Journal of
Water & Climate Change

© 2024 The Authors

Journal of Water and Climate Change Vol 15 No 2, 431 doi: 10.2166/wcc.2023.403

Water quality prediction: a data-driven approach exploiting advanced machine learning algorithms with data augmentation

Karthick K[a,*], S. Krishnan[b] and R. Manikandan[c]

[a] Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh, India
[b] Department of EEE, Mahendra Engineering College (Autonomous), Namakkal, Tamil Nadu, India
[c] Department of ECE, Panimalar Engineering College, Chennai, India
*Corresponding author. E-mail: karthick.k@gmrit.edu.in; kkarthiks@gmail.com

KK, 0000-0001-7755-5715

ABSTRACT

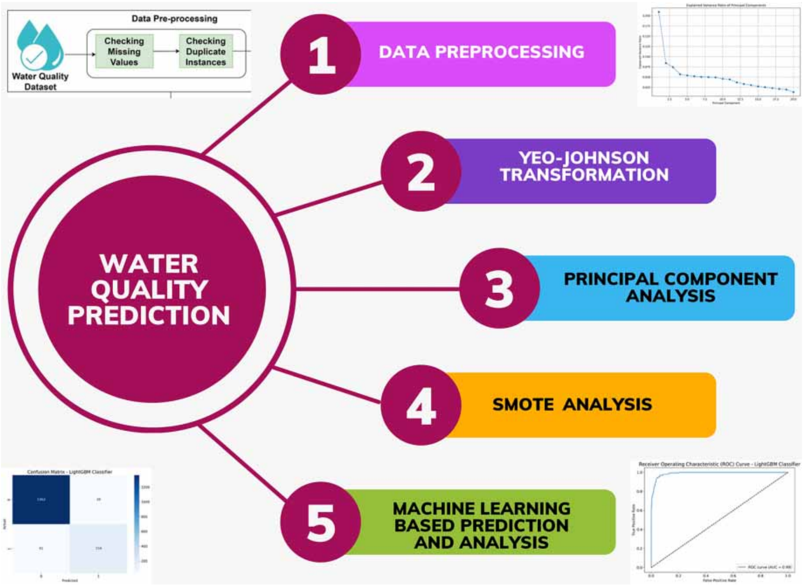
Water quality assessment plays a crucial role in various aspects, including human health, environmental impact, agricultural productivity, and industrial processes. Machine learning (ML) algorithms offer the ability to automate water quality evaluation and allow for effective and rapid assessment of parameters associated with water quality. This article proposes an ML-based classification model for water quality prediction. The model was tested with 14 ML algorithms and considers 20 features that represent various substances present in water samples and their concentrations. The dataset used in the study comprises 7,996 samples, and the model development involves several stages, including data preprocessing, Yeo–Johnson transformation for data normalization, principal component analysis (PCA) for feature selection, and the application of the synthetic minority over-sampling technique (SMOTE) to address class imbalance. Performance metrics, such as accuracy, precision, recall, and F1 score, are provided for each algorithm with and without SMOTE. LightGBM, XGBoost, CatBoost, and Random Forest were identified as the best-performing algorithms. XGBoost achieved the highest accuracy of 96.31% without SMOTE and had a precision of 0.933. The application of SMOTE enhanced the performance of CatBoost. These findings provide valuable insights for ML-based water quality assessment, aiding researchers and professionals in decision-making and management.

Key words: machine learning, principal component analysis (PCA), synthetic minority over-sampling technique (SMOTE), water quality, water quality classification

HIGHLIGHTS

- Valuable contribution to the field of water quality prediction.
- Proposed ML classification model for water quality prediction with 20 features and tested 14 ML algorithms with hyperparameter optimization.
- Applied Yeo–Johnson transformation for data normalization, PCA for feature selection, and SMOTE to address class imbalance.

GRAPHICAL ABSTRACT


Data Pre-processing
Water Quality Dataset
Checking Missing Values
Checking Duplicate Instances
WATER QUALITY PREDICTION
1
DATA PREPROCESSING
2
YEO-JOHNSON TRANSFORMATION
3
PRINCIPAL COMPONENT ANALYSIS
4
SMOTE ANALYSIS
5
MACHINE LEARNING BASED PREDICTION AND ANALYSIS
Explained Variance Ratio of Principal Components
Confusion Matrix - LightGBM Classifier
Receiver Operating Characteristic (ROC) Curve - LightGBM Classifier

1. INTRODUCTION

Water quality encompasses the chemical, physical, biological, and radiological characteristics of water, which determine its suitability for various purposes, including drinking, irrigation, industrial use, and ecosystem health (Saleh & Hassan 2021). Water quality assessment is essential for people's health, the environment, crop production, and industrial operations. Access to clean and safe drinking water is essential for maintaining human health. Poor water quality can contain harmful substances such as bacteria, viruses, heavy metals, pesticides, and chemical pollutants, leading to waterborne diseases and health problems (Lin *et al.* 2022). Water quality also has a significant impact on aquatic ecosystems and biodiversity. Contaminated water can harm aquatic plants, animals, and organisms, disrupt ecosystem balance, and degrade habitats (Amoatey & Baawain 2019). In the agricultural sector, water quality is vital for irrigation and crop growth. Poor-quality water can have negative effects on crop growth, soil fertility, and overall agricultural productivity (Dotaniya *et al.* 2023). Furthermore, various industries rely on water for their operations. If the water quality is poor, it can damage machinery, reduce efficiency, and pose risks to industrial processes. Assessing water quality is essential to ensure that water is suitable for industrial purposes (Akhtar *et al.* 2021).

Machine learning (ML) algorithms offer the capability to automate water quality assessment, eliminating the need for manual and time-consuming analysis. This leads to efficient and rapid evaluation of water quality parameters (Kaddoura 2022). By analysing large datasets, ML algorithms can identify intricate patterns and relationships among different water quality parameters, uncovering hidden patterns that traditional analytical methods may miss (Ghobadi & Kang 2023). These algorithms are adept at detecting anomalies and deviations from normal water quality patterns, enabling early identification of potential issues or contamination events (Zehra *et al.* 2023). This proactive approach allows for prompt interventions and measures to ensure water safety. By training on historical water quality data, ML models can predict future water quality conditions, facilitating proactive planning and resource allocation for effective water resource management (Varadharajan *et al.* 2022). Moreover, by integrating ML algorithms into decision support systems, real-time water quality monitoring and decision-making support can be provided to water managers, policymakers, and stakeholders Giupponi & Sgobbi (2013). This integration enhances the ability to make informed decisions and take appropriate actions to protect and manage water resources.

The major contributions of this research are as follows:

✓ In this research work, we propose an ML-based classification model for water quality prediction. We tested 14 ML algorithms and considered a total of 20 features for model development. These features represent various substances present in water samples, and their concentrations are used to predict water safety. The inclusion of multiple features enhances the accuracy and robustness of the model.

✓ The classification model was developed using 7,996 data samples and involved several stages. These stages include data preprocessing, and Yeo–Johnson transformation to stabilize variance, improve the normality of the data, and mitigate the impact of outliers. This leads to better model interpretability and performance.

✓ Principal component analysis (PCA) was used for feature selection, and the synthetic minority over-sampling technique (SMOTE) was employed to address class imbalance in the dataset. SMOTE helps improve the performance of the ML models, particularly in correctly classifying instances of the minority class.

✓ The paper provides detailed performance metrics such as accuracy, precision, recall, and F1 score for each algorithm, both with and without the application of SMOTE.

✓ To ensure optimal performance on the water quality prediction task, we employed RandomizedSearchCV, a technique for hyperparameter optimization with five-fold cross-validation. This helps us find the best combination of hyperparameters for each ML algorithm, fine-tuning the models.

The methodology and findings presented in the article make a valuable contribution to the existing knowledge in the field of water quality prediction. They offer insights into the application of ML algorithms and their effectiveness in water quality assessment. These results have practical implications for researchers and practitioners engaged in water quality management and decision-making processes. By utilizing ML algorithms, this research provides a more efficient and automated approach to evaluating water quality parameters. The performance metrics and evaluation of different algorithms allow for informed decision-making regarding the selection of appropriate models for water quality prediction.

1.1. Related work

Samsudin *et al.* (2019) aimed to construct prediction models for the marine water quality index (MWQI) in the mangrove and estuarine zones of West Malaysia using advanced ML algorithms. The study utilized data from six monitoring stations in West Malaysia from 2011 to 2015, comprising 13 parameters. Spatial discriminant analysis (SDA) was employed to identify the significant parameters for MWQI development, and artificial neural network (ANN) and multiple linear regression (MLR) models were used for prediction. The SDA identified seven significant parameters for MWQI development, namely dissolved oxygen (DO), total suspended solids (TSS), oil and grease (O&G), phosphate (PO_4), cadmium (Cd), chromium (Cr), and zinc (Zn). These parameters were used to develop prediction models for MWQI using ANN and MLR. The SDA–ANN model exhibited higher R^2 values for training (0.9044) and validation (0.7113) compared with the SDA–MLR model. Additionally, the SDA–ANN model demonstrated a lower root mean square error (RMSE) than the SDA–MLR model.

Ahmed *et al.* (2019) discussed the development of a water quality prediction model for the Johor River Basin. The study proposed the implementation of artificial intelligence (AI) techniques to create a flexible mathematical structure capable of identifying non-linear and complex relationships between input and output data. The authors evaluated different modelling approaches, including the adaptive neuro-fuzzy inference system (ANFIS), radial basis function artificial neural networks (RBF-ANNs), and multi-layer perceptron artificial neural networks (MLP-ANNs). The study focused on three primary water quality parameters: ammoniacal nitrogen (AN), suspended solids (SS), and pH. Three evaluation techniques or assessment processes are employed to assess the impacts of the models. The first assessment process involves partitioning neural network connection weights to determine the significance of each input parameter. The second and third assessment processes determine the most effective input parameter for constructing the models, using either a single parameter or a combination of parameters, respectively. Two scenarios, Scenario 1 and Scenario 2, are introduced based on the value of the 12 input parameters. Scenario 1 predicts water quality parameters at each station, while Scenario 2 develops a prediction model based on the value of the same parameter at the previous station (upstream). The recommended model by them satisfactorily predicts all water quality parameters, with R^2 values equal to or greater than 0.9.

Chou *et al.* (2018) presented an ML approach for predicting water quality in reservoirs using Carlson's trophic state index (CTSI). The study utilized data collected over ten years from 20 reservoirs in Taiwan. Four AI techniques, consisting of ANNs, support vector machines (SVMs), classification and regression trees, and linear regression, are employed to analyse the data and predict the CTSI. The article also discusses ensemble models, which combine multiple individual models to improve predictive performance. Different ensemble methods, including voting, bagging, stacking, and tiering, are employed. The tiering method divides the samples into two classes based on the CTSI values, and a weighted average is used to evaluate prediction performance. Furthermore, a hybrid metaheuristic regression model is proposed, combining the firefly algorithm (MetaFA) and least squares support vector regression (LSSVR), to optimize the hyperparameters of the LSSVR model. The results indicate that the ensemble models and the hybrid metaheuristic regression model outperform the individual baseline models, achieving an RMSE of 4.707 and an MAPE of 7.9 for the voting-based approach.

