Ensemble Machine Learning for Reliable Water Potability Prediction with Optimized Physicochemical Feature Engineering

Rohit Srivastava
CSE Department, NIIT University, Neemrana, Rajasthan, India
Rohit.Srivastava@niituniversity.in

Keshav Sinha
School of Computer Science, UPES, Dehradun, India
Keshav.sinha@ddn.upes.ac.in

Dheeraj Samanta
School of Computer Science, UPES, Dehradun, India
dheeraj.105410@stu.upes.ac.in

Reham Alsabet
College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia | Automated Systems and Computing Lab (ASCL), Prince Sultan University, Riyadh, Saudi Arabia
ralsabet@psu.edu.sa (corresponding author)

Walid El-Shafai
College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia | Automated Systems and Computing Lab (ASCL), Prince Sultan University, Riyadh, Saudi Arabia | Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt
welshafai@psu.edu.sa

Ahmad Taher Azar
College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia | Automated Systems and Computing Lab (ASCL), Prince Sultan University, Riyadh, Saudi Arabia
aazar@psu.edu.sa

Received: 25 November 2025 | Revised: 28 December 2025, 3 February 2026, and 12 February 2026 | Accepted: 13 February 2026

 DOI: https://doi.org/10.48084/etasr.16495

ABSTRACT

Access to safe drinking water is a global challenge for both health and sustainability. This study presents a comprehensive examination of supervised machine learning models for water potability classification using physicochemical water quality characteristics, and a leakage-safe preprocessing pipeline incorporating stratified data splitting, mean imputation, and feature standardization. Class imbalance is handled by the Synthetic Minority Over-Sampling Technique (SMOTE), and feature selection and normalization were used to preprocess the dataset. Grid search and 5-fold cross-validation were used to train and optimize models, namely Random Forest (RFC), Logistic Regression (LR), Decision Tree (DTC), AdaBoost, K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), and Gradient Boosting (GBC). Accuracy, precision, recall, and F1 score on the test set were used to assess performance, showing that ensemble approaches, namely RFC and GBC, performed better, having outstanding stability and prediction accuracy. Preprocessing contributes to increasing model sensitivity to the minority class, while balancing trade-offs between interpretability and model complexity are discussed. The findings contribute to the

development of reliable, data-driven water quality monitoring systems that can align with sustainable development and public health goals. Since the limited dataset size may affect the generalizability to other regions, future studies should validate these findings on more diverse and comprehensive datasets.

Keywords-water potability; water quality; machine learning; feature selection; class imbalance; SMOTE; ensemble methods; mutual information; variance inflation factor; recursive feature elimination

I. INTRODUCTION

Water plays a crucial role in the life of every living being. The preservation of human health, the promotion of sustainable development, and the protection of the environment depend on clean and safe water. In addition to direct consumption, sanitation, agriculture, and industries require clean water [1]. Poor or contaminated water supplies are a major contributor to global waterborne diseases. According to the World Health Organization (WHO), more than two billion people worldwide lack access to safe drinking water supplies. In low and middle-income countries, waterborne diseases are a leading source of morbidity and mortality. Worldwide, the burden of disease due to unsafe water is immense, as infected water sources serve as the primary vehicle for pathogens, leading to gastrointestinal diseases that claim 1.8 million lives annually. A significant portion of these deaths can be prevented through improved water quality and sanitation measures [2]. Beyond direct health consequences, the impact of poor water quality extends to broader societal dimensions, as it adversely affects educational attainment, economic productivity, and collective well-being. These challenges are particularly acute within susceptible communities, where a lack of resources compounds the effects.

The barriers to universal access to clean water are becoming more and more complex. The rapid increase in population, urbanization, industry, and the consequences of climate change are putting strain on water supplies, which worsens due to the increased risks of biological and chemical contamination. Low-income groups and slum communities are disproportionately affected by inconsistent water supply due to unaffordability and lack of facilities in cities, particularly in the global South. Outdated infrastructure, economic disparity, and environmental risk factors remain obstacles to progress towards Sustainable Development Goal 6, which calls for universal access and sustainable management of water and sanitation by 2030. Traditional laboratory-based water quality assessment techniques are typically time-consuming, costly, and have limited capabilities for large-scale real-time monitoring [3], limiting the efficiency of water quality management.

State-of-the-art data-driven technology, particularly Machine Learning (ML), is a solution in response. ML models can analyze large datasets effectively and efficiently, discovering underlying patterns in the data and enabling automatic real-time water quality assessments. Such approaches can enhance the scalability and effectiveness of water monitoring systems, allowing proactive decision-making and early identification of potential pollution. Recent studies have increasingly explored the potential of ML in predicting water quality, demonstrating its superiority over conventional techniques in handling nonlinear and complex environmental data. The following works highlight diverse approaches, datasets, and ML models applied to assess water quality in different geographical and environmental contexts.

The study in [1] aimed to address the uncertainty of different classification schemes in assessing the Water Quality Index (WQI), which results in conflicting interpretations of water quality. Four ML classifiers, Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (KNN), were used to predict water quality classes from WQI models. The KNN and XGBoost algorithms performed better in accurately predicting water quality. This study showed that the Weighted Quadratic Mean (WQM) and Unweighted Root Mean Square (RSM) models integrated with ML classifiers can offer a practical and reliable coastal water quality assessment. In [2], evolutionary and ensemble ML methods were examined to assess water quality in the Indus River basin of Pakistan. The models were developed using 360 temporal readings and various input variables to determine water quality at the Bisham Qilla and Doyian Stations. The uncertainty analysis, employing quartile regression methods, demonstrated that both models had a low level of uncertainty, with prediction interval coverage probability greater than 85% at all phases.

In [3], different ML algorithms were used to assess water quality in Nainital Lake, Uttarakhand, India. This study focused on the degradation of water quality due to tourism and waste disposal. The three main parameters were identified as pH, total suspended solids, and turbidity. The results demonstrated that RF was best suited for regression analysis, and SVM, RF, and Stochastic Gradient Descent (SDG) all predicted water quality with equal accuracy. In [4], ML applications were reviewed in different contexts of water quality assessment, comparing ML data-driven models and conventional approaches. ML can be used to solve complex nonlinear problems using efficient means for water-related science. This study investigated the use of supervised, unsupervised, and reinforcement learning techniques in a variety of water settings, such as drinking water, sewage, ocean, groundwater, and surface water. This study highlighted that ML provides an effective set of tools to solve complex water quality problems.

According to [5], conventional water quality assessment techniques based on laboratory analysis face high expenses, time requirements, and poor spatial resolution. ML models provide a data-driven solution that can easily manage complicated nonlinear relationships between water quality factors. This study examined several ML as well as Deep Learning (DL) approaches, emphasizing the value of preprocessing procedures such as the Synthetic Minority Over-Sampling Technique (SMOTE) to address class imbalance in water quality data. The efficiency of Explainable AI (XAI) techniques was also demonstrated in real-time monitoring using IoT and remote sensing data to extend the scope of ML applications.

