Water Quality Research Journal

Water Quality Research Journal Vol 57 No 3, 152 doi: 10.2166/wqrj.2022.004

Machine learning for water quality classification

Saleh Y. Abuzir [a] and Yousef S. Abuzir [b,*]

[a] Civil and Environmental Engineering, Brescia University, Brescia, Italy
[b] Faculty of Technology and Applied Sciences, Al-Quds Open University, Ramallah, Palestine
*Corresponding author. E-mail: yabuzir@qou.edu

SYA, 0000-0002-9825-3473; YSA, 0000-0002-1220-1411

ABSTRACT

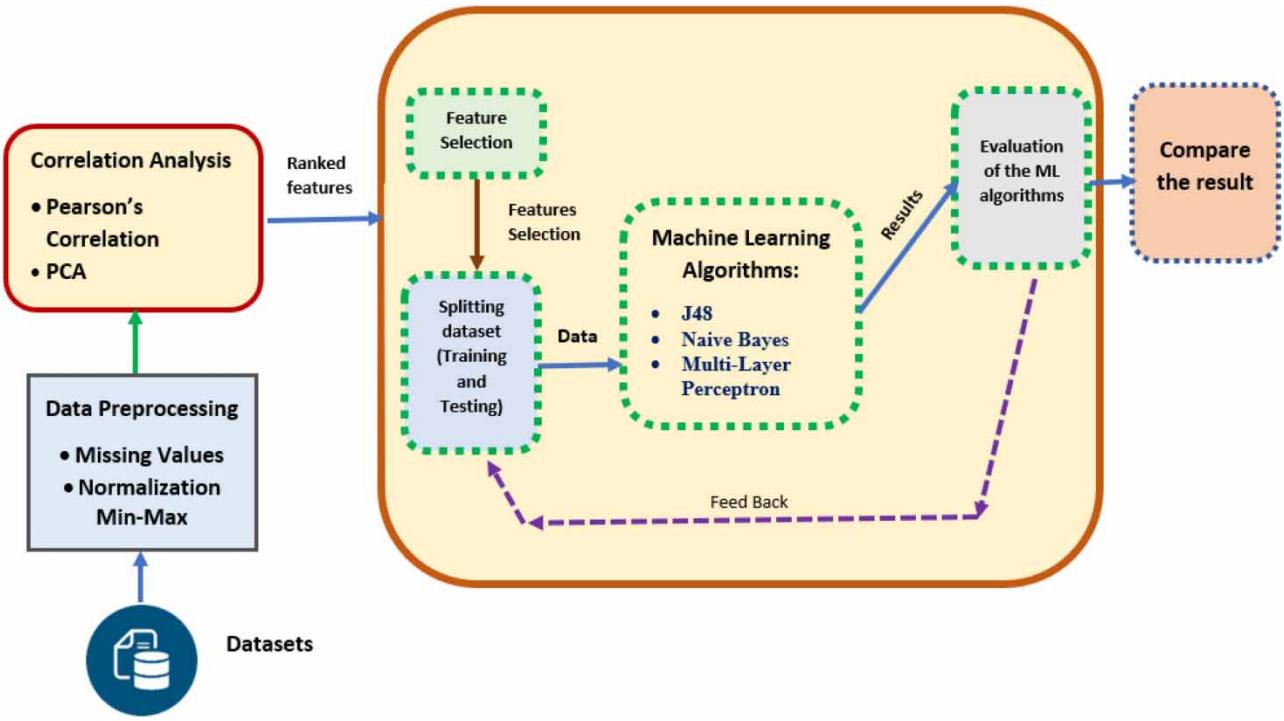
In the past years, there has been a lot of interest in water quality and its prediction as there are many pollutants that affect water quality. The techniques provided herein will help us in controlling and reducing the risks of water pollution. In this study, we will discuss concepts related to machine learning models and their applications for water quality classification (WQC). Three machine learning algorithms, J48, Naïve Bayes, and multi-layer perceptron (MLP), were used for WQC prediction. The dataset used contains 10 features, and in order to evaluate the machine's algorithms and their performance, some accuracy measurements were used. Our study showed that the proposed models can accurately classify water quality. By analyzing the results, it was found that the MLP algorithm achieved the highest accuracy for WQC prediction as compared to other algorithms.

Key words: data mining, machine learning, multi-layer perceptron algorithm, Pearson's correlation coefficient, principal component analysis (PCA), water quality classification (WQC)

HIGHLIGHTS

- Machine learning concept was adopted to analyze the water quality.
- The accuracy of multi-layer perceptron (MLP), is higher than other machine learning algorithms for water quality classification (WQC).
- Extraction of useful and relevant features increases classification accuracy using principal component analysis (PCA).
- The PCA was used for dimensionality reduction and extracts the most dominant water quality features.

GRAPHICAL ABSTRACT



ABBREVIATIONS

AI	artificial intelligence
ANN	artificial neural network
EC	electrical conductivity
GA	genetic algorithm
IoT	Internet of Things
K-NN	K-nearest neighbor
MLP	multi-layer perceptron
PCA	principal component analysis
PRC	precision–recall curve
ROC	receiver operating characteristic curve
SAR	sodium absorption ratio
SVM	support vector machine
TDS	total dissolved solids
TH	total hardness
TOC	total organic carbon
WQC	water quality classification
WQI	water quality index
WQS	water quality status

INTRODUCTION

Water is one of the most important natural resources on which the planet depends, as it constitutes 71% of the Earth's area. Water is of great importance to the life of everything on earth, as we know that all living creatures cannot live without water. Water is the basis of human, animal, and plant life, and its use is not only limited to drinking but is also considered as an important resource in industry, agriculture, and global trade through the seas and oceans. Because of the importance of water for human life, research has focused on water quality and its preservation from pollution based on international

standards (Cosgrove & Loucks 2015) in order for pollution not to exceed the standard limits and threaten human life and living creatures with disease or death.

