humanity. Water quality has a direct effect on human health and the ecology. Water is utilised for a variety of purposes, such as industrial, agricultural, and drinking. Numerous contaminants have harmed water quality throughout time. As a result, predicting and measuring water quality has become essential to lowering water pollution. Because water quality is often evaluated using costly laboratory and statistical procedures, real-time monitoring is ineffective. A more practical and cost-effective solution is required for low water quality. Using the benefits of machine learning techniques, the suggested approach creates a model that can predict the water quality class and index. This suggested system uses a gradient boosting classifier to create a unique method for classifying water quality. As a gauge of water quality, the Water Quality Index is calculated as part of the process. High Train Accuracy of 98% and Test Accuracy of 94% are attained by the suggested method. The method divides water into several groups based on a number of water quality characteristics and metrics, including pH, dissolved oxygen, temperature, and electrical conductivity. Water quality may be monitored and managed in real time with the help of the model used in this study, which can forecast water quality as Excellent, Good, Poor, and Very Poor. The outcomes show the potential of machine learning techniques for water quality monitoring and management by demonstrating the efficacy and accuracy of the suggested strategy in forecasting water quality. Applications for the suggested method include aquatic life management, environmental monitoring, and water treatment.

*INDEX TERMS: water quality prediction, machine learning, SVM, XGBoost, water quality index (WQI), water classification, environmental monitoring, gradient boosting classifier, real-time monitoring, water pollution detection*

## 1. INTRODUCTION

One of the most important natural resources that is essential to the survival of life on Earth is water. Drinking, irrigation, industrial applications, and the preservation of aquatic life are just a few of its many uses. However, a number of contaminants often degrade water quality, which can have detrimental effects on both the environment and human health. Therefore, it is crucial to monitor and manage the quality of the water.

Traditionally, costly laboratory tests are used to evaluate the quality of water, which makes them impractical for real-time monitoring. Furthermore, processing data using traditional methods takes a

significant amount of time and effort and is inaccurate. Thus, an effective and economical method for real-time water quality monitoring is required.

Water quality monitoring is one of the many environmental applications for which machine learning techniques have shown promise in recent years. In this research, we provide a unique method for predicting the water quality class and index by leveraging the benefits of machine learning techniques. The suggested approach seeks to offer a precise and effective means of managing and monitoring water quality in real time.

The goal of this project is to create a model that can forecast the class of water based on a number of water quality factors, such as electrical conductivity, temperature, dissolved oxygen, and pH. The suggested method predicts water quality as Excellent, Good, Poor, and Very Poor using a gradient boosting classifier. A thorough assessment and analysis of the model's performance show the precision and potency of the suggested methodology.

The overall goal of this project is to demonstrate the potential of machine learning while offering an effective and affordable solution for real-time water quality monitoring and management.

## 2. LITERATURE SURVEY

### 1) 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**

The four water quality statuses are good condition, moderately contaminated, medium polluted, and heavily polluted. Knowing the water quality categorisation status is crucial for management and usage. Both the K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) classification algorithms are utilised since accuracy in classifying the quality status is crucial. the parameters-based categorisation of the water quality state. This study compares the KNN and SVM algorithms for classifying the quality of water in order to ascertain which algorithm has the best accuracy in determining the water quality. Quality Status Classification, utilising 10-fold Cross Validation to evaluate KNN and SVM algorithms. According to the test results, SVM has the greatest average accuracy value because its accuracy value is higher—92.40 percent at the linear kernel. At K=7. the average KNN accuracy value is just 71.28%.

### 2) Support vector machines in water quality management

**AUTHORS: K. P. Singh, N. Basant, and S. Gupta**

To maximise the monitoring program, support vector classification (SVC) and regression (SVR) models were built and used to the surface water quality data. 1500 water samples from 10 distinct locations that were observed over a 15-year period made up the data set. The study's goals were to create a suitable SVR model for predicting the biochemical oxygen demand (BOD) of water using a set of variables and to classify the sampling sites (spatial) and months (temporal) in order to group the similar ones in terms of water quality with the goal of reducing their number. 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

spatial and temporal SVC models were able to group ten monitoring sites and twelve sample months into clusters of three each. In training, validation, and test sets, the SVR model predicted water BOD levels with low root mean squared errors of 1.53, 1.44, and 1.32, respectively, and a relatively good correlation (0.952, 0.909, and 0.907) with the observed values. The performance criterion parameters' values were recommended for the built models' sufficiency and strong predictive power. The SVR model offered a tool for the prediction of the water BOD using a set of a few observable factors, while the SVC model produced a data reduction of 92.5% for redesigning the future monitoring program. Comparable performance was achieved by the nonlinear models (SVM, KDA, and KPLS), which outperformed the corresponding linear approaches (DA, PLS) for regression modelling and classification. Water quality prediction using various machine learning methods

