Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

***E-ISSN: 2687-6167*** ***Number 58, September 2024***

***RESEARCH ARTICLE***

*Receive Date: 07.01.2024* *Accepted Date: 15.07.2024*

# Comparative analysis of machine learning techniques for detecting potability of water

Vahid Sinap[a,*]

[a]*Ufuk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, 06805, Ankara, Türkiye, ORCID: 0000-0002-8734-9509*

**Abstract**

This research aims to evaluate the effectiveness of machine learning algorithms in determining the potability of water. In the study, a total of 3276 water samples were analyzed for 10 different features that determine the potability of water. Besides that, the study's consideration is to evaluate the impact of trimming, IQR, and percentile methods on the performance of machine learning algorithms. The models were built using nine different classification algorithms (Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, AdaBoost, and Bagging Classifier). According to the results, filling the missing data with the population mean and handling outliers with Trimming and IQR methods improved the performance of the models. Random Forest and Decision Tree algorithms were the most accurate in determining the potability of water. The findings of this research are of high importance to sustainable water resource management and serve as a crucial input for the decision-making process on the quality of water. The study also offers an example for researchers working on datasets that contain missing values and outliers.

*Keywords:* Water quality; potability analysis; machine learning; data processing; classification

* Corresponding author. Tel.: +90-312-586-7378.
*E-mail address:* vahidsinap@gmail.com

## 1. Introduction

Water is an essential resource for sustaining life. Regular monitoring of water quality is important for the health of ecosystems and the human population [1]. Potable water is the water quality that is mainly used for drinking, cooking, and hygiene practices [2]. This water has been stripped of toxic substances and microorganisms through different cleaning and treatment processes. Water is not only a liquid that a human being needs to drink for them to survive. It is also an invaluable commodity in the agricultural, manufacturing, and power generation sectors. So, the safeguarding of water standards is indispensable not only for the health of the people but also for the economic and ecological sustainability [3]. Besides that, some medical practices like kidney dialysis and lens cleaning may require the use of higher quality water. At this point, the protection and monitoring of water quality becomes even more important.

A high-quality drinking-water-supply system greatly enhances people's life quality, whether in small or large towns. The most apparent gain weights the reduction of illnesses caused by polluted water for these initiatives to gain success. Diarrhea, cholera, and typhoid are quite common in rural areas so the availability of clean drinking water can minimize them [4]. In addition, easier access to safe water contributes to the healthy growth and development of children, increasing school attendance rates and expanding educational opportunities. In economic terms, access to safe water supplies allows agricultural and industrial activities to be more efficient and sustainable [5]. This contributes to the development of the local economy. In addition, easier access to clean water resources encourages people to use environmentally friendly water. In this way, natural resources are protected. Increasing access to safe water positively affects the overall well-being, health, and environmental sustainability of a society [6].

Considering the importance of water for life on earth, access to clean drinking water is an important health and development issue at global, national, and local levels. Investments in water supply and sanitation also provide economic benefits through reduced risk of disease and lower health expenditures. This applies to large-scale investments in water supply infrastructure as well as domestic water treatment methods. To protect water quality, supply risks should be assessed, and a comprehensive strategic plan should be developed [7]. This strategy includes the systematic assessment of water-related risks at all stages from the point of supply to the consumer and the development of solutions to mitigate these risks. Several techniques are used to ensure daily assessment of water quality. These techniques analyze the chemical, physical and microbiological properties of water and assess its compliance with drinking and potable water standards [8].

The measurement of dissolved oxygen is essential for the aquatic life among chemical analyses. A low level of dissolved oxygen results in a dangerous situation for aquatic life. The pH meter is among the important monitoring apparatus of drinking water and healthy ecosystems as it can tell if the water is acidic, neutral or alkaline [9]. The TSS test is used to determine the mass of total solids in water and then the number of solids which are either filtered or unfiltered is also determined [10]. Heavy metal analysis is aimed at finding out whether the water has lead, mercury, and arsenic which are some of the harmful heavy metals [11]. With regard to the physical dimension of analysis, there are methods that are used to measure the color and odor properties of water. Color and odor are the direct indicators of the organic and inorganic pollutants, respectively [12]. Temperature measurement is another parameter besides others that can be used to determine ecosystem health as it influences biological activities [13]. Suspended matter refers to the measurement of the number of suspended solid particles in water [14]. Furthermore, microbiological analyses are carried out to check the drinking water quality too. The coliform bacteria that are the indicators of fecal pollution are a big piece of information for the potability of the water. Among the methods used are the analyses of enterococci and Escherichia coli (E. coli) which are the indicators of microbiological pollution in water [15].

Among water quality measurement methods that are currently in use, some have certain advantages as well as some limitations. These methods, in general, operate only on water samples that are collected at a specific moment in time. But then it can be said that water quality is a constantly changing phenomenon. Thus, water quality can be

influenced by many factors such as seasonal changes, weather conditions, and the human impact on the environment [16]. These methods have their limitations in reflecting the instantaneous water quality changes. On the contrary, most of the techniques used for water quality measurement are exorbitantly priced and consume a lot of time. The length of the time and the number of resources required for the whole process of the collection, transportation, and analyzing the water samples for laboratory analysis is considerable [17]. This causes a lot of problems when trying to keep the water quality monitoring system operating continuously and everywhere. Besides that, the water quality measurement techniques used in the past only dealt with samples that were collected at specific points or within certain areas [18]. This makes it impossible to detect local discrepancies in water quality and carry out comprehensive studies over large areas. The other limitation is that traditional methods of analysis tend to focus on a restricted number of parameters and also have limitations in measuring a wider range of parameters [19]. Consequently, these strategies are potentially ineffective in spotting the situations that require prompt help. In order to solve these problems, artificial intelligence (AI) methods are being implemented.

Machine learning's superiority over traditional methods in water quality monitoring is due to the fact that it can use continuous monitoring, large-scale data collection and rapid analysis. Machine learning is the one which employs a more progressive and flexible procedure than the classical ones. Traditional methods are limited in the ability to effectively capture the dynamic changes of the water quality because they get samples at some specific time points [17]. Machine learning techniques can detect sudden spikes of pollution in water, and as a result, monitoring can be done without involving humans [20]. In addition, machine learning is the procedure to follow when it comes to the collection of large-scale data. Instantaneous data streams from extensive geographical areas can be attained through sensors, cameras, and data collection tools [21]. Then the possibility of total monitoring of water quality and comparing quality differences between different regions in detail comes to the surface. Beyond that, machine learning has the capacity to analyze data and thus detect water quality changes quicker than laboratory tests [22]. The unique feature of machine learning is that it can both identify a definite pattern in the water quality and also adjust to environmental changes. This capacity can be employed to find the possible sources of pollution and constantly improve the water quality. The blending of machine learning and remote sensing techniques provides a huge benefit for water quality remote monitoring [23]. The remote sensing data collected from satellite and sensor networks can be processed by machine learning algorithms to get a complete picture of the water quality.

Along with the important benefits of machine learning techniques, there are also some challenges in the case of their usage. Datasets utilized for water quality monitoring are usually very complex and voluminous. Machine learning algorithms suffer from two main problems: firstly, the data may be incomplete or inaccurate, and secondly, the integrity of the data may be compromised. Moreover, the relevant parameters concerning water quality can be very different and hence be required to be checked together for many parameters. The combination of this result with the complexity of the model often leads to the problem of feature selection. The first series of data preprocessing steps is the remedy to these challenges. Moreover, a technique that is particularly developed for feature selection or dimensionality reduction should be utilized to easily cope with the multiple parameters.

In this paper, nine different classification algorithms were used to determine water quality and potability and performance comparisons were made. The methods under consideration are Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Gradient Boosting Decision Tree (XGBoost), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), AdaBoost (ADA) and Bagging Classifier (BAG). The performance of each algorithm is considered in detail in the report for the water quality problem. Apart from this, a major topic of discussion in the paper is the problems machine learning techniques face in water quality assessment as stated in the literature. The research study, however, deals with the performance of machine learning algorithms on water potability analysis, which is too little, and there is a lack of studies in the literature focusing on the performance difference with regard to outlier treatment methods. To overcome these challenges and the lack of knowledge in the literature, we discuss the relationship between different strategies such as feature engineering, outlier detection and processing, data normalization and standardization, and missing values evaluation used in the data preprocessing stages and their effect on the performance of the algorithms. These modern techniques will help

to make the datasets used in the process of water quality monitoring better and the machine learning models more accurate and reliable.

The performance comparisons in the paper provide important insights for decision makers, researchers, and practitioners in the field of water quality monitoring and management. The aim of these analyses is to direct the right choice of algorithms in machine learning so that water resources could be saved and managed more efficiently. The data preprocessing steps in the paper are meant to be a guide to show how to analyze large, multi-parametric, and dynamic datasets.

## 2. Related work

The literature contains some significant studies focused on water quality classification based on the application of machine learning algorithms. One study conducted using AI techniques in Tayra River in Southwest Iran to predict water quality components showed positive results, using ANN and SVM algorithms. The findings indicated that SVM was the most precise in the aspect of water quality prediction [25]. Another study by [26] developed machine learning algorithms for water quality classification to control water pollution. The three algorithms SVM, KNN and NB were used on a dataset comprising of 7 parameters. Out of the methods used, SVM algorithm recorded the highest accuracy in predicting the water quality. In a study by [27], a dataset with pH, dissolved oxygen, biological oxygen, and electrical conductivity parameters was used to verify the water quality assessment models. Machine learning algorithms such as SVM, DT, and NB were administered in the research. The evaluation results indicated that the DT algorithm had the highest accuracy. Besides, the study done by [28] also confirmed that the DT algorithm was the most successful algorithm with the same result. The performance of classification algorithms was analyzed to classify and compare the water quality of Kinta River in Perak Malaysia. NB, J48, and BAG were the algorithms that were used. NB was the best of the three models [29].

[30] sought to uncover the relationship between agricultural chemicals and the Salton Sea's water quality degradation over time. Regression and machine learning algorithms including linear regression, random forest, SVM, and Long Short-Term Memory (LSTM) were employed for the estimation of salinity and other parameters. LSTM was utilized here as it provided flexibility and accuracy in its output and was a major factor in the management of freshwater.

