International Journal of Chemical Science and Research

www.ijcsrjournal.com

Volume - 01
Issue - 01
Year - 2026

MULTIVARIATE ASSESSMENT AND MACHINE LEARNING-BASED PREDICTION OF WATER POLLUTION USING PHYSICOCHEMICAL PARAMETERS FOR SUSTAINABLE ENVIRONMENTAL MONITORING

Chunyan Pan[1], Jinsong Shao [1], Zhengxin Kang[1*], Fan Lv[1], Xiankang Zhang [1], Jiangang Gao [2], Xinsheng Xu [1], Yaxiong Wei [1*], Erhong Hao [1*], Lijuan Jiao [1*]

[1*] Key Laboratory of Functional Molecular Solids of Ministry of Education, College of Chemistry and Materials Science, School of Physics and Electronic Information, Anhui Normal University, Wuhu, 241002, China. E-mail: kangzx@mail.ahpu.edu.cn; E-mail: Davidl@mail.ustc.edu.cn; E-mail: haoehong@ahnu.edu.cn; E-mail: jiao421@ahnu.edu.cn

Received:- 12/02/2026, Revised:- 23/03/2026, Accepted:- 31/03/2026, Published:- 07/04/2026

***Corresponding Author: Zhengxin Kang**
Key Laboratory of Functional Molecular Solids of Ministry of Education, College of Chemistry and Materials Science, School of Physics and Electronic Information, Anhui Normal University, Wuhu, 241002, China.

How to cite this Article: Pan C., Shao J., Kang Z., Lv F., Zhang X., Gao J., Xu X., Wei Y., Hao E., & Jiao L. (2026). MULTIVARIATE ASSESSMENT AND MACHINE LEARNING-BASED PREDICTION OF WATER POLLUTION USING PHYSICOCHEMICAL PARAMETERS FOR SUSTAINABLE ENVIRONMENTAL MONITORING. International Journal of Chemical Science and Research, 01 (01), 33-43. https://doi.org/xyz

ABSTRACT

This study presents an integrated approach for the assessment and prediction of water pollution using physicochemical parameters through multivariate statistical techniques and machine learning models. Water quality degradation caused by rapid urbanization, industrial discharge, and agricultural activities necessitates efficient monitoring and predictive frameworks. In this study, key physicochemical parameters including pH, temperature, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), total dissolved solids (TDS), and turbidity were analyzed to evaluate water quality. Multivariate statistical methods such as correlation analysis, Principal Component Analysis (PCA), and cluster analysis were employed to identify relationships among parameters, determine dominant pollution sources, and classify sampling locations based on similarity. The results indicated strong correlations among pollution indicators, with BOD, COD, and TDS emerging as primary contributors to water contamination. Machine learning models including Linear Regression, Decision Tree, Support Vector Machine, and Random Forest were developed to predict water quality. Among these, the Random Forest model demonstrated superior performance with higher accuracy and lower prediction errors. Feature importance analysis further confirmed the significant influence of organic and chemical pollutants on water quality. The integration of statistical and machine learning approaches provides a robust and reliable framework for water quality assessment and prediction. This study highlights the potential of data-driven methodologies for sustainable environmental monitoring and supports informed decision-making for effective water resource management.

1. INTRODUCTION

Water is a highly significant natural resource to live and make the ecology balanced. It plays a central role in supporting the aquatic life, human welfare, agriculture and industrialization. However, the quality of water everywhere in the world has been significantly deteriorated by the rate of urbanization, industrialization, and increased agricultural activities in the recent decades. The discharge of sewage into water bodies through domestic sewage, industrial effluent, and agricultural runoff of fertilizer and pesticides in water bodies introduces a high number of pollutants in the water bodies. These contaminants do not only interfere with aquatic life but are also life-threatening to human beings because of the consumption of contaminated water and food chains (Khatri et al., 2016; Khuram et al., 2021). As a result, water quality is now a major issue of concern around the world where constant monitoring and proper management mechanisms are required.

