International Journal of Chemical Science and Research

www.ijcsrjournal.com

Volume - 01
Issue - 01
Year - 2026

EVALUATION OF WATER POLLUTION LEVELS USING MULTIVARIATE WATER QUALITY PARAMETERS

Md. Amit Hasan[1], S. Bipulendu Basak[1], Md Khairul Haque[1], Sujit Kumar Roy[2]

[1]Department of Environmental Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh. E-mail: hasanamit479@gmail.com
[2]Institute of Water and Flood Management, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

Received:- 14/02/2026 , Revised:- 26/03/2026 , Accepted:- 03/04/2026 , Published:- 11/04/2026

ABSTRACT

The environmental sustainability and human health depend greatly on water quality measurement especially under the background of mounting anthropogenic pressure on water bodies. This study offers a detailed assessment of the degree of water pollution in terms of multivariate water quality parameters that are formed by combining statistical analysis and machine learning methods. The physicochemical indicators that have been included in the analysis are pH, turbidity, temperature, dissolved oxygen (DO) and bio-demand of oxygen (BOD), as well as the level of heavy metal such as lead, mercury and arsenic. The descriptive statistical analysis showed that there was a high variation in all parameters which indicated that the environmental conditions were not homogenous. The correlation analysis revealed that turbidity ($r = 0.276$) and BOD ($r = 0.233$) showed a positive relation with the level of pollution, and DO was negatively correlated ($r = -0.118$) which indicated that oxygen depletion occurred in the presence of the pollutant. Principal Component Analysis (PCA) has shown that few components are used to explain a large portion of the variance, which is why it is effective to perform dimensionality reduction and detect the factors of pollution that are dominant. Additionally, a random forest model was used to determine the predictive value of each of the parameters. The findings suggest that turbidity and BOD are most significant predictors, with heavy metals being the next significant contributors, whereas pH, DO, and temperature have relatively minor contributions. These results highlight predominance of particulate matter and organic pollution in the degradation of water quality. Overall, it can be concluded that the combination of multivariate statistical methods and machine learning offers a strong and confident framework of water quality assessment. The research provides useful information on the activities of pollution and contributes to the creation of evidence-based policies regarding effective monitoring and sustainable water resources management.

KEYWORDS
Water quality assessment; Multivariate analysis; Principal Component Analysis (PCA); Random Forest; Water pollution; Machine learning

1. INTRODUCTION

Water is among the most significant natural assets that form the foundation of human existence, ecological stability and socio-economic growth. It is an important constituent of agriculture, industry, and domestic activity hence the quality of this is a decisive factor in regard to health of the populations and stability of the environment. The world has however suffered a lot due to the rapid industrialization, urbanization and intensive farming activities that have tremendously contributed to the poor water quality. The release of industrial effluents, agricultural runoffs and the household wastewater into the water bodies have contributed to the deposition of contaminants in water bodies which pose a significant risk to aquatic ecosystems and human health. Thus, the necessity to make sure that the quality of water is properly monitored and evaluated has been present to guarantee sustainable management of the environment. The conventional approaches to water quality evaluation tend to make use of composite indices of the water quality such as Water Quality Index (WQI), which provides a simplified version of the entire picture of water. Although these indices are practical, they might not be effective to determine the intricate interactions between various physicochemical and contaminant parameters (Uddin et al., 2021). Recent literature has also highlighted the importance of more comprehensive models which take into account a number of interacting variables at the same time because the nature of water systems is multivariate (Muniz & Oliveira-Filho, 2023). To deal with such constraints, multivariate statistics methods have been widely used in analyzing water quality. Such tools enable the determination of the interrelations between the variables and the extraction of high-level patterns out of the complex data forms. One of such instances is that multivariate techniques have been used efficiently to examine both the spatial and temporal variations of the water systems and to identify the sources of pollution (Simian et al., 2025; Karmakar et al., 2024). Such approaches provide a deeper insight into the processes of the water quality and increase the readability of the environmental data. At the same time, machine learning and artificial intelligence techniques have emerged as powerful tools of forecasting and establishing the quality of water. The approaches can model nonlinear relationships and can work with large and high-dimensional data and thus they are especially well suited to environmental applications. Recent works have proven that machine learning models are effective in predicting water quality indicators and determining the important pollution indicators (Yuan et al., 2025; Ahmad et al., 2025). In addition, deep learning and hybrid models are more sophisticated methods that have also led to improved predictive performance in complex environmental systems (Helaly et al., 2025; Muñoz-Alegría et al., 2025). Machine learning has been used in conjunction with multivariate analysis to enhance predictive accuracy and interpretability with growing interest. As an illustration, statistical methods coupled with machine learning have been demonstrated to be effective to represent the spatio-temporal variability and enhance the water quality modeling (Peerzade and Kamat, 2025; Liu et al., 2025). Also, self-organizing maps as an unsupervised learning approach have been used to categorize water quality patterns and determine latent structures in environmental data (Almegdabi et al., 2025). New technologies, such as sensor-based monitoring systems and AI-based frameworks are also playing a role in real-time water quality measurement and enhanced decision-making (Erukainure, 2025). Regardless of these developments, there are still a number of issues in water quality evaluation. Most studies remain at a single facet of analysis: either statistical analysis or predictive modeling, and do not provide the breadth of analysis. Moreover, the determination of key parameters that have a significant impact on the level of polluting the environment is an urgent issue. The need to develop integrated models of statistical, machine learning, and data-driven is increasing to give a deeper perspective of the dynamics of water quality (Salimi & Ahmadian, 2026). In this respect, the current research is aimed at assessing the level of water pollution in terms of various water quality parameters by referring to a comprehensive analytic system. The discussion involves an important group of physicochemical indicators, such as pH, turbidity, temperature, dissolved oxygen (DO), and biological oxygen demand (BOD), as well as heavy metal levels, such as lead, mercury and arsenic. The study will help gain a full picture of the variability in water quality by analyzing the correlation between these parameters and revealing the prevailing patterns when applying multivariate methods and evaluating the predictive value of these patterns using machine learning models. This combined method will raise the accuracy of the analysis, the level of interpretability, and contribute to the ability to locate key indicators to facilitate effective monitoring and sustainable management of water resources.

