

Article

Water Quality Prediction Using Explainable Machine Learning Models

Majed Alwateer 

Department of Computer Science, College of Computer Science and Engineering, Taibah University,
Yanbu 46421, Saudi Arabia; mwater@taibahu.edu.sa

Abstract

Water pollution is a common global challenge, with significant impacts for ecosystem preservation, human health, and sustainable development. Addressing this issue requires a comprehensive understanding of the complex relationships between environmental factors and water quality outcomes. This study investigates the application of machine learning models to enhance water quality prediction and environmental management. The study utilizes robust machine learning models, including Adaboost, Random Forest, and Decision Tree classifiers, to uncover patterns within multidimensional water quality datasets. The Extra Trees classifier combined with the SMOTE-ENN resampling strategy achieved an accuracy of 87.25%, a recall of 95.26%, and an ROC of 94.52%. The Explainable AI (XAI) method is used to determine the impact of each parameter to the predictions of the model. This study identified parameters such as turbidity, solids, sulfate, hardness, and pH as some of the most influential factors in determining water potability. The identified factors provide valuable insights that can inform policy decisions and targeted interventions to reduce water pollution. The use of machine learning techniques provides a strong foundation for enhancing water quality evaluation and prediction, therefore facilitating sustainable water resource management.

Keywords: water pollution; quality prediction for water; ecosystem preservation; random forest; decision tree; machine learning

1. Introduction

Water is necessary for life, but pollution is making it less healthy, which is very harmful for people's health and for ecological systems [1]. Water pollution, defined as the deterioration of water quality, makes it unsuitable for various uses, including drinking, sanitation, and recreational activities [2,3]. This challenge is a global concern, exacerbated by the release of untreated wastewater into natural rivers and lakes, which negatively affects aquatic ecosystems.

The pollution of water can be caused by a variety of sources, including chemical runoff, waste disposal, and biological contaminants [4]. Agricultural practices are a primary contributor, as they utilize over 70% of the planet's freshwater resources, significantly impacting water quality. These practices are recognized as major drivers of water pollution in rivers and streams. Approximately 80% of the world's wastewater flows without treatment, leading to the contamination of rivers, lakes, and oceans. Pollution presents an issue to both human health and aquatic organisms that depend on clean water for survival [5,6]. Contaminated water and poor sanitation are major contributors to the spread of numerous diseases, including cholera, hepatitis A, dysentery, diarrhea, typhoid,



Academic Editor: Andrea G.
Capodaglio

Received: 7 January 2026

Revised: 1 February 2026

Accepted: 4 February 2026

Published: 7 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and polio. The lack, insufficiency, or mismanagement of sanitation and water services leaves people vulnerable to preventable health risks. This issue is especially alarming in healthcare facilities, where inadequate water, sanitation, and hygiene services increase the risk of infection and disease for both patients and staff. Worldwide, around 15% of patients acquire infections during hospital stays, with this rate being notably higher in low-income countries.

Access to safe and readily available water is vital for public health, fulfilling essential needs for drinking, food production, domestic use, and recreational activities [7,8]. Improving access to water and sanitation, coupled with effective management of water resources, can substantially drive economic growth and contribute to poverty reduction across nations. The global challenge of access to safe water highlights that approximately 2.1 billion individuals lacked safe water at home in 2015 [9]. Specifically, it notes that 263 million people spent over 30 min on average per round trip to collect water, emphasizing the significant time burden associated with water procurement. Additionally, 159 million individuals relied on surface sources, such as streams or lakes, for their drinking water, which poses health risks due to potential contamination. Moreover, 844 million individuals lacked access to enhanced drinking water services [10].

The consequences of unsafe water are profound, resulting in more fatalities annually than conflict and violence combined [11,12]. Furthermore, less than one percent of Earth's freshwater is accessible for human consumption. Without proactive measures, these challenges are projected to intensify, with global freshwater demand expected to increase by one-third by 2050.

Causal inference in machine learning represents a sophisticated methodological approach aimed at understanding and quantifying cause-and-effect relationships within complex datasets, transcending traditional correlational analyses. This computational strategy seeks to disentangle the underlying mechanisms that generate observed patterns by employing advanced statistical techniques, probabilistic models, and interventional frameworks such as potential outcomes and graphical models. Researchers leveraging causal inference methodologies attempt to estimate the precise impact of specific interventions or treatments by controlling for confounding variables, selection bias, and other potential distortions that might obscure genuine causal mechanisms. By integrating domain expertise, experimental design principles, and sophisticated algorithmic techniques, this field enables data scientists to construct more nuanced and interpretable models that can predict not just statistical associations but also the fundamental causal pathways driving observed phenomena across diverse domains, including healthcare, economics, social sciences, and policy research.

In support of sustainable water resource management and evidence-based environmental decision-making, this study is guided by the following research questions:

- RQ1: Can a reproducible machine learning pipeline reliably predict water potability under class imbalance conditions while ensuring stable and robust performance?
- RQ2: To what extent do ensemble-based classifiers, including Random Forest, Extra Trees, AdaBoost, and Decision Tree models, enhance predictive performance for water quality assessment compared to single-model approaches?
- RQ3: Which physicochemical and environmental parameters contribute most significantly to water potability predictions, as revealed through explainable artificial intelligence methods, and how can these insights support sustainable water quality monitoring and management strategies?

To tackle the complexities of water quality data, this study explores the application of machine learning algorithms, such as Random Forest and Decision Tree [13–18], to uncover the relationships between environmental factors and water quality outcomes. By analyzing

extensive datasets that incorporate diverse water quality metrics and climatic conditions, this research seeks to offer valuable insights into the factors that influence water quality.

This study investigates the application of machine learning techniques to enhance water quality prediction and environmental management. The study utilizes robust machine learning models, including Random Forest, Extra Trees, Adaboost, and Decision Tree classifiers, to uncover patterns within multidimensional water quality datasets. The results demonstrate the efficacy of the proposed framework, with the optimized Extra Trees model achieving an accuracy of 87.25% after hyperparameter tuning. The Explainable AI (XAI) method was used to quantify the impact of each parameter to the model's predictions. This study identified parameters such as turbidity, solids, sulfate, hardness, and pH as some of the most influential factors in determining water potability. The identified factors provide valuable insights that can inform policy decisions and targeted interventions to reduce water pollution.

The key contributions of this study are summarized as follows:

- This research introduces an interpretable and reproducible machine learning framework tailored for sustainable water quality assessment, integrating data preprocessing, imbalance correction via SMOTE-ENN, model optimization, and explainable artificial intelligence techniques.
- A comprehensive comparative analysis of multiple tree-based and ensemble learning models is conducted, demonstrating that the optimized Extra Trees classifier achieves superior predictive performance, with an accuracy of 87.25% and a ROC-AUC of 0.9452, thereby offering a reliable tool for long-term water quality monitoring.
- By employing SHAP-based explainability, the study identifies critical water quality parameters, such as turbidity, total dissolved solids, sulfate concentration, hardness, and pH, that significantly influence potability predictions, providing actionable insights to support sustainable policy formulation, targeted pollution mitigation, and resilient water resource management.

2. Related Work

Numerous studies have investigated various aspects of water quality, including its parameters, impacts, and management strategies. This literature review gathers key findings from recent research on water quality and its implications. The quality of drinking water is a crucial factor influencing public health, environmental sustainability, and economic development.

The physicochemical characteristics of water significantly influence its suitability for human consumption and ecosystem health. Parameters such as pH, hardness, turbidity, and the presence of contaminants (e.g., chloramines, trihalomethanes) have been extensively studied. According to WHO [19], pH levels outside the range of 6.5 to 8.5 can adversely affect both human health and aquatic life. Similarly, hardness has been linked to various health outcomes, with some studies suggesting a correlation between hard water and reduced cardiovascular disease incidence [20].

The relationship between water quality and public health is well-documented. Contaminated water sources are associated with a range of diseases, including cholera, dysentery, and typhoid fever [21,22]. A systematic review by Prüss-Ustün et al. [23,24] quantified the health impacts of unsafe water and inadequate sanitation, estimating that 2 million deaths annually can be attributed to waterborne diseases. Furthermore, the presence of trihalomethanes and chloramines has raised concerns regarding their potential carcinogenic effects [25,26].

