EPH–International Journal of Science & Engineering
Published by: E-Pub Solutions Private Limited
ISSN (Online): **2454-2016**
Volume 12- Issue- 02 May, 2026

Machine Learning-Based Water Pollution Prediction for Sustainable Environmental Monitoring

Dr. Shubham Tiwari[1*], Dr. Kavya Singh[2], Dr. Nitin Rajan[3], Dr. Tanvi Gokhale[4]

[1*]Department of Environmental Science, Banaras Hindu University, Varanasi, Uttar Pradesh, India
[2]Department of Computer Science and Engineering, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh, India
[3]School of Water Resources Engineering, Anna University, Chennai, Tamil Nadu, India
[4]Department of Civil Engineering, Visvesvaraya National Institute of Technology Nagpur, Maharashtra, India
*Corresponding Author Email: waterquality.ai.research@gmail.com

Article History:

Article Type: **Research**
Received Date: **18/02/2026**
Revised Date: **19/03/2026**
Accepted Date: **20/04/2026**
Published Date: **24/05/2026**

Keywords: Water Potability Prediction, Machine Learning, Environmental Monitoring, Water Quality Assessment, Sustainable Water Management

ABSTRACT

Pollution of water and the decreasing supply of clean drinking water have now become one of greatest environmental and human health issues in the world. The conventional methods of water quality assessment are usually costly, time-consuming and ineffective in continuous environmental monitoring. Here, the current paper examines application of machine learning techniques in predicting water potability using physicochemical water quality indicators. The authors used a publicly available dataset of 3276 water samples with nine environmental factors, including conductivity, organic carbon, trihalomethanes, pH, hardness, solids, sulphate, chloramines, and turbidity. Several monitored ML algorithms, which included; LR, Decision Tree, Random Forest, Support Vector Machine and K-Nearest Neighbor, were comparatively tested in terms of predictive performance. Mean imputation, feature scaling, and exploratory data analysis methods were applied to the data to prepare it before the development of the model. The results showed that SVM had highest accuracy, whereas the Decision Tree model had the most balanced classification results in terms of the F1-score. The most influential predictors of water potability were found to be pH, sulfate, solids, and hardness through feature importance analysis. The findings emphasize the suitability of ML methods in assisting intelligent environmental surveillance, sustainable water resources, and systems of early contamination. The research has a contribution to the increasing role of artificial intelligence to the environment and offers meaningful implications to policy-makers, smart cities, and health agencies.

Abstract

1. Introduction

Access to clean and safe drinking water is one of the biggest environmental and health problems facing the world today. High rate of industrialization, urbanization, agricultural runoffs, and climate change, have increased the water pollution issues in both developed and developing countries. Polluted water sources have negative effects on human health, aquatic life and sustainable development projects. Recent environmental research efforts suggest that the growing levels of chemical pollution, microbial pollution, heavy metals, and toxic industrial wastes are increasingly causing threats to freshwater resources around the world (Lukić Bilela et al., 2023). The freshwater crisis has thus augmented the need to have effective water quality assessment and monitoring systems that can guarantee availability of safe drinking water and sustainable environmental management.

Water potability evaluation is important in environmental surveillance since water quality directly affects health of people and sustainability of ecosystems. Physicochemical parameters of water including pH, hardness, sulfate level, chloramines, dissolved solids, conductivity, turbidity, and organic carbon are commonly accepted as significant factors of water safety and level of contamination. The conventional methods of testing water quality in laboratories though reliable, tend to be costly, time consuming and require a lot of resources and hence are not effective in real time environmental monitoring and quick decision making. The growing severity of environmental pollution has thus brought about the necessity of automated, smart, and scalable monitoring systems that have the ability to enable sustainable water management practices (Sharma et al., 2021).

Recent advancements in the domains of artificial intelligence (AI), machine learning (ML), and environmental informatics have provided new opportunities to improve water quality prediction and monitoring systems. ML algorithms are useful to analyze extensive datasets in the environment, to find latent relationships between water quality parameters and create water potability prediction models (Tran, 2022). ML algorithms can learn nonlinear interactions and complex patterns among environmental data, which are not available to conventional statistical techniques, and enhance predictive accuracy and decision support. The combination of AI-based technologies and environmental surveillance systems has facilitated the promotion of predictive analytics and smart resource management systems to a considerable extent (Subramaniyaswamy et al., 2025).

A few recent studies have highlighted the growing significance of machine learning techniques in water quality research and sustainable environmental monitoring. The study by Hossain et al. revealed that the use of ML algorithms with explainable AI methods has the potential to enhance the diagnostics of water potability and to offer contextual explanations in the process of making environmental decisions (Hossain et al., 2024). In the same line of thought, Rachid et al. also focused on the efficiency of the ML approaches in forecasting the potability of water based on the physicochemical water quality indicators (Rachid et al., 2025). Recent developments in deep learning, sensor technologies, and IoT-enabled environmental systems have also increased the usage of intelligent water quality assessment frameworks (Das et al., 2025).

Predictive environmental analytics and decision support systems have also been increasingly used in ML. New computational models can facilitate the imputation of data, classification, and prediction of data in complicated environmental records (Zhang et al., 2023). Environmental systems based on AI assist in identifying the risks of pollution early, optimizing the process of water treatment, and real-time environmental management (Islam, 2025). Moreover, the combination of predictive models and sustainable smart city infrastructures has enhanced the advancement of intelligent environmental governance structures that can enhance the sustainability and health outcomes of cities (Subramaniyaswamy et al., 2025).