In [6], a supervised learning method was introduced to develop models capable of accurately predicting water quality. This study emphasized physicochemical and microbiological parameters to determine the suitability of water for drinking. The issue was defined as a binary classification problem, assessing various classifier models such as NB, Logistic Regression (LR), KNN, Tree-based classifiers, and ensemble methods. One of the primary features was the balancing of the dataset with SMOTE to prevent possible biases. In [7], ML techniques were applied to anticipate water quality in fish farming environments. This study highlighted the need to monitor water quality factors, including temperature, pH, dissolved oxygen, alkalinity, nitrogen compounds, electric conductivity, and meteorological data, in high-quality fish farming. The findings indicated that RF was able to predict all of these features effectively.

In [8], several ML methods, such as RF, Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Adaptive Boosting (AdaBoost), were utilized for water quality classification using a WQI threshold. Several regression models, such as KNN, Decision Tree (DT), SVM, and Multi-Layer Perceptron (MLP), were applied for WQI prediction. In addition to ML models, many DL models, such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), were also used. The results showed that DL can outperform traditional methods in terms of prediction accuracy and lower error rates. In [9], ML models, specifically ensemble models like XGBoost and RF, offered a precise and scalable solution for predicting water potability. These methods overcome the limitations of conventional approaches and facilitate timely interventions to reduce waterborne diseases. In addition, discussing model optimization, data quality enhancement, and interpretability can enhance and advance ML-based water quality prediction systems. In [10], ML models were utilized to predict water quality efficiently using multiple physical, chemical, and biological parameters. Previous studies have demonstrated that ML models can accurately predict water quality indices and classes. A combination of ML with the Internet of Things (IoT) and remote sensing can lead to real-time water quality monitoring and management.

Despite the extensive use of ML techniques to predict water quality and potability, several limitations can still be observed. Existing works focus primarily on improving the accuracy of prediction, and ML models are compared without systematically analyzing the impact of preprocessing choices, feature relevance, and multicollinearity on model performance. Several studies employ ensemble or DL models but do not explicitly distinguish between learning paradigms or evaluate them under consistent experimental conditions. Performance evaluation in previous research is often limited to accuracy-based metrics, with limited use of statistical significance testing to validate model comparisons.

This study is motivated by these gaps to provide a practically structured and reproducible framework. It integrates advanced preprocessing, multicriteria feature selection, taxonomy-aware ensemble evaluation, and statistical validation. By explicitly addressing these overlooked aspects, it seeks to enhance the reliability, interpretability, and robustness of ML-based water potability prediction. The primary goal is to develop and validate a strong ML model that uses physicochemical characteristics to predict water potability. This work uses advanced preprocessing approaches, such as feature selection, standardization, and class imbalance handling, to optimize model accuracy and reliability. A detailed comparison of many supervised learning models, both ensemble and interpretable, is carried out to determine the most effective technique to predict water quality [4]. The focus is on the development of a scalable, intelligent water monitoring system that will protect public health and enable the sustainable use of water resources worldwide, in line with global development goals. The key contributions of the study are as follows:

- Adopts an ML pipeline to integrate preprocessing, feature scaling, and class imbalance handling for reliable water potability prediction.
- Uses multicriteria features to assess information content, redundancy, and stability using Mutual Information (MI), correlation analysis, and Recursive Feature Elimination (RFE).
- Uses bagging-based and boosting-based learning paradigms in identical experimental conditions.
- Uses statistical validation mechanisms based on cross-validated F1-scores and hypothesis testing to ensure robustness and accuracy.

II. RESEARCH METHODOLOGY

An end-to-end experimental design was adopted to examine preprocessing, feature extraction, ensemble learning behavior, and statistical validation within a single framework. The ML pipeline provides a systematic and reproducible way to convert raw data into valuable insights. The pipeline comprises several stages, ranging from data acquisition and preprocessing to model training and testing, for the final performance and trustworthiness of the model, as shown in Figure 1. The step-by-step process taken to develop, train, and test ML models is as follows:

- Data collection forms the backbone of the entire ML process. Sources can be databases, APIs, files, or sensors. The quality of the data collected heavily influences performance at later stages.
- Exploratory Data Analysis (EDA) helps in understanding data, finding patterns, revealing anomalies, and developing hypotheses. Various techniques are used to understand the distribution of the data and relationships between variables and possible problems such as outliers and missing values. Preprocessing is used to enable an ML model recognize and uncover underlying patterns in the data.
- Data imbalance check helps identify if the classes are balanced or if one heavily dominates the others, which can affect model performance by introducing bias.
- Oversampling is used to create synthetic samples for the minority class, help balance the classes, and enhance model performance.

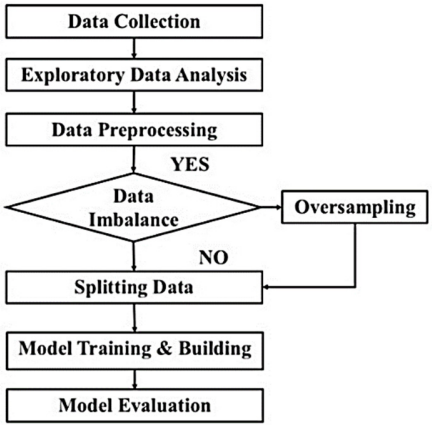

Fig. 1. ML pipeline.

A. Data Description

The Water Potability Dataset [17] was used in this study, which includes water quality measures of 3276 distinct water bodies. This dataset is widely used for environmental data science research and provides valuable insights into water potability (whether the water is safe for human consumption or not). The dataset contains several physicochemical properties of water, as shown in Table I.

TABLE I. STANDARD WATER QUALITY METRICS FOR SAFE WATER

Parameter	Standard values
pH	6.5–8.5
Hardness	60–120 mg/L
Solids	500–1000 mg/L
Sulphate	3–30 mg/L
Chloramines	Up to 4mg/L
Conductivity	<400 µS/cm
Trihalomethanes	Up to 80 ppm
Organic Carbon	< 2 mg/L
Turbidity	1–5 NTU
Potability	1-safe, 0-unsafe

The dataset consists of unique samples, each characterized by nine key attributes that help assess water safety. It also includes a binary classification label that indicates whether a sample is potable (1) or not (0). Water potability is formulated as a binary classification problem, where potable water is defined as the positive class (label = 1) and non-potable water as the negative class (label = 0). The dataset is class-imbalanced, with non-potable samples constituting the majority class and potable samples representing the minority class. To maintain class proportions, stratified sampling was used during splitting, training, and testing.

B. Data Preprocessing Techniques

In ML, several preprocessing techniques help make the dataset consistent, reliable, and effective, so models can effectively learn and understand data patterns and try to capture them to produce better results when tested on unseen real-world data. The following preprocessing techniques were implemented sequentially.

1) Handling Missing Values

Missing values were identified in several features, such as pH, Sulphate, and Trihalomethanes. These were handled with mean imputation, as the distribution of the features was approximately normal (between -0.5 and 0.5). Mean imputation is commonly preferred for features that follow a normal distribution. Although median imputation is another good option, it is generally preferred for moderately skewed features (between -1 and -0.5 or 0.5 to 1), as it is more robust to outliers.

2) Handling Skewness

Most of the dataset's properties, including pH, Hardness, Chloramines, Sulphate, Organic Carbon, Trihalomethanes, and Turbidity, exhibit distributions that are very symmetrical and centered, closely matching a normal distribution. Figure 2 represents the median, interquartile range, and prevalence of outliers in 3,276 samples per feature, emphasizing the spread and magnitude of extreme values that are important for the following data preprocessing and modeling activities.