Access to safe water is also pivotal for economic growth. The World Bank [27] emphasizes that improved water supply and sanitation can enhance productivity and contribute significantly to poverty alleviation. Investments in water infrastructure yield substantial

economic returns, as they reduce healthcare costs associated with waterborne diseases and improve workforce productivity [28,29].

Effective management strategies are critical for safeguarding water quality. Integrated Water Resources Management (IWRM) is recognized as a comprehensive approach that fosters the coordinated development and management of water, land, and related resources [30,31]. Furthermore, technological innovations, including remote sensing and machine learning, are increasingly being utilized to monitor water quality and predict pollution events [32,33].

The synthesis of the existing literature highlights the multifaceted challenges posed by water quality issues. Continued research is essential to develop innovative solutions and effective management practices that ensure safe water access for all, thereby safeguarding public health and promoting sustainable development [34,35].

3. Methodology

This section outlines the proposed framework employed to investigate the factors influencing water quality using machine learning algorithms. The approach involves data collection, the preprocessing of that data, training of the machine learning models, the evaluation of those models, and interpretation of results as shown in Figure 1. The figure illustrates the comprehensive workflow employed in the analysis of water quality data, focusing on the classification of water as potable (1) or non-potable (0). The process begins with the collection and pre-processing of the dataset, which involves cleaning and preparing the data to ensure its suitability for subsequent analysis. Following this, the dataset is partitioned into training and testing sets, with 30% of the data reserved for testing purposes.

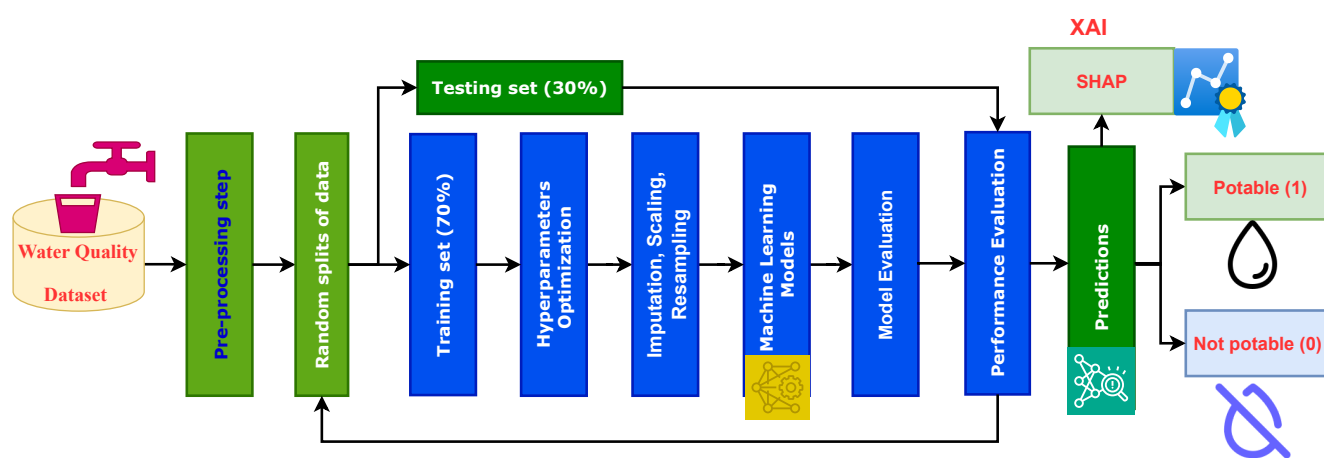


Figure 1. The framework employed in the analysis of water quality.

3.1. Water Quality Dataset

The dataset used for this study is gathered from multiple reliable databases that provide comprehensive information on water quality parameters [36]. The dataset includes the following parameters for each water sample: pH, chloramines, sulfate, hardness, trihalomethanes, conductivity, organic carbon, total dissolved solids, turbidity, potability (binary: 0 for non-potable, 1 for potable). The dataset is accessible online from the Kaggle website (<https://www.kaggle.com/code/alberthronaldy/water-quality-prediction>, accessed on 6 January 2026) and Github (<https://github.com/TusharPaul01/Water-Quality-IoT-ML>, accessed on 6 January 2026).

The dataset utilized in this study comprises various physicochemical parameters that are critical for assessing water quality. Table 1 presents the structure of dataset and the description of water quality features. Each feature is defined as follows: (1) pH: This

variable represents the acidity or alkalinity of water, measured on a scale from 0 to 14. A pH value of 7 indicates neutrality, while values below and above this threshold signify acidic and basic conditions, respectively. (2) Hardness: This parameter quantifies the capacity of water to precipitate soap, expressed in milligrams per liter (mg/L). Hardness is primarily attributed to the presence of calcium and magnesium ions and is an important indicator of water quality [37]. (3) Solids: This feature denotes the total dissolved solids in parts per million (ppm), which encompass a variety of inorganic and organic substances dissolved in water. (4) Chloramines: This variable measures the concentration of chloramines, a group of chemical compounds formed when chlorine is used for disinfection, reported in parts per million (ppm). Elevated levels can indicate water treatment practices. (5) Sulfate: This parameter reflects the concentration of sulfates dissolved in water, measured in milligrams per liter (mg/L). Sulfates can originate from natural sources and anthropogenic activities and affect water taste and quality.

Table 1. Database structure and description of water quality features.

Feature	Unit	Plausible Range / Unit	Missing
pH	–	0–14	491
Hardness	mg/L	0–300	0
Total Dissolved Solids	ppm	0–50,000	0
Chloramines	ppm	0–15	0
Sulfate	mg/L	0–500	781
Conductivity	$\mu\text{S}/\text{cm}$	0–2000	0
Organic Carbon	ppm	0–30	0
Trihalomethanes	$\mu\text{g}/\text{L}$	0–120	162
Turbidity	NTU	0–10	0
Potability	Binary	{0,1}	0

Furthermore, (6) Conductivity: This variable indicates the electrical conductivity of water, expressed in microsiemens per centimeter ($\mu\text{S}/\text{cm}$). Conductivity is a useful measure for assessing the ion concentration and overall water quality. (7) Organic Carbon: This feature represents the concentration of organic carbon in parts per million (ppm). Organic carbon content is a key indicator of water quality, as it can influence microbial activity and the presence of pollutants. (8) Trihalomethanes: This parameter quantifies the amount of trihalomethanes, which are byproducts of water chlorination, measured in micrograms per liter ($\mu\text{g}/\text{L}$). Their presence is pertinent to evaluating potential health risks associated with drinking water. (9) Turbidity: This variable measures the clarity of water, expressed in Nephelometric Turbidity Units (NTU). Turbidity indicates the presence of suspended particles, impacting both aesthetic quality and potential health risks. Finally, (10) Potability: This binary variable represents water safety for human consumption, with a value of 1 indicating potable water and a value of 0 signifying non-potable water.

Table 2 presents descriptive statistics for various water quality parameters, encompassing a total of 2785 to 3276 observations across different metrics. Key statistics include the mean pH value of 7.08, indicating a slightly acidic to neutral water quality, with a standard deviation of 1.59, suggesting variability in pH levels. Hardness averages at 196.37 mg/L, while total dissolved solids have a mean of 22,014.09 mg/L, reflecting the overall mineral content of the samples. Chloramines, sulfate, and conductivity are also reported, with means of 7.12 mg/L, 333.78 mg/L, and 426.21 $\mu\text{S}/\text{cm}$, respectively. The data highlights significant variability, as evidenced by the standard deviations and the range

from minimum to maximum values across all parameters. Notably, the potability status shows a mean of 0.39, indicating that approximately 39% of the samples were deemed potable, thereby emphasizing the necessity for ongoing monitoring and assessment of water quality to ensure public health and safety.

Table 2. Water Quality Parameters Descriptive Statistics.