Elkiran *et al.* (2019) discussed the application of various AI models and ensemble techniques for modelling DO levels in the Yamuna River, India. The study compared the performance of three single AI models: back propagation neural network (BPNN), ANFIS, and SVM, along with a linear auto-regressive integrated moving average (ARIMA) model. Additionally, three ensemble techniques – simple average ensemble (SAE), weighted average ensemble (WAE), and neural network ensemble (NNE) – were employed for multi-step ahead modelling. The study used data on DO, biological oxygen demand (BOD), chemical oxygen demand (COD), discharge (Q), pH, ammonia (NH_3), and water temperature (WT) recorded at three different stations along the Yamuna River. The performance of the models was evaluated using the determination coefficient (DC) and RMSE. The results indicated that the ANFIS model performed the best among the single models, with an average increase in performance accuracy of up to 7% and 19% for two of the stations. For the third station, the SVM model performed better. Among the ensemble techniques, the NNE approach achieved the highest improvement in performance, increasing the average accuracy by up to 14% in the verification phase. Table 1 provides the survey on water quality classification models.

Class imbalance is a prevalent issue in water quality prediction, where one class (e.g., safe or not safe) is significantly underrepresented compared with the other. Most of the existing prediction models do not explicitly address this issue, leading to biased models with poor performance on the minority class. Additionally, many existing articles tend to focus on limited or specific parameters such as DO, pH, and SS. However, to develop a comprehensive water quality prediction model, it is crucial to consider a wide range of parameters that capture the complexity of water quality characteristics. Another research issue that we identified in some existing works is the lack of emphasis on identifying appropriate features.

In our research work, we addressed these unaddressed issues by employing the SMOTE approach to tackle class imbalance in the dataset. The proposed model was developed by incorporating 20 features that represent various substances present in water samples and their concentrations, resulting in a more comprehensive representation of water quality parameters. We carefully selected these features using PCA. To evaluate the performance of water quality prediction, we assessed 14 ML algorithms, including LightGBM, XGBoost, CatBoost, and Random Forest. By incorporating these advanced algorithms, we conducted a more thorough assessment of their effectiveness compared with the existing studies.

Table 1 | Survey on water quality classification models

Author	Model	Features	Accuracy	F1 score	Precision
Nasir *et al.* (2022)	Logistic regression	Includes a total of seven features: DO, conductivity, nitrate, BOD, faecal coliform, and total coliform	0.7291	0.7249	0.7247
	SVM		0.8068	0.80601	0.81302
	Decision tree		0.81623	0.8156	0.8169
	XGBoost		0.8807	0.8804	0.8836
	MLP		0.8863	0.8864	0.8890
	Random Forest		0.9393	0.9394	0.9397
	CATBoost		0.9451	0.9449	0.9458
	Ensemble approach – Meta XGBoost		0.9621	0.96208	0.96215
Juna *et al.* (2022)	Logistic regression	Includes a total of nine features: pH, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, and turbidity	0.48	0.48	0.48
	Support vector classifier		0.52	0.47	0.54
	Decision tree		0.72	0.72	0.72
	Random Forest		0.79	0.79	0.79
	kNN classifier		0.57	0.56	0.55
	XGBoost		0.76	0.76	0.76
Hmoud Al-Adhaileh & Waselallah Alsaade (2021)	Feed-forward neural network (FFNN)	Includes a total of eight features: temperature, DO, pH, conductivity, biochemical oxygen demand, nitrate and nitrite levels, faecal coliform, and total coliform	1.00	–	0.99961
	kNN		0.8063	–	0.8250

2. METHODOLOGY

The proposed ML-based classification model is illustrated in Figure 1. The model encompasses several stages, including data preprocessing, Yeo–Johnson transformation for data normalization, PCA for feature selection, and the creation of a balanced dataset using SMOTE analysis on the training set. The performance of the ML classification model is evaluated using both the original raw dataset (which is imbalanced) and the SMOTE-based balanced dataset. The ML algorithms were tuned with hyperparameters using RandomizedSearchCV and five-fold cross-validation. The optimized models were trained on the training set and evaluated on the test set. The following sections will provide a comprehensive explanation of each stage in the process.

2.1. Data

The water quality dataset was accessed from the Kaggle open-source digital library (Data Link 2021). The dataset consists of 7,999 instances and 21 attributes in total. Table 2 provides information on the levels of various substances in the water samples. The objective is to predict the safety of the water samples based on these features. Each attribute in the dataset represents a substance found in the water samples, and its threshold value determines whether the concentration of that substance is hazardous or not. The 'is_safe' attribute serves as the target variable for this classification model, aiming to predict the safety of the water sample based on the concentrations of the different compounds.

Table 3 presents a comprehensive summary of the dataset, including various statistical measures for each attribute. Each row corresponds to a specific attribute, while the columns provide information such as the count, mean, standard deviation, minimum value, quartiles (25%, 50%, 75%), and maximum value.

The count column indicates the number of non-null values available for each attribute, giving us an indication of data completeness. The mean represents the average value of the attribute, providing a measure of its central tendency. The standard deviation provides information about the spread or variability of the attribute's values.

The minimum and maximum values provide insights into the range of values observed in each attribute. The quartiles (25%, 50%, 75%) offer information about the data distribution and help identify the spread of values around the median.

Analysing these statistical measures allows us to gain a better understanding of the characteristics and variations present in the dataset. This information is crucial for further analysis, modelling, or making informed decisions based on the data.

2.2. Data preprocessing

The dataset has been examined to identify the presence of missing values in each feature. Missing values and duplicate values can introduce noise and inconsistencies, leading to a negative impact on the performance and accuracy of the ML model. By removing these values, we ensure the integrity and quality of the data (Emmanuel *et al.* 2021). Additionally, missing values and duplicate values can increase the computational complexity and training time required for the model.

In the specific 'ammonia' attribute of this dataset, there are three missing values. However, no duplicate instances have been identified. After the removal of the missing values, the updated dataset consists of 7,996 instances with 21 features.

Figure 2 presents the distribution of the cleaned dataset based on the 'is_safe' attribute. It illustrates that 88.6% of the dataset has a value of 0, indicating 'not safe,' while 11.4% of the dataset has a value of 1, indicating 'safe.'

2.3. Yeo–Johnson transformation

Figure 3(a)–3(t) displays the box plots of feature distributions in the dataset. Box plots are excellent visualizations that provide a summary of the value distribution for a specific attribute. They show the lowest, first quartile, median, third quartile, and highest values, as well as any potential outliers.

These box plots are significant in the context of the dataset as they allow us to gain insights into the distribution and variability of each attribute. By examining the box plots, we can observe the range of values, detect the presence of outliers, and understand the overall shape of the distribution for each attribute.

Box plots enable us to understand the variation, central tendency, and skewness of the data. They can identify potential data concerns such as extreme values, data asymmetry, or the presence of outliers, which can impact the analysis or modelling process. They also help identify features that may require additional preprocessing, normalization, or outlier handling before performing data analysis or developing ML models.

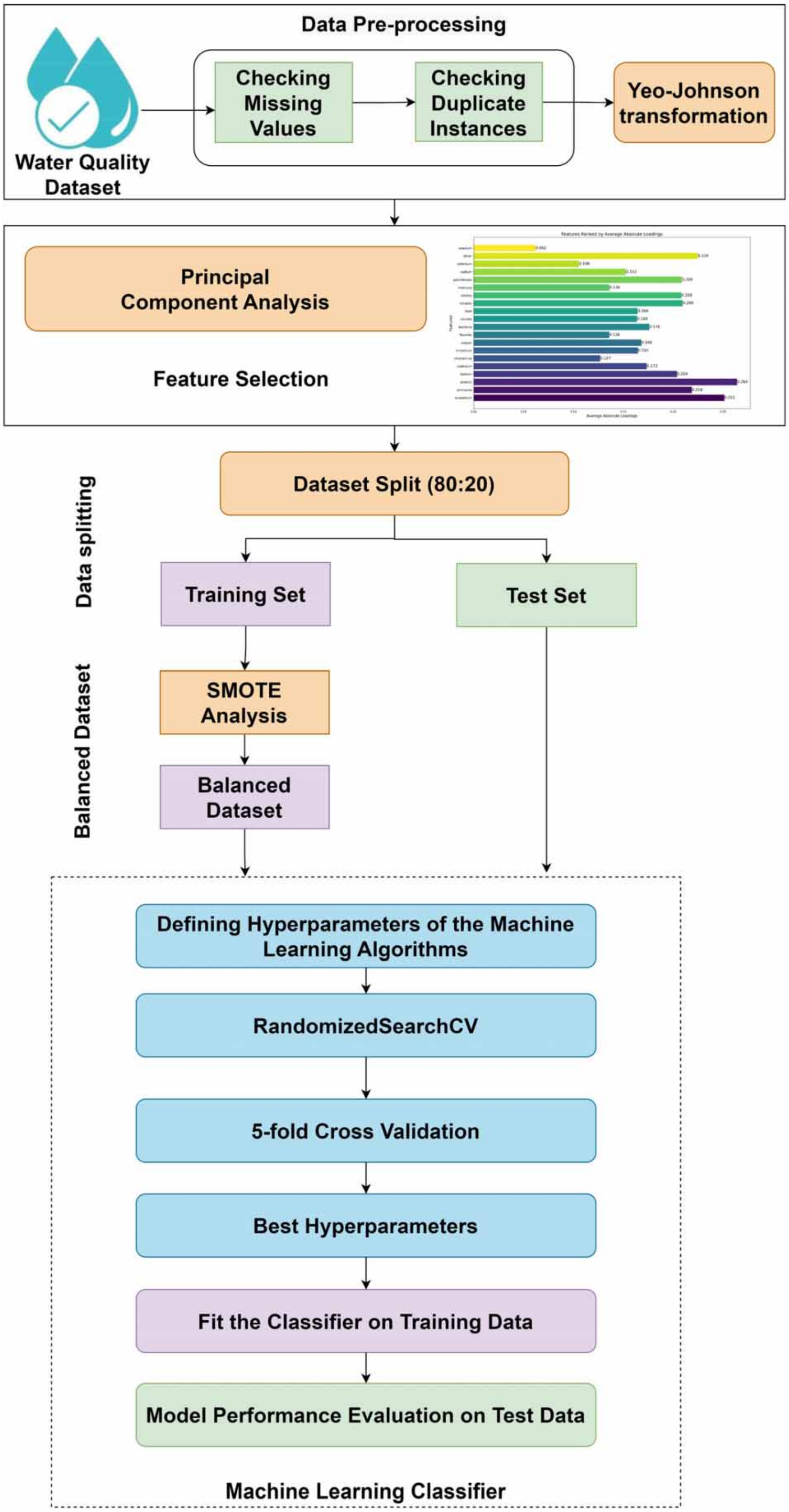

Data Pre-processing
Water Quality Dataset
Checking Missing Values
Checking Duplicate Instances
Yeo-Johnson transformation
Principal Component Analysis
Feature Selection
Features Ranked by Average Absolute Loadings
Average Absolute Loadings
Features
Data splitting
Dataset Split (80:20)
Training Set
Test Set
Balanced Dataset
SMOTE Analysis
Balanced Dataset
Defining Hyperparameters of the Machine Learning Algorithms
RandomizedSearchCV
5-fold Cross Validation
Best Hyperparameters
Fit the Classifier on Training Data
Model Performance Evaluation on Test Data
Machine Learning Classifier

Figure 1 | Block diagram of the proposed approach.

