ISSN: 1648-6897 / eISSN: 1822-4199

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of ENVIRONMENTAL ENGINEERING & LANDSCAPE MANAGEMENT

2024
Volume 32
Issue 2
Pages 117–127
https://doi.org/10.3846/jeelm.2024.20979

EVALUATING THE PERFORMANCE OF MACHINE LEARNING APPROACHES IN PREDICTING ALBANIAN SHKUMBINI RIVER'S WATERS USING WATER QUALITY INDEX MODEL

Lule BASHA[1]✉, Bederiana SHYTI[2], Lirim BEKTESHI[3]

[1]Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania
[2]Department of Mathematics, Faculty of Natural Sciences, University of Elbasan "Aleksandër Xhuvani", Elbasan, Albania
[3]Department of Chemistry, Faculty of Natural Sciences, University of Elbasan "Aleksandër Xhuvani", Elbasan, Albania

Highlights:

- calculating the Shkumbini River's WQI scores based on Brown et al.'s (1972) methodology;
- identifying water quality classes through a classification scheme;
- a statistical summary of WQI models input;
- selection of four machine learning classifier models on the acquired dataset;
- methodological description of each of the models, with their advantages and disadvantages;
- selecting the best model by considering factors such as precision, mean square error, and RSquare;
- this study showed that ensemble tree-based (XGBoost and Random Forest) approaches outperform other algorithms based on model errors in terms of stability and reliability;
- determining important indicators and their relative rankings in this stage;
- in this study, we discovered that BOD, HCO_3, and TP had the greatest positive effects on the water quality of the Shkumbini River, according to all the models;
- the confusion matrices for the four models are constructed;
- the performance of the classifiers was evaluated using validation measures, including accuracy, sensitivity, and F1 score.

Article History:

- received 28 August 2023
- accepted 06 February 2024

Abstract. A common technique for assessing the overall water quality state of surface water and groundwater systems globally is the water quality index (WQI) method. The aim of the research is to use four machine learning classifier algorithms: Gradient boosting, Naive Bayes, Random Forest, and K-Nearest Neighbour to determine which model was most effective at forecasting the various water quality index and classes of the Albanian Shkumbini River. The analysis was performed on the data collected during a 4-year period, in six monitoring points, for nine parameters.

The predictive accuracy of the models, XGBoost, Random Forest, K-Nearest Neighbour, and Naive Bayes, was determined to be 98.61%, 94.44%, 91.22%, and 94.45%, respectively. Notably, the XGBoost algorithm demonstrated superior performance in terms of F1 score, sensitivity, and prediction accuracy, the lowest errors during both learning (RMSE = 2.1, MSE = 9.8, MAE = 1.13) and evaluating (RMSE = 0.0, MSE = 0.01, MAE = 0.01) stages. The findings highlighted that Biochemical oxygen demand (BOD), Bicarbonate (HCO_3), and Total Phosphor had the most positive impact on the Shkumbini River's water quality. Additionally, a statistically significant, strong positive correlation (r = 0.85) was identified between BOD and WQI, emphasizing its crucial role in influencing water quality in the Shkumbini River.

Keywords: Water Quality Index model, Shkumbini River, machine learning classifier, model accuracy.

✉Corresponding author. E-mail: *lule.hallaci@fshn.edu.al*

1. Introduction

The future problem for the planet is to preserve "high level of water quality" because freshwater is a crucial bio-indicator for living creatures in any aquatic habitat. To secure a higher standard of living, industrialization and urbanisation have intensified day by day. As a result, over many years, freshwater consumption has dramatically increased. A lot of factors, including an institutional framework, a qualified workforce, regulatory limitations, financial flexibility, and the availability of resources, go into the control of water resources, which is a critical activity. A variety of tools and techniques have been developed for monitoring water quality. One of them is the index for water quality.

Recently, this technique has been used often to evaluate the water quality. Its use has expanded quickly because it may use straightforward mathematical operations to transform a sizable amount of data on water quality into a unitless numerical statement. Previous studies have shown that using the mathematical model WQI for conventional water quality assessment has limitations. To obtain a final index value using the WQI mathematical model, complex calculations are required. While existing WQI models have applied various statistical approaches to identify crucial water quality indicators (Uddin et al., 2021, 2022a, 2022b, 2022c), recent studies have underscored their limitations in effectively selecting key indicators (Sutadian et al., 2018; Uddin et al., 2022a). Uddin et al. (2021) conduct a comprehensive analysis of widely used Water Quality Index (WQI) models, aiming to assess their structures, components, and applications. The study reviews 21 global WQI models by critically examining 110 published manuscripts. Seven fundamental WQI models, influencing the development of others, undergo detailed scrutiny. The paper delves into the history, basic structure, and critical elements of WQI models, provides detailed insights into the seven primary models. The study revealed that, despite similar overall structures, WQI models differed significantly in finer components. Issues of eclipsing and uncertainty in model development were identified. WQI models typically involve four stages but exhibit region-specific variability based on waterbody type, intended uses, local guidelines, and data availability. There is substantial diversity in parameters, weightings, and criteria among WQI models, hindering comparability across study areas. Streamlining with international guideline values may enhance their utility. Model updating is crucial, incorporating new parameters of interest. Eclipsing and uncertainty impact model accuracy, with reliance on expert opinions. Mathematical techniques and computer-based systems can mitigate uncertainty, promoting more accurate index computations. Model uncertainty assessment should be a standard practice in WQI applications, ensuring reliability and precision.

Ravindra et al. (2023) study the groundwater quality assessment was conducted in Guntur district, Andhra Pradesh, India, focusing on Water Pollution Index (WPI) and Water Quality Index (WQI). Results indicated alkaline and very hard groundwater with high concentrations of various ions. TDS, TH, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^{-}, Cl^{-}, $NO_3^{\ -}$, and F^{-} exceeded recommended limits for drinking water in a significant percentage of samples. WQI classified 75% as poor and 25% as very poor groundwater quality, covering 85.84% and 14.06% of the study area, respectively. The groundwater was deemed unsuitable for drinking, recommending preventive measures for human health protection, including safe water supply, desalinization, defluoridation, denitrification, calcium-rich diet, and rainwater harvesting. Subba Rao et al. (2022) investigates groundwater pollution causes in a rural region of Wanaparthy district, Telangana, India, comparing pre-monsoon and post-monsoon seasons. Results reveal elevated values of various chemical elements after the monsoon. The Overall Water Quality Index indicates an increase in moderate (53.86%) and very low (20.18%) groundwater quality zones post-monsoon, compared to pre-monsoon (35.22% and 4.77%, respectively). Piper's diagram shows a transition from freshwater to mixed water type (70%) predominating post-monsoon. Principal component analysis identifies human-induced contamination in post-monsoon groundwater, supporting the need for regular monitoring of index-wells to manage pollution and protect the aquifer system cost-effectively.

