Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Comparative Analysis of Machine Learning Algorithms for Water Quality Index Prediction

D. Dasu[1], Research Scholar, Prof. P. Suresh Varma[2], Professor

[1,2] Dept. of CSE, Adikavi Nannaya University, Rajamahendravaram, Andhra Pradesh

Article History:

Received: 12-11-2024

Revised: 17-12-2024

Accepted: 06-01-2025

Abstract:

Access to clean and safe water is essential for environmental sustainability and economic progress. The Water Quality Index (WQI) serves as an important metric, synthesizing complex water quality data into a single value to assess its suitability for various uses, including drinking, agriculture, and aquaculture. In the recent years, Machine Learning (ML) algorithms have revolutionized WQI prediction, offering efficient and scalable solutions to analyse complex datasets. A comparative study of the performance of multiple ML algorithms—Naïve Bayes, XGBoost, K-Nearest Neighbors, Decision Tree, AdaBoost, and Logistic Regression for WQI prediction is explored in this paper. Experimental results demonstrate that Naïve Bayes outperforms other classifiers, highlighting its suitability for handling probabilistic relationships in water quality data. The findings underscore the growing significance of ML to enhance real-time water quality monitoring and sustainable resource management, particularly for small-scale water systems. By leveraging ML, this research provides a foundation for more accurate and reliable water quality assessments.

Keywords: Machine Learning, Water Quality Index, Naïve Bayes, Water Quality Assessment.

INTRODUCTION

Safe water is an essential resource. It is indispensable for preventing waterborne diseases and maintaining ecosystems. Additionally, access to clean water is crucial for overall economic development. Almost all major industrial manufacturing processes rely on water for production, making it a key resource for economic growth. Furthermore, agriculture heavily depends on water for irrigation, highlighting the importance of clean water for food security and livelihoods.

Recognizing the critical role of water quality in various sectors, there arose a need for a simplified, standardized tool to assess and communicate the condition of water resources effectively. This need led to the development of the Water Quality Index (WQI), a comprehensive metric to evaluate water suitability for different uses. The concept of the Water Quality Index (WQI) was first developed in the early 1970s by Horton (1965), who proposed a method "to simplify the complex water quality data into a single index"[1]. This was further simplified by Brown *et al.* in 1970[2]. This approach was further refined and popularized by the United Nations Environment Programme (UNEP) in collaboration with various researchers[3].

The WQI methodology was designed to provide a standardized metric for assessing water quality by merging numerous parameters into a single value that reflects overall water conditions, making it easier for decision-makers and the public to understand and use in resource management.

The Water Quality Index (WQI) is a pivotal metric, synthesizing various parameters of water quality into a single value to assess the water for diverse uses, including drinking, agriculture, and industrial applications. The WQI thus helps us to make informed decisions regarding water management and its utilization. By monitoring and improving water quality, communities can ensure sustainable development and protect the health of both people and ecosystems. The importance of quality water cannot be stressed in the field of aquaculture where water quality directly impacts the growth and health of aquatic organisms. Maintaining high water quality is essential for maximizing production and minimizing disease outbreaks in aquaculture systems.

The prediction and evaluation of water quality have been completely transformed in recent years by the incorporation of Machine Learning (ML) methods. ML models improve decision-making in water resource management by effectively analysing complicated datasets, spotting trends, and producing precise water quality predictions. Numerous ML algorithms have been shown in studies to be effective in accurately predicting WQI. Additionally, it has been demonstrated that by utilising the advantages of several models, ensemble techniques like Random Forest and Gradient Boosting enhance prediction performance.

In addition to expediting the prediction process, the use of ML in water quality evaluation enables real-time monitoring and early warning systems. This capacity is essential for guaranteeing sustainable water management and reducing the negative impacts of water pollution. The use of machine learning and artificial intelligence in environmental monitoring and management is set to grow as the availability of massive datasets and processing capacity increases, offering reliable and scalable solutions for water quality prediction.