Table 2 | Dataset description

Feature	Attribute type	Description
Aluminium	Chemical	The concentration of aluminium in the water sample. It is considered dangerous if the concentration is greater than 2.8 mg/L
Ammonia	Chemical	The concentration of ammonia in the water sample. It is considered dangerous if the concentration is greater than 32.5 mg/L
Arsenic	Chemical	The concentration of arsenic in the water sample. It is considered dangerous if the concentration is greater than 0.01 mg/L
Barium	Chemical	The concentration of barium in the water sample. It is considered dangerous if the concentration is greater than 2 mg/L
Cadmium	Chemical	The concentration of cadmium in the water sample. It is considered dangerous if the concentration is greater than 0.005 mg/L
Chloramine	Chemical	The concentration of chloramine in the water sample. It is considered dangerous if the concentration is greater than 4 mg/L
Chromium	Chemical	The concentration of chromium in the water sample. It is considered dangerous if the concentration is greater than 0.1 mg/L
Copper	Chemical	The concentration of copper in the water sample. It is considered dangerous if the concentration is greater than 1.3 mg/L
Flouride	Chemical	The concentration of fluoride in the water sample. It is considered dangerous if the concentration is greater than 1.5 mg/L
Bacteria	Biological	The concentration of bacteria in the water sample. It is considered dangerous if the concentration is greater than 0 CFU/mL
Viruses	Biological	The concentration of viruses in the water sample. It is considered dangerous if the concentration is greater than 0 PFU/mL
Lead	Chemical	The concentration of lead in the water sample. It is considered dangerous if the concentration is greater than 0.015 mg/L
Nitrates	Chemical	The concentration of nitrates in the water sample. It is considered dangerous if the concentration is greater than 10 mg/L
Nitrites	Chemical	The concentration of nitrites in the water sample. It is considered dangerous if the concentration is greater than 1 mg/L
Mercury	Chemical	The concentration of mercury in the water sample. It is considered dangerous if the concentration is greater than 0.002 mg/L
Perchlorate	Chemical	The concentration of perchlorate in the water sample. It is considered dangerous if the concentration is greater than 56 mg/L
Radium	Radiological	The concentration of radium in the water sample. It is considered dangerous if the concentration is greater than 5 mg/L
Selenium	Chemical	The concentration of selenium in the water sample. It is considered dangerous if the concentration is greater than 0.5 mg/L
Silver	Chemical	The concentration of silver in the water sample. It is considered dangerous if the concentration is greater than 0.1 mg/L
Uranium	Radiological	The concentration of uranium in the water sample. It is considered dangerous if the concentration is greater than 0.3 mg/L
is_safe	Class	The attribute is the class attribute that indicates the safety of the water sample. It has two values: 0 represents 'not safe' and 1 represents 'safe'

Upon analysing the box plots, it is observed that the 'aluminium' and 'arsenic' features have outliers. The skewness and kurtosis values for 'aluminium' are found to be 2.013463 and 2.723089, respectively, while for the 'arsenic' feature, they are 1.985241 and 2.684839, respectively. To stabilize variance and improve the normality of the data, the Yeo–Johnson transformation (Yeo & Richard 2000) is applied. This transformation addresses the non-normality or heteroscedasticity in the selected features, mitigates the impact of outliers, reduces skewness, and improves the overall interpretability and

Table 3 | Dataset characteristics

Index	Count	Mean	SD	Min	25%	50%	75%	Max
Aluminium	7,999	0.666	1.265	0	0.04	0.07	0.28	5.05
Arsenic	7,999	0.161	0.253	0	0.03	0.05	0.1	1.05
Barium	7,999	1.568	1.216	0	0.56	1.19	2.48	4.94
Cadmium	7,999	0.043	0.036	0	0.008	0.04	0.07	0.13
Chloramine	7,999	2.177	2.567	0	0.1	0.53	4.24	8.68
Chromium	7,999	0.247	0.271	0	0.05	0.09	0.44	0.9
Copper	7,999	0.806	0.654	0	0.09	0.75	1.39	2
Flouride	7,999	0.772	0.435	0	0.405	0.77	1.16	1.5
Bacteria	7,999	0.320	0.329	0	0	0.22	0.61	1
Viruses	7,999	0.329	0.378	0	0.002	0.008	0.7	1
Lead	7,999	0.099	0.058	0	0.048	0.102	0.151	0.2
Nitrates	7,999	9.819	5.541	0	5	9.93	14.61	19.83
Nitrites	7,999	1.330	0.573	0	1	1.42	1.76	2.93
Mercury	7,999	0.005	0.003	0	0.003	0.005	0.008	0.01
Perchlorate	7,999	16.460	17.687	0	2.17	7.74	29.48	60.01
Radium	7,999	2.921	2.323	0	0.82	2.41	4.67	7.99
Selenium	7,999	0.050	0.029	0	0.02	0.05	0.07	0.1
Silver	7,999	0.148	0.144	0	0.04	0.08	0.24	0.5
Uranium	7,999	0.045	0.027	0	0.02	0.05	0.07	0.09

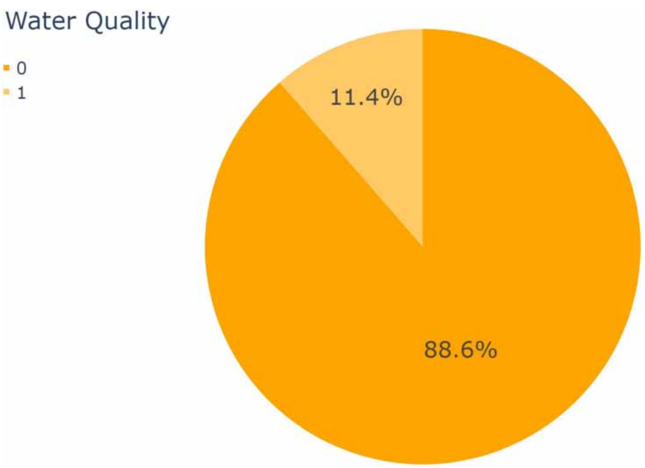


Figure 2 | Distribution of safe and not-safe samples in the dataset.

performance of the models applied to the dataset. The Yeo–Johnson transformation is defined in Equation (1):

$$\psi(y,\lambda)=\begin{cases}\dfrac{(y+1)^{\lambda}-1}{\lambda} & y\geq 0 \text{ and } \lambda \neq 0\\ \log(y+1) & y\geq 0 \text{ and } \lambda = 0\\ \dfrac{(-y+1)^{(2-\lambda)}-1}{2-\lambda} & y<0 \text{ and } \lambda \neq 2\\ -\log(-y+1) & y<0 \text{ and } \lambda = 2\end{cases} \tag{1}$$

Here ψ is concave in 'y' for power parameter $\lambda < 1$ and convex for power parameter $\lambda > 1$.

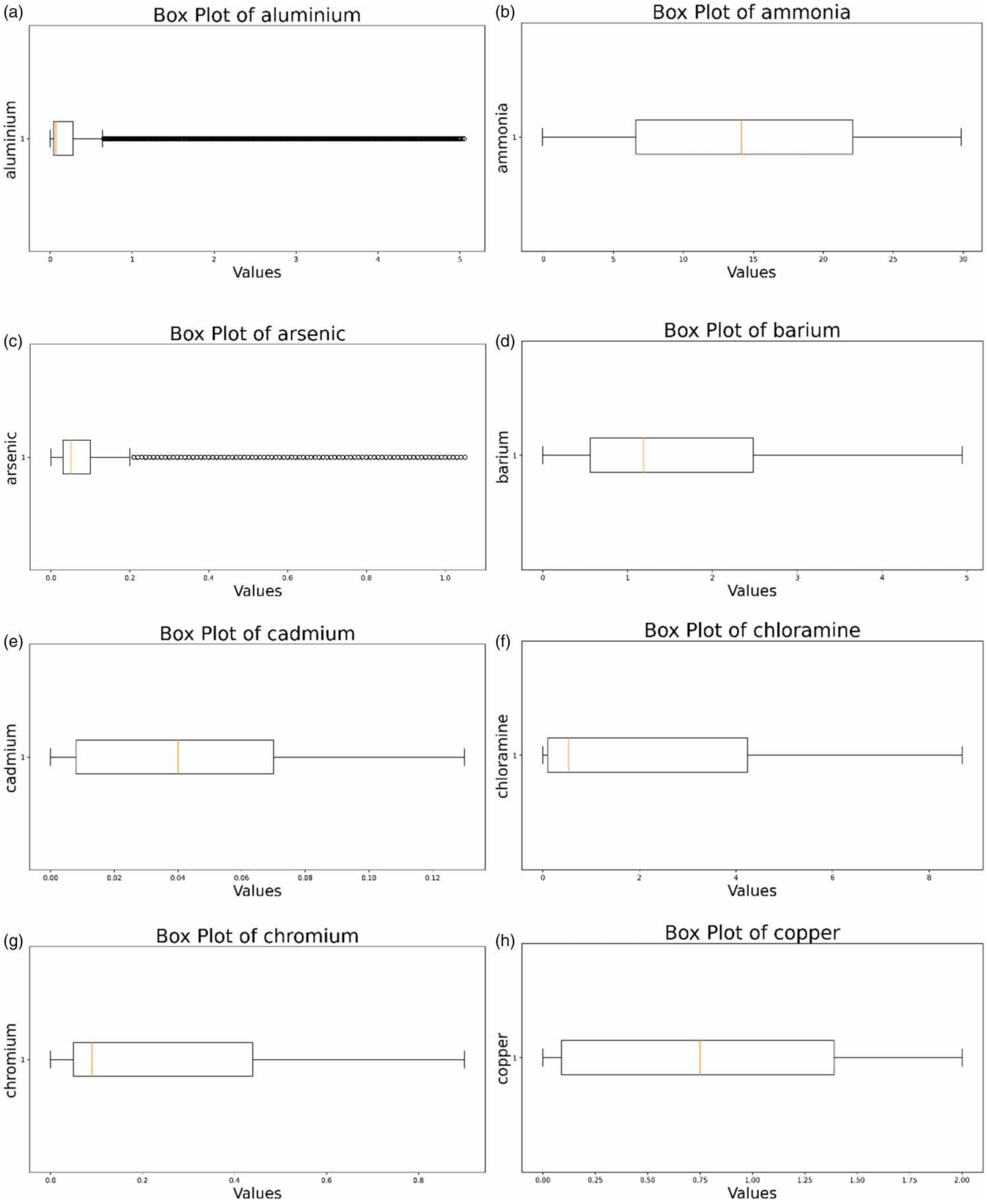


Figure 3 | Box plots of feature distributions in the dataset: (a) aluminium, (b) ammonia, (c) arsenic, (d) barium, (e) cadmium, (f) chloramine, (g) chromium, (h) copper, (i) fluoride, (j) bacteria, (k) viruses, (l) lead, (m) nitrates, (n) nitrites, (o) mercury, (p) perchlorate, (q) radium, (r) selenium, (s) silver, and (t) uranium. (*continued*)