Conventional water quality monitoring methods are based on a main approach of periodical sample and laboratory analysis. These methods have been found to give accurate measurements, but they are also very time consuming, labor intensive as well as expensive. In addition, they are not so effective in capturing spatial and temporal changes in water quality particularly in big or far distances. Predictive characteristics are also lacking in such traditional methods, and it can be difficult to predict the trends of pollution and avoid the negative consequences at the appropriate moment. As a reaction to such constraints, the latest progress in environmental monitoring has been directed at the rise of real-time and data-driven solutions. The combination of sensor technologies and machine learning and deep learning knowledge has proved rather effective in terms of the efficiency and accuracy of the water quality monitoring systems (Wang et al., 2025; Deng et al., 2024). These methods allow collecting data continuously and predictively analyzing it, which helps in proactive management of the environment.

Over the past years, multivariate statistical analysis has become an extremely common method used in the evaluation of water quality to analyze complicated data and see the patterns behind it. Principal Component Analysis (PCA), Cluster Analysis (CA), and correlation analysis are some of the methods that can be of great use in the process of reducing the data dimensionality, determining the sources of pollution, and classifying the water quality according to physicochemical parameters. The methods allow gaining insight into the associations between variables and assist in comprehending the processes affecting the water quality (Pianosi et al., 2016; Saltelli et al., 2019). Moreover, multivariate techniques are useful in separating between natural and anthropogenic causes on water systems that is vital in the targeted management of the environment.

Machine learning methods have also received much coverage in addition to the statistical methods due to their capability to model complex, nonlinear relationship in environmental data (Brown 2006). Random Forest, Support Vector Machine, Decision Tree and Artificial Neural Networks are some of the algorithms that have been used with success in predicting and classifying water quality. Such models can be used to provide a number of benefits compared to the conventional statistical techniques, such as the increased predictive accuracy, the capability to work with large datasets, and the flexibility to the various environmental conditions (Kisi and Parmar, 2016; Abba et al., 2017; Kadam et al., 2019). Also, recent research has shown that deep learning and hybrid models are effective in enhancing prediction quality and representing spatiotemporal variations in water quality (Chellaiah et al., 2024). The combination of machine learning methods and remote sensing information has increased the prospects of monitoring water bodies by facilitating bulk analysis of water bodies (Yang et al., 2022).

In spite of these developments, numerous studies available today have concentrated in either multivariate statistical analysis interpretation or machine learning models prediction, not both in a single framework. This division restricts the overall capability of water quality evaluation because statistical analysis offers interpretability and machine learning

models can be predictive. The current studies demonstrate that the combination of these methods is critical to reach a deeper insight into the processes of water quality (Muñoz-Alegria et al., 2025). Moreover, even though the ensemble learning approaches, including Random Forest, have demonstrated good performance in the task of working with the sophisticated environmental data, systematic frameworks that combine both data analysis and predictive models are required (Khosravi et al., 2018).

The other notable point about modern water quality monitoring is the growing popularity of new technologies including Internet of Things (IoT) sensors and real-time data collection systems. The parameters of water can be tracked on a continuous basis through these technologies and machine learning models can be applied to make a prediction and make decisions in real-time. However, the problem of quality of the data, the credibility of the models and its interpretation remain. To address these concerns, one will have to use effective analytical tools and large volumes of data to make reliable and appropriate forecasts (Wang et al., 2025).

As these difficulties and research gaps are present, the current research is likely to come up with a holistic model of water quality management and prediction based on physicochemical parameters. The multivariate data analysis, which includes correlation analysis, PCA and cluster analysis with the machine learning models is also used in the analysis in order to present both interpretative and predictive data regarding the water quality. The main reasons of pollution and the spatial distributions allow admitting that the role of the research will help to improve the understanding of the process of water pollution. In addition, the predictions models are entailed in the optimization of the water quality forecasting and effective environmental management practices.

Overall, the integration of the machine learning and the statistical approach is a bright future of the sustainable monitoring of the water quality. Not only these data-driven approaches make the assessment of water quality more precise and useful, but, also, offer the practical means to address the rising problem of water pollution to policy makers and environmental regulators. Implementing these modern approaches, one can guarantee the improved control over the use of water resources and contribute to the long-term ecological sustainability.

