Grenze International Journal of Engineering and Technology, January Issue

# A Data-Driven Approach for Microbial Contamination Prediction in Indian Rivers using BOD with Interpretable Machine Learning

Hrishikesh Tendulkar[1] and Sachin Bhoite[2]
[1-2]Dr. Vishwanath Karad MIT-World Peace University Department of Computer Science and Applications, Pune, India
Email: hrishikesh.tendulkar@mitwpu.edu.in, sachin.bhoite@mitwpu.edu.in

***Abstract*— Water quality research is essential to the sustainability of the environment and the safety of people, especially in countries like India where surface water bodies are under increasing stress from human interference. This study analyzes the relationship between physicochemical parameters such as pH, temperature, (DO) dissolved oxygen, (BOD) biochemical oxygen demand, and microbial contamination (as measured by Total Coliform counts).and electrical conductivity, in India's major rivers and states. The study examines regional variation in important water quality indicators and pinpoints regions of concern where water is most susceptible to contamination using a dataset that includes 534 observations from 18 Indian states. According to the findings, states like Maharashtra, Bihar, Uttar Pradesh, and Karnataka have higher monitoring data frequencies, which may reflect in-creased industrialization, urbanization, and awareness of water quality problems. States like Delhi, Chhattisgarh, and Kerala, on the other hand, have many fewer reported observations, which may indicate either underreporting or a lack of monitoring infrastructure. Low DO values were seen in regions with high BOD and Total Coliform levels, suggesting oxygen depletion brought on by microbial and organic pollution. The findings highlight the immediate need for real-time data gathering systems, a standardized national framework for water monitoring, and legislative actions to address disparate regional capacity. This study highlights how important it is to combine microbiological evaluations with physicochemical studies to provide a complete picture of water quality.**

***Keywords*— Water Quality Monitoring, Physicochemical Parameters, Microbial Contamination, XAI, explainable AI, SHAP, Lime.**

## I. Introduction

Water quality is a significant environmental and public health issue affecting aquatic life, hu-man health, agriculture, and industrial operations. Monitoring factors like temperature, dis-solved oxygen (DO), pH, conductivity, biochemical oxygen demand (BOD), nitrate and nitrite concentrations, and microbial contamination indicate the suitability of water for various applications and potential pollution levels. Rivers, vital for ecological services, livelihoods, and biodiversity, are facing declining water quality due to industrialization, urbanization, agricultural runoff, and household garbage discharge. Major Indian rivers like the Ganga, Godavari, Yamuna, Krishna, and Brahmaputra are among the most polluted, highlighting the need for pollution control management and sustainable water resource management.

*Grenze ID: 01.GIJET.12.1.187*

Monitoring the water quality of these rivers is necessary for pollution control and sustainable water resource management.
The water quality in Indian rivers varies significantly due to various factors such as temperature, oxygen solubility, conductivity, biochemical oxygen demand (BOD), and nitrate and nitrite levels. Dissolved oxygen, an essential measure of water quality, ranges from 1.2 mg/L in Delhi Yamuna River regions to higher levels of 9.0 mg/L in some Ganga locations. pH values usually range between 7.0 to 8.5, which is neutral to slightly alkaline for freshwater systems.
Conductivity, a measure of water's electrical properties, is correlated with ion concentration, with higher values in industrial areas indicating pollution and dissolved salts. Biochemical ox-ygen demand (BOD) shows how much oxygen microbes need to break down organic materials, with heavy organic contamination often from sewage and industrial waste. Nitrate levels can range from very low at some stations to as high as 9.0 mg/L at others, causing eutrophication and harmful algal blooms. Urban and downstream regions have a high level of microbial pollu-tion, as measured by faeces and total coliform levels. The regional variation in water quality metrics highlights the impact of both point and non-point pollution sources, such as industrial effluents, agricultural runoff, untreated sewage, and religious or cultural activities. Microbial contamination is a major problem in urban areas, but nutrient burdens are higher in rural re-gions. Pollutant concentration and dilution are also influenced by sea-sonal variations and riv-er flow dynamics.
Effective water quality management requires implementing treatment and conservation measures, identifying contamination sources, and conducting thorough monitoring. Policies that lower industrial discharge, enhance sewage treatment facilities, encourage sustainable farming methods, and increase public awareness are essential. Scientific study and the applica-tion of cutting-edge technology like real-time sensors, GIS mapping, and remote sensing can enhance water quality evaluation and management.
A prediction system using machine learning and explainable artificial intelligence (XAI) is a powerful tool for forecasting environmental indicators like river water quality. The system us-es regression algorithms like Linear Regression, SVR, Decision Tree, Random Forest, and XGBoost to identify patterns and relationships in historical water quality data. eXplainable AI techniques like SHAP and LIME are integrated to assure transparency and trust in predictions. These tools help interpret the model's outputs, highlighting the contribution of each feature to the prediction. This approach allows stakeholders like environmental agencies and policymak-ers to rely on accurate forecasts and understand their reasoning, promoting informed decision-making and early intervention strategies.

