Asian Journal of Water, Environment and Pollution, Vol. 21, No. 5 (2024), pp. 19-26.
DOI 10.3233/AJW240056

Water Quality Assessment using Machine Learning: A Focus on Coliform Prediction in Water

Ishleen Kaur, Archa Gulati[1], Puneet Singh Lamba*, Achin Jain[2], Harsh Taneja[3] and Jessica Singh Syal[2]

Sri Guru Tegh Bahadur Khalsa College, University of Delhi, New Delhi, India
[1]Ramjas College, University of Delhi, New Delhi, India
[2]Bharati Vidyapeeth's College of Engineering, New Delhi, India
[3]Department of Computer Science & Engineering, Graphic Era (Deemed to be University), Dehradun, India
✉ singhs.puneet@gmail.com

Received January 27, 2024; revised and accepted May 15, 2024

Abstract: Water quality assessment is essential for safeguarding public health and protecting water resources. This study focused on predicting water quality, specifically the presence of total coliforms, using various machine-learning techniques. The present study utilises a publicly available dataset encompassing the geographical area of India consisting of various physical water quality parameters. Various regression techniques were applied to the dataset after appropriate pre-processing including feature selection and normalisation. The findings demonstrate that gradient boosting regression outperforms other methods, achieving high accuracy with mean absolute error (MAE) of 0.0349, mean squared error (MSE) of 0.0038, and root mean squared error (RMSE) of 0.0620. Conductivity and temperature emerged as the most influential factors in total coliform prediction, as revealed by feature importance analysis. These results contribute to water quality understanding, aiding water resource management for public health protection. By accurately predicting total coliform presence, proactive measures can be taken timely to mitigate and minimise health risks associated with microbial contamination.

Key words: Water quality index, conductivity, temperature, machine learning, total coliform, regression.

Introduction

Water is the most essential resource to sustain life on earth. However, due to discharge of wastewater directly into the water bodies from industries, domestic households, agricultural fields, etc., has rendered the surface and ground water severely contaminated. Consequently, this polluted water poses a serious threat to human health and is unfit for human consumption and usage (Madhav et al., 2020). Therefore, water pollution is a matter of grave concern globally. Thus, immediate steps are required to assess the quality of the water source and develop strategies for water remediation. One such important indicator of water quality is total coliform content. The main source of coliform is human and animal waste and untreated industrial waste. Total coliform content is a significant facet of monitoring water quality as it signifies the potential microbial contamination in water (Zidan et al., 2022). Therefore, the timely prediction of total coliform is essential for ensuring the safety of the water body and determining its usage.

Conventionally, the total coliform content is determined by electrochemical detection (Mittelmann

*Corresponding Author

et al., 2002), membrane-based methods (Gafri et al., 2019), chromogenic or fluorogenic media (Kadyan et al., 2020), incubation and dilution method (Avigliano et al., 2015). However, these methods have various drawbacks associated with them, such as they are cumbersome to perform, labour-intensive, long incubation periods and limited sample throughput. To overcome these challenges, there has been a growing interest in developing machine learning (ML) models that can estimate total coliform levels based on various physicochemical and environmental parameters. Machine learning and related fields have been utilised in various areas pertaining to healthcare, environmental science, software, agriculture, and finance among other areas (Benoset al., 2021; Kauret al., 2017; Liu et al., 2023; Shekoohiyan et al., 2023). Its outstanding adaptability has recently shown its potential as a tool in the environmental science and engineering areas. For instance, Nunno et al. (Di Nunno et al., 2023) developed a simple and accurate machine learning model, using neural networks and random forests, for forecasting lake surface water temperature using daily air temperature as input. The model outperformed other models, had good forecasting capabilities for various time horizons, and was reliable for different temperature ranges. Another study by Kang et al. (2023) developed an autoencoder-based model using molecular structure embedding and machine learning to predict micropollutant treatability in drinking water treatment plants achieving a high accuracy. Machine learning models, thus, have high accuracy and are facile to execute, thus, making them excellent alternatives of the mechanical methods.