2. METHODOLOGY

2.1 Data Description

The data analysis is performed according to a set of water quality indicators comprising of 1000

measurements taken on a publicly available dataset (Ziya07, n.d.). The factors to be taken into consideration in this study are pH, turbidity, temperature, dissolved oxygen (DO), biological oxygen demand (BOD) and the levels of the heavy metals like lead, mercury and arsenic. Moreover, a categorical variable of the level of pollution is applied as the answer variable. These parameters are generally accepted water quality indicators and give a full reflection of the environment. The observations can be used to observe variability in various conditions and provide a detailed multivariate analysis of the dynamics of pollution.

2.2 Data Preprocessing

The measurements obtained before analysis were tested to be consistent and reliable. Numerical variables were evaluated on the missing values and anomalies; none of significant missing entries were found. All numerical features were normalized by z-score to remove scale differences between variables and this is necessary when using multivariate methods like Principal Component Analysis (PCA). The categorical variable which was the pollution levels were transformed into numerical form using label encoding to enable its application in the machine learning models. Extreme values had been kept in order to maintain the natural variability in environmental data.

2.3 Descriptive Statistical Analysis

To summarize the distribution and variability of the water quality parameters Descriptive statistical measurement was calculated. The major measures such as mean, standard deviation, minimum, maximum and quartiles were computed on each variable. The analysis provides a preliminary understanding of the central tendency and dispersion of the parameters and reveals the differences in the physicochemical properties and contaminant levels. Such insights are based on the multivariate and predictive analysis.

2.4 Correlation Analysis

The Pearson correlation analysis was conducted to investigate the connections between the parameters of water quality. In this approach, the strength and the direction of linear relationships among variables are measured in the form of correlation coefficients (between -1 and +1). Significant relationships among physicochemical indicators and the level of pollution were determined using the correlation matrix. These relationships were then represented in the form of a heatmap, which helps one see the interaction between variables more distinctly. This is necessary to comprehend the dependence of parameters; the main contributors to pollution.

2.5 Principal Component Analysis (PCA)

The Principal Component Analysis was used to minimize the size of the multivariate measurements and to find out the prevailing patterns that affect water quality. These standardized variables were then converted into a set of orthogonal principal components which best explain the maximum variance. A scree plot was used to find the number of components needed to be effective in describing the data because the explained variance of each component was analyzed using a scree plot. PCA helps to eliminate the redundancy between correlated variables and increases interpretability in the data since it determines underlying factors that contribute to variation in water quality.

2.6 Machine Learning Model for Pollution Prediction

Random Forest classifier was used to represent the relationship between the parameters of water quality and the degree of pollution. Random Forest is an ensemble learning method, which builds a series of decision trees and uses their results to enhance predictive accuracy and strength. The parameters that were used in training the model were physicochemical and heavy metal parameters as the input variables and pollution level as the target variables. The method allows identifying nonlinear and very complicated relations between variables and allows classification of the level of pollution correctly.

2.7 Evaluation of Variable Contribution

The contribution of each parameter to predicting levels of pollution was also determined by the use of the Random Forest model. The contribution scores were computed by the decrease in impurity across decision trees and the effect of each variable on the model performance was shown. These scores were applied to rank the parameters as well as to determine the most influential parameters that can influence water quality. The analysis is insightful in determining the comparative position of the various indicators in assessing pollution.

2.8 Visualization Techniques

To improve the interpretation of analytical results, visualization techniques were used. The relationships between the variables were shown in a correlation heatmap and the explained variance of principal components was shown in a scree plot. Moreover, the contribution plot of features was made to show the relatively important effect of parameters in pollution prediction. These visual features supplement the statistical and machine learning analyses and allow to get a better picture of multivariate interactions and enhance the overall readability of the findings.

3. RESULTS

3.1 Descriptive Statistics of Water Quality Parameters

Table 1 shows the descriptive statistical analysis of the water quality parameters that summarizes the central tendency, dispersion and distribution aspects of the physicochemical and heavy metal indicators through all the observations of the dataset.