Statistic	pH	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic Carbon	Trihalomethanes	Turbidity	Potability
Count	2785	3276	3276	3276	2495	3276	3276	3114	3276	3276
Mean	7.08	196.37	22,014.09	7.12	333.78	426.21	14.28	66.40	3.97	0.39
Std Dev	1.59	32.88	8768.57	1.58	41.42	80.82	3.31	16.18	0.78	0.49
Min	0	47.43	320.94	0.35	129	181.48	2.20	0.74	1.45	0
25%	6.09	176.85	15,666.69	6.13	307.70	365.73	12.07	55.84	3.44	0
50%	7.04	196.97	20,927.83	7.13	333.07	421.88	14.22	66.62	3.96	0
75%	8.06	216.67	27,332.76	8.11	359.95	481.79	16.56	77.34	4.50	1
Max	14	323.12	61,227.20	13.13	481.03	753.34	28.30	124	6.74	1

Figure 2 illustrates the distribution of water samples based on their potability status, represented as a bar chart. Two categories are depicted: non-potable (0) and potable (1) samples. The non-potable category comprises approximately 60.99% of the total count, reflecting a predominant proportion of the samples that do not meet safety standards for consumption. In contrast, the potable category accounts for 39.01%, indicating that a smaller subset of the samples is considered safe for drinking. This visual representation effectively highlights the disparity between potable and non-potable water sources, underscoring the critical need for continued monitoring and intervention to enhance water quality for public health purposes.

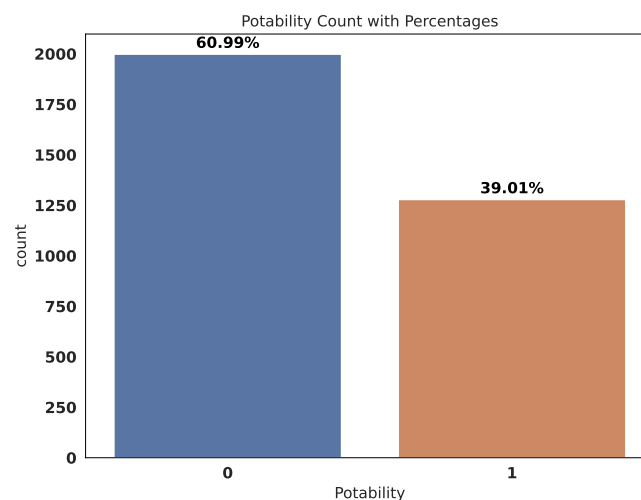


Figure 2. The distribution of water samples.

3.2. Data Preprocessing

Preprocessing is crucial to ensure the quality and usability of the dataset: (1) Data Cleaning: Identify and address missing values, outliers, and inconsistencies. Missing values can be handled using techniques such as mean/mode imputation or removal of records with excessive missing data. (2) Normalization: Normalize the data to bring all features within a similar scale. This can be done using Min-Max scaling or Z-score standardization, which is essential for algorithms sensitive to the range of data. (3) Feature Engineering: Create additional features if necessary, such as interaction terms or polynomial features, to capture complex relationships between parameters. (4) Data Splitting: Divide the dataset into training (70%), and test (30%) subsets to ensure robust model evaluation and avoid overfitting.

Table 1 summarizes the counts of missing data for various water quality parameters, highlighting the extent of incomplete observations across key metrics. Notably, the pH parameter shows a significant gap with 491 missing entries, while sulfate exhibits the highest level of missing data at 781. In contrast, other parameters, including hardness, solids, chloramines, conductivity, organic carbon, turbidity, and potability, report no missing values. Additionally, trihalomethanes have 162 absent entries. To address the missing data in the pH, sulfate, and trihalomethanes columns, mean values are utilized for imputation. This approach allows for the mitigation of missing data issues while preserving the integrity of the relationships among the various parameters.

3.3. Model Selection

The methodology provides a comprehensive framework for analyzing water quality using machine learning. By leveraging advanced analytical techniques, the study used various machine learning algorithms to identify the most effective model for predicting water quality. The selected algorithms include: Random forest is an ensemble learning model that constructs multiple decision trees and integrates their outputs to enhance performance and reduce overfitting.

The decision tree is a simple and interpretable model that divides data into subsets according to feature values. In contrast, Gradient Boosting Machines (XGB) utilize an ensemble approach, constructing decision trees in a sequential manner. Each subsequent tree aims to address the errors made by its predecessor, thereby enhancing the overall predictive performance of the model. This iterative process allows for improved accuracy and adaptability to complex data patterns.

3.4. Model Training

The framework incorporates the application of various machine learning models, which are optimized through hyperparameter tuning to enhance their predictive accuracy. The grid search technique is used to optimize hyperparameters for each algorithm. The performance of these models is rigorously evaluated using established metrics to determine their effectiveness in classifying water quality. Additionally, the framework used the SHapley Additive exPlanations (SHAP), a method established in Explainable AI (XAI), to provide interpretable insights into the model's decision-making process. This approach ensures transparency and aids in understanding the factors influencing the classification outcomes. Overall, the framework demonstrates an integrated and accurate approach to water quality assessment, using state-of-the-art machine learning algorithms and interpretability tools.

3.5. Model Evaluation

The evaluation of the trained models was conducted using various metrics, including: Accuracy: The ratio of correct predictions to total predictions, particularly important for binary classification tasks.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Precision and Recall: To evaluate the model's ability to accurately distinguish between potable and non-potable water.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score: The harmonic mean of precision and recall, offering a single comprehensive measure of model performance.

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

To ensure a robust and fair evaluation under class imbalance, multiple complementary performance metrics are employed. In addition to accuracy, precision, recall, F₁-score, and ROC-AUC, this study incorporates the Area Under the Precision-Recall Curve (PR-AUC), Matthews Correlation Coefficient (MCC), and Balanced Accuracy.

PR-AUC is particularly suitable for imbalanced classification problems, as it focuses on the trade-off between precision and recall for the positive (potable water) class and is insensitive to the abundance of true negatives. Balanced Accuracy is defined as the average of sensitivity (true positive rate) and specificity (true negative rate), ensuring equal importance is assigned to both classes. The Matthews Correlation Coefficient provides a single scalar measure that accounts for all elements of the confusion matrix and remains reliable even in highly skewed class distributions [38].

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

The MCC ranges from -1 to $+1$, where a value of $+1$ indicates perfect classification with complete agreement between predicted and true labels. A value of 0 corresponds to performance equivalent to random guessing, while a value of -1 reflects total disagreement, indicating that the predicted labels are entirely opposite to the ground truth.

3.6. Interpretation of Results

After evaluating the models, the best-performing model is selected for further analysis. The interpretation of results includes: (1) Feature Importance Analysis: Employing techniques like SHAP values to assess the contribution of each feature to the model's predictions. (2) Visualizations: Creating visual representations of key findings, such as feature importance plots, confusion matrices, and ROC curves, to communicate the results effectively.

4. Experiments

The analysis of the dataset provided critical insights into the factors that influence water quality and potability. The machine learning models employed demonstrated various levels of accuracy in predicting whether water was safe for human consumption. The machine learning models were evaluated based on their ability to accurately predict water potability using the provided physicochemical parameters. The primary models selected for this analysis include decision tree, Extra Trees, AdaBoost, XGBoost, and random forest.

4.1. Comparison of Resampling Methods

Table 3 summarizes the predictive performance of various classifiers under different data balancing strategies, expressed as percentages. In the baseline scenario without resampling, linear models such as logistic regression achieved moderate accuracy (59.4%) but low precision (49.0%) and F1-score (44.7%), reflecting limited capability in detecting minority-class instances. Distance-based models, exemplified by K-Nearest Neighbors, slightly improved both accuracy (61.8%) and F1-score (60.0%), while ensemble approaches, including Random Forest and Extra Trees, reached accuracies above 66% and F1-scores above 64%. Support Vector Classification displayed the most competitive baseline perfor-

mance, with an accuracy of 67.7% and an F1-score of 63.5%, indicating better discrimination in the unbalanced dataset.

Table 3. Comprehensive comparative performance of classifiers under different resampling and hybrid strategies.