[31] had a goal of using machine learning algorithms such as SVR and XGBoost to predict water quality factor and assess the accuracy of those algorithms. The two algorithms were used to forecast nine separate factors with the accuracies in the range of 79% to 99%.

[32] proposed an innovative technique of water quality forecasting which is a Long Short-Term Memory Neural Network (LSTM NN). For training, the monthly mean values of the water quality indicators of Lake Taihu from 2000 to 2006 were used. The method was compared with other techniques and the results showed that the LSTM NN outperformed the Back Propagation Neural Network (BP NN) and Online Sequential Extreme Learning Machine (OS-ELM) in water quality prediction.

[33] focused on the use of AI techniques for the optimization of the water supply and sanitation systems, the control of the water quality standards compliance and the efficient operation of monitoring drinking water for the sustainable, environmentally friendly use. As for the study, the adaptive neuro-fuzzy inference system (ANFIS) algorithm was then used for WQI (water quality index) prediction and feed-forward neural network (FFNN) and KNN (K Nearest Neighbor) algorithms for water quality classification. The results showed that ANFIS model was the most precise in predicting the values of WQI with an accuracy of 96.17%, while the FFNN algorithm completely classified water quality data with 100% accuracy. The research revealed that the advanced AI technique proposed by the research team can immensely help in the activities of water treatment and management.

[34] aimed at employing a very sophisticated AI algorithm to approximate and assess water quality. The study set up models for WQI prediction and classified water quality using the artificial neural networks (NARNET and LSTM) for WQI prediction and SVM, KNN, and NB for WQC prediction. The dataset consisted of seven main

parameters and the models were evaluated according to statistical criteria. The results indicated that NARNET slightly outperformed LSTM in WQI prediction, whereas SVM gained the best accuracy (97.01%) for WQC prediction. Both NARNET and LSTM had almost the same accuracy when testing, with a small difference in regression coefficients.

[35] proposed a feature selection method that combines the weighted entropy and the Pearson correlation coefficient to estimate the water quality. This approach developed the prediction to be more precise and secure by taking into consideration the information content and the feature correlation. They examined various machine learning algorithms for the water quality prediction and found out that SVM was the one that performed well in the DO prediction, while MLP was the one that was successful in the nonlinear modelling. The RF and XGBoost methods were quite weak, but the LSTM method was very good at capturing dynamic patterns.

[36] aimed to predict WQI and Water Quality Classification (WQC) using machine learning models. They optimized parameters for models like RF, XGBoost, and others. Data preprocessing included mean imputation and normalization. The dataset had 7 features and 1991 instances. GB model achieved 99.50% accuracy in WQC prediction, while MLP regressor model had 99.8% R2 for WQI prediction, outperforming other models.

[37] aimed to predict river water quality and categorize the WQI based on water quality standards using machine learning models. Data from eleven sampling stations along the Bhavani River were used, considering 27 parameters. MLP regressor predicted WQI efficiently with a root mean squared error of 2.432, and MLP classifier classified the WQI with 81% accuracy.

[38] utilized machine learning models to predict total dissolved solids (TDS), sodium absorption ratio (SAR), and total hardness (TH) in the Karun River. Models included multiple linear regression (MLR), M5P model tree, support vector regression (SVR), and random forest regression (RFR), with principal component analysis (PCA) for variable reduction. Results showed RFR, SVR, and MLR had the lowest errors in predicting TDS, SAR, and TH, respectively, indicating the effectiveness of machine learning models in water quality prediction.

[39] emphasized the importance of water quality prediction due to water pollution's increasing impact. They developed a model using machine learning algorithms to predict the WQI and quality class based on four parameters: temperature, pH, turbidity, and coliforms. Multiple regression algorithms are effective in predicting WQI, while ANN is the most efficient in classifying water quality.

[40] focused on developing deep learning algorithms to predict WQI and WQC. They used the long LSTM algorithm to predict WQI and a convolutional neural network (CNN) for WQC. The study considered seven water quality parameters: DO, pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. Experimental results demonstrated that LSTM predicted water quality with superior robustness, achieving a 97% accuracy in WQI prediction.

[41] developed a machine learning model using adaptive boosting to evaluate drinking water quality. They used a Kaggle dataset and experimented with different machine learning techniques, finding that their ensemble model achieved 96.4% accuracy, outperforming individual models like LR (88.6%), CHAID (93.1%), XGBoost tree (94.3%), and multi-layer perceptron (95.3%).

Table 1 summarizes the key studies, the algorithms used, the datasets or parameters involved, and the results obtained.

Table 1. Summary of studies on water quality classification.

| Study | Algorithms used | Parameters/Dataset | Best algorithm | Accuracy/Result |
|---|---|---|---|---|
| [25] | ANN, SVM | Tayra River, Iran | SVM | SVM gave the most accurate results |
| [26] | SVM, KNN, NB | Dataset with 7 parameters | SVM | Highest accuracy in water quality prediction |
| [27] | SVM, DT, NB | pH, DO, BOD, EC | DT | DT achieved the highest accuracy |

| Study | Algorithms used | Parameters/Dataset | Best algorithm | Accuracy/Result |
|---|---|---|---|---|
| [28] | DT | - | DT | DT was the most successful algorithm |
| [29] | NB, J48, BAG | Kinta River, Malaysia | NB | NB was the best model |
| [30] | LR, RF, SVM, LSTM | Salton Sea, USA | LSTM | LSTM provided flexible and accurate predictions |
| [31] | SVR, XGBoost | Various water quality factors | - | Success rates ranging from 79% to 99% |
| [32] | LSTM NN | Taihu Lake, China | LSTM NN | Outperformed BP NN and OS-ELM |
| [33] | ANFIS, FFNN, KNN | Water supply and sanitation systems | ANFIS (WQI), FFNN (classification) | ANFIS: 96.17%, FFNN: 100% accuracy |
| [34] | NARNET, LSTM, SVM, KNN, NB | 7 parameters | NARNET (WQI), SVM (WQC) | NARNET outperformed LSTM, SVM achieved 97.01% |
| [35] | SVM, MLP, RF, XGBoost, LSTM | Feature selection with entropy weighting and Pearson correlation | SVM (DO), MLP (nonlinear), LSTM (dynamic patterns) | - |
| [36] | RF, XGBoost, GB, MLP | 7 features, 1991 instances | GB (WQC), MLP (WQI) | GB: 99.50%, MLP: 99.8% R2 |
| [37] | MLP | Bhavani River, India | MLP | RMSE: 2.432, accuracy: 81% |
| [38] | MLR, M5P, SVR, RFR | Karun River, Iran | RFR (TDS), SVR (SAR), MLR (TH) | - |
| [39] | Multiple regression, ANN | 4 parameters: temperature, pH, turbidity, coliforms | ANN | Effective for classifying water quality |
| [40] | LSTM, CNN | 7 parameters: DO, pH, conductivity, BOD, nitrate, fecal coliform, total coliform | LSTM | LSTM: 97% accuracy in WQI prediction |
| [41] | LR, CHAID, XGBoost, MLP, adaptive boosting | Kaggle dataset | Adaptive boosting | 96.4% accuracy |

In general, most previous studies have focused on sea or river water classification, and their impact on drinking water and the importance of variables have not been sufficiently examined. Since drinking water is of vital importance, especially for human life, it is evaluated under more diverse parameters. Therefore, classifying drinking water quality is seen as a more challenging task. Moreover, the target class that studies try to predict is usually based on the WQI value. This study, on the other hand, directly tries to predict the potability or non-potability of water. Furthermore, none of the studies have focused on the effect of outliers on the performance of machine learning algorithms. This study will focus on different outliers processing methods and examine their impact on machine learning algorithms.

## 3. Material and methods

In this section, explanations of the machine learning algorithms used in the research, the performance criteria used in the comparison of the algorithms, the characteristics of the dataset and information about the data preparation process are given.

### *3.1. Algorithms used*

Machine learning is a sub-branch of AI and involves computers making intelligent decisions by learning from data [42]. Machine learning uses various methods depending on the nature of the data and the objectives of the features. Among these methods, three approaches can be singled out such as supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is an approach that has the machine learning model to come to terms with the relationship between the input and the target output [43]. Supervised learning can be classified basically into two types: classification and regression. The purpose of classification is to find out if an input belongs to a given category. Regression deals with the continuous numerical output of a prediction related to the input data. In this research, the context is given to using supervised classification algorithms for water quality assessment. Classification of water quality is done through a physical, chemical, and biological inspection of water samples. Through this assessment, the potential of the water is decided.

Algorithms for supervised classification could title the samples in a short time using the information from the historic data [44]. This is a feature that makes it possible to quickly and correctly classify the water samples collected in the water quality assessment. Water quality assessment plays a critical role in water resource protection and environmental sustainability. Therefore, it is foreseen that using machine learning algorithms will contribute to the creation of a powerful system for the effective assessment of water quality and protection of water resources.

### *3.2. Supervised classification algorithms*

In this study, a total of nine supervised classification algorithms were used for water quality assessment. These algorithms include Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, AdaBoost and Bagging Classifier.

#### *3.2.1. Logistic regression*

LR is a statistical modelling technique and is used to estimate the probability of a two-category dependent variable. LR is a widely used technique, especially in classification problems. LR limits linear regression by using the logit transformation of the probability distribution [45].

The mathematical formula of LR is given in Equation 1. When the equation is analyzed, $P(Y=1)$ represents the probability that the dependent variable is 1. $e$ is the Euler number and the coefficients $b_0$, $b_1$, $b_2$, ... $b_k$ are the parameters estimated by the model. $X_1$, $X_2$, ... $X_k$ are the independent variables.

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_kX_k)}} \quad (1)$$

#### *3.2.2. Decision trees*

DT algorithm, as a machine learning method, is used to classify or regress a dataset by dividing it according to features. The algorithm first creates a tree structure. Each internal node is associated with a feature in the dataset and each leaf node is associated with a class. The tree divides the dataset into homogeneous subsets and makes predictions in this way [46]. DT use various criteria to select features that best summarize the information in the data. These criteria include concepts such as entropy, gain ratio and Gini index [47].

