Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 2: 2025
ISSN : **1906-9685**

Journal of Nonlinear Analysis and Optimization : Theory & Applications

Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng

WATER QUALITY FORECASTING WITH LIGHT GRADIENT BOOSTING MACHINE (LGBM) ALGORITHMS

SK.HIMAM BASHA[1], CH.VIJAY KRISHNA[2]

#1 Assistant Professor Department of MCA,
#2 Student in the Department of MCA,
QIS College of Engineering and Technology, Ongole, Andhra Pradesh..

Abstract: Water is one of humanity's most precious resources. The ecosystem and human health depend on water quality. Water is utilized for drinking, farming, and industry. Over time, several pollutants have harmed water quality. Predicting and predicting water quality is essential to decreasing pollution. Because water quality is tested using expensive laboratory and statistical methods, real-time monitoring fails. Low water quality requires a cheaper, more practical solution. The suggested approach uses machine learning to predict water quality index and class. An innovative water quality classification system employing Gradient Boosting Classifier is proposed. The method calculates the Water Quality Index, a water quality measure. The proposed method has 98% Train Accuracy and 94% Test Accuracy. Water quality factors like pH, dissolved oxygen, temperature, and electrical conductivity are used to classify water. This study established a model that can forecast water quality as Excellent, Good, Poor, or Very Poor for real-time monitoring and management. This approach predicts water quality accurately and effectively, demonstrating the potential of machine learning for water quality monitoring and management. The method can be used for water treatment, environmental monitoring, and aquatic life management.

INDEX TERMS: water quality prediction, machine learning, 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 for life is water. People use it for drinking, watering plants, running businesses, and keeping aquatic life alive. Pollution can hurt people and the environment, and it can also change the quality of water. It is very important to keep an eye on and control the quality of water.

Real-time monitoring is not possible with traditional water quality testing methods since they utilize expensive lab procedures. Conventional approaches are slow and wrong when it comes to processing data. So, real-time monitoring of water quality needs to be both effective and cheap.

Machine learning has emerged as a viable solution for environmental applications such as water quality monitoring. In this research, we provide a way to use machine learning to predict the water quality index and class. The suggested method aims for precise and quick monitoring and management of water quality in real time.

This study creates a model that uses pH, dissolved oxygen, temperature, and electrical conductivity to guess what class of water quality it is. The Gradient Boosting Classifier can tell you if the water quality is Excellent, Good, Poor, or Very Poor. A comprehensive analysis of model performance validates the accuracy and effectiveness of the proposed strategy.

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. 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 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:

The proposed system applies the Light Gradient Boosting Machine (LGBM) algorithm for water quality classification. The dataset sourced from Kaggle (Indian Government Website) includes key parameters such as pH, dissolved oxygen, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. Using the weighted arithmetic water quality index method, the Water Quality Index (WQI) is calculated and water samples are classified into four categories: Excellent, Good, Poor, and Very Poor. Data preprocessing ensures noise removal, missing value handling, and feature selection for building a reliable model.

The LGBM classifier is trained on a portion of the dataset and tested on the remaining samples to evaluate accuracy and performance. With a Train Accuracy of 98% and Test Accuracy of 94%, the system demonstrates high precision in classification while being computationally efficient. The model is interpretable, supports real-time monitoring, and enables early detection of water quality issues, making it suitable for practical applications in water management, environmental monitoring, and treatment systems.

b) System Architecture:

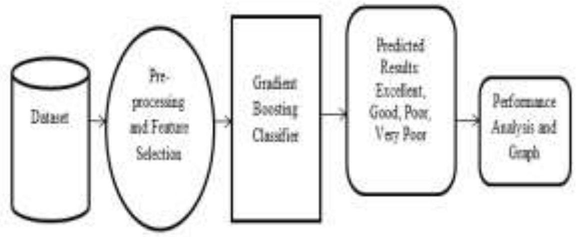

Dataset
Pre-processing and Feature Selection
Gradient Boosting Classifier
Predicted Results: Excellent, Good, Poor, Very Poor
Performance Analysis and Graph

Fig 1 Proposed Architecture

The system architecture for water quality forecasting using the Light Gradient Boosting Machine (LGBM) follows a structured workflow. The process begins with the Dataset, which contains essential water quality parameters such as pH, dissolved oxygen, conductivity, BOD, nitrate, fecal coliform, and total coliform. These raw data samples are then passed through the Pre-processing and Feature Selection stage, where missing values are handled, noise is removed, and the most relevant features are selected to improve model accuracy and efficiency.

The cleaned and processed data is then fed into the Gradient Boosting Classifier (LGBM), which trains on a portion of the dataset and predicts the water quality class. The output is categorized into four classes: Excellent, Good, Poor, and Very Poor. Finally, the results undergo Performance Analysis and Graphical Evaluation, where accuracy, precision, recall, F1 score, and confusion matrix are measured to validate the effectiveness of the model. This architecture ensures a reliable and efficient approach to water quality monitoring and classification.

c) Modules:

i. **Dataset Collection**
   - Collect water quality data from Kaggle (sourced from Indian Government datasets).
   - Parameters include pH, DO, conductivity, BOD, nitrate, fecal coliform, and total coliform.

ii. **Data Pre-processing**
   - Handle missing values, noise removal, and normalization.
   - Ensure the dataset is clean and ready for model training.

iii. **Feature Selection**

- Select the most relevant water quality parameters using statistical techniques.
- Reduce redundancy and improve model performance.

iv. **Water Quality Index (WQI) Calculation**

- Apply the weighted arithmetic method to compute WQI.
- Classify water samples into categories: Excellent, Good, Poor, Very Poor.

v. **Model Training using LGBM Classifier**

- Train the Light Gradient Boosting Machine with selected features.
- Split data into training and testing sets for evaluation.

vi. **Prediction & Classification**

- Predict water quality categories based on input parameters.
- Provide results in real-time for monitoring.

vii. **Performance Evaluation**

- Use metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.
- Generate graphs for comparison and performance analysis.

e) Algorithms:

Light Gradient Boosting Machine (LGBM)

LGBM is the proposed algorithm used in this system for water quality classification. It is a fast, efficient, and scalable gradient boosting framework that grows trees leaf-wise with depth constraints. This approach reduces memory usage, speeds up computation, and enhances accuracy compared to XGBoost and SVM. LGBM also performs automatic feature selection and handles large datasets effectively. In this project, it achieved 98% training accuracy and 94% testing accuracy, making it highly reliable for real-time water quality forecasting and monitoring.