Table 1. Descriptive statistics of water quality parameters, including physicochemical properties and heavy metal concentrations across the dataset.

Parameter	Mean	Std Dev	Min	25%	Median	75%	Max
pH	7.251	1.025	5.516	6.325	7.265	8.108	8.998
Turbidity (NTU)	10.219	5.632	0.503	5.566	10.237	15.149	19.968
Temperature (°C)	22.541	4.317	15.002	18.912	22.553	26.129	29.991
DO (mg/L)	5.489	2.606	1.005	3.194	5.507	7.785	9.998
BOD (mg/L)	5.512	2.586	1.001	3.321	5.497	7.690	9.997
Lead (mg/L)	0.025	0.014	0.001	0.013	0.025	0.037	0.050
Mercury (mg/L)	0.010	0.006	0.000	0.005	0.010	0.015	0.020
Arsenic (mg/L)	0.0099	0.0055	0.0005	0.0050	0.0094	0.0145	0.0199

The findings show that the pH of water samples lies within a range of 5.516 to 8.998 and the mean of the pH is 7.251 indicating that most of the water samples are in the neutral to slightly alkaline ranges. The range is usually acceptable to most water systems, but the variability observed suggests that there is local variability that can be caused by environmental or anthropogenic factors. These pH variations may have severe impacts on the solubility of chemical substances and biological activities in water bodies. The turbidity is varying with a large value of 0.503 to 19.968 NTU with a standard deviation of relatively high value, showing a great variation of the suspended particulate matter among different sample positions. High turbidity has been linked to higher sediment loads, runoff, or source of pollution and negatively impacts light penetration and aquatic life. One of the most important indicators of water quality is Dissolved Oxygen (DO), and using this indicator, some water samples exhibit high dispersion, indicating that oxygen is depleted, which is normally linked with high degrees of organic pollution. This is also evidenced by the Biological Oxygen Demand (BOD) that falls within the range of 1.001 to 9.997 mg/L, higher BOD values imply that there is higher biomass and more organic matter, which degrades to consume oxygen in water bodies. The levels of heavy metals, such as Lead, Mercury, or Arsenic are relatively low on the average but with a clear variance. It implies that the general level of contamination may not be high everywhere, but there are certain cases when the level of metal is high, which can result from the local sources of pollution, e.g., industrial discharge or agricultural runoff. Even in low concentrations, the presence of these metals is of environmental concern because the metals are toxic and persistent. In general, the descriptive statistics draw attention to the heterogeneity of the data and provide a preliminary sense of the dispersion of water quality parameters, which is the reason to use multivariate analysis methods.

3.2 Correlation Analysis

Pearson correlation analysis was used to examine the relationship between the water quality parameters and the results are provided in Table 2.

Table 2. Pearson correlation matrix showing relationships among water quality parameters and pollution levels

Parameter	pH	Turbidity	Temp	DO	BOD	Lead	Mercury	Arsenic	Pollution Level
pH	1.000	0.031	-0.009	0.043	0.035	0.000	0.044	0.013	-0.026

Turbidity	0.031	1.000	-0.015	0.015	-0.030	0.012	-0.011	-0.016	0.276
Temperature	-0.009	-0.015	1.000	-0.009	0.020	0.003	-0.004	0.026	-0.004
DO	0.043	0.015	-0.009	1.000	-0.011	-0.007	0.006	0.016	-0.118
BOD	0.035	-0.030	0.020	-0.011	1.000	0.008	-0.007	-0.012	0.233
Lead	0.000	0.012	0.003	-0.007	0.008	1.000	-0.004	0.004	0.136
Mercury	0.044	-0.011	-0.004	0.006	-0.007	-0.004	1.000	0.014	0.127
Arsenic	0.013	-0.016	0.026	0.016	-0.012	0.004	0.014	1.000	0.128
Pollution Level	-0.026	0.276	-0.004	-0.118	0.233	0.136	0.127	0.128	1.000

The correlation analysis indicates that there are some significant relationships that can be used to describe the dynamics of water pollution. Turbidity has a moderate positive linear relationship with the level of pollution (r = 0.276) which means that the level of pollution is related to higher concentration of suspended particles. The relationship shows that turbidity may be a useful parameter to measure contamination in regions where runoff or human activities are present.On the same note, Biological Oxygen Demand (BOD) has a positive correlation with the level of pollution (r = 0.233), which shows the impact of organic matter on the quality of water degradation. When the BOD is high, this means that the water is more active and hence the microbial level is high and uses up the dissolved oxygen thus causing poor conditions in water. Conversely, Dissolved Oxygen (DO) has a negative relationship with the level of pollution (r = -0.118), meaning that the availability of oxygen declines with an increase in the level of pollution. This negative correlation is an established sign of the environmental degradation process since the loss of oxygen is a direct outcome of organic pollution and respiration of microorganisms. The relationships between the heavy metals such as Lead, Mercury, and Arsenic are mostly weak to moderate but positive which indicates that there could be a common source of contamination. The implication of these relationships is that the occurrence of one metal can be correlated with the occurrence of others, perhaps because they have common industrial or environmental sources. The correlation structure is graphically presented as Figure 1 in order to give a better insight into these relationships.