Nevertheless, ongoing monitoring and proper forecasting of Due to environmental variations, water quality is still challenging, incomplete data and the dynamics of interactions between water quality indicators. Current monitoring systems are often limited with respect to delayed lab tests, inadequate automation, and expensive operations. In addition, the traditional analytical techniques might be ineffective to express the multidimensional attributes of the environmental data (Tiwari & Darbari,

2025). These issues underscore the need to come up with effective ML-based predictive models that can improve environmental sustainability and water quality monitoring (Sharmila et al., 2024).
In this regard, the current study examine how potability is impacted by physicochemical water quality characteristics utilising supervised machine learning techniques. The research also aims to come up with predictive models that can classify potable and non-potable water samples using environmental data analytics methods. Several ML algorithms are relatively tested since classification performance measures to decide which predictive algorithm is most appropriate to predict potable water. The study helps to develop intelligent and sustainable water quality management systems by combining ML techniques with environmental monitoring solutions.
The research is likely to have practical implications to the environmental agencies, public health authorities, policymakers, and smart environmental monitoring initiatives. The results can be used to develop cost-efficient and automated decision-support systems to manage water resources sustainably and implement better water protection strategies.

2. Literature Review

The growing threats of freshwater contamination and unsafe drinking water distributions have turned water quality assessment into a critical element of sustainable environmental management. Earlier research has highlighted that the physicochemical parameters like pH, hardness, sulfate level, conductivity, turbidity, chloramine, and organic carbon play an important role in determining water safety and environmental sustainability. Conventional water quality evaluation methods predominantly depended on laboratory tests and statistical methods of measuring the level of contamination and drinking water appropriateness. Nevertheless, they tend to be lengthy, expensive, and inefficient in large-scale or real-time environmental monitoring applications (Abba et al., 2017). As a result, academics are becoming more interested in clever computational techniques for regulating and forecasting water quality.
The concept of ML has become one of the most effective tools of environmental science because of its capacity to identify intricate nonlinear correlations between environmental variables and analyse sizable data sets. According to Bui et al., hybrid machine learning algorithms can greatly enhance the prediction of water quality indexes and the environmental decision-making process (Bui et al., 2020). Equally, Mosavi et al. emphasized the usefulness of ML models in the field of groundwater salinity vulnerability mapping and environmental risk assessment (Mosavi et al., 2021). Deep learning and artificial intelligence have recently advanced predictive analytics, increasing its application in environmental monitoring systems.
The application of AI to assess and predict water quality has been a subject of a number of studies. Banda and Kumarasamy designed a water quality index model of river water based on the artificial neural network to analyze spatiotemporal changes in river water quality (Banda & Kumarasamy, 2024). Their results showed that AI-based models are efficient and accurate in their environmental predictions. Rejini et al. also expanded on deep learning models that learn with transformers to predict the suitability of water in agricultural environments and showed that novel AI architectures are increasingly applicable to environmental surveillance (Rejini et al., 2025).
Physicochemical indicators of water quality that are usually used as significant predictive variables in studies on water quality prediction include pH, hardness, turbidity, sulfates, conductivity, organic carbon, dissolved solids and chloramines. These parameters give useful data on chemical content, pollution status, and the general water usability to the human body and environmental sustainability. The unpredictability of the environment and interactions between these indicators render ML methods especially appropriate when it comes to predictive modelling.
Past studies have applied various supervised ML models to tasks of environmental prediction such as Logistic Regression (LR), Decision Tree, random forest (RF), Support Vector machine (SVM), K-nearest neighbor, and XGBoost. Decision Tree and RF models are popular due to their interpretability and classification capabilities, whereas Support Vector Machine (SVM) models are useful in the nonlinear environment data (Bui et al., 2020). In the more recent past, smart environmental treatment

systems and intelligent monitoring frameworks using AI have become the center of interest in the study of sustainable water management (Wang et al., 2026).

Although ML has increasingly been used in environmental science, there is relatively little comparative research on the topic of water potability prediction as a subset of a sustainability-oriented environmental monitoring system by multiple classification algorithms. Available literature tends to focus more on prediction accuracy or environmental analysis separately, as opposed to developing a comprehensive research void about integrated ML-based sustainable water monitoring systems. Thus, the current research paper tries to fill this gap by comparatively assessing various ML algorithms to predict water potability based on physicochemical water quality indicators.

3. Materials And Methods

3.1 Research Design

This work employed a quantitative, predictive, and comparative research methodology to model machine learning in forecasting water potability as a subset of the larger, sustainable environmental monitoring. Supervised ML methods of classification designed to determine whether water samples were potable or non-potable based on physicochemical parameters of water quality formed the basis of the research framework. The research took a quantitative analytical method where environmental variables were processed, examined and applied in both performance evaluation and the creation of predictive models. Research design is consistent with the growing trend of AI and environmental informatics integration in the new sustainability studies. Through ML, the research aims to make contributions towards effective and smart environmental surveillance systems that have the potential to facilitate the safety of the masses and sustainable management of water resources. The comparative modeling approach was used to compare predictive performance of various ML models and identify best model to be used in water quality assessment.

3.2 Dataset Description

The publicly accessible water potability dataset was utilised in the study with 3276 observations and several physicochemical water quality characteristics related to drinking water safety. The data was created to be analyzed in binary classification, with the target variable being the water potability status. All the observations were associated with a water sample with environmental and chemical indicators which are typically utilised in environmental monitoring systems and water quality assessments. Data has nine independent variables, and they include pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. All these variables are important pointers that affect water safety, the level of contamination as well as general appropriateness to be consumed by human beings. Potability was the dependent variable and was modeled as a binary variable where a score of 1 meant that the water is potable and that it can be used to drink, whereas a score of 0 signified that the water is non-portable (Aditya Kadiwa, 2020).

