Grenze International Journal of Engineering and Technology, June Issue

Effective Hyper-Parameter Tuning of Machine Learning Model for Analysis Drinking Water Quality

Ms.Neha Tanwar[1] and Dr. Anil V Turukmane[2]
[1-2]School of Computer Science & Engineering, VIT-AP University. Amaravati, Vijayawada, AP India
Email: neha.23phd7088@vitap.ac.in, anil.turukmane@vitap.ac.in

***Abstract*— The water quality of the river Krishna was carried out. Thermal conductivity, dissolved oxygen, total particulates in solution, biochemical oxygen demand, and pH levels were among these parameters. For a thorough knowledge of the Krishna River's overall water quality, each of these measurements is essential. This comprehensive technique enables a detailed evaluation that considers several elements of the river's environmental conditions and composition. When evaluating the river's water quality, several factors are crucial. In the Indian state of Karnataka, scientists measured the water quality at six different locations along the Krishna River. The Krishna River, which originates in the Bay of Bengal and travels around 1400 kilometers, is the fourth biggest river in India. Only the 483 km or so of the Krishna River that flows through the state of Karnataka was taken into consideration for our analysis. Rapid industrialization has recently been applied to the mineral-rich river basin, causing pollution. Unusable for human use, river water is inevitably going to get contaminated by a variety of contaminants, including industrial, agricultural, and municipal waste. The current manual procedure, which is conventional, is a very sluggish operation. Staff must collect water samples from the location, transport them to the laboratory, and analyze the samples for a variety of water characteristics. This is an expensive and time-consuming procedure. Therefore, the people living in the river basin region are not able to get timely information regarding the quality of the water. This feature makes it perfect for real-time water quality analysis of samples collected from the Krishna River. Water quality has a profound influence on the health of ecosystems and the welfare of human beings. Both people and animals may get sick from water pollution. Rivers are now one of the most widely utilized natural water sources on the planet, but in the last ten years, human activity and climate change have increased river pollution.**

***Index Terms*— Machine learning, Water quality index, Decision tree, Random Forest.**

I. INTRODUCTION

A vital component of life on Earth, water comes from both surface and subsurface sources, including precipitation and glacier melting. Water is formed via the hydrological cycle, which emphasizes the water's vital role in supporting life. Water is a major resource that is greatly relied upon by industries, which are essential for the growth of society. The health of water bodies is impacted when around 50% of the waste from industries end up in rivers and oceans. This makes a close examination of the Krishna River's water quality necessary. The formula known as the Water sanitation Index (WQI) was employed to assess the water's sanitation. WQI provides key stakeholders and policymakers with an efficient and user-friendly means of receiving complete information on the

Grenze ID: 01.GIJET.10.2.503

overall state of water quality. Water quality has a significant influence on the environment, as shown by the direct link between it and public health. Water is a vital resource for many uses, including drinking, farming, and industrial operations. Notably, the increase in leisure activities involving water has turned them into a draw for travelers. Rivers like the Krishna are readily available and often used in the formation of human cultures, but caution must be used. Using other sources of water, such as saltwater and groundwater, may sometimes help solve problems. But if management is not done carefully, problems might occur, such over-extraction that causes land to sink. River engineering is the resultant discipline that focuses on sediment movement, water quality, morphological changes, and pollutant transmission mechanisms. For sustainable and ethical use of water resources in the Krishna River basin and beyond, a thorough investigation of these factors is essential. For the health of humans, animals, and our environment, the rivers, lakes, and seas must all have clean water. Human activity, agriculture, and the environment are just a few of the many variables that may contaminate water. In particular, climate change, more heavy rain that causes floods, farming runoff, and garbage are some of the things that can make water dangerous. Human activity and climate change have increased river pollution throughout the last ten years. According to estimates, drinking water pollution kills 485 thousand people annually due to diarrhea. Water may be toxic not only when drunk but also when bathed in or swimmer in due to pollution, which can cause life-threatening diseases. Entering contaminated coastal waters, for example, causes about 120 million instances of gastrointestinal sickness