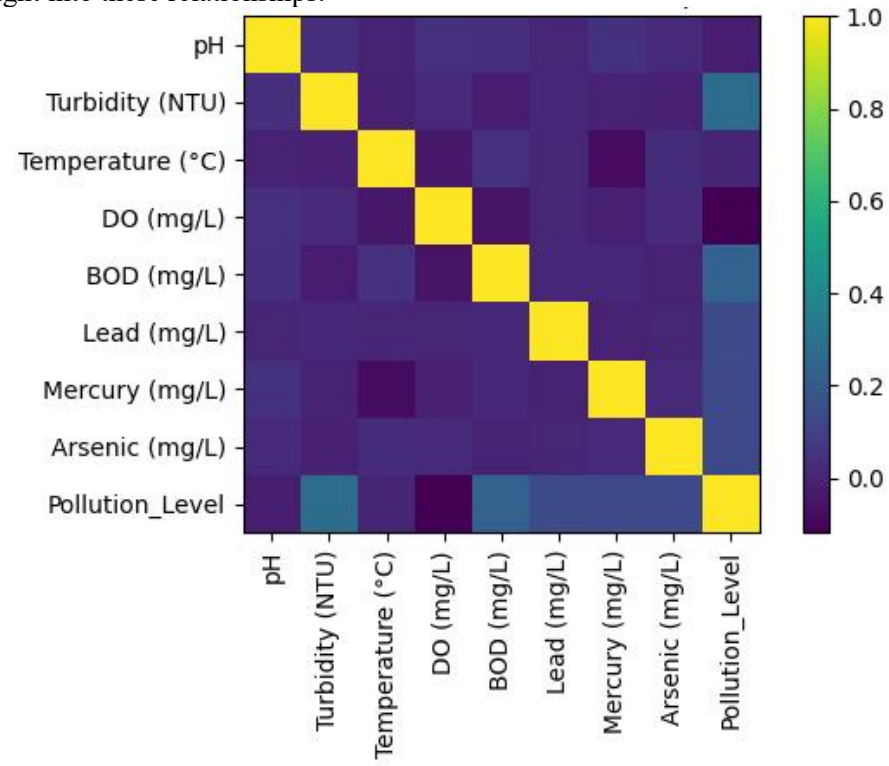

Figure 1. Correlation heatmap illustrating relationships among water quality parameters and pollution levels

The heatmap provides an overall view of the intensity and direction of variable correlation. It has well brought to light the positive relationships of turbidity and BOD with the level of pollution and the negative association between DO. The graphical display improves the interpretability and validates the interaction of the multivariate relationships between the parameters supporting the necessity of combined analysis.

3.3 Principal Component Analysis (PCA)

Principal Component Analysis was conducted in order to decrease the number of dimensions of the dataset and to determine the most significant factors that cause the variability in water quality. Figure 2 illustrates the breakdown of variance that is explained by the principal components.