*3.2.3. Random forest*

RF is a machine learning algorithm used to solve classification and regression problems. This algorithm is constructed by combining multiple decision trees. Each decision tree is used to generate randomly selected subsets (bootstrap samples). RF aims to create more reliable and accepted models by taking the prediction of the majority of these trees. By training each tree on different subsets, the RF algorithm reduces overfitting and increases the generalizability of the model [48]. In addition, by averaging the predictions of each tree, the errors of the individual trees are balanced. Mathematically, the prediction formula of the RF algorithm is given in Equation 2.

$$\acute{Y} = \frac{1}{B}\sum_{j=1}^{B} f_i(X) \tag{2}$$

In the formula in Equation 2, Ý is the predicted value, $B$ is the number of trees, $f_i(X)$ is the prediction number $j$ of each tree.

*3.2.4. Gradient boosting decision tree*

Gradient Boosting Decision Tree (XGBoost) creates a powerful forecasting model by successively adding decision trees. The first tree attempts to predict the dataset, and then an error correction process is performed on the errors of this tree, allowing the next tree to focus on them. This process continues by adding a series of trees, each tree correcting previous errors and improving the overall performance of the model. XGBoost learns iteratively, focusing on correcting the errors of the previous trees as each tree is added [49]. The formula of the XGBoost algorithm is given in Equation 3.

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{3}$$

In Equation 3, $F(x)$ is the total prediction model, $M$ is the number of trees, $\gamma_m$ is the learning rate of each tree and $h_m(x)$ is the prediction of each tree.

*3.2.5. Naive bayes*

NB works based on Bayes' Theorem and accepts the assumption of independence between features when classifying. This assumption states that the features are independent of each other, hence the name "naive" [50]. Basically, NB uses the probabilities of features to determine the class to which a data point belongs. Using these probabilities, NB makes predictions for each class and selects the class with the highest probability. The basic mathematical formula of NB is given in Equation 4.

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \tag{4}$$

When the above equation is analyzed:

- $P(C \mid X)$ is the probability that a given class $(C)$ is determined by the given feature set $(X)$,
- $P(X \mid C)$ is the probability that the feature set $(X)$ belongs to the given class $(C)$,
- $P(C)$ is the probability of belonging to a given class $(C)$,

- $P(X)$ is the probability of observing the feature set $(X)$.

*3.2.6. K-Nearest neighbors*

The basic principle of the KNN algorithm is to use the influence of its k-nearest neighbors to determine the class or value of a data point. The algorithm measures the distances of each point in the dataset to each other according to their position in the feature space and then determines the class or value of a given data point by the label or value of its k-nearest neighbor [51]. The KNN algorithm uses measurement metrics such as Euclidean Distance or Manhattan Distance to measure the distances of data points. The user-specified value of k is determined interactively, and it is generally preferred that it is not an even number. The basic mathematical formula of KNN is given in Equation 5.

$$\acute{Y} = \text{argmax} \left( \sum_{i=1}^{k} I(y_i = j) \right) \tag{5}$$

In Equation 5, Ý represents the predicted class, $k$ represents the number of neighbours specified by the user, $y_i$ represents the class of the $i$ -th neighbor, and $j$ represents the class index.

*3.2.7. Support vector machine*

The main purpose of the SVM algorithm is to classify data points into two or more classes using a hyperplane. SVM performs particularly effectively on datasets that cannot be separated linearly. The algorithm selects a hyperplane to classify data points and places this hyperplane in such a way as to maximize the margin between classes [52]. The mathematical formula of the algorithm is expressed in Equation 6.

$$f(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \tag{6}$$

In Equation 6, $f(x)$ represents the estimation function, **w** represents the weight vector, **x** represents the input feature vector, and $b$ represents a constant term or plane.

*3.2.8. AdaBoost*

AdaBoost (Adaptive Boosting) is an ensemble learning algorithm for classification problems. ADA creates a strong classifier by combining weak learners together. Its basic principle is to iteratively strengthen the model by adjusting the weights of the examples in the dataset, focusing on the points where a learner is weak. In each iteration, the weight of the misclassified examples is increased so that the next learner focuses more on these examples [53]. The mathematical formulation of the algorithm is expressed in Equation 7.

$$F(x) = \sum_{t=1}^{T} \alpha_t f_t(x) \tag{7}$$

When Equation 7 is analyzed, $F(x)$ is the total prediction model, $T$ is the number of iterations, is the weight of each learner, $f_t(x)$ is the prediction of each learner.

### *3.2.9. Bagging classifier*

Bagging Classifier (Bootstrap Aggregating) is an ensemble learning algorithm used for classification problems. Its main goal is to create a more powerful and generalizable classifier by aggregating many weak learners trained on different subsets [54]. The algorithm trains each learner using different bootstrap samples (repeated sampling). In this way, learners trained on different samples increase the diversity of the model and reduce the risk of overfitting. BAG is based on weak learners such as SVM or KNN. Since each learner is trained on different subsets of the dataset, it contributes to making the model more general and stable [55].

### *3.3. Data validation method*

In the context of water potability prediction, the k-fold cross-validation method was used. The dataset is divided into k parts and uses each of them as a test set respectively. The remaining k-1 parts are considered the training set. This is done k times, and all the parts are used once as a test set. The model performance is evaluated based on the average of the results.

In this study, the dataset is divided into 5 parts, which is a common practice in k-fold cross-validation where k is set to 5. The fiber count k is which intuitively assures computational efficiency and statistical robustness. When k=5, every fold has a data part that corresponds to 20%. This ensures that all the samples are diverse in all folds while still having enough data to train and test. In addition, 5-fold cross-validation method enables researchers to reliably estimate the model's performance on different subsets of the data and thereby assess its generalization capability. The flowchart of the model designed based on k-fold cross-validation is shown in Figure 1.

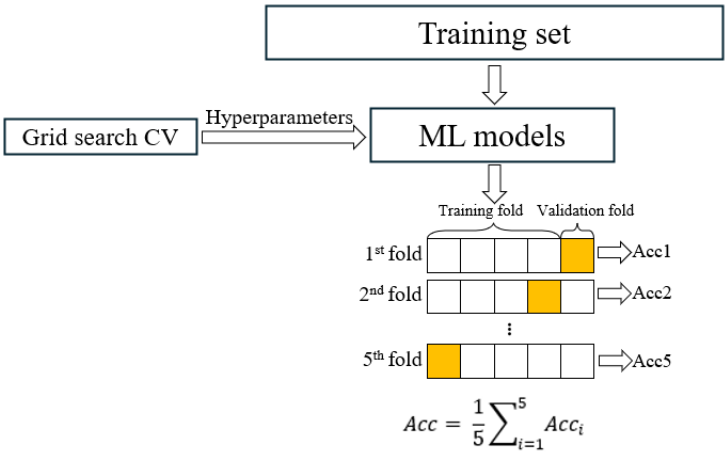


Fig. 1. K-fold cross-validation method.

### *3.4. Performance metrics*

The process of checking the performance of machine learning models is extremely crucial to the process of building and putting models into practice. In this regard, performance metrics are utilized for the consideration and appraisal of model quality regarding the accuracy, precision, sensitivity, and performance of a particular model. The measures that are fundamental such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) provide metrics that help understand the model behavior characteristics and overall quality of model

results respectively. Additionally, techniques such as ROC curves, confusion matrices, and F1-Score can give a different view by combining various metrics. A thorough grasp of this data is of fundamental importance to the fact that the model is not only functioning well but is also very dependable.

### *3.4.1. Accuracy*

Accuracy is the measure that reflects how close the prediction made by a classification model would be to the actual class. Thus, accuracy is the extent of the number of correct decisions to the total decision number [56]. The formula for calculating accuracy is expressed in Equation 8.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

### *3.4.2. Precision*

Precision is a measure of how many of the cases that a classification model predicts as positive are actually true positive. Precision is the ratio of false positive predictions to the total number of positive predictions. A high precision value indicates that the model's positive predictions are reliable, while a low precision value indicates a high rate of false positive predictions. This metric can be expressed using Equation 9 in its formula.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

### *3.4.3. Recall*

Recall is a quality index that is used to find out how well the positive samples have been recognized by the model. Recall is the proportion of true positive predictions and total positive samples. A high recall value points to the fact that the model is very good at detecting positive instances. Recall formula is given in Equation 10.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

### *3.4.4. F-Measure/F1-Score*

The F-measure or F1-score is a gauge utilized to determine the success of a classification model by fusing two measures of precision and recall. The F-measure, which combines these two measures at a balanced point, evaluates how well the model can both minimize the number of false positive predictions and accurately identify the true positives. F-measure is calculated by the following formula:

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{11}$$

### *3.4.5. Receiver operating characteristic curve*

Receiver Operating Characteristic (ROC) Curve a tool in graphical form is which used for assessment of model performance, particularly in classification problems. The ROC curve demonstrates the relationship between the probability of the correct classification (rate of recall) and the probability of the incorrect classification (rate of false positives). This is a means of determining whether the efficiency of the model is governed more by sensitivity or specificity. The ROC curve is a graph showing the effectiveness of the model when using various cut-off values. It goes from the corner where the perfect case is to the line at 45 degrees which is random guessing. A perfect model's ROC curve will rest in the upper left section of the graph, indicating that the true positive rate is high while the false negative rate is low [57]. The ROC curve has two inputs - True Positive Rate (TPR) and False Positive Rate (FPR). True Positive Rate (TPR) is determined the use of the Equation 12 given and whereas for False Positive Rate (FPR) the formula in Equation 13 is provided.

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

### *3.4.6. The area under the curve*

The Area Under the Curve (AUC) can be considered as the measurement of the area under the ROC curve. AUC measures the model performance for various thresholds by looking at the size of the area under the ROC curve [58]. AUC can take on any value from 0 to 1. A high AUC value indicates that the model has a good trade-off between high true positives and low false negatives. This means that the model is quite skilled in classifying the data.