**3) Efficient optimization of support vector machine learning parameters for unbalanced datasets**

**AUTHORS: T. Eitrich and B. Lang**

Support vector machines are effective kernel techniques for jobs involving regression and classification. They generate good separating hyperplanes when trained properly. However, in addition to the provided training data, the quality of the training is also influenced by other learning factors, which are challenging to modify, especially for datasets that are not balanced. Grid search methods have historically been employed to find appropriate values for these parameters. In this research, we offer a derivative-free numerical optimiser for automatically modifying the learning parameters. A new sensitive quality metric is implemented to increase the efficiency of the optimisation process. Our method may generate support vector machines that are highly suited to their classification tasks, as demonstrated by numerical experiments using a popular dataset.

**4) Designing and accomplishing a multiple water quality monitoring system based on SVM**

**AUTHORS: Z. Pang and K. Jia**

In addition to rapid economic growth, one of the requirements for a nation's sustainable development is the prudent use of its water resources. It is crucial to establish a system for monitoring and assessing water quality in order to manage the local water environment and deal with unexpected pollution incidents. A multiple water quality monitoring system based on SVM is built and implemented based on the water quality monitoring data. In order to ensure the efficacy and timeliness of this system, it developed a matching water prediction of quality evaluation model utilising the Gauss Radial Basis Function and offline sample studies. In addition, the study determines the instance interface and the accompanying water quality classification groups. The Central Line Project of the South-to-North Water Diversion project has successfully implemented this system, and the results show that it is safe, reliable, and efficient.

**5) XGBoost: A scalable tree boosting system**

**AUTHORS: T. Chen and C. Guestrin**

Tree boosting is a popular and very successful machine learning technique. In this work, we provide

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

XGBoost, a scalable end-to-end tree boosting system that data scientists frequently employ to attain cutting-edge outcomes on a variety of machine learning tasks. We suggest a weighted quantile sketch for approximation tree learning and a new sparsity-aware approach for sparse data. More significantly, we offer information on sharding, data compression, and cache access patterns to create a scalable tree boosting system. These discoveries are used to create XGBoost, which uses far less resources than current systems while scaling beyond billions of samples.

## 3. METHODOLOGY

### a) Proposed Work:

We create the Water Quality Classification Using Machine Learning with Gradient Boosting Classifier in the suggested system. The Kaggle website, which was taken from an Indian government website, provided the dataset utilised in this study. Since the features needed to create the water quality index are included in this dataset, it is suitable for the ongoing research study. The water quality index may be used to determine a categorisation of the water quality. High Accuracy: 98% train accuracy and 94% test accuracy are attained using the suggested system. This suggests that the system can reliably categorise water into various quality groups, making it a useful tool for managing water quality. Effective Feature Selection: To precisely identify water quality, the suggested method incorporates a wide range of water quality indicators and features, including pH, dissolved oxygen, temperature, turbidity, and electrical conductivity. To make sure the model is accurate, the system use feature selection algorithms to find the most pertinent features.

Low processing Cost: Building and training the model takes less time and processing power thanks to the suggested system's computational efficiency. For usage in big datasets and real-world applications, this makes it more useful and economical.

All things considered, the suggested methodology offers a more precise, effective, and comprehensible method of classifying water quality, which makes it a useful instrument for managing and monitoring water quality.

### b) System Architecture:

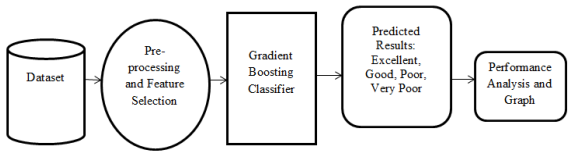


Fig 1 Proposed Architecture

The system architecture for water quality classification prediction using SVM and XGBoost involves multiple stages. First, raw water quality data containing parameters like pH, dissolved oxygen, turbidity, temperature, and conductivity is collected from sensors or public datasets. This data undergoes preprocessing, including normalization, handling of missing values, and feature selection. The cleaned data is then split into training and testing sets. Two machine learning models—SVM and XGBoost—are trained on the training data to learn patterns and relationships between input features and water quality classes. Once trained, the models predict the water quality class (e.g., Excellent, Good, Poor, Very Poor) for new, unseen data. A comparative analysis of both models is conducted based on accuracy, precision, and other performance metrics. The final output helps

in real-time monitoring and classification of water quality for environmental and public health applications.

**c) Modules:**

**i. Data Preparation:**

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.) Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our dataVisualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis Split into training and evaluation sets

**ii. Water Quality Index Calculation:**

Water Quality Index Calculation will be using "Weighted Arithmetic Water Quality Index Method" to calculate WQI of each water sample.