Method	Classifier	Accuracy	Precision	Recall	F1	ROC AUC	PR AUC	Balanced Acc	MCC
Baseline	LogisticRegression	0.594	0.490	0.594	0.447	0.496	0.416	0.499	−0.010
	KNeighborsClassifier	0.618	0.604	0.618	0.600	0.630	0.534	0.577	0.170
	DecisionTreeClassifier	0.626	0.626	0.626	0.626	0.611	0.630	0.611	0.223
	SVC	0.677	0.699	0.677	0.635	0.705	0.630	0.616	0.311
	RandomForestClassifier	0.662	0.657	0.662	0.641	0.709	0.619	0.618	0.268
	GradientBoostingClassifier	0.634	0.623	0.634	0.603	0.677	0.599	0.582	0.197
Random Oversampling	LogisticRegression	0.506	0.506	0.506	0.506	0.514	0.533	0.506	0.011
	KNeighborsClassifier	0.607	0.607	0.607	0.607	0.660	0.651	0.607	0.214
	DecisionTreeClassifier	0.692	0.693	0.692	0.691	0.692	0.772	0.692	0.385
	SVC	0.660	0.667	0.660	0.656	0.727	0.721	0.660	0.327
	RandomForestClassifier	0.744	0.746	0.744	0.744	0.804	0.840	0.744	0.490
	GradientBoostingClassifier	0.656	0.656	0.656	0.655	0.723	0.725	0.656	0.312
Random Undersampling	LogisticRegression	0.497	0.497	0.497	0.497	0.522	0.551	0.497	−0.006
	KNeighborsClassifier	0.595	0.596	0.595	0.595	0.611	0.615	0.595	0.191
	DecisionTreeClassifier	0.608	0.608	0.608	0.607	0.608	0.710	0.608	0.216
	SVC	0.647	0.648	0.647	0.646	0.683	0.717	0.647	0.295
	RandomForestClassifier	0.649	0.649	0.649	0.649	0.699	0.719	0.649	0.298
	GradientBoostingClassifier	0.637	0.637	0.637	0.636	0.698	0.698	0.637	0.274
SMOTE	LogisticRegression	0.511	0.511	0.511	0.511	0.502	0.511	0.511	0.022
	KNeighborsClassifier	0.629	0.631	0.629	0.628	0.681	0.673	0.629	0.260
	DecisionTreeClassifier	0.643	0.643	0.643	0.643	0.643	0.730	0.643	0.286
	SVC	0.647	0.648	0.647	0.647	0.722	0.712	0.647	0.295
	RandomForestClassifier	0.722	0.722	0.722	0.722	0.788	0.800	0.722	0.444
	GradientBoostingClassifier	0.671	0.672	0.671	0.670	0.729	0.714	0.671	0.342
ADASYN	LogisticRegression	0.473	0.459	0.473	0.450	0.455	0.498	0.465	−0.077
	KNeighborsClassifier	0.635	0.639	0.635	0.627	0.668	0.656	0.629	0.269
	DecisionTreeClassifier	0.615	0.615	0.615	0.615	0.614	0.725	0.614	0.228
	SVC	0.639	0.643	0.639	0.632	0.690	0.675	0.633	0.277
	RandomForestClassifier	0.699	0.701	0.699	0.696	0.782	0.798	0.696	0.397
	GradientBoostingClassifier	0.621	0.622	0.621	0.617	0.671	0.665	0.617	0.240

Traditional resampling methods yielded mixed outcomes. Random oversampling enhanced the performance of ensemble classifiers: Random Forest and Decision Tree achieved accuracies of 74.4% and 69.2%, respectively, with corresponding F1-scores exceeding 74.4% and PR-AUC values above 84%. Conversely, random undersampling generally reduced model effectiveness, particularly for linear and distance-based learners, likely due to information loss from majority-class reduction. Synthetic oversampling methods, including SMOTE, ADASYN, and Borderline-SMOTE, provided moderate gains for ensemble learners, improving both accuracy and F1-scores, with Random Forest consistently performing at the top (up to 69.9% with ADASYN and 72.2% with SMOTE).

Table 4 presents the performance of various classifiers under multiple advanced resampling techniques, including Borderline-SMOTE, Tomek Links, SMOTE-ENN, SMOTE-Tomek, and ADASYN-Tomek. Each method was evaluated using seven classifiers: logistic regression, K-Nearest Neighbors, Decision Tree, Support Vector Classifier (SVC), Random Forest, Gradient Boosting, and Extra Trees. Performance metrics include accuracy, precision, recall, F1-score, and ROC-AUC.

Under Borderline-SMOTE, logistic regression achieved moderate performance with 51.0% accuracy and an F1-score of 51.0%, while K-Nearest Neighbors demonstrated im-

proved results with 66.5% accuracy and 66.4% F1-score. Ensemble-based models, particularly Random Forest and Extra Trees, outperformed other classifiers, attaining 73.3% and 77.0% accuracy respectively, with corresponding F1-scores of 73.3% and 77.0%, and ROC-AUC values exceeding 80%.

Table 4. Comprehensive comparative performance of classifiers under different resampling and hybrid imbalance-handling methods.

Method	Classifier	Accuracy	Precision	Recall	F1	ROC	Samples Train	Samples Test	Class Distribution
Borderline SMOTE	LogisticRegression	0.5104	0.5104	0.5104	0.5104	0.5074	2797	1199	0: 1998, 1: 1998
	KNeighborsClassifier	0.6647	0.6672	0.6647	0.6635	0.7140	2797	1199	0: 1998, 1: 1998
	DecisionTreeClassifier	0.6230	0.6230	0.6230	0.6230	0.6230	2797	1199	0: 1998, 1: 1998
	SVC	0.6455	0.6481	0.6455	0.6440	0.7130	2797	1199	0: 1998, 1: 1998
	RandomForestClassifier	0.7331	0.7331	0.7331	0.7331	0.8068	2797	1199	0: 1998, 1: 1998
	GradientBoostingClassifier	0.6397	0.6402	0.6397	0.6394	0.6935	2797	1199	0: 1998, 1: 1998
	ExtraTreesClassifier	0.7698	0.7700	0.7698	0.7698	0.8396	2797	1199	0: 1998, 1: 1998
Tomek Links	LogisticRegression	0.5800	0.6967	0.5800	0.4335	0.5130	2100	900	0: 1722, 1: 1278
	KNeighborsClassifier	0.6189	0.6132	0.6189	0.6135	0.6395	2100	900	0: 1722, 1: 1278
	DecisionTreeClassifier	0.5856	0.5883	0.5856	0.5867	0.5794	2100	900	0: 1722, 1: 1278
	SVC	0.6589	0.6816	0.6589	0.6198	0.7058	2100	900	0: 1722, 1: 1278
	RandomForestClassifier	0.6567	0.6531	0.6567	0.6439	0.6878	2100	900	0: 1722, 1: 1278
	GradientBoostingClassifier	0.6433	0.6434	0.6433	0.6184	0.6631	2100	900	0: 1722, 1: 1278
	ExtraTreesClassifier	0.6567	0.6556	0.6567	0.6390	0.6887	2100	900	0: 1722, 1: 1278
SMOTE ENN	LogisticRegression	0.5956	0.4945	0.5956	0.4609	0.5247	1060	455	0: 603, 1: 912
	KNeighborsClassifier	0.8110	0.8201	0.8110	0.8033	0.9047	1060	455	0: 603, 1: 912
	DecisionTreeClassifier	0.7385	0.7392	0.7385	0.7388	0.7285	1060	455	0: 603, 1: 912
	SVC	0.7846	0.7896	0.7846	0.7764	0.8446	1060	455	0: 603, 1: 912
	RandomForestClassifier	0.8308	0.8314	0.8308	0.8279	0.9068	1060	455	0: 603, 1: 912
	GradientBoostingClassifier	0.7670	0.7736	0.7670	0.7560	0.8241	1060	455	0: 603, 1: 912
	ExtraTreesClassifier	0.8725	0.8767	0.8725	0.8699	0.9452	1060	455	0: 603, 1: 912
SMOTE Tomek	LogisticRegression	0.5035	0.5035	0.5035	0.5035	0.5216	2636	1130	0: 1883, 1: 1883
	KNeighborsClassifier	0.6602	0.6651	0.6602	0.6576	0.7223	2636	1130	0: 1883, 1: 1883
	DecisionTreeClassifier	0.6097	0.6101	0.6097	0.6094	0.6097	2636	1130	0: 1883, 1: 1883
	SVC	0.6442	0.6447	0.6442	0.6440	0.7193	2636	1130	0: 1883, 1: 1883
	RandomForestClassifier	0.7425	0.7425	0.7425	0.7425	0.8190	2636	1130	0: 1883, 1: 1883
	GradientBoostingClassifier	0.6381	0.6381	0.6381	0.6380	0.7140	2636	1130	0: 1883, 1: 1883
	ExtraTreesClassifier	0.7611	0.7612	0.7611	0.7610	0.8552	2636	1130	0: 1883, 1: 1883
ADASYN Tomek	LogisticRegression	0.4879	0.4879	0.4879	0.4878	0.4822	2806	1203	0: 1998, 1: 2011
	KNeighborsClassifier	0.6409	0.6416	0.6409	0.6404	0.6725	2806	1203	0: 1998, 1: 2011
	DecisionTreeClassifier	0.6193	0.6195	0.6193	0.6191	0.6192	2806	1203	0: 1998, 1: 2011
	SVC	0.6309	0.6353	0.6309	0.6277	0.6875	2806	1203	0: 1998, 1: 2011
	RandomForestClassifier	0.7249	0.7249	0.7249	0.7248	0.8064	2806	1203	0: 1998, 1: 2011
	GradientBoostingClassifier	0.6492	0.6496	0.6492	0.6489	0.6943	2806	1203	0: 1998, 1: 2011
	ExtraTreesClassifier	0.7581	0.7585	0.7581	0.7580	0.8382	2806	1203	0: 1998, 1: 2011