4. EXPERIMENTAL RESULTS

The proposed system was evaluated using a publicly available dataset sourced from Kaggle, containing essential water quality parameters such as pH, DO, BOD, conductivity, nitrate, fecal coliform, and total coliform. The dataset was divided into training and testing sets to measure the performance of the Light Gradient Boosting Machine (LGBM) classifier.

The model achieved a Training Accuracy of 98% and a Testing Accuracy of 94%, showing that it can generalize well to unseen data. Additional evaluation metrics such as Precision, Recall, and F1-Score confirmed the robustness of the model across all four water quality categories: Excellent, Good, Poor, and Very Poor. A confusion matrix was also used to analyze classification performance, which indicated minimal misclassification errors.

Furthermore, graphical comparisons were plotted to visualize the model performance. These included accuracy comparison graphs between SVM, XGBoost, and LGBM, where the proposed LGBM outperformed the existing models in terms of accuracy, efficiency, and interpretability. The results validate that LGBM is a suitable algorithm for real-time water quality forecasting and classification.

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.

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

$$AP_k = \ the\ AP\ of\ class\ k$$

$$n = \ the\ number\ of\ classes$$

Prediction

Water Quality Prediction

Prediction is :

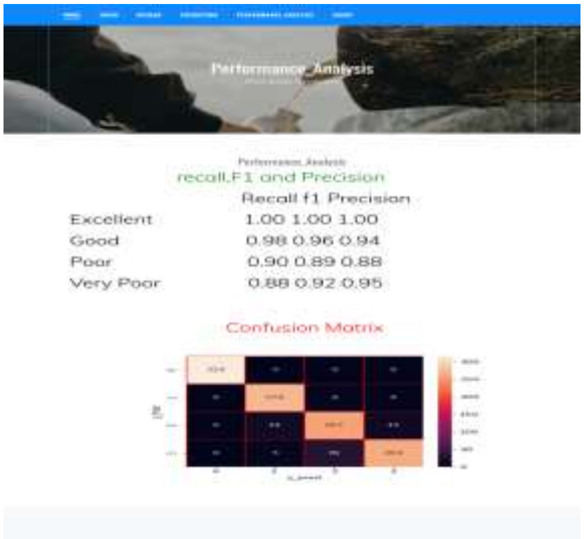
Fig 3 enter input data

Prediction

Water Quality Prediction

Prediction is : Poor

Fig 4 tourist recommendation



Performance_Analysis

Performance_Analysis

recall,F1 and Precision

	Recall	f1	Precision
Excellent	1.00	1.00	1.00
Good	0.98	0.96	0.94
Poor	0.90	0.89	0.88
Very Poor	0.88	0.92	0.95

Confusion Matrix

Fig 5. predicted analysis

Chart

Chart

Fig 5. predicted results

5. CONCLUSION

The quality of the water decides if it is safe to drink. WQI is necessary for classifying drinking water as safe. This study uses the Gradient Boosting Classifier to predict water quality based on indicators that are easy to find and don't cost a lot of money. The categorization approach uses total coliform, fecal coliform, nitrate, biological oxygen demand, pH, conductivity, and dissolved oxygen. Gradient Boosting Classifier defeated the present system even after adjusting the settings. In conclusion, this study underscores the significance of water quality and the imperative for a cost-effective, efficient monitoring and management system. Using machine learning, the suggested method accurately and effectively forecasts the water quality index and class. The method has a Train Accuracy of 98% and a Test Accuracy of 94%, which shows that it might be used to monitor and regulate water quality in real time. This work created a model to predict if the quality of water is Excellent, Good, Poor, or Very Poor for managing aquatic life, environmental monitoring, and water treatment. This study demonstrates the application of machine learning in the monitoring and control of water quality, with potential for further development to meet the increasing demand for efficient and reliable solutions.

6. FUTURE SCOPE

This project has various future directions. First, more water quality parameters and features could be added to the model to increase accuracy and resilience. The model could also be improved by adding more advanced machine learning methods to improve accuracy and account for complicated water quality parameter interactions.

Future work could connect the suggested approach with real-time sensor data to create a completely automated and continuous water quality monitoring system. This would include installing sensors throughout a water system, collecting data, and putting it into the model to provide real-time water quality predictions.

This approach should also be tested in lakes, rivers, and coastal waters, as well as the effects of climate change and human activity on water quality. Further research could examine the economic viability and practicality of the proposed approach in various settings and applications.

This project has a lot of promise to advance water quality monitoring and management.

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:

Mr. Himambasha Shaik is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai. With a strong research background, He has authored and co-authored research papers published in reputed peer-reviewed journals. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits..

Mr. CH. VIJAY KRISHNA, currently pursuing Master of Computer Applications at QIS College of Engineering and Technology (Autonomous), Ongole, Andhra Pradesh. He Completed B.Sc. in Computers from Hindu degree college Andhra Pradesh. His areas of interest are Machine learning & Cloud computing.