This study intends to compare the effectiveness of a number of machine learning algorithms in forecasting the Water Quality Index, including SVM, XGBoost, KNN, Decision Tree, Naive Bayes, AdaBoost, and Logistic Regression. The aim is to determine the best methods for evaluating the quality of water by examining their accuracy, computational effectiveness, and applicability. This will help to create more dependable and efficient water management plans.

RELATED WORK

Predicting and managing water quality has been greatly improved by recent developments in machine learning. In 2018, in order to estimate the water quality of the Tireh River in Iran, Abbas Parsaie, Ali Heidar Nasrolahi, and Amir Hamzeh Haghiabi assessed Artificial Neural Networks (ANN), Support Vector Machines (SVM), and the Group Method of Data Handling (GMDH). Using characteristics like pH and dissolved oxygen, SVM proved to be the most accurate model, especially when combined with the Radial Basis Function (RBF) kernel. Despite its processing efficiency, GMDH showed poorer accuracy than ANN, which also performed well.[4]

Xiaohui Yan et al. (2024) conducted a comprehensive review of 170 studies, highlighting the superior performance of Random Forest, Gradient Boosting, and neural networks compared to traditional methods. Enhanced by remote sensing and satellite data, these models provided scalable predictions but faced challenges related to model uncertainty and regional specificity.[5]

Priyanka Gupta and colleagues (2024) compared Logistic Regression, Decision Tree, Random Forest, and SVM, finding Random Forest to be the most effective with an accuracy of 68.58%. The study emphasized Random Forest's ability to handle complex datasets and mitigate overfitting.[6]

In the U.S., Yinpu Li et al. (2023) evaluated models such as Linear Regression, XGBoost, and LightGBM for pH prediction. Ensemble methods, especially LightGBM, achieved the highest accuracy and demonstrated the potential to challenge assumptions about the necessity of spatial and temporal data.[7]

Md. Mehedi Hassan et al. (2021) explored models like Multinomial Logistic Regression (MLR) and Random Forest for predicting the Water Quality Index (WQI) in India. MLR achieved the highest accuracy of 99.83%, highlighting the effectiveness of supervised learning techniques for water quality classification.[8]

Francesca Razzano et al. (2024) introduced a novel approach for turbidity monitoring using the CatBoost algorithm integrated with atmospherically corrected Sentinel-2 satellite data. The model achieved high accuracy, with an RMSE of 0.24, demonstrating the potential of combining remote sensing with machine learning for large-scale water quality monitoring.[9]

Similarly, Dhruti Dheda and Ling Cheng (2020) developed Long Short-Term Memory (LSTM) models for predicting dissolved oxygen levels in Australia's Burnett River. The single-step LSTM model achieved an error of 0.01 mg/L, proving highly effective for modeling sequential data and long-term dependencies.[10]

Thair S.K., Abdul Hameed M.J., and Ayad S.M. (2014) used ANN to predict Total Dissolved Solids (TDS) in the Euphrates River, utilizing data from 1999 to 2013 across six stations. ANN outperformed Multiple Linear Regression (MLR), achieving a correlation coefficient of 0.928, RMSE of 319.5, and MAPE of 21.26%. Discharge and station distance were identified as the most influential factors, with relative importance of 75% and 61%, respectively. The study emphasized ANN's capability to model complex relationships in water quality data and underscored the need for robust monitoring systems to manage upstream irrigation and pollution impacts.[11]

Al Shehhi, Maryam R and Abdullah Kaya (2020) explored SARIMA, regression, and neural networks for forecasting water quality parameters such as chlorophyll-a (Chl-a), fluorescence line height (FLH), and sea surface temperature (SST) using MODIS satellite data from 2003 to 2012 in the Arabian Gulf. SARIMA was most effective for predicting SST and FLH, achieving R^2 values over 0.98, while neural networks excelled at predicting Chl-a in shallow, turbid waters. Regression models performed well in deep, less turbid conditions, demonstrating the adaptability of ML models to diverse water environments and their potential for real-time monitoring.[12]

Folorunso Taliha et al. (2019) applied ANN to predict WQI in aquaculture systems using parameters such as dissolved oxygen (DO), temperature, pH, alkalinity, and conductivity. Among various activation functions, the tangent sigmoid function performed best, achieving an MSE of 0.00245, RMSE of 0.0495, and an R-value of 0.998. The ANN models outperformed Multilinear Regression (MLR), demonstrating superior accuracy and reliability for water quality management in aquaculture systems.[13]

Much of the existing research in water quality prediction using machine learning focuses predominantly on satellite-derived data or generalized approaches not specifically tailored for a particular industry or purpose of utilization.