However, the Solids and Conductivity features exhibit a clear right skew, with a progressive tail extending toward higher values and a larger concentration of data points at lower values. This imbalance is most noticeable in the Solids histogram, where the distribution is skewed to the left and extends to the right; similarly, conductivity has a larger upper tail. These patterns show that although most of the variables are roughly normal, the positive skewness of Solids and Conductivity causes them to depart from this trend. To correct for this, the Yeo-Johnson transformation was applied. This transformation is an extension of the Box-Cox transformation and can handle both positive and negative skewness. This helped in handling the skewness and bringing the distribution closer to normal, which was best suited for ML algorithms.

3) Handling Outliers

Data outliers were handled by using the Interquartile Range (IQR) approach. IQR estimates the difference between the first (Q1) and third (Q3) quartiles of data. Any data point falling outside this distance was capped to reduce its impact. This approach was used to normalize outlying values and preserve their overall form.

Figure 3 represents the underlying data asymmetry, as quantified by the skewness metric, highlighting distributional irregularities relevant for subsequent preprocessing and modeling steps. It represents several features, such as Solids, Conductivity, and Trihalomethanes, that contain individual points significantly beyond the boxplot whiskers, indicating the presence of extreme values or outliers. If these outliers are not addressed, they could distort model training and negatively impact predicted accuracy. To manage these high values, IQR-based capping was applied, which helped maintain the stability of the dataset and made it suitable for further ML analysis.

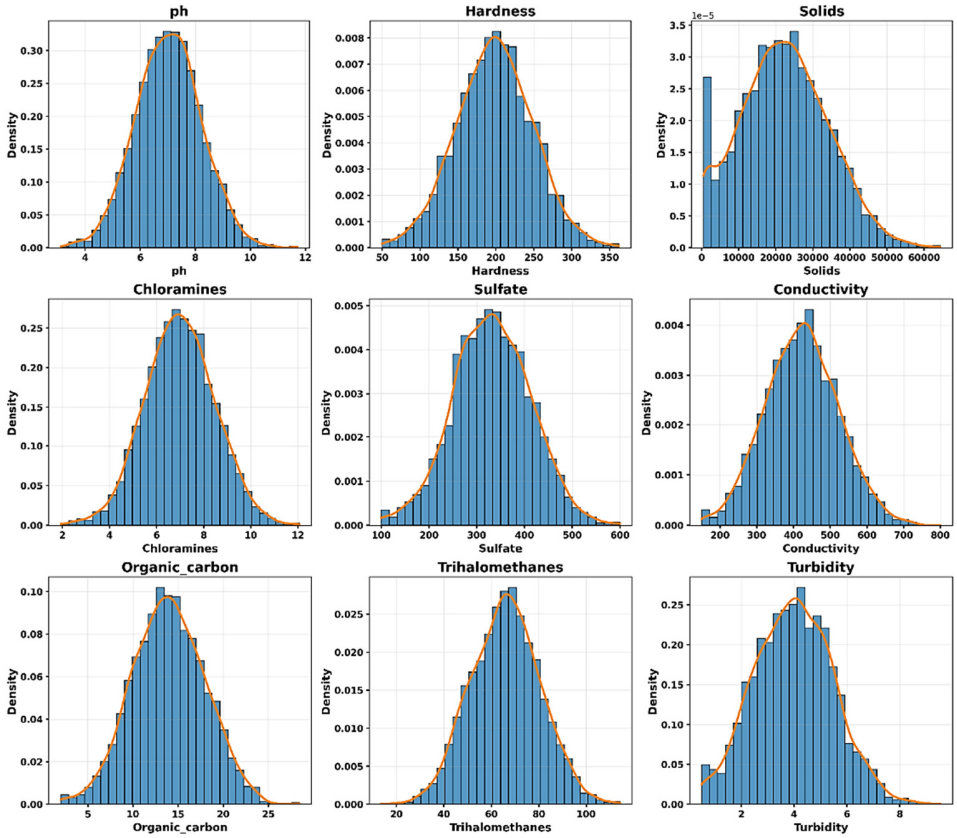

Fig. 2. Distributional profiles of raw water quality features.

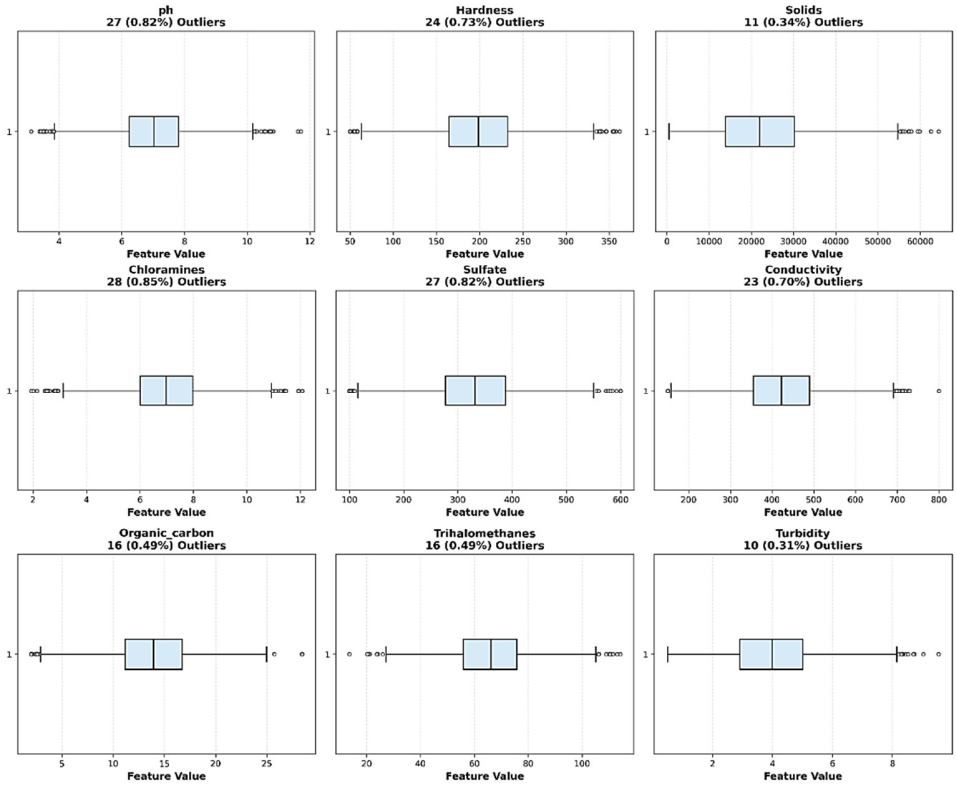

Fig. 3. Raw water quality characteristics box plots.

C. Mathematical Formulation of Preprocessing

Let the full dataset be expressed as:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

where $x_i \in \mathbb{R}^d$ represents the physicochemical feature vector and $y_i \in \{0,1\}$ denotes the potability label.