The application of Tomek Links yielded modest gains for certain classifiers. Logistic Regression reached 58.0% accuracy but exhibited limited F1 performance (43.3%), whereas Random Forest and Extra Trees maintained consistent effectiveness with accuracies of 65.7% and F1-scores of 64.4% and 63.9%, respectively. SVC performance remained competitive, with 65.9% accuracy and 61.9% F1-score.

Hybrid resampling approaches demonstrated substantial improvements. SMOTE-ENN notably enhanced performance for K-Nearest Neighbors (81.1% accuracy, 80.3% F1-score) and Extra Trees (87.3% accuracy, 87.0% F1-score), achieving the highest ROC-AUC values across all classifiers (up to 94.5%). SMOTE-Tomek also yielded favorable results, with Random Forest reaching 74.3% accuracy and 74.3% F1-score, while Extra Trees attained 76.1% for both metrics. Similarly, ADASYN-Tomek improved ensemble

classifier performance, with Extra Trees achieving 75.8% accuracy and F1-score above 75.8%, and Random Forest maintaining 72.5% accuracy.

Table 5 presents the comparative performance of various classifiers following different feature selection strategies on the dataset. The baseline feature generation, performed over ten and twenty generations, converged on five features with a best fitness of 62.48%, indicating that this subset achieved a moderate balance between dimensionality reduction and classification potential.

Table 5. Comparative performance of classifiers under different feature selection strategies.

Category	Classifier	Accuracy	Precision	Recall	F1-score	ROC	Features Selected	Reduction Rate
Genetic Algorithm	Logistic Regression	0.6104	0.3726	0.6104	0.4627	0.5276	5	0.44
	K-Nearest Neighbors	0.5951	0.5745	0.5951	0.5759	0.5601	5	0.44
	Decision Tree	0.5778	0.5780	0.5778	0.5779	0.5564	5	0.44
	SVC	0.6531	0.6776	0.6531	0.5807	0.6181	5	0.44
	Random Forest	0.6236	0.6051	0.6236	0.6007	0.6015	5	0.44
	Gradient Boosting	0.6409	0.6316	0.6409	0.5869	0.6046	5	0.44
	Extra Trees	0.6389	0.6227	0.6389	0.6106	0.6099	5	0.44
Particle Swarm Optimization	Logistic Regression	0.6104	0.3726	0.6104	0.4627	0.5302	4	0.56
	K-Nearest Neighbors	0.5626	0.5409	0.5626	0.5456	0.5258	4	0.56
	Decision Tree	0.5493	0.5496	0.5493	0.5494	0.5265	4	0.56
	SVC	0.6236	0.6393	0.6236	0.5121	0.5687	4	0.56
	Random Forest	0.5982	0.5717	0.5982	0.5685	0.5725	4	0.56
	Gradient Boosting	0.6063	0.5691	0.6063	0.5397	0.5760	4	0.56
	Extra Trees	0.5921	0.5637	0.5921	0.5613	0.5661	4	0.56
Mutual Information	Logistic Regression	0.6104	0.3726	0.6104	0.4627	0.5301	9	0.00
	K-Nearest Neighbors	0.6165	0.5969	0.6165	0.5939	0.6120	9	0.00
	Decision Tree	0.5809	0.5793	0.5809	0.5801	0.5575	9	0.00
	SVC	0.6684	0.6944	0.6684	0.6086	0.6536	9	0.00
	Random Forest	0.6623	0.6584	0.6623	0.6236	0.6518	9	0.00
	Gradient Boosting	0.6684	0.6809	0.6684	0.6174	0.6202	9	0.00
	Extra Trees	0.6633	0.6631	0.6633	0.6205	0.6652	9	0.00
Chi-Square	Logistic Regression	0.6104	0.3726	0.6104	0.4627	0.5301	9	0.00
	K-Nearest Neighbors	0.6165	0.5969	0.6165	0.5939	0.6120	9	0.00
	Decision Tree	0.5809	0.5793	0.5809	0.5801	0.5575	9	0.00
	SVC	0.6684	0.6944	0.6684	0.6086	0.6536	9	0.00
	Random Forest	0.6623	0.6584	0.6623	0.6236	0.6518	9	0.00
	Gradient Boosting	0.6684	0.6809	0.6684	0.6174	0.6202	9	0.00
	Extra Trees	0.6633	0.6631	0.6633	0.6205	0.6652	9	0.00

When employing the Genetic Algorithm, five out of nine features were selected. Classifier performance under this configuration varied notably: Logistic Regression achieved an accuracy of 61.04% with an F1-score of 46.27%, while Extra Trees demonstrated comparatively superior performance with an accuracy of 63.89% and an F1-score of 61.06%. Support Vector Classifier and Gradient Boosting achieved intermediate results, highlighting the differential sensitivity of classifiers to the selected features.

The Particle Swarm Optimization (PSO) method selected four features, representing the most aggressive reduction. Under PSO, classification accuracy ranged from 54.93% (Decision Tree) to 62.36% (Logistic Regression), while F1-scores were generally lower compared to the Genetic Algorithm, suggesting that overly aggressive feature reduction can impair predictive balance.

Feature selection based on Mutual Information retained all nine features, yielding performance comparable to the baseline. Logistic Regression maintained 61.04% accuracy, and Extra Trees reached 66.33% accuracy with an F1-score of 62.05%. Similarly, Chi-Square

selection preserved the full feature set, producing results consistent with Mutual Information, indicating that these statistical approaches did not eliminate predictive information.

4.2. Performance Evaluation of Extra Trees with SMOTE-ENN

Table 6 shows the evaluation metrics obtained by the Extra Trees classifier when trained using the SMOTE-ENN resampling strategy. The results indicate strong and balanced predictive performance across both potable and non-potable water classes. For the non-potable class, the model achieved a precision of 0.91 and a recall of 0.75, resulting in an F1-score of 0.82 over 181 samples. This outcome suggests a low rate of false positive predictions for non-potable instances, while maintaining a reasonable level of sensitivity. In contrast, the potable class exhibited a precision of 0.85 and a recall of 0.95, corresponding to an F1-score of 0.90 across 274 samples. The higher recall reflects the model's strong ability to correctly identify potable water samples, which is particularly important for public health applications.

Table 6. Classification report for the Extra Trees classifier using the SMOTE-ENN method.

Class	Precision	Recall	F1-Score	Support
Not potable	0.91	0.75	0.82	181
Potable	0.85	0.95	0.90	274
Accuracy			0.87	455
Macro avg	0.88	0.85	0.86	455
Weighted avg	0.88	0.87	0.87	455

At the aggregate level, the classifier attained an overall accuracy of 0.87 on the test set. The macro-averaged precision, recall, and F1-score were 0.88, 0.85, and 0.86, respectively, indicating consistent performance across classes without dominance from class imbalance. Similarly, the weighted averages closely aligned with the overall accuracy, confirming that the predictive capability remains stable when accounting for class support. These findings demonstrate that the integration of SMOTE-ENN with the Extra Trees model yields reliable and well-balanced classification outcomes for water quality prediction tasks.

