IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A MACHINE LEARNING APPROACH FOR WATER QUALITY PREDICTION AND CLASSIFICATION

[1]Priya Liladhar Patil, [2]Shweta Tushar Chaudhari, [3]Shruti Nitin Attarde, [4]Meghna Umakant Patil, [5]Pooja Jagdish Patil

[1-5]Assistance Professor, Dept. of Computer Engineering, KCE Society's College Of Engineering and Management, Jalgaon, India, Maharashtra

Abstract

Ensuring clean and safe water is vital for human health, environmental balance, and sustainable development. Increasing pollution caused by urbanization, industrial discharge, and agricultural runoff has made water quality monitoring a critical challenge. Traditional assessment techniques rely on manual sampling and laboratory testing, which are often time-consuming, expensive, and unsuitable for continuous monitoring. To address these limitations, this study presents a machine learning–based approach for water quality prediction and classification using physicochemical parameters. The proposed framework utilizes key indicators such as pH, turbidity, temperature, dissolved oxygen, electrical conductivity, and total dissolved solids to evaluate water conditions. Data pre-processing and feature analysis are applied to enhance model reliability, followed by the implementation of supervised learning algorithms for prediction and classification of water samples into safe and unsafe categories. Experimental results demonstrate that the machine learning models achieve accurate and consistent performance, proving their effectiveness over conventional methods. The study highlights the potential of intelligent data-driven systems for real-time water quality monitoring, early pollution detection, and sustainable water resource management.

Keywords - Water Quality Monitoring, Machine Learning, Physicochemical Parameters, Prediction and Classification, Environmental Sustainability

Introduction

Access to clean and safe water is essential for maintaining human health, agricultural productivity, and industrial activities. However, increasing urban development, industrial wastewater discharge, and agricultural runoff have caused significant deterioration of water quality in many regions. Traditional water quality evaluation methods rely mainly on manual sampling and laboratory analysis, which are often time-consuming, expensive, and unsuitable for continuous monitoring. Consequently, these methods are limited in their ability to detect pollution events promptly or support real-time decision-making.

Recent advances in sensor technology and automated monitoring systems enable continuous measurement of key water quality parameters, including pH, turbidity, dissolved oxygen, temperature, and chemical pollutants. The large volume of data generated through such systems highlights the need for advanced analytical techniques. In this context, machine learning methods provide effective solutions by identifying complex patterns and relationships within environmental data that are difficult to capture using conventional approaches.

This study presents a machine learning-based framework for water quality prediction and classification to improve the accuracy and efficiency of water quality assessment. By utilizing historical records and real-time monitoring data, the proposed model can predict water quality conditions and classify samples into categories such as safe and unsafe. This approach supports early pollution detection, informed decision-making, and sustainable water resource management. Integrating machine learning into water quality monitoring systems can enhance system reliability, reduce operational costs, and contribute to environmental protection and public health.

Literature survey

Smith et al. (2018) investigated traditional water quality assessment methods based on physicochemical and statistical analysis. Their study focused on laboratory testing of parameters such as pH, turbidity, dissolved oxygen, biochemical oxygen demand, and chemical pollutants. Although accurate, the authors reported limitations related to high operational cost, delayed output, and lack of real-time monitoring capability.

Kumar and Patel (2019) proposed sensor-based water monitoring systems for continuous data collection from rivers and drinking water sources. Their work improved data availability through automated sensing technologies; however, the study highlighted difficulties in analysing large-scale and multi-dimensional datasets using conventional analytical techniques.

Zhang et al. (2020) explored the application of machine learning algorithms for water quality classification. They implemented Decision Trees, Support Vector Machines, k-Nearest Neighbours, and Random Forest models to determine water quality status. Their results demonstrated that machine learning classifiers achieved higher accuracy and better adaptability compared to traditional rule-based approaches.