1) Stratified Train-Test Split

The dataset is first partitioned into disjoint training and test sets and represented as

$$D_{train} \cup D_{test} = D, \quad D_{train} \cap D_{test} = \emptyset \quad (2)$$

such that the class distribution of y is preserved and represented as

$$P(y \mid D_{train}) \approx P(y \mid D_{test}) \quad (3)$$

2) Mean Imputation

Let $X_{train} = [x_1, x_2, \ldots, x_n]$ denote the feature matrix of the training set. For each feature $j \in \{1, \ldots, d\}$, the mean is computed only on the training data and expressed as:

$$\mu_j = \frac{1}{|D_{train}|} \sum_{x_i \in D_{train}} x_{ij} \quad (4)$$

Missing values are imputed as:

$$\bar{x}_{ij} = \begin{cases} x_{ij}, & if\ x_{ij} \text{ is observed} \\ \mu_j, & if\ x_{ij} \text{ is missing} \end{cases} \quad (5)$$

The same μ_j values are applied to both the training and test sets.

3) Feature Standardization

For each feature j, the standard deviation is computed on the training set, represented as:

$$\sigma_j = \sqrt{\frac{1}{|D_{train}|} \sum_{x_i \in D_{train}} \left(x_{ij} - \mu_j\right)^2} \quad (6)$$

Standardization is then applied as:

$$z_{ij} = \frac{\bar{x}_{ij} - \mu_j}{\sigma_j}, \forall x_i \in D_{train} \cup D_{test} \quad (7)$$

Thus, no statistics from the test set influence the transformation.

4) SMOTE Applied Only Within Training Folds

During k-fold cross-validation, the training set is further split into folds represented as:

$$D_{train} = \bigcup_{k=1}^{K} D_{train}^{(k)} \quad (8)$$

For each fold k, the SMOTE algorithm is applied only to the training fold, as expressed in:

$$D_{train}^{(k)} \xrightarrow{SMOTE} \widehat{D}_{train}^{(k)} \quad (9)$$

The validation fold and test set remain unchanged and are expressed as:

$$D_{val}^{(k)} \cap \widehat{D}_{train}^{(k)} = \emptyset \quad (10)$$

5) Model Training and Evaluation

The classifier f_θ is trained on the resampled data, expressed as

$$f_\theta : \widehat{D}_{train}^{(k)} \rightarrow \{0, 1\} \quad (11)$$

and evaluated on unseen validation and test data.

D. Feature Selection

Feature selection was performed to determine the most essential features to utilize for the prediction of water potability. Correlation analysis was performed to determine whether there were redundant features, and Mutual Information (MI) was utilized to determine the dependency of the features and the target variable. MI is a nonlinear approach that determines the dependency of features and the target, and helps determine nonlinear dependencies.

1) Encoding Target Variable

The target variable Potability was in the binary format itself, i.e., is the water potable or not (potable = 1, non-potable = 0), which is the best format for binary classification models.

2) Handling Class Imbalance

The dataset shows class imbalance in the target variable (Potability). To overcome this issue, SMOTE was used, as it helps models to learn from all classes and reduce bias towards the majority class.

3) Standardization of Features

Standardization achieves zero-mean and unit-variance to enable the models to converge more quickly and avoid the domination of the learning process by features with a larger scale. In addition, it is essential, particularly when features exist with various units. The dataset was transformed into a clear, balanced, and well-structured format by applying preprocessing steps systematically, as discussed above. This improved the robustness, interpretability, and generalizability of the ML models developed.

III. ANALYTICAL METHODS

A. Descriptive Statistical Analysis

This is an initial step in understanding the nature and quality of the dataset before applying predictive models. It comprises different statistical measures, such as central tendency, dispersion, and distribution, and provides a detailed summary of the dataset. Table II presents a statistical summary of the data after preprocessing, including count, mean, standard deviation, minimum, first quartile (25%), median (50%), third quartile (75%), and maximum values. The following can be observed from Table II:

- pH: The mean pH is 7.08, indicating slightly acidic or alkaline water. The pH varies from slightly acidic (3.89) to slightly alkaline (10.26), indicating normal variation in water quality.
- Hardness: The average hardness is 196.39, which is typical for municipal water supplies. The hardness varies between 117.13 and 276.39, indicating moderate variation in the data.

- Solids and Conductivity: These two features exhibit wide variation. The Solids range between -2.70 and 2.69, and conductivity ranges from -2.89 to 2.88. This wide range suggests that normalization might be required for these features to maximize model performance.
- Chloramines and Sulphate: Both these parameters have reasonably normal distributions with very low standard deviations, which indicates a very small range of values. The means for Chloramines and Sulphate are 7.12 and 333.79, respectively, indicative of average concentration in potable water.
- Trihalomethanes: At a mean of 66.42, Trihalomethanes show slight positive skewing, probably because of a couple of high values with high concentrations. Values vary between 26.62 and 106.70.
- Turbidity: The turbidity varied from 1.85 to 6.09, and the mean value was 3.97, which shows the water quality in general follows drinking water standards.

TABLE II. DESCRIPTIVE STATISTICS OF WATER QUALITY FEATURES

Features	Count	Mean	Std	Min	25%	50%	75%	Max
pH	3276	7.08	1.38	3.89	6.28	7.08	7.87	10.26
Hardness	3276	196.39	32.02	117.13	176.85	196.97	216.67	276.39
Solids	3276	0.00	0.99	-2.70	-0.68	-0.03	0.67	2.69
Chloramines	3276	7.12	1.54	3.15	6.13	7.13	8.11	11.10
Sulphate	3276	333.79	31.77	267.16	317.09	333.78	350.39	400.32
Conductivity	3276	0.00	1.00	-2.89	-0.72	0.00	0.72	2.88
Organic Carbon	3276	14.28	3.29	5.33	12.07	14.22	16.56	23.30
Trihalomethanes	3276	66.42	15.49	26.62	56.65	66.40	76.67	106.70
Turbidity	3276	3.97	0.78	1.85	3.44	3.96	4.50	6.09

B. *Correlation and Multicollinearity*

1) *Correlation Analysis*

This is an important step in understanding the relationships between features in the dataset, as it helps identify features that have a strong positive or negative correlation and influence on the performance of predictive models [12]. Figure 1 presents the correlation matrix of water quality features in the raw dataset, indicating pairwise relationships and the degree of linear association among variables as follows:

- Potability and Other Features: The target variable, Potability, exhibits a weak correlation with most features, with the highest correlation being with Chloramine (0.024) and Solids (0.03). This indicates that no single feature can individually predict portability, but it requires complex relationships or feature combinations.
- Solids and Sulphate had a moderate negative correlation (–0.14). This shows that as the concentration of total dissolved solids increases, the Sulphate concentration tends to decrease, or vice versa.
- pH and Hardness display a positive correlation (0.089), indicating that more alkaline waters may have slightly higher hardness levels, although the effect is minimal.