II. Relevant Water Quality Research, (2000–2023)

Water quality greatly affects human and ecological health. Pollution in lakes, rivers, and coastal waterways may degrade ecosystems, kill animals, and make people sick. Labs test water samples for quality, which is expensive and time-consuming. Machine learning for water quality forecasting is becoming more popular due to its speed, accuracy, and affordability. It also lets users publish notifications about toxic lakes, rivers, and beaches. This portion will critically assess water quality studies, its implications on human health and the ecosystem, and machine learning forecasts. Sulfate, pH, hardness, organic carbon, trihalomethanes, turbidity, conductivity, and TDS were included in the dataset. Processing employed the dataset's most essential attributes to explore and scale features. The dataset's correlations were shown as a correlation heatmap. The dataset was divided into training and assessment datasets, with 30% for testing and 70% for training. The Decision Tree predictor was crucial to decision-making. In addition to data classification, the decision tree is a useful tool for monitoring and learning. Its applicability as a problem-solving and analytical tool in the study was shown by its insights and solutions. Because it could anticipate and learn, the decision tree may assist organize and interpret huge water quality statistics.[1]. Coastal water quality is assessed using case studies. The Irish EPA monitors port water quality with 32 stations, focusing on 29. Water quality indicators are statistically analyzed and given recommended values in Table 1. The table shows ammonia nitrogen, molybdate reactive phosphorus, transparency (TRAN), chlorophyll a (CHL) for algal measurement, total organic nitrogen (TON), MRP, and dissolved oxygen (DOX). Seven of 10 water quality measures were necessary to achieve research goals. Table.3 suggests a coastal water quality classification method. This method classifies water quality using these indicators. Table 4 shows the outcomes of trial-and-error classifier evaluation to find a reliable algorithm. Existing WQI models evaluate water quality using diverse methods, therefore model interpretation and evaluation are crucial. Two predictive classifier models are explored. Supervisory machine learning, particularly the Naïve Bayes (NB) classifier, is a popular and effective approach for binary and multiclass classification. Adding emphasis. It classifies by closeness.

A. The Environment's and Human Health's Reactions to Water Quality

Human health may be put at risk by contaminated water, which can spread illnesses that can lead to fatalities or severe illnesses. The issue of maintaining clean, pollutant-free water sources is one that faces people everywhere. among metropolitan areas, low-quality drinking and bathing water may cause illness, disease outbreaks, and even fatalities among people. [1] A number of WQIs, including the surface quality water index, the ground quality water index, and others, have been established by various nations and organizations to achieve particular objectives. A number of factors have prompted criticism of this method, such as (i) uncertainties, (ii) the dependability of the models, (iii) the lack of transparency, and (iv) the sensitivity of the models. Recent research has shown that the WQI model created a significant amount of ambiguity in the final result because of its Furthermore, a number of studies have recently shown that the fact that the WQI index model employs many categorization methods means that it does not accurately reflect the real condition of water quality.[2] It is advised to interpret the WQI score using a variety of qualitative metrics, such as "good," "bad," "very bad," "poor," marginal, higher, lower, etc. Because of this, there is a great deal of ambiguity in the accurate categorization of water quality due to the many approaches that provide multiple interpretations for the same water qualities [3]. These kinds of issues might be dealt with as categorization "megaphoning problems." A rising number of people are getting concerned about these issues as the WQI model progressively gains traction. Recent research indicates that because of these problems, the existing WQI model gives the water resources management unclear information, which prevents bodies from acting promptly enough. The primary cause of the declining water quality in metropolitan areas is now human-generated municipal and industrial wastewater. Research on surface water quality and machine learning has grown more popular [4].

B. A Critical Evaluation of Current Techniques, Algorithms, and Outcomes

Table 1 critiques present methods and their results. Most models performed well, however supervised learning regression approaches like Extreme Gradient Boosting, Multiple Linear Regression, Support Vector Machines, and Random Forest Regression yield accurate results. Back-propagation neural networks and Gaussian process regression models exhibited higher MSEs and RMSEs than other approaches but R-Squares over 94%. Even with poor R-squared results in their support vector machine, random forest, and polynomial regression models, the research's 699-row sample size should be considered.