The chosen dataset suits the environmental monitoring studies on ML since it includes various chemical and physical indicators of the water quality factors, which directly affect the safety of drinking water. The data also promotes predictive environmental analytics by facilitating the creation of smart-classification systems that can help government agencies, environmental authorities, and water management organizations in detecting unsafe water conditions. The description of the physicochemical water quality variables in ML-based prediction of potability of water is shown in Table 1.

Table 1. Description of Dataset Variables

Variable	Description
pH	Measures the acidic or basic nature of water
Hardness	Represents mineral concentration in water
Solids	Indicates the amount of dissolved solids
Chloramines	Concentration of disinfectant chemicals
Sulfate	Sulfate ion concentration in water

Conductivity	Electrical conductivity of water
Organic Carbon	Concentration of organic matter
Trihalomethanes	Chemical by-products formed during water treatment
Turbidity	Measurement of water clarity
Potability	Target variable indicating drinking suitability

3.3 Data Preprocessing

Preprocessing of the data was done to enhance quality of data, reduce inconsistency and improve predictive capability of the ML models. The environmental datasets often have missing values, outliers and different numerical scales hence Before the models were implemented, several pre-processing techniques were used.

First, missing values in the variables like pH, Sulfate and Trihalomethanes were found and replaced with mean imputation. Gaps in the observations were filled in with the mean of the relevant variable to maintain the general design of the dataset and to prevent significant loss of information. This was done to guarantee completeness of the set of data but also to provide statistical uniformity across observations. This was followed by processing feature scaling and normalization since the variables were measured in different units of measurement and numeric scales. The data used was standardized to convert them to a single scale with a zero mean and a unit variance. It was especially essential to this process in distance-based and optimization-based algorithms like SVMand K-Nearest Neighbor, which are sensitive to changes in feature magnitude.

The standardization process can be mathematically expressed as:

$$Z = \frac{X - \mu}{\sigma}$$

$$x$$

$$\mu$$

$$\sigma$$

$$z = \frac{x - \mu}{\sigma} \approx 1.2$$

$$\Phi(z) \approx 88.5\%$$

where Xdenotes the original feature value, μrepresents the mean of the feature, and σindicates the standard deviation.

Boxplot visualization and distribution analysis techniques were also used to analyze the outlier analysis. Environmental data inherently has variations in their ecological and chemical conditions; hence, there was no complete removal of extreme values. Rather, the techniques employed to scale down the effects of outliers in model training and predicting were employed.

To measure the predictive performance objectively the dataset was separated into the training and testing subsets. The model training was done on approximately 80 percent of the observations and the remaining 20 percent was used in testing and validation. Randomized sampling process was employed whereby potable and non-portable classes were represented unbiasedly in both subsets.

3.4 Exploratory Data Analysis

The preliminary analysis of the ML model involved Exploratory Data Analysis (EDA) to gain insight into statistical properties and the distribution of variables and the correlation between variables to water quality indicators. EDA helped to discover patterns, trends, and possible irregularities in the data and gave a background knowledge of the environmental variables influencing water potability.

The central tendency and dispersion characteristics of the variables of the dataset were first summarized by descriptive statistical analysis. All numerical features were computed using statistical measures, such as mean, median, minimum, maximum, and standard deviation. These measures offered valuable information about the variance and distribution of environmental parameters related to water quality. Pearson correlation coefficient was then applied to analyze linear relationships between the variables based on correlation analysis. The correlation table helped to determine the

positive and negative relationships between the environmental indicators and the target variable. The strength and direction of these relationships were represented as a correlation heatmap.
Pearson coefficient of correlation is denoted as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

In addition, the analysis of distribution was performed with the help of histogram plots and boxplots to analyze the data symmetry, skewness, and variability of features. These graphical methods helped to identify the abnormal distributions and contributed to the measurement of the overall behavior of environmental indicators in potable and non-potable water samples.
The EDA step was thus crucial in providing insights that were needed to assist in effective feature interpretation and shape the ML modelling process that followed.

3.5 Machine Learning Models

To come up with a predictive framework that will be useful in assessing the potability of water, several supervised ML classification algorithms were executed and compared. The algorithms that have been selected are determined by the popularity of the algorithms in predictive analytics, environmental monitoring and classification-based research studies.

3.5.1 Logistic Regression

A baseline binary classification model was used to estimate the likelihood of water being potable using physicochemical indicators, and this was done using LR. The algorithm uses the logistic sigmoid function to convert linear combinations of predictor variables into probability values between 0 and 1. Logistic function is denoted by:

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

where $P(Y = 1)$represents the probability of water being potable.

3.5.2 Decision Tree

To come up with hierarchical decision rules to predict water potability, Decision Tree classification was applied. The model recursively splits the dataset into subsets according to level of feature importance and reduction of impurity, thus creating interpretable classification structures to analyze the environment.

3.5.3 Random Forest

An ensemble ML method called RF was applied to enhance the predictive performance and minimize overfitting. The algorithm builds more than one decision tree and aggregates their forecasts by majority voting, thus improving the stability of model and classification accuracy on environmental datasets.

3.5.4 Support Vector Machine

SVM was used to determine the best hyperplanes that could be used to differentiate between potable and non-portable water samples in the multidimensional feature space. SVM has been shown to be efficient especially in nonlinear classification problems and large dimensional environment data due to its maximization of classification margins.

3.5.5 K-Nearest Neighbor

One of the instance-based classification methods used was the K-Nearest Neighbor (KNN) algorithm predicting water potability based on similarity measures of surrounding observations. Identification of the nearest neighboring samples in the dataset was done using Euclidean distance.
Euclidean distance formula can be expressed as:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

3.6 Model Evaluation Metrics

ML models were tested on the predictive performance based on the commonly used metrics of classification performance used in environmental prediction and research in artificial intelligence. These evaluation metrics allowed the thorough evaluation of the model accuracy, reliability and classification capability.