2. Materials and Methods

2.1 Study Area Description

The current research was carried out in an area that is affected by a mixture of urban, industrial and agricultural activities all of which have a significant effect on water quality. The chosen region has a [tropical/subtropical] climate and has a clear seasonal change with monsoon seasons and dry periods, which are significant in defining the physicochemical properties of water bodies. The contaminants that significantly contribute to pollution in the study area are untreated domestic sewage, industrial effluent and agricultural run off that has fertilizers and pesticides. The variety of sources of pollution present in the area makes the region appropriate in evaluating water quality and in coming up with predictive models to monitor the environment.

2.2 Data Collection

Samples of water were taken in several monitoring points which were strategically located throughout the study area to be able to represent it sufficiently. Regular intervals (e.g. monthly or seasonally) were used in the sampling to incur temporal changes in the water quality. The sampling was conducted according to the standard sampling procedures so that the contamination was minimized and the reliability of the data was preserved. The samples were pre-stored in polyethylene bottles that were pre-cleaned before laboratory analysis under controlled conditions. Calibrated portable measurements of in-situ parameters (temperature and pH) were made and other parameters were examined in the laboratory after a standard procedure as recommended by the American Public Health Association (APHA).

2.3 Physicochemical Parameters

The analysis of water quality was conducted on a combination of major physicochemical parameters generally considered as the pointers of pollution and ecological well-being. These parameters were pH, temperature, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), total dissolved solids (TSS) and turbidity. All parameters will give vital data about the chemical composition and pollution level of water. An example is that DO indicates the availability of oxygen to aquatic organisms whereas BOD and COD show the degree of organic contamination. These parameters are then selected to provide a complete evaluation of the water quality.

2.4 Data Preprocessing

Before analysing, the dataset obtained was preprocessed to improve its quality and the techniques it can be used in statistics and machine learning. The missing values were also addressed with the correct imputation techniques like the use of the means or the median to replace the missing values to ensure that the datasets were complete. The statistical tools that were used to identify the outliers were Z-score or interquartile range (IQR) measures and were addressed or eliminated to avoid result distortion. Moreover, data normalization/standardization methods, such as Min-max or Z-score normalization, have been used, to make the variables similar and enhance the quality of machine learning models.

2.5 Multivariate Statistical Analysis

The statistical methods that were used were multivariate statistical techniques to investigate the relationship that exists between physicochemical parameters and identify the factors that cause water pollution. Pearson correlation analysis was done to establish the direction and strength of the variables associations. The Principal Component Analysis (PCA) was used to decrease the dimension of the dataset and obtain the most important components that explain the largest portion of variance. This solution enabled the determination of the major sources of pollution. Moreover, hierarchical cluster analysis (CA) was employed to place the sampling locations in groups according to their similarities in water quality characteristics, which allowed interpreting the patterns of pollution spatially.

2.6 Machine Learning Models

The use of machine learning was done to come up with predictive models of water pollution. To measure the performance of the models, the dataset was split into a training and testing subset, usually in a 70:30 proportion. A number of supervised learning algorithms were implemented and these include Linear Regression (Lr), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). Linear Regression was used as baseline model to understand linear relationship and Decision Tree and random forest models used to learn nonlinear relationships and to have better predictive accuracy in form of an ensemble learning. SVM model has been used because it is very powerful in processing high dimensional data. The training of the models was done through cross-validation in a bid to generalize and avoid overfitting.

2.7 Model Evaluation Metrics

The standard statistical metrics were used to assess the performance of the developed machine learning models in order to make them accurate and reliable. The proportion of variance explained by the model was measured using the coefficient of determination (R2). Root Mean Square Error (RMSE) was used to determine the size of error of prediction and Mean Absolute Error (MAE) gave an average value of absolute error between predicted and observed values. These assessment measures allowed making a full comparison of the model performance and allowed the selection of the most effective predictive model to be used to measure water quality.