C. Current Water Quality Forecasting Machine Learning Algorithms

Machine learning (ML) replicates human behavior using data and algorithms and is linked to statistics and computer science. ML is increasingly utilized in environmental studies to predict water quality as machine learning capabilities improve. Machine learning has been used to forecast water quality in several studies. emphasis on classification methods for Indian Narmada River water quality forecasts [1]. They compare Naive Bayes, Decision Trees, and Support Vector Machine algorithms using temperature, BOD, turbidity, pH, DO, nitrate, and other parameters. When few data points are mislabeled, Decision Trees outperform Support Vector Machines. The writers argue that Naive Bayes predicts water quality lowest and misclassifies the most data.[2] Suma S et al. (2023) employed SVMs to anticipate river water quality, specifically India's Krishna River. The strongest predictor was suspended solids with 93.6% R Squared. They predicted ammonia nitrogen, oxygen dissolved in water, chemical oxygen demand, biochemical oxygen demand, and suspended particles using six quality parameters. That 148-record dataset performed nicely for SVMs.5 There are additional techniques to assess water quality than categorization. throughout their probe [3]. Ensemble methods and neural networks may forecast estuary water quality. Forecasting water quality is important since estuaries may pollute coastal regions. The 824 records in this study were used to create training and test datasets. the best at forecasting estuarine ammonia nitrogen, but like earlier study, the dataset was small and no other factors, like as climate or human activity, were considered, which may have improved the model's accuracy. [4] machine learning classification approaches to find marine contaminant-exposed dolphin gene expression differences. The ability to create additional water limits population expansion and area development. They tested ANN and SVM algorithms based on many sensors for real-time water quality monitoring in 2007. Both models performed well after running recognition rates for the two water types. SVM outperformed ANN in stability [5].

D. Summary and Highlighted Deficiencies

The literature review suggests ensemble techniques, neural networks, regression, and classification for water quality forecasting. Random Forests and Support Vector Machines are common supervised machine learning algorithms. Forest Regression, Multiple Linear Regression, and Extreme Gradient Boosting forecast water quality accurately, however Neural Networks have performed inconsistently across experiments. While many studies receive high scores, the number of datasets used may affect the results. In prior studies on drinking water quality, Muharemi et al. proposed the K-Nearest Neighbor technique and Logistic Regression-based Neural Network Classification [1]. Haghiabi et al. tested support vector machines, group data management systems, and artificial neural networks to forecast water quality in Tireh River, Iran [2]. Zhang et al. created a twofold movement window technique and statistical model to detect water quality abnormalities [3]. The method outperformed AD and ADAM in pH anomaly detection [6-10].

TABLE I: CRITICAL EVALUATION OF CURRENT TECHNIQUES FOR PREDICTING WATER QUALITY

Algorithm	Evaluation	Result	Author
ANN	R-Squared, RMSE	92%, 1.35	Wang et al
Random Forest	R-Squared, RMSE	91%, 1.15	Ahmed et al.

Multiple linear regression	R-Squared, RMSE	94%, 1.14	Ahmed et al.
XG Boost	R-Squared, RMSE	2.31, 9.57	Thoe et al.
Support Vector Machine	MAE, MSE, RMSE, R-Squared	3.09, 67%	Radhakrishnan
Polynomial Regression	MAE, MSE, RMSE, R-Squared	2.44, 10.63	Chafloque et al.
Support Vector Machine (Classification)	R-Squared	2.73, 12.73	Su et al
Naïve Bayes (Classification)	Accuracy	87%	Lopez et al
Decision Trees (Classification)	Accuracy	75%	Rani et al