Proportion of correctly classified observations over the number of predictions made by the model was measured using accuracy.

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

where TP denotes True Positives, TN represents True Negatives, FP indicates False Positives, and FN denotes False Negatives.

Precision was calculated to evaluate the proportion of correctly predicted positive observations among all predicted positive cases.

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

Recall was utilized to measure the capability of model to correctly identify actual positive observations.

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

The F1-score was computed as the harmonic mean of precision and recall to provide a balanced assessment of classification performance.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Moreover, Receiver Operating Characteristic-Area Under Curve (ROC-AUC) analysis was done to assess the discriminative potential of the classification models. The ROC curve plots True Positive rate versus False Positive rate at various threshold levels whereas the AUC value is used to measure the total predictive ability of the model. The increased values of ROC-AUC are the signs of better classification and greater ability to differentiate between potable and non-potable water samples.

4. Results

4.1 Descriptive Statistics Results

The sample size was 3276 water samples having nine physicochemical water quality indicators and one target variable, Potability. The descriptive statistics indicated a lot of variability between the water quality parameters. The average pH was 7.0808, which meant that the water samples were mostly near to neutral. Hardness mean was 196.3695 with the highest variability being in Solids with a mean of 22014.0925 with a standard deviation of 8768.5708. The high variation indicates that the dissolved solid concentration of the water samples varied significantly.

The mean values of chloramines were 7.1223, Sulfate was 333.7758 and Conductivity had the mean value of 426.2051. The mean values of Organic Carbon, Trihalomethanes and Turbidity were 14.2850, 66.3963 and 3.9668 respectively. The average Potability was 0.3901, which means that the percentage of potable samples was 39.01, and that of non-potable samples was 60.99. This indicates a moderate imbalance of classes in the data. Table 2 shows descriptive statistics of water quality variables such as mean, standard deviation, minimum and maximum values.

Table 2. Descriptive Statistics of Water Quality Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
pH	7.0808	1.5943	0.0000	14.0000
Hardness	196.3695	32.8798	47.4320	323.1240

Solids	22014.0925	8768.5708	320.9426	61227.1960
Chloramines	7.1223	1.5831	0.3520	13.1270
Sulfate	333.7758	41.4168	129.0000	481.0306
Conductivity	426.2051	80.8241	181.4838	753.3426
Organic Carbon	14.2850	3.3082	2.2000	28.3000
Trihalomethanes	66.3963	16.1750	0.7380	124.0000
Turbidity	3.9668	0.7804	1.4500	6.7390
Potability	0.3901	0.4878	0.0000	1.0000

The pH, Sulfate, and Trihalomethanes had missing values. The pH variable had 491 missing values; Sulfate had 781 missing values and Trihalomethanes had 162 missing values. Mean imputation was used to fill in these missing values. Following the preprocessing, there were no missing values in the dataset.

4.2 Correlation Analysis

The analysis of linear relationship between the water quality parameters and Potability was done using correlation analysis. The results showed that all independent variables had very weak correlations with target variables. Positive correlation between Solids and Potability was the strongest, but very low at 0.0337. The weak positive correlation of chloramines was also 0.0238.
Organic Carbon had the highest negative correlation with Potability with a value of -0.0300. There were also weak negative relationships between Potability and Sulfate, Hardness, Conductivity and pH. These findings suggest that none of the physicochemical parameters has a strong independent explanation of potability of water.
The low values of correlation indicate that the potability of water is the result of the joint and potentially nonlinear effect of several indicators of water quality. This is why ML models can be suitable to the study since they can identify complex relationships that could be overlooked by simple linear correlation. The correlation heatmap in figure 1 indicates the association between the physicochemical water quality parameters and the water potability.