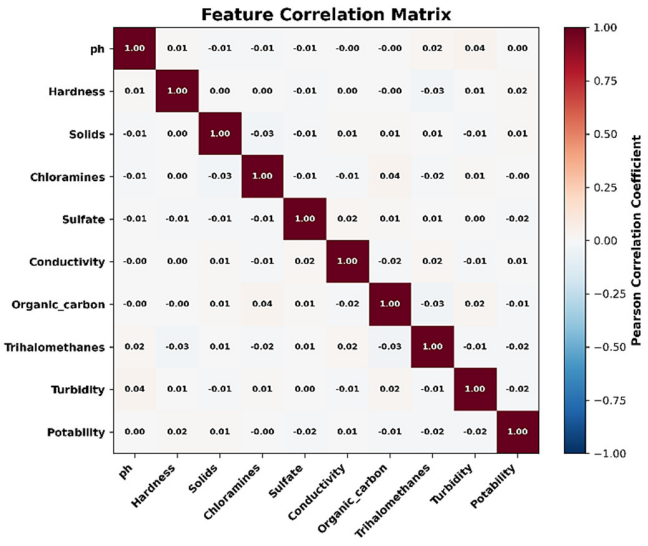

Fig. 4. Feature correlation matrix.

2) *Mutual Information (MI)*

MI is a method used to quantify the dependency between two random variables [13]. Unlike linear correlation coefficients, MI is capable of capturing both linear and nonlinear dependencies. MI is suitable for environmental datasets where relationships between physicochemical parameters and water potability are often complex and nonlinear. The MI between a predictor variable X and the target variable Y is:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (12)$$

where $p(x,y)$ denotes the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ represent the marginal distributions. The analysis is based on the higher value of MI that represents the strong dependency between the target and feature variables. Table III shows the MI scores.

TABLE III. MUTUAL INFORMATION OF PREDICTOR VARIABLES

Features	MI scores
pH	0.00170
Hardness	0.00180
Solids	0.00084
Chloramines	0.00230
Sulphate	0.00810
Conductivity	0.00076
Organic carbon	0.00053
Trihalomethanes	0.00025
Turbidity	0.00003

- Sulphate possesses the highest MI value (0.00810) for Potability, indicating a moderate relationship.
- Chloramines and hardness both possess relatively higher MI values (0.00230 and 0.00180, respectively), suggesting that they can play a part in determining potability.

- pH, Solids, and Conductivity have very low MI values, indicating that they do not contribute much information toward water potability.

3) Multicollinearity Between Predictor Variables

Multicollinearity can negatively affect regression stability and interpretability by increasing variances in coefficient estimators. Multicollinearity was identified by computing the Variance Inflation Factor (VIF) of all features after standardizing the data for numerical stability [14]. The VIF of a predictor variable quantifies the extent to which linear connections with other predictors inflate the variance of its predicted regression coefficient. Table IV displays the VIF calculated for each of the nine features and shows that there is very little multicollinearity among them.

TABLE IV. VARIANCE INFLATION FACTOR OF PREDICTOR VARIABLES

Features	VIF scores
pH	1.02
Hardness	1.02
Solids	1.04
Chloramines	1.01
Sulphate	1.03
Conductivity	1.00
Organic carbon	1.00
Trihalomethanes	1.00
Turbidity	1.00

- All predictor variables in the dataset have VIF values ranging from 1.00 to 1.04, indicating a lack of multicollinearity. This means that none of the features is strongly correlated with another, and each provides unique information to the model.
- The low VIF scores indicate that the dataset is in good condition for constructing ML models. No variable deletion or transformation is required because of multicollinearity issues, making the feature selection task more manageable.
- The independence between the predictor variables is helpful for a variety of ML models, including both linear and tree-based models. This feature will most probably enhance the interpretability and generalizability of models built for water potability prediction.

C. Feature Selection Techniques

Feature selection is an elementary process to train ML models only with the most relevant features while eliminating redundant, irrelevant, or highly collinear variables [15]. This not only helps in dimensionality reduction but also helps in improving the model's generalizability. This study systematically applied both filter-based and wrapper-based methods to identify relevant features to predict water potability.

1) Filter-Based Method

Filter methods assess feature importance using statistical criteria independently of any ML algorithm. Three filter-based approaches were utilized:

- The Correlation Matrix (Figure 4) was used to identify highly correlated features and redundancy.
- Mutual Information (MI - Table III) measures the dependency of each feature on the target variable. Features like pH, Sulphate, chloramines, and hardness show the highest MI scores, indicating strong relevance to water potability. MI values closer to zero are considered to have weak relevance, while higher values (e.g., >0.001) imply stronger relationships.
- Variance Inflation Factor (VIF - Table IV) measures multicollinearity among features. All features were in the vicinity of 1. Since the widely accepted threshold is 5 or 10, this indicates very minimal collinearity and supports their inclusion.

2) Wrapper-Based Methods

Wrapper-based methods evaluate different subsets of features on model performance, providing better results compared to filter-based methods at the cost of computational complexity. In this study, Recursive Feature Elimination (RFE) was employed with tree-based classifiers such as RF and XGBoost.

D. Comparative Summary and Inferences

Table V illustrates a comparative analysis of water quality attributes based on various feature selection metrics such as MI scores, VIF, and RFE based on RF and XGBoost algorithms.

TABLE V. WATER QUALITY FEATURE SELECTION USING MI, VIF, AND RFE

Feature	MI score	VIF	RFE (RF)	RFE (XGB)	Selected
pH	0.00173	1.02	✓	✓	✓
Hardness	0.00180	1.02	✓	✓	✓
Solids	0.00084	1.04	✓	✓	✓
Chloramines	0.00231	1.01	✓	✓	✓
Sulphate	0.00812	1.03	✓	✓	✓
Conductivity	0.00076	1.00	✓	✓	✓
Organic Carbon	0.00053	1.00	✓	✓	✓
Trihalomethanes	0.00025	1.00			
Turbidity	0.00003	1.00			

- Priority was given to features with MI values above 0.001 and VIF levels below 5, since they are essential predictors of water potability and are both meaningful and independent of predictors.
- Both RFE with RF and XGBoost consistently selected the following variables: pH (MI: 0.00173, VIF: 1.02), Hardness (MI: 0.00180, VIF: 1.02), Solids (MI: 0.00084, VIF: 1.04), Chloramines (MI: 0.00231, VIF: 1.01), Sulphate (MI: 0.00812, VIF: 1.03), and Organic Carbon (MI: 0.00053, VIF: 1.00). The consistent selection of these features by various wrapper methods highlights their importance and validity. Conductivity (MI: 0.00076, VIF: 1.00), Trihalomethanes (MI: 0.00025, VIF: 1.00), and Turbidity (MI: 0.00003, VIF: 1.00) were not chosen by either RFE method, implying limited predictive utility for water potability in this scenario.

- This comprehensive, multi-criteria approach enhances model interpretability, predictive ability, and generalizability in water quality assessment, ensuring that the final feature subset is trustworthy and informative.

E. Machine Learning Models

Multiple ML models were used to make water potability predictions, such as those presented in [16]. All of these models possess varying strengths and weaknesses, which were considered when their performance was being evaluated. Simple models, such as LR, and ensemble models, such as RF and GB, were used to determine their generalization ability on the data using F1-score, accuracy, and recall. Hyperparameter tuning was considered for each model.

