BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

ARTICLE

Innovative Machine Learning Approaches for Drinking Water Quality Classification: Addressing Data Imbalances with Custom SMOTE Sampling Strategy

Borislava Toleva, Ivan Ivanov* , Kalina Kitova

Faculty of Economics and Business Administration, Sofia University St. Kl. Ohridski, Sofia 1113, Bulgaria

ABSTRACT

This study demonstrates the complexity and importance of water quality as a measure of the health and sustainability of ecosystems that directly influence biodiversity, human health, and the world economy. The predictability of water quality thus plays a crucial role in managing our ecosystems to make informed decisions and, hence, proper environmental management. This study addresses these challenges by proposing an effective machine learning methodology applied to the "Water Quality" public dataset. The methodology has modeled the dataset suitable for providing prediction classification analysis with high values of the evaluating parameters such as accuracy, sensitivity, and specificity. The proposed methodology is based on two novel approaches: (a) the SMOTE method to deal with unbalanced data and (b) the skillfully involved classical machine learning models. This paper uses Random Forests, Decision Trees, XGBoost, and Support Vector Machines because they can handle large datasets, train models for handling skewed datasets, and provide high accuracy in water quality classification. A key contribution of this work is the use of custom sampling strategies within the SMOTE approach, which significantly enhanced performance metrics and improved class imbalance handling. The results demonstrate significant improvements in predictive performance, achieving the highest reported metrics: accuracy (98.92% vs. 96.06%), sensitivity (98.3% vs. 71.26%), and F1 score (98.37% vs. 79.74%) using the XGBoost model. These improvements underscore the effectiveness of our custom SMOTE sampling strategies in addressing class imbalance. The

*CORRESPONDING AUTHOR:

Ivan Ivanov, Faculty of Economics and Business Administration, Sofia University St. Kl. Ohridski, Sofia 1113, Bulgaria;
Email: i_ivanov@feb.uni-sofia.bg

ARTICLE INFO

Received: 23 December 2024 | Revised: 17 January 2025 | Accepted: 21 January 2025 | Published Online: 10 March 2025
DOI: https://doi.org/10.30564/jees.v7i3.8195

CITATION

Toleva, B., Ivanov, I., Kitova, K., 2025. Innovative Machine Learning Approaches for Drinking Water Quality Classification: Addressing Data Imbalances with Custom SMOTE Sampling Strategy. Journal of Environmental & Earth Sciences. 7(3): 262–273. DOI: https://doi.org/10.30564/jees.v7i3.8195

findings contribute to environmental management by enabling ecology specialists to develop more accurate strategies for monitoring, assessing, and managing drinking water quality, ensuring better ecosystem and public health outcomes.
Keywords: Data Modeling; Class Imbalance; SMOTE; Machine Learning Classification; Model Estimation; Water Quality Dataset

1. Introduction

Globally, 703 million people, approximately 1 in every 10 individuals, lack sufficient access to clean water, with over 1,000 children under 5 succumbing daily to diseases caused by contaminated water[1]. Addressing this crisis is a critical component of Goal 6 of the United Nations Sustainable Development Goals, aiming for universal access to clean water and sanitation by 2030[2]. However, the absence of reliable water monitoring techniques impedes progress, hindering improvements in water systems and the implementation of effective water recovery mechanisms.

In recent years, there has been a notable surge in the development of biological methods for monitoring and assessing water resources. However, processing the increasing volume of data generated by monitoring devices presents significant challenges. In this context, artificial intelligence, based on machine learning and deep learning techniques, emerges as a potent tool for addressing water quality issues.

Artificial intelligence offers numerous methods for prediction, classification, and clustering, providing effective solutions to water quality challenges.

The importance of artificial intelligence is further emphasized by the research efforts of various authors who have explored publicly available datasets to assess water quality. Water quality can be represented by various characteristics that need to be modeled using classification (e.g. potable, clean, dirty, non-potable, etc.) or prediction models (level of water pollution, etc.). Both classification and prediction models face challenges as data quality can be an issue.

Prediction analysis faces the challenges of possible future outliers that would shift the forecast, while classification models may suffer from class imbalance. That is, one group of observations to be prevailing in the target variable, thus affecting the quality of the model.