Specific quality standards are available to indicate the quality of different water sources, such as groundwater, springs, rivers, lakes, and streams, and there are specific water quality criteria for agricultural, industrial, human, or other water uses. For instance, drinking water should be fresh and unpolluted, while irrigation water should not be too saline and toxic, and water standards for industrial uses have different characteristics depending on the nature of the industrial processes. The different practices and activities of humans, industrial processes, and natural processes affect the quality of water resources in a significant and alarming manner, especially for humans (Jury & Vaux 2005; WHO 2011; Adebesin *et al.* 2018). These activities and practices will cause pollution by leaving most wastes and pollutants without adequately being treated. This, in turn, affects natural water resources. Industrial plants and vehicles cause acidic conditions to develop in surface water and groundwater sources by lowering pH levels, decreasing acid-neutralizing capacity, and increasing aluminum concentrations (Jury & Vaux 2005). This then affects precipitation, surface water, and groundwater, as well as degrading ecosystems (Abuzir & Abuzir 2021). Sewage and runoff from farms, farmlands, and gardens can contain pesticides and nutrients, such as nitrogen and phosphorus, that cause excessive aquatic plant growth. These pollutants, which enter the water body through various channels, have become a great source of various dangers to the environment (WHO 2011; Adebesin *et al.* 2018).

Water quality is determined by features such as pH value, hardness caused by calcium and magnesium salts, total dissolved solids (TDS), chloramines, sulfate, electrical conductivity (EC), total organic carbon (TOC), turbidity, and trihalomethanes. Machine learning algorithms first need to pre-process the data and manage the missing data using weka.filters.unsupervised.attribute.ReplaceMissingValues, feature correlation, apply classification machine learning, and analyze the value of the feature selection (Islam Khan *et al.* 2021).

Water is one of the most vital resources for the sustainability of life on earth. There are many laboratory tests in research centers and universities to check the quality of water. In these research centers, water quality is assessed through laboratory and statistical analyses that are time-consuming and expensive. These analyses require the collection of samples, transportation to laboratories for examination, and a great deal of time and calculation. The results of these tests are very important to detect whether the water is contaminated with water quality features or not. These conditions call for covert treatment to detect water pollution in a faster and cheaper way (Thienen *et al.* 2018). The primary contribution of the current study is analyzing the performance of machine learning algorithms in predicting the water quality classification (WQC). Three machine learning algorithms, namely J48, Naïve Bayes, and multi-layer perceptron (MLP), were applied to predict the WQC. Performing measurement analysis using metrics, such as root mean squared error, recall, precision, receiver operating characteristic (ROC) area, and precision–recall curve (PRC) area, was used in this investigation.

In this study, we presented the topics according to the following structure: first, the background and the problem are presented in the section 'Literature Review', followed by the relevant literature in the section 'Materials and Methods', and classification of the data to estimate the state of water quality features based on machine learning algorithms in the section 'Results Analysis and Discussion'. Next, we used these models to analyze and utilize them as a tool to aid in interpreting our results and decision-making and concluded with a discussion of the usefulness of the methodology and tools developed in this study.

LITERATURE REVIEW

Water is one of the most important elements for the existence of life. Drinking water safety and accessibility are urgent issues around the world. There is extensive work on using machine learning in the water quality index (WQI), WQC, and wastewater treatment (Asadi *et al.* 2017; Szeląg *et al.* 2017; Güller *et al.* 2019; Qiu *et al.* 2021). In the study of Hassan *et al.* (2021), various machine learning techniques such as random forests (RF), neural network (NN), multinomial logistics regression (MLR), support vector machine (SVM), and bagged tree models (BTM) have been applied to classify a dataset of water quality in India. Their results showed that nitrate, pH, conductivity, dissolved oxygen (DO), total coliform (TC), and biological oxygen demand (BOD) are the main features that affect WQC.

Ahmed *et al.* (2019) used three different machine learning algorithms such as gradient boosting, MLP, and polynomial regression to predict water quality. They used four different features, namely, pH, TDS, temperature, and turbidity. The results showed that MLP has the highest classification accuracy of 85.07%, with a configuration of (3, 7). Aldhyani *et al.* (2020) used

three different machine learning algorithms to predict WQC, namely SVM, K-nearest neighbor (K-NN), and Naïve Bayes. The result showed that the SVM algorithm has the highest classification accuracy of 97.01%. In order to determine the WQI, they used two artificial intelligence (AI) techniques, namely, nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM). The NARNET technique showed slightly better performance than the LSTM.

The study by Sillberg *et al.* (2021) used a machine learning algorithm called the SVM algorithm and attribute realization (AR) to classify the water quality of the Chao Phraya River. Their results showed that AR-SVM has achieved 0.86–0.95 accuracy when applying three to six features to classify river water's quality. The study by Azad *et al.* (2017) used machine learning techniques to select the quality features mentioned in their dataset for the Gorganroud River water. They used the following three machine learning algorithms: ant colony optimization for continuous domains (ACOR), genetic algorithm (GA), and adaptive neuro-fuzzy inference system (ANFIS) for evaluating the quality features of the Gorganroud River water. The ANFIS model showed the best performance in the prediction of the features (EC, sodium absorption ratio (SAR), and total hardness (TH)) mentioned in the training stage.