This study aims to investigate the feasibility of using machine learning algorithms to predict the total coliform content in water samples. By analysing a comprehensive dataset of Indian rivers comprising water quality parameters, we aim to develop an accurate and robust prediction model. Indian rivers are a vital part of the country's water resources, playing a crucial role in supporting various ecosystems, agriculture, and human settlements. However, the water quality of Indian rivers faces significant challenges due to pollution from industrial, agricultural, and domestic sources. Figure 1 shows the important rivers of India. Water pollution in Indian rivers is a complex and multidimensional problem that requires continuous monitoring, assessment, and remediation measures. In this study, regression methods have been used for the prediction of coliform content in the water of various Indian states. The study also aims to identify the significant features in predicting the pollutant level in water. Through precise total coliform presence prediction, policymakers can adopt proactive strategies to minimise health hazards linked to microbial contamination.

The rest of the study is organised as follows: Section: Related Work presents the background study and related work for predicting water quality in different areas. The methodology adopted in the study is discussed in detail in Section: Materials and Methods. The results are discussed in Section: Results and Discussion, followed by the concluding remarks in Section: Conclusion.

Related Work

The Water Quality Index (WQI) has gained widespread attention for water usage globally and has a longstanding history. Researchers have presented various methods for computing the WQI in the existing literature. For example, Pesce and Wunderlin (Pesce et al., 2000) proposed a comprehensive formula that incorporates normalized values and corresponding weights for each parameter. This approach is consistent with methodologies employed in numerous other studies (Misaghi et al., 2017; Wuet al., 2018; Wuet al., 2021). As highlighted in a study by Bui et al. (2020), conventional calculations of the WQI suffer from drawbacks. These include the need to measure multiple parameters, inconsistency arising from the use of various equations, and errors that may occur during the derivation of sub-indices. To address these challenges, an alternative approach based on Machine Learning (ML) has been proposed and proven effective in predicting the WQI, as demonstrated in previous research.

In the realm of machine learning applications, prediction accuracy is influenced by two primary factors: selecting an appropriate model and ensuring the quality of the training dataset. Artificial Neural Networks (ANN) and Support Vector Machines (SVM) have demonstrated outstanding performance in forecasting water quality components (Liu and Lu, 2014). The accuracy of machine learning models in prediction is also influenced by the choice of features used for training. Dissolved oxygen (DO) is a crucial parameter in surface water quality assessment, reflecting the health of aquatic ecosystems and their ability to support organisms. In a study on the Danube River, a linear polynomial neural network (PNN) model was employed to predict DO concentration (Tomić et al., 2018). Tripathi and Singal (2019) focussed on the first step of developing a Water Quality Index (WQI), which is parameter selection. By employing Principal Component Analysis (PCA), the study reduced the

Figure 1: Important rivers of India.

number of parameters from 28 to 9, namely Dissolved Oxygen, pH, Conductivity, Biological Oxygen Demand, Total Coliform, Chlorides, Magnesium, Sulphate, and Total Dissolved Solids.

Materials and Methods

Dataset

A dataset is one of the most significant factors in any machine learning-based prediction model. Water quality can be predicted using several important characteristics. However, physical parameters like temperature, pH levels, etc. can help in an efficient yet reliable prediction model for water quality. The present study employs a publicly available dataset of Indian territory (Kaggle Datasets). Table 1 defines the attributes used for the prediction model. Some of the attributes in the original dataset like the state and locations of the water sample were excluded from the prediction and used only for the exploratory data analysis.

Total coliform bacteria are a group of microorganisms commonly found in the environment, including soil, vegetation, and faeces of animals and humans. While not necessarily harmful themselves, their presence in water can indicate potential faecal contamination and the possible presence of pathogenic microorganisms. The dataset used in the present study for the water quality prediction includes data from different locations of Indian rivers and lakes. Figure 2 shows an analysis of the average coliforms state wise. One striking observation is the exceptionally high average count of total coliforms in Uttar Pradesh and Delhi. These states require immediate attention to address water quality issues and ensure access to clean drinking water for their populations. These results from Figure 2 provide insights into the relative levels of total coliforms present in water samples from different states in India, highlighting variations in water quality and potential contamination risks across the regions.