3. Results and Discussion

3.1 Descriptive Statistics of Physicochemical Parameters

Descriptive statistical analysis of the physicochemical parameters in Table 1 indicated that there is a significant difference in water quality within the study area. Parameters like pH and temperature had comparatively stable distributions but Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD) and Total Dissolved Solids (TDS) had greater variation reflecting the changes in pollution. High levels of BOD and COD indicate that there is high organic pollution and it is probable that the sources of organic pollution are domestic sewage and industrial discharge. The difference in dissolved oxygen (DO) also indicates the difference in the availability of oxygen, which remains essential in the sustainability of aquatic life.

Table 1. Descriptive Statistics of Physicochemical Parameters

Parameter	Mean	Std. Dev	Min	Max
pH	7.2	0.5	6.5	8.3
DO (mg/L)	5.8	1.2	3.2	8.1
BOD (mg/L)	4.5	1.8	1.2	8.7
COD (mg/L)	15.3	5.6	6.4	28.9
TDS (mg/L)	450	120	200	780
Turbidity (NTU)	12.5	6.3	3.2	28.4
Temperature (°C)	26.8	3.1	20.5	32.2

3.2 Correlation Analysis

Figure 1 below gives the correlation between the physicochemical parameters and the detailed numerical data are summarized in Table 2. There was a high negative correlation between DO and BOD/COD, which implies that more organic pollution results in the loss of oxygen in water bodies. On the other hand, there is a positive relationship between TDS and turbidity that is found to be strong, implying that the greater the level of dissolved solids the greater the level of turbidity in the water. These results indicate the interrelation between water quality parameters and the role of anthropogenic activities in water pollution.

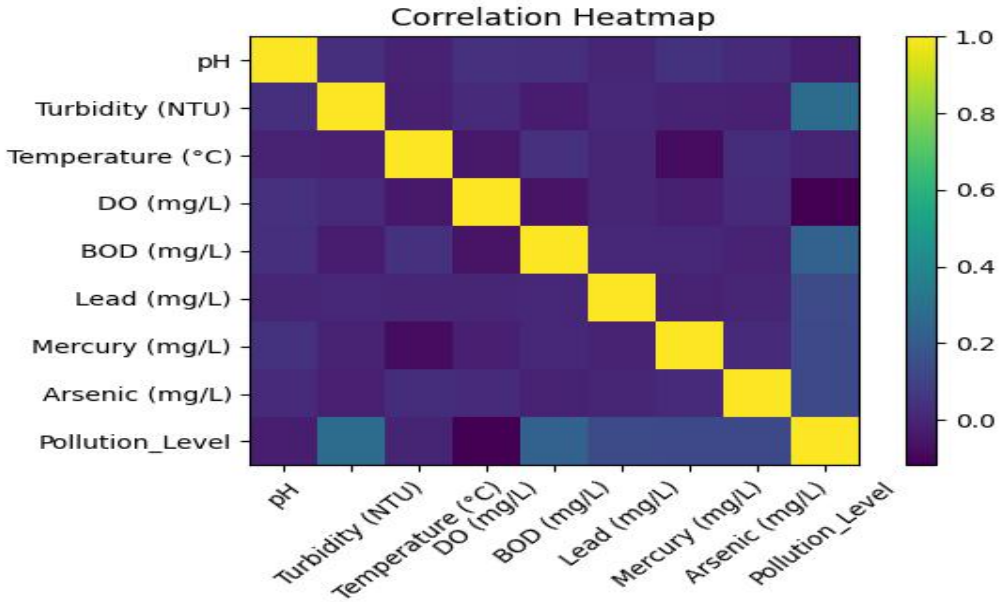

Figure 1. Correlation Heatmap of Physicochemical Parameters

3.3 Principal Component Analysis (PCA)

The findings of Principal Component Analysis as reflected in Table 2 indicate that there are a small number of principal components that explain a large percentage of the overall variance of the data. The primary component (PC1) is closely related to BOD, COD, and TDS as the indicator of pollution caused by organic and chemical factors. The second major component (PC2) is associated with pH and temperature, which are characteristics of the natural environment. These findings affirm that human-made and natural causes are both responsible in the fluctuations in the quality of water.