While these studies have demonstrated the efficacy of advanced algorithms in monitoring large water bodies and regional-scale environments, they often lack the granularity and precision required for addressing the unique challenges faced during the determination of WQI for industrial or agricultural purposes.

Furthermore, a review of the literature reveals a significant gap in studies that fine-tune models for the specific conditions of water usage such as: aquaculture ponds or effluents from a specific industry. This dearth of targeted research underscores the need for machine learning approaches designed explicitly for small-scale, water utilization systems to ensure their sustainability and efficiency.

METHODOLOGY

a) Comparison of selected ML algorithms

XGBoost

Extreme Gradient Boosting, or XGBoost for short, is a decision tree-based ensemble learning technique. By concentrating on the residuals, each new model it creates fixes the mistakes of the one before it. In order to manage model complexity and guarantee better generalisation and less overfitting, XGBoost uses a regularised objective function that incorporates both loss function and penalty terms.[14]

In water quality prediction, XGBoost is often used to predict continuous WQI values due to its ability to handle missing values, outliers, and non-linear relationships. Its fast computation and scalability make it ideal for large datasets derived from extensive monitoring systems. Moreover, XGBoost provides feature importance rankings, enabling researchers to identify critical parameters impacting water quality.

K-Nearest Neighbors (KNN)

KNN is a non-parametric and instance-based learning algorithm used for classification and regression. The algorithm assigns a data point to the majority class among its k-nearest neighbors, determined using distance metrics such as Euclidean, Manhattan, or Minkowski. It assumes that similar data points exist in close proximity, relying on the distance function to measure similarity. KNN does not require a training phase, making it a lazy learning algorithm.[15]

KNN is particularly useful for water quality classification tasks, where it assigns water samples to quality categories based on the proximity of their parameter values to previously labelled samples. Its simplicity and effectiveness make it a popular choice for preliminary analyses in regions with limited computational resources. However, its sensitivity to noise and high-dimensional data necessitates careful pre-processing for accurate results.

Decision Tree

Decision Trees are hierarchical, tree-like structures that recursively split the data into subsets based on feature values. At each node, the algorithm chooses a feature that maximizes information gain or minimizes impurity measures like Gini index or entropy. The process continues until the data is split

into pure subsets or reaches a stopping criterion. The resulting structure is intuitive and easy to interpret.[16]

In water quality studies, Decision Trees are used to identify critical thresholds for water parameters, such as pH or dissolved oxygen, that influence WQI classifications. They are particularly advantageous for exploratory analysis, as they provide a clear representation of decision rules. However, their tendency to overfit can be mitigated through pruning or by using ensemble methods like Random Forest.

Naive Bayes

The Bayes theorem is the foundation of the probabilistic algorithm known as Naive Bayes. With the premise that all characteristics are independent, it determines the posterior probability of a class given the feature values (naive assumption). Despite this presumption, Naive Bayes works effectively in a variety of real-world situations, particularly when dealing with categorical data. The model assigns the class with the highest probability to a given data point after calculating the probabilities for each class.[17]

When classifying water quality using categorical data, like labelling samples according to discrete parameter ranges (e.g., low, medium, high contamination levels), Naive Bayes is helpful. It is perfect for real-time applications in monitoring systems because to its simplicity and computational efficiency. It might, however, have trouble with qualities that contain intricate connections or are highly connected.