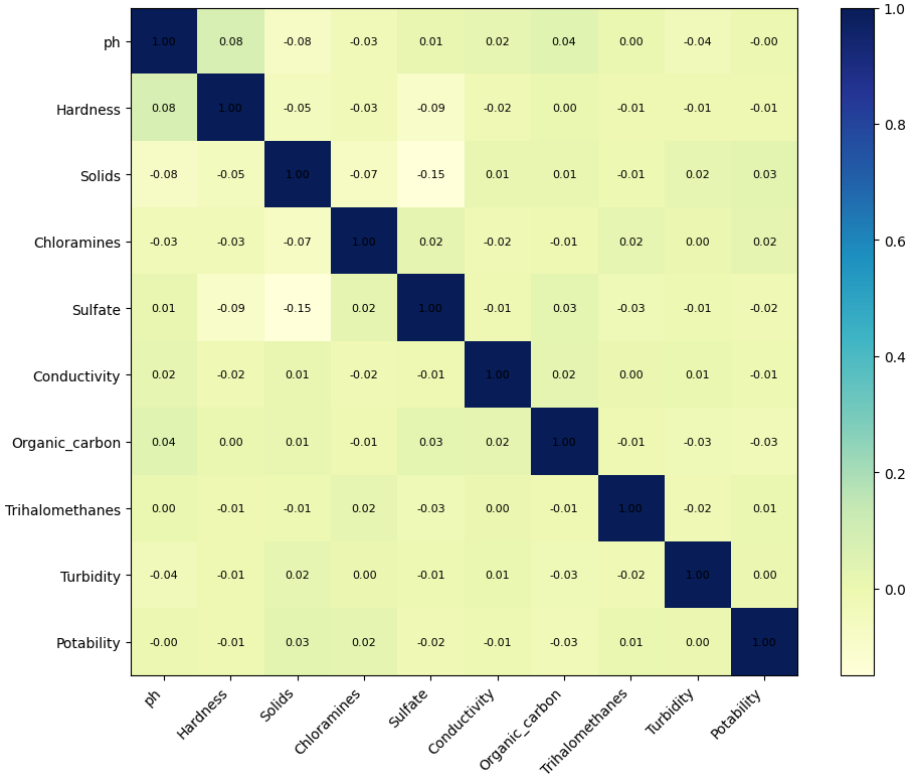

Figure 1. Correlation Heatmap of Water Quality variables

4.3 Machine Learning Model Performance

Evaluation of five supervised ML models was carried out: LR, Decision tree, RF, SVM and K-nearest neighbor. Accuracy, precision, recall, F1-score, and ROC-AUC were used to compare the models.
SVM had the best overall accuracy of 0.6692. RF had an accuracy of 0.6555, and K-Nearest Neighbor, LR, and Decision Tree had accuracy of 0.6113, 0.6098 and 0.6037, respectively. Nevertheless, the accuracy was not deemed to be an adequate criterion to select the model since the dataset was moderately imbalanced.
LR gave zero values of the precision, recall and F1-score of the potable class. This was due to the fact that the model was not accurate in identifying potable water samples and instead it mostly identified observations as non-potable. Hence, even though the LR had a predictive accuracy of 0.6098, this was not very useful in prediction.
The potable class had the highest F1-score of 0.4672 with the Decision Tree model. It means that Decision Tree offered the most appropriate balance between precision and recall as compared to the other models tested. RF had the largest ROC-AUC of 0.6548, which means that it has a better overall discriminatory power at all classification levels. SVM was also doing well in ROC-AUC, having a value of 0.6481. Performance of ML models is compared in table 3 based on the accuracy, precision, recall, F1-score, and ROC-AUC.

Table 3. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Interpretation
Logistic Regression	0.6098	0.0000	0.0000	0.0000	0.5484	Failed to identify potable water samples due to class imbalance
Decision Tree	0.6037	0.4914	0.4453	0.4672	0.5752	Best balanced classification performance
Random Forest	0.6555	0.6172	0.3086	0.4115	0.6548	Highest ROC-AUC and strong discrimination ability
Support Vector Machine	0.6692	0.6970	0.2695	0.3887	0.6481	Highest overall accuracy but low potable-water recall
K-Nearest Neighbor	0.6113	0.5031	0.3203	0.3914	0.5991	Moderate overall performance

The results of the classification indicate that SVM and RF were more effective in distinguishing non-portable samples, but their recall scores of potable water were relatively low. Decision Tree was not as accurate in general but offered a better-balanced classification performance in detecting potable water.

4.4 Best Performing Model

Decision Tree model was chosen as most effective model to use in this study based on the F1-score. Even though SVMwas the highest accuracy, its potable water sample recall was only 0.2695. Recall and F1-score are significant in water quality prediction since the model should be able to distinguish between drinkable and non-drinkable water samples. Thus, Decision Tree was deemed as the most appropriate since it had the best F1-score of 0.4672.
Confusion matrix of Decision Tree model revealed that it was able to correctly categorize 282 non-potable and 114 potable samples. It however, misclassified 118 non-portable samples as potable and 142 potable samples as non-portable. This shows that the model had moderate predictive performance but still needs to be enhanced to be useful in practice with environmental monitoring. Table 4 shows the importance of features ranking of the Decision Tree model to determine the most significant predictors of water potability.

Table 4. Feature Importance Ranking of Decision Tree Model

Rank	Feature	Importance
1	pH	0.1393
2	Sulfate	0.1329
3	Solids	0.1283
4	Hardness	0.1261
5	Chloramines	0.1249
6	Turbidity	0.1096
7	Trihalomethanes	0.1008
8	Organic Carbon	0.0703
9	Conductivity	0.0678

The analysis of the importance of features revealed that the pH is the most significant variable and the importance score was 0.1393. The second feature that was significant was Sulfate then Solids, Hardness and Chloramines. The lowest score was on conductivity. These results suggest that water potability forecasting is affected by a combination of various physicochemical parameters, but not a single variable. The feature of importance plot of Decision Tree model (Figure 2) indicates relative contribution of each water quality parameter to the prediction of potability.

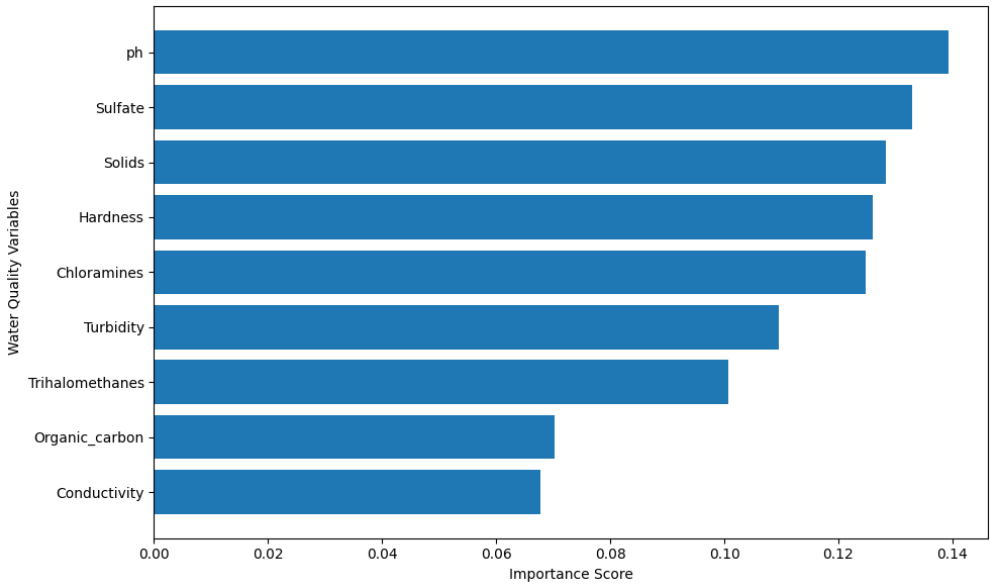

Figure 2. Feature Importance Plot

4.5 ROC Curve Analysis

Capacity of respective models to distinguish between water samples that are portable and those that are not at different threshold levels was assessed using ROC-AUC. SVMis the one with the highest ROC-AUC of 0.6481, then RF with 0.6548. K-Nearest Neighbor had a ROC-AUC of 0.5991, whereas Decision Tree and LR had a ROC-AUC of 0.5752 and 0.5484 respectively.