The ability of machine learning algorithms to simulate intricate interactions between variables has led to their success in the environmental field. Machine learning algorithms have the ability to decrease computation period, expenses, and inaccuracies in the categorization of water quality, the forecasting of parameters related to water quality, and the forecasting of water quality indices. Machine learning methods have been extensively employed in recent years for the evaluation of river water quality, involving the calculation of WQI (Nearing et al., 2021). In studies of water resource management, these methods have shown to be efficient modelling tools for complex non-linear phenomena (Shamsuddin et al., 2022). The ability of gradient boosting has been described with regard to the modelling and forecasting of water quality. Extreme gradient boosting (XGB) is used by Bedi et al. (2020) to test how well it can forecast contamination levels from sparse data with relationships that are not linear. According to Naloufi et al. (2021), the random forest model performed the best when comparing the performance of six models to predict the microbial quality of surface waters. Water quality indices have also been predicted and sorted using deep learning algorithms. Machine learning algorithms to estimate the water quality index for the La Buong River in Vietnam was used by Khoi et al. (2022). The findings of this study demonstrate that all models performed well in forecasting the water quality index, albeit the XGBoost model had the best accuracy. The potential of machine learning models to anticipate the water quality index is evaluated.

Uddin et al. (2022a), aimed to enhance coastal water quality assessment by developing an improved Water Quality Index (WQI) model. Employing the XGBoost algorithm, the study objectively ranked water quality indicators, determined sub-index weightings using the rank order centroid method, and tested various aggregation functions. Key findings include XGBoost's effectiveness in ranking indicators, the sensitivity of XGBoost rankings to the desired output, and the usefulness of the rank order centroid weighting method. The study recommends a weighted quadratic mean or unweighted arithmetic mean for aggregation functions. The improved WQI model, applied to Cork Harbour, Ireland, selects indicators based on importance to water quality, employs objective ranking and weighting methods, and suggests optimal aggregation functions. In Uddin et al.'s (2023b) study, the aims were to evaluate WQI model performance, correct classification using machine learning, and introduce a new coastal water quality assessment scheme. The study's goals were achieved by archiving WQI scores for coastal water quality, employing four predictive classifier models, identifying the best model, and evaluating performance using machine learning metrics. Utilizing support vector machines, Naïve Bayes, random forest, k-nearest neighbor, and gradient boosting, the study found that KNN (100% correct) and XGBoost (99.9% correct) outperformed in accurately predicting water quality classes for seven WQI models. XGBoost was identified as the superior classifier based on model validation results. Georgescu et al. (2023), aims to forecast Water Quality Index (WQI) time series data using Cascade-forward network (CFN) models, with Radial Basis Function Network (RBF) as a benchmark. Using 19 initial water quality features, CFN models, refined by Random Forest (RF) algorithm, outperform RBF models. CFNs provide accurate short-term forecasting for the first and fourth quarters, demonstrating their computational ability to predict water quality status. Research contributions include demonstrating CFN's efficacy in time-series prediction, optimizing model architecture, refining input data using RF, determining relevant thresholds for water quality indicators, and proposing a direct prediction model as an alternative to WQI calculation methods. The proposed approach offers advantages in model size and prediction accuracy compared to existing literature. Uddin et al. (2023a), addresses uncertainties in various stages of WQI application and presents a robust methodology. Evaluating eight WQI models, the Monte Carlo simulation (MCS) and Gaussian Process Regression (GPR) techniques estimate and predict model uncertainties. Sub-index functions contribute significantly to uncertainty, emphasizing the need for careful selection. Water quality indicator selection and weighting processes show low uncertainties. Statistical differences among aggregation functions are notable, with the weighted quadratic mean (WQM) function identified as providing a plausible and reduced uncertainty assessment of coastal water quality. The study suggests the potential use of the unweighted root mean squared (RMS) aggregation function for coastal water quality assessment. These findings have implications for decision-makers, researchers, and agencies involved in water quality monitoring and management.

Albania is a country full of precious water resources. The water surface of Albania is a natural asset presented by a wide network of rivers and lakes, and other sources of groundwater. There are over 150 streams and rivers in Albania that flow from east to west. Urban areas, agriculture, aquaculture, recreation, electricity, and industry all need river water. Major rivers' higher reaches flow across steep terrain, which has a considerable impact on both alluvium deposits and erosion in the western flatlands and highland areas to the east. The water quality index (WQI) of the Canadian Council of Ministries of the Environment (CCME), which was used for evaluating water quality, was employed in Damo and Icka's (2013) study on the portability of the water in Pogradec, Albania. Sulce et al. (2018) provide a basic review of the problems with the quality of Albania's surface and ground water and talk about the origins and controlling mechanisms. With the help of principal component analysis and cluster dendogram, Zela et al. (2020) assess the environmental quality of the Seman River water in Southern Albania. This work was completed utilising a five-year monitoring scheme that included 14 factors to assess the waterbody's environmental condition.

Shkumbini River is one of the main rivers of Albania, to which various initiatives have been taken to prevent its pollution. The goal of this study was to assess the efficiency of models developed using machine learning in order to ascertain how the Shkumbini River's WQI categorization is affected by water quality indicators. The following steps were undertaken to meet the study's objectives:

- Calculating the Shkumbini River's WQI scores based on Brown et al.'s (1972) methodology.
- Identifying water quality classes through a classification scheme.
- Standardizing and splitting data variables into training and testing sets.
- Employing four machine learning classifier models on the acquired dataset.
- Selecting the best model by considering factors such as accuracy, precision, sensitivity, mean square error, and RSquare.
- Determining important indicators and their relative rankings in this stage.
- The efficacy of the models was evaluated for each model and water quality class using the confusion matrix.
- Implementing the best predictive model to forecast water quality class.

This comprehensive approach allowed for a thorough examination of the relationship between water quality indicators and the Shkumbini River's WQI categorization, ultimately leading to the identification of the most effective predictive model. The five main sections of the paper are as follows. A quick summary of the study is given in the first part. The second part included a variety of tools and methods used to evaluate the performance of models. Section three, where the results are provided, also includes information on how to select the best model by using model performance metrics. The conclusions and consequences of the research's findings are presented in section four.