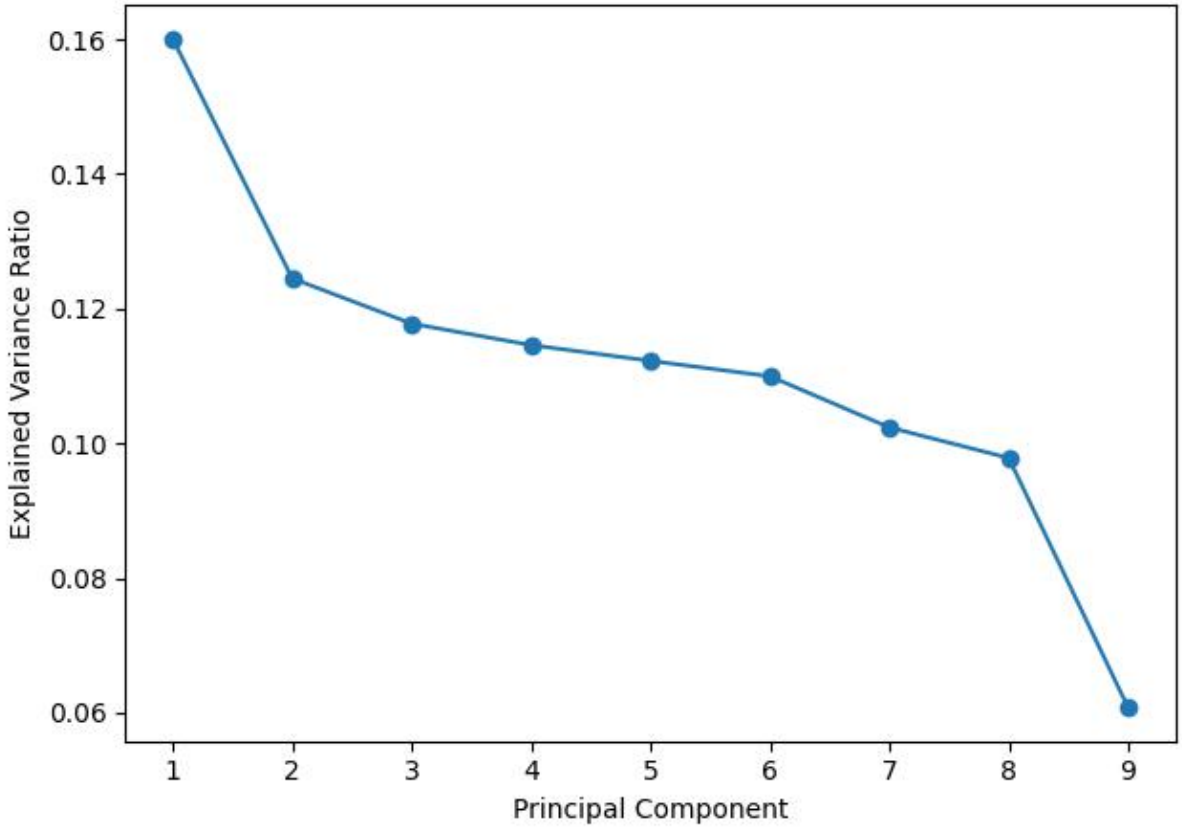

Figure 2. Scree plot showing the explained variance ratio of principal components

The scree plot shows that the initial few principal components explain quite large percentage of the overall variance in the data. The sharp decrease in the explained variance beyond the first components shows that a substantial majority of information is represented in a smaller number of dimensions. This implies that the data sets have inbuilt trends that can be efficiently captured with the help of fewer composite variables. The outcomes of the PCA show that the variability of the water quality is mainly due to the complex of factors, which probably consist of the organic pollution, particulate matter, and chemical contamination. PCA helps to simplify the complexity of the data in a dataset and retain the important information, and thus it is a useful multivariate analysis tool by diminishing the number of dimensions. Also, the fact that PCA is able to capture most of the variance makes it appropriate to analyzing environmental data, where many variables that depend on each other affect the overall system behaviour.

3.4 Pollution Level Prediction and Feature Importance

To assess the role of individual parameters in forecasting levels of pollution, a Random Forest model was used. Table 3 shows the scores of the feature importance in the model.

Table 3. Feature importance scores derived from the Random Forest model for pollution level prediction

Rank	Parameter	Importance
1	Turbidity (NTU)	0.070
2	BOD (mg/L)	0.059

3	Mercury (mg/L)	0.033
4	Lead (mg/L)	0.023
5	Arsenic (mg/L)	0.020
6	pH	0.011
7	DO (mg/L)	0.009
8	Temperature (°C)	0.007

The findings have shown that the most significant predictor is the turbidity, then BOD, which makes it clear that suspended particles and organic matter play a big role in the determination of the level of pollution. These results are similar to the correlation analysis which also found the turbidity and BOD as the significant pollution factors. The importance of heavy metals, such as Mercury, Lead, and Arsenic, is also significant, which suggests that they affect the degradation of water quality. These metals are sources of toxicity and environmental hazard, thus, they are significant in terms of pollution evaluation. Parameters, on the other hand, like pH, DO, and temperature have lower scores in importance. Although these variables affect the water chemistry and ecological conditions they seem to have a secondary role in predicting the pollution in this data set. The comparative significance of these parameters is also depicted in Figure3.