Kakkar *et al.* (2021) collected data using the Internet of Things (IoT) and employed a machine learning technology neural network to forecast the amount of water pollution in residential overhead tanks. Khan *et al.* (2021) utilized the principal component regression (PCR) technique to select the most dominant WQI features. Regression algorithms are used for the principal component analysis (PCA) output and utilized gradient boosting classifiers to classify the water quality status (WQS). The study by Lerios & Villarica (2019) aimed to utilize data mining techniques for pattern extraction and model prediction of water quality in water reservoirs using different features and the WQI. The result indicated that the WQI was mostly in fair and marginal rank, providing an indication that water quality was being threatened by different water pollutants. The main objective of the study by Solanki *et al.* (2015) was to use deep learning for accurate predictions of water quality features using the WEKA tool. The evaluation in their approach was based on metrics, such as mean absolute error and mean square error, to examine the error rate of prediction.

MATERIALS AND METHODS

In this paper, machine learning algorithms were used to classify water quality with a model based on several algorithms. In the following paragraphs, we explained the appropriate methodology to implement our approach (Figure 1):

Data collection: the main step in machine learning manipulation is collecting data in digital format. We can utilize it by linking the features of the data to make predictions and interpretations of our results based on our machine learning algorithms.
Missing values replacement: during data preparation, it is common to replace all missing values for nominal and numeric features in a dataset with the modes and means from the training data.
Normalization: in this method, we normalized all number values in the dataset.

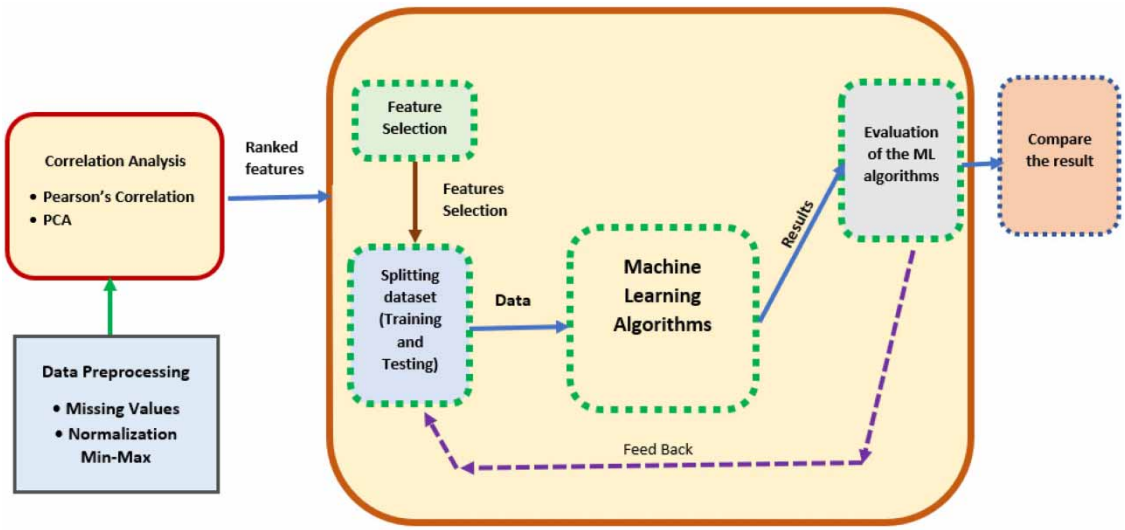


Figure 1 | Framework of the proposed approach.

Feature selection: we used the two techniques, Parsons' analysis and PCA, to extract significant features from the dataset.
Data split: the dataset is divided into two sets: training and testing with 10-folds for cross-validation.
Statistical analysis of the features: this method describes the significant distribution and correlation coefficient between the features of the dataset.
Dimension reduction: this method is used for the selection of the most important features for WQC.
Classification algorithms: we used different machine learning algorithms including J48, Naïve Bayes, and MLP.
Evaluation: significant measurement methods are used in order to evaluate our machine learning algorithms.

The following algorithms were employed in our study such as Decision Tree, Naïve Bayes, and MLP.

A **Decision Tree (DS)** is a supervised machine learning technique where the input data (X) and the output or class label (terminal node) are in the training data. It can be used in machine learning applications for both classification and regression. The entropy is used to determine the root variable and, accordingly, is oriented towards the values of other features. We do not know if the parent entropy or the entropy of a particular node has decreased or not. In order to find that, we used a new metric called 'information gain,' which tells us how much the parent entropy has decreased after splitting it with some features. We can calculate entropy which is the amount of uncertainty in the dataset and information gain by using the following formula:

$$E(S) = \sum_{i=1}^{c} -p_i \log p_i \tag{1}$$

where S is the subset of the training example and p is the probability that the tuple belongs to class C.

The information required for exact classification after portioning is given by the formula:

$$E(T.X) = \sum_{c \epsilon X} P(c)E(c) \tag{2}$$

where $P(c)$ is the weight of partition. This information represents the information needed to classify the dataset D on portioning by X.

Information gain is the difference between the original and expected information that is required to classify the tuples of dataset D.

$$Information\ Gain(T,\ X) = E(T) - E(T, X) \tag{3}$$

Naïve Bayes is a probabilistic classifier based on the Bayes theorem and is used in machine learning applications that require classification tasks. Simplicity, speed, accuracy, and reliability are among the features of Naïve Bayes. This theorem is used to calculate conditional probabilities. There are different Naïve Bayes, such as Multinomial Naïve Bayes Classifier, Bernoulli Naïve Bayes Classifier, and Gaussian Naïve Bayes Classifier. The general form is

$$P(X_1,\ \ldots, X_n|Y) = \prod_{j=1}^{n} P(X_j|Y) \tag{4}$$

Given X_j are conditionally independent given Y.

An MLP is an artificial neural network (ANN) algorithm used for classification and regression. It is composed of multiple layers of perceptrons with a threshold activation function. The architecture of MLP consists of three layers, two of them are visible layers, the input layer and output layer, and the hidden layers.

The algorithm of MLP iterates using training data and generalizes the output model by calculating and updating the weight on each node of each layer. For classification or prediction, we use the training model with weights to decide what units to activate based on the input (Figure 2).