### *3.4.7. Area under the precision-recall curve*

The Area Under the Precision-Recall Curve (AUPRC) is a metric that tells the degree to which a classification model strikes a balance between accuracy and recall. The Precision-Recall curve is showing an image of the model's precision and recall over the entire range of cut-off points. The trick is AUPRC, which is the process of measuring the model's performance at different cut-off points, and this is done by getting the area under the curve [59]. High AUPRC of the model shows its ability to pinpoint true positives as well as false positives which indicates that the learner is capable of detecting true positive predictions as well as true positives. AUPRC is a method that is mainly applied to imbalanced classification problems.

## *3.5. Dataset*

In this research we utilized “Water Quality and Potability” dataset available online, a renowned dataset disseminated in the Kaggle platform. The accuracy of features in the dataset predicting the potability of water was verified by two faculty members working in Departments of Water Science Engineering consulting them. They stated that the samples in the dataset meet the standards of WHO (World Health Organization). The dataset consists of 3276 water samples and 10 features that provide essential information about the water quality parameters. These parameters include:

1. pH Value: Indicates the water's acidity or alkalinity, with values within WHO standards (6.52–6.83).
2. Hardness: Reflects the water's calcium and magnesium salt content, impacting its raw hardness.
3. TDS (Total Dissolved Solids): Reflects the water's mineralization, adhering to WHO limits (desirable: 500 mg/l, maximum: 1000 mg/l).
4. Chloramines: Result from ammonia added to chlorine for water disinfection, with safe levels up to 4 mg/L.
5. Sulfate: Naturally occurring in minerals, with concentrations within the 3–30 mg/L range in freshwater.
6. Conductivity: Reflects water's ion concentration, following WHO's 400 µS/cm limit.
7. Organic Carbon: Measures carbon content in water's organic compounds, complying with US EPA standards (< 2 mg/L).
8. THMs (Trihalomethanes): Found in chlorinated water, with levels within the safe limit of 80 ppm.
9. Turbidity: Reflects solid matter quantity in water, with values meeting WHO's recommended limit (0.98 NTU).
10. Potability: Binary indicator (1 for Potable, 0 for Not Potable), determining water's safety for human consumption.

Figure 2 shows the potability distribution of the water samples in the dataset. There is an imbalance in the target variable in the dataset. This should be taken into account in the modeling process. The imbalanced distribution of the target variable (Potable (1) or Not Potable (0)) requires a careful approach to avoid biasing the model towards the majority class.

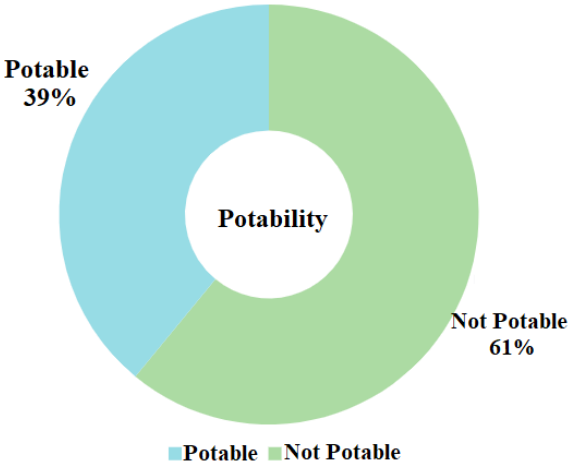


Fig. 2. Potability distribution of water samples in the dataset.

Figure 3 demonstrates the distribution of missing values across different features in the dataset. Notably, a significant number of missing values are observed in the 'ph', 'Sulfate', and 'Trihalomethanes' features. Missing values can negatively affect the accuracy and reliability of the model. Effectively handling these missing values and optimizing the dataset allows the model to learn more robustly and reliably. Solving the missing value problem relies on strategies such as selecting appropriate imputation methods or choosing algorithms that are sensitive to missing values.

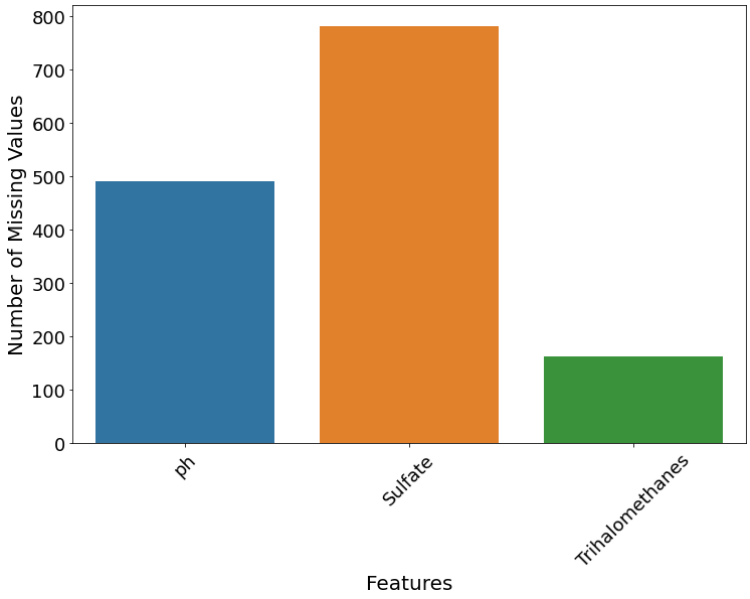


Fig. 3. Missing values in the dataset.

Figure 4 shows the correlation matrix of the dataset. According to the correlation matrix, there is no linear relationship between the features that can explain the target variable. Therefore, linear models may not be effective on this problem. Considering this situation, it would be more appropriate to experiment with probabilistic models. Such models can handle the complexity and relationships in the dataset in a more flexible way, and therefore solve the problem better.

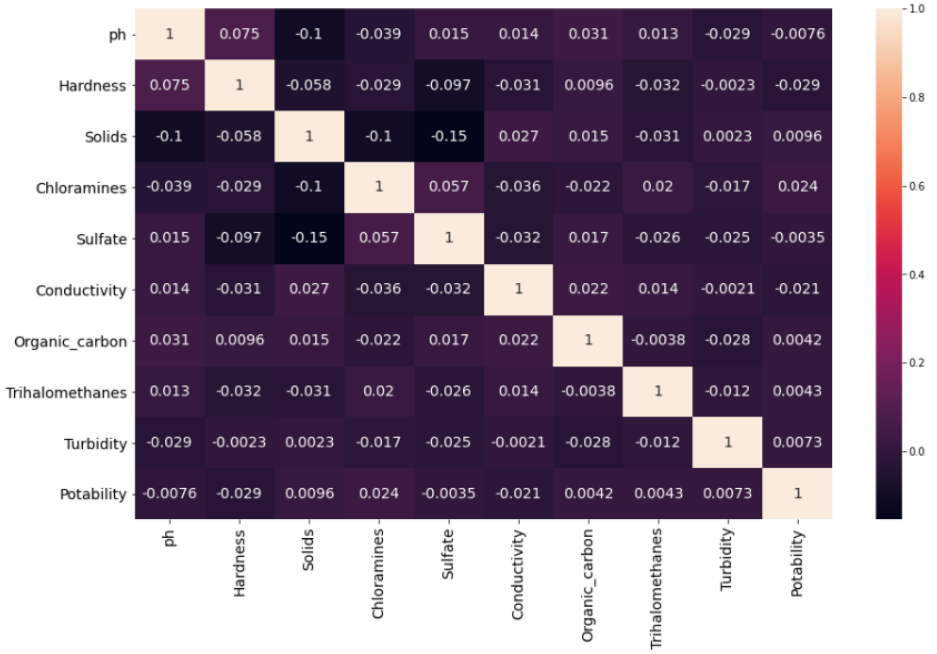


Fig. 4. Correlation matrix.

### *3.6. Data preparation*

Data preparation is a fundamental step in the machine learning process and, when performed correctly, can significantly impact the performance of the model. In this context, three main topics come to the fore: the imbalanced structure of the dataset, missing values and outliers. In the research, the imbalance of the dataset was first addressed and resampling was performed. As the next step, missing values were identified and completed using appropriate imputation techniques. Then, outliers were identified, and solutions were generated with various methods. These three topics are discussed in more detail below.

#### *3.6.1. Class imbalance reduction strategy*

To address class imbalance, instances with values 0 and 1 in the 'Potability' column were treated separately. From the DataFrame of instances with value 0 (“zero”) and a DataFrame of instances with value 1 (“one”), we found two instances. Resampling was then used to increase the instances of the minority class “one” to a larger size. Here, the technique used is to increase the number of instances in the minority class which is then used to balance the classes. Through this step, the samples of the minority class are randomly and accordingly selected and then this process is repeated. Ultimately, the resulting augmented minority class is put together with the majority class so that the new balanced DataFrame (the “df”) can be created. This balancing technique seeks to enhance the model’s performance through the equalization of the learning curves between the classes.

*3.6.2. Review and processing of missing values*

According to features like pH, Sulfate, and Trihalomethanes that were analyzed, missing values were detected. Missing values stand for data values that are not saved for certain observations in a dataset which can cause different problems in statistical analyses and modeling processes. In the case of the dataset being too correlative and the analysis being too confident, missing data can make the dataset incompletable and thus lower the quality, generalizability, and reliability of the results [61]. It can be a cause of such problems as analysis results and modeling accuracy becoming misleading and decision-making being faulted [62]. In turn, proper imputation methods and data preparation techniques should be adopted to effectively cope with missing data.

Since the missing values in the dataset used in this study were covered both classes, these missing values were taken as the population mean. The population mean is a statistical method that shows the sum of the values of a population divided by the population size [63]. Mathematically, the population mean ($\mu$) is expressed as:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{14}$$

This step adds a homogeneity level to the dataset. Thus, making the data more uniform by substituting missing values with population averages of estimates. Also, different statistical techniques for imputing missing data were used. In mean imputation, the absent values were replaced by the average of the whole dataset. Additionally, median and mode are used to finish off the missing entries. In case of median imputation missing values were replaced with the median. The median is the middle number when the numbers are arranged in order. If the mode imputation was employed, the missing values would be replaced by the mode which is the most common value in the dataset. These methods were selected due to the various distributions of the variables. For instance, pH values are typically nearly a normal distribution. Hence, the mean was used for missing values. This is due to the mean being a good representation of central tendency in a normal distribution. However, for variables with more complicated distributions, like Sulfate or Trihalomethanes, it was more appropriate to use the median or mode instead. The median entails the median of the data thus it is more robust than the mean. The mode is the most common value in the dataset and thus shows the most typical position of the distribution. Therefore, it was decided that the mean would be appropriate for pH and the median or mode for Sulfate or Trihalomethanes when filling in missing values. The process of imputing the median with missing values for an $X$ variable can be expressed as shown in Equation 15:

$$\hat{X}_{median} = \text{Median}(X) \tag{15}$$

For a variable $X$ with missing values, the mode imputation process can be expressed as:

$$\hat{X}_{mode} = \text{Mode}(X) \tag{16}$$

In Equation 16 $\hat{X}_{mode}$ is the imputed value and $\text{Mode}(X)$ is the mode of the observed values of $X$.