2. Materials and methods

In southeast Albania, the Valamara mountain range is where the Shkumbini River rises. From its source to its delta in the Adriatic Sea, this significant watercourse is 181 kilometres long. In Central Albania, the Shkumbin River drains an area of 2,444 square kilometres in an east-west direction. The river discharges 61.5 cubic metres per second on average. Only 40% of the river's yearly flow comes from subsurface sources, making up the other 60%. Therefore, erosion's impacts have a significant impact on river pollution. Conduct evaluations of the Shkumbini River's biological, chemical, and physical characteristics in Albania, and educate the local administration and populace about the significance of enhancing the river ecosystems. The river Shkumbini had environmental issues. It should be noted the significant contribution to river pollution made by sediments originating in ultramafic zones and remnants of the Elbasan Metallurgical Combine (Roba et al., 2016).

2.1. Data collection

The water quality monitoring data from the years 2018, 2019, 2020, and 2021 were used for the purposes of this study. Based on the parameters of data accessibility and covering of the entire extents of the Shkumbini River, six monitoring sites (Qukës, Librazhd, Xibrak, Papr, Bishqem, and Rrogozhin) were taken into consideration for this research (Figure 1). Also the data have been collected in three different periods of the year: March, July and October.

To prevent unexpected changes in the water's qualities, plastic holders that had been acid-washed were employed to gather data at each monitoring site. Containers with volumes of 2L, 1.5L, and 0.5L were utilised for the examination of the nine factors that affect the quality of water. The sampling techniques were chosen in accordance with ISO 5667-4 and ISO 5667-6. The samples were submitted to the Regional Directorate of Public Health's laboratory in Elbasan, Albania, for examination.

The parameters for the conventional WQI model are not specified in a systematic manner. The selection of the WQI model's parameters appears to have taken into account a few common water quality concerns, including oxygen availability, eutrophication, health factors, physical as well as chemical phenomena, and dissolved components (Verma et al., 2019).

The Albanian Standards for drinking water which fit with those of EU, applied the calculation of the WQI based on nine parameters: pH, Total Dissolved Solids (TDS), General Hardness (GH), Biochemical oxygen demand (BOD), Dissolved Oxygen (DO), Chloride (Cl), Total Phosphor (TP), Thermotolerant Coliforms, Bicarbonate (HCO_3). According to the relative significance that each parameter has in determining the quality of drinking water, the nine parameters are each assigned a weight (mg/l). According to the relative significance that each parameter has in determining the quality of drinking water, the nine parameters are each assigned a weight (mg/l), where: TDS 6.5 mg/l; GH 20 mg/l; BOD 1.5 mg/l; pH 7.5 mg/l; DO 5.8 mg/l; Cl 20 mg/l; HCO_3 200 mg/l; TP 0.09 mg/l; and Thermotolerant Coliforms 600 mg/l. Our data table consists of nine physico-chemical water quality indicators, water quality index value and water quality classification.

2.2. Water quality index calculation

The WQI model is one of several tools and methods that are used to evaluate the water quality with the goal to handle water resources. This method is frequently employed to evaluate the quality of water, including groundwater, surface water, etc. Even though WQI methods have only been in existence for the past 50 years, water quality indices have been used to classify water quality since the mid-1800s (Abbasi & Abassi, 2012). The WQI values were computed using a variety of WQI models. Due to its straightforward mathematical operations and user-friendliness, its applicability has gradually risen. Horton developed the initial WQI model in the 1960s, basing it on 10 water quality indicators that were deemed crucial in most lakes and rivers (Horton, 1965). The component collection and weighting for Brown's NSF-WQI, a more exacting variant of Horton's WQI model developed with support from the National Sanitation Foundation, was reviewed by a group of 142 water quality specialists. Later, Steinhart et al. (1982) developed the Environmental Quality Index system to assess the water quality in the Great Lakes environments. To date, various countries and/or organisations throughout the world have deployed over thirty-five WQI methods to evaluate the condition of surface waters (Dadolahi-Sohrab et al., 2012).

In the realm of water resource management, diverse techniques and tools are employed to evaluate water quality, with the Water Quality Index (WQI) model standing out as a prominent method. This model facilitates the

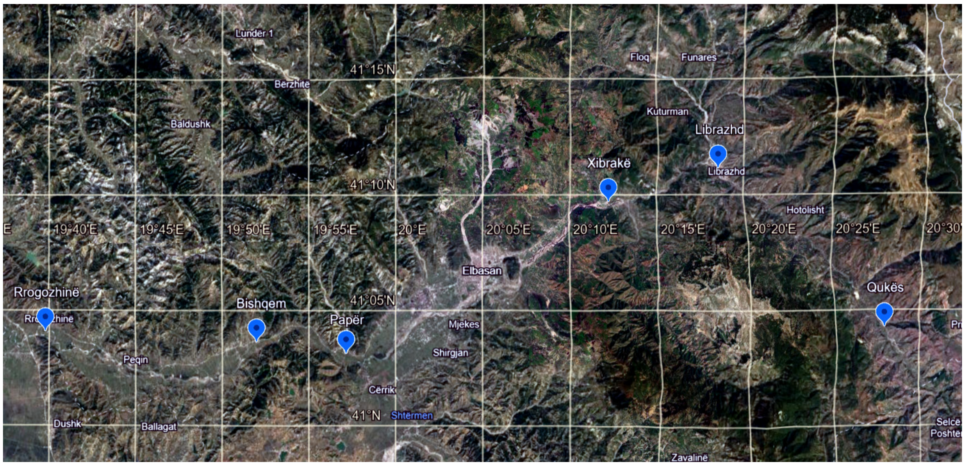


Figure 1. The points of data collection map

conversion of extensive water quality data into a singular numerical value known as the index score (Parween et al., 2022; Uddin et al., 2021, 2022a). Its growing popularity can be attributed to its user-friendly nature and straightforward mathematical operations, distinguishing it from more complex hydrological tools (Uddin et al., 2021, 2022a). Given the significance of this issue and considering that this study marks the inaugural exploration of the Skumbini River, the WQI model was chosen as the initial framework. This decision was based on insightful discussions with experts in the field, emphasizing the need for a robust starting point in this unique context.

The WQI was computed using the World Health Organization's (2017) suggested water for consumption quality requirement. The Water Quality Index has been calculated using the weighted numerical calculation method, which was first put forth by Horton (1965) and modified by Brown et al. (1972). The following is how the weighted arithmetic water quality index (WQI) is displayed:

$$WQI = \sum_{i=1}^{n} W_i Q_i \,/ \sum_{i=1}^{n} W_i, \tag{1}$$

where n is the quantity of variables or factors, the parameter's weight in units is W_i, and the parameter's score for quality (sub-index) is known as Q_i is the parameter's quality rating (sub-index). The recommended requirements for the associated metrics are inversely correlated with W_i of the several water quality tests.

$$W_i = K / S_n, \tag{2}$$

where K is a fixed percentage and S_n is the expected value for the parameter i. The numerical value of the quality rating is determined using the following formula, based on Brown et al. (1972):

$$Q_i = 100\left[(A_0 - A_i) / \left(S_n - A_i\right)\right], \tag{3}$$

where A_i is the i-th parameter's desired level in the case of pure water, and A_0 is the i-th parameter's actual value as measured at the designated sampling location.