Table 1: Dataset description

Attribute	*Description*	*Statistics*
Temp	Temperature of the water sample, an important parameter affecting microbial growth and activity.	Min = 10.5 Max = 33.8 Average = 25.24
DO	Dissolved Oxygen level in the water, indicating the availability of oxygen for aquatic life and influencing microbial populations.	Min = 0 Max = 16.3 Average = 6.39
pH	Measure of the acidity or alkalinity of the water, impacting the survival and growth of microorganisms, including coliform bacteria.	Min = 6.3 Max = 14.7 Average = 7.78
Conductivity	Electrical conductivity of the water, indicating the presence of dissolved salts and ions, which can affect microbial activity.	Min = 39 Max = 24062 Average = 684.98
BOD	Biochemical Oxygen Demand, representing the amount of oxygen required by microorganisms to decompose organic matter, influencing microbial populations.	Min = 0.2 Max = 75.6 Average = 5.33
Nitrate_N_Nitrite_N	Concentration of nitrate and nitrite in the water, which can act as nutrients for microbial growth and potentially affect total coliform levels.	Min = 0 Max = 45.5 Average = 1.38
Total Coliform	The output variable represents the presence of total coliform bacteria, commonly used as an indicator of microbial contamination in water.	Min = 1 Max = 23816667 Average = 124397

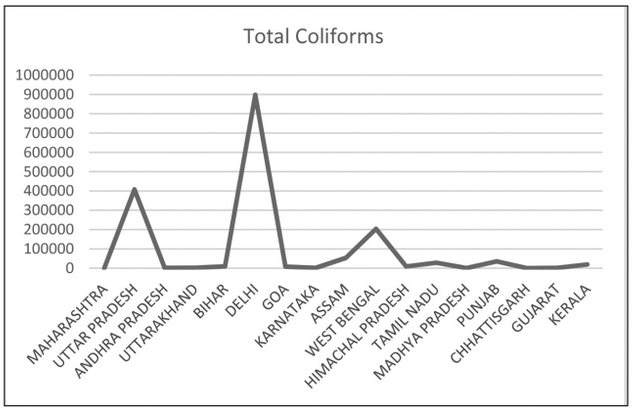

Figure 2: State wise average coliforms.

Data Pre-Processing

The dataset used in the analysis has a wide range of values for different attributes (Table 1). For instance, the values of Nitrate_N_Nitrite_N lie between 0 and 45.5, while conductivity ranges from 39 to 24062. Such a wide range of values in the dataset requires appropriate scaling, handling of outliers, and careful consideration of feature importance during the training and evaluation of machine learning models. Models might give more weight or importance to attributes with larger ranges, which can lead to biased predictions. Therefore, after the removal of insignificant attributes like the state and locations of the water sample from the dataset, the rest of the numeric values have been normalised using scaler and minmax functions. In addition to scaling, normalisation techniques were applied to ensure that the attribute distributions adhere to certain assumptions of the machine learning algorithms. For example, attributes like pH may need normalisation to have a symmetric distribution, as some algorithms assume Gaussian distribution for optimal performance. The formula for min-max scaling is given in Equation 1.

$$X_{\text{normalised}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X is the original value of the feature, $X_{\min}$ is the minimum value of the feature in the dataset, $X_{\max}$ is the maximum value of the feature in the dataset, and $X_{\text{normalised}}$ is the normalised value of the feature (between 0 and 1).