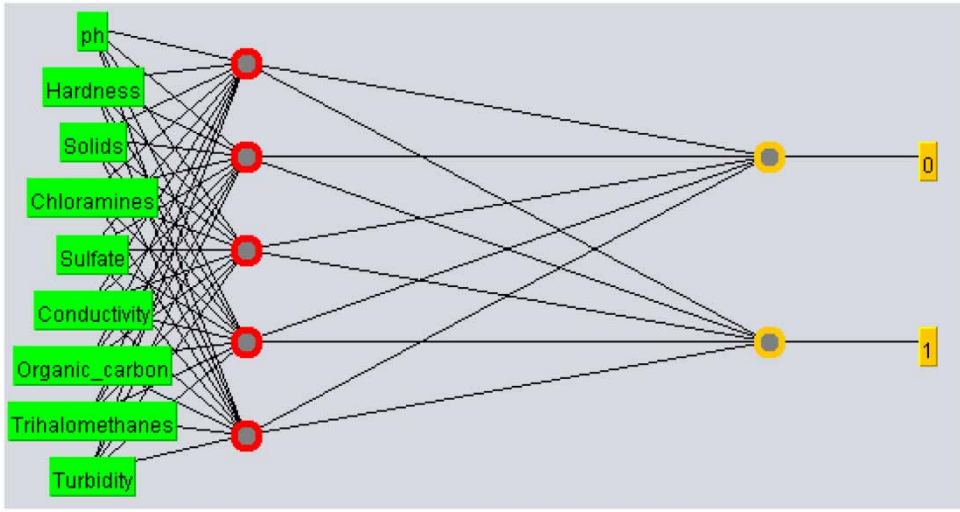


Figure 2 | An MLP architecture based on our model.

We can apply the different machine learning algorithms and our framework to predict water quality in different countries as water bodies can be very different. Our predictive and classification approach could be applied to datasets from different countries.

Dataset

To achieve the process of building a machine learning model, we used the water_potability.csv file that contains water quality metrics for 3,276 different water bodies. Available, comprehensive, and rigorous water quality datasets are key in conducting machine learning application techniques for the purpose of scientific research. The data are comprehensive and used for research purposes, especially in machine learning applications. The dataset was downloaded from Kaggle (https://www.kaggle.com/adityakadiwal/water-potability). There are 10 features in the dataset: pH value, hardness caused by calcium and magnesium salts, TDS, chloramines, sulfate, EC, TOC, turbidity, trihalomethanes, and potability. The classification feature is potability with {0,1} values. Table 1 shows these features with basic statistical analysis.

Data normalization

Normalization is used as an initial stage in order to prepare the data for machine learning. This process will change the values of numeric features in the dataset to be on a similar scale, without distorting differences in value ranges or losing information.

Table 1 | Basic statistical analysis for water quality features

Feature	Unit WHO	Min	Max	Mean	SD
pH	6.52–6.83	0	14	7.081	1.47
Hardness	Hardness caused by calcium and magnesium salts	47.432	323.124	196.369	32.88
TDS	500–1,000 mg/L	320.943	61,227.196	22,014.093	8,768.571
Chloramines	4 mg/L or 4 ppm	0.352	13.127	7.122	1.583
Sulfate	3–30 mg/L	129	481.031	333.776	36.143
EC	200–400 μS/cm	181	753.343	426.205	80.824
TOC	EPA <2 mg/L as TOC in treated/drinking water and <4 mg/L in source water which is use for treatment	2.2	28.3	14.285	3.308
Trihalomethanes	5.00 NTU	0.738	124	66.396	15.77
Turbidity	THM levels up to 80 ppm	1.45	6.739	3.967	0.78
Potability	0 or 1				

Therefore, the new scale will make the largest value for each feature equal to 1 and the smallest value equal to 0 (Figure 3). In general, we apply the concept of normalization when we do not know the distribution of the data or when we know that the distribution is not a normal one. In machine learning, we called min–max normalization (linear scaling), the formula is given below

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

Feature selection

Feature selection is divided into two parts:

- *Features relationship:* an evaluation method for evaluating each feature in the dataset based on the output variable and the relationship that the features have with each other. In our study, we used Pearson and PCA to perform this stage in order to achieve a shortlist of the selected features.
- *Test method:* to test the selected features using the different machine learning algorithms.

Using the Pearson correlation coefficient, we can calculate the correlation between each feature and the output variable and the order of the correlation. Based on the calculations, we only identify features that have a high positive or negative correlation (close to −1 or 1), i.e., features that are closely related to the output variable whether it is positive or negative correlation. For those features with low correlation (the value is close to zero), we can just drop them.

Experimental setup

The aim of this research is to detect or determine water quality using historical data from Kaggle. Three machine learning algorithms were adopted in order to achieve the results. These algorithms are J48, Naïve Bayes, and MLP. This section discusses experiments, results, and evaluation of our approach. In order to implement our model, we used the data mining tool WEKA 3.9.