In accordance with Brown et al. (1972), the following categories apply to the water quality index (WQI): WQI value between 0–25 belongs in the rating class "Excellent"; WQI value between 26–50 belongs in the rating class "Good"; WQI value between 51–75 belongs in the rating class "Poor"; WQI value between 76–100 belongs in the rating class "Very Poor"; WQI value >100 belongs in the rating class "Unsuitable".

2.3. Data pre-processing

Standardising data variables is crucial before machine learning algorithms are trained. This strategy is frequently used in machine learning (ML) to uniformly scale all data variables in order to reduce model training mistakes (Rahman, 2020). The z score normalisation procedure was used in this study to standardise the variables used to measure water quality. This is one way to portray a z score:

$$z = \frac{x_i - \overline{x}}{\sigma}, \tag{4}$$

where x_i is the i-th sample factor, z is the standardise score, $\overline{x}$ is the average of the data variable, and σ denotes the standard deviation of the data.

Data was split into training (80%) and testing (20%) sets prior to training the machine learning algorithms. After the data had been divided into training and testing sets, four machine learning algorithms were trained and evaluated. Throughout both stages, the models' efficacy was evaluated.

2.4. Machine learning algorithms

Recently, machine learning techniques have been used across a wide range of academic fields. For instance, studies on the prediction of water quality have shown that the ML algorithm may be superior to other conventional methods in this regard (Uddin et al., 2022). For the WQC forecasting, Aldhyani et al. (2020), employ K-Nearest Neighbour and Naive Bayes algorithm. Based on four water characteristics and the advantages of machine learning techniques, Azrour et al. (2021) create a model that can predict the water quality category and then the water quality index. This study used four machine learning techniques to find effective formulas for forecasting the WQI index in the Shkumbini river. The models that were employed in this investigation are briefly described in the section below.

(a) XGBoost algorithm

The most common type of ensemble learning, boosting, creates a powerful algorithm by merging numerous weak learners (Ferreira & Figueiredo, 2012). The advantage of boosting lies in the serial structure of its learning, which produces great approximation and generalisation. Numerous boosting strategies have already been put forth. By changing a few of the steps in the overall boosting scheme, each one enhances classification performance. GradientBoost was first launched in 2016 by Chen, and XGBoost is an enhanced version of it. By implementing certain efficient methods to control split discovery, manage inaccurate information, and handling overfitting in the learning stage, XGBoost has enhanced the conventional GradientBoost (Chen & Guestrin, 2016). The goal (minimization) function, regulates the model's complexity to avoid overfitting and consists of two parts: a normalisation component and a loss rate:

$$obj = \sum_{i=1}^{n} l\left(y_i, f_i\right) + \sum_{m=1}^{M} \Omega\left(f^m\right), \tag{5}$$

where $\Omega\left(f^m\right)$ is the regularization term. In XGBoost, the objective function is optimised using a second-order approximation. As a result, in order to select the optimal tree for each iteration, Eq. (6) is applied:

$$obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T, \tag{6}$$

where T is the number of leaf nodes. The j-th leaf node data' first and second order gradient metrics on the loss function, correspondingly, are added up to form G_j and H_j. The regularisation coefficients are λ and γ. As a result, under this method, each iteration of the tree's complexity is determined and managed separately, and the number of leaves does not remain constant (Nayan et al., 2020).

(b) Random forest (RF)

Decision-trees are used in combination to create Random Forests (RF). It enhances the accuracy in classification of just one tree classifier by incorporating the bootstrap aggregating strategy and randomization in choosing of data nodes during the construction of a decision tree. The feature space is divided into K regions using a decision tree with K leaves, where $R_k, 1 \le k \le K$ (Sain, 1996; El Bilali et al., 2021). The forecasting formula is specified as follows for each tree:

$$\sum_{k=1}^{K} c_k \Pi\left(x, R_k\right), \tag{7}$$

where R_k is a region suited to k, and c_k is a constant appropriate to k, where K denotes how many zones there are in the feature area:

$$\Pi\left(x, R_k\right) = \begin{cases} 1, \text{ if } x \in R_k \\ 0, \text{ otherwise} \end{cases}. \tag{8}$$

The final classification judgement is based on the majority vote or average of all trees.

(c) K-Nearest Neighbour (KNN)

Applying prior learning methods, the KNN classifier can be employed to quickly evaluate unknown sample class data because it is a simpler and dependable technique. Without any prior knowledge of data distribution, it may be quickly incorporated into any framework for machine learning. The KNN classifier first calculated the distance between each sample point, then it produced novel groups according to the closest sample groups. By considering the nearest neighbours of the most samples, the categorization of the new sample collection is determined (Altman, 1992; Cunningham & Delany, 2007). The K-Nearest Neighbour technique assumes that members of each class are mostly present around each example of that class. As an outcome, it is given a scalar k and a set for learning samples in the area of features. The Euclidean distance is the most often employed measure for determining the separation between instances among the numerous others. This approach uses the Euclidean distance, which is characterised by the equation below:

$$L\left(x_i, x_j\right) = \left(\sum_{i,j=1}^{n} \left(\left|x_i - x_j\right|\right)^2\right)^{\frac{1}{2}}, \; X \in R^n. \tag{9}$$

(d) Naïve Bayes (NB)

The Bayesian technique forecasts and classifies datasets using probability statistics. The Bayesian technique, which integrates pre and posterior probability, avoids both the supervisor's bias and the overfitting issue related to depending only on sample data. The autonomous nature of feature requirements and the Bayes theorem serve as the foundation for this Naive Bayes classification algorithm. Attributes are assumed to be conditionally independent of each other when the goal value is given. The Bayesian method's complexity is significantly reduced by this approach. The goal of Naïve Bayes Classifier is to calculate conditional probability:

$$p(C_k \mid x_1, x_2, \ldots, x_n). \tag{10}$$

Given that the category C_k is present, the "aïve" conditional independence rules are put into action. These hypotheses posit that every attribute in x are mutually independent (Shafi et al., 2018). The discussion up to this moment led to the independent feature approach, also known as the naïve Bayes probability model. The aforementioned model and a decision rule are combined by the Naive Bayes algorithm. The analogous classification algorithm, a Bayes classifier, is the function that determines the next category label:

$$\hat{y} = \underset{k \in \{1,..,K\}}{\operatorname{argmax}} \, p\left(C_k\right) \prod_{i=1}^{n} p(x_i \mid C_k). \tag{11}$$