Machine Learning Model

In this study, we employed three different regression algorithms, namely linear regression, random forest regression, and gradient boosting regression, to predict the levels of total coliforms in the water samples. Each technique has its unique characteristics and is widely used in various domains for regression tasks (Kaur and

Kapoor, 2016). Linear regression is a well-established and widely used technique for regression analysis. It assumes a linear relationship between the input features and the target variable. In our analysis, we applied linear regression to establish a baseline prediction model for total coliforms. Random Forest regression is an ensemble learning method that combines the predictions from multiple decision trees to make accurate predictions. It mitigates the limitations of individual decision trees by leveraging the concept of bagging and feature randomness. Within the random forest model, we instantiated 1000 decision trees. The large number of trees allows for capturing complex relationships between the input features and the target variable. Gradient boosting regression is another ensemble learning technique that aims to minimise the errors of the previous models by focussing on the data points that were not well predicted. In gradient boosting regression, the algorithm starts with an initial weak model, often a shallow decision tree. It then fits subsequent models to the residual errors of the previous models. Each new model is trained to learn from the mistakes of the previous models and improve overall prediction accuracy. For our analysis, we utilised the gradient-boosting regression algorithm with specific parameter settings (Jasti et al., 2022; Singh and Singh, 2023). The learning rate of 0.01 controls the contribution of each weak model to the ensemble. The min_samples_split parameter was set to 5, specifying the minimum number of samples required to split an internal node during tree construction. Finally, we used a total of 1000 estimators, which indicates the number of weak models to be sequentially added to the ensemble.

In this study, we employed an 80-20 split evaluation measure to assess the performance of the regression models. The dataset was divided into two subsets: an 80% portion used for training the models and a 20% portion held out for evaluation purposes.

Evaluation Metrics

Evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were utilised for assessing the performance of the regression models. The formulae for each of the metrics are given as equations 2, 3 and 4, respectively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad (2)$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

By employing MAE, MSE, and RMSE as the evaluation metrics, we were able to comprehensively assess the performance of the regression models in predicting the levels of total coliforms in the water samples. These metrics provided insights into the accuracy, precision, and goodness-of-fit of the models, enabling us to compare and select the most suitable model for the task. The methodology applied in this study is summarised in Figure 3.

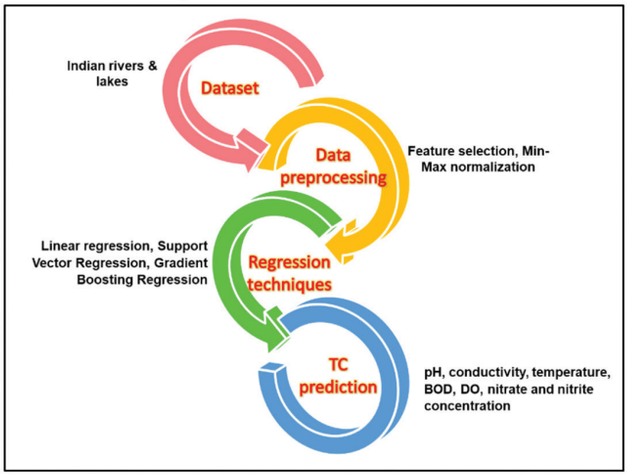

Figure 3: Schematic representation of methodology.

Results and Discussion

The prediction of water quality parameters, particularly the presence of total coliforms, is crucial for maintaining public health and ensuring the safety of water resources. In this research, we employed various machine learning techniques to predict total coliforms using a dataset containing attributes such as temperature, dissolved oxygen (DO), pH, conductivity, biochemical oxygen demand (BOD), and nitrate-nitrite (Nitrate_N_Nitrite_N). After pre-processing the dataset and scaling the attributes for normalization, we compared the performance of linear regression, support vector machines (SVM), and gradient boosting regression. Table 2 depicts the results obtained for each of the techniques.

Table 2: Prediction results

Technique/Parameter	*MAE*	*MSE*	*RMSE*
Linear Regression	0.035	0.0038	0.062
Support Vector Regression	0.033	0.0036	0.06
Gradient Boosting Regression	0.009	0.0002	0.015

Figure 4 depicts the results obtained for each of the techniques. Among the three techniques, gradient boosting regression yielded the highest accuracy in predicting total coliforms. However, to gain a deeper understanding of the predictive power of the attributes, we also assessed their importance in the prediction. Feature importance provides insights into the relative significance of the attributes and their contribution to the predictive model. When compared with existing studies, the proposed study aims to predict the water quality with respect to predicting the number of coliforms or related quality parameters. Existing studies use datasets corresponding to the presence of coliforms only, not the amount of such parameters in water, which itself can help in understanding the extent of water quality issues. Also, the gradient boosting approach is an ensemble method that builds a strong predictive model by iteratively adding weak models which has not been explored in the existing studies.