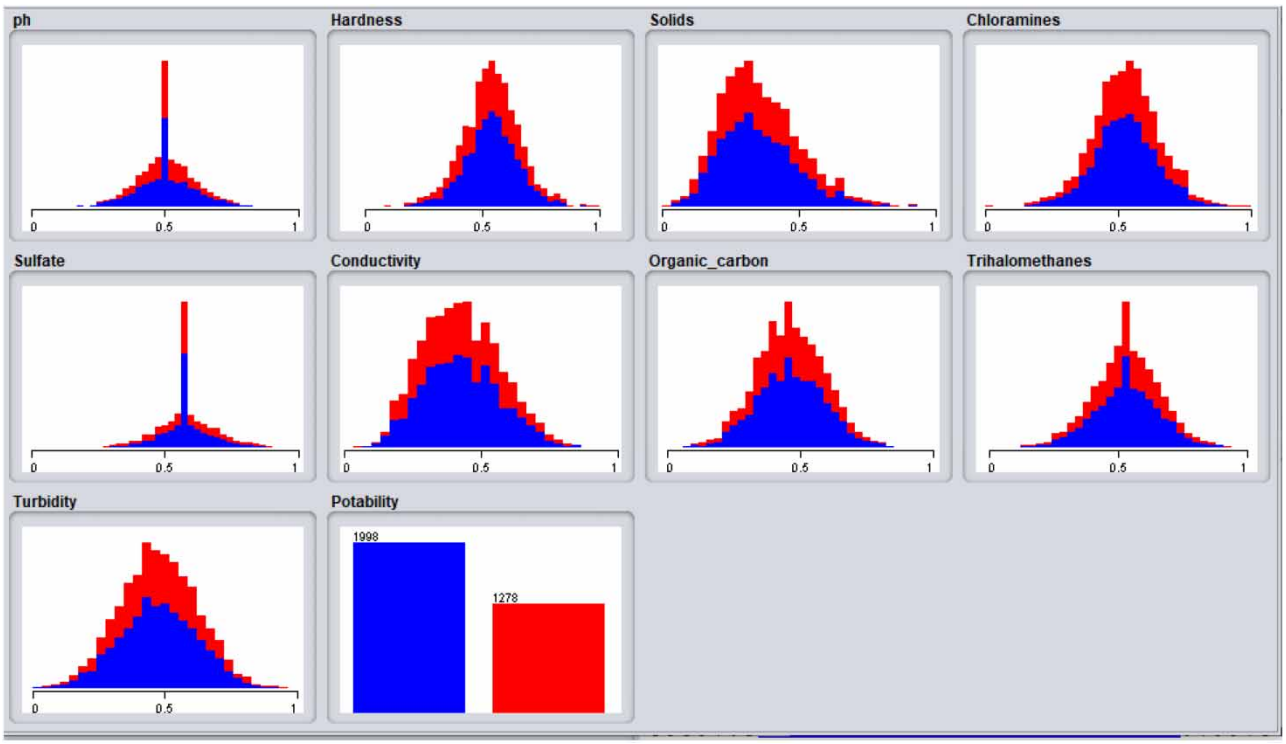


Figure 3 | Normalization of the dataset.

Data analysis

After performing all the basic operations we needed to manipulate the data with the aim of analyzing it. In our research, we applied three machine learning algorithms to determine WQC using all features and the fewest or specified number of features depending on the feature selection techniques. Before applying the machine learning algorithms, in the early steps of data preparation, we used correlation analysis, data segmentation, and feature selections to prepare the data as input to the three machine learning algorithms.

Correlation analysis

In the early steps of our model, we filled in the missing values and applied a normalization algorithm to our dataset. The number of the features in our dataset is 10, and not all of these features are related and useful for predicting the classification label (output feature). In machine learning, the feature selection task can be useful in finding dependent features in our model. Correlation analysis can be used in order to reduce the dimension of the data and infer the possible relationships between the input features. In this work, we analyzed and applied the Pearson correlation as a measure to identify and select the relevant features from the list of features in Table 1. We removed the irrelevant features and applied our model using the values of the selected features.

The purpose of correlation analysis is to measure, analyze, and determine the degree of the relationship between two features and the dependence between features. In order to determine the relationship between the features, we used decimal values, known as the correlation coefficient. If the value of the correlation coefficient is greater than 0 (positive sign), this indicates that the relationship between the features is significant. The negative value of the correlation coefficient indicates that the relationship between the features is weak. The correlation coefficient of a value of $+1$ indicates that the two features have a strong correlation. On the other hand, the correlation coefficient of a value of -1 indicates otherwise. There is no correlation if the correlation coefficient is 1.

This relationship can be used to identify the strongest features that can have a significant effect (a strong predictor) and more efficiently predict the outcomes. In this research, we used two measurements, namely, Pearson's coefficient and PCA, to find the relevant features and the relationships between the features themselves and the labeled feature. These techniques are often used to determine the strong predictor for water quality features.

Pearson's correlation coefficient can be calculated using the following formula:

$$R_{x,y} = \frac{\text{Covariance}(x,\ y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{6}$$

where $C_{x,y}$ is the correlation coefficient, Covariance (x,y) is the covariance, and σ_x and σ_y are the standard deviations of x and y, respectively. x_i is the value of the x-variable in a sample, $\bar{x}$ is the mean of the values of the x-variable, y_i is the value of the y-variable in a sample, and $\bar{y}$ is the mean of the values of the y-variable.

Weka can perform the correlation-based feature selection (Pearson's correlation coefficient) using the CorrelationAttributeEval. It requires the use of a Ranker search method. Table 2 shows our results by using the different methods. If we use 0.01 as our cut-off (threshold value) for relevant features, then we only keep TDS, TOC, chloramines, sulfate, and hardness caused by calcium and magnesium salts. The remaining features could be removed (Table 3).

In machine learning, the PCA is applied in dimensionality reduction (DR). PCA is a common technique used in feature selection to reduce the dimension of the features by keeping and considering only the relevant six features. The widely used official methods to calculate the principal components (PCs) are based on solving the covariance matrix, eigen values, and eigen vector or using singular value decomposition (SVD). As a better approach, we applied the PCA before fitting our data to a model. Table 4 shows the correlation between the features themselves.

For PCA, positive values greater than 0 show correlation, while negative values less than 0 represent no correlation between the features. As the magnitude of the positive values increases they were more strongly correlated with the other features. The analysis of the correlation values in Table 4 indicates that:

- pH is not strongly or poorly related with hardness, sulfate, EC, and TOC and not correlated with TDS, chloramines, and turbidity.
- Hardness is weakly correlated with pH and very loosely related with the other features.

