Innovating Water Purity Analysis with Gradient Boosting Classification Techniques

Jatin Sharma[1], Kanwarpartap Singh Gill[1], Mukesh Kumar[2], Ruchira Rawat[3]

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1]

Computer Science & Engineering, Graphic Era Hill University, Dehradun, Uttarakhand, India[2]

Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India[3]

Corresponding author: Kanwarpartap Singh Gill, Email: kanwarpartap.gill@chitkara.edu.in

Ensuring water quality is crucial for protecting both human health and aquatic ecosystems. In this research, we delve into the effectiveness of gradient boosting classifiers for predicting water quality parameters. Our study evaluates the potential of this approach in assessing groundwater quality, comparing it against three other gradient boosting-based algorithms, and contrasts its performance with traditional machine learning methods. We also analyze feature importance to identify the key factors influencing water quality. The findings reveal an accuracy rate of approximately 78%. Moreover, the identified significant features provide valuable insights into the factors driving water quality changes. This research advances predictive modeling techniques in water quality assessment, aiding in proactive management strategies for sustainable water resource use and ecosystem conservation. These models hold significant implications for informed decision-making in agricultural water management and resource allocation within the region. As the population grows and the demand for resources increases, managing our lives becomes more challenging. In this struggle, we sometimes resort to using poor or contaminated water sources, endangering our health. According to a recent World Health Organization (WHO) survey, over 2.2 billion people in India suffer from issues related to unsafe drinking water, and 21% of diseases are linked to impure water.

Keywords: Machine Learning, Water Quality, Gradient Boosting Classifier, Classifier Evaluation, Artificial Intelligence, Predictive Modelling Techniques, Diseases, Population, Resources, Resources.

2024. In Mukesh Saraswat & Rajani Kumari (eds.), *Applied Intelligence and Computing*, 125–130. Computing & Intelligent Systems, SCRS, India. DOI: https://doi.org/10.56155/978-81-955020-9-7-14

1 Introduction

No living organism on Earth could survive without water. Water from lakes and rivers supports the fishing industry and human well-being, either directly or indirectly. However, the annual growth of various industries, driven by increasing demand, has led to the routine disposal of hazardous waste into these water bodies, causing significant pollution. This water pollution results in millions of deaths each year, incalculable financial losses, and the degradation of agricultural land. Recent research indicates a drastic decline in groundwater quality across most countries. In this study, we examine the effectiveness of gradient boosting classifiers in predicting water quality. In recent years, prediction models like artificial neural networks, multiple linear regression, and decision trees have been crucial in water quality management systems. However, most of these models, including decision trees, artificial neural networks, wavelet neural networks, and recurrent neural networks, require numerous input parameters and substantial processing power, making them costly to develop. Gradient boosting, a powerful ensemble learning method in machine learning, is known for creating highly accurate predictive models. It operates by sequentially training a series of decision trees, with each new tree aiming to correct the errors of its predecessors. The term "gradient" refers to the direction in which model parameters are adjusted to minimize prediction errors, focusing on reducing the loss function through an iterative process. By combining the predictions of several weak models, gradient boosting produces a strong, robust model capable of handling complex data interactions. Its adaptability, performance, and efficiency make it a compelling choice for tasks ranging from regression to classification in various industries, including banking and healthcare. This article also introduces the Recurrent Neural Network (RNN), a neural network architecture that excels in handling sequential information, such as text and time-series data. Unlike traditional generative neural networks, where inputs and outputs are independent, RNNs use feedback channels to process sequential input, employing the output from the previous time step as the input for the current time step. This capability to effectively model data sequences makes RNNs suitable for tasks like speech recognition, natural language processing, and time series prediction.