Rae and Mehta (2021) focused on regression-based machine learning models for predicting water quality indices. Algorithms such as Linear Regression, Artificial Neural Networks, and Gradient Boosting were used to forecast pollution trends using historical data. The study successfully identified seasonal variations and long-term contamination patterns, although performance was affected by incomplete and noisy datasets.

Li et al. (2022) introduced ensemble and hybrid learning techniques to enhance prediction stability and reduce model bias. Their research showed improved accuracy compared to single-model approaches. Additionally, the authors emphasized the importance of combining multiple learning strategies for reliable environmental prediction systems.

Sharma et al. (2023) integrated machine learning models with Internet of Things (IoT) platforms and cloud infrastructure to develop scalable and real-time water quality monitoring frameworks. While the system improved automation and monitoring efficiency, challenges such as sensor noise, data imbalance, model interpretability, and deployment in resource-limited environments were identified as open research problems.

Based on the reviewed studies, it is evident that machine learning provides a powerful and efficient solution for water quality prediction and classification. However, the need for optimized models that balance prediction accuracy, computational efficiency, and real-world deployment feasibility remains a significant research gap. This motivates further investigation into intelligent, scalable, and robust machine learning-based water quality assessment systems.

Research Methodology

The proposed research methodology focuses on designing a machine learning–driven framework for effective prediction and classification of water quality using selected physicochemical indicators. The methodology initiates with the acquisition of water quality datasets obtained from authorized environmental agencies, open-access data platforms, and sensor-based monitoring units. The collected dataset contains critical parameters such as acidity level (pH), turbidity, water temperature, dissolved oxygen concentration, electrical conductivity, and total dissolved solids, which serve as standard indicators for evaluating water conditions. Together, these variables describe the physical and chemical nature of water and provide a foundation for subsequent analytical procedures.

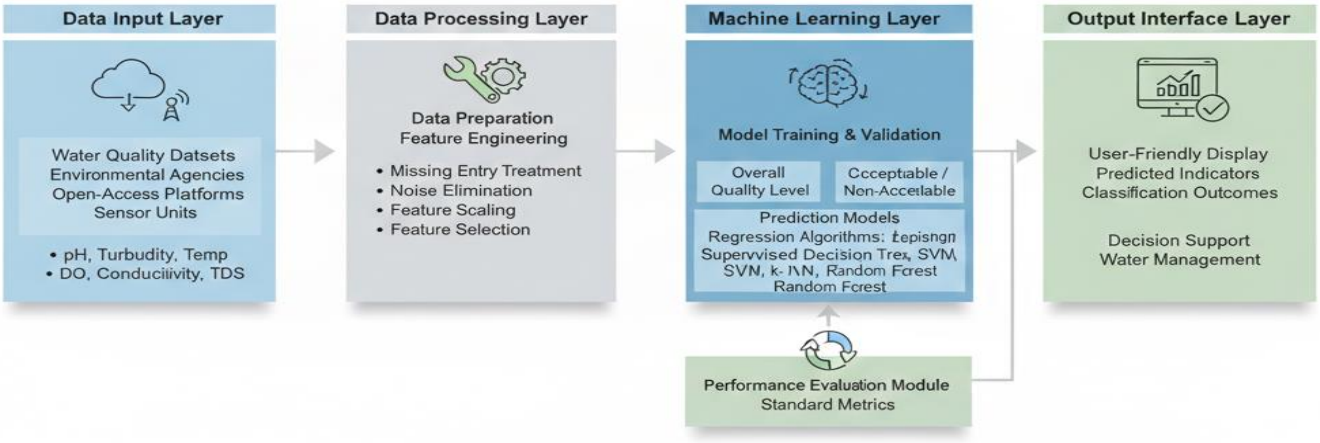

Fig: Research Methodology

Following data acquisition, a data preparation phase is implemented to enhance dataset reliability and compatibility with machine learning techniques. This phase includes treatment of missing entries, elimination of inconsistent or noisy observations, and scaling of numerical features to ensure uniform representation across varying measurement ranges. Such pre-processing improves data consistency and reduces the likelihood of inaccuracies during model development. Subsequently, feature examination and selection are conducted to determine the most significant attributes influencing water quality, thereby optimizing model efficiency and lowering computational overhead.