*3.6.3. Review and processing of outliers*

The very nature of the outlier problem in machine learning may lead to the disrupted functioning of a dataset. According to some researchers, if data with unprocessed outliers is included in the dataset, the model's performance results can drop by more than half [64]. So the detection and treatment of the outliers are of the primary importance of the model building process.

Trimming and Standard Deviation methods were used to remove outliers. Trimming involves removing a certain percentage of outliers to improve the accuracy of the results [65]. Standard Deviation is a statistical technique that shows how closely the values of a dataset are to the average. Outlier values correspond to the data beyond a specified standard deviation [66]. In the research, a few features that were outlier indicators, for example, 5% trimming was applied for 'Sulfate'. In this manner, 2.5% values were added up to both extremes and the rest of the data, which was 95%, were used. More so, outlier computation was done using the deviation calculation method for the 'hardness' feature in the dataset. The threshold value was determined by the average hardness value and adding two times the standard deviation for the two values to finish the process.

The second treatment for outliers in this research is the use of IQR (Inter-Quartile Range) method. As per the method, the real data of the highest and lowest values of the dataset are replaced with a certain threshold value, thus, the extreme values will not have a great influence on the model performance [67]. So, according to the IQR method, the data points that are outside the normal range are determined by setting a certain threshold value. This threshold is usually associated with a certain percentile. After that, all the values beyond the threshold are replaced with this threshold value. It is worth noting the fact that in analyses of water quality, the pH values are restricted to a specific range limiting the extreme values to a certain threshold. Thus, it is aimed to increase the certainty of the results by minimizing the probability of extreme values affecting the model's performance. While examining the pH values in the original dataset, a specific threshold value (pH = 9) was chosen and IQR was applied.

The percentile method was also applied to deal with outliers. This method is used to detect outliers by setting a threshold value at a certain percentile [68]. Here, the threshold value may be set at the range of 2%-98%, thus, any value beyond that will be regarded as a misfit. Nevertheless, the application of such a method produced some obstacles in doing the experiment, which resulted in not getting the expected correct results. The method of the percentile requires human instruction to set a particular number and set the limit. Nevertheless, the dataset is complicated and has had different values over time, which makes it hard to find the appropriate threshold value. The disparity in the dataset is illustrated by the water samples drawn from diverse geographical locations. These samples coming from urban water treatment systems to natural water sources in grouped areas are as such. The fact that the water properties are different makes it more difficult to study the effect of the factors involved. An example is the pH of the water in the city which is different from water in the rural area. The complexity is that every dataset feature is a result of many factors which differ in their independence. Hardness of water can be a mix-up of the geographical and the seasonal variability. This complexity makes the assignment of the threshold value of each characteristic more difficult, as the factors influencing them are expected to vary more widely. Furthermore, the percentile method uses a specific percentile range when identifying outliers. However, the threshold value set in the dataset, especially on the TDS (Total Dissolved Solids) value, reflects the mineralization of the water samples. Geographical differences of water sources and intended uses can cause significant variations in TDS values. Natural water sources in rural areas may have higher mineral content, while urban water supply systems may have lower levels of mineralization. This means that although the TDS threshold value is in accordance with generally established WHO standards, it may not be fully compatible with the specific characteristics of water samples from geographical areas. Therefore, it can be difficult to achieve an accurate agreement in defining or interpreting the TDS value as an outlier in water samples from specific geographical areas.

## 4. Experimental study and findings

In this study, nine supervised classification algorithms including LR, DT, RF, XGBoost, NB, KNN, SVM, ADA and BAG were used to determine the potability of water. GridSearchCV hyperparameter optimization technique was used to determine the hyperparameters of the models. Hyperparameters are manually determined parameters that affect the performance of the model. GridSearchCV is one tool for hyperparameter tuning. It is a systematic way to tune hyperparameters so that machine learning models can be better able to perform. This hyperparameter optimization process consists of certain steps to obtain the best performance of the machine learning models [69].

First, we specify the hyperparameters to be optimized along with a set of values for them. Next, the combination of hyperparameter values is generating the different models and then all the models are evaluated using a predefined metric. The selection of the best performing combination of hyperparameter values is made through a process of trying out different options and sorting out the models. A last model is constructed using the optimal hyperparameter values, and this model is now ready to be utilized on a more extensive dataset. Cross-validation (CV) is the technique used in GridSearchCV, facilitating the model's being tested on different portions of the data, thus improving its generalizability [70]. This process makes the model aware of such tendencies as over-fitting and under-fitting thereby providing reliable and superior results.

While creating the model, the dataset was divided into two parts, 75% for training and 25% for testing. In all algorithms used, the random state was set to 42. The number of trees in RF was defined as 9. The number of neighbors in KNN was set as 5. In the XGBoost algorithm, the learning rate was set as 0.01, the number of predictors as 8 and the number of seeds as 25. The kernel used in SVM is Radial Basis Function (RBF) and C parameter is set as 2. In SVM optimization, the C parameter indicates the extent to which misclassification of each training sample is prevented. For the ADA algorithm, 100 weak learners and a learning rate of 0.1 were used. For the BAG algorithm, 50 base classifiers and sampling strategies were determined. Figure 5 shows the confusion matrices of the algorithms used in the research.

When Figure 5 is analyzed, confusion matrices are used to examine the performance of each algorithm in detail. The BAG algorithm achieved the highest success in correctly identifying the potability status of water, exhibiting the highest True Negative value. On the other hand, the DT algorithm showed a significant performance in correctly classifying the potential risks related to the potability of the water, achieving the highest True Positive value. The LR algorithm has higher False Negative values. This indicates that it performs poorly in classifying the potability of water. Concerning the RF algorithm, it exhibits a well-balanced performance, effectively identifying both potable and non-potable water samples. The RF algorithm's contributions to True Positive and True Negative values contribute to its overall success in classifying water samples.


Fig. 5. Confusion matrices (based on trimming method).

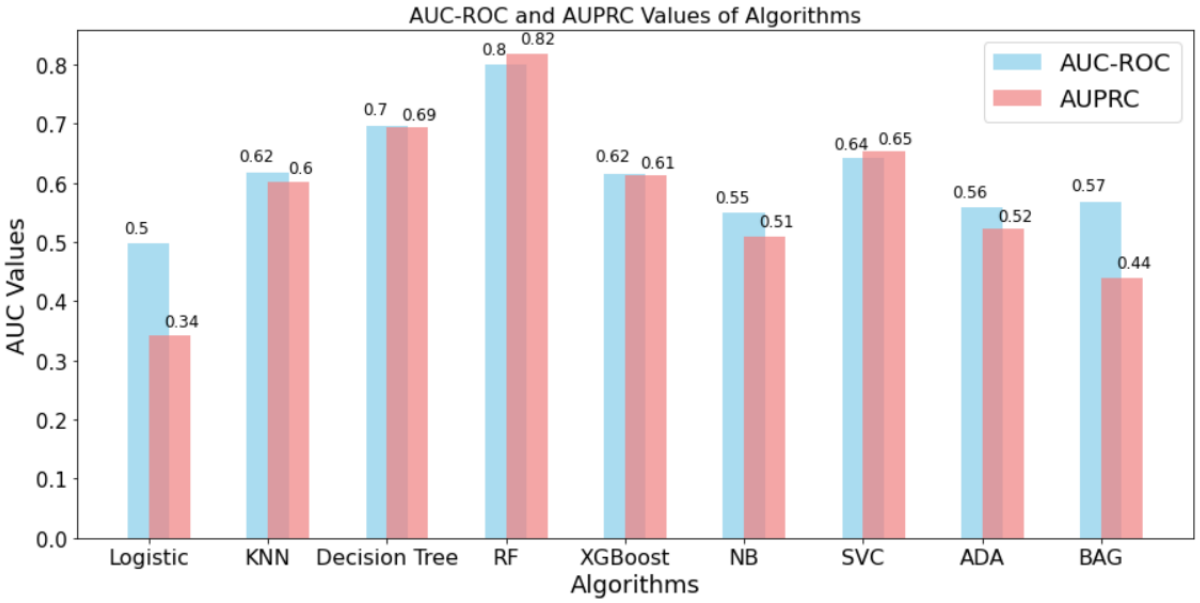


Fig. 6. AUC-ROC and AUPRC values of algorithms (based on trimming method).

Figure 6 shows the AUC-ROC and AUPRC scores of the water potability detection system that was used machine learning. These scores are very important benchmarks to be used for the assessment of the quality of classification models. AUC-ROC is the computation of the trade-off between the false positive rate and the true positive rate and a representation of the capacity of the model to be able to correctly classify non-potable water and at the same time having minimal false positives. AUC-ROC is a key metric with a higher score than the others indicative of better performance classification [71]. Moreover, AUPRC is a metric that investigates the relationship between precision and recall and it is particularly important for the case of the analysis of classification problems with imbalanced data. A high AUPRC score implies that the model can detect the non-potable water samples successfully and thus, it will reduce the probability of false positives [72].

The graph shows the obvious result that the RF algorithm has the highest AUC-ROC (0.80) and AUPRC (0.82) scores and the other algorithms have lower ones. In this instance, it can be stated that the RF algorithm is the top one in both accurately classifying potable water cases and the non-potable ones. In the same way, the DT algorithm is the one that has the most outstanding ability in demonstrating the exact results of the water samples with the AUC-ROC (0.70) and AUPRC (0.69) scores. On the other hand, the LR algorithm is the one that has some possible drawbacks in classifying the water samples correctly, as it has lower AUC-ROC (0.50) and AUPRC (0.34) scores than the other algorithms.