2 Literature Review

Ensuring water quality prediction is vital for the safety and sustainability of water resources, impacting public health and environmental management significantly. Recently, machine learning (ML) algorithms have become essential tools for predicting water quality, enhancing forecasting accuracy and efficiency [1]. This literature review examines the latest advancements in ML techniques for water quality prediction, synthesizing recent research findings. Raheja, Goel, and Pal (2024) present a novel method for predicting groundwater quality using gradient boosting-based algorithms, demonstrating their effectiveness in capturing complex relationships within groundwater quality datasets for more accurate predictions [2] [3]. Similarly, Helm et al. (2024) develop a gradient boosting-assisted ML model for predicting free chlorine residual, showcasing its effectiveness in enhancing predictive performance [4]. These studies highlight the utility of gradient boosting algorithms in improving water quality prediction models. Mohseni et al. (2024) use ensemble machine learning models to predict the weighted arithmetic water quality index for urban water quality assessment. Their research shows the potential of ensemble techniques in integrating diverse data sources and optimizing prediction accuracy [3]. Additionally, Krishnan and Manikandan (2024) employ advanced ML algorithms with data augmentation to predict water quality, demonstrating the effectiveness of ensemble approaches in managing complex datasets. These studies emphasize the robustness and versatility of ensemble ML models for water quality prediction tasks [5]. Díaz-González and Aguilar-Rodríguez (2024) introduce Aqua-P, a ML-based tool for water quality assessment, highlighting the importance of data-driven approaches in informing decision-making processes [6]. Similarly, Ghosh et al. (2023) propose a predictive ML framework for water quality assessment, underscoring the role of data analytics in tackling water management challenges. These studies illustrate the transformative potential of data-driven ML techniques in advancing water quality prediction and management practices [7]. Leggesse et al. (2023) investigate the integration of remote sensing data with ML algorithms for predicting optical

water quality indicators in the tropical highlands of Ethiopia [8]. Their research demonstrates the feasibility of using remote sensing technology to enhance water quality prediction capabilities, especially in data-scarce regions [9]. Similarly, Rawat et al. (2023) conduct a comprehensive analysis of ML algorithms for predicting water quality, emphasizing the importance of integrating diverse data sources for improved prediction performance. These studies highlight the significance of incorporating remote sensing data into ML-based water quality prediction models [10].
The following points cover the methodology:

- The research utilizes a gradient boosting classifier for the proposed material.
- Various optimization techniques are employed for classification purposes.
- Multiple water parameters are considered.
- Research benefits significantly from such classification techniques.

The second part of the research is outlined as follows: section III focuses on input, section IV discusses the proposed methodology, section V presents the results, and the conclusion and references are at the end.

3 Input Dataset

The water quality dataset contains information on various parameters collected from 3,276 distinct water sources. These parameters are vital indicators of the water's suitability for human consumption and environmental well-being [11]. The pH values, which indicate the acidity or alkalinity of the water, fall within the World Health Organization (WHO) standards, ranging from 6.52 to 6.83. Water hardness, primarily due to calcium and magnesium salts, is influenced by the length of time the water is in contact with hardness-causing materials. Total dissolved solids (TDS), reflecting the mineral content in the water, should not exceed 500 mg/L for drinking water according to WHO guidelines, with higher values indicating increased mineralization [12]. Chloramines, a byproduct of water treatment, and naturally occurring sulfates are monitored within safe limits. Conductivity, which measures the water's ability to conduct electrical current due to ion concentration, is kept within the WHO-recommended limit of 400 µS/cm. Total organic carbon (TOC) is an indicator of organic compounds in the water, with strict standards set by the US Environmental Protection Agency (EPA) for both treated and source water. Trihalomethanes (THMs), formed during chlorine treatment, are regulated to remain within safe levels. Turbidity, which indicates the presence of suspended solid matter, is assessed against WHO guidelines. The final attribute in the dataset, potability, classifies water as safe (1) or unsafe (0) for human consumption, emphasizing the critical importance of providing access to safe drinking water as a fundamental human right essential for health and development at all levels. The dataset's input parameters are illustrated in Figure 1.

ph	Hardness	Solids	Chloramine	Sulfate	Conductivit	Organic_ca	Trihalomet	Turbidity	Potability
	204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
8.842464	229.9644	7839.319	10.50881	278.4283	370.0852	15.79098	77.16662	4.576729	1
7.721078	208.4386	17248.62	7.68537	286.4035	269.0136	11.75735	56.88453	3.223951	1
9.089421	208.9147	32238.08	6.895014	321.0805	449.26	12.51464	89.21307	4.430116	1
1.757037	147.5818	41538.24	7.728177	376.0129	428.4448	10.8287	65.00584	2.967554	1
7.725192	213.9513	20461.76	5.340319		338.2143	17.03247	58.58746	4.163326	1
6.729191	178.4936	31991.96	5.63294		341.7748	14.8052	64.12915	5.112254	1