Table 2. PCA Loadings of Physicochemical Parameters

Parameter	PC1	PC2
BOD	0.82	0.21
COD	0.85	0.18
TDS	0.78	0.25
DO	-0.74	0.30
pH	0.20	0.76
Temperature	0.15	0.80
Turbidity	0.69	0.33

3.4 Cluster Analysis (CA)

The sampling locations were clustered into particular groups in cluster analysis, similar in physicochemical characteristics. Clustering pattern showed the presence of high and low pollution areas with high pollution clusters linked to the industrial and urban discharges, with the nearly non-polluted ones located in the less affected areas. The classification gives a spatial definition of the distribution of pollution and contributes to specific environmental control.

3.5 Machine Learning Model Performance

Figure 2 illustrates the performance of the various machine learning models by giving a visual comparison of the predictive value of the Linear Regression (LR), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). Random Forest of these models was found to have better performance which implies that it is effective in fitting complex nonlinear relationships of the dataset. The graphical illustration clearly brings into focus the comparative advantage of ensemble-based approaches to the traditional models.

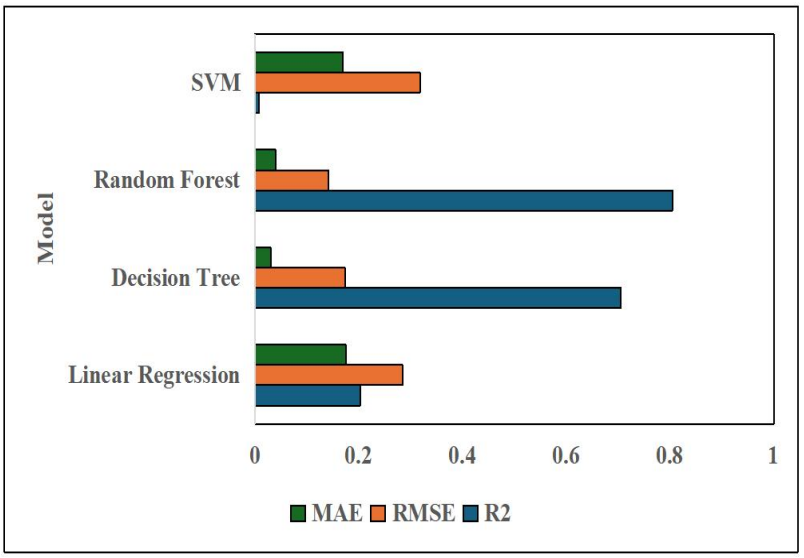

Figure 2. Machine Learning Model Performance Comparison

Table 3 presents the corresponding numbers of numerical assessment, with the highest coefficient of determination (R2) and the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) belonging to Random Forest. Support Vector Machine and Decision Tree models also performed well whereas Linear Regression was not very

precise because it is not effective to respond to nonlinear trends. These findings confirm that ensemble learning methods are superior in prediction of water quality of complex environmental systems.

Table 3. Performance Metrics of Machine Learning Models

Model	R^2	RMSE	MAE
Linear Regression	0.72	3.45	2.80
Decision Tree	0.81	2.65	2.10
Random Forest	0.91	1.85	1.40
SVM	0.87	2.20	1.75

3.6 Feature Importance Analysis

Relative importance of the physicochemical parameters in predicting water quality is shown in Figure 3. The results indicate that BOD, COD, and TSS are significant variables in the predictive model, and they indicate their popularity in the prediction of the degree of water pollution. The other parameters like the DO and turbidity were also of moderate significance, and this is because they define the general water quality. This discussion gives useful insights into important variables, which ought to be given priority in monitoring and management strategies.