Ensemble learning techniques were used to improve predictive robustness. This study used two categories of ensemble methods: (i) bagging-based and (ii) boosting-based approaches. The bagging-based ensemble method used is RF, where multiple decision trees are trained independently, and the bootstrap samples of the data and predictions are aggregated through majority voting. This approach reduces the variance and mitigates overfitting, making it well-suited for noisy environmental datasets. Boosting-based ensemble techniques were represented by GB and AdaBoost classifiers, which construct ensembles sequentially, and each weak learner focuses on correcting the errors of previously trained models. GB optimizes a differentiable loss function in a stage-wise manner, while AdaBoost adaptively adjusts sample weights to emphasize misclassified instances. Boosting approaches are effective in capturing complex nonlinear relationships present in physicochemical water quality data. This study did not consider advanced gradient-boosting variants, such as XGBoost, LightGBM, stacking, and blending ensemble strategies, as it focused primarily on interpretable methods in standard ML frameworks. These techniques maintain reproducibility, interpretability, and consistency with prior water quality prediction studies.

1) Logistic Regression (LR)

LR was selected as the baseline model due to its ease of interpretation, simplicity, and computational efficiency. It estimates the probability of water being potable using a sigmoid function that maps a linear combination of features. Although it makes a linearity assumption between the input features and the logarithmic odds of the target, it provides useful feature importance information in the form of its coefficient. LR is computationally efficient and well-suited to applications where speed and interpretability matter. Hyperparameter tuning here is mostly adjusting the regularization hyperparameter to prevent overfitting, maintaining its predictive power.

2) Support Vector Classifier (SVC)

SVC with Radial Basis Function (RBF) kernel was used to deal with non-linearly separable data. SVC is robust in high-dimensional space and works well when the data is not linearly separable. It is susceptible to feature scaling, which was resolved in preprocessing. Important hyperparameters in the context of SVC are the regularization parameter and the kernel for the generalization capability of the model and overfitting.

3) Decision Tree Classifier (DTC)

DT is a non-parametric classifier that partitions the feature space into regions according to decision rules learned from data. The output is interpretable as it follows the tree data structure, and the path can be traced from the root node to the leaf node. However, it is prone to overfitting, especially for deep trees. Hyperparameters like *max_depth*, *min_samples_split*, and *min_samples_leaf* were adjusted to restrict the depth of the tree and avoid overfitting. Pruning techniques like pre- and post-pruning were also used to avoid overfitting.

4) Random Forest Classifier (RFC)

RF is an ensemble learning technique that averages the predictions of many decision trees. It improves generalizability and reduces overfitting by averaging multiple trees. RFE was also used to select features. To get the ideal balance between complexity and generalizability, hyperparameters such as *max_depth*, *min_sample_split*, and the number of trees (*n_estimators*) are utilized for tuning.

5) Gradient Boosting Classifier (GBC)

GB constructs trees one at a time, sequentially, and each tree attempts to fix the mistakes of the previous one. This ensemble approach is very systematic in identifying intricate patterns and providing high accuracy. Hyperparameter tuning involves the learning rate, boosting iterations, and tree depth. There is a conviction that a lower learning rate and more trees generally give the best result.

6) AdaBoost Classifier (ABC)

AdaBoost is another boosting algorithm that combines weak classifiers into a strong model by repeatedly modifying the weights of the misclassified instances. It is particularly well-suited for skewed datasets. Hyperparameters like the number of estimators and the learning rate were tuned to improve the model's performance.

7) K-Nearest Neighbors (KNN)

KNN is a learner that makes predictions for new instances by looking at the majority class of their closest neighbors in feature space. KNN is very sensitive to the k (number of neighbors) value and needs feature standardization to ensure all features are equally important. KNN is expensive to run at inference time, but tuning hyperparameters like k can optimize its predictive performance.

IV. RESULTS AND ANALYSIS

A. Experimental Setup

All models were developed and evaluated using Python 3.12.1 with the necessary ML libraries such as scikit-learn. The experiments were carried out on a machine with an Intel Core i5 processor and 16 GB of RAM. The preprocessing of the dataset was performed by standardization, missing values imputation, and stratified splitting (80:20) to preserve the class distribution in the training and testing sets. RFE and correlation-based feature selection were used to minimize multicollinearity.

B. Evaluation Metrics

Several metrics were used to assess the performance of the ML models, which signify various aspects of classification quality, particularly valuable given the presence of imbalanced datasets and sensitive applications such as water potability prediction.

- Accuracy calculates the number of correctly classified samples (both positive and negative) as a proportion of total samples using:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (13)$$

Although accuracy is intuitive and most used, it can be deceptive in imbalanced datasets where the majority class overpowers. For instance, if 90% of the sample falls into one class, a naïve model will predict the majority class with 90% accuracy but miss minority class instances.

- Precision measures the ratio of correctly predicted positive samples to all samples as positive using:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (14)$$

High precision means that when the model classifies a sample as potable, it is most likely to be correct. This is important when false positives have serious consequences.

- Recall is particularly critical in water quality applications because it indicates the model's ability to identify water potability accurately. Low recall implies that numerous potable samples are not identified, which might result in unsafe water being misclassified as safe.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (15)$$

- F1 score is the harmonic mean of precision and recall, measuring the predictive performance using:

$$\text{F1} = 2 * \frac{\text{Recall}*\text{Precision}}{\text{Recall+Precision}} \quad (16)$$

It is especially useful when there is an unbalanced class distribution because it will penalize extreme class imbalance between precision and recall.

C. Hyperparameter Tuning

Hyperparameter tuning is a crucial step in the optimization of ML models to achieve the optimal predictive ability. In this study, every classifier was subjected to systematic hyperparameter tuning with Grid Search along with 5-fold cross-validation on the training set. This approach ensures that selected hyperparameters will generalize to new data by measuring model performance over every validation fold. The hyperparameter search space of each model was set by a mix of defaults established by prior research and a preliminary study to compromise between computation expense and extensive exploration. The actual parameter grids that each model explored are outlined below.

1) Logistic Regression (LR)

For the LR model, the C value for regularization was varied between 0.1, 1, and 10 to determine how various degrees of regularization affect the model. A low value of C adds more penalty to large coefficients and avoids overfitting, while a high value provides more freedom for the model to fit the training data. The penalty was to be $l2$, imposing ridge regularization, which is efficient at managing model complexity and can be helpful when features can be correlated. The solver used was *liblinear*, which can deal well with small to moderate-sized datasets and supports L2 regularization. These parameter choices were organized in a Python dictionary for efficient hyperparameter tuning.

$$para_{grid} = \{C: [0.1,1,10], penalty: ['l2'], \\ solver: ['liblinear']\}$$

2) Support Vector Classifier (SVC)

For the SVC, the regularization parameter C was tested with values 0.1, 1, and 10 to maintain the trade-off between opening up a larger margin and reducing classification errors. The kernel coefficient *gamma*, which controls the impact of specific data points on the decision boundary, was tested using the values *scale* (the default option), 0.01, and 0.001. These choices enable the model to be flexible about changing levels of data complexity. The kernel function was specified as *rbf* (radial basis function) to capture nonlinear relationships in the water quality dataset accurately. The selected hyperparameters were structured in a Python dictionary for systematic tuning.

$$para_{grid} = \{C: [0.1, 1.10], \gamma: ['0.01', '0.001'], \\ kernel: ['rbf']\}$$

3) Decision Tree Classifier (DTC)