### *A. Motivation*

The quality of India's water is a major problem, particularly for its surface water bodies, which are essential for industry, agriculture, and drinking. The combined effect of chemical proper-ties and bacterial growth poses major hazards to the environment and public health.
To improve monitoring systems and guide resource and policy management strategies.

### *B. Problem definition*

The impact of physicochemical characteristics on microbiological contamination in surface water is not well understood. Inconsistent water quality monitoring in India's many states ex-acerbates the issue. By examining observational data, this study aims to close that gap by iden-tifying trends and forecasting correlations between microbial loads and water quality indica-tors.

### *C. Innovative content*

Integrating machine learning with data-driven research to analyse water quality metrics. For-mulation of a Water Quality Index (WQI) that integrates microbial and physicochemical in-formation based on BOD. Model interpretability is achieved through the application of SHAP and LIME approaches to comprehend feature contributions.

## II. Related Work or Literature Studies

Ref.[1] A 2023 survey analysed machine learning (ML) and deep learning (DL) models for wa-ter quality prediction, emphasizing their role in classifying physicochemical parameters such as pH, electrical conductivity, BOD and DO.
The study highlighted preprocessing and feature extraction as critical steps for improving model accuracy, with DL architectures like convolu-tional neural networks (CNNs) outperforming traditional ML models in handling nonlinear da-ta patterns.

Ref.[2] A survival study (2022) reviewed 40 papers on AI-driven water quality prediction, in-cluding ANN, fuzzy logic, and support vector machines (SVM). The analysis revealed that ANN models achieved 85–92% accuracy in predicting contaminants in groundwater, while SVM excelled in classifying surface water quality.
Ref.[3] A 2023 survey applied supervised ML techniques to predict parameters like alkalinity and phosphate levels. The study found that gradient-boosted decision trees outperformed line-ar regression by 18–22% in multivariate analysis, emphasizing the importance of feature selec-tion.
Ref.[4] A 2025 study combined PCA and Lasso regression with stacking ensemble models to predict pH values across water stations in Georgia. The hybrid approach reduced mean abso-lute error (MAE) by 15% compared to standalone models, demonstrating the efficacy of fea-ture engineering in improving predictions.
Ref.[5] A 2020 study employed random forest models to predict DO levels in the Potomac River. Feature importance analysis revealed water temperature and pH as top predictors, with RF reducing root mean square error (RMSE) by 27% compared to linear regression.
Ref.[6] employed Multivariate Linear Regression (MLR) to predict BOD levels in wastewater using parameters such as DO, Nitrogen, Faecal Coliform, and Total Coliform. The model achieved an RMSE of 6.74 mg/L and a correlation coefficient (r) of 0.60, indicating moderate predictive capability.
Ref.[7] reviewed various AI applications in water quality assessment across different coun-tries. They highlighted that models like Decision Trees (DT), Extreme Gradient Boosting (XGBoost), and Random Forests (RF) provided robust results in predicting Water Quality Indi-ces (WQIs), with XGBoost achieving an $R^2$ of 0.989 in Vietnam's La Buong River.
Ref.[8] A study proposed a soft sensor model integrating ML algorithms for real-time BOD es-timation using edge computing. This approach addressed challenges in sensor accuracy and re-liability, offering a cost-effective solution for continuous water quality monitoring
Ref.[9] Explainable AI (XAI) is necessary for trust and the development of machine learning models. It provides transparency in how models make decisions, allowing researchers and us-ers to discover potential biases and limitations. This paper examines two XAI methods: Shap-ley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). The study aims to examine the decision-making process of a machine learning model and its development. The paper discusses the approaches and frameworks of both LIME and SHAP and evaluates their performance in predicting the model's outcomes.
Ref.[10] This research explores the current state of Explainable Artificial Intelligence (XAI), examining existing literature, methods, and building trustworthiness. It proposes methodolog-ical suggestions for future research and highlights critical issues for further exploration. The research is an ongoing endeavour, subject to evolve and adapt as the field progresses.