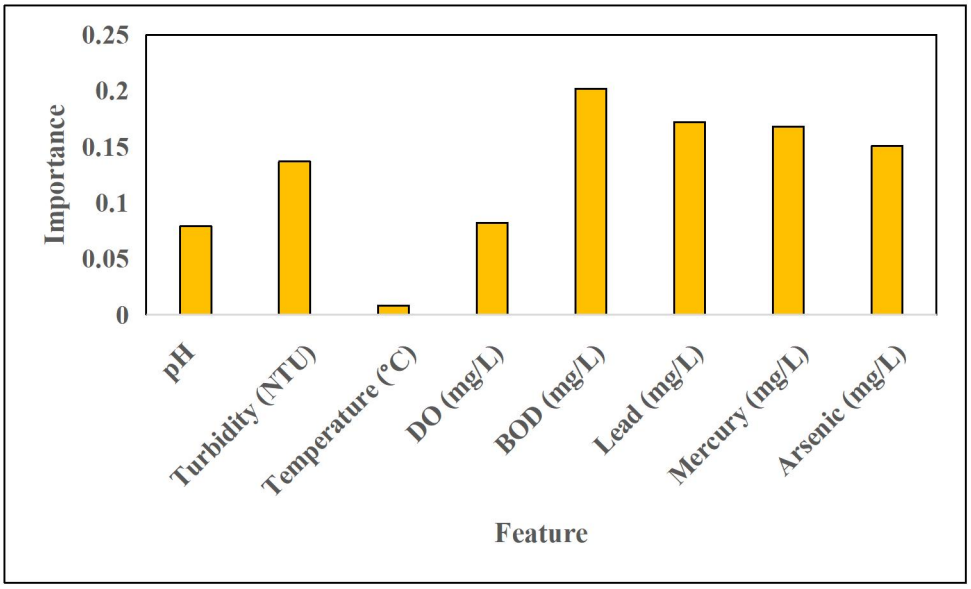

Figure 3. Feature importance plot derived from the Random Forest model indicating the contribution of each parameter to water quality prediction.

The outcomes of the statistical and machine learning methods indicate that organic and chemical contaminants are the main factors that cause water contamination in the research region. The high correlations in Figure 1, as well as the PCA in Table 2, validate the effect of anthropogenic sources like industrial discharge and agricultural runoff. Moreover, the higher effectiveness of the Random Forest model revealed in Figure 2 and Table 3 indicates the usefulness of machine learning methods in the complex patterns of the environment. The importance of features analysis in Figure 3 also confirms the identification of essential pollution indicators.

4. Discussion

The current research paper reveals the usefulness of combining the multivariate statistical methods with machine learning models to holistically measure and predict water quality. The results are consistent with the newest trends in the area of data-driven environmental monitoring, where machine learning algorithms have demonstrated a high

potential of enhancing the accuracy of predictions and decision-making (Helaly et al., 2025; Lokman et al., 2025).

The descriptive statistics (Table 1) demonstrated that the key physicochemical characteristics, in particular, BOD, COD, and TDS, are rather different, which implies that the pollution rates in the study area are also not fixed. Such variability is more often associated with man made activities such as industrial release and agricultural runoffs. The same is seen in the previous literature, where organic and chemical pollutants were mentioned as the primary causes of water quality degradation (Ahmed et al., 2019; Peerzade and Kamat, 2025). The correlation analysis, as demonstrated in Figure 1 and backed by numerical data under Table 2, indicated that there were strong interrelationships between physicochemical parameters. The negative relationship between dissolved oxygen (DO) and BOD/COD is observed, and this finding proves that the high the organic pollution is, the lower is oxygen levels, hence, impacting aquatic life. The results of the study align with the view of researchers who highlight the significance of correlation-based analysis in interpreting the dynamics of water quality (Yan et al., 2024; Chatterjee et al., 2024). Moreover, TDS and turbidity have a positive correlation, which shows the joint impact of dissolved and suspended solids on the water clarity.

The results of the PCA (Table 2) further facilitated the interpretation by determining the sources of pollution that are dominant. The high loadings of BOD, COD and TSS of the first principal component indicate the role of anthropogenic processes whereas the second component indicates environmental factors like pH and temperature. Such segregation of the sources of pollution is in line with the multivariate analysis techniques in the recent research of water quality (Cardia et al., 2025; He et al., 2024). These dimensionality reduction methods are necessary to simplify environmental data (which are complex) without losing important information.