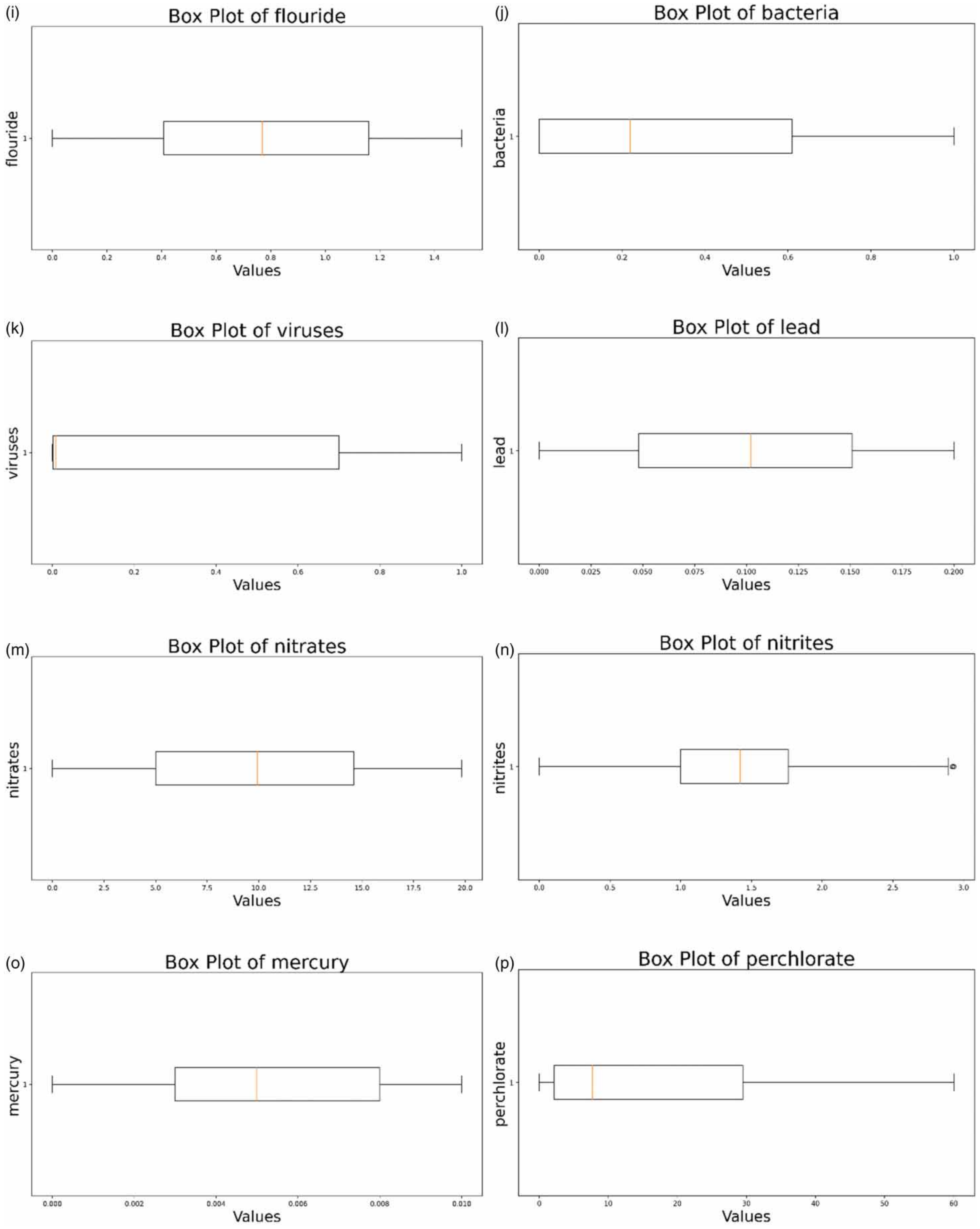


Figure 3 | Continued.

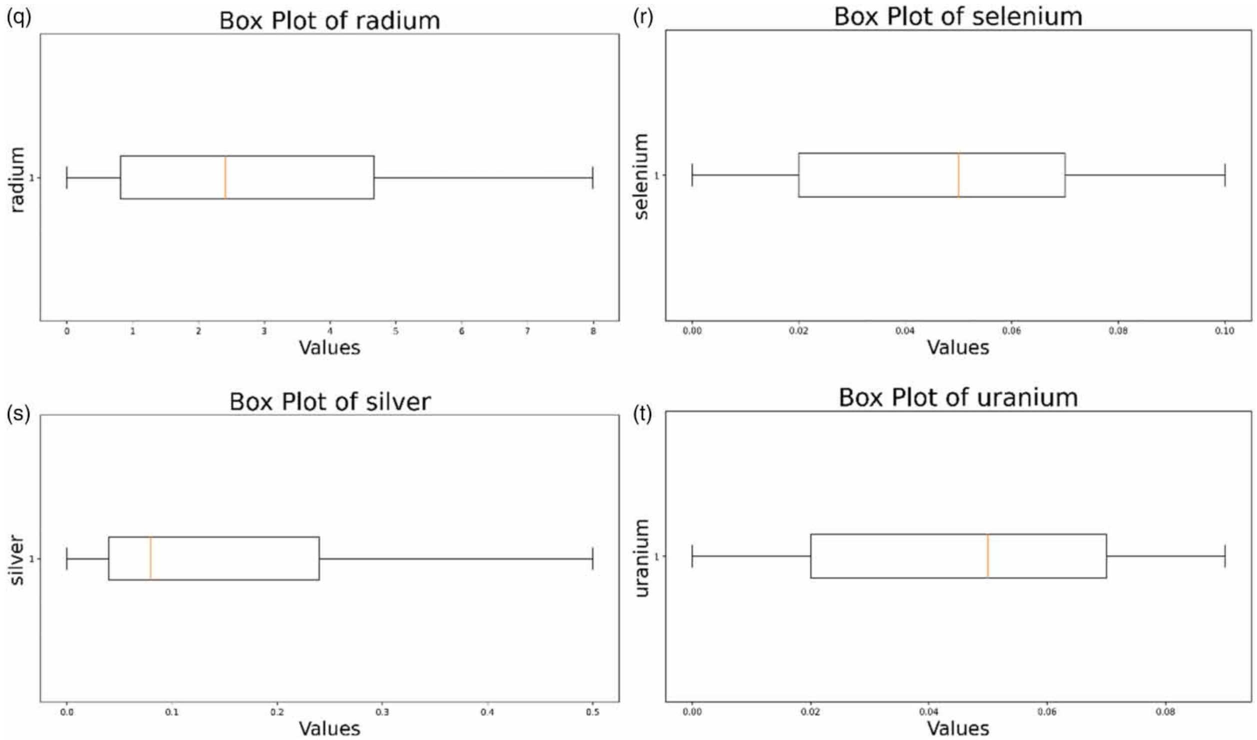


Figure 3 | Continued.

Table 4 provides the skewness and kurtosis values of the features in the dataset. Skewness measures the asymmetry of the distribution, while kurtosis measures the peakedness or flatness of the distribution (Cain *et al.* 2017). After applying the Yeo–Johnson transformation, it is observed that the skewness and kurtosis values of the 'aluminium' and 'arsenic' features have been reduced. The skewness value of 'aluminium' is now 1.055551, and the kurtosis value is −0.56711. Similarly, for 'arsenic', the skewness value is 0.825192, and the kurtosis value is −0.60079. These reduced values indicate a decrease in skewness and a shift towards a more normal distribution for both features.

2.4. Principal component analysis

The PCA method has been employed to identify the most influential features for water quality prediction. Figure 4 shows the explained variance ratio of each principal component (PC). PCs are linear combinations of the original features that capture the maximum amount of variance in the dataset. The explained variance ratio of each PC indicates the proportion of the total variance in this water quality data that is captured by that particular PC (Jolliffe & Cadima 2016).

PC1 is the principal component that explains the highest amount of variance in the data. It represents the direction in the feature space along which the data vary the most. The explained variance ratio of PC1 (0.2090 in this case) indicates the proportion of the total variance in the data that is accounted for by PC1.

Similarly, PC2 represents the second direction in the feature space that captures the next highest amount of variance in the data. The explained variance ratio of PC2 (0.0841 in this case) indicates the proportion of the total variance accounted for by PC2. The same pattern continues for the remaining 20 principal components, as indicated in Figure 4.

However, it is important to note that the explained variance ratio values provided in the PCA results do not directly indicate the importance or strength of individual features in PC1.

To assess the average influence of each feature, Figure 5 shows the average absolute loading of all the features in the dataset. The average of the absolute loadings for each feature has been calculated across all principal components. This provides a measure of the average influence of each feature (Peres-Neto *et al.* 2003). By comparing the values, it is observed that the 'uranium' feature has the lowest average absolute loading value. Therefore, it has been decided to discard this feature from the ML-based classification model development.