AdaBoost

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that combines multiple weak classifiers to create a strong classifier. It assigns weights to data points and iteratively updates these weights based on misclassifications. The weak classifiers focus more on difficult cases in subsequent iterations, and the final model aggregates their weighted predictions to improve overall accuracy. [18]

AdaBoost is applied in water quality prediction to enhance the performance of base classifiers like Decision Trees or Logistic Regression. Its ability to adapt to different types of datasets makes it suitable for heterogeneous water quality data. However, it is sensitive to noise and outliers, which can adversely affect its performance, necessitating careful data pre-processing.

Logistic Regression

Logistic Regression is a linear model used for binary classification. It represents the relationship between the dependent binary variable and one or more independent variables using a logistic function. The logistic function maps input values to probabilities, making it suitable for predicting the likelihood of outcomes. The model estimates coefficients using maximum likelihood estimation.[19]

In water quality analysis, Logistic Regression is often used to classify water samples as safe or unsafe based on a threshold WQI value. Its simplicity and interpretability make it a preferred choice for understanding the influence of specific parameters on water quality. However, its linear decision boundary may limit its applicability in datasets with complex non-linear relationships.

b) Water Quality Index Calculation & Classification

The fundamental goal of computing the water quality index (WQI) is to convert the complicated data on water quality into information that is simple to comprehend and useful. The "Weighted Arithmetic Water Quality Index Method" is used to calculate WQI of each water sample.

The formula to calculate WQI is -

$WQI = \sum qi \times wi / \sum wi$

Here wi - Unit weight of ith parameter

qi - Quality estimate scale of each parameter, it is calculated with the formula -

$qi = 100 \times (Vi - VIdeal / Si - VIdeal)$

Here Vi - Measured value of ith parameter

Videal - Ideal value of ith parameter in pure water

Si - Standard value recommended for ith parameter

wi is calculated by the formula -

$wi = K / Si$

Here K is proportionality constant which is -

$K = 1 / \sum Si$

Table 1: WQI Classification

WQI Score	**WQI Class**
90-100	Excellent
70-89	Good
50-69	Fair
25-49	Poor
0-24	Very Poor

Table 2: WQI Ideal Values

Parameter	**Ideal Value (Videal)**
Dissolved Oxygen	14.6 mg/L
pH	7.0
Conductivity	0 µS/cm
Biological Oxygen Demand	0 mg/L
Nitrate	0 mg/L
Fecal Coliform	0/100 mL
Total Coliform	0/100 mL

Internationally, the ideal WQI values are determined by World Health Organization[20]. Bureau of Indian Standards (BIS) determines the ideal WQI values for India[21]. However, most of the studies are based on the values recommended by United States Environmental Protection Agency (US EPA) and United Nations Environment Programme (UNEP)[22]. Furthermore, it must be noted that these values change based on the purpose of water utilisation and public policies in effect.

c) Dataset

The parameters—Dissolved Oxygen (DO), pH, Conductivity, Biological Oxygen Demand (BOD), Nitrate, and Faecal Coliform—are essential for comprehensively assessing water quality. DO reflects the oxygen available for aquatic life and indicates pollution or nutrient loading. pH captures the water's acidity or alkalinity, influencing chemical and biological processes. BOD highlights organic pollution and oxygen depletion, often linked to sewage or industrial waste. Nitrate levels reveal nutrient loading and eutrophication risks from agricultural runoff, and Faecal Coliform indicates microbial contamination and potential waterborne disease risks. Together, these parameters provide a robust dataset for predicting Water Quality Index (WQI) and identifying key water quality challenges effectively.

The United Nations, through the United Nations Environment Programme (UNEP) and the World Health Organization (WHO), emphasizes the importance of monitoring key water quality parameters that align closely with those selected for this study. By focusing on these specific parameters, the study aligns with global water quality monitoring standards.