III. Machine Learning Applications For Various Water Environments

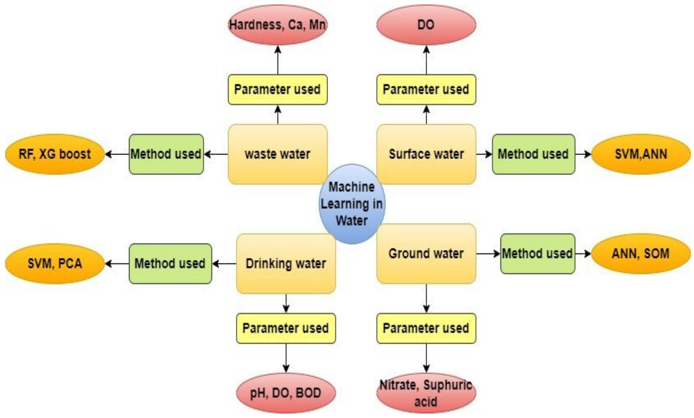

Figure 1 Methods used for Water quality prediction

A. Surface water applications

Surface water, abundant in lakes, rivers, and reservoirs, is a vital resource with many applications. Relevance affects many elements of human and environmental health. Its main duty is providing clean water to numerous communities worldwide. Surface water supplies require substantial treatment to become crucial reservoirs that serve communities with drinking water. Agriculture relies on surface water for crop irrigation [10-14]. Dependence on rivers and canals ensures global food security and agricultural output. Businesses utilize surface water for energy generating, cooling, and manufacturing due to its accessibility and cost. The renewable energy business benefits from rivers and reservoirs' hydropower. Dams provide sustainable energy by converting water's kinetic energy into electricity. Surface water bodies are popular outdoor recreation and tourism locations as well as utilitarian uses. Lakes and rivers sustain regional economies and offer swimming, fishing, and boating. Most importantly, surface water nurtures aquatic ecosystems by providing homes for plants and animals. Wetlands are crucial to biodiversity and numerous species. Since they accept wastewater from households, businesses, and farms, surface water bodies are essential to wastewater treatment and disposal. Treatment is necessary to prevent pollution and protect surface water quality. Traditional rivers and lakes were natural channels for navigation, making it simpler to transfer people and goods and sustaining trade routes. Flood-prone areas need excellent surface water management to reduce and manage flooding [15-20].

B. Groundwater applications

Earth water is valuable. Therefore, groundwater safety is essential to human health. Machine learning can predict contamination source and quality in groundwater studies. Subsurface groundwater is essential to environmental and biological activities. Every community relies on groundwater for clean drinking water. Subsurface water from wells and aquifers is often treated to fulfill quality requirements. Groundwater is needed for agricultural irrigation, civilization's basis. The dependability and accessibility of groundwater wells assure agricultural production and food security. Energy, cooling, and manufacturing use groundwater. Thermally stable groundwater is ideal for industrial cooling. Energy companies employ renewable geothermal energy from groundwater. Groundwater supports ecosystems and is useful. Having numerous plants and animals helps protect wetlands. Because groundwater-dependent ecosystems are more vulnerable to water level fluctuations, biodiversity protection requires sustainable groundwater management. Groundwater is vital for infrastructure construction because to its stability and dependability. A drought buffer, groundwater supplies water when surface water is low [20-25].

C. Drinking water applications

Machine learning improves drinking water management by forecasting and analyzing source water quality for pollution control, providing early alerts, improving treatment procedures, and increasing efficiency. It enables real-time analysis, anomaly identification, and chemical dosage adjustment for safer water and resource use.

D. Waste water applications

Machine learning has transformed wastewater treatment, from plant management and technological optimization to monitoring and forecasting. Main uses include water quality forecasts and monitoring. Historical and real-time sensor data is used by machine learning to anticipate water quality changes. Wastewater treatment systems work better when they can estimate pollutant levels. Treatment technology improvement incorporates machine learning. The algorithms examine complex links between treatment parameters and results to boost aeration, chemical doses, and other operational elements [27-30]. Increased efficiency and energy and resource savings make wastewater treatment more sustainable. Machine learning enhances wastewater treatment facility operations. These systems assess sensor data during therapy for real-time decision-making. Predictive maintenance models can predict equipment breakdowns and speed up repairs [31-32]. Machine learning also helps create adaptive control systems that increase WWTP performance in different scenarios. Residential and commercial wastewater contains several contaminants, making water quality evaluation before treatment necessary. Machine learning detects and quantifies contaminants to assess water quality and adjust treatment. Advanced, data-driven, and adaptable machine learning algorithms improve wastewater treatment, protecting the environment and public health from assessment to treatment [33-35].

IV. METHODS ON WATER QUALITY

TABLE II PARAMETERS ALONG WITH THEIR "WHO" STANDARD LIMITS: -

S.NO.	Parameters	Method Adopted	Indian Standard
1	Potential of hydrogen	Electrometric Method	6.5-8.5
2	Dissolved oxygen (milligrams per liter)	Azide alteration	7.6-7.0
3	Biological oxygen demand (milligrams per liter)	Azide modification	< 30
4	COD (mg/L)	Dichromate reflux	< 250
5	Calcium concentration (mg/L)	EDTA Titration Method	75
6	Magnesium concentration per liter	EDTA Titration Method	30
7	Hardness in (mg/L)	EDTA Titration Method	300
8	Chloride concentrations (mg/L)	Argentometric Titrimetric method	250
9	Sulfuric acid (mg/L)	colorimetry turbidimetry method	200
10	Nitrates (mg/L)	Colorimetric Turbidimetric technique	45