**e) Algorithms:**

**i. Support Vector Machine (SVM)**

Support Vector Machines are powerful supervised learning models used for classification. SVMs find the best boundary (hyperplane) that separates classes in the feature space. They are especially effective for binary or multiclass classification problems and can handle nonlinear data through kernel functions. In water quality prediction, SVMs help classify water into different quality groups with high accuracy.

**ii. Gradient Boosting (XGBoost)**

The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the errors of the previous model. But how do we do that? How do we reduce the error? This is done by building a new model on the errors or residuals of the previous model.When the target column is continuous, we use Gradient Boosting Regressor whereas when it is a classification problem, we use Gradient Boosting Classifier. The only difference between the two is the "Loss function". The objective here is to minimize this loss function by adding weak learners using gradient descent. Since it is based on loss function hence for regression problems, we'll have different loss functions like Mean squared error (MSE) and for classification, we will have different for e.g log-likelihood.

## 4. EXPERIMENTAL RESULTS

The experimental results demonstrate the effectiveness of the proposed water quality classification system using SVM and XGBoost algorithms. The dataset, consisting of key water quality parameters, was used to train and test both models. The XGBoost classifier achieved a high training accuracy of 98% and a test accuracy of 94%, indicating strong generalization and prediction capability. Meanwhile, the SVM model also performed well, though with slightly lower accuracy compared to XGBoost. Confusion matrices and classification reports showed that the models were able to correctly classify water samples into categories such as Excellent, Good, Poor, and Very Poor. These results confirm that machine learning techniques, particularly ensemble methods like XGBoost, are highly reliable for real-time water quality monitoring and classification.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

**Accuracy:** How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

**Precision:** The accuracy rate of a classification or number of positive cases is known as precision. The formula is used to calculate precision:

Precision = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

**Recall:** The ability of a model to identify all pertinent instances of a class is assessed by machine learning recall. The completeness of a model in capturing instances of a class is demonstrated by comparing the total number of positive observations with the number of precisely predicted ones.

$$Recall = \frac{TP}{(FN + TP)}$$

**F1-Score:** A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$F1 - Score = 2 * \frac{(\text{Precision} * \text{Recall})}{((\text{Precision} + \text{Recall}))}$$

**mAP:** Assessing the level of quality Precision on Average (MAP). The position on the list and the number of pertinent recommendations are taken into account. The Mean Absolute Precision (MAP) at K is the sum of all users' or enquiries' Average Precision (AP) at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