Table 4 | Skewness and kurtosis values of the features

Feature	Skewness	Kurtosis
Aluminium	1.055551	– 0.56711
Ammonia	0.026804	– 1.22892
Arsenic	0.825192	– 0.60079
Barium	0.660957	– 0.7038
Cadmium	0.478226	– 0.99193
Chloramine	0.888123	– 0.68213
Chromium	1.028188	– 0.37287
Copper	0.253558	– 1.35389
Flouride	– 0.03973	– 1.17328
Bacteria	0.554417	– 1.1403
Viruses	0.424567	– 1.59092
Lead	– 0.0606	– 1.15475
Nitrates	– 0.04206	– 1.1886
Nitrites	– 0.49821	– 0.35877
Mercury	– 0.08173	– 1.17095
Perchlorate	0.937767	– 0.4951
Radium	0.548391	– 0.92515
Selenium	0.010495	– 1.09943
Silver	1.029489	– 0.29239
Uranium	– 0.02704	– 1.17055

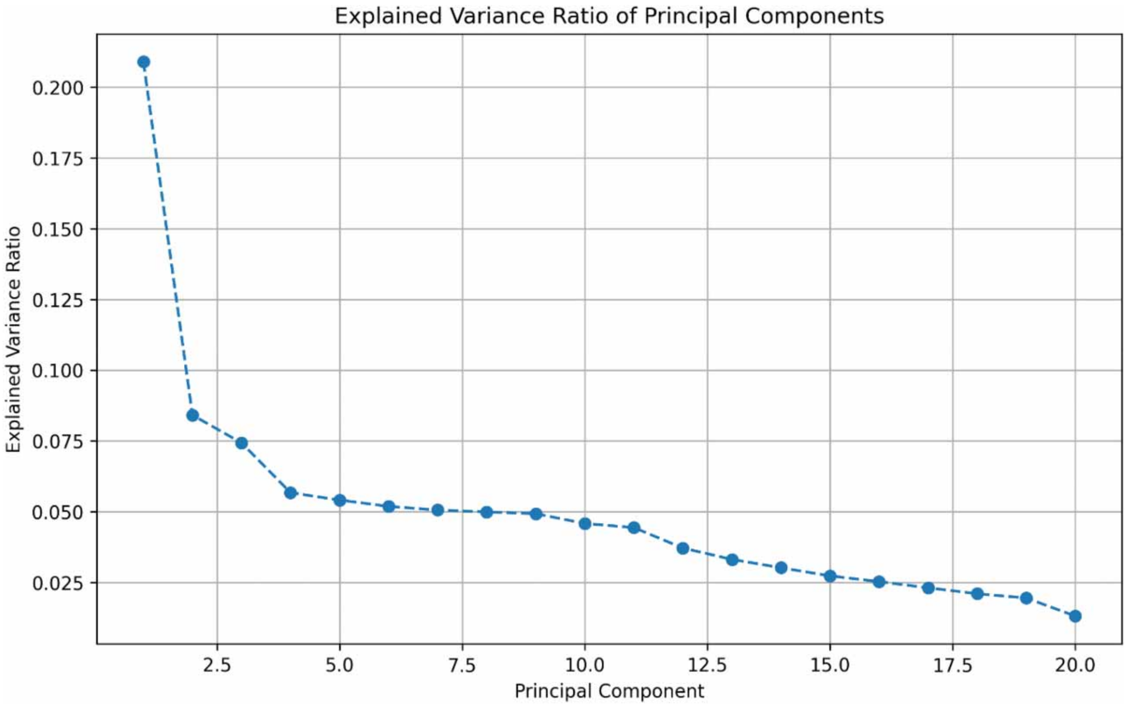


Figure 4 | Explained variance ratio of principal components.

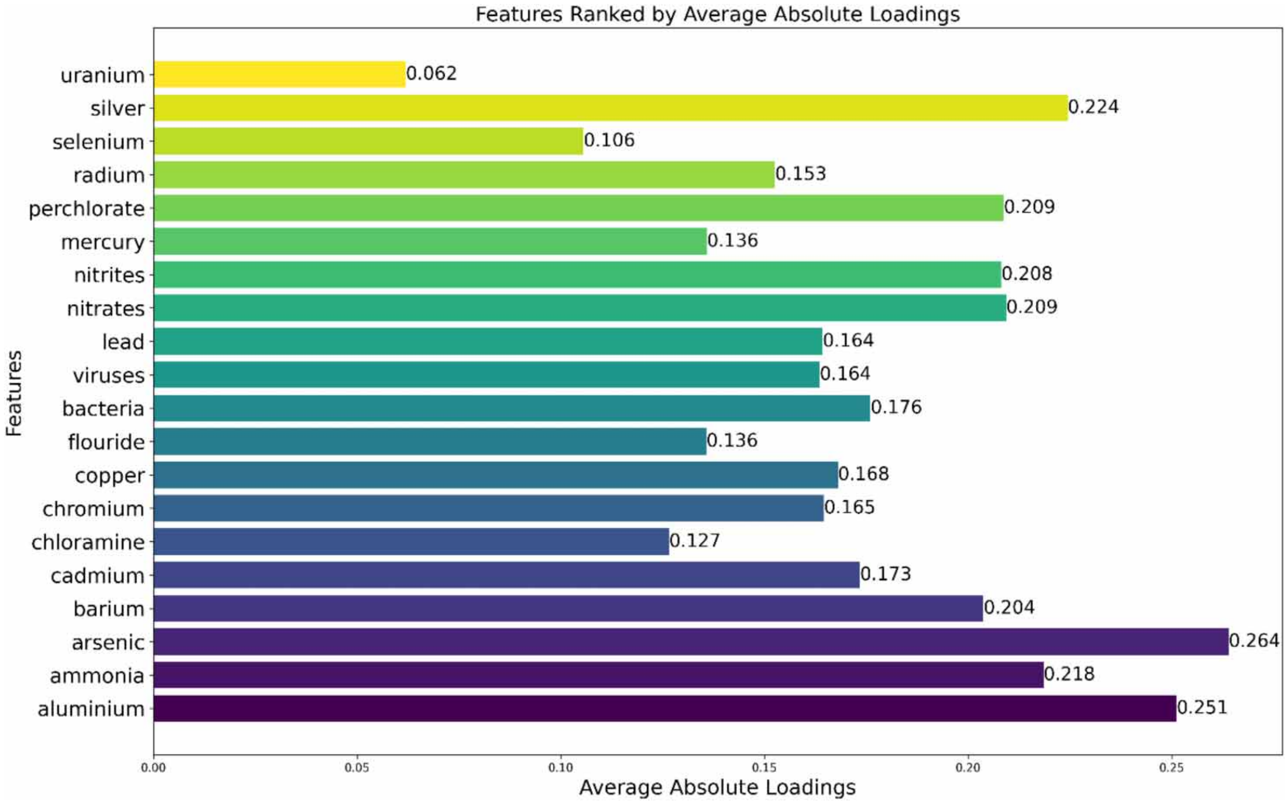


Figure 5 | Average absolute loading of all the features of the dataset.

2.5. Dataset split up

The dataset was split into a training set and a test set using an 80:20 ratio. The random_state parameter was set to 42 to ensure reproducibility by fixing the random seed. The test set provides us with unseen data to evaluate the performance of our trained classification model. The split between the training and test sets is essential for assessing the model's performance, determining its generalization capability, and guiding the model improvement process (Xu & Goodacre 2018).

The water quality data training set consists of 6,396 instances, while the test set consists of 1,600 instances. A total of 20 features have been considered for model development.

2.6. SMOTE analysis

SMOTE is an algorithm used for data augmentation, particularly in addressing class imbalance. Class imbalance occurs when one class, known as the minority class, has significantly fewer instances compared with the other class, known as the majority class. This imbalance can present challenges for ML algorithms as they tend to be biased towards the majority class, leading to poor predictive performance for the minority class (Li *et al.* 2021).

SMOTE is an over-sampling technique that tackles class imbalance by generating synthetic samples of the minority class, thereby increasing its representation in the dataset (Guo *et al.* 2018). In the training set, there are 5,684 instances of Class 0 (not safe) and 712 instances of Class 1 (safe). This indicates a binary class imbalance. By applying SMOTE analysis, the class counts for both safe and not-safe instances are balanced to 5,684 in the training set.

After fitting the data to the training set, the model's performance will be evaluated on the test set, which comprises 1,600 instances.

2.6.1. Steps involved in SMOTE on training data

The SMOTE transformation on the training data involves the following steps:

Step 1: Selection of Minority Class Instances

SMOTE begins by selecting instances belonging to the minority class, which in this dataset corresponds to Class 1 (safe). These instances serve as the starting point for generating synthetic samples.

Step 2: Identification of Nearest Neighbours

SMOTE identifies the k nearest neighbours for each selected minority class instance. The neighbours are determined based on the feature space, considering the attributes of the instances.

Step 3: Generation of Synthetic Samples

Synthetic samples are created by interpolating between the selected minority class instance and its k nearest neighbours. The interpolation process generates new instances that bridge the gap between existing instances, effectively expanding the minority class representation.

Step 4: Determination of Synthetic Sample Count

The number of synthetic samples to generate is determined based on the difference between the minority class and majority class instances. In this case, the difference is 4,972, and SMOTE ensures that enough synthetic samples are created to match the majority class count.

Step 5: Achieving Class Balance

The training set is transformed to achieve an equal representation of both the minority class (safe) and majority class (not safe). The class counts for both safe and not-safe instances are balanced, resulting in a count of 5,684 instances for each class in the training set.

The utilization of the SMOTE-based balanced dataset helps to address the issue of class imbalance, resulting in a more equitable and representative dataset for training the ML model (Douzas *et al.* 2018). By mitigating the imbalance, the model can effectively learn from both classes and make more accurate predictions.

2.7. Machine learning

2.7.1. LightGBM

LightGBM is a gradient-boosting framework that utilizes tree-based learning algorithms. It is specifically designed to be efficient, providing faster training speeds and higher accuracy compared with other boosting algorithms (Machado *et al.* 2019). LightGBM employs a leaf-wise growth strategy and incorporates features such as parallel learning and exclusive feature bundling, which contribute to its speed and performance. It is particularly effective when working with large datasets and has gained popularity in various domains due to its ability to efficiently handle high-dimensional data.

In the direction of the gradient space G, from the input space X: a training set is assumed with instances such as x_1, x_2, and up to x_n, where every attribute is a vector in the space X with s dimensions. All loss function negative gradients corresponding to the output model are represented as $g_1, g_2,..., g_n$ in each restatement of a gradient boosting.

Let O represent a set of data for training of a decision tree; the mean squared error of dividing measure j at a point d is stated as:

$$V_{j|o}(d) = \frac{1}{n_o}\left(\frac{\left(\sum\{X_i \in O{:}X_{ij\leq d}\}g_i\right)^2}{n^j_{l|o}(d)} + \frac{\left(\sum\{X_i \in O{:}X_{ij>d}\}g_i\right)^2}{n^j_{r|o}(d)}\right) \quad (2)$$

where $n_o = \sum I[X_i \in o]$, $n^j_{l|o}(d) = \sum I[X_i \in o{:}\ X_{ij\leq d}]$, and $n^j_{r|o}(d) = \sum I[X_i \in O{:}X_{ij>d}]$.

In the context of water quality prediction, LightGBM achieved an accuracy of 96.25% and a recall of 0.74 without utilizing SMOTE. This indicates that LightGBM was effective in accurately classifying instances of water quality, minimizing false negatives. The algorithm's optimized hyperparameters, obtained through RandomizedSearchCV, contributed to its robust predictive capabilities (Li 2020).