3. Results

3.1. A statistical summary of WQI models input (indicators)

The Z statistics for pH, Total Dissolved Solids (TDS), General Hardness (GH), Biochemical oxygen demand (BOD), Dissolved Oxygen (DO), Chloride (Cl), Total Phosphor, Thermotolerant Coliforms, Bicarbonate (HCO_3) and the WQI, are displayed in Figure 2. A nice graphical representation of the data density may be seen in Whisker's box-plot, which is made up of the minimum value, the first and third quartile, the median, and the maximum value. The Pearson's correlation test was used to examine the significant relationships between the water quality indicators at a 99% confidence level in order to determine their correlation. Figure 3 displays the findings of the association between the indicators.

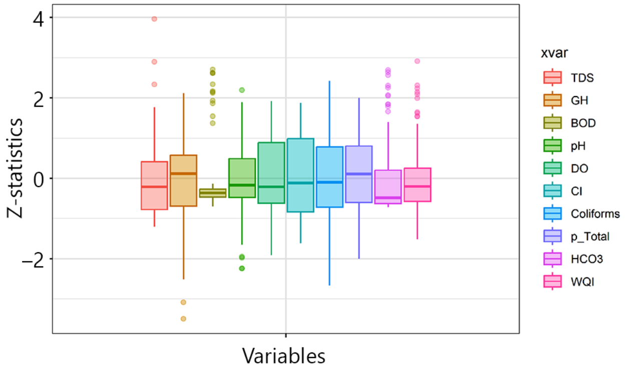


Figure 2. Z statistics of water quality indicators and WQI

With a mean and median value of 10.32 mg/l and 8.7 mg/l, respectively, over the course of the research time frame, the highest and lowest TDS values were determined to be 40.0 mg/l and 1.3 mg/l, respectively, indicating a positive skew (mean > median) in the data set (Figure 2). With a mean value of 4.457 mg/l, a median value of 7.75 mg/l, and a strong moderately favourable correlation with the WQI (r = 0.85), BOD likewise had a positive skew (Figure 3). At Uddin et al. (2022). The data for BOD showed a higher median value (1.86 mg/l) than the mean value (1.74 mg/l), and it varied between 0.00 mg/l and 3.55 mg/l, at Uddin et al. (2022). The GH had a skew that was negative and a variability of 10.5 mg/l to 28.52 mg/l, with a mean value of 21.73 mg/l. The DO, however, was 10.59 mg/l at all of the monitoring sites. The pH measurements in the dataset had a positive skew and fell between 6.8 mg/l to 8.3 mg/l. At Uddin et al. (2022) WQI showed a significant moderate positive relationship with water pH (r = 0.60, p < 0.01).

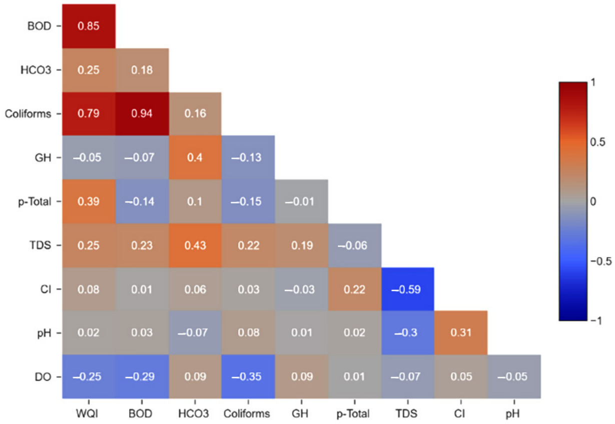


Figure 3. Pearson's correlation of indicators

Thermotolerant coliforms varied considerably between the observation areas in this study. With a range of 130 to 1450 and a mean value of 409.65 mg/l (Figure 2), it was irregular. BOD (r = 0.94) and WQI (r = 0.79) exhibited a statistically significant, strong positive correlation with Thermotolerant coliforms (Figure 3). The range for Cl was between 10.6 mg/l and 26.4 mg/l, with a median value of 17.96 mg/l and a mean value of 17.4 mg/l (Figure 2). With mean and medial values of 221 mg/l and 218.6 mg/l, correspondingly, HCO_3 likewise displayed positive skewness throughout the monitoring locations. With a mean of 0.062 mg/l, Total Phosphor had a range of 0.042 mg/l to 0.082 mg/l. TDS and Cl were shown to have a poor connection in this study (r = –0.59) (Figure 3). Through the course of the investigation, there were outliers in the data for TDS, GH, BOD, pH, and HCO_3.

3.2. Comparative evaluation of several models

The WQI values of the Shkumbini River were estimated in this work applying four machine learning algorithms. Cross-validation techniques were used to verify the prediction outcomes of different ML algorithms. Cross-validation (CV) stands out as one of the most common procedures for evaluating machine learning models, particularly in the case of small datasets. In the current study, the random CV technique was employed to assess the performance of the machine learning predictive model. Specifically, a 10-fold CV technique was utilized, incorporating four widely used evaluation criteria: mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R2). This approach allowed for a comprehensive comparison of the model's performance, providing valuable insights into its predictive capabilities.

Table 1. Hyper-parameters of various ML models

Model parameters	XGBoost	Random Forest	K-Nearest Neighbour	Naive Bayes
n_estimators	100	100	–	300
learning_rate	0.4	–	–	–
max.depth	20	20	–	–
gamma	0	–	–	–
booster	gbtree	–	–	–
subsample	1	–	–	–
bootstrap	True	True	–	–
Objective	reg.linear	–	–	–
criterion	–	Squared_error	–	–
max_leaf_nodes	–	4	30	–
min_samples_leaf	–	1		–
n_neighbors	–	–	5	–
weight	–	–	gaussian	–
metrics	–	–	minkowski	–
power_parameters	–	–	3	–