## III. Research Methodology and Analysis

The methodology adapted from Ref.[11] involves a structured ML pipeline comprising data preprocessing, model selection based on prior accuracy benchmarks, and training with robust evaluation. While originally applied to image-based garbage classification, this approach can be adapted for regression tasks in water quality prediction using biochemical indicators such as BOD. The structured pipeline supports data-driven optimization, and by integrating SHAP and LIME, the model outputs become interpretable, ensuring transparency in identifying key water quality contributors.

### *A. Area of study*

Surface water bodies in 18 main Indian states, which span a range of geographical, meteoro-logical, and human-induced conditions, are the research subject. These include less regulated or rural areas like Kerala, Chhattisgarh, and Delhi, as well as highly urbanised and industrial-ised regions like Maharashtra, Bihar, Uttar Pradesh, and Karnataka. Millions of people depend on the rivers under study—the Ganga, Godavari, Yamuna, Krishna, and Brahmaputra—but they are also heavily impacted by industrial, urban, and agricultural pollution.

### *B. Data Collection*

534 observations from several locations for monitoring along major Indian rivers make up the dataset. Temperature, pH, DO, BOD, electrical conductivity, and microbial contamination measured by Total Coliform counts are all included in each observation.
Spatial and temporal diversity in water quality across the nation is reflected in the data, which cover a variety of environmental contexts. Furthermore, the dataset exhibits unequal geographical representa-tion, with higher data frequencies from states with enhanced monitoring systems or larger-scale pollution issues

*C. Data Preprocessing*

Many essential steps were taken during data preparation to ensure the dataset's dependability and usability:

- Handling Missing Values: To maintain data integrity for upcoming modelling and analysis, rows with null or missing values were eliminated.
- Data Cleaning: To avoid distorted findings, outliers and inconsistent entries were exam-ined and, if required, eliminated or fixed.
- Standardisation and Scaling: To guarantee comparability and make it easier to calculate composite indices like the Water Quality Index (WQI), variables like pH, DO, BOD, and conductivity were normalised or scaled for some analyses.
- Feature Engineering: Using well-established formulas that combine scaled values of the physicochemical and microbiological characteristics, a Water Quality Index column was calculated, offering a single metric for evaluating the overall quality of the water.
- Data Structuring: A structured tabular format was created using the cleaned and processed data, with columns signifying distinct water quality metrics and each row indicating a dis-tinct sampling occurrence

*D. Modelling*

Based on physicochemical factors, regression models were used to forecast microbial contam-ination (Total Coliform counts).

Cleaning, scaling, and creating a WQI were all part of the data preprocessing procedure.

R2 and RMSE ratings were used to assess models such as Linear Re-gression and sophisticated machine learning approaches. Model validation was aided by visu-alization tools such as real vs. predicted graphs and correlation heatmaps.