Figure 1. Different Parameters describing the water quality features

4 Proposed Methodology

In the realm of machine learning, boosting stands out as a powerful technique for tackling regression and classification challenges. Boosting is a type of ensemble learning where the model is iteratively trained, with each new iteration building upon the previous one to enhance performance [13]. This process transforms a collection of weak learners into a formidable ensemble. Among the most prevalent boosting algorithms are AdaBoost and Gradient Boosting [14]. Gradient Boosting employs gradient descent to iteratively train new models, focusing on minimizing a loss function, such as mean squared error or cross-entropy, from the previous iteration. This technique effectively refines the model by computing the gradient of the loss function with respect to the current ensemble's predictions [15] [16]. A new weak learner is then trained to minimize this gradient. The predictions from this new model are incorporated into the ensemble, and this process repeats until a predefined stopping criterion is met [17] [18]. Through this systematic approach, Gradient Boosting transforms a series of weak learners into a robust predictive model, offering significant improvements in performance for various machine learning tasks [19] [20] (Figure 2).

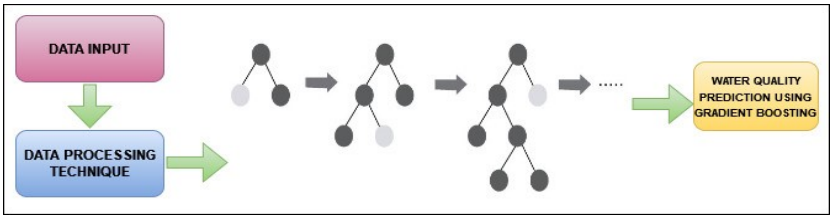

Figure 2. Proposed methodology for the gradient boosting classifier

5 Results

5.1 Confusion Matrix Analysis

The confusion matrix's off-diagonal cells represent the count of misclassifications. For instance, the cell located in the second row and first column reveals that the model mistakenly identified 8 instances as class 0, whereas they truly belonged to class 1 (see Figure 3).

Figure 3. Confusion Matrix Analysis

5.2 Classification Report Analysis

The model demonstrates superior performance for class 0 in comparison to class 1, evident through higher precision, recall, and F1-score metrics for class 0. The support column further highlights this, showing a greater number of data points in class 0 (510) than in class 1 (309). This phenomenon, known as class imbalance, poses a challenge for machine learning models. Despite the class imbalance, the model achieves an average precision (weighted avg) of 0.79 and a macro average F1-score of 0.77, indicating reasonably strong performance across both classes. An intriguing observation for class 1 is the higher precision (0.67) compared to recall (0.80). Precision measures the accuracy of positive predictions, while recall assesses the model's ability to identify actual positives. The elevated precision for class 1 implies that a significant portion of the model's predictions for this class were correct, yet it also suggests that many actual positives were missed (lower recall). Overall, the data indicates a potential bias towards class 0, and the model might benefit from strategies to counteract class imbalance. It excels in precision for class 1 but overlooks a considerable number of true positive instances in this class (see Figure 4).

	precision	recall	f1-score	support
0	0.86	0.77	0.81	510
1	0.67	0.80	0.73	309
accuracy			0.78	819
macro avg	0.77	0.78	0.77	819
weighted avg	0.79	0.78	0.78	819

Figure 4. Classification Report Analysis

6 Conclusion

In summary, this study examines the effectiveness of gradient boosting classifiers in predicting groundwater quality. By contrasting the performance of these classifiers with traditional machine learning algorithms and assessing their consistency across various datasets, the research showcases promising results with approximately 78% accuracy. Furthermore, the analysis of feature importance sheds light on the key factors influencing water quality variations, advancing predictive modeling techniques in this field. These findings hold significant implications for proactive management strategies aimed at sustainable water resource use and ecosystem preservation. In regions like India, where millions are affected by unsafe drinking water, these models provide valuable tools for informed decision-making in agricultural water management and resource allocation. By leveraging ensemble learning techniques and machine learning algorithms, this research highlights the transformative potential of data-driven approaches in tackling critical water management issues. Looking ahead, the continued exploration and application of advanced ML techniques, as illustrated in this study, promise to enhance water quality prediction and management practices, ultimately contributing to the protection of human health and aquatic ecosystems.