The significance of the learning rate, along with other parameters, cannot be overstated in the development of prediction/forecasting models. It plays a pivotal role in determining the model's convergence speed, precision, and accuracy. The learning rate governs how swiftly the model reaches its optimal solution and influences the delicate balance between accuracy and computational efficiency. A well-chosen learning rate is instrumental in preventing oscillations, steering clear of local minima, and shaping the model's sensitivity to initial parameters. The meticulous tuning of this hyperparameter is crucial, given its direct impact on the efficiency, accuracy, and stability of the prediction model. Consequently, it stands as a fundamental factor in the entire model development process. The enhancement of model accuracy in machine learning often involves the tuning of hyperparameters. Various methods are employed in ML approaches to fine-tune predictive models, with grid search and random search techniques being commonly utilized in the literature. In alignment with this trend, the current research employs the grid search technique to optimize the model parameters during the training phase. Table 1 provides an overview of

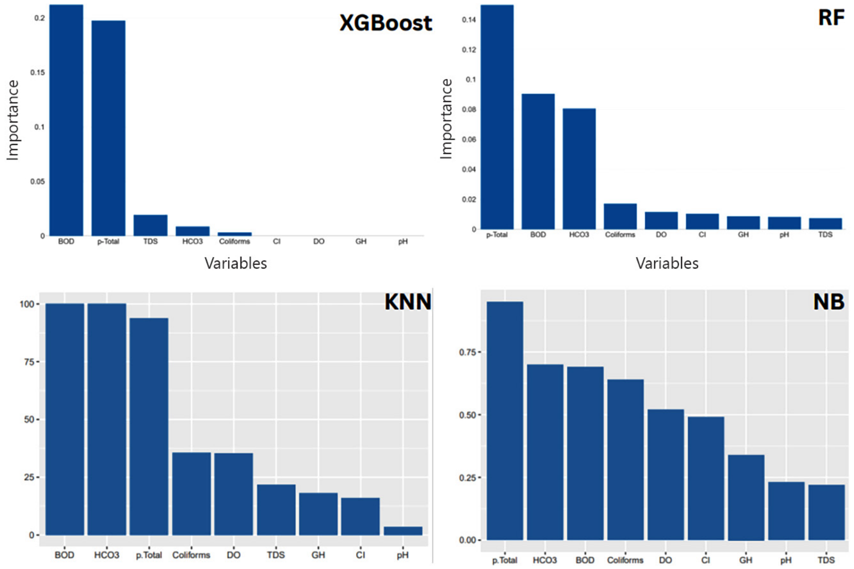


Figure 4. Variable importance according all models

the hyperparameters associated with different ML models throughout the model training process.

The current study additionally used the calculation of the coefficient (Rsquare) to assess the model performance in order to identify the best methods. The R2 frequently refers to the association as well as consistency between predictors and response factors in order to choose the best algorithm. According to the results of the Tukey's HSD test, there were no statistically significant differences between any algorithm's expected and real WQI scores in terms of bias among true and forecasted WQI.

This study showed that ensemble tree-based (XGBoost and Random Forest) approaches outperform other algorithms based on model errors in terms of stability and reliability. In terms of lowering the uncertainty associated with the WQI model, these models may be reliable and effective at forecasting WQI. On the other hand, determining whether algorithm is "better" or "worst" is difficult.

Given the cross-validation data, the XGBoost and Random Forest models had the best prediction accuracy of the approaches. The Naive Bayes and K-Nearest Neighbour algorithms have the worst prediction errors. Training (RMSE = 2.1, and MAE = 1.13) and testing (RMSE = 0.1 and MAE = 0.12) have the lowest error rates for the XGBoost model. Greater prediction errors were seen for the K-Nearest Neighbour in comparison to the best techniques (RMSE = 4.03 and MAE = 3.2) during the model training stage and (RMSE = 2.03 and MAE = 2.5) during the model validation phase, Table 2. While it would be inaccurate to categorize the performance of the other models as poor, the obtained results indicate that these models exhibit similar levels of accuracy. However, it is noteworthy that among these models, there are two specific models that outperform the others in terms of performance.

Table 2. Model performance

Model	Training		Testing	
	MAE	RMSE	MAE	RMSE
XGBoost	1.13	2.1	0.12	0.1
Random Forest	1.7	2.33	0.34	0.27
K-Nearest Neighbour	3.2	4.03	2.5	2.03
Naive Bayes	2.84	3.01	1.99	2.0

The current study used the aforementioned method to undertake a relative relevance analysis to ascertain the impact of water quality indicators on classification. In this study, we discovered that BOD, HCO_3, and Total Phosphor had the greatest positive effects on the water quality of the Shkumbini River, according to all the models. The important indicators and their relative rankings are shown in Figure 4.

3.3. Confusion matrix outcomes

The current study examines the performance of the four machine learning classifiers in order to identify the most effective techniques for precise classification. The performance of the classifiers on the unbalanced dataset was evaluated using validation measures, including accuracy, sensitivity, and F1 score. One of these measurements was confusion matrices. The confusion matrix for the four models is displayed in Figure 5. Observations from three classifications – "poor", "very poor," and "unsuitable" – were utilised in this research to predict the classification.

(a) XGBoost model

As seen in Figure 5a, 54.16% of "very poor" water quality is accurately assessed, whereas 1.39% of poor is wrongly classified as "very poor". The "unsuitable" water quality class, on the other hand, is accurately identified at 16.67%.

For each class of water quality, an average of 98.61% of the observations are properly categorised. According to the results of the confusion matrices, the XGBoost outperformed the confusion matrices of the other four classifiers in predicting the proper categorization of water quality.

(b) RF model

The RF results demonstrate that "very poor" water quality is wrongly classified as "poor," 2.78 percent, and that "poor" water quality is incorrectly rated as "very poor," 2.78 percent. Figure 5b shows that an average of 94.44% of observations are correctly identified across all water quality classes.

(c) KNN model

According to the KNN statistics, "very poor" water quality is wrongly classed as "poor," making up 5.56% of the total, and "poor" water quality is incorrectly labelled as "very poor," making up 2.77%. Figure 5c shows that only 91.22% of the data were correctly categorised into the poor class, with the remaining observations receiving incorrect classification.