V. Water Quality Classification Based On Architectural Design

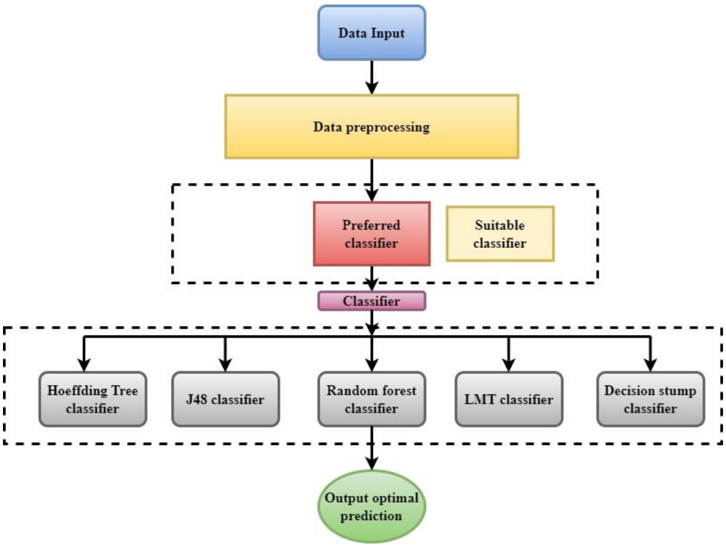

Figure 2 Prediction Methods

A. Classifier Comparison of Water Quality Predictions: Accuracy Assessment

This decision tree classifier handles streaming data. The decision tree develops as branches become available. The method detects split tree nodes using Hoeffding bounds. Hoffding constraints let the algorithm make confident judgments with fewer data. Hoeffding Tree is used in financial, sensor, and internet traffic monitoring. Hoeffding Tree handles big data sets with less memory and processing. System flexibility to shifting data distributions is another benefit [31-34].

J 48- J48 is a popular C4.5 decision tree classifier. J48 uses training data to generate greedy, top-down decision trees. Every tree node recursively divides data by its most important attribute. The tree will deepen when subgroups are homogenous. J48 studies water quality and more [35,36].

Random Forest - The Random Forest Classifier is used in multiple-decision trees. Overfitting and accuracy are reduced using ensemble classifiers. Training data and characteristics are randomly selected at each forest node to build decision trees. Conclusion is based on several decision tree predictions. Feature relevance is also scored for classification. Finance, biology, and image recognition use Random Forest. Random Forest classifies noisy, high-dimensional data better than standard methods. less outlier-prone. The categorization work aspects are also effectively detailed [36-37].

D LMT Classifier: Decision trees and logistic regression are used. Based on training data, LMT predicts tree nodes using logistic regression models. Pruning the model reduces overfitting and simplifies the tree. LMT helps with noisy data and multi-attribute classification. Bioinformatics, finance, and environmental studies use it. Decisions benefit from LMT's accuracy and interpretability [32-37].

Decision-Stump Classifier: Decision stump classifier has one root and two leaf nodes. Data is divided by a single attribute using the majority class subset. When there is only one meaningful data point or a quick judgment is needed, the decision stump strategy may work. Since it evaluates one classification criteria, it is less accurate than more elaborate decision tree algorithms. Text categorization and picture recognition use decision stumps. A fast and simple decision stump algorithm helps when processing power is limited [39-40].

VI. Conclusion

Hyper-parameter tuning is crucial to establishing accurate machine learning models for drinking water quality analysis. This simple step-by-step tutorial will help you understand the process, improve your model, and ensure safe and clean drinking water for everybody. Ping accurate drinking water quality machine learning models. This simple step-by-step tutorial will help you understand the process, improve your model, and ensure safe and clean drinking water for everybody. Water quality is assessed using decision tree categorization. Water quality may be predicted using many input parameters. This strategy may enhance water management and decision-making to guarantee environmental sustainability and offer safe, clean drinking water. Decision tree methods for water quality categorization include Decision Stump, J48, Hoeffding Tree, LMT, and random forests. The suggested strategy involves gathering and splitting data. Making a decision tree model, measuring its performance, and improving it with training and testing sets, putting it on an appropriate platform, and monitoring and updating. It is important to note that a classification model's precision and dependability depend on several aspects, including feature selection, methodologies, parameter optimization, and data quality and quantity. Ensemble techniques and deep learning may improve water quality classification models.