Figure 3 illustrates the confusion matrix obtained using the Extra Trees classifier combined with the SMOTE-ENN resampling strategy. The performance of the Extra Trees classifier combined with the SMOTE-ENN resampling strategy was assessed using standard classification metrics derived from the confusion matrix. Out of 455 evaluated samples, the model correctly identified 261 (True Positives (TP)) potable instances and 136 (True Negatives (TN)) non-potable instances, while producing 45 False Positives (FP) and 13 False Negatives (FN). This resulted in an overall accuracy of 87.25%, indicating that the majority of samples were classified correctly.

Table 7 summarizes the classification performance of the Extra Trees classifier using multiple evaluation metrics. The model achieves an overall accuracy of 0.882, indicating strong general predictive capability. Notably, the recall score is exceptionally high (0.968), demonstrating the model's effectiveness in correctly identifying positive instances, which is particularly important in risk-sensitive or safety-critical applications such as water quality monitoring. The precision score of 0.854 and the corresponding F1 of 0.907 reflect a well-balanced trade-off between false positives and false negatives. Furthermore, the high PR-AUC value of 0.972 suggests robust performance under class imbalance conditions. The balanced accuracy of 0.860 confirms consistent performance across classes, while the Matthews Correlation Coefficient (0.7566) indicates a strong overall correlation between predicted and true labels. The optimal decision threshold of 0.35 highlights that performance is maximized at a non-default cutoff, emphasizing the importance of threshold tuning in practical deployment scenarios.

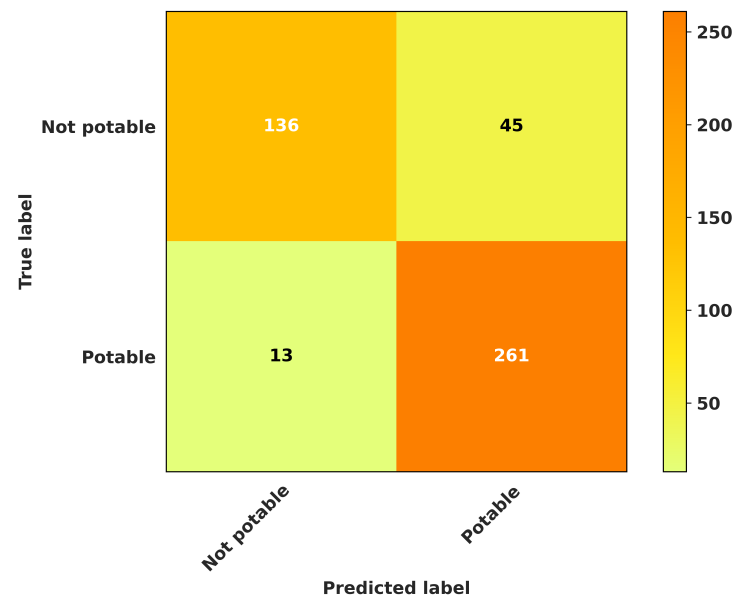


Figure 3. The confusion matrix of the Extra Trees classifier combined with the SMOTE-ENN.

Table 7. The performance metrics of the Extra Trees classifier combined with the SMOTE-ENN.

Metric	Value
Accuracy	0.882
Precision	0.854
Recall	0.968
F1-score	0.907
PR-AUC	0.972
Balanced Accuracy	0.860
Matthews Correlation Coefficient (MCC)	0.7566
Optimal Decision Threshold	0.35

Figure 4 illustrates a comprehensive evaluation of the Extra Trees classifier through complementary threshold-independent and threshold-dependent analyses. The Precision–Recall curve demonstrates consistently high precision across a wide range of recall values, yielding a PR-AUC of 0.972, which indicates excellent performance under class imbalance and strong reliability in positive class identification. The ROC curve further confirms robust discriminative capability, with a ROC-AUC of 0.957 and a clear separation from the random baseline, reflecting high true positive rates at low false positive rates. The metrics-versus-threshold analysis reveals the inherent trade-off between precision and recall, with the F1 peaking around a mid-range threshold, highlighting an optimal balance between the two. This behavior is reinforced by the optimal threshold selection plot, where a decision threshold of 0.35 is identified as optimal, satisfying a minimum precision constraint of 0.7 while maintaining high recall.

4.3. XAI Results

The robust performance of the optimized Random Forest model, combined with the interpretability provided by the feature importance analysis, underscores the value of integrating machine learning techniques with domain knowledge for water quality assessment. These findings offer valuable insights that can inform targeted interventions and policy decisions to address water pollution and ensure safe water access for all. To gain a deeper understanding of the relative importance of the input features, a feature importance analysis was conducted. The SHAP method was used to quantify the contribution of each parameter to the model's predictions. This analysis identified parameters such as turbidity,

solids, sulfate, hardness, and pH as some of the most influential factors in determining water potability.

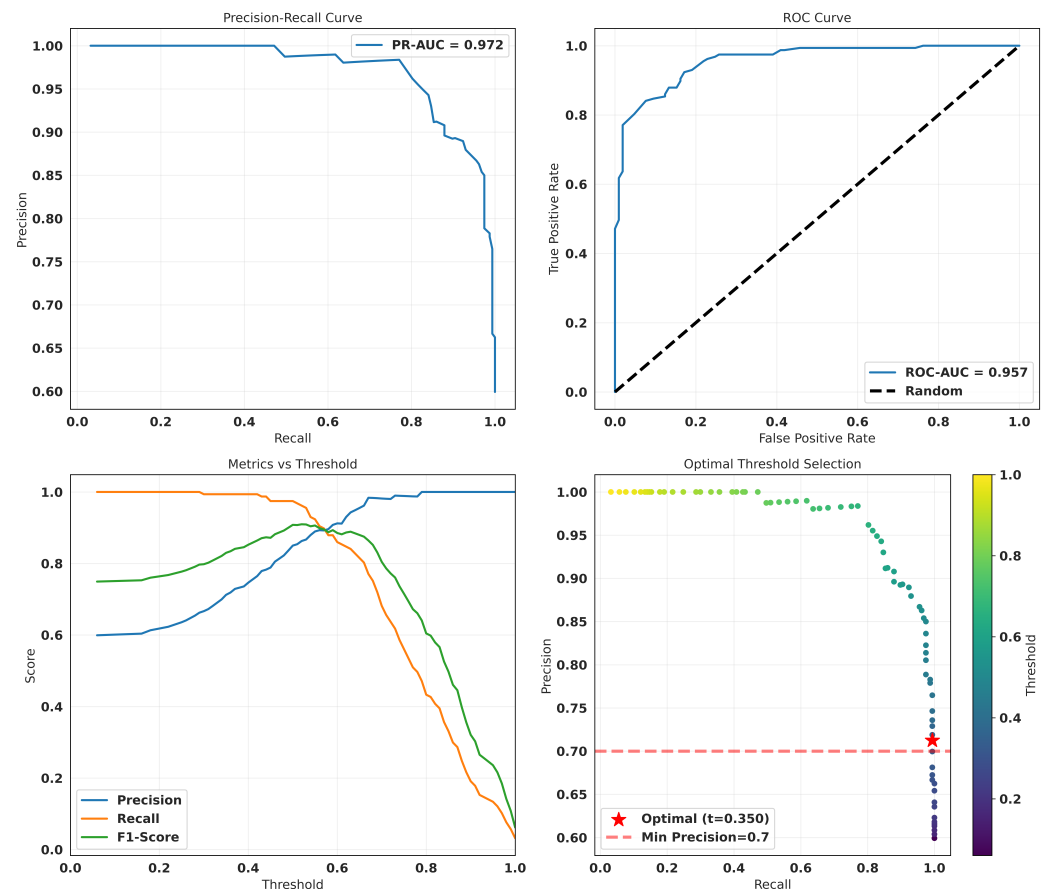


Figure 4. The PR-AUC, ROC-AUC, and optimal threshold selection for the Extra Trees classifier using SMOTE-ENN.