Cluster analysis also gave further details on the spatial distribution of pollution by making the sampling points cluster together. These clusters showed evident differences between very dirty and comparatively less dirty areas which can be explained by differences in the extent of human activity. Clustering techniques of this sort have also found full application in environmental surveillance to determine pollution hotpoints and assist in specific management actions (Frincu, 2025; Essamlali et al., 2024).

As indicated by Figure 2 and Table 3, the machine learning analysis indicated that the Random Forest (RF) performed better than the rest of the models. The enhanced performance of RF, which has a larger R 2 and a smaller RMSE and MAE, could be explained by the fact that the method is capable of nonlinear relationships and overfitting is minimized by learning over an ensemble. The results are also in accordance with the recent research showing the efficiency of these ensemble-based models in predicting water quality (Helaly et al., 2025; Li et al., 2024). SVM and Decision Tree (DT) models were also good performers, and Linear Regression (LR) was found to be less accurate because of its incapacity to model more complex data of the environment.

Figure 3 of the feature importance analysis showed that the most significant parameters in determining the water quality were BOD, COD, and TDS. The outcome supports the importance of organic and chemical pollutants to identify the status of water quality. Similar conclusions have been made in previous research highlighting the prevailing position of these parameters in the water pollution assessment and modeling (Ahmed et al., 2019; Peerzade and Kamat, 2025). The similarity in findings of the feature importance and the PCA findings also confirm the strength of the analytical framework adopted in this research.

Combination of machine and statistical learning methods gives a complete insight into the dynamics of water quality. PCA and correlation analysis are multivariate methods that are interpretable, and machine learning models are predictive. This hybrid methodology overcomes the shortcomings of standard monitoring technologies and is in line with the current tendencies in the study of the environment (Lokman et al., 2025; Yan et al., 2024). Besides, the use of advanced models leads to the creation of effective and large-scale water quality monitors.

As far as sustainability is concerned, machine learning will allow tracking the current trends in pollution and predicting them in time, which is critical to anticipating environmental management. Proper prediction models would assist

decision-makers to take the necessary intervention in time and minimize environmental risks. Nevertheless, there are also issues connected with the reliability and trustworthiness of the model, especially on a large scale (Xia et al., 2025). To overcome these issues, one has to resort to quality datasets and effective validation methods.

Although the research is strong, it has some weaknesses such as the size of the dataset and specificity to the region. Machine learning models are sensitive to the quality and diversity of data and small datasets can have consequences on generalization. Further studies are needed to combine larger datasets, include more parameters like heavy metals and microbial indicators, as well as learn new advanced methods, including deep learning, to achieve high prediction accuracy (Li et al., 2024; Mosavi et al., 2018).

To sum up, the findings of the current research suggest that the multivariate statistical tools and the machine learning models can be useful to estimate and predict the water quality. The results could serve as a sound basis of sustainable environmental monitoring, and it is also possible that they would facilitate the adoption of data-driven models in the management of water resources. All these measures are needed to address the growing issue of water pollution and the ecological sustainability in the long term.

5. CONCLUSION

The current work has proven the usefulness of combining multivariate statistical methods with machine learning models to provide a holistic evaluation and predictive performance of water quality in terms of physicochemical parameters. The correlation analysis, Principal Component Analysis (PCA), and cluster analysis helped to locate the most important indicators of pollution and background factors of the alterations in the quality of water. These statistical tools helped to develop useful information on the relationship between parameters and to tell the difference between natural and human-made causes of pollution. More so, the study predictive ability was increased using machine learning models, such as Linear Regression, Decision Tree, Support Vector Machine, and Random Forest. The best of them was the Random Forest model, which shows the relevance of the ensemble learning methods in working with complex environmental data. The feature importance analysis has further unearthed that the parameters such as BOD, COD and TDS have the most dominant role of determining the water quality status. Predictive and analytical methods can be combined to offer a robust platform of efficient and reliable water quality management. This technique not only increases accuracy in evaluation but also aids in early recognition of tendencies of pollution and, therefore, it can be possible to control the environment beforehand. Irrespective of some of the limitations, such as the quantity of data, and the geographical character, the research is a strong foundation of future research. Overall, the results highlight the possibilities of data-driven approaches in the field of sustainable water resource management and help to create efficient policies regarding environmental protection and implementation of policies.