2.7.2. XGBoost

XGBoost is an optimized gradient-boosting algorithm renowned for its speed and performance. It combines the benefits of gradient boosting and regularization techniques, leading to highly accurate models (Tarwidi *et al.* 2023). XGBoost has gained popularity in ML competitions owing to its capability to handle diverse data types and capture intricate patterns.

The XGBoost algorithm is adopted from and the objective function is defined as

$$o = \sum_{i=1}^{n} L(y_i,\ F(x_i)) + \sum_{k=1}^{t} R(f_k) + C \tag{3}$$

where $R(f_k)$ represents the regularization term at the kth iteration and C is a constant, which can be excluded selectively, and $R(f_k)$ is denoted as:

$$R(f_k) = \alpha H + \frac{1}{2}\ \eta \sum_{j=1}^{n} w_j^2 \tag{4}$$

where α denotes the leaf complexity, H represents the number of leaves, η represents the penalty variable, and w_j is each leaf node output result.

In water quality prediction, XGBoost achieved the highest accuracy of 96.31%, which was slightly higher than LightGBM. However, it exhibited excellent precision with a score of 0.933. This indicates that XGBoost was highly precise in identifying positive instances of water quality. The optimized hyperparameters of the algorithm, obtained through RandomizedSearchCV, significantly contributed to its performance.

2.7.3. CatBoost

CatBoost is an efficient gradient-boosting algorithm that is specifically designed to handle categorical features. It utilizes an ordered boosting algorithm, which takes into account the order of categorical variables, leading to improved accuracy in

Table 5 | Best performing ML algorithms on water quality prediction

	Without SMOTE				**With SMOTE**			
Model name	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
LightGBM	0.9625	0.949	0.74	0.831	0.9506	0.807	0.795	0.801
XGBoost	0.9631	0.933	0.76	0.837	0.9475	0.805	0.765	0.785
CatBoost	0.9544	0.938	0.68	0.788	0.9468	0.821	0.735	0.775
Random Forest	0.9525	0.955	0.65	0.774	0.935	0.737	0.745	0.741

Table 6 | Performance of other ML algorithms on water quality prediction

	Without SMOTE			**With SMOTE**		
Model name	**Accuracy**	**ROC AUC**	**F1 Score**	**Accuracy**	**ROC AUC**	**F1 Score**
DecisionTreeClassifier	0.94	0.86	0.94	0.92	0.85	0.92
QuadraticDiscriminantAnalysis	0.91	0.83	0.92	0.89	0.87	0.90
GaussianNB	0.87	0.78	0.88	0.82	0.79	0.84
LinearDiscriminantAnalysis	0.92	0.78	0.93	0.85	0.82	0.86
SGDClassifier	0.93	0.76	0.92	0.82	0.80	0.85
AdaBoostClassifier	0.93	0.76	0.92	0.88	0.80	0.89
ExtraTreeClassifier	0.89	0.76	0.89	0.85	0.77	0.86
LogisticRegression	0.92	0.74	0.91	0.84	0.82	0.86
BernoulliNB	0.86	0.74	0.87	0.73	0.74	0.77
RidgeClassifier	0.90	0.61	0.88	0.85	0.82	0.86

predictions (Hancock & Khoshgoftaar 2020). CatBoost also offers built-in handling of missing values and robustness to outliers, enhancing its performance in real-world datasets.

In the water quality prediction dataset, CatBoost achieved an accuracy of 95.44% without the use of SMOTE. It demonstrated improvements in precision, recall, and F1 score compared with the other algorithms, indicating a balanced performance across multiple evaluation metrics. The optimized hyperparameters, obtained through RandomizedSearchCV, played a crucial role in enhancing the algorithm's overall performance.

2.7.4. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It constructs a multitude of decision trees and then aggregates their predictions to obtain the final result (Breiman 2001). Random Forest is well known for its robustness, scalability, and ability to handle high-dimensional data, making it a popular choice in various domains.

In the water quality prediction dataset, Random Forest achieved an accuracy of 95.25% without the use of SMOTE. Although it had a slightly lower accuracy compared with LightGBM and XGBoost, it still performed well. Random Forest

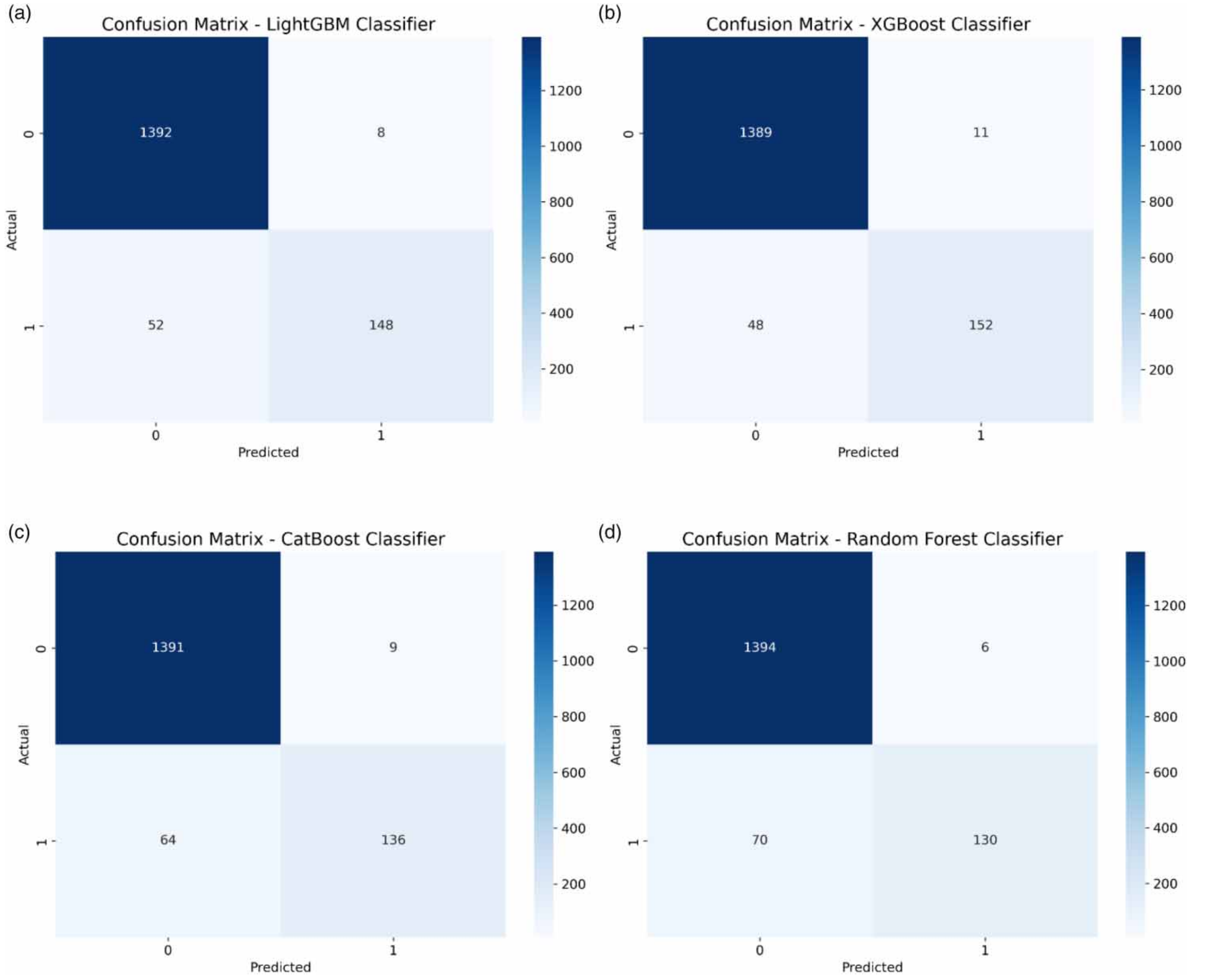


Figure 6 | Confusion matrix of best-performing ML algorithms without SMOTE on water quality prediction: (a) LightGBM, (b) XGBoost, (c) CatBoost, and (d) Random Forest.

demonstrated the highest precision score of 0.955, indicating its ability to accurately classify positive instances of water quality. The optimized hyperparameters, obtained through RandomizedSearchCV, significantly contributed to its overall performance.

3. RESULTS AND DISCUSSION

The ML algorithms were initially configured with hyperparameters, and RandomizedSearchCV was employed to discover the optimal hyperparameters using five-fold cross-validation. RandomizedSearchCV is a technique used to optimize hyperparameters for ML models. In the context of the water quality dataset, RandomizedSearchCV aids in finding the best combination of hyperparameters for each ML algorithm. By exploring a random subset of the hyperparameter space, RandomizedSearchCV efficiently explores different configurations and selects the ones that yield the best performance. This process ensures that the models are finely tuned and capable of achieving optimal results when applied to the water quality prediction task.

Based on these optimized hyperparameters, the ML classifiers were trained using the training set data. The performance of each model was then evaluated using the test set data.

In this prediction model, we tested 14 ML algorithms. Table 5 presents the performance metrics of the best-performing ML algorithms for water quality prediction, both with and without SMOTE. The evaluated metrics are accuracy, precision, recall, and F1 score.