REFERENCES

[1] Kengnal P, Megeri MN, Giriyappanavar BS, Patil RR (2015) Multivariate analysis for the water quality assessment in rural and urban vicinity of Krishna River (India). Asian J Water Environ Pollut 12:73–80

[2] Loganathan K, Ahamed AJ (2017) Multivariate statistical techniques for the evaluation of groundwater quality of Amaravathi River Basin: South India. Appl Water Sci 7:4634649

[3] Roy PK, Pal S, Banerjee G et al (2014) Variation of water quality parameters with siltation depth for river Ichamati along international border with Bangladesh using multivariate statistical techniques.

[4] Shah KA, Joshi GS (2017) Evaluation of water quality index for River Sabarmati, Gujarat, India.

[5] Sharma D, Kansal A (2011) Water quality analysis of River Yamuna using water quality index in the national capital territory, India

[6] Water N, Monitoring Q, Pradesh A, et al. (2000) Water quality status of river Krishna Water quality of river Krishna is monitored regularly on monthly basis under National Water Quality Monitoring project (NWMP) by Andhra Pradesh Pollution Control Board

[7] Zhao Y, Xia XH, Yang ZF, Wang F (2012) Assessment of water quality in Baiyangdian Lake using multivariate statistical techniques.

[8] Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. Environ. Sustain. India

[9] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., 2019, E ffi cient Water Quality Prediction Using Supervised 1–14.

[10] Aldhyani, T.H.H., Al-Yaari, M., Alkahtani, H., Maashi, M., 2020. Water Quality Prediction Using Artificial Intelligence Algorithms. Appl. Bionics Biomech. 2020.

[11] River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. J. Environ. Chem. Eng. 9,

[12] Babbar, R., Babbar, S., 2017. Predicting river water quality index using data mining techniques. Environ.

[13] Brown, C.A., Nelson, W.G., 2015. A method to identify estuarine water quality exceedances associated with ocean conditions. Environ. Monit. Assess.

[14] Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., Ren, H., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data.

[15] Danades, A., Pratama, D., Anggraini, D., Anggriani, D., 2017, Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status. Proc. 2016

[16] Dezfooli, D., Hosseini-Moghari, S.M., Ebrahimi, K., Araghinejad, S., 2018. Classification of water quality status based on minimum quality parameters: application of machine learning techniques. Model.

[17] EPA, 2001. Parameters Of Water Quality: Interpretation and Standards. EnvironmentalProtection AGENCY. Ireland.

[18] Garabaghi, F.H., 2021, Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classi cation of Water Quality Indices Based on Different Subset of Features.

[19] Gikas, G.D., Sylaios, G.K., Tsihrintzis, V.A., Konstantinou, I.K., Albanis, T., Boskidis, I., 2020. Comparative evaluation of river chemical status based on WFD methodology and CCME water quality index. Sci. Total Environ

[20] pta, S., Gupta, S.K., 2021. A critical review on water quality index tool: Genesis, evolution and future directions.

[21] Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A., 2018. Water quality prediction using machine learning methods. Water Qual

[22] Sherif, M., Sefelnasr, A., El-Shafie, A., 2022. Application of Soft Computing in Predicting Groundwater Quality Parameters. Front. Environ

[23] Hartnett, M., Nash, S., 2015. An integrated measurement and modeling methodology for estuarine water quality management. Water Sci.

[24] Islam Khan, M.S., Islam, N., Uddin, J., Islam, S., Nasir, M.K., 2021. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. J. King Saud.

[25] Kaur, A., Khurana, M., Kaur, P., Kaur, M., 2021. Classification and Analysis of Water Quality Using Machine Learning Algorithms, Lecture Notes in Electrical Engineering. Springer, Singapore

[26] Khoi, D.N., Quan, N.T., Linh, D.Q., Nhi, P.T.T., Thuy, N.T.D., 2022. Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River.

[27] Khullar, S., Singh, N., 2021. Machine learning techniques in river water quality modelling: A research travelogue. Water Sci. Technol. Water Supply 21.

[28] Leong, W.C., Bahadori, A., Zhang, J., Ahmad, Z., 2021. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). Int. J. River Basin Manag.
[29] Malek, N.H.A., Yaacob, W.F.W., Nasir, S.A.M., Shaadan, N., 2022. Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water
[30] Modaresi, F., Araghinejad, S., 2014. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water Resour.