Above results suggest that RF was most discriminative in general with Decision Tree offering the best balance in precision and recall in classification of potable water. Thus, the findings are to be interpreted with caution. RF is a better choice in case the goal is general ranking and discrimination. But, when the goal is to classify potable-class balanced, then the Decision Tree is more appropriate. The comparison of ROC curves of the assessed ML models used to classify water as potable or not is shown in Figure 3.

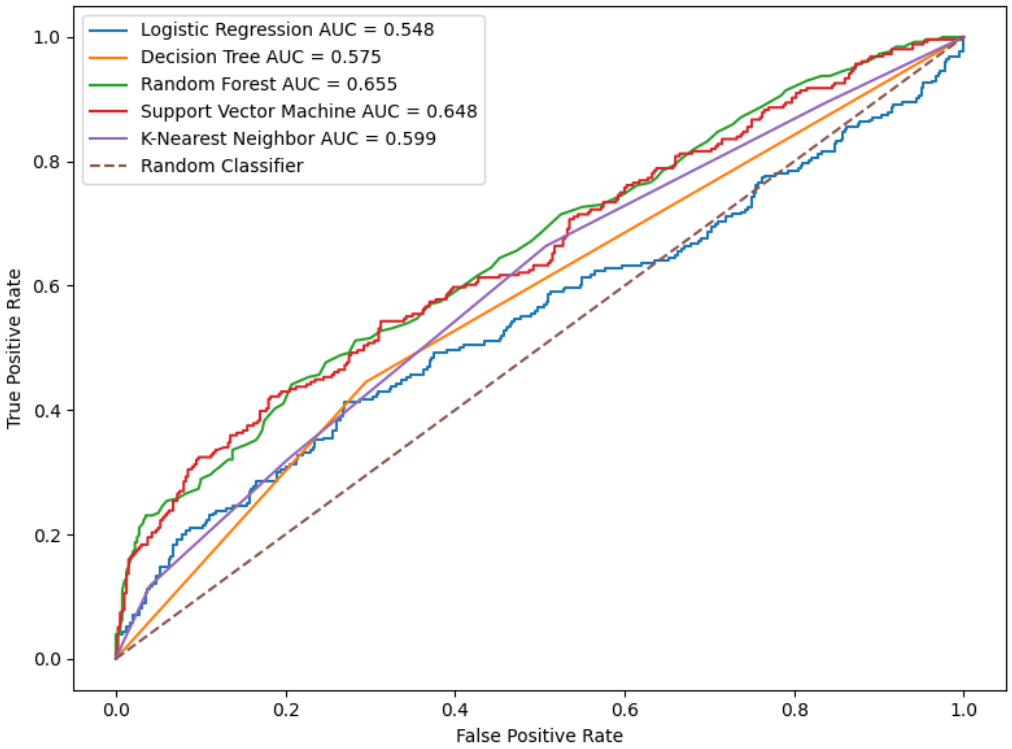

Figure 3. ROC Curve Comparison

In general, the findings indicate that ML is applicable in predicting the water potability based on physicochemical quality parameters of water. Nevertheless, the average values of performance reveal that the dataset poses classification difficulties as there are weak linear associations and class imbalance. This can be improved in the future by class balancing using SMOTE, hyperparameter optimization, ensemble learning, and more sophisticated models like Gradient Boosting or XGBoost.

5. Discussion

Results of current study indicate that ML methods could be successfully used to predict potability of water based on physicochemical water quality indicators. Some measured variables included pH, sulfate, dissolved solids, hardness, and chloramines, which proved to be the strongest predictors of potability of water. These results show that the state of water safety is a complex interdependence of several chemical and physical parameters and not one strong indicator. Likewise, it has been observed in the environmental monitoring literature where hydrochemical variables played a prominent role in the process of assessing water quality and sensing pollution (Kumar et al., 2024). The importance of features analysis also supported the idea that multidimensional environmental analysis is necessary to predict the potability of the water because the process of water contamination is rather complex.

The relative assessment of the ML algorithms showed that models had different predictive abilities. SVM was the most accurate on the overall and Decision Tree was the most balanced in terms of the F1-score. RF was the best in terms of discrimination based on ROC-AUC analysis. These findings indicate that ensemble and nonlinear ML methods are better applicable to environmental prediction issues with complex and weakly correlated data. The average performance figures in this research also depict the natural challenge of robustly forecasting the potability of water utilizing physicochemical indicators. Past research has also found that ML models are more effective in different datasets based on the characteristics of the dataset, imbalance of classes, and environmental variability (Campos et al., 2026).

Results of current study are aligned with previous studies that focus on increased significance of AI and predictive analytics in environmental monitoring systems. Sensing technologies enhanced with AI- and Internet of Things will be able to enhance sustainable water quality monitoring systems to a considerable degree (Das et al., 2025). Likewise, predictive systems with ML reinforcements enhance

early warning systems against aquatic diseases and enhance the capacity of the environment to make decisions (Hridoy et al., 2026). The present research reinforces these findings by showing that environmental data analytics can help predict patterns relating to potable and non-potable water conditions using ML algorithms.