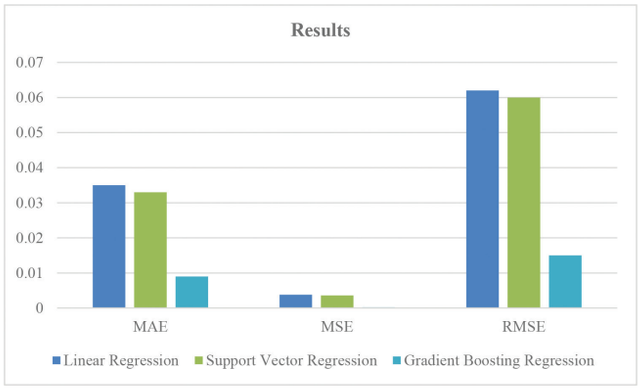

Figure 4: Comparison of results.

In this study, we employed Mean Decrease Impurity (MDI) as a measure of feature importance for the gradient-boosting regression model. The feature importance analysis (Figure 5) revealed that conductivity and temperature were the most influential attributes in predicting total coliforms, with conductivity being the most significant. The importance of conductivity suggests that the electrical conductivity of water, which is an indicator of the concentration of dissolved salts and ions, plays a crucial role in determining the presence of total coliforms. High conductivity levels can indicate the presence of pollutants or contaminants, which can promote the growth and survival of coliform bacteria. Monitoring and maintaining optimal conductivity levels can aid in preventing microbial contamination and ensuring water safety. Temperature was identified as the second most important attribute in the prediction. It has a significant impact on microbial growth and activity. Warmer temperatures can accelerate bacterial growth and increase the likelihood of microbial contamination. Therefore, accurate temperature monitoring and control are essential for preventing the proliferation of coliform bacteria and maintaining water quality. The third most important attribute, BOD, reflects the amount of dissolved oxygen consumed by microorganisms in the water. High BOD levels can indicate the presence of organic matter, which can serve as a nutrient source for coliform bacteria. By understanding the relative importance of these attributes, water management authorities can prioritise and focus their efforts on monitoring and controlling the most influential factors to ensure water quality.

The superior performance of gradient boosting regression compared to linear regression and SVM can be attributed to its ability to capture complex interactions and nonlinear relationships present in the dataset. By combining weak predictive models, such as decision trees, into a strong ensemble learner, gradient boosting effectively leverages the strengths of individual models and improves prediction accuracy. This finding highlights the potential of ensemble learning techniques in water quality prediction, particularly for complex and nonlinear relationships between attributes and total coliform presence.

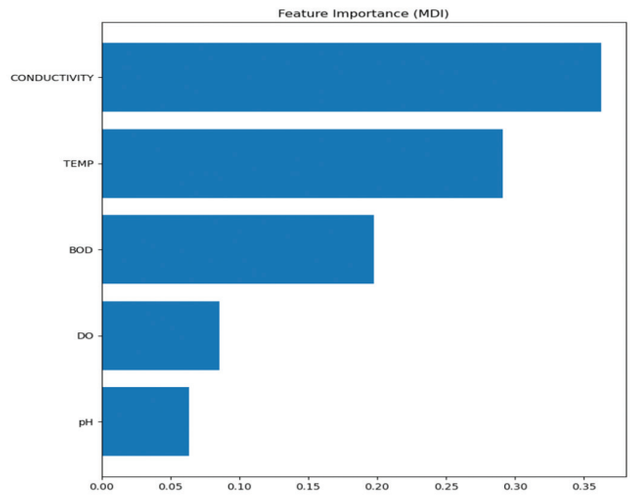

Figure 5: Important features for water quality.