Figure 5 illustrates the mean absolute SHAP values for various water quality features, providing insights into their average impact on model output magnitude across two distinct classes. The x-axis represents the mean absolute SHAP values, which quantify the average contribution of each feature to predictions, while the y-axis lists the features, including chloramines, conductivity, organic carbon, trihalomethanes, turbidity, solids, sulfate, hardness, and pH. The comparison between Class 0 and Class 1 highlights the varying importance of these features in influencing model outcomes. Through the analysis of mean absolute values, it becomes clear which parameters have the most significant effect on the model's predictions, thereby enhancing understanding of the relationships between water quality indicators and the targeted outcomes within the two classes.

Figure 6 presents a summary of SHAP values, illustrating the impact of various water quality features on model output. Each feature, including trihalomethanes, turbidity, conductivity, chloramines, hardness, organic carbon, solids, pH, and sulfate, is plotted along the y-axis, while the x-axis indicates the SHAP value, which reflects the contribution of each feature to the model's predictions. The distribution of points over the x-axis illustrates the range of feature values, differentiating between low and high values across the different features.

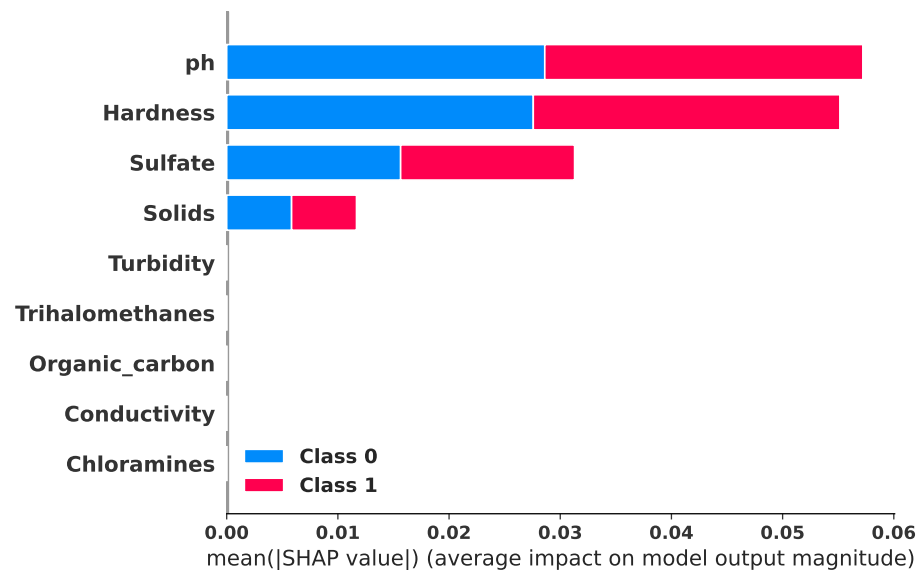


Figure 5. The SHAP summary of the Random Forest model.

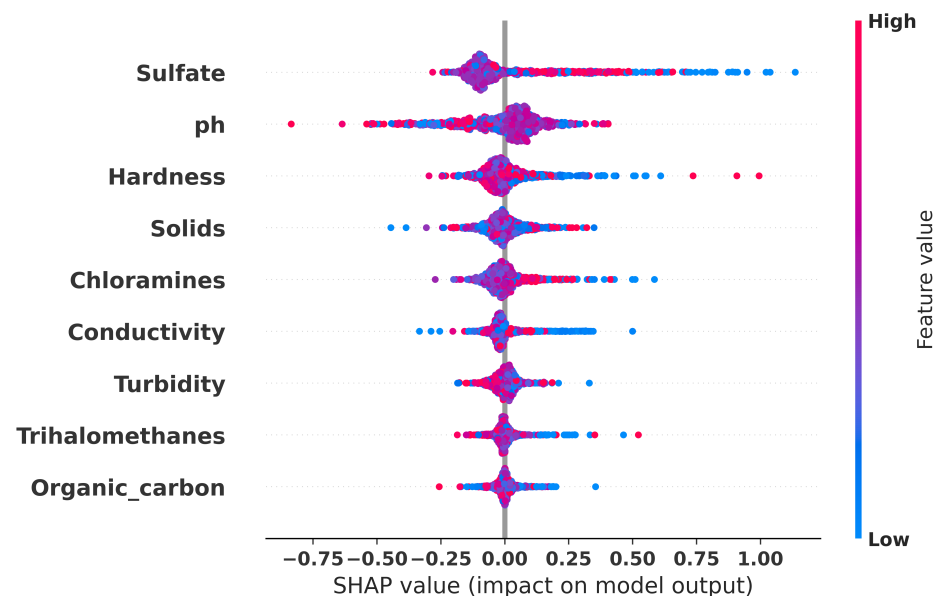


Figure 6. The SHAP summary of the XGB model.

Feature importance analysis shows that key parameters such as sulfate, pH, hardness, and the concentration of total dissolved solids were significant predictors of water quality. Specifically, pH levels were found to have a strong negative correlation with non-potable classifications, indicating that lower pH values significantly increased the likelihood of water being unsafe. Hardness also emerged as a critical factor, with higher hardness levels correlating with a greater probability of water being classified as non-potable. The concentration of total dissolved solids was similarly influential, with elevated levels associated with compromised water quality.

5. Discussion

This study investigated the effectiveness of machine learning models for water potability classification using a set of physicochemical water quality indicators. The results demonstrate that incorporating imbalance-aware evaluation metrics and strict data handling protocols substantially improves the reliability and interpretability of model performance.

In particular, the use of PR–AUC, balanced accuracy, and Matthews Correlation Coefficient (MCC) provides a more faithful assessment than accuracy alone, which can be misleading under class imbalance.

The experimental findings indicate that models optimized using resampling strategies applied exclusively to the training data achieve improved discrimination of the potable class without compromising the integrity of the test set. This confirms that careful separation between training and evaluation data is critical to avoid optimistic bias and information leakage. The observed improvements in MCC and PR–AUC suggest that the proposed framework achieves a better balance between sensitivity and specificity, which is essential in water quality assessment, where false negatives may pose public health risks.

Despite the promising results, this study has several limitations that should be acknowledged. First, the dataset used in this work is derived from publicly available sources and may not fully capture the diversity of water quality conditions encountered across different geographic regions or climatic contexts. Consequently, model generalizability to unseen environments may be limited without further validation on independent datasets. Second, the study relies exclusively on physicochemical parameters and does not incorporate microbiological, biological, or temporal variables, which are often critical for comprehensive water quality assessment. The absence of time-series analysis also limits the ability to capture seasonal variations or long-term trends in water quality.

Most importantly, the proposed models are intended strictly for research and decision-support purposes. They have not been verified or validated as standalone diagnostic or regulatory tools and should not be used for clinical screening, regulatory compliance, or the analysis of blood or other clinical samples unless formally evaluated and approved by appropriate regulatory authorities, such as the U.S. Food and Drug Administration (FDA). Future work should focus on external validation using geographically diverse datasets, integration of additional data modalities, and collaboration with regulatory and domain experts to assess the feasibility of transitioning from decision-support systems to certified operational tools.