References

[1] Raheja, H., Goel, A. and Pal, M., 2024. A novel approach for prediction of groundwater quality using gradient boosting-based algorithms. ISH Journal of Hydraulic Engineering, pp.

[2] Olatinwo, S.O., Joubert, T.H. and Olatinwo, D.D., 2024. Water Quality Assessment Tool for On-site Water Quality Monitoring. IEEE Sensors Journal.

[3] Mohseni, U., Pande, C.B., Pal, S.C. and Alshehri, F., 2024. Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model. Chemosphere, p.141393.
[4] Helm, W., Zhong, S., Reid, E., Igou, T. and Chen, Y., 2024. Development of gradient boosting-assisted machine learning data-driven model for free chlorine residual prediction. Frontiers of Environmental Science & Engineering, 18(2), p.17.
[5] Krishnan, S. and Manikandan, R., 2024. Water quality prediction: A data-driven approach exploiting advanced machine learning algorithms with data augmentation. Journal of Water and Climate Change.
[6] Díaz-González, L. and Aguilar-Rodríguez, R.A., Aqua-P: A Machine Learning-Based Tool for Water Quality Assessment. Available at SSRN 4796085.
[7] Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S. and Mohanty, S.N., 2023, February. Water Quality Assessment Through Predictive Machine Learning. In International Conference on Intelligent Computing and Networking (pp. 77-88). Singapore: Springer Nature Singapore.
[8] Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., 2023. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Safety and Environmental Protection, 169, pp.808-828.
[9] Leggesse, E.S., Zimale, F.A., Sultan, D., Enku, T., Srinivasan, R. and Tilahun, S.A., 2023. Predicting optical water quality indicators from remote sensing using machine learning algorithms in tropical highlands of Ethiopia. Hydrology, 10(5), p.110.
[10] Rawat, P., Bajaj, M., Sharma, V. and Vats, S., 2023, March. A comprehensive analysis of the effectiveness of machine learning algorithms for predicting water quality. In 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) (pp. 1108-1114). IEEE.
[11] Ding F, Zhang W, Cao S, Hao S, Chen L, Xie X, Li W, Jiang M. Optimization of water quality index models using machine learning approaches. Water Research. 2023 Sep 1;243:120337.
[12] Wang, X., Li, Y., Qiao, Q., Tavares, A. and Liang, Y., 2023. Water quality prediction based on machine learning and comprehensive weighting methods. Entropy, 25(8), p.1186.
[13] Khan, M.S.I., Islam, N., Uddin, J., Islam, S. and Nasir, M.K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences, 34(8), pp.4773-4781.
[14] Malek, N.H.A., Wan Yaacob, W.F., Md Nasir, S.A. and Shaadan, N., 2022. Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. Water, 14(7), p.1067.
[15] Reshan, M.S.A., Gill, K.S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A. and Shaikh, A., 2023, May. Detection of pneumonia from chest X-ray images utilizing mobilenet model. In Healthcare (Vol. 11, No. 11, p. 1561). MDPI.
[16] Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., 2023. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Safety and Environmental Protection, 169, pp.808-828.
[17] Alnahit, A.O., Mishra, A.K. and Khan, A.A., 2022. Stream water quality prediction using boosted regression tree and random forest models. Stochastic Environmental Research and Risk Assessment, 36(9), pp.2661-2680.
[18] Gill, K.S., Tuteja, G., Anand, V., Gupta, R. and Hsiung, P.A., 2023, December. Water Quality Prediction using Classification techniques on XGBoost, KNeighbors, SVC, Random Forest, AdaBoost, and GaussianNB Classifier. In 2023 Global Conference on Information Technologies and Communications (GCITC) (pp. 1-4). IEEE.
[19] Alnahit, A.O., Mishra, A.K. and Khan, A.A., 2022. Stream water quality prediction using boosted regression tree and random forest models. Stochastic Environmental Research and Risk Assessment, 36(9), pp.2661-2680.
[20] Lu, H. and Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, p.126169.