(d) NB model

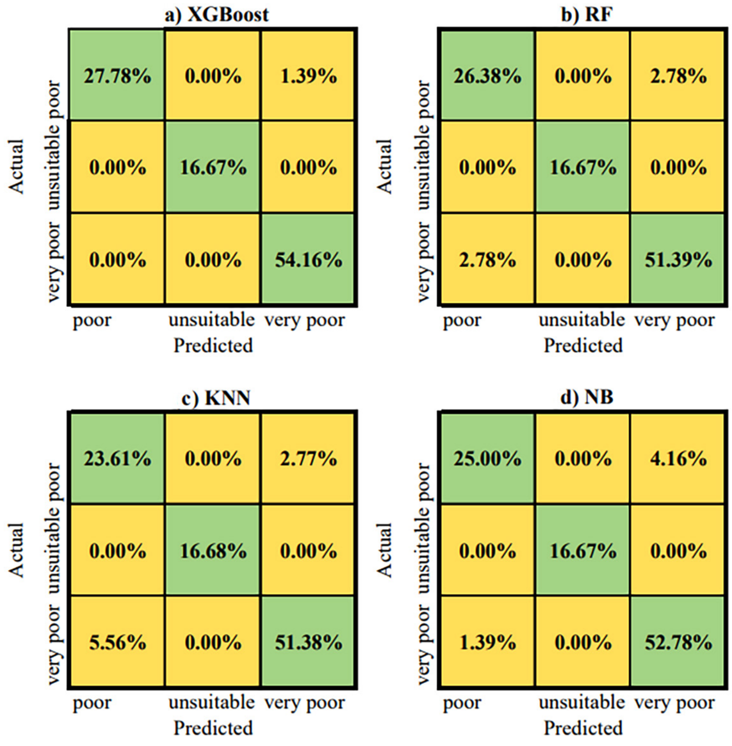
In the Shkumbini River, the NB has accurately assigned 94.45% of the water quality classes. According to the KNN results, "poor" water quality is wrongly identified as "very poor," 1.39%, and "poor" water quality is incorrectly labelled as "very poor," 4.16% (Figure 5d).

The current study examined the used models utilising their accuracy, sensitivity and F1 scores for the evaluation of classifier performance. In this study, the XGBoost, RF, KNN, and NB models' predicted accuracy was determined to be 98.61%, 94.44%, 91.22%, and 94.45%, respectively. The XGBoost algorithm was shown to have great performance when compared to models. The XGBoost model had the highest precision, sensitivity, and F1 scores, while the KNN model scored poorly in these terms. The results of the performance metrics demonstrate the accuracy with which the XGBoost algorithm can forecast the categorization of water quality.

4. Conclusions

Given the Shkumbini River's significance as one of Albania's largest rivers, extensively utilized by Central Albanian residents for drinking water post-treatment, numerous studies have focused on determining the Water Quality Index (WQI) as a crucial measure of water quality. This work addresses the challenge of model unpredictability and aims to identify the optimal machine learning method for forecasting the WQI. Four machine learning models were rigorously evaluated and validated for predicting the Shkumbini River's Water Quality Index, utilizing various validators such as RMSE, MSE, MAE, accuracy, precision, and sensitivity.

The study's results highlight the exceptional performance of XGB, with the lowest errors observed during both the learning (RMSE = 2.1, MSE = 9.8, MAE = 1.13) and evaluating (RMSE = 0.0, MSE = 0.01, MAE = 0.01) stages. Furthermore, the research delves into the influence of nine physico-chemical water quality indicators on the WQI, revealing that Biochemical Oxygen Demand (BOD), Bicarbonate (HCO_3), and Total Phosphor have the most positive impact on the Shkumbini River's water quality.



a) XGBoost

	poor	unsuitable	very poor
poor	27.78%	0.00%	1.39%
unsuitable	0.00%	16.67%	0.00%
very poor	0.00%	0.00%	54.16%

b) RF

	poor	unsuitable	very poor
poor	26.38%	0.00%	2.78%
unsuitable	0.00%	16.67%	0.00%
very poor	2.78%	0.00%	51.39%

c) KNN

	poor	unsuitable	very poor
poor	23.61%	0.00%	2.77%
unsuitable	0.00%	16.68%	0.00%
very poor	5.56%	0.00%	51.38%

d) NB

	poor	unsuitable	very poor
poor	25.00%	0.00%	4.16%
unsuitable	0.00%	16.67%	0.00%
very poor	1.39%	0.00%	52.78%

Figure 5. Results of confusion matrices

While acknowledging the need for additional research using more indicators and diverse predictive classifier algorithms, the study's findings contribute to the proper categorization of water quality. Despite its limitations, particularly in terms of evaluating WQI performance, if only there were more observation points, parameters and more measurements, maybe if they were monthly, the results offer valuable insights for mitigating model uncertainty and providing useful information for researchers and policymakers.

References

Abbasi, T., & Abbasi, S. A. (2012). Water-quality indices: Looking back, looking ahead. In *Water quality indices* (pp. 353–356). Elsevier. https://doi.org/10.1016/B978-0-444-54304-2.00016-6

Aldhyani, T. H. H., Al-Yaari, M., Alkahtani H., & Maashi, M. (2020). Retraction: Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics*, *2020*, Article 6659314. https://doi.org/10.1155/2020/6659314

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–85. https://doi.org/10.1080/00031305.1992.10475879

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz A., & Aziz, F. (2021). Machine learning algorithms for efficient water quality prediction. *Modeleling Earth Systems and Environment*, *8*, 2793–2801. https://doi.org/10.1007/s40808-021-01266-6

Bedi, S., Samal, A., Ray, C., & Snow, D. (2020). Comparative evaluation of machine learning models for groundwater quality assessment. *Environmental Monitoring and Assessment*, *192*, Article 776. https://doi.org/10.1007/s10661-020-08695-3

Brown, R. M., Mccleiland, N. J., Deiniger R. A., & O'Connor, M. F. (1972, June 18–23). Water quality index-crossing the physical barrier. In *Proceedings of the International Conference on Water Pollution Research* (pp. 787–797), Jerusalem. https://doi.org/10.1016/B978-0-08-017005-3.50067-0

Chen, T., & Guestrin, C. (2016, August 13–17). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794), San Francisco. https://doi.org/10.1145/2939672.2939785

Cunningham, P., & Delany, S. J. (2007). k-Nearest neighbour classifiers. *ACM Computing Surveys*, *54*(6), 1–25. https://doi.org/10.1145/3459665

Dadolahi-Sohrab, A., Arjomand, F., & Fadaei-Nasab, M. (2012). Water quality index as a simple indicator of watersheds pollution in southwestern part of Iran. *Water and Environment Journal*, *26*(4), 445–454. https://doi.org/10.1111/j.1747-6593.2011.00303.x

Damo, R., & Icka, P. (2013). Evaluation of water quality index for drinking water. *Polish Journal of Environmental Studies*, *22*(4), 1045–1051.