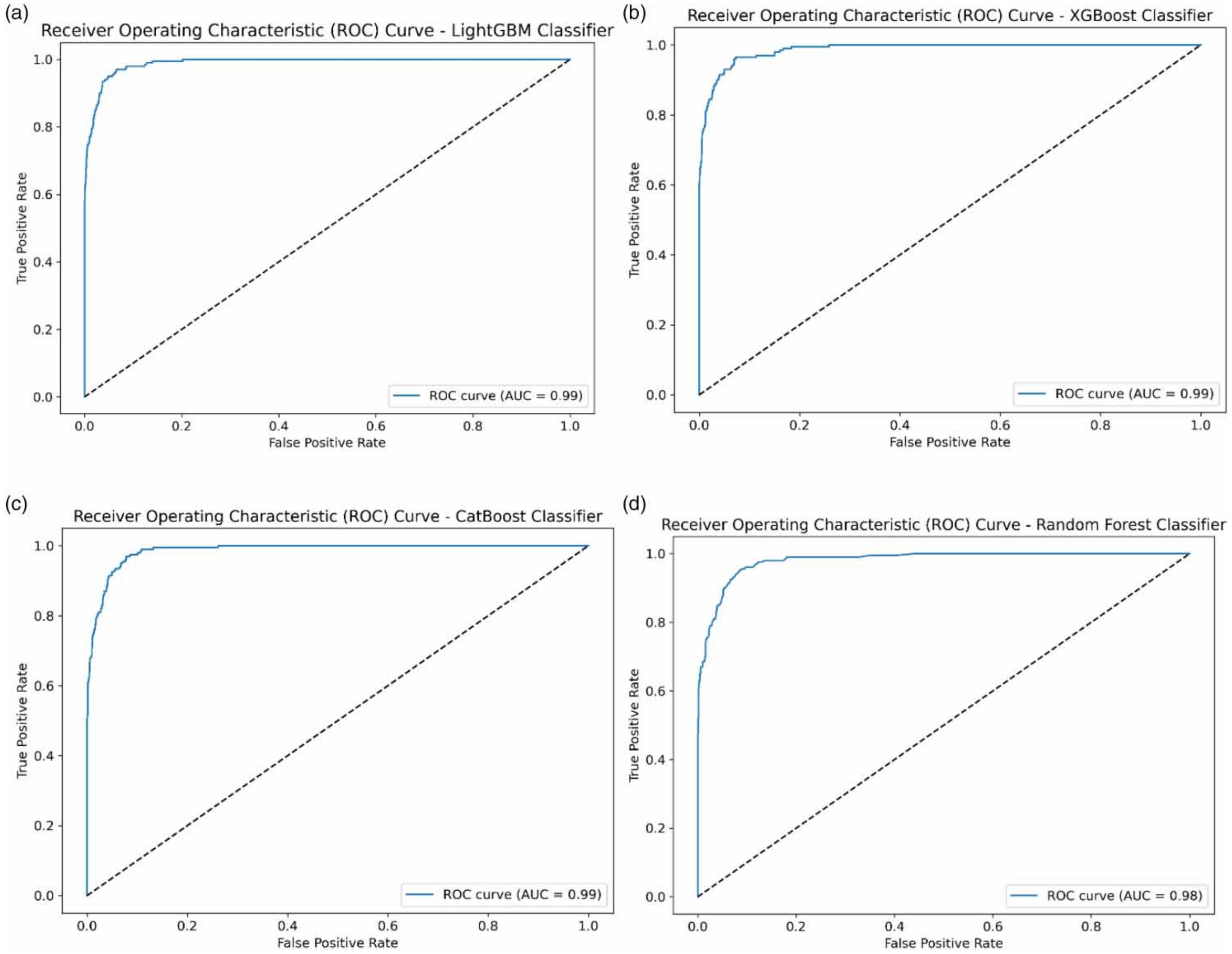


Figure 7 | ROC curves of best-performing ML algorithms without SMOTE on water quality prediction: (a) LightGBM, (b) XGBoost, (c) CatBoost, and (d) Random Forest.

The classification model's accuracy is assessed using performance measures from the confusion matrix. Accuracy, precision, recall, and F1 score can be defined as shown in Equations (5)–(8), respectively De Diego *et al.* (2022):

$$\%\ \text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \times 100 \tag{5}$$

$$\text{Precision } (p) = \ = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{6}$$

$$\text{Recall } (r) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times 100 \tag{7}$$

$$\text{F1 score} = \frac{(2 \times p \times r)}{(p + r)} \tag{8}$$

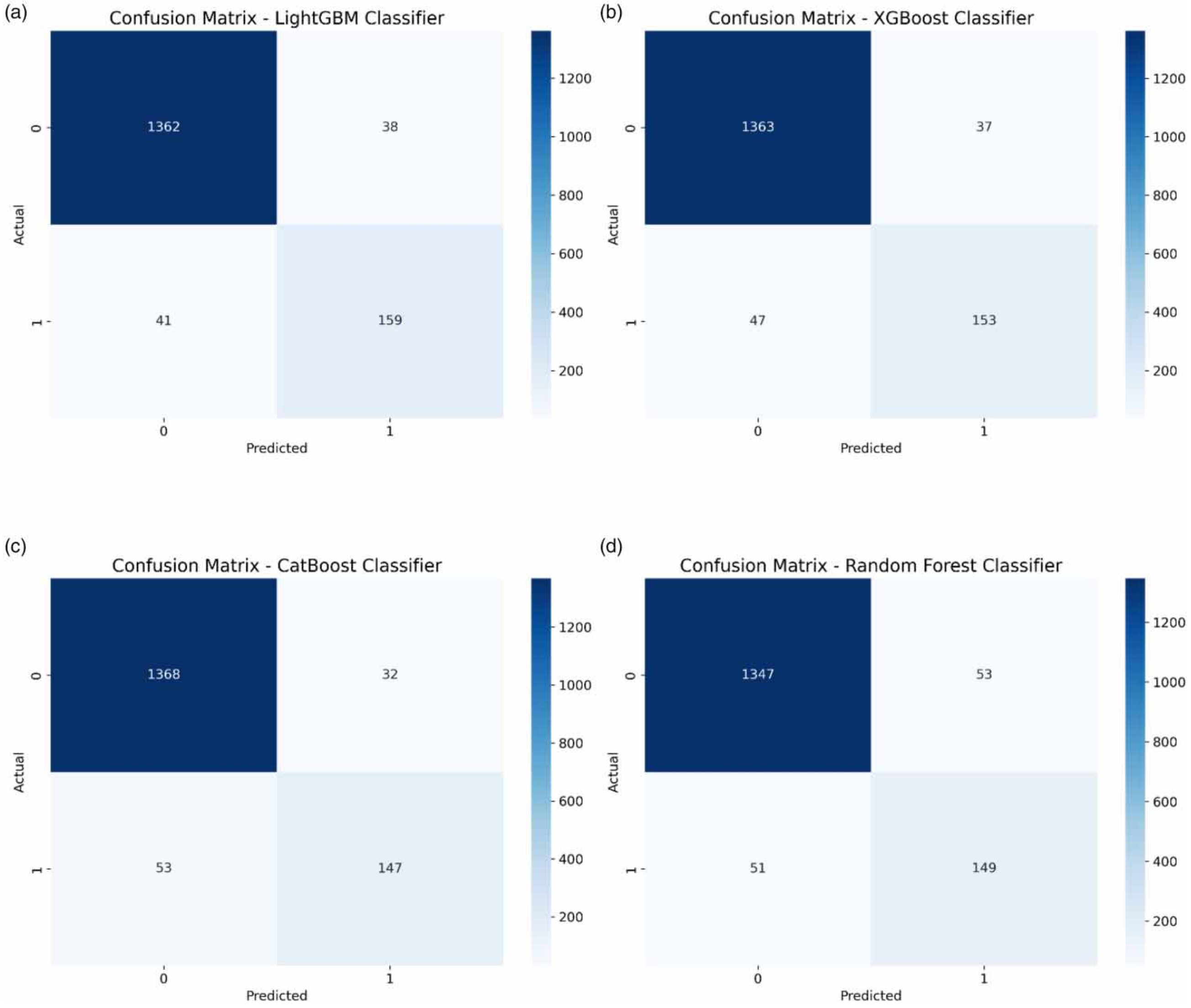


Figure 8 | Confusion matrix of best-performing ML algorithms with SMOTE on water quality prediction: (a) LightGBM, (b) XGBoost, (c) CatBoost, and (d) Random Forest.

Without SMOTE, XGBoost had a slightly higher accuracy at 96.31% compared with LightGBM's accuracy of 96.25%. In terms of precision, Random Forest attained the highest score of 0.955, while XGBoost demonstrated the highest recall of 0.76. Additionally, XGBoost exhibited the highest F1 score of 0.837.

When SMOTE was applied to the algorithms, there was a slight decrease in performance. However, LightGBM remained the top-performing algorithm with an accuracy of 95.06% and a precision of 0.807. CatBoost showed improvement in precision, recall, and F1 score compared with the other algorithms.

Table 6 presents the performance metrics of other ML algorithms for water quality prediction, both with and without SMOTE. The metrics are accuracy, ROC AUC (area under the curve), and F1 score. Without SMOTE, DecisionTreeClassifier achieved an accuracy of 94% and had the highest ROC AUC of 0.86. LogisticRegression showcased balanced performance across multiple metrics. When SMOTE was applied, DecisionTreeClassifier attained the highest accuracy of 92%, and QuadraticDiscriminantAnalysis exhibited the highest ROC AUC of 0.87.

The confusion matrices and ROC curves for the best-performing algorithms, both with and without SMOTE, are depicted in Figures 6–9. These visualizations provide a comprehensive understanding of the classification performance for each algorithm.

Overall, the best-performing algorithms for water quality prediction were LightGBM and XGBoost without SMOTE. However, the application of SMOTE helped enhance the performance of CatBoost. LightGBM and XGBoost demonstrated high

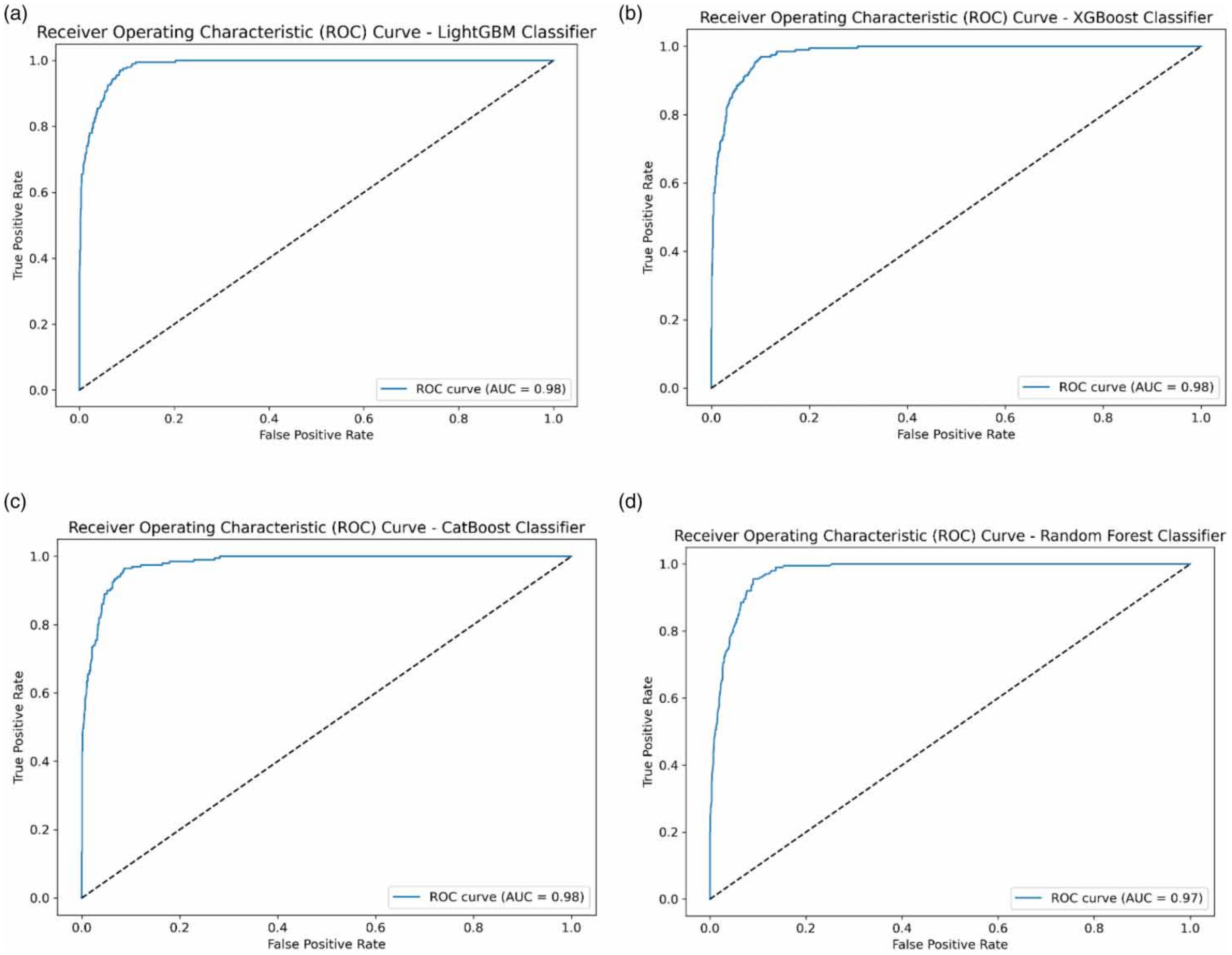


Figure 9 | ROC curves of best-performing ML algorithms with SMOTE on water quality prediction: (a) LightGBM, (b) XGBoost, (c) CatBoost, and (d) Random Forest.