Table 2. Performance metrics according to the outlier treatment methods.

| Classifier | Accuracy | | | Precision | | | Recall | | |
|---|---|---|---|---|---|---|---|---|---|
| | Trim. | IQR | Per. | Trim. | IQR | Per. | Trim | IQR | Per. |
| LR | 0.62 | 0.47 | 0.51 | 0.30 | 0.47 | 0.40 | 0.01 | 0.49 | 0.25 |
| KNN | 0.66 | 0.65 | 0.58 | 0.54 | 0.65 | 0.55 | 0.46 | 0.65 | 0.45 |
| DT | 0.72 | 0.74 | 0.65 | 0.56 | 0.70 | 0.68 | 0.73 | 0.82 | 0.50 |
| RF | **0.83** | 0.80 | 0.75 | **0.84** | 0.81 | 0.72 | 0.67 | **0.79** | 0.75 |
| XGBoost | 0.69 | 0.75 | 0.54 | 0.65 | 0.74 | 0.60 | 0.34 | 0.78 | 0.50 |
| NB | 0.63 | 0.53 | 0.51 | 0.50 | 0.54 | 0.40 | 0.23 | 0.42 | 0.38 |
| SVC | 0.71 | 0.67 | 0.60 | 0.70 | 0.66 | 0.55 | 0.38 | 0.66 | 0.60 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ADA | 0.63 | 0.56 | 0.52 | 0.49 | 0.56 | 0.45 | 0.30 | 0.56 | 0.38 |
| BAG | 0.64 | 0.77 | 0.53 | 0.62 | 0.75 | 0.65 | 0.03 | **0.80** | 0.52 |

Table 3. Performance metrics according to the outlier treatment methods.

| Classifier | F1-Score | | | AUC-ROC | | | AUPRC | | |
|---|---|---|---|---|---|---|---|---|---|
| | Trim | IQR | Per. | Trim | IQR | Per. | Trim | IQR | Per. |
| LR | 0.02 | 0.48 | 0.45 | 0.50 | 0.47 | 0.30 | 0.34 | 0.48 | 0.42 |
| KNN | 0.49 | 0.65 | 0.50 | 0.62 | 0.65 | 0.55 | 0.60 | 0.60 | 0.53 |
| DT | 0.63 | 0.75 | 0.60 | 0.70 | 0.74 | 0.68 | 0.69 | 0.66 | 0.58 |
| RF | 0.75 | **0.80** | 0.70 | **0.80** | **0.80** | 0.72 | **0.82** | 0.74 | 0.65 |
| XGBoost | 0.44 | 0.76 | 0.45 | 0.62 | 0.75 | 0.60 | 0.61 | 0.69 | 0.55 |
| NB | 0.32 | 0.47 | 0.48 | 0.55 | 0.53 | 0.35 | 0.51 | 0.51 | 0.41 |
| SVC | 0.49 | 0.66 | 0.55 | 0.64 | 0.66 | 0.55 | 0.65 | 0.61 | 0.50 |
| ADA | 0.37 | 0.56 | 0.45 | 0.56 | 0.56 | 0.40 | 0.52 | 0.53 | 0.42 |
| BAG | 0.05 | 0.77 | 0.48 | 0.57 | 0.77 | 0.60 | 0.44 | 0.70 | 0.51 |

Table 2 and Table 3 demonstrates the effects of machine learning algorithms for evaluating water quality and its potability. Various outlier treatment methods were applied for the calculation of the performance of the algorithms. “Trimming” refers to the process of clearing the dataset usually by deleting the points from the lowest and highest percentiles. The “IQR” method is the process of limiting the outliers in the dataset and bringing the values above or below a certain threshold to that threshold. The method of the “Percentile” is the process of the ranking of the values in the dataset according to certain percentiles and intervening on values that are either within or outside a certain percentile. In particular, the RF and DT algorithms have strong potential in this area. RF was the most successful algorithm with an accuracy score of 0.83 with the trimming method. According to a precision score of 0.84, the recall is 0.67 (F1-Score = 0.75). The trimming method's recall of the RF algorithm is 0.67 while IQR has a score of 0.79. Because the trimming technique is mainly concerned with low values, the model gets more sensitive to low values and thus has a smaller recall. A lower recall means that occurrences, particularly in the positive class, are found with a success rate that is lower. These scores show the skill of the RF in the differentiation of the potable and non-potable water samples. The model was confirmed to be effective with respect to the classification task and prediction reliability by high values of precision and recall even for imbalanced datasets. DT was the second-best algorithm, thus, via IQR method, the accuracy score of 0.74 was reached. The DT algorithm, which was used to classify water samples correctly, recorded the performance which was represented by a precision score of 0.70 a recall score of 0.82 and an F1-Score of 0.75. The RF algorithm demonstrated high competence in the water potability detection task when the trimming method was applied. The IQR method increased the recall value; however, a balanced adjustment may be needed to increase this value further. Both IQR and trimming methods are efficient methods that can be used to control the outliers that degrade the performance of the model. These methods are important in balancing the model's sensitivity to certain values and its success rate. When the Percentile method was applied, a decrease in the overall performance of the algorithms was observed.

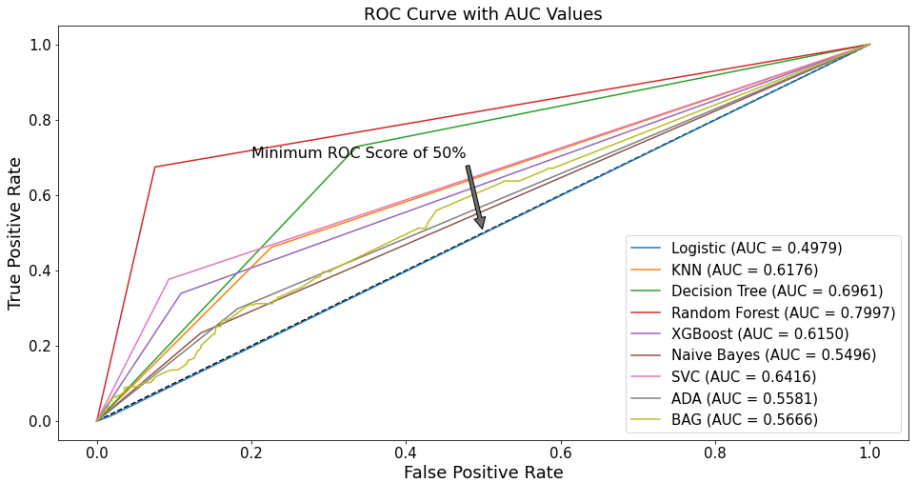


Fig. 7. ROC curve with AUC values (based on trimming method).

When the ROC curve plot is analyzed, it can be seen that certain classifiers perform strongly. The ROC curve provides a visual representation of how a model's sensitivity (true positive rate) and specificity (true negative rate) change at various thresholds. High AUC values indicate the ability of the models to effectively discriminate between different classes [73]. In this context, RF is the algorithm that stands out with a significant difference compared to other algorithms (AUC = 0.80). RF algorithm has demonstrated its competence in data classification by obtaining a very high AUC score. The AUC score obtained emphasizes the effectiveness in classifying the potability of water while maintaining the balance between precision and recall.

## 5. Discussion

This study compares the performance of various machine learning algorithms to determine water potability status. When the literature is examined, the target class that is tried to be estimated is usually on the WQI value. While many studies aim to estimate the WQI value using machine learning algorithms, this research addresses the practical need to classify water samples as potable or non-potable, which is crucial for ensuring safe drinking water. Furthermore, the multitude of parameters defining water potability, the variability of datasets, and the difficulty in comparing findings across different datasets add complexity to water quality research. This study's unique contribution lies in its use of a dataset that, to our knowledge, has not been previously employed in the literature. The dataset's novelty presents both opportunities and challenges. On one hand, it allows for the exploration of water quality classification using diverse parameters specific to the dataset, potentially uncovering insights that may not be evident in studies using different datasets. On the other hand, this uniqueness makes direct comparisons with other studies challenging, as each dataset may have its own characteristics and biases. Although studies directly focusing on binary classification of water potability may be limited, comparisons have been made between the findings of this study and existing literature based on specific features.

In this study, it was found that algorithms such as RF and DT are effective in accurately determining the potability status of water. A study by [74] suggests that Gradient Boosting algorithms are effective for a similar water quality analysis. Which algorithm to choose may vary depending on the intended use in a particular context, the characteristics of the dataset, and the problem domain to be solved. Algorithms such as RF and DT can achieve successful results, especially on high-dimensional and complex datasets, while Gradient Boosting algorithms can

learn more complex relationships [75]. Therefore, a decision should be made by considering the requirements of the model to be applied and the characteristics of the analyzed water samples.

When the findings from other studies focusing on water quality analysis in different geographical regions are examined, a study conducted in South America showed high accuracy of the KNN algorithm in determining the potability of water [76]. Geographical factors such as climatic conditions, rainfall, type of water sources are important parameters affecting water quality. Therefore, it is important to evaluate whether an algorithm shows the same success in another geographical region. Furthermore, the characteristics of the dataset used to evaluate differences in water quality analysis performance between geographical regions are important [77]. Factors such as regional characteristics, minerals in the water, pollution levels, etc. can affect the success of the analysis algorithms. Furthermore, geographic region-specific parameter settings may be required to optimize the success of the algorithm.

[30] used LSTM to estimate salinity in Salton Lake, emphasizing its accuracy and flexibility. Nevertheless, this research revealed that RF was the most efficient algorithm for the purpose of classifying drinking water. The difference in the observed results is due to the peculiarities of the datasets and the characteristics of the variables used. Chawla et al. concentrated on salinity, which may have different predictors than water potability. Exactly the same, [31] and [32] succeeded in high rates using SVR, XGBoost, and LSTM to forecast the water quality factors. In this work, the results of the SVC and XGBoost were low, however, the RF outcome was high. The different performance of SVC, XGBoost, and RF algorithms in this research and the studies [31] and [32] may be due to several things. Initially, the datasets used in each research may have had differences in size, complexity, and feature selection. RF is recognized for its strong performance on high-dimensional and complex datasets as well as noisy features, which might be the case in this research. In contrast, SVC and XGBoost may be more appropriate for the datasets for which they obtained the highest success rates in [31] and [32]. In addition, the differences in the hyperparameters selected for each algorithm, data preprocessing methods, and model evaluation techniques may have also played a role in the performance differences. RF's performance on such a diverse array of feature types and its strong generalization capacity have likely made it outperform SVC and XGBoost in this study. It should be emphasized that the performance of machine learning methods can be different in relation to the particular features of the dataset and the problem being tackled.