El Bilali, A., Taleb, A., & Brouziyne, Y. (2021). Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricultural Water Management*, *245*, Article 106625. https://doi.org/10.1016/j.agwat.2020.106625

Ferreira, A. J., & Figueiredo, M. A. (2012). Boosting algorithms: A review of methods, theory, and applications. In *Ensemble machine learning* (pp. 35–85). Springer. https://doi.org/10.1007/978-1-4419-9326-7_2

Georgescu, P.-L., Moldovanu, S., Iticescu, C., Calmuc, M., Calmuc, V., Topa, C., & Moraru, L. (2023). Assessing and forecasting water quality in the Danube River by using neural network approaches. *The Science of the Total Environment*, *879*, Article 162998. https://doi.org/10.1016/j.scitotenv.2023.162998

Horton, R. K. (1965). An index number system for rating water quality. *Journal of the Water Pollution Control Federation*, *37*(3), 303–306.

International Organization for Standardization. (2018). *Water quality – Sampling – Part 4: Guidance on sampling from lakes, natural and man-made* (ISO Standard No. 5667-4). https://standards.iteh.ai/catalog/standards/sist/a1a7bb26-7c03-462f-a7ae-7619d48945e2/sist-iso-5667-4-2018

International Organization for Standardization. (2015). *Water quality – Sampling – Part 6: Guidance on sampling of rivers and streams* (ISO 5667-6). https://standards.iteh.ai/catalog/standards/sist/b8b8c606-00fc-46fb-a38f-109c197cc3b9/sist-iso-5667-6-2015

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, Vietnam. *Water*, *14*(10), Article 1552. https://doi.org/10.3390/w14101552

Naloufi, M., Lucas F. S., Souihi, S., Servais, P., Janne, A., & Wanderley Matos De Abreu, T. (2021). Evaluating the performance of machine learning approaches to predict the microbial quality of surface waters and to optimize the sampling effort. *Water*, *13*(18), Article 2457. https://doi.org/10.3390/w13182457

Nayan, A.-A., Kibria, M. G., Rahman, M. O., & Saha, J. (2020, November 28–29). River water quality analysis and prediction using GBM. In *Proceedings of the 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)* (pp. 219–224). IEEE. https://doi.org/10.1109/ICAICT51780.2020.9333492

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), Article e2020WR028091. https://doi.org/10.1029/2020WR028091

Parween, S., Siddique, N. A., Mahammad Diganta, M. T., Olbert, A. I., & Uddin, Md. G. (2022). Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. *Environmental and Sustainability Indicators*, *16,* Article 100202. https://doi.org/10.1016/j.indic.2022.100202

Rahman, A. (2020). *Statistics for data science and policy analysis.* Springer. https://doi.org/10.1007/978-981-15-1735-8

Ravindra, B., Subba Rao, N., & Dhanamjaya Rao, E. N. (2023). Groundwater quality monitoring for assessment of pollution levels and potability using WPI and WQI methods from a part of Guntur district, Andhra Pradesh, India. *Environment, Development and Sustainability*, *25*, 14785–14815. https://doi.org/10.1007/s10668-022-02689-6

Roba, C., Rosu, C., Pistea, I., Baciu, C., Costin, D., & Ozunu, A. (2016). Transfer of heavy metals from soil to vegetables in a mining/smelting influenced area (Baia Mare – Ferneziu, Romania). *Journal of Environmental Protection and Ecology*, *16*, 891–898.

Sain, S. R. (1996). The nature of statistical learning theory. *Technometrics*, *38*(4), 409. https://doi.org/10.2307/1271324

Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M., & Khurshid, H. (2018, October 8–10). Surface water pollution detection using internet of things. In *Proceedings 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)* (pp. 92–96). IEEE. https://doi.org/10.1109/HONET.2018.8551341

Shamsuddin, I. I. S., Othman, Z., & Sani, N. S. (2022). Water quality index classification based on machine learning: A case from the Langat River Basin model. *Water*, *14*(19), Article 2939. https://doi.org/10.3390/w14192939

Steinhart, C. E., Schierow, L. J., & Sonzogni, W. C. (1982). An environmental quality index for the great lakes. *Journal of the American Water Resources Association*, *18*(6), 1025–1031. https://doi.org/10.1111/j.1752-1688.1982.tb00110.x

Subba Rao, N., Sunitha, B., Das, R., & Anil Kumar, B. (2022). Monitoring the causes of pollution using groundwater quality and chemistry before and after the monsoon. *Physics and Chemistry of the Earth, 128*, Article 103228. https://doi.org/10.1016/j.pce.2022.103228

Sulce, S., Rroco, E., Malltezi, J., Shallari, S., Libohova, Z., Sinaj, S., & Qafoku, N. P. (2018). Water quality in Albania: An overview of sources of contamination and controlling factors. *Albanian Journal of Agricultural Sciences*, *2* (Special edition – Proceedings of ICOALS), 279–297.

Sutadian, A. D., Muttil, N., Yilmaz, A. G., & Perera, B. J. C. (2018). Development of a water quality index for rivers in West Java Province, Indonesia. *Ecological Indicators*, *85*, 966–982. https://doi.org/10.1016/j.ecolind.2017.11.049

Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators, 122*, Article 107218. https://doi.org/10.1016/j.ecolind.2020.107218

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2022a). A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. *Water Research*, *219*, Article 118532. https://doi.org/10.1016/j.watres.2022.118532

Uddin, M. G., Nash, S., Mahammad Diganta, M. T., Rahman, A., & Olbert, A. I. (2022b). Robust machine learning algorithms for predicting coastal water quality index. *Journal or Environmental Management, 321*, Article 115923. https://doi.org/10.1016/j.jenvman.2022.115923

Uddin, G., Nash, S., & Olbert, A. I. (2022c). *Optimization of parameters in a water quality index model using principal component analysis* [Conference presentation]. Proceedings of the 39th IAHR World Congress, Granada, Spain. https://doi.org/10.3850/IAHR-39WC2521711920221326

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023a). A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. *Water Research*, *229*, Article 119422. https://doi.org/10.1016/j.watres.2022.119422

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023b). Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Safety and Environmental Protection*, *169*, 808–828. https://doi.org/10.1016/j.psep.2022.11.073

Verma, R. K., Murthy, S., Tiwary, R. K., & Verma, S. (2019). Development of simplified WQIs for assessment of spatial and temporal variations of surface water quality in upper Damodar river basin, eastern India. *Applied Water Science*, *9*, Article 21. https://doi.org/10.1007/s13201-019-0893-0

World Health Organization. (2017). *Guideline for drinking water quality* (4th ed., incorporating the 1st addendum). https://www.who.int/publications/i/item/9789241549950

Zela, G., Demiraj, E., Marko, O., Gjipalaj, J., Erebara, A., Malltezi, J., Zela, E., & Bani, A. (2020). Assessment of the water quality index in the Semani River in Albania. *Journal of Environmental Protection*, *11*(11), 998–1013. https://doi.org/10.4236/jep.2020.1111063