For DTC, the *max_depth* parameter was varied with values of 3, 5, and 10 to control the depth of the tree and manage the balance between model complexity and overfitting. The criterion parameter was tested with both *gini* and *entropy* options to determine which impurity measure, Gini index or information gain, provides better splits for the dataset. These hyperparameters were organized in a Python dictionary for systematic tuning.

$$para\_grid = \{\max\_depth: [3,5,710], \\ criterion: ['gini','entropy']\}$$

4) Random Forest Classifier (RFC)

For RFC, *n_estimators* were chosen as 100 and 200 to compare the influence of ensemble size on prediction accuracy and stability. The *max_depth* parameter was varied from 5 to 10 to regulate the depth of individual trees, hence preventing overfitting. The *min_samples_split* parameter, which determines the minimum number of samples for splitting an internal node, was set at 2 and 5 to control tree growth and model generalization. The grid parameter was set as:

$$para_{grid}\{n\_estimators: [100,200], \\ max_{depth}: [5, 10], min\_sample\_split: [2,5]\}$$

5) Gradient Boosting Classifier (GBC)

For GBC, *n_estimators* were kept at 100 and 200 to study the effect of ensemble size on model performance. The learning rate, which controls how much each tree contributes, was kept at 0.05 and 0.1 so that learning speed and threat of overfitting could be balanced. The *max_depth* parameter was kept between 3 and 5 to regulate the complexity of each tree in the ensemble individually. The hyperparameter grid was organized as:

$$para\_grid = \{n\_estimators : [100,200], \\ learning_{rate}: [0.05,0.1], max\_depth : [3,5]\}$$

6) AdaBoost Classifier (ABC)

For ABC, the ensemble size (*n_estimators*) of the weak learners decreased from 100 to 50 to evaluate the impact of the ensemble size. The learning rate, which adjusts the contribution of each weak learner, was examined at 0.01, 0.1, and 1.0 levels to find the best rate for convergence and accuracy of the model. The grid parameter was set as:

$$para\_grid = \{n\_estimators : [50,100], \\ learning\_rate : [0.01,0.1,1] \}$$

7) K-Nearest Neighbors (KNN)

For KNN, neighbors (*n_neighbors*) were tried at 3, 5, 7, and 9 to determine the best neighborhood size for classification. The parameter weights were used with either *uniform*, where all neighbors are equally weighted, or *distance*, where neighboring points closer to the query point are given higher weights in their contribution to predictions. The hyperparameter grid was:

$$para\_grid = \{neighbors : [3,5,7,9], \\ weight : ['uniform', distance]\}$$

D. Model Performance and Comparison

Table VI shows the performance of the ML models trained on the original raw dataset without any normalization or class imbalance treatment. Using no standardization and class balancing, RFC performed better than all other models in all evaluation indicators, with the highest accuracy (0.669), precision (0.649), recall (0.637), and F1-score (0.640). KNN and GBC also performed quite well, with KNN performing slightly better than GBC both in accuracy and F1-score. The SVC was moderately efficient, outperforming DTC, ABC, and LR, which had the lowest ranking, demonstrating its poor applicability to raw, imbalanced data. In general, the results point out that ensemble techniques, especially RF, are stronger and better than single classifiers when used in raw water quality datasets with no preprocessing.

TABLE VI. PERFORMANCE COMPARISON OF ML MODELS ON THE TEST SET TRAINED IN RAW DATA

Model	ACC	PREC	REC	F1
LR	0.524	0.528	0.529	0.520
SVC	0.637	0.618	0.618	0.618
DTC	0.585	0.563	0.562	0.562
RFC	0.669	0.649	0.637	0.640
GBC	0.648	0.627	0.623	0.624
ABC	0.613	0.605	0.609	0.604
KNN	0.652	0.637	0.638	0.637

Table VII shows that models trained on standardized data using class-imbalance correction trends improve or remain steady in all performance metrics compared to the baseline variant trained with raw data. LR experiences significant gains in precision and F1-score, reflecting improved classification accuracy after these preprocessing operations. Both RFC and KNN have high precision and recall values, indicating better detection of minority class instances and model stability. The SVC gains the same accuracy as its baseline, but its precision, recall, and F1-score all increase, demonstrating that these preprocessing methods make models more predictive and reliable overall.

TABLE VII. PERFORMANCE OF ML MODELS ON THE TEST SET TRAINED ON STANDARDIZED AND CLASS-BALANCED DATA

Model	ACC_SC	PREC_SC	REC_SC	F1_SC
REG	0.524	0.552	0.524	0.530
SVC	0.637	0.637	0.637	0.637
DTC	0.585	0.584	0.585	0.584
RFC	0.669	0.662	0.669	0.662
GBC	0.648	0.643	0.648	0.645
ABC	0.613	0.627	0.613	0.617
KNN	0.652	0.655	0.652	0.654

Table VIII reports the mean and standard deviation of F1-scores obtained from stratified k=5 cross-validation for each classifier. The results of paired statistical tests are compared with each model against the RFC. For each classifier, fold-wise F1-scores from the same cross-validation splits were used, ensuring paired comparisons on identical data partitions. The paired differences between each model and RFC were evaluated using both the paired t-test and the Wilcoxon signed-rank test. The paired t-test assesses the differences in the mean F1-score. The Wilcoxon signed-rank test provides a non-parametric alternative that does not rely on distributional assumptions. The results indicate that KNN achieves the highest mean F1-score (0.714) with relatively low variability. The paired t-test p-value (0.005) suggests a statistically significant difference compared to RFC. REG, DTC, and ABC also exhibit statistically significant differences under the paired t-test, although their mean F1-scores are lower than that of RFC. In contrast, SVC and GBC do not show statistically significant differences under the paired t-test.

TABLE VIII. STATISTICAL SIGNIFICANCE OF MODEL PERFORMANCE COMPARED TO RFC (CROSS-VALIDATED F1-SCORE)

Model	Mean Acc.	Standard Acc.	t-test p	Wilcoxon p
LR	0.493	0.016	0.003	0.062
SVC	0.678	0.030	0.071	0.125
DTC	0.624	0.054	0.027	0.062
GBC	0.683	0.053	0.051	0.062
ABC	0.587	0.044	0.000	0.062
KNN	0.714	0.038	0.005	0.062

The Wilcoxon signed-rank test p-values are consistently greater than 0.05 across all models. This behavior arises from the small number of cross-validation folds. The Wilcoxon results are interpreted conservatively and used for parametric findings. These results highlight some models that differ in

mean performance in relation to RFC under parametric testing. The non-parametric analysis indicates more conservative conclusions, which underscore the stability of the RFC baseline.

V. DISCUSSION

A. Impact of Preprocessing on Model Performance

Feature standardization and SMOTE helped to handle the class imbalance and improve the predictive accuracy of most models. From Tables VI and VII, it can be observed that models trained on standardized and balanced data tend to have higher precision, recall, and F1 scores than similar models trained from raw data. This is particularly important for environmental datasets where minority classes, such as potable water samples, are undersampled. For instance, LR showed improved precision and F1-score following preprocessing, reflecting enhanced classification of drinkable water cases and fewer false positives. Similarly, RFC and KNN demonstrated notable improvements in recall and precision, showing more robust classification boundaries and enhanced sensitivity to the minority class.