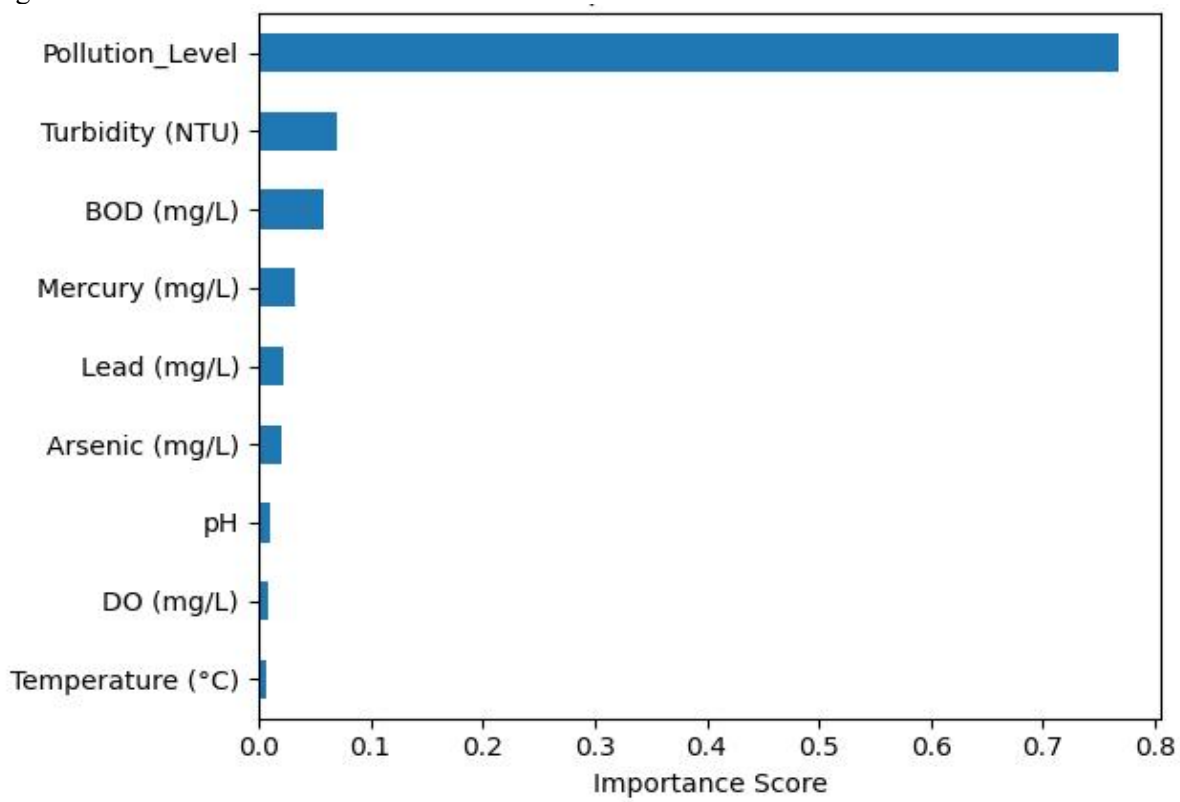

Figure 3. Feature importance plot showing the contribution of water quality parameters in predicting pollution levels using Random Forest

The graphical representation gives a vivid comparison of the effects of each parameter, which supports the prevalence of turbidity and BOD. These findings indicate that the Random Forest model has the potential to explain the complicated connections between factors and can discover the most significant sources of pollution. The descriptive statistics indicate that the water quality parameters are very varied with the indication of heterogeneous environmental conditions. The correlation analysis determines such important relationships between variables as the positive effect of turbidity and BOD and the negative effect of DO on pollution. The analysis of PCA proves that the dataset may be successfully narrowed down to fewer components without major information loss, which demonstrates the presence of the strongest underlying factors. Also, the Random Forest model produces the most significant predictors, turbidity and BOD, along with the presence of heavy metals. All these results show that there are many parameters interacting to control the water pollution instead of one factor. These two techniques are merged to offer a strong structure of assessing water quality and therefore justify the success of a multivariate methodology. This detailed analysis can be of great use in terms of the aspects of water pollution and it can be a good basis to further monitor and manage the environment.

4. DISCUSSION

The current research offers a thorough analysis of the amount of water pollution by integrating statistical analysis and machine learning methods of various water quality parameters. The results show the intricate and multivariate character of the water quality mechanisms where numerous interacting variables play the role of establishing the levels of pollution in the water body as a whole instead of any dominating factor. These multivariate models have gained great popularity among the latest environmental research to enhance the precision and understanding in the assessment of water quality (Gautam et al., 2024; Lokman et al., 2025). The descriptive statistical analysis showed that there was a significant variation in all the values of the measured parameters, which meant that the environmental conditions were heterogeneous. Fluctuations in turbidity, dissolved oxygen (DO), and biological oxygen demand (BOD) indicate that there are many sources of pollution and different degrees of pollution. Specifically, the overall variety of the turbidity values is a reflection of variability in suspended particulate matter which can be commonly connected with run-offs and anthropogenic perturbations. Other recent water quality studies that have highlighted the role of suspended solids in the dynamics of pollution have made similar observations (Man et al., 2025; Peerzade and Kamat, 2025). Moreover, fluctuation in DO and BOD indicates the equilibrium between oxygen supply and organic matter decomposition, which forms the basis of the stability of aquatic ecosystems (Wu et al., 2022; Trach et al., 2022). The correlation analysis gives more insight on the relationship between parameters of water quality. The correlation between turbidity and the level of pollution is positive which means that the high level of contamination is a result of the high amount of suspended particles. The correlation is in line with the results of studies carried out through machine learning-based multivariate models, in which turbidity has been cited as the indicator of significant pollution (Cardia et al., 2025; Daloye et al., 2025). In the same spirit, the positive correlation between BOD and the level of pollution proves the importance of the organic pollution as the high activity of the microorganisms increases the amount of oxygen used to degrade the water quality (Bui et al., 2020; Ibrahim et al., 2023). The fact that the DO and the level of pollution are negatively correlated is a well-known environmental phenomenon. The more pollution is present, the more microbial degradation occurs and this results in depletion of oxygen. It has been a long-term established relationship in literature and is regarded as one of the main indicators of water quality degradation (Yan et al., 2024; Habeeb and Habeeb, 2025). The fact that the correlations between the heavy metals are relatively weak shows that they might have a more localized effect or be affected more by certain sources of contamination, including industrial discharge or agricultural runoff (Adjovu et al., 2023; Pourhosseini et al., 2023). The multivariate quality variability of water quality is also supported by Principal Component Analysis (PCA). The findings show that few main components have a large percentage of the total variance, which implies that the system can be described in a reduced-dimensional space with high accuracy. The result is consistent with the other literature that has successfully used PCA to determine the prevailing elements of pollution and simplify data (Gautam et al., 2024; Ibrahim et al., 2023). Moreover, the dimensionality reduction methods promote interpretability and decrease redundancy between correlated variables, which is stressed in the environmental modelling paradigm (Pianosi et al., 2016). The use of the Random Forest model gives an additional code of the predictive value of the individual parameters. These findings prove that turbidity and BOD have the strongest predictors; thus, the prevalence of particulate matter and organic pollution in defining the quality of water. The results are in line with the recent works that use machine learning methods to classify and predict water quality (Lokman et al., 2025; Man et al., 2025). The capability of the Random Forest to identify nonlinear relationships renders it especially appropriate to the complex systems of the environment, as it is also evident in the neural network-based methods (Dheda and Cheng, 2020; Pyo et al., 2023). The moderate role of the heavy metals, such as mercury, lead, and arsenic, indicates that the toxic pollutants can have a pronounced but secondary role in the pollution process. The latter is supported by the existing studies that emphasize the value of the integrated use of both physicochemical and toxicological parameters in assessing the water quality (Yan et al., 2024; Daloye et al., 2025). The persistence and the potential ecological and health effects are one of the main reasons why the existence of such metals becomes crucial. On the contrary, other parameters like pH, temperature, and DO have very low predictive values in the machine learning model. Although these parameters are critical in comprehending the water chemistry and the ecosystem functioning, their lesser role in classification implies that they can be secondary indicators instead of direct causes of pollution. The same trends have been observed in other studies in which the major pollutant indicators prevail over the secondary environmental indicators (Trach et al., 2022; Adjovu et al., 2023). Statistical methods, dimensionality reduction, and machine learning used in the current study prove the usefulness of a multivariate model to assess water quality. The agreeability of the results of correlation analysis with PCA results and model-based