The results obtained in this study have practical implications for water resource management and public health protection. By identifying the most influential attributes in predicting total coliforms, water management authorities can prioritise monitoring efforts and allocate resources effectively. For instance, continuous monitoring of conductivity, temperature, BOD, DO, and pH can provide early warnings of potential microbial contamination events, allowing for swift intervention and preventive measures. Additionally, the findings highlight the importance of maintaining optimal conditions for these attributes, such as controlling pollution sources, implementing proper wastewater treatment, and ensuring appropriate aeration and circulation of water. Furthermore, the feature importance analysis can guide future data collection efforts. By understanding the relative importance of different attributes, researchers and water quality professionals can focus on collecting accurate and representative data for the most influential factors. This can lead to the development of more refined and robust predictive models, ultimately enhancing water quality assessment capabilities.

Conclusion

Water quality assessment plays a crucial role in maintaining public health and ensuring the safety of water resources. This study examined the prediction of water quality, specifically the presence of total coliforms, using various machine-learning techniques. The results indicate that gradient boosting regression outperforms other techniques, providing the highest accuracy with mean absolute error (MAE) of 0.0349, mean squared error (MSE) of 0.0038, and root mean squared error (RMSE) of 0.0620. The feature importance analysis revealed that conductivity and temperature were the most influential attributes in predicting total coliforms. These findings contribute to the understanding of water quality assessment and provide valuable insights for water resource management and public health protection. By accurately predicting total coliform presence, decision-makers can take proactive measures to mitigate health risks associated with microbial contamination. Continued research and development in this field, including the incorporation of additional attributes and exploration of alternative algorithms, will further advance water quality assessment capabilities, ultimately benefiting communities worldwide.

References

Avigliano, E. and N.F. Schenone (2015). Human health risk assessment and environmental distribution of trace elements, glyphosate, fecal coliform and total coliform in Atlantic Rainforest Mountain rivers (South America). *Microchemical Journal*, **122:** 149-158. Doi: https://doi.org/10.1016/j.microc.2015.05.004

Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D. and D. Bochtis (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, **21(11):** 3758. doi: 10.3390/s21113758

Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H. and N. Kazakis (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, **721:** 137612. doi: https://doi.org/10.1016/j.scitotenv.2020.137612

Di Nunno, F., Zhu, S., Ptak, M., Sojka, M. and F. Granata (2023). A stacked machine learning model for multi-step ahead prediction of lake surface water temperature. *Science of The Total Environment*, **890:** 164323. doi: https://doi.org/10.1016/j.scitotenv.2023.164323

Gafri, H.F., Zuki, F.M., Aroua, M.K. and M.M. Bello (2019). Enhancing the anti-biofouling properties of polyethersulfone membrane using chitosan-powder activated carbon composite. *Journal of Polymers and the Environment*, **27:** 2156-2166. Doi: 10.1007/s10924-019-01505-z

Jasti, V.D.P., Kumar, G.K., Kumar, M.S., Maheshwari, V., Jayagopal, P., Pant, B., Karthick, A. and M. Muhibbullah (2022). Relevant-based feature ranking (RBFR) method for text classification based on machine learning algorithm. *Journal of Nanomaterials*, **2022(1):** 1-12. Doi: 10.1155/2022/9238968

Kadyan, S., Kumar, N., Lawaniya, R., Sharma, P.K., Arora, B. and N. Tehri (2020). Rapid and miniaturized method for detection of hygiene indicators, *Escherichia coli* and coliforms, in dairy products. *Journal of Food Safety*, **40(5):** 12839 doi: 10.1111/jfs.12839

Kaggle Datasets. Available at: https://www.kaggle.com/datasets

Kang, J.K., Lee, D., Muambo, K.E., Choi, J.W. and J.E. Oh (2023). Development of an embedded molecular structure-based model for prediction of micropollutant treatability in a drinking water treatment plant by machine learning from three years monitoring data. *Water Research*, **239:** 120037. https://doi.org/10.1016/j.watres.2023.120037

Kaur, I. and N. Kapoor (2016). Token based approach for cross project prediction of fault prone modules. *In:* 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), New Delhi, India, pp. 215-221, doi: 10.1109/ICCTICT.2016.7514581.