The findings are also consistent with literature that highlights importance of AI-based monitoring systems in sustainable environmental management and the creation of smart infrastructure. The combination of AI agents and IoT systems can make the process of environmental monitoring more efficient and enable real-time assessment of climate and water quality (Miller et al., 2025). Moreover, Industry 4.0 technologies, such as AI-based smart monitoring systems, will help to make the environment more sustainable and transform smart infrastructure towards ESG (Kim et al., 2023). In this regard, the current research study is relevant to sustainability-based environmental monitoring by showing the suitability of ML methods to smart water quality prediction systems.

This study has significant implications on environmental sustainability. Predicting water quality systems developed with ML can be used to aid in early warning systems to detect contamination, minimize the reliance on expensive laboratory tests, and enhance real-time monitoring. These systems will help environmental agencies and policymakers to monitor any possible risks of pollution and put in place preventive water management measures. With the growing effects of industrial contaminants and the rise of new contaminants like PFAS in aquatic environments, smart environmental monitoring systems have become a requisite to protect water resources in a sustainable manner (Lukić Bilela et al., 2023).

Practically, the results can be used to inform the use of the findings in governmental water management authorities, smart cities, governmental health agencies, and IoT-based environmental monitoring networks. Water quality systems with AI can enhance resource distribution, environmental risk evaluation and sustainable urban governance systems. The practicality of predictive water quality analytics is further emphasized by the growing presence of automation and AI technologies in Industry 5.0 and smart environmental systems (Punitha et al., 2026). Also, computational AI systems are being used in the industrial and environmental domains more often to enhance automated decision support systems (Campos et al., 2026; Natarajan, 2025).

The study has various limitations despite its contributions. To start with, the study used secondary data that might not be a true reflection of environmental variability in the real world. Second, the data lacked real-time measurements of IoT sensors and time-dependent environmental dynamics, which restricted the use of the models in continuous monitoring systems. Third, the data were not geographically diverse, and this can decrease the applicability of the research findings to various environmental conditions and locations.

The next round of research should then be on incorporating deep learning models, hybrid AI models, IoT based sensor networks and real time systems for environmental monitoring to increase forecast precision. Further enhancements of sustainable water quality assessment and smart decision-making systems can be achieved through advanced methods that combine explainable AI, automated ML, and smart environmental infrastructures.

6. Conclusion

The current study examined use of ML methods in predicting water potability based on physicochemical parameters of water quality in the framework of sustainable environmental surveillance. Main aims of study were to examine impact of water quality indicators on potability, to create predictive ML models, compare them, and to assess their feasibility in intelligent environmental monitoring systems. The study employed a variety of supervised ML models such as LR, Decision Tree, RF, SVM and K-Nearest Neighbor. The results indicated that the prediction of potability of water relies on the joint effects of a few physicochemical parameters, especially pH, sulfate concentration, dissolved solids, hardness, and chloramines. The Decision Tree algorithm was the most balanced in terms of classification according to the F1-score, whereas SVM was the most accurate in general and the ROC-AUC performance of the RF was best. These findings affirm the usefulness of ML methods to aid predictive environmental analytics and water quality evaluation.

The research adds to the sustainable environmental monitoring by showing how AIcan be used to facilitate automated water quality predictions, early contamination, and data-informed environmental decisions. The results have significant implications to government agencies, public health authorities, and smart city projects in search of cost-effective and intelligent water monitoring solutions. Altogether, study emphasizes the increasing potential of ML-based environmental monitoring systems in ensuring sustainable water resource management and enhancing the protection of human health with the help of intelligent predictive analytics.

References

1. Abba, S. I., Hadi, S. J., & Abdullahi, J. (2017). River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Computer Science*, *120*, 75–82.
2. Aditya Kadiwa. (2020). *Water Quality*. https://www.kaggle.com/datasets/adityakadiwal/water-potability
3. Banda, T. D., & Kumarasamy, M. (2024). Artificial neural network (ANN)-based water quality index (WQI) for assessing spatiotemporal trends in surface water quality—A case study of South African river basins. *Water*, *16*(11), 1485.
4. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, *721*, 137612.
5. Campos, D., Galvão, V., de Rezende, M. L., Braga, A., Bodini, M., Aires, U. R., Yonaba, R., & Goliatt, L. (2026). Automated machine learning achieves accurate water quality prediction with reduced parameter requirements. *Scientific Reports*. https://www.nature.com/articles/s41598-025-34448-8
6. Das, S., Khondakar, K. R., Mazumdar, H., Kaushik, A., & Mishra, Y. K. (2025). AI and IoT: Supported Sixth Generation Sensing for Water Quality Assessment to Empower Sustainable Ecosystems. *ACS ES&T Water*, *5*(2), 490–510. https://doi.org/10.1021/acsestwater.4c00360
7. Hossain, M. M., Rahman, M. H., Rahman, M. A., & Ahmed, H. (2024). *Machine Learning for Diagnosing Water Potability and Explainable AI for Contextual Insights*. https://www.researchsquare.com/article/rs-4557533/latest
8. Hridoy, M. A. A. M., Pastorino, P., Bordin, C., Goliatt, L., Schneider, P., Shawkat, A. I., Rahman, M. S., Uddin, M., Bodini, M., & Ditthakit, P. (2026). *Machine Learning Enhanced Prediction of TDS for Strengthening Aquatic Disease Early Warning Systems*. https://www.researchsquare.com/article/rs-8199017/latest
9. Islam, F. S. (2025). A comprehensive analysis of air pollution in Dhaka City, Bangladesh, and the application of artificial intelligence and machine learning for enhanced management and forecasting. *International Journal of Applied and Natural Sciences*, *3*(1), 131–167.
10. Kim, H., Quan, Y.-J., Jung, G., Lee, K.-W., Jeong, S., Yun, W.-J., Park, S., & Ahn, S.-H. (2023). Smart factory transformation using industry 4.0 toward ESG perspective: A critical review and future direction. *International Journal of Precision Engineering and Manufacturing-Smart Technology*, *1*(2), 165–185.
11. Kumar, V., Alam, A., Kumar, J., Thakur, V. R., Kumar, V., Srivastava, S. K., Jha, D. N., & Das, B. K. (2024). Water Quality Assessment, Possible Pollution Source Identification from Anthropogenically Stressed River Yamuna, India using Hydrochemical, Water Quality Indices and Multivariate Statistics Analysis. *Water, Air, & Soil Pollution*, *235*(12), 820. https://doi.org/10.1007/s11270-024-07649-6
12. Lukić Bilela, L., Matijošytė, I., Krutkevičius, J., Alexandrino, D. A., Safarik, I., Burlakovs, J., Gaudêncio, S. P., & Carvalho, M. F. (2023). *Impact of per-and polyfluorinated alkyl substances (PFAS) on the marine environment*. https://run.unl.pt/entities/publication/3cc3f6c1-7a02-4b71-8514-e74a926426c2
13. Miller, T., Durlik, I., Kostecka, E., & Kozlovska, P. (2025). *Łobodzi nska, A.; Sokołowska, S.; Nowy, A. Integrating Artificial Intelligence Agents with the Internet of Things for Enhanced*