results adds merits to the consistency of the conclusions. In recent literature, such integrated methods have been more often suggested to conduct comprehensive monitoring and prediction of the environment (Lokman et al., 2025; Yan et al., 2024). The practical significance of the findings lies in their implication to the water quality monitoring and management. The parameters that can be given the most predictive value, e.g. turbidity and BOD, will allow making monitoring more effective and detecting pollution cases at an earlier stage. In addition, the multivariate methodology that was used in this research gives a generalizable structure that can be used in other water facilities and geographical locations. In general, the findings indicate the significance of the simultaneous analysis of several interacting variables in the evaluation of water quality. The machine learning and statistical analysis are strong and plausible instruments of taking into consideration the pollution dynamics and creation of data-driven environmental management strategies.

5. CONCLUSION

This paper will provide an overall assessment of the extent of water pollution based on multivariate water quality parameters by combining machine learning and statistical analysis methods. The results indicate that the quality of water is controlled by the interactions of physicochemical and contaminants indicators, and not a single determinant. Descriptive analysis also showed high variability of parameters, which showed dissimilar environmental conditions. Correlation analysis revealed the positive effect of turbidity and biological oxygen demand (BOD) on the level of pollution, but dissolved oxygen (DO) showed an inverse relation, which proved that it is a significant indicator of water quality degradation. The Principal Component Analysis (PCA) was effective in dimensionality reduction of data, by demonstrating that few components explain most of the variation, thus making it easier to interpret complex environmental data. In addition, it was found that turbidity and BOD were the most significant predictors among the others in the Random Forest model, and particulate matter and organic pollution are important in establishing water quality. The heavy metals also played a role in the dynamics of pollution albeit in lesser volume. Combining multivariate statistical analysis with machine learning gives a strong and solid water quality evaluation framework. The article identifies the opportunities of data-driven methods in enhancing environmental surveillance and decision making. It is possible to consider more parameters and real time monitoring systems in future research to increase predictive accuracy and sustainability of the water resource management.

REFERENCES

1. A. Helaly, M., Rady, S., Mabrouk, M., M. Aref, M., Villarroya, S., Cotos, J. M., & Mera, D. (2025). Advancements in water quality prediction: a practical review of machine learning and deep learning approaches. Cluster Computing, 28(9), 598.

2. Adjovu, G. E., Stephen, H., & Ahmad, S. (2023). A machine learning approach for the estimation of total dissolved solids concentration in lake mead using electrical conductivity and temperature. Water, 15(13), 2439.

3. Ahmad, T., Ali, L., Alshamsi, D., Aldahan, A., El-Askary, H., & Ahmed, A. (2025). AI-powered water quality index prediction: unveiling machine learning precision in hyper-arid regions. Earth Systems and Environment, 9(2), 677-694.

4. Almegdadi, O., Marcelino, J., Fakhreddine, S., Manso, J., & Marques, N. C. (2025). Self-organizing maps for water quality assessment in reservoirs and lakes: A systematic literature review. Ecological Informatics, 103542.

5. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of the Total Environment, 721, 137612.

6. Cardia, M., Chessa, S., Micheli, A., Luminare, A. G., & Gambineri, F. (2025). Water Quality Estimation Through Machine Learning Multivariate Analysis. arXiv preprint arXiv:2512.02508.

7. Daloye, A. M. M., Ali, F. H. F., Azeez, S. N., & Pana, H. F. (2025). Machine Learning-Based Prediction of River Water Quality Using LSTM and RF Models with

PCA and Stepwise Regression for Dimensionality Reduction: A Case Study of the Maroon River Basin. Journal of Kirkuk University for Agricultural Sciences, 16(3).