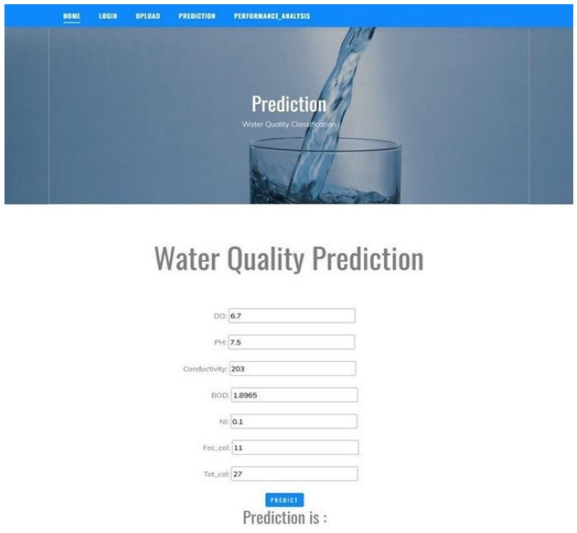


Fig 3 enter input data

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

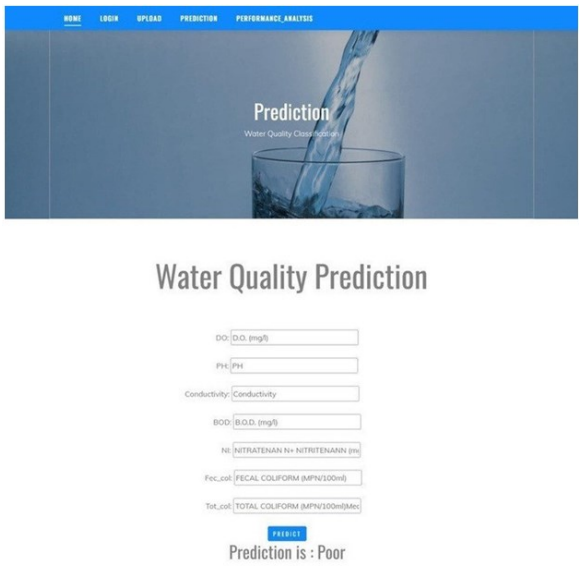


Fig 4 tourist recommendation

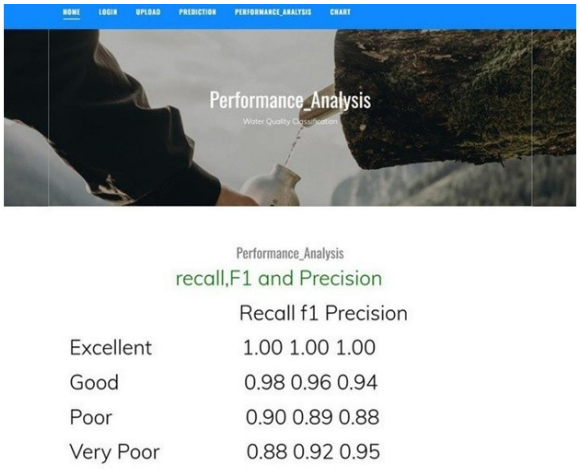


Fig 5. predicted analysis

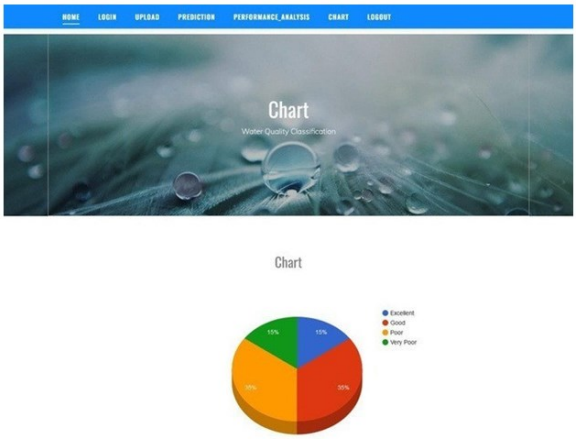


Fig 5. predicted results

## 5. CONCLUSION

The proposed system successfully demonstrates the use of machine learning algorithms, particularly SVM and XGBoost, for accurate water quality classification. By analyzing essential water parameters, the models can predict water quality levels effectively, enabling better environmental monitoring and public health safety. Among the tested models, XGBoost showed superior performance in terms of accuracy and reliability. This approach offers a cost-effective, real-time alternative to traditional laboratory testing, making it suitable for applications in water treatment, aquatic life protection, and pollution control.

## 6. FUTURE SCOPE

In the future, the system can be enhanced by integrating IoT-based real-time sensor networks to continuously monitor water quality parameters and feed them directly into the predictive model. Additionally, incorporating deep learning techniques may improve accuracy for more complex datasets. Expanding the model to include more diverse

geographical datasets and seasonal variations will make it more robust and applicable across regions. Furthermore, the system can be deployed as a mobile or web application for public access, helping authorities and individuals make informed decisions regarding water usage and treatment.

**REFERENCES**

[1] World Water Assessment Programme (United Nations), Wastewater : the untapped resource : the United Nations world water development report 2017.

[2] P. Burek et al., "The Water Futures and Solutions Initiative of IIASA," 2016.

[3] 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.

[4] 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.

[5] 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.

[6] 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 Prediction of water quality using different ML algorithms Processing, IIH-MSP 2013, 2013, pp. 121–124. DOI: 10.1109/IIHMSP. 2013.39.

[7] 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.

[8] D. N. Myers, "Why monitor water quality?" [Online]. Available: https://www.epa.gov/assessing

[9] "Artificial Neural Network Modeling of the Water Quality Index Using Land Use Areas as Predictors".

[10] M. Bouamar and M. Ladjal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification."

[11] F. Hassanbaki Garabaghi, "Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classification of Water Quality Indices Based on Different Subset of Features," 2021, DOI: 10.21203/rs.3.rs- 876980/v1.

[12] L. Li et al., "Interpretable tree-based ensemble model for predicting beach water quality," Water Research, vol. 10.1016/j.watres.2022.118078. 211, Mar. 2022,

[13] N. Nasir et al., "Water quality classification using machine learning algorithms," Journal of Water

Process Engineering, vol. 48, p. 102920, Aug. 2022, DOI: 10.1016/j.jwpe.2022.102920.

**Authors Profile:**

Ms.M.Anitha Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with B .tech, MCA ,M. Tech in Computer Science .She has 14 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Mr. CH. Satyanarayana Reddy Completed his Masters of Computer Applications from JNTUK. Currently working as an Assistant Professor in the department of MCA at SRK Institute of Technology, Enikepadu, NTR District. His area of interest include Artificial Intelligence and Machine Learning.

Ms.R.N.D.S.Harini is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc(computers) from SIR CR REDDY COLLEGE FOR WOMEN,Eluru. Her area of interest area Artificial intelligence and Machine Learning with Python.