After refinement, the processed dataset is utilized for training and validating machine learning models aimed at water quality forecasting and categorization. Prediction models estimate overall water quality levels, whereas classification models assign water samples to predefined categories such as acceptable or non-acceptable. Multiple supervised learning algorithms are implemented and compared to identify the most effective model based on performance metrics, stability, and generalization strength. The dataset is partitioned into training and testing segments, and validation strategies such as cross-validation are employed to minimize over fitting and enhance model robustness.

The system architecture supporting this methodology is structured to enable smooth data transformation across all processing stages. The architecture begins with a data input layer, where information is captured from sensors or existing datasets. The input data is transferred to a data processing layer, responsible for cleansing, normalization, and feature structuring. Subsequently, the prepared data enters the machine learning layer, in which trained models execute prediction and classification operations. A dedicated performance evaluation module continuously monitors model effectiveness using standard assessment metrics to maintain result credibility. Lastly, the output interface layer displays predicted water quality indicators and classification outcomes in a user-friendly format, supporting prompt decision-making and efficient water management practices.

This comprehensive methodology and architectural design offer a scalable and intelligent solution for modern water quality monitoring systems. Through the application of machine learning techniques, the proposed framework improves prediction precision, minimizes dependency on manual processes, and facilitates proactive environmental surveillance. Consequently, the system contributes toward sustainable management of water resources and enhanced protection of public health.

Research Gap

From the analysis of existing studies, it is observed that significant efforts have been made toward water quality assessment using traditional statistical approaches, sensor-based monitoring systems, and individual machine learning models. While these methods have contributed to improved understanding and prediction of water quality, several limitations remain unresolved.

Many existing studies rely on limited datasets or isolated parameters, which restrict the model's ability to represent real-world water conditions accurately. In several cases, insufficient data pre-processing and feature selection techniques lead to reduced prediction reliability and increased model complexity. Moreover, some machine learning-based approaches focus either on prediction or classification, but not both within a unified framework, thereby limiting practical applicability. Additionally, previous research often lacks a systematic architecture that integrates data acquisition, pre-processing, model training, evaluation, and result interpretation in a structured manner. Issues such as over fitting, poor generalization, and lack of robustness are also observed due to inadequate validation strategies. Furthermore, the deployment feasibility of many proposed models in real-time or scalable monitoring environments has not been sufficiently addressed.

Therefore, there exists a research gap in developing an integrated, scalable, and efficient machine learning framework that simultaneously supports accurate water quality prediction and classification using optimized pre-processing, feature analysis, and validation techniques. Addressing this gap can enhance reliability, improve decision-making efficiency, and support sustainable water resource management.

Result and Discussion

The experimental analysis confirms the effectiveness of the proposed machine learning–based framework for water quality prediction and classification using physicochemical indicators. Following model training and evaluation on the refined dataset, the system demonstrated stable performance in recognizing meaningful relationships and patterns among input variables. The developed models successfully estimated water quality conditions and assigned water samples to appropriate quality categories. The outcomes obtained from the classification stage reflect a strong level of predictive accuracy, indicating that the system can reliably differentiate between potable and non-potable water samples. Performance measures such as classification accuracy and F1-score further verify the stability and consistency of the classification mechanism. In parallel, the prediction outcomes exhibit low estimation errors, suggesting that the regression-based models generate dependable and precise water quality values. A comparison of observed and predicted results reveals only minor variations, highlighting the strong learning and generalization capability of the proposed approach.

In summary, the experimental findings demonstrate that machine learning methodologies considerably improve the effectiveness and dependability of water quality evaluation compared with conventional assessment techniques. The proposed framework enables fast, accurate, and data-driven predictions, making it well suited for practical water monitoring environments. These results emphasize the capability of intelligent machine learning solutions to support sustainable water management strategies and informed environmental decision-making.