*E. Model Evaluation*

The model's performance was evaluated using:

- The strength of the Total Coliform prediction based on a few characteristics is indicated by R2 scores.
- To measure the accuracy of a prediction, use the Root Mean Squared Error (RMSE).
- SHAP values and residual analysis are used to investigate variable significance and model behaviour.
- Edge Regression: By that include L2 regularization, this linear model with diminishing co-efficients avoids excessive fitting. It works best when qualities are related.
- Like Ridge, Lasso Regression uses L1 regularization, which essentially selects features by making some coefficients smaller to zero.
- Decision Tree: A DT regression is a non-linear model that separates the data into branches based on feature values. Overfitting is a possibility even though it is simple to compre-hend.
- Random Forest Regressor: A collection of decision trees that minimizes overfitting and averages forecasts to enhance generalization.
- XGBoost Regressor: A strong and effective gradient boosting model that fixes prior faults by building trees sequentially. In prediction tasks, it frequently produces the most effec-tive results.
- The R^2 score (Coefficient of Determination): This measure shows the extent to which the model explains the variance of the target variable. An R^2 value closer to 1 shows better performance.
- Root Mean Squared Error: RMSE measures the average size of prediction errors. Lower numbers show more accurate forecasts.
- To guarantee a fair comparison of models, these metrics must be computed on the test da-taset. The model with the lowest RMSE and the highest R2 performed most well

*F. Model Interpretation*

- Model interpretation increases a machine learning model's transparency and reliability by illuminating how and why it makes predictions. The best-performing model in this study will be analysed using two potent model interpretation techniques: SHAP and LIME.
- Features like BOD and DO contribute to microbial contamination, according to SHAP and LIME research.
- Increased levels of coliform, a sign of organic pollution, have been connected to low DO and high BOD.
- Additionally important factors are conductivity and pH, which identify areas impacted by chemical runoff and industrial waste.
- The model can direct real-time monitoring and intervention tactics and aid in the identifi-cation of key variables.

## IV. Implemented Methodologies

Firstly, dataset is imported to be used for the project and sample is displayed along with Fea-ture Correlation Heatmap.

| | STATION CODE | LOCATIONS | STATE | TEMP | DO | pH | CONDUCTIVITY | BOD | NITRATE_N_NITRITE_N | FECAL_COLIFORM | TOTAL_COLIFORM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1312 | GODAVARI AT JAYAKWADI DAM, AURNAGABAD,MAHARASHTRA | MAHARASHTRA | 29.2 | 6.4 | 8.1 | 735.0 | 3.4 | 2.00 | 3.0 | 73.0 |
| 1 | 2177 | GODAVARI RIVER NEAR SOMESHWAR TEMPLE. | MAHARASHTRA | 24.5 | 6.0 | 8.0 | 270.0 | 3.1 | 2.00 | 72.0 | 182.0 |
| 2 | 2182 | GODAVARI RIVER AT SAIKHEDA. | MAHARASHTRA | 25.8 | 5.5 | 7.8 | 355.0 | 4.2 | 9.00 | 59.0 | 133.0 |
| 3 | 2179 | GODAVARI RIVER AT HANUMAN GHAT, NASHIK CITY. | MAHARASHTRA | 24.8 | 5.5 | 7.8 | 371.0 | 5.6 | 3.55 | 90.0 | 283.0 |
| 4 | 2183 | GODAVARI RIVER AT NANDUR- MADMESHWAR DAM. | MAHARASHTRA | 25.7 | 5.7 | 7.9 | 294.0 | 3.2 | 2.69 | 45.0 | 132.0 |

Figure 1. Shows us a sample of data present in the dataset showing various features.

The table displays data on water quality along Maharashtra's Godavari River, including tem-perature, conductivity, pH level, dissolved oxygen, nitrate nitrogen, nitrite nitrogen, faecal col-iform, and total coliform. The water at Jayakwadi Dam has the highest temperature, ranging from 24.5°C to 29.2°C. Moderate oxygen levels are indicated by dissolved oxygen values ranging from 5.5 mg/L to 6.4 mg/L. Organic pollution levels are indicated by BOD readings ranging from 2.0 to 5.6 mg/L. There is significant variation in nitrate and nitrite amounts, with a peak nitrite value of 9.00 mg/L at Saikheda indicating potential wastewater contamination or agricultural runoff. Total coliform and fecal coliform counts indicate different levels of micro-bial contamination, with the highest total coliform concentration in Nashik City near Hanuman Ghat.