Table 2 | Ranked features based on Pearson's correlation coefficient

Feature number	Feature	Pearson's correlation coefficient
3	Solids	0.03374
7	Organic_carbon	0.03
4	Chloramines	0.02378
5	Sulfate	0.02062
2	Hardness	0.01384
6	Conductivity	0.00813
8	Trihalomethanes	0.00696
1	pH	0.00329
9	Turbidity	0.00158

Table 3 | Feature selection using Pearson's correlation coefficient

Methods	Threshold value	Features
Pearson's correlation coefficient	Threshold value = 0.01	TDS, TOC, chloramines, sulfate, and hardness caused by calcium and magnesium salts.
	Threshold value = 0.02	TDS, TOC, chloramines, and sulfate.
	Threshold value = 0.03	TDS and TOC.

Table 4 | Correlation coefficient between the input features of WQC (correlation matrix PCA)

Features	pH	Hardness	TDS	Chloramines	Sulfate	EC	TOC	Trihalomethanes	Turbidity
pH	1	0.08	− 0.08	− 0.03	0.01	0.02	0.04	0	− 0.04
Hardness	0.08	1	− 0.05	− 0.03	− 0.09	− 0.02	0	− 0.01	− 0.01
TDS	− 0.08	− 0.05	1	− 0.07	− 0.15	0.01	0.01	− 0.01	0.02
Chloramines	− 0.03	− 0.03	− 0.07	1	0.02	− 0.02	− 0.01	0.02	0
Sulfate	0.01	− 0.09	− 0.15	0.02	1	− 0.01	0.03	− 0.03	− 0.01
EC	0.02	− 0.02	0.01	− 0.02	− 0.01	1	0.02	0	0.01
TOC	0.04	0	0.01	− 0.01	0.03	0.02	1	− 0.01	− 0.03
Trihalomethanes	0	− 0.01	− 0.01	0.02	− 0.03	0	− 0.01	1	− 0.02
Turbidity	− 0.04	− 0.01	0.02	0	− 0.01	0.01	− 0.03	− 0.02	1

- TDS is correlated with EC, TOC, and turbidity and loosely correlated with pH, hardness, TDS, chloramines, sulfate, and trihalomethanes.
- Chloramines are correlated with sulfate and trihalomethanes and is not correlated with pH, hardness, TDS, chloramines, EC, TOC, and turbidity.
- Sulfate is correlated with chloramines and trihalomethanes and is not correlated with pH, hardness, TDS, sulfate, EC, TOC, and turbidity.
- EC is correlated with pH, TDS, TOC, and turbidity and is not correlated with hardness, chloramines, sulfate, EC, and trihalomethanes.

- TOC is correlated with pH, TDS, sulfate, and EC and is not correlated with hardness, chloramines, TOC, trihalomethanes, and turbidity.
- Trihalomethanes are correlated with chloramines and is not correlated with pH, hardness, TDS, sulfate, EC, TOC, trihalomethanes, and turbidity.
- Turbidity is correlated with TDS and EC and is not correlated with pH, hardness, chloramines, sulfate, TOC, trihalomethanes, and turbidity.

RESULTS ANALYSIS AND DISCUSSIONS

To illustrate the machine learning models in this study, we used a water quality dataset and 10-fold cross-validation. The resulting training set consisted of 3,276 different water bodies, 1,278 of them classified as non-potability and the rest 1,998 as potability. The performance measure for evaluating the models was the different measurements as root mean squared error (RMSE), precision, recall, *F*-measure, area under the ROC curve, and PRC area. The previously mentioned methods and measurements can be calculated using the following formulas:

$$RMSE = \sqrt{\frac{\sum (x_{obs} - x_{pred})^2}{n}} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

An ROC curve is a plot of the false positive rate (TPR) on the *x*-axis and the true positive rate on the *y*-axis. The term referred to as the sensitivity or the recall.

$$True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{11}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \tag{12}$$

We can think of the plot as the fraction of correct predictions for the positive class (*y*-axis) versus the fraction of errors for the negative class (*x*-axis).

To achieve our goal, we used three machine learning methods to classify water quality classes based on the datasets. We utilized J48, Naïve Bayes, and MLP to achieve our model. We applied each machine learning algorithm with the complete list of features and after applying feature selection. Initially, all the features were considered (Table 5). We used ROC area as the accuracy measurement for utilizing these algorithms. Among these algorithms, MLP performed better than the other and has the highest accuracy of 0.661 for ROC and 0.734 for PRC, while Naïve Bayes has 0.589 for ROC and 0.684 for PRC. The J48

Table 5 | Classification results using all features

ML algorithm	Feature selection	Correctly classified instances	Incorrectly classified instances	RMSE	Precision	Recall	*F*-measure	ROC area	PRC area
J48	All features	2,074 63.3089%	1,202 36.6911%	0.49	0.637	0.928	0.755	0.565	0.652
Naïve Bayes	All features	2,040 62.2711%	1,236 37.7289%	0.4841	0.639	0.878	0.740	0.589	0.684
MLP	All features	2,172 66.3004%	1,104 33.6996%	0.4712	0.673	0.869	0.759	0.661	0.734

algorithm has the lowest accuracy with 0.565 for ROC and 0.652 for PRC. We also obtained the same order of accuracy when we applied feature selection in our model when we used four to five features in our model (Tables 6 and 7), whereas Naïve Bayes performed better in prediction or classification of WQC (Table 8).

The ROC curve is only defined for binary classification problems. If we have a score of around 0.5 that means the system is essentially randomly guessing. Anything above 0.5 means that the system is performing better than random guessing. Anything below means the system is not classifying correctly. When we iterated through our results, we can see that J48 algorithms do not classify when the number of features is reduced to two features (Table 8) in our model (Figure 4).