In the case of the study at hand, just like in [39], it can be noted that ensemble methods like Random Forest (RF) were found to be bieng excellent for drinking water quality prediction. This is a good example of the effectiveness of ensemble models for several reasons. One of the advantages of ensemble models is that they make the predictions more stable and reduce overfitting by combining several base learners, thus, they are suitable for complex datasets with non-linear relationships, e.g. water quality data. The ensemble models' high performance might come from the fact that they are capable of capturing the diverse and sometimes conflicting patterns that are present in the data. Through pooling the predictions of multiple models, ensemble models can successfully reduce both bias and variance and consequently, get more accurate and precise predictions. Also, ensemble models are able to deal with the different kinds of features and data distributions, thus, they are flexible and can be used for different datasets. Here, the ensemble model may have combined the information from temperature, pH, turbidity, and coliforms which consequently led to a more precise water quality prediction. Furthermore, the effectiveness of ensemble models can be related to some characteristics of the dataset and the problem domain. The dataset used in the current research may have had intricate relationships and patterns that are suitable for community modeling.

This research also analyzed the impact of outlier treatment methods (Trimming, IQR, Percentile) on performance of the algorithms. The best results of 83% accuracy were achieved via the RF algorithm implemented with the trimming method. This means that the RF algorithm is an expert in differentiating between samples of water that can be consumed and those that cannot be consumed. Meanwhile, it was noted that the recall value of the trimming method increased from 67% to 79%. The reason being, the trimming method mainly excludes a few values, hence the model becomes more sensitive to low values resulting in a lower recall value. The DT algorithm had 74% accuracy with the IQR method, but higher accuracy with the trimming method. In these cases, the IQR and trimming

methods obviously controlled the effect of outliers on the algorithm's performance. Nevertheless, the algorithms performance on the percentile method was decreased when the method was applied. The reason for this phenomenon is that assigning ranks to the values in the dataset according to some percentiles and intervening according to this is less effective than other methods.

A study analyzing the influence of outlier handling methods on algorithms shows that specifically, trimming can increase the performance by increasing the sensitivity of the model to low values [78]. This also means that trim can reveal the presence of crucial components at low concentrations of water, especially in the quality analysis of water. However, the high recall value with the IQR method may ignore critical features at low concentrations in the case of water quality analysis [79]. Consequently, the application of the IQR method should be exercised with care measuring the characteristics to be analyzed. The clear performance deterioration of the Percentile method agrees with some results in the literature [80]. Researchers have pointed out that this approach usually corrupts the overall configuration of a dataset and not so useful for complex datasets. In other words, applying the percentile method may result in the natural structure, distribution, or pattern of the dataset being changed. This is possible when the percentile method does its magic by ranking the values in the dataset according to some percentiles. Taking this into account, this method frequently has the tendency of limiting or modifying the values in the dataset to a certain percentile ranking. Disassembly of data in this way can have unintended consequences, particularly when it is necessary to grasp or keep intact the intricate structures that are part of the dataset. In case of the water potability analysis, applying concentrating or diluting of water properties to a certain percentile may produce results that are less relevant to real-world conditions.

## 6. Conclusion

This research was conducted to evaluate the performance of nine different machine learning algorithms in determining the potability of water. Through the study of the accuracy of the algorithms in water sample potability detection, it became evident that the purpose of the study was fulfilled successfully.

Through data preparation stages, the procedures performed in this study are the main steps to ensure reliable and effective training of machine learning models. Taking care of the issue of class imbalance, deciding on how to treat missing values, cleaning potential mistakes in the dataset and properly treating outliers were pointed out as main strategic decisions to augment the accuracy of analysis and modeling. To balance the dataset's classes, two sets of samples with different data values (0 and 1) for the 'Potability' column were treated separately. Resampling was utilized to magnify the number of instances in the minority category, and the augmented minority group was coupled with the majority group to make a balanced Data Frame.

Analysis of missing values was carried out to find out the missing values of critical features such as pH, sulfate, and trihalomethanes. For the gaps in this regard, statistical imputation techniques chosen according to how each variable was distributed were used. For instance, the pH values were normally distributed so the mean was used as a central measure. On the other hand, the commonly used median and mode were more appropriate in the case of variables with complex distributions like sulfate or trihalomethanes. In these situations, the median is the value in the middle and the mode is the value that is the most frequent, thus giving more reliable solutions to the mean. The mean was used for pH and the mode or median for sulfate or trihalomethanes. They were successfully tackled this way which consequently led to model reliability improvement.

The process of outlier examination and treatment is the crucial one that is related to the irregularities of the dataset removal. The methods of Trimming, IQR, Standard Deviation, and Percentile were used to remove outliers. In this way, the machine learning model is less likely to be influenced by the outliers thus, the model becomes more generalizable. The dataset received through the data preparation steps provided the machine learning algorithms that were used for the determination of water potassium successful training.

The RF and DT algorithms were discovered to be efficient in potable water detection as a result of the analysis. The RF algorithm was the best method in terms of accuracy and AUC-ROC scores that cutting and trimming were

used for potable water and non-potable water sample classification. DT algorithm has been successful with IQR outlier processing, but the result is not as high as RF. In contrast, the Percentile method often resulted in inferior algorithm performance. The importance of these outcomes shows the potential of machine learning algorithms to be able to distinguish between potable and non-potable water reliably and accurately, as well as identify the outlier treatment methods used in pre-processing of the dataset. Moreover, it was revealed that the algorithms' performance can be distorted by the geographical factors when the comparison is made with other studies in the literature and in the other geographical regions. These are the issues that illustrate the effects of geographical differences in water quality analysis and should be considered in model selection.

The results give crucial data for people who make decisions on the preservation of water resources and the management of water quality. It is suggested that future studies should carry out such analyses in different regions of the world, and the performance of the various algorithms should be compared in more detail and the use of more advanced machine learning models for the water quality analysis should be developed. Moreover, more research needs to be carried out on issues like the efficient handling of the missing values, the inclusion of the imbalanced distribution of the dataset, and the further investigation of the model's sensitivity to the geographical differences. Such recommendations would bring about more stable and transferable results of machine learning-based methods in the field of water quality analysis.

## Acknowledgements

The study did not receive specific financing from any grant agencies in the public, commercial, or non-profit sectors.

## References

[1] X. Wen et al., "Microbial indicators and their use for monitoring drinking water quality—A review," *Sustainability*, vol. 12, no. 6, pp. 2249, 2020.
[2] S. E. Hrudey and E. J. Hrudey, *Safe Drinking Water*. IWA publishing, 2004.
[3] W. J. Cosgrove and D. P. Loucks, "Water management: Current and future challenges and research directions," *Water Resources Research*, vol. 51, no. 6, pp. 4823-4839, 2015.
[4] H. G. Peterson, "Rural drinking water and waterborne illness," *Saskatoon, SK: Safe Drinking Water Foundation*, pp. 162-91, 2001.
[5] T. Russo, K. Alfredo, and J. Fisher, "Sustainable water management in urban, agricultural, and natural systems," *Water*, vol. 6, no. 12, pp. 3934-3956, 2014.
[6] S. A. Esrey, "Water, waste, and well-being: a multicountry study," *American Journal of Epidemiology*, vol. 143, no. 6, pp. 608-623, 1996.
[7] World Health Organization**,** "Guidelines for drinking-water quality (Vol. 1)," *World Health Organization*, 2004.
[8] J. DeZuane, *Handbook of Drinking Water Quality*, John Wiley & Sons, 1997.
[9] S. J. Kulkarni, "A review on research and studies on dissolved oxygen and its affecting parameters," *International Journal of Research and Review*, vol. 3, no. 8, pp. 18-22, 2016.
[10] C. Jingsheng, Y. Tao, and E. Ongley, "Influence of high levels of total suspended solids on measurement of COD and BOD in the Yellow River, China," *Environmental Monitoring and Assessment*, vol. 116, pp. 321-334, 2006.
[11] S. Morais, F. G. Costa, and M. D. L. Pereira, "Heavy metals and human health," *Environmental Health–Emerging Issues and Practice*, vol. 10, no. 1, pp. 227-245, 2012.
[12] A. K. Singh and R. Chandra, "Pollutants released from the pulp paper industry: Aquatic toxicity and their health hazards," *Aquatic Toxicology*, vol. 211, pp. 202-216, 2019.
[13] P. Nannipieri, S. Greco, and B. Ceccanti, "Ecological significance of the biological activity in soil," *Soil Biochemistry*, pp. 293-356, 2017.
[14] D. Eisma, *Suspended Matter in the Aquatic Environment*, Springer Science & Business Media, 2012.
[15] S. Some, R. Mondal, D. Mitra, D. Jain, D. Verma, and S. Das, "Microbial pollution of water with special reference to coliform bacteria and their nexus with environment," *Energy Nexus*, vol. 1, pp. 100008, 2021.
[16] I. Delpla, A. V. Jung, E. Baures, M. Clement, and O. Thomas, "Impacts of climate change on surface water quality in relation to drinking water production," *Environment International*, vol. 35, no. 8, pp. 1225-1233, 2009.
[17] T. Dube, O. Mutanga, K. Seutloali, S. Adelabu, and C. Shoko, "Water quality monitoring in sub-Saharan African lakes: a review of remote sensing applications," *African Journal of Aquatic Science*, vol. 40, no. 1, pp. 1-7, 2015.
[18] D. T. E. Hunt and A. L. Wilson, *The Chemical Analysis of Water: General Principles and Techniques (Vol. 2)*, Royal Society of Chemistry, 1986.