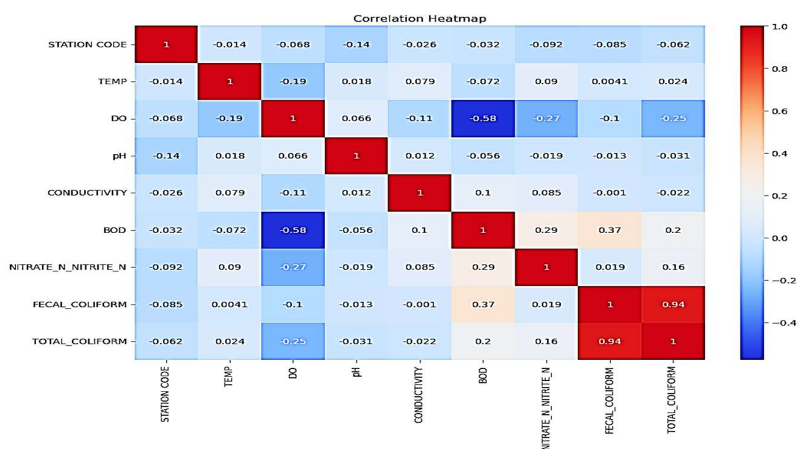


Figure 2: Feature Correlation Heatmap

The correlation heatmap provides illuminates the connections between different physicochem-ical characteristics and markers of microbial contamination in surface water. A pair of varia-bles' Pearson correlation coefficient, which ranges from -1 (strong -ve correlation) to +1 (strong +ve correlation), is represented by each cell in the matrix.

| | Model | R2_Score | RMSE |
|---|---|---|---|
| 0 | Linear Regression | 0.372720 | 6.918645 |
| 1 | SVR | 0.100965 | 8.282825 |
| 2 | Decision Tree | 0.770773 | 4.182375 |
| 3 | Random Forest | 0.696172 | 4.815083 |
| 4 | XGBoost | 0.832302 | 3.577290 |

Figure 3. R^2 Score and RMSE outcomes

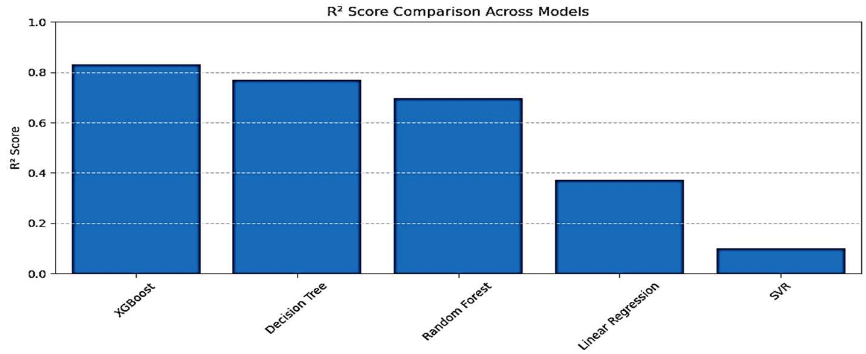


Figure 4. R^2 Score Graph

A comparison of five machine learning models for predicting microbial contamination based on physicochemical factors was conducted. XGBoost, with the highest R2 score (~0.83), explained over 83% of variation in microbial contamination. Decision Tree and Random Forest showed high but marginally inferior predictive power, with Random Forest often generalizing properly. Linear regression trailed with a lower R2 score (~0.37), indicating its inability to capture nonlinear relationships between the target variable and features. Support Vector Regression (SVR) was the worst due to its sensitivity to hyperparameters and difficulty with high-dimensional or nonlinearly separable data without correct kernel adjustment.