precision but exhibited varied recall scores, indicating a trade-off between false positives and false negatives. The high precision (94.9% in LightGBM and 93.3% in XGBoost) suggests a low false-positive rate, meaning that when the model predicts unsafe water, it is often correct. This trade-off is evident in the recall scores (74% in LightGBM and 76% in XGBoost), signifying a moderate ability to identify actual safe water samples – a common characteristic in imbalanced datasets. On the other hand, CatBoost and Random Forest showed similar accuracy but diverged in precision and recall. CatBoost demonstrated high precision (93.8%) but lower recall (68%), emphasizing accurate predictions of unsafe water with a potential risk of missing safe samples. Random Forest, with high precision (95.5%) but lower recall (65%), exhibited a comparable trade-off. SMOTE aims to address class imbalance by over-sampling the minority class (safe water samples). Both LightGBM and XGBoost exhibited improvements in recall (LightGBM: 79.5%, XGBoost: 76.5%), suggesting that SMOTE contributed to a better identification of safe water samples.

4. CONCLUSION

The article presents an ML-based classification model for water quality prediction, utilizing 20 features that represent various substances and their concentrations in water samples. The dataset used for model development consists of 7,996 samples. The study highlights that LightGBM, XGBoost, CatBoost, and Random Forest were the best-performing algorithms. XGBoost achieved the highest accuracy of 96.31% without SMOTE and had a precision of 0.933. The application of SMOTE improved the performance of CatBoost.

The research findings contribute valuable insights into the application of ML algorithms for water quality assessment. The proposed model demonstrates the potential to automate water quality assessment and enhance the efficiency of evaluating water quality parameters. These results hold relevance for researchers and practitioners involved in water quality management and decision-making processes.

However, it is important to acknowledge a limitation of the developed model. The study relies on a specific dataset comprising 7,996 samples and 20 features, which may not fully represent the entire spectrum of water quality variations across different regions or time periods. For future research, it is recommended to consider the collection and incorporation of a larger and more diverse dataset that encompasses different regions, time periods, and sources of water samples. This approach would provide a more comprehensive understanding of water quality assessment in various contexts.

AUTHOR CONTRIBUTIONS

KK: Conceptualization; Roles/Writing – original draft; Investigation; Methodology, Formal analysis; SK: Supervision; Validation; Software; Visualization; RM: Data curation; Writing – review & editing.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: https://www.kaggle.com/datasets/mssmartypants/water-quality?select=waterQuality1.csv.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M. & Elshafie, A. 2019 Machine learning methods for better water quality prediction. *Journal of Hydrology* **578**, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084.

Akhtar, N., Syakir Ishak, M. I., Bhawani, S. A. & Umar, K. 2021 Various natural and anthropogenic factors responsible for water quality degradation: a review. *Water* **13** (19), 2660. https://doi.org/10.3390/w13192660.

Amoatey, P. & Baawain, M. S. 2019 Effects of pollution on freshwater aquatic organisms. *Water Environment Research* **91**, 1272–1287. https://doi.org/10.1002/wer.1221.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32. https://doi.org/10.1023/A:1010933404324.

Cain, M. K., Zhang, Z. & Yuan, K. H. 2017 Univariate and multivariate skewness and kurtosis for measuring nonnormality: prevalence, influence and estimation. *Behavioral Research Methods* **49**, 1716–1735. https://doi.org/10.3758/s13428-016-0814-1.

Chou, J.-S., Ho, C.-C. & Hoang, H.-S. 2018 Determining quality of water in reservoir using machine learning. *Ecological Informatics* **44**, 57–75. https://doi.org/10.1016/j.ecoinf.2018.01.005.

Data Link 2021. Water Quality: Dataset for Water Quality Classification. Available from: https://www.kaggle.com/datasets/mssmartypants/water-quality?select=waterQuality1.csv.

De Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J. & Moguerza, J. M. 2022 General Performance Score for classification problems. *Appl. Intell.* **52**, 12049–12063. https://doi.org/10.1007/s10489-021-03041-7.

Dotaniya, M. L., Meena, V. D., Saha, J. K., Dotaniya, C. K., El Din Mahmoud, A., Meena, B. L., Meena, M. D., Sanwal, R. C., Meena, R. S., Doutaniya, R. K., Solanki, P., Lata, M. & Rai, P. K. 2023 Reuse of poor-quality water for sustainable crop production in the changing scenario of climate. *Environment, Development and Sustainability* **25**, 7345–7376. https://doi.org/10.1007/s10668-022-02365-9.

Douzas, G., Bacao, F. & Last, F. 2018 Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences* **465**, 1–20. https://doi.org/10.1016/j.ins.2018.06.056.

Elkiran, G., Nourani, V. & Abba, S. I. 2019 Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *Journal of Hydrology* **577**, 123962. https://doi.org/10.1016/j.jhydrol.2019.123962.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. & Tabona, O. 2021 A survey on missing data in machine learning. *Journal of Big Data* **8**, 140. https://doi.org/10.1186/s40537-021-00516-9.

Ghobadi, F. & Kang, D. 2023 Application of machine learning in water resources management: a systematic literature review. *Water* **15** (4), 620. https://doi.org/10.3390/w15040620.

Giupponi, C. & Sgobbi, A. 2013 Decision support systems for water resources management in developing countries: learning from experiences in Africa. *Water* **5** (2), 798–818. https://doi.org/10.3390/w5020798.

Guo, H., Zhou, J. & Wu, C.-A. 2018 Imbalanced learning based on data-partition and SMOTE. *Information* **9** (9), 238. https://doi.org/10.3390/info9090238.

Hancock, J. T. & Khoshgoftaar, T. M. 2020 CatBoost for big data: an interdisciplinary review. *Journal of Big Data* **7**, 94. https://doi.org/10.1186/s40537-020-00369-8.

Hmoud Al-Adhaileh, M. & Waselallah Alsaade, F. 2021 Modelling and prediction of water quality by using artificial intelligence. *Sustainability* **13** (8), 4259. https://doi.org/10.3390/su13084259.

Jolliffe, I. T. & Cadima, J. 2016 Principal component analysis: a review and recent developments. *Philosophical Transactions Series A, Mathematical, Physical, and Engineering Sciences* **374** (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A. & Ashraf, I. 2022 Water quality prediction using KNN imputer and multilayer perceptron. *Water* **14** (17), 2592. https://doi.org/10.3390/w14172592.

Kaddoura, S. 2022 Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability* **14** (18), 11478. https://doi.org/10.3390/su141811478.

Li, B. 2020 *Random Search Plus: A More Effective Random Search for Machine Learning Hyperparameters Optimization*. Master's thesis, University of Tennessee, Knoxville, TN, USA. Available from: https://trace.tennessee.edu/utk_gradthes/5849.

Li, J., Zhu, Q., Wu, Q. & Fan, Z. 2021 A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences* **565**, 438–455. https://doi.org/10.1016/j.ins.2021.03.041.

Lin, L., Yang, H. & Xu, X. 2022 Effects of water pollution on human health and disease heterogeneity: a review. *Frontiers in Environmental Science* **10**, 880246. doi:10.3389/fenvs.2022.880246.

Machado, M. R., Karray, S. & de Sousa, I. T. 2019 LightGBM: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In: *2019 14th International Conference on Computer Science & Education (ICCSE)*, IEEE, Piscataway, NJ, USA, pp. 1111–1116. doi:10.1109/ICCSE.2019.8845529.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A. & Al-Shamma'a, A. 2022 Water quality classification using machine learning algorithms. *Journal of Water Process Engineering* **48**, 102920. https://doi.org/10.1016/j.jwpe.2022.102920.

Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. 2003 Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology* **84** (9), 2347–2363. Available from: http://www.jstor.org/stable/3450140.

Saleh, H. M. & Hassan, A. I. 2021 Water quality standards. In: *Applied Water Science, Volume 1: Fundamentals and Applications* (Inamuddin., Ahamed, M. I., Boddula, R. & Rangreez, T. A., eds), Wiley, Hoboken, NJ, USA, pp. 441–468. https://doi.org/10.1002/9781119725237.ch17.

Samsudin, M. S., Azid, A., Khalit, S. I., Sani, M. S. A. & Lananan, F. 2019 Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones. *Marine Pollution Bulletin* **141**, 472–481. https://doi.org/10.1016/j.marpolbul.2019.02.045.

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D. & Apri, M. 2023 An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX* **10**, 102119. https://doi.org/10.1016/j.mex.2023.102119.

Varadharajan, C., Appling, A. P., Arora, B., Christianson, D. S., Hendrix, V. C., Kumar, V., Lima, A. R., Müller, J., Oliver, S., Ombadi, M., Perciano, T., Sadler, J. M., Weierbach, H., Willard, J. D., Xu, Z. & Zwart, J. 2022 Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrological Processes* **36** (4), e14565. https://doi.org/10.1002/hyp.14565.

Xu, Y. & Goodacre, R. 2018 On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing* **2**, 249–262. https://doi.org/10.1007/s41664-018-0068-2.

Yeo, I. & Johnson, R. A. 2000 A new family of power transformations to improve normality or symmetry. *Biometrika* **87** (4), 954–959. https://doi.org/10.1093/biomet/87.4.954.

Zehra, S., Faseeha, U., Syed, H. J., Samad, F., Ibrahim, A. O., Abulfaraj, A. W. & Nagmeldin, W. 2023 Machine learning-based anomaly detection in NFV: a comprehensive survey. *Sensors* **23** (11), 5340. https://doi.org/10.3390/s23115340.

First received 9 July 2023; accepted in revised form 6 December 2023. Available online 20 December 2023