Table 3: Dataset used for comparing algorithms

Dissolved Oxygen (mg/L)	pH	Conductivity (µS/cm)	BOD (mg/L)	Nitrate (mg/L)	Faecal Coliform (count/100mL)	WQI Class
7.5	7.2	800	2.5	10	50	Good
3.0	9.0	1500	7.0	40	500	Poor
8.0	6.8	500	1.0	5	30	Excellent
4.0	8.5	1200	4.5	20	200	Fair
2.5	8.0	1600	8.0	50	1000	Very Poor

EXPERIMENTAL RESULTS

The experiments aimed to predict the Water Quality Index (WQI) using various machine learning classifiers. The results demonstrated that Naïve Bayes outperformed other ML classifiers in accuracy and reliability for WQI prediction. This highlights Naïve Bayes' suitability for handling categorical data and simple probabilistic relationships, making it an effective choice for water quality assessments.

Table 4: Accuracy Comparison of Algorithms

S.No	Model	Accuracy Score
1	Naïve Bayes	0.788540
2	XGBoost	0.770980
3	KNeighbours	0.753420
4	Decision Tree	0.745102

5	AdaBoost	0.734011
6	Logistic Regression	0.728466

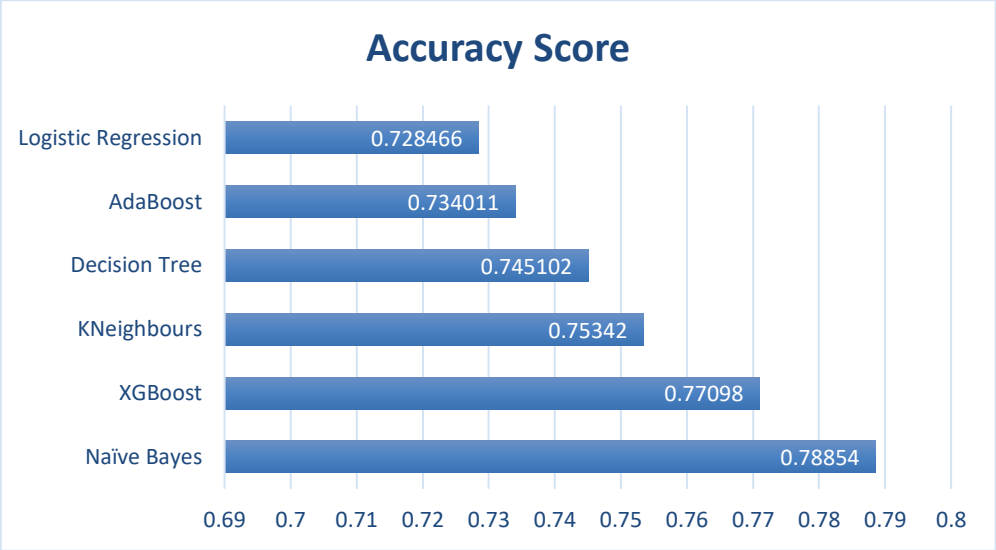

Fig1: Accuracy Comparison of Algorithms

CONCLUSION

In forecasting and controlling the Water Quality Index (WQI), a crucial indicator for guaranteeing clean and safe water that is necessary for human health, environmental sustainability, and economic growth, this conversation highlights the revolutionary potential of machine learning (ML). It is clear from evaluating different machine learning classifiers—such as Decision Trees, K-Nearest Neighbours, Naïve Bayes, XGBoost, AdaBoost, Logistic Regression, and Random Forests—that these models provide reliable, scalable, and accurate solutions for a range of applications, from large-scale industrial wastewater monitoring to aquaculture systems.

Interestingly, research shows that Naïve Bayes performs exceptionally well, capturing intricate patterns in water quality data. The integration of machine learning (ML) techniques with water quality management will open the door to improved real-time monitoring, predictive analytics, and sustainable resource utilisation, especially in understudied areas of water utilisation for industrial and individual consumption, as computing power and data availability continue to increase.

REFERENCES

[1]. Horton, R. K. (1965). An index number system for rating water quality. *Journal of Water Pollution Control Federation, 37*(3), 300–306.

[2]. Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). A water quality index—Do we dare? *Water and Sewage Works, 117*(10), 339–343.