Kaur, I., Narula, G.S. and V. Jain (2017). Differential analysis of token metric and object oriented metrics for

fault prediction. *International Journal of Information Technology*, **9:** 93-100. doi: 10.1007/s41870-017-0004-0

Liu, G., Tian, S., Xu, G., Zhang, C. and M. Cai (2023). Combination of effective color information and machine learning for rapid prediction of soil water content. *Journal of Rock Mechanics and Geotechnical Engineering*, **15(9):** 2441-2457. doi: https://doi.org/10.1016/j.jrmge.2022.12.029

Liu, M. and J. Lu (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? *Environmental Science and Pollution Research*, **21:** 11036-11053. doi: 10.1007/s11356-014-3046-x

Madhav, S., Ahamad, A., Singh, A.K., Kushawaha, J., Chauhan, J.S., Sharma, S. and P. Singh (2020). Water Pollutants: Sources and Impact on the Environment and Human Health. *In:* Pooja, D., Kumar, P., Singh, P. and S. Patil (eds). Sensors in Water Pollutants Monitoring: Role of Material. Advanced Functional Materials and Sensors. Springer, Singapore. https://doi.org/10.1007/978-981-15-0671-0_4

Misaghi, F., Delgosha, F., Razzaghmanesh, M. and B. Myers (2017). Introducing a water quality index for assessing water for irrigation purposes: A case study of the Ghezel Ozan River. *Science of the Total Environment*, **589:** 107-116. https://doi.org/10.1016/j.scitotenv.2017.02.226

Mittelmann, A.S., Ron, E.Z. and J. Rishpon (2002). Amperometric quantification of total coliforms and specific detection of *Escherichia coli*. *Analytical Chemistry*, **74(4):** 903-907. doi: 10.1021/ac0156215

Pesce, S.F. and D.A. Wunderlin (2000). Use of water quality indices to verify the impact of Córdoba City (Argentina) on Suquía River. *Water Research*, **34(11):** 2915-2926. https://doi.org/10.1016/S0043-1354(00)00036-1

Shekoohiyan, S., Hadadian, M., Heidari, M. and H. Hosseinzadeh-Bandbafha (2023). Life cycle assessment of Tehran municipal solid waste during the COVID-19 pandemic and environmental impacts prediction using machine learning. *Case Studies in Chemical and Environmental Engineering*, **7:** 100331. https://doi.org/10.1016/j.cscee.2023.100331

Singh, P. and D.P. Singh (2023). Comparative Analysis of Machine Learning Classifiers for Heart Disease Prediction in Cloud Environment. *In:* 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 552-556.

Tomić, A.Š., Antanasijević, D., Ristić, M., Perić-Grujić, A. and V. Pocajt (2018). A linear and non-linear polynomial neural network modeling of dissolved oxygen content in surface water: Inter-and extrapolation performance with inputs' significance analysis. *Science of the Total Environment*, **610-618:** 1038-1046. https://doi.org/10.1016/j.scitotenv.2017.08.192

Tripathi, M. and S.K. Singal (2019). Use of principal component analysis for parameter selection for development of a novel water quality index: A case study of river Ganga India. *Ecological Indicators*, **96(1):** 430-436. https://doi.org/10.1016/j.ecolind.2018.09.025

Wu, Z., Lai, X. and K. Li (2021). Water quality assessment of rivers in Lake Chaohu Basin (China) using water quality index. *Ecological Indicators*, **121:** 107021. doi: https://doi.org/10.1016/j.ecolind.2020.107021

Wu, Z., Wang, X., Chen, Y., Cai, Y. and J. Deng (2018). Assessing river water quality using water quality index in Lake Taihu Basin, China. *Science of the Total Environment*, **612:** 914-922. https://doi.org/10.1016/j.scitotenv.2017.08.293

Zidan, K., Sbahi, S., Hejjaj, A., Ouazzani, N., Assabbane, A. and L. Mandi (2022). Removal of bacterial indicators in on-site two-stage multi-soil-layering plant under arid climate (Morocco): Prediction of total coliform content using K-nearest neighbor algorithm, *Environmental Science and Pollution Research*, **29:** 75716-75729. Doi: 10.1007/s11356-022-21194-x