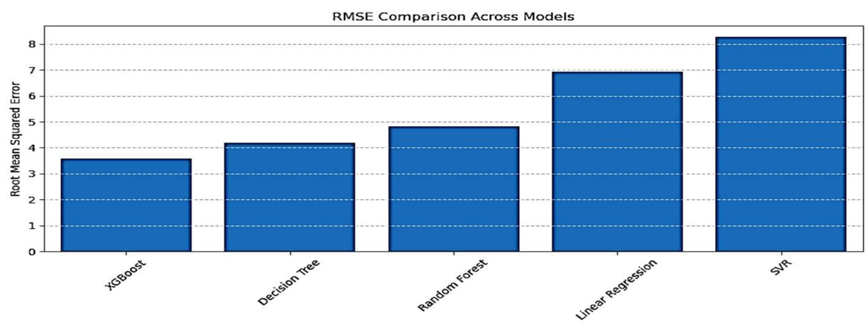


Figure 5. RMSE Graph

The Root Mean Squared Error (RMSE) graphic illustrates how well various models predict outcomes:

- On average, XGBoost produces the best accurate predictions because it has the lowest RMSE (~3.6).
- Random Forest and Decision Tree come next, with RMSE values somewhat rising.Poor model performance is shown by the much higher RMSE values of SVR and linear regression, particularly SVR (>8). This confirms that XGBoost is the best at minimizing prediction error and has good generalization.

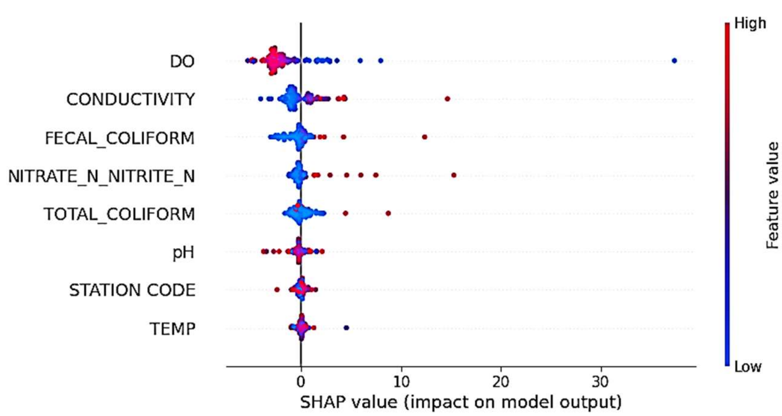


Figure 6. SHAP Value Graph

Features are ranked in this SHAP plot according to the quantity that they contribute to the model's out-put:

- ➢ The factors with the most effects involve DO, conductivity, faecal coliform, and nitrate/nitrite.
- ➢ Higher contamination projections have been linked to lower DO values (blue), which validates established environmental correlations.
- ➢ This interpretability increases confidence in the model's predictions and aids in its validation.

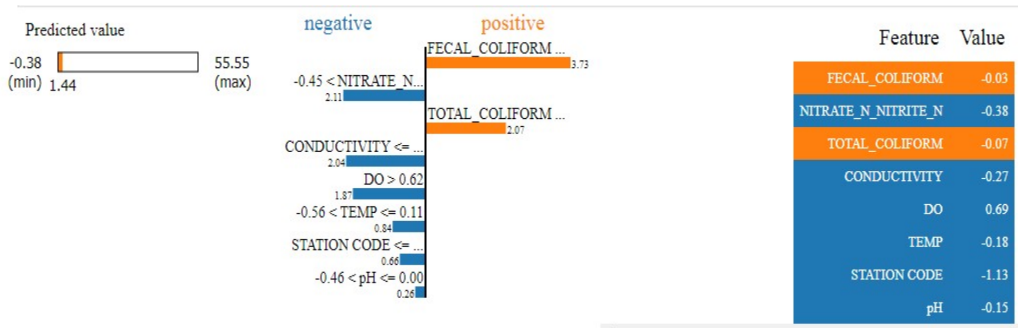


Figure 7. LIME implementation

Mean feature weights are visualized via LIME (Local Interpretable Model-Agnostic Explana-tions):
Once more, the most predictive factors are DO, nitrate/nitrite, and fecal coliform.DO < -0.49 and high nitrate levels have a considerable beneficial impact on predicting contamination, ac-cording to the direction and range of weights.