[3]. United Nations Environment Programme [UNEP]. [n.d.]. Ambient water quality indicators for Sustainable Development Goal [SDG] 6.3.2. Retrieved from https://www.unep.org/explore-topics/water/what-we-do/world-water-quality-alliance-wwqa-partnership-effort/faqs-water

[4]. Parsaie, A., Nasrolahi, A. H., & Haghiabi, A. H. [2018]. Water quality prediction using machine learning methods. *International Research Journal of Natural Sciences, 53*[1], 3–12.

[5]. Yan, X., Zhang, T., Du, W., Meng, Q., Xu, X., & Zhao, X. [2024]. A comprehensive review of machine learning for water quality prediction over the past five years. *Journal of Marine Science and Engineering, 12*[1].

[6]. Gupta, P., Chidrawar, S., Adhav, M., Pawar, D., & Chaudhari, H. [2024]. Comparative analysis of machine learning algorithms for water quality prediction. *International Journal of Research Publication and Reviews, 5*[2], 224–229.

[7]. Li, Y., Mao, S., Yuan, Y., Wang, Z., Kang, Y., & Yao, Y. [2023]. Beyond tides and time: Machine learning's triumph in water quality forecasting. *arXiv preprint*. https://arxiv.org/abs/2309.16951

[8]. Hassan, M. M., Hassan, M. M., Akter, L., et al. [2021]. Efficient prediction of water quality index [WQI] using machine learning algorithms. *Human-Centric Intelligent Systems, 1*[1].

[9]. Razzano, F., Mauro, F., Di Stasio, P., et al. [2024]. Monitoring water contaminants in coastal areas through ML algorithms leveraging atmospherically corrected Sentinel-2 data. *arXiv preprint*. https://arxiv.org/abs/2401.03792

[10]. Dheda, D., & Cheng, L. [2020]. A multivariate water quality parameter prediction model using recurrent neural network. *arXiv preprint*. https://arxiv.org/abs/2003.11492

[11]. Thair, S. K., Hameed, A. M. J., & Ayad, S. M. [2014]. Prediction of water quality of Euphrates River using artificial neural network model: A spatial and temporal study. *International Research Journal of Natural Sciences, 2*[4], 45–58.

[12]. Al Shehhi, M. R., & Kaya, A. [2020]. Time series and machine learning to forecast water quality from satellite data. *arXiv preprint*. https://arxiv.org/abs/2003.11923

[13]. Folorunso, T. A., Aibinu, M. A., Kolo, J. G., Sadiku, S. O. E., & Orire, A. M. [2019]. Water quality index estimation model for aquaculture system using artificial neural network. *International Journal of Artificial Intelligence Research, 8*[1], 12–25.

[14]. T. Chen, University of Washington, C. Guestrin, and University of Washington, "XGBoost: A Scalable Tree Boosting System," *Terra Swarm Research Center*, 2016, [Online]. Available: https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf

[15]. G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," in *Lecture notes in computer science*, 2003, pp. 986–996. doi: 10.1007/978-3-540-39964-3_62.

[16]. L. Rokach and O. Maimon, "Decision Trees," in *Springer eBooks*, 2006, pp. 165–192. doi: 10.1007/0-387-25465-x_9.

[17]. F. Sabry, *Naive Bayes classifier: Fundamentals and Applications*. One Billion Knowledgeable, 2023.

[18]. Bonaccorso, G. (2018). Mastering Machine Learning Algorithms: Expert Techniques to Implement Popular Machine Learning Algorithms and Fine-tune Your Models. Germany: Packt Publishing.

[19]. Kleinbaum, D. G. (2013). Logistic Regression: A Self-Learning Text. Germany: Springer New York.

[20]. World Health Organization [WHO]. [2017]. Guidelines for drinking-water quality [4th ed.]. Retrieved from https://www.who.int/publications/i/item/9789241549950

[21]. Bureau of Indian Standards. (2012). *IS 10500:2012 Drinking Water—Specification (Second Revision)*. Retrieved from https://law.resource.org/pub/in/bis/S06/is.10500.2012.pdf

[22]. United States Environmental Protection Agency (US EPA). (1970s). Development of the National Water Quality Index. *Environmental Studies Report*.