It is clear from the results that machine learning is sensitive to irrelevance and number of features. As shown in the results, the reduced predictive performance was related to the number of selected features. The results showed for all algorithms, especially MLP.

From the results, feature reduction makes sense to provide these algorithms with minimal features that provide acceptable results (Tables 5–7). In some cases, removing features can reduce the cost of acquiring data or improve the productivity of the software used to make predictions. It is generally better to have fewer features in the dataset file.

The use of a variety of machine learning algorithms in our study together to predict WQ gives better results than using a single model. There are several proposed methodologies for predicting WQ. These methodologies include different statistical methods and machine learning algorithms. In order to determine the correlation and relationship between different water

Table 6 | Classification results using five features

ML algorithm	Features selection	Correctly classified instances	Incorrectly classified instances	RMSE	Precision	Recall	*F*-measure	ROC area	PRC area
J48	Threshold value = 0.01	2,041 62.3016%	1,235 37.6984%	0.4874	0.626	0.948	0.754	0.534	0.631
Naïve Bayes	Threshold value = 0.01	2,028 61.9048%	1,248 38.0952%	0.4861	0.636	0.877	0.737	0.576	0.659
MLP	Threshold value = 0.01	2,046 62.4542%	1,230 37.5458%	0.4857	0.629	0.937	0.753	0.568	0.650

Table 7 | Classification results using four features

ML algorithm	Features selection	Correctly classified instances	Incorrectly classified instances	RMSE	Precision	Recall	*F*-measure	ROC area	PRC area
J48	Threshold value = 0.02	2,041 62.3016%	1,235 37.6984%	0.4873	0.625	0.953	0.755	0.533	0.629
Naïve Bayes	Threshold value = 0.02	2,036 62.149%	1,240 37.851%	0.4856	0.634	0.895	0.743	0.562	0.651
MLP	Threshold value = 0.02	2,071 63.2173%	1,205 36.7827%	0.484	0.633	0.942	0.757	0.568	0.647

Table 8 | Classification results using two features

ML algorithm	Features selection	Correctly classified instances	Incorrectly classified instances	RMSE	Precision	Recall	*F*-measure	ROC area	PRC area
J48	Threshold value = 0.03	1,998 60.989%	1,278 39.011%	0.4878	0.610	1.000	0.758	0.499	0.499
Naïve Bayes	Threshold value = 0.03	1,984 60.5617%	1,292 39.4383%	0.488	0.611	0.976	0.751	0.512	0.619
MLP	Threshold value = 0.03	1,999 61.0195%	1,277 38.9805%	0.4893	0.611	0.994	0.757	0.509	0.620

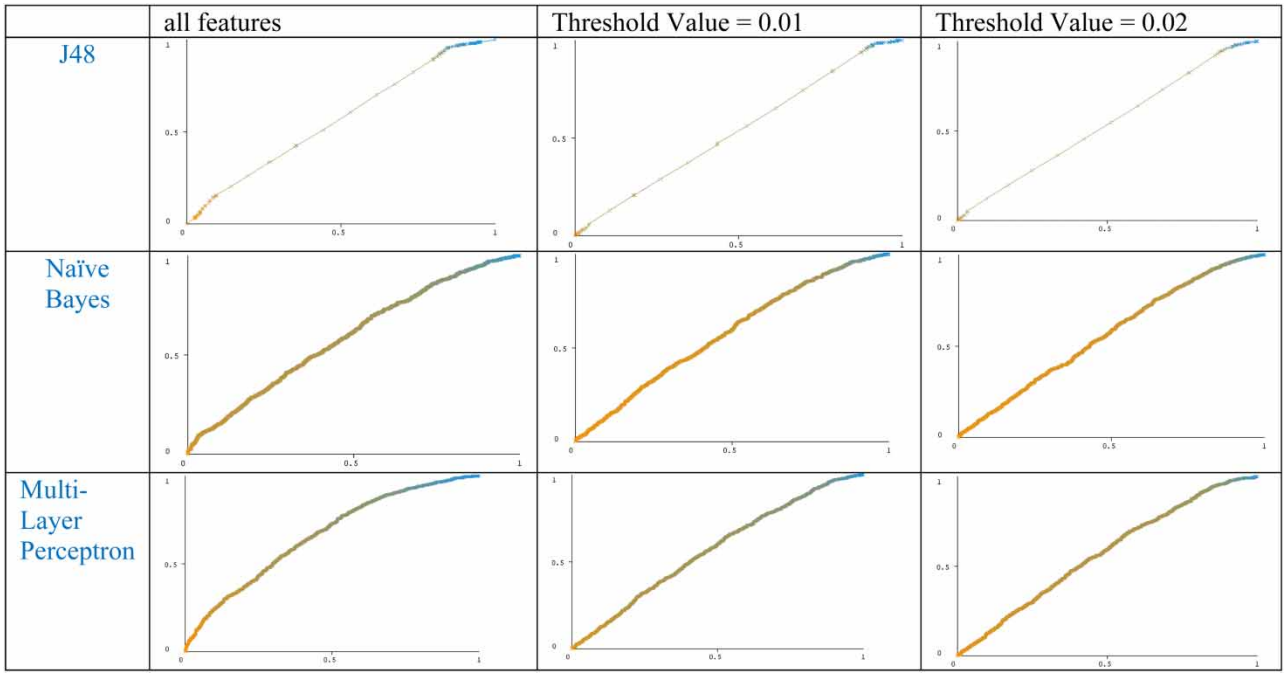


Figure 4 | Representation of ROC area with all features and 4–5 features.

quality features, we applied the PCA to the dataset. The PCA is a statistical analysis used for dimensionality reduction and extracts the most dominant water quality features. For predicting the WQ, three different machine learning algorithms were used. Significant results have been observed after using different statistical analysis methods (PCA), accuracy measurement methods (ROC), and machine learning algorithms. Finally, the best prediction model is selected by comparing the performances of such models.