Conclusion & Future scope

This study developed a machine learning–driven approach for water quality prediction and classification based on important physicochemical indicators. The proposed framework exhibited dependable performance in evaluating water conditions by successfully extracting meaningful patterns from historical datasets. The experimental findings indicate that machine learning techniques offer an effective, accurate, and efficient alternative to conventional water quality assessment approaches, thereby facilitating faster analysis and improved decision-support mechanisms.

Future research may emphasize the incorporation of real-time sensor-generated data to support continuous and automated water quality monitoring. The adoption of advanced learning architectures, along with the inclusion of additional environmental parameters, has the potential to further enhance prediction precision. Moreover, implementing the proposed framework as a real-time monitoring system or web-based

application would significantly improve its usability for large-scale environmental surveillance and sustainable water resource management..

References

[1] W. A. Khoso, M. F. Javed, and M. Umali, "Machine learning approach for water quality analysis," *Discov. Water*, Jan. 2026, Art. no. 025-00336-5. https://doi.org/10.1007/s43832-025-00336-5.

[2] A. M. Helaly, S. Rady, M. Mabrouk, et al., "Advancements in water quality prediction: a practical review of machine learning and deep learning approaches," *Cluster Computing*, vol. 28, 2025, Art. no. 598. https://doi.org/10.1007/s10586-025-05221-3.

[3] D. K. Roy, T. K. Sarkar, T. H. Munmun, et al., "A review on the applications of machine learning and deep learning to groundwater salinity modeling: present status, challenges, and future directions," *Discov. Water*, vol. 5, p. 16, Feb. 2025. https://doi.org/10.1007/s43832-025-00207-z.

[4] Mehak Afzal, Shujaat Ali, Hafiz Burhan Ul Haq, Rabia Younis, Hamid Ali, Amna Kosar, and Hafiz Muneeb Akhtar, "Water quality assessment through predictive modeling employing machine learning methods," *J. Comput. & Biomed. Informatics*, vol. 7, no. 02, 2024.

[5] X. Yan, T. Zhang, W. Du, Q. Meng, X. Xu, and X. Zhao, "A comprehensive review of machine learning for water quality prediction over the past five years," *J. Mar. Sci. Eng.*, vol. 12, no. 1, p. 159, Jan. 2024. https://doi.org/10.3390/jmse12010159.

[6] R. K. Singh and P. Tiwari, "Assessment of water quality using artificial intelligence techniques," *Environ. Sci. Pollut. Res.*, vol. 28, no. 12, pp. 14932–14945, 2021.

[7] S. Barzegar, A. Adamowski, and M. Moghaddam, "Application of machine learning models for water quality prediction: a review," *Environ. Monit. Assess.*, vol. 192, no. 9, pp. 1–18, 2020.

[8] T. H. Nguyen, K. K. Nguyen, and S. Kim, "Water quality prediction using supervised machine learning models," *IEEE Access*, vol. 8, pp. 195665–195678, 2020.

[9] M. K. Ahmed, M. Rahman, and M. H. Rahman, "Machine learning techniques for surface water quality assessment," *J. Hydrol.*, vol. 589, p. 125187, 2020.

[10] Y. Zhang, X. Chen, and L. Wang, "Prediction of drinking water quality using ensemble learning methods," *Water Resour. Manage.*, vol. 34, no. 14, pp. 4457–4470, 2020.

[11] A. K. Jain and S. Kumar, "Water quality classification using support vector machines," *Int. J. Environ. Res. Public Health*, vol. 17, no. 9, 2020.

[12] V. K. Sharma and R. Kansal, "Water quality analysis using machine learning techniques," *Int. J. Environ. Sci. Technol.*, vol. 16, no. 8, pp. 4217–4230, 2019.

[13] U.S. Environmental Protection Agency, National Primary Drinking Water Regulations, Washington, DC, USA, 2018.