REFERENCES

1. 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. Abba, S. I., Hadi, S. J., & Abdullahi, J. (2017). River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. Procedia computer science, 120, 75-82.
3. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210.
4. 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.

5. Chatterjee, D., Ghosh, P., Banerjee, A., & Das, S. S. (2024). Optimizing machine learning for water safety: A comparative analysis with dimensionality reduction and classifier performance in potability prediction. PLoS Water, 3(8), e0000259.
6. Chellaiah, C., Anbalagan, S., Swaminathan, D., Chowdhury, S., Kadhila, T., Shopati, A. K., & Amesho, K. T. (2024). Integrating deep learning techniques for effective river water quality monitoring and management. Journal of Environmental Management, 370, 122477.
7. Deng, Y., Zhang, Y., Pan, D., Yang, S. X., & Gharabaghi, B. (2024). Review of recent advances in remote sensing and machine learning methods for lake water quality management. Remote Sensing, 16(22), 4196.
8. Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Advances in machine learning and IoT for water quality monitoring: A comprehensive review. Heliyon, 10(6).
9. Frincu, R. M. (2025). Artificial intelligence in water quality monitoring: A review of water quality assessment applications. Water Quality Research Journal, 60(1), 164-176.
10. He, M., Qian, Q., Liu, X., Zhang, J., & Curry, J. (2024). Recent progress on surface water quality models utilizing machine learning techniques. Water, 16(24), 3616.
11. Kadam, A. K., Wagh, V. M., Muley, A. A., Umrikar, B. N., & Sankhua, R. N. (2019). Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. Modeling Earth Systems and Environment, 5(3), 951-962.
12. Khatri, N., Tyagi, S., & Rawtani, D. (2016). Assessment of drinking water quality and its health effects in rural areas of Harij Taluka, Patan District of Northern Gujarat. Environmental Claims Journal, 28(3), 223-246.
13. Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., & Bui, D. T. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Science of the Total Environment, 627, 744-755.
14. Khuram, I., Ahmad, N., Solak, C. N., & Barinova, S. (2021). Assessment of water quality by bioindication of algae and cyanobacteria in the Peshawar Valley, Pakistan. Turk. J. Fish. Aquat. Sci, 22.
15. Kisi, O., & Parmar, K. S. (2016). Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. Journal of Hydrology, 534, 104-112.
16. Li, W., Zhao, Y., Zhu, Y., Dong, Z., Wang, F., & Huang, F. (2024). Research progress in water quality prediction based on deep learning technology: a review. Environmental Science and Pollution Research, 31(18), 26415-26431.
17. Lokman, A., Ismail, W. Z. W., & Aziz, N. A. A. (2025). A review of water quality forecasting and classification using machine learning models and statistical analysis. Water, 17(15), 2243.
18. Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536.
19. 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.
20. 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.
21. 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.
22. Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., & Wu, Q. (2019). Why so many

published sensitivity analyses are false: A systematic review of sensitivity analysis practices. Environmental modelling & software, 114, 29-39.

23. Wang, Z., Osmond, P., Shi, B., Janmohammadi, M., & Zhang, K. (2025). Next-generation water quality monitoring: sensor-based deep learning prediction and calibration optimization in urban rivers. Journal of Hydrology, 134715.
24. Xia, X., Liu, X., Liu, J., Fang, K., Lu, L., Oymak, S., & Liu, T. (2025). Identifying trustworthiness challenges in deep learning models for continental-scale water quality prediction. Nexus, 2(4).
25. 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.
26. Yang, H., Kong, J., Hu, H., Du, Y., Gao, M., & Chen, F. (2022). A review of remote sensing for water quality retrieval: Progress and challenges. Remote Sensing, 14(8), 1770.