Environmental Monitoring: Applications in Water Quality and Climate Data. https://mlgp4climate.com/uploads/MLGP%20Library/Useful%20Documents/English/923.pdf

14. Mosavi, A., Sajedi Hosseini, F., Choubin, B., Taromideh, F., Ghodsi, M., Nazari, B., & Dineva, A. A. (2021). Susceptibility mapping of groundwater salinity using machine learning models. *Environmental Science and Pollution Research*, *28*(9), 10804–10817. https://doi.org/10.1007/s11356-020-11319-5
15. Natarajan, V. (Ed.). (2025). *Computational Artificial Intelligence and Methods for industries: A Machine-Generated Literature Overview*. Springer Nature Singapore. https://doi.org/10.1007/978-981-96-5277-8
16. Punitha, A., Syedakbar, S., & Jeyasudha, S. (2026). *Advanced Pathways in Electrical, Communication, and Automation: Reconfigurable Systems, Smart Energy, and AI for Industry 5.0*. CRC Press. https://books.google.com/books?hl=en&lr=&id=W7HNEQAAQBAJ&oi=fnd&pg=PP13&dq=Reddy,+V.+K.,+Kumar,+P.,+%26+Eswar,+P.+(2025).+Water+potability+detection+using+machine+learning.+International+Journal+of+Advanced+Research+in+Computer+Science,+16(1),+45%E2%80%9358.&ots=Jc1IKJxOGV&sig=6yR7yUFhEDC1VXJ--t_NqpVHWWw
17. Rachid, E.-B., Abderrahim, S., Hafid, A., & Souad, R. (2025). Predicting water potability using a machine learning approach. *Environmental Challenges*, *19*, 101131.
18. Rejini, K., Visumathi, J., & Genitha, C. H. (2025). Application of transformer-based deep learning models for predicting the suitability of water for agricultural purposes. *Water*, *17*(9), 1347.
19. Sharma, K., Raizada, P., Hasija, V., Singh, P., Bajpai, A., Nguyen, V.-H., Rangabhashiyam, S., Kumar, P., Nadda, A. K., & Kim, S. Y. (2021). ZnS-based quantum dots as photocatalysts for water purification. *Journal of Water Process Engineering*, *43*, 102217.
20. Sharmila, V., Kannadhasan, S., Kannan, A. R., Sivakumar, P., & Vennila, V. (2024). *Challenges in Information, Communication and Computing Technology: Proceedings of the 2nd International Conference on Challenges in Information, Communication, and Computing Technology (ICCICCT 2024), April 26th & 27th, 2024, Namakkal, Tamil Nadu, India*. CRC Press. https://books.google.com/books?hl=en&lr=&id=ORZEEQAAQBAJ&oi=fnd&pg=PP15&dq=Chowdary,+G.,+%26+Reddy,+V.+(2024).+Water+potability+prediction+using+ensemble+machine+learning+methods.+Journal+of+Environmental+Informatics,+39(2),+211%E2%80%93225.&ots=BDeQpNa7oC&sig=Cwyeg7kgk3M_IDQjZcZ2PlCXhy0
21. Subramaniyaswamy, V., Kshetri, N., Ravi, L., Revathy, G., & Thillaiarasu, N. (2025). *Deep Learning and Blockchain Technology for Smart and Sustainable Cities*. Auerbach Publications. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.1201/9781003476047&type=googlepdf
22. Tiwari, A., & Darbari, M. (2025). *Emerging trends in computer science and its application*. CRC Press Boca Raton, FL, USA. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.1201/9781003606635&type=googlepdf
23. Tran, K. P. (2022). *Machine learning and probabilistic graphical models for decision support systems*. CRC Press. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.1201/9781003189886&type=googlepdf
24. Wang, Z., Tang, R., Chen, G., Li, H., Deng, Y., Shen, J., & Li, D. (2026). A Review of Artificial Intelligence-Driven Smart Treatment of Aquaculture Effluent: Technical Framework, Application Scenarios, and Development Outlook. *Water*, *18*(4), 470.
25. Zhang, W., Li, R., Zhao, J., Wang, J., Meng, X., & Li, Q. (2023). Miss-gradient boosting regression tree: A novel approach to imputing water treatment data. *Applied Intelligence*, *53*(19), 22917–22937. https://doi.org/10.1007/s10489-023-04828-6