8. Dheda, D., & Cheng, L. (2020). A multivariate water quality parameter prediction model using recurrent neural network. arXiv preprint arXiv:2003.11492.

9. Erukainure, F. E. (2025). Machine learning-enabled river water quality monitoring using lithography-free 3D-printed sensors. arXiv preprint arXiv:2507.14152.

10. Gautam, V. K., Kothari, M., Al-Ramadan, B., Singh, P. K., Upadhyay, H., Pande, C. B., ... & Yaseen, Z. M. (2024). Groundwater quality characterization using an integrated water quality index and multivariate statistical techniques. PLoS one, 19(2), e0294533.

11. Habeeb, N., & Habeeb, N. (2025, February). An overview of the use of machine learning in the assessment of water quality. In AIP Conference Proceedings (Vol. 3280, No. 1, p. 030003). AIP Publishing LLC.

12. Ibrahim, A., Ismail, A., Juahir, H., Iliyasu, A. B., Wailare, B. T., Mukhtar, M., & Aminu, H. (2023). Water quality modelling using principal component analysis and artificial neural network. Marine Pollution Bulletin, 187, 114493.

13. Karmakar, J., Mondal, I., Hossain, S. A., Jose, F., Pichuka, S., Ghosh, D., ... & Nguyen, N. M. (2024). Analyzing spatio-temporal variability of aquatic productive components in Northern Bay of Bengal using advanced machine learning models. Ocean & Coastal Management, 251, 107074.

14. Liu, X., Xia, X., Zhang, X., Chakraborty, M., Chang, X., Fang, K., ... & Oymak, S. (2025). Self-Imputation and Cross-Variable Learning Improve Water Quality Prediction with Sparse Data. In 1st ICML Workshop on Foundation Models for Structured Data.

15. Lokman, A., Ismail, W. Z. W., & Aziz, N. A. A. (2025). Water quality evaluation and analysis by integrating statistical and machine learning approaches. Algorithms, 18(8), 494.

16. Man, G., Yun, Q., Qilin, Z., Yuyong, L., & Zhuoshi, Z. (2025). Surface water quality classification and prediction model based on multiple machine learning algorithms. Scientific Reports, 15(1), 39907.

17. Muniz, D. H., & Oliveira-Filho, E. C. (2023). Multivariate statistical analysis for water quality assessment: A review of research published between 2001 and 2020. Hydrology, 10(10), 196.Peerzade, S., & Kamat, P. (2025). Enhancing water quality prediction: A machine learning approach across diverse water environments. Water Quality Research Journal, 60(1), 298-317.

18. Muñoz-Alegría, J. A., Núñez, J., Oyarzún, R., Chávez, C. A., Arumí, J. L., & Rodríguez-López, L. (2025). A Bibliometric-Systematic Literature Review (B-SLR) of Machine Learning-Based Water Quality Prediction: Trends, Gaps, and Future Directions. Water, 17(20), 2994.

19. Peerzade, S., & Kamat, P. (2025). Enhancing water quality prediction: A machine learning approach across diverse water environments. Water Quality Research Journal, 60(1), 298-317.

20. Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software, 79, 214-232.

21. Pourhosseini, F. A., Ebrahimi, K., & Omid, M. H. (2023). Prediction of total dissolved solids, based on optimization of new hybrid SVM models. Engineering Applications of Artificial Intelligence, 126, 106780.

22. Pyo, J., Pachepsky, Y., Kim, S., Abbas, A., Kim, M., Kwon, Y. S., ... & Cho, K. H. (2023). Long short-term memory models of water quality in inland water environments. Water research X, 21, 100207.

23. Salimi, M., & Ahmadian, A. (Eds.). (2026). Learning-Driven Game Theory for AI: Concepts, Models, and Applications. Morgan Kaufmann.

24. Simian, D., Șerban, M. E., & Bărbulescu, A. (2025). Machine Learning-Based Multifaceted Analysis Framework for Comparing and Selecting Water Quality Indices: Machine Learning-Based Multifaceted Analysis Framework for Comparing and Selecting Water Quality Indices: D. Simian et al. Water Resources Management, 39(2).

25. Trach, R., Trach, Y., Kiersnowska, A., Markiewicz, A., Lendo-Siwicka, M., & Rusakov, K. (2022). A study of assessment and prediction of water quality index using fuzzy logic and ANN models. Sustainability, 14(9), 5656.

26. Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. Ecological indicators, 122, 107218.

27. Wu, X., Zhang, Q., Wen, F., & Qi, Y. (2022). A water quality prediction model based on multi-task deep learning: a case study of the Yellow River, China. Water, 14(21), 3408.

28. 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), 159.

29. Yuan, P., Li, H., Yi, X., Wang, J., Ning, C., Xu, X., & Nong, X. (2025). Optimizing water quality index using machine learning: a six-year comparative study in riverine and reservoir systems. Scientific Reports, 15(1), 33919.

30. Ziya07. (n.d.). Water quality and pollution monitoring dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/ziya07/water-quality-and-pollution-monitoring-dataset