CONCLUSION

The study used the dataset from Kaggle for training and testing. Three different machine learning algorithms such as J48, Naïve Bayes, and MLP models were used. The machine learning models are trained and tested using the default configuration of the features of the machine learning classifier. We analyzed and compared the results of our models, especially the accuracy measurement that is based on ROC for classifying water quality conditions into potable (1) or non-potable (0) classes. We studied the accuracy results for each model and compared their performance in classifying our data with different configurations of the number of selected features. The performance of our model showed that MLP performs better than the other two algorithms with all features and with a good number of selected features. Naïve Bayes predicts better with fewer numbers of features selected. We used ROC area as the accuracy measurement for utilizing these algorithms.

In future work, we recommend using IoT technology and integrating it with an online monitoring system using the required sensors and machine learning algorithms. Other algorithms and techniques can be proposed like the deep learning approach to improve the efficacy of the classification process. Another proposal is to use other different machine learning algorithms, namely ensemble learning, SVM, and K-NN. Further research can be done on management/measures or practices, as well as the natural characteristics of water bodies that will influence water quality.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories. Water Quality: Drinking water potability https://www.kaggle.com/adityakadiwal/water-potability.

CONFLICT OF INTEREST STATEMENT

The authors declare there is no conflict.

REFERENCES

Abuzir, S. & Abuzir, Y. 2021 Data mining for CO_2 emissions prediction in Italy. *The Journal of Engineering Sciences and Researches* **3** (1), 59–68. doi:10.46387/bjesr.862179.

Adebesin, B. O., Oluyori, A. P., Adelani-Akande, T. A., Dada, A. O. & Oreofe, T. A. 2018 Water pollution: effects, prevention, and climatic impact. In: *Water Challenges of an Urbanizing World* (Glavan, M., ed.). IntechOpen. doi:10.5772/intechopen.72018.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R. & García-Nieto, J. 2019 Efficient water quality prediction using supervised machine learning. *Water* **11** (11), 2210. https://doi.org/10.3390/w11112210.

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

Asadi, A., Verma, A., Yang, K. & Mejabi, B. 2017 Wastewater treatment aeration process optimization: a data mining approach. *Journal of Environmental Management* **203**, 630–639. doi:10.1016/j.jenvman.2016.07.047.

Azad, A., Karami, H., Farzin, S., Saeedian, A., Kashi, H. & Sayyahi, F. 2017 Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River). *KSCE Journal of Civil Engineering* **22** (7), 2206–2213. doi:10.1007/s12205-017-1703-6.

Cosgrove, W. J. & Loucks, D. P. 2015 Water management: current and future challenges and research directions. *Water Resources Research* **51** (6), 4823–4839. https://doi.org/10.1002/2014WR016869.

Güller, S., Silahtaroğlu, G. & Akpolat, O. 2019 Analysis waste water characteristics via data mining: a Muğla province case and external validation. Communications in statistics: case studies. *Data Analysis and Applications*, 1–14. doi:10.1080/23737484.2019.1604192.

Hassan Md., M., Akter, L., Rahman Md., M., Zaman, S., Hasib Md., K., Jahan, N., Smrity, R. N., Farhana, J., Raihan, M. & Mollick, S. 2021 Efficient prediction of water quality index (WQI) using machine learning algorithms. *Human-Centric Intelligent Systems* **1** (3–4), 86–97. https://doi.org/10.2991/hcis.k.211203.001.

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. *Journal of King Saud University – Computer and Information Sciences*. doi:10.1016/j.jksuci.2021.06.003.

Jury, W. A. & Vaux, H. 2005 The role of science in solving the world's emerging water problems. *Proceedings of the National Academy of Sciences* **102** (44), 15715–15720. doi:10.1073/pnas.0506467102.

Kakkar, M., Gupta, V., Garg, J. & Dhiman, S. 2021 Detection of water quality using machine learning and IoT. International Journal of Engineering Research & Technology *(*IJERT*)* **10** (11), 73–75.

Lerios, J. L. & Villarica, M. V. 2019 Pattern extraction of water quality prediction using machine learning algorithms of water reservoir. *International Journal of Mechanical Engineering and Robotics Research* **8** (6), 992–997.

Qiu, J., Lü, F., Zhang, H., Shao, L. & He, P. 2021 Data mining strategies of molecular information for inspecting wastewater treatment by using UHRMS. *Trends in Environmental Analytical Chemistry* **31**, e00134. doi:10.1016/j.teac.2021.e00134.

Sillberg, C. V., Kullavanijaya, P. & Chavalparit, O. 2021 Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya river. *Journal of Ecological Engineering* **22** (9), 70–86. https://doi.org/10.12911/22998993/141364.

Solanki, A., Agrawal, H. & Khare, K. 2015 Predictive analysis of water quality parameters using deep learning. *International Journal of Computer Applications* **125** (9), 29–34.

Szeląg, B., Barbusiński, K., Studziński, J. & Bartkiewicz, L. 2017 Prediction of wastewater quality indicators at the inflow to the wastewater treatment plant using data mining methods. *E3S Web of Conferences* **22**, 00174. doi:10.1051/e3sconf/20172200174.

Thienen, P., Alphen, H.-J., Brunner, A., Fujita, Y., Hillebrand, B., Sjerps, R., Summeren, J., Verschoor, A. & Wullings, B. 2018 Explorations in Data Mining for the Water Sector. Water Research Institute (KWR), Delft, The Netherlands, Report BTO 2018.085.

World Health Organization WHO 2011 *Guidelines for Drinking-Water Quality*, 4th edn. WHO Library Cataloguing-in-Publication Data, Geneva.

First received 23 January 2022; accepted in revised form 12 May 2022. Available online 30 May 2022