## V. Conclusion

The study explores the relationship between microbiological contamination in surface water in 18 Indian states and physicochemical characteristics. It uses machine learning algorithms and environmental data analysis to understand water quality dynamics. The XGBoost model was found to be the most accurate and reliable, capturing complex interactions between microbial markers and factors like DO, BOD, conductivity, and nitrate/nitrite concentrations. The study also found that monitoring capacity varied by location, with states like Delhi and Kerala hav-ing less accessible data. This highlights the need for community involvement, uniform national water quality standards, and real-time monitoring infrastructure. The study integrates envi-ronmental science and machine learning to provide a foundation for forecasting and under-standing water contamination, encouraging the development of proactive water management technologies and strategies. The results can guide initiatives to improve water quality, protect ecosystems, and public health in India by identifying high-risk areas and important pollution indicators.

## Future Work

- ➢ To provide a more comprehensive national picture and enhance model generalizability, data from other states and smaller bodies of water, such as lakes, reservoirs, and rural streams, can be added.
- ➢ Enhancing response to pollution events can be achieved by integration with IoT-based sensors and remote sensing technologies, which can facilitate real-time data collecting and early-warning systems for contamination.
- ➢ Newer contaminants that are having an increasing impact on aquatic ecosystems and hu-man health, such as microplastics, heavy metals, and pharmaceutical residues, may also be included in future research.

## References


[1] S. K. et al., “A Survey of Machine Learning Models for Water Quality Prediction,” IEEE Access, 2023, [Online]. Available: https://ieeexplore.ieee.org/document/10192728

[2] P. M. et al., “Survival Study on Different Water Quality Prediction Methods Using Machine Learning,” NEPT Journal, 2022, [Online]. Available: https://neptjournal.com/upload-images/(32)B-3875.pdf

[3] T. Nguyen and L. Wang, “Survey on Water Quality Prediction,” IEEE Sens J, 2023, [Online]. Available: https://ieeexplore.ieee.org/document/10113069

[4] Y. C. et al., “Ensemble Learning-Based Approach for Water Quality Prediction,” in SPIE Proceedings, 2025. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13486/3055821/

[5] L. Z. et al., “Prediction of the Concentration of Dissolved Oxygen in Running Water by Employing A Ran-dom Forest Machine Learning Technique,” Preprints (Basel), 2020, [Online]. Available: https://www.preprints.org/manuscript/202004.0342/v1
[6] I. K. Mutai, K. Van Laerhoven, N. W. Karuri, and R. K. Tewo, “Using Multivariate Linear Regression for Biochemical Oxygen Demand Prediction in Waste Water,” arXiv preprint arXiv:2209.14297, 2022.
[7] S. Uddin and others, “Artificial intelligence in water quality monitoring: A review of water quality assess-ment applications,” Water Quality Research Journal, vol. 60, no. 1, pp. 164–182, 2022.
[8] B. S. Pattnaik, A. S. Pattanayak, S. K. Udgata, and A. K. Panda, “Machine learning based soft sensor model for BOD estimation using intelligence at edge,” Complex and Intelligent Systems, vol. 7, no. 2, pp. 961–976, Apr. 2021, doi: 10.1007/s40747-020-00259-9.
[9] O. Khode, S. Koli, K. Bhat, and P. Korade, “Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability,” 2024.
[10] M. Arunika, S. Saranya, S. Charulekha, S. Kabilarajan, and G. Kesavan, “A Survey on Explainable AI Using Machine Learning Algorithms Shap and Lime,” in 2024 15th International Conference on Computing Com-munication and Networking Technologies, ICCCNT 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICCCNT61001.2024.10725120.
[11] S. J. Buchade and S. Bhoite, “Comparative Study of ML Algorithms for Garbage Classification,” Jan. 30, 2024. doi: 10.21203/rs.3.rs-3903806/v1.