[19] C. E. Hatch, A. T. Fisher, J. S. Revenaugh, J. Constantz, and C. Ruehl, "Quantifying surface water–groundwater interactions using time series analysis of streambed thermal records: Method development," *Water Resources Research*, vol. 42, no. 10, pp. 1-14, 2006.
[20] I. Yaroshenko et al., "Real-time water quality monitoring with chemical sensors," *Sensors*, vol. 20, no. 12, pp. 3432, 2020.
[21] H. B. Glasgow, J. M. Burkholder, R. E. Reed, A. J. Lewitus, and J. E. Kleinman, "Real-time remote monitoring of water quality: A review of current applications, and advancements in sensor, telemetry, and computing technologies," *Journal of Experimental Marine Biology and Ecology*, vol. 300, no. 1-2, pp. 409-448, 2004.
[22] K. T. Peterson, V. Sagan, P. Sidike, E. A. Hasenmueller, J. J. Sloan, and J. H. Knouft, "Machine learning-based ensemble prediction of water-quality variables using feature-level and decision-level fusion with proximal remote sensing," *Photogrammetric Engineering & Remote Sensing*, vol. 85, no. 4, pp. 269-280, 2019.
[23] L. F. Arias-Rodriguez et al., "Integration of Remote Sensing and Mexican Water Quality Monitoring System Using an Extreme Learning Machine," *Sensors*, vol. 21, no. 12, pp. 4118, 2021.
[24] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1-20, 2017.
[25] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. Garc´ıa-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, pp. 2210, 2019.
[26] S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S Kateb, "Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)," *Applied Water Science*, vol. 11, no. 12, pp. 190, 2021.
[27] J. P. Nair and M. S. Vijaya, "Predictive models for river water quality using machine learning and big data techniques - a Survey," in *Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, Coimbatore, India, March 2021.
[28] M. M. Hassan, M. M. Hassan, L. Akter et al., "Efficient prediction of water quality index (WQI) using machine learning algorithms," *Human-Centric Intelligent Systems*, vol. 1, no. 3-4, pp. 86–97, 2021.
[29] B. Charbuty and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.
[30] P. Chawla, X. Cao, Y. Fu, C. M. Hu, M. Wang, S. Wang, and J. Z. Gao, "Water quality prediction of Salton Sea using machine learning and big data techniques," *Int. J. Environ. Anal. Chem.*, vol. 103, no. 18, pp. 6835–6858, 2023.
[31] K. Joslyn, "Water quality factor prediction using supervised machine learning," *REU Final Reports*, vol. 6, 2018.
[32] Y. Wang, J. Zhou, K. Chen, Y. Wang, and L. Liu, "Water quality prediction method based on LSTM neural network," in *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Nov. 2017, pp. 1-5.
[33] M. Hmoud Al-Adhaileh and F. Waselallah Alsaade, "Modelling and prediction of water quality by using artificial intelligence," *Sustainability*, vol. 13, no. 8, pp. 4259, 2021.
[34] T. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water quality prediction using artificial intelligence algorithms," *Applied Bionics and Biomechanics*, 2020.
[35] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang, "Water quality prediction based on machine learning and comprehensive weighting methods," *Entropy*, vol. 25, no. 8, pp. 1186, 2023.
[36] M. Y. Shams, A. M. Elshewey, E. S. M. El-kenawy, A. Ibrahim, F. M. Talaat, and Z. Tarek, "Water quality prediction using machine learning models based on grid search method," *Multimedia Tools and Applications*, pp. 1-28, 2023.
[37] J. P. Nair and M. S. Vijaya, "River water quality prediction and index classification using machine learning," *Journal of Physics: Conference Series*, vol. 2325, no. 1, pp. 012011, Aug. 2022.
[38] A. Nouraki, M. Alavi, M. Golabi, and M. Albaji, "Prediction of water quality parameters using machine learning models: A case study of the Karun River, Iran," *Environmental Science and Pollution Research*, vol. 28, no. 40, pp. 57060-57072, 2021.
[39] M. Azrour, J. Mabrouki, G. Fattah, et al., "Machine learning algorithms for efficient water quality prediction," *Model. Earth Syst. Environ.*, vol. 8, pp. 2793-2801, 2022.
[40] S. Dharshini, "Deep learning approach for prediction and classification of potable water," *Analytical Sciences*, vol. 39, pp. 1179-1189, 2023.
[41] S. Dalal, E. M. Onyema, C. A. T. Romero, L. C. Ndufeiya-Kumasi, D. C. Maryann, A. J. Nnedimkpa, and T. K. Bhatia, "Machine learning-based forecasting of potability of drinking water through adaptive boosting model," *Open Chemistry*, vol. 20, no. 1, pp. 816-828, 2022.
[42] Z. H. Zhou, *Machine Learning*. Springer Nature, 2021.
[43] V. Sinap, "Prediction of Counter-Strike: Global Offensive round results with machine learning techniques," *Journal of Intelligent Systems: Theory and Applications*, vol. 6, no. 2, pp. 119-129, 2023, doi: 10.38016/jista.1235031.
[44] S. Keskin, O. Sevli, and E. Okatan, "Comparative analysis of the classification of recyclable wastes," *Journal of Scientific Reports-A*, vol. 055, pp. 70-79, 2023.
[45] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 197-200, 1992.
[46] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature Biotechnology*, vol. 26, no. 9, pp. 1011-1013, 2008.
[47] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease," *Design Automation for Embedded Systems*, vol. 22, pp. 225-242, 2018.
[48] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31-39, 2017.
[49] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.
[50] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive bayes: aggregating one-dependence estimators," *Machine Learning*, vol. 58, pp. 5-24, 2005.
[51] L. E. Peterson, "K-Nearest neighbor," *Scholarpedia*, vol. 4, no. 2, pp. 1883, 2009.

[52] H. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 1, no. 10, pp. 185-189, 2012.
[53] A. Taherkhani, G. Cosma, and T. M. McGinnity, "AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning," *Neurocomputing*, vol. 404, pp. 351-366, 2020.
[54] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.
[55] X. Zhu, C. Bao, and W. Qiu, "Bagging very weak learners with lazy local learning," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1-4.
[56] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412-424, 2000.
[57] N. R. Cook, "Use and misuse of the receiver operating characteristic curve in risk prediction," *Circulation*, vol. 115, no. 7, pp. 928-935, 2007.
[58] J. Myerson, L. Green, and M. Warusawitharana, "Area under the curve as a measure of discounting," *Journal of the Experimental Analysis of Behavior*, vol. 76, no. 2, pp. 235-243, 2001.
[59] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, 2013, pp. 451-466.
[60] Kaggle, *Water Quality and Potability*, 2021 [Online]. Available: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability.
[61] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, vol. 793. John Wiley & Sons, 2019.
[62] T. D. Pigott, "A review of methods for missing data," *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353-383, 2001.
[63] G. Rose and S. Day, "The population mean predicts the number of deviant individuals," *BMJ: British Medical Journal*, vol. 301, no. 6759, pp. 1031, 1990.
[64] R. K. Pearson, "Outliers in process modeling and identification," *IEEE Transactions on Control Systems Technology*, vol. 10, no. 1, pp. 55-63, 2002.
[65] V. Tkachev, M. Sorokin, C. Borisov, A. Garazha, A. Buzdin, and N. Borisov, "Flexible data trimming improves performance of global machine learning methods in omics-based personalized oncology," *International Journal of Molecular Sciences*, vol. 21, no. 3, pp. 713, 2020.
[66] N. E. Huang, M. L. C. Wu, S. R. Long, S. S. Shen, W. Qu, P. Gloersen, and K. L. Fan, "A confidence limit for the empirical mode decomposition and hilbert spectral analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 459, no. 2037, pp. 2317-2345, 2003.
[67] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in *Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA*, Springer Singapore, pp. 511-518, 2018.
[68] N. Aravind, S. Nagajothi and S. Elavenil, "Machine learning model for predicting the crack detection and pattern recognition of geopolymer concrete beams," *Construction and Building Materials*, *297*, pp. 123785, 2021.
[69] D. Kartini, D. T. Nugrahadi and A. Farmadi, A, "Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, pp. 390-395, Sep. 2021.
[70] C. Schaffer, "Selecting a classification method by cross-validation," *Machine Learning*, vol. 13, p.135-143, 1993.
[71] S. Narkhede, "Understanding AUC-ROC curve," *Towards Data Science*, vol. 26, no. 1, pp. 220-227, 2018.
[72] V. J. Lei et al., "Model performance metrics in assessing the value of adding intraoperative data for death prediction: Applications to noncardiac surgery," in *MedInfo*, 2019, pp. 223-227.
[73] J. A. Hanley and B. J. McNeil, "the meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.
[74] M. Durairaj and T. Suresh, "Enhanced gradient boosting tree classifier using optimization technique for water quality prediction," *Annals of the Romanian Society for Cell Biology*, pp. 3860-3873, 2021.
[75] T. Kavzoglu and A. Teke, "Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost)," *Arabian Journal for Science and Engineering*, vol. 47, no. 6, pp. 7367-7385, 2022.
[76] D. Dezfooli et al., "Classification of water quality status based on minimum quality parameters: Application of machine learning techniques," *Modeling Earth Systems and Environment*, vol. 4, pp. 311-324, 2018.
[77] S. Shrestha and F. Kazama, "Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River Basin, Japan," *Environmental Modelling & Software*, vol. 22, no. 4, pp. 464-475, 2007.
[78] V. Tkachev, M. Sorokin, C. Borisov, A. Garazha, A. Buzdin and N. Borisov, "Flexible data trimming improves performance of global machine learning methods in omics-based personalized oncology," *International Journal of Molecular Sciences*, vol. 21 no. 3, pp. 713, 2020.
[79] P. Ukkonen and A. Mäkelä, "Evaluation of machine learning classifiers for predicting deep convection," *Journal of Advances in Modeling Earth Systems*, vol. 11 no. 6, pp. 1784-1802, 2019.
[80] C. Mantel, F. Villebro, G. A. dos Reis Benatto, H. R. Parikh, S. Wendlandt, K. Hossain, ... and S. Forchhammer, "Machine learning prediction of defect types for electroluminescence images of photovoltaic panels," in *Applications of Machine Learning*, vol. 11139, SPIE, p. 1113904, Sep. 2019.