B. Model Comparison and Suitability

Statistical significance was used to determine whether the differences found between the models were significant or due to random variation. The application of paired t-tests and Wilcoxon signed-rank tests to cross-validated F1 scores showed that some models, such as GBC and KNN, achieved marginally higher mean F1 Scores compared to RFC, but these differences did not show consistent statistical significance in both tests. This result suggests that with the variability of the data, the RFC remains a strong and reliable choice for solving this classification problem. Ensemble models like RFC and GBC consistently outperformed all other models on all evaluation metrics. Their capabilities to identify and capture nonlinear correlations between water quality features make them a good choice for this classification problem. Despite maintaining the same accuracy levels, preprocessing helped the SVC perform better by striking a better balance between precision and recall.

C. Practical Implications

Preprocessing improved models' recall and F1-scores, demonstrating their usefulness in practical applications, where avoiding false negatives, i.e., incorrectly labeling hazardous water as safe, is crucial for maintaining public health. The findings are consistent with recent research that emphasizes the necessity of preprocessing tasks involving unbalanced classification in the context of environmental health [16].

D. Limitations and Future Work

Although the results of this study are promising, the small size and limited diversity of the dataset could limit the applicability of the model to other water sources and geographic locations. Improving the robustness and practicality of the model will require further studies to employ larger and more diverse datasets that cover a wider range of places, conditions, and water quality metrics. Further, incorporating XAI techniques, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), may improve interpretability and assist decision-makers in comprehension and trust towards the model output. Additional research must also address overcoming deployment issues and improving the model for real-time use, particularly in low-resource environments. Advances in these areas will help develop robust, scalable, and efficient water quality monitoring systems that can facilitate timely and informed interventions in various environmental contexts. In general, these results demonstrate the critical role of appropriate preprocessing and model selection in developing reliable water quality prediction systems.

VI. CONCLUSION

This study investigated the prediction of water potability based on physicochemical water quality attributes with a variety of supervised ML models. The models were tuned to produce stable and generalizable performance through strict preprocessing, such as handling class imbalance, and systematic hyperparameter tuning. The results showed that ensemble methods such as RFC and GBC outperformed other models on unseen test data in terms of accuracy, precision, recall, and F1-score. Preprocessing significantly improved sensitivity to the minority class. Simpler models, such as LR, also offered practical interpretability to prove their application in scenarios where transparency is the priority. The results support the merging of sophisticated ML methods to improve the estimation of water quality. Future research will prioritize a larger and more diverse and larger dataset that covers more geographical areas and develop explainable AI models to further improve predictive accuracy and model interpretability. These developments may lead to real-time water quality monitoring systems, thereby enhancing public health and environmental safety.

DATA AVAILABILITY

All data is available upon request from the corresponding author.

ACKNOWLEDGMENTS

This paper is derived from a research grant funded by the Research, Development, and Innovation Authority (RDIA), Kingdom of Saudi Arabia, with grant number 13382-psu-2023-PSNU-R-3-1-EI-. The authors would like to acknowledge the support of Prince Sultan University, Riyadh, Saudi Arabia, in paying the article processing charges of this publication. This research is supported by the Automated Systems and Computing Lab (ASCL), Prince Sultan University, Riyadh, Saudi Arabia.

REFERENCES

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

[2] A. Aldrees, M. F. Javed, A. T. Bakheit Taha, A. Mustafa Mohamed, M. Jasiński, and M. Gono, "Evolutionary and ensemble machine learning predictive models for evaluation of water quality," *Journal of Hydrology: Regional Studies*, vol. 46, Apr. 2023, Art. no. 101331, https://doi.org/10.1016/j.ejrh.2023.101331.

[3] M. Koranga, P. Pant, T. Kumar, D. Pant, A. K. Bhatt, and R. P. Pant, "Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand," *Materials Today: Proceedings*, vol. 57, pp. 1706–1712, 2022, https://doi.org/10.1016/j.matpr.2021.12.334.

[4] M. Zhu *et al.*, "A review of the application of machine learning in water quality evaluation," *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, June 2022, https://doi.org/10.1016/j.eehl.2022.06.001.

[5] P. William, O. J. Oyebode, G. Ramu, M. Gupta, D. Bordoloi, and A. Shrivastava, "Artificial Intelligence based Models to Support Water Quality Prediction using Machine Learning Approach," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Aug. 2023, pp. 1496–1501, https://doi.org/10.1109/ICCPCT58313.2023.10245020.

[6] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," *Computation*, vol. 11, no. 2, Jan. 2023, Art. no. 16, https://doi.org/10.3390/computation11020016.

[7] A. F. Zambrano, L. F. Giraldo, J. Quimbayo, B. Medina, and E. Castillo, "Machine learning for manually-measured water quality prediction in fish farming," *PLOS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256380, https://doi.org/10.1371/journal.pone.0256380.

[8] M. Y. Shams, A. M. Elshewey, E. S. M. El-kenawy, A. Ibrahim, F. M. Talaat, and Z. Tarek, "Water quality prediction using machine learning models based on grid search method," *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 35307–35334, Sept. 2023, https://doi.org/10.1007/s11042-023-16737-4.

[9] S. Patel, K. Shah, S. Vaghela, M. Aglodiya, and R. Bhattad, "Water Potability Prediction Using Machine Learning." In Review, May 25, 2023, https://doi.org/10.21203/rs.3.rs-2965961/v1.

[10] N. S. Pagadala, M. Marri, A. Myla, B. Abburi, and K. S. Ramtej, "Water Quality Prediction Using Machine Learning Techniques," in *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, Mar. 2023, pp. 358–362, https://doi.org/10.1109/SPIN57001.2023.10117415.

[11] A. Kadiwal, "Water Quality." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability.

[12] P. Kamath B., G. Sharma, A. Bongale, D. Dharrao, and M. Seitshiro, "Exploratory Data Analysis and Water Potability Classification using Supervised Machine Learning Algorithms," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20898–20903, Apr. 2025, https://doi.org/10.48084/etasr.8904.

[13] R. J. May, G. C. Dandy, H. R. Maier, and J. B. Nixon, "Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems," *Environmental Modelling & Software*, vol. 23, no. 10, pp. 1289–1299, Oct. 2008, https://doi.org/10.1016/j.envsoft.2008.03.008.

[14] M. Na, X. Liu, Z. Tong, B. Sudu, J. Zhang, and R. Wang, "Analysis of water quality influencing factors under multi-source data fusion based on PLS-SEM model: An example of East-Liao River in China," *Science of The Total Environment*, vol. 907, Jan. 2024, Art. no. 168126, https://doi.org/10.1016/j.scitotenv.2023.168126.

[15] N. G. Rezk, S. Alshathri, A. Sayed, and E. El-Din Hemdan, "EWAIS: An Ensemble Learning and Explainable AI Approach for Water Quality Classification Toward IoT-Enabled Systems," *Processes*, vol. 12, no. 12, Dec. 2024, Art. no. 2771, https://doi.org/10.3390/pr12122771.

[16] A. Zaghini *et al.*, "A Pragmatic Approach for Chlorine Decay Modeling in Multiple-Source Water Distribution Networks Based on Trace Analysis," *Water*, vol. 16, no. 2, Jan. 2024, Art. no. 345, https://doi.org/10.3390/w16020345.