6. Conclusions

This study demonstrates the efficacy of machine learning models in predicting water quality and assessing potability based on various physicochemical parameters. The results indicate that the Extra Trees model outperformed other algorithms, achieving a notable accuracy of 0.8725 following hyperparameter optimization. Key parameters such as pH, hardness, and organic carbon content emerged as significant indicators of water potability. The integration of advanced analytical techniques not only enhances predictive capabilities but also provides insights into the underlying factors affecting water quality. Furthermore, even though machine learning has great potential in water quality prediction, there are several limitations related to data quality, model interpretability, environmental complexity, and computational resources. These challenges need to be addressed in future work to improve the practical application and robustness of water quality prediction models.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tong, S.; Bambrick, H.; Beggs, P.J.; Chen, L.; Hu, Y.; Ma, W.; Steffen, W.; Tan, J. Current and future threats to human health in the Anthropocene. *Environ. Int.* **2022**, *158*, 106892.
2. Singh, J.; Yadav, P.; Pal, A.K.; Mishra, V. Water pollutants: Origin and status. In *Sensors in Water Pollutants Monitoring: Role of Material*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 5–20.
3. Sharma, K.; Rajan, S.; Nayak, S.K. Water pollution: Primary sources and associated human health hazards with special emphasis on rural areas. In *Water Resources Management for Rural Development*; Elsevier: Amsterdam, The Netherlands, 2024; pp. 3–14.
4. Akhtar, N.; Syakir Ishak, M.I.; Bhawani, S.A.; Umar, K. Various natural and anthropogenic factors responsible for water quality degradation: A review. *Water* **2021**, *13*, 2660.
5. Jan, S.; Mishra, A.K.; Bhat, M.A.; Bhat, M.A.; Jan, A.T. Pollutants in aquatic system: A frontier perspective of emerging threat and strategies to solve the crisis for safe drinking water. *Environ. Sci. Pollut. Res.* **2023**, *30*, 113242–113279.
6. Madhav, S.; Ahamad, A.; Singh, A.K.; Kushawaha, J.; Chauhan, J.S.; Sharma, S.; Singh, P. Water pollutants: Sources and impact on the environment and human health. In *Sensors in Water Pollutants Monitoring: Role of Material*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 43–62.
7. World Health Organization. *Safe Drinking-Water from Desalination*; Technical report; World Health Organization: Geneva, Switzerland, 2011.
8. Reader, G.T. Access to drinking water, food security and adequate housing: Challenges for engineering, past, present and future. In *Proceedings of the Symposium on Responsible Engineering and Living*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–41.
9. Unicef, U. Strategy for Water, Sanitation and Hygiene 2016–2030. 2016. Available online: <https://www.unicef.org/documents/unicef-strategy-water-sanitation-and-hygiene-2016-2030> (accessed on 6 January 2026).
10. Cassivi, A.; Johnston, R.; Waygood, E.; Dorea, C. Access to drinking water: Time matters. *J. Water Health* **2018**, *16*, 661–666.
11. Yakubu, A.; Gano, Z.S.; Ahmed, O.U.; Shuwa, S.M.; Atta, A.Y.; Jibril, B.Y. Application of hydrophobic deep eutectic solvent for the extraction of aromatic compounds from contaminated water. *Korean J. Chem. Eng.* **2022**, *39*, 1299–1306.
12. Wolf, A.T.; Kramer, A.; Carius, A.; Dabelko, G.D. Managing water conflict and cooperation. In *State of the World 2005*; Routledge: Oxfordshire, UK, 2017; pp. 106–125.
13. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **2020**, *249*, 126169.
14. Gad, I.; Elmezain, M.; Alwateer, M.M.; Almaliki, M.; Elmarhomy, G.; Atlam, E. Breast cancer diagnosis using a machine learning model and swarm intelligence approach. In *Proceedings of the 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
15. Gakii, C.; Jepkoech, J. A Classification Model for Water Quality Analysis Using Decision Tree. 2019. Available online: <http://repository.embuni.ac.ke/handle/embuni/2203> (accessed on 6 January 2026).
16. Fang, X.; Li, X.; Zhang, Y.; Zhao, Y.; Qian, J.; Hao, C.; Zhou, J.; Wu, Y. Random forest-based understanding and predicting of the impacts of anthropogenic nutrient inputs on the water quality of a tropical lagoon. *Environ. Res. Lett.* **2021**, *16*, 055003.
17. Alfahaid, A.; Alalwany, E.; Almars, A.M.; Alharbi, F.; Atlam, E.; Mahgoub, I. Machine Learning-Based Security Solutions for IoT Networks: A Comprehensive Survey. *Sensors* **2025**, *25*, 3341.
18. Alwateer, M.; Atlam, E.S.; Abd El-Raouf, M.M.; Ghoneim, O.A.; Gad, I. Missing data imputation: A comprehensive review. *J. Comput. Commun.* **2024**, *12*, 53–75.
19. World Health Organization. *Guidelines for Drinking-Water Quality*; World Health Organization: Geneva, Switzerland, 2002.
20. Bhatnagar, A.; Maziak, W.; Eissenberg, T.; Ward, K.D.; Thurston, G.; King, B.A.; Sutfin, E.L.; Cobb, C.O.; Griffiths, M.; Goldstein, L.B.; et al. Water pipe (hookah) smoking and cardiovascular disease risk: A scientific statement from the American Heart Association. *Circulation* **2019**, *139*, e917–e936.
21. Bartram, J.; Brocklehurst, C.; Fisher, M.B.; Luyendijk, R.; Hossain, R.; Wardlaw, T.; Gordon, B. Global monitoring of water supply and sanitation: History, methods and future challenges. *Int. J. Environ. Res. Public Health* **2014**, *11*, 8137–8165.
22. Eniolorunda, M.F. Faecal Contamination of Drinking Water Sources in Chanchaga Local Government AREA, Minna and Its Public Health Implications. Ph.D. Thesis, Federal University of Technology Minna, Minna, Nigeria, 2021.
23. Prüss-Ustün, A.; Wolf, J.; Bartram, J.; Clasen, T.; Cumming, O.; Freeman, M.C.; Gordon, B.; Hunter, P.R.; Medlicott, K.; Johnston, R. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low-and middle-income countries. *Int. J. Hyg. Environ. Health* **2019**, *222*, 765–777.
24. Prüss-Üstün, A.; Bos, R.; Gore, F.; Bartram, J. *Safer Water, Better Health: Costs, Benefits and Sustainability of Interventions to Protect and Promote Health*; World Health Organization: Geneva, Switzerland, 2008.
25. Richardson, S.D. Environmental mass spectrometry: Emerging contaminants and current issues. *Anal. Chem.* **2010**, *82*, 4742–4774.
26. Richardson, S.D. Disinfection By-Products: Formation and Occurrence in Drinking Water. *Encycl. Environ. Health* **2011**, *2*, 110–136.
27. Andrés, L.; Joseph, G.; Rana, S. *The Economic and Health Impacts of Inadequate Sanitation*; Oxford University Press: Oxford, UK, 2021. <https://doi.org/10.1093/acrefore/9780199389414.013.561>.

28. Hutton, G.; Haller, L.; Water, S.; World Health Organization. *Evaluation of the Costs and Benefits of Water and Sanitation Improvements at the Global Level*; Technical report; World Health Organization: Geneva, Switzerland, 2004.
29. Hulton, G.; World Health Organization. *Global Costs and Benefits of Drinking-Water Supply and Sanitation Interventions to Reach the MDG Target and Universal Coverage*; Technical report; World Health Organization: Geneva, Switzerland, 2012.
30. Fromageau, E. The global water partnership: Between institutional flexibility and legal legitimacy. *Int. Organ. Law Rev.* **2011**, *8*, 367–395.
31. Ewuzie, U.; Bolade, O.P.; Egbedina, A.O. Application of deep learning and machine learning methods in water quality modeling and prediction: A review. In *Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*; ACM: New York, NY, USA, 2022; pp. 185–218.
32. Cheng, B.; Zhang, Y.; Xia, R.; Wang, L.; Zhang, N.; Zhang, X. Spatiotemporal analysis and prediction of water quality in the Han River by an integrated nonparametric diagnosis approach. *J. Clean. Prod.* **2021**, *328*, 129583.
33. Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* **2022**, *1*, 107–116.
34. Muigua, P.D. Safeguarding Human Health Through Health in All Policies Approach to Sustainability. 2023. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4802419 (accessed on 6 January 2026).
35. Takhumova, O.; Goncharova, M. Innovative approaches to water resource management in achieving sustainable development goals. In *Proceedings of the E3S Web of Conferences, Malang, Indonesia, 21 October 2025*; EDP Sciences: Occitania, France, 2025; Volume 614, p. 04019.
36. Paul, T.; Raj, R.; Garg, P.; Jain, S. Real Time Monitoring of Water Quality for Rural Areas: A Machine Learning and Internet of Things Approach. In *Proceedings of the 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 9–11 May 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. <https://doi.org/10.1109/iciem59379.2023.10165824>.
37. Mequanint, G.; Argew, H.; Tilahun, H. Antimicrobial Solid Soap Production. Ph.D. Thesis, 2018.
38. Chicco, D.; Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **2023**, *16*, 4. <https://doi.org/10.1186/s13040-023-00322-4>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.