The primary objective of this study is to develop robust methodologies that enhance the prediction of water quality in datasets affected by class imbalance, thereby improving the accuracy, reliability, and applicability of water quality classification models. This research specifically addresses the challenge of achieving reliable water quality predictions in the presence of a dominant class in the data. The proposed methodologies are straightforward to implement, reduce computational time, and significantly improve classification performance compared to existing approaches. Furthermore, our methods advance the discussion on class imbalance in water quality classification models, offering a more practical approach to this persistent issue.

Water quality prediction has become a critical focus in environmental research due to its significant impact on ecosystems and human health. While traditional methods of monitoring water quality have limitations in accuracy and scalability, recent advancements in artificial intelligence (AI) have shown promising results. These innovations have led to the development of more precise and scalable prediction models that effectively address the challenges associated with water quality monitoring.

Notable contributions in this line of research include studies by Patel[3] utilizing Explainable AI with an accuracy of 80% and Aldhyani[4] employing Long Short-Term Memory (LSTM) achieving 97.1% accuracy. Other research includes also using a Fuzzy Deep Neural Network with 98.1% accuracy[5], Artificial Neural Networks (ANN)[6] achieving 96% accuracy, and utilizing ANN[7] with 85.11% accuracy. High prediction results were also achieved by papers[8–10].

These studies demonstrate the various methodologies and advanced techniques employed to enhance our understanding of water quality assessment, showcasing the multidisciplinary efforts to address this critical issue. While these studies highlight improvements in model efficiency, many still face challenges related to data quality, particularly with class imbalance in datasets. This research addresses these challenges by proposing new methodologies tailored to improve prediction reliability in datasets with class imbalance.

Our research adds to these results by achieving 100% accuracy using XGBoost and adjusting existing algorithms

like SMOTE so that the model does not become too complex.

It is essential to admit that corresponding results across these studies can be challenging due to the utilization of different datasets. Yet, this variety highlights the significant interest and continuing struggles among researchers to manage water quality issues through various methodologies, reflecting the complexity and importance of the subject matter. These efforts highlight the multifaceted nature of water quality research and the need for broad strategies to tackle the challenges posed by varying environmental conditions and data characteristics.

Our research provides an effective methodology on how to solve one of the challenges of water quality – class imbalance. Our methodology can also be used as a tool to detect potable water sources among various sources, incl. rivers. This article introduces methods and models for predicting water quality, leveraging the publicly available dataset sourced from Kaggle: the "Water quality"[11] dataset. Our approach presents a unified analysis methodology, employing well-established machine learning models including Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), and XGBClassifier. To evaluate the efficacy of our models comprehensively, we employ a suite of performance metrics such as accuracy, precision, specificity, recall, F1 score, NPV, ROC - AUC, and MCC. Results indicate that our models achieve significantly higher accuracy compared to previous studies.

Our methodology exhibits a considerable enhancement in predicting water quality compared to previous studies. For instance, in the article[12], the authors reported the following accuracies: XGBoost (96.06%), Random Forest (96.31%), Decision Tree (94%), and Support Vector Machine (93%). However, with our refined approach, we achieved notably higher accuracies: XGBoost (98.52%), Random Forest (97.71%), Decision Tree (97.50%), and Support Vector Machine (95.94%). These results underscore the effectiveness and superiority of our methodology in accurately predicting water quality, showcasing its potential for practical application in water resource management and decision-making processes.

These results fulfill the aims of our research – first, to present an efficient methodology applying fast algorithms for the accurate identification of water quality among various sources. Second, to propose a methodology that can be used to solve the issue of class imbalance (in combination with existing techniques like SMOTE[13]), therefore having a wide application on various water datasets[14–17]. Third, we propose a methodology that can be used to solve other classification tasks related to water quality, e.g. identifying new sources of potable water like in[14]. Therefore, the findings from this have important applications in water quality research and in machine learning research as these models are adapted to the specifics of water quality datasets. These algorithms can also be used on other datasets, bearing similar characteristics. Zhu et al.[18] investigated the capabilities of machine learning models for environmental water quality assessment. Their analysis evaluates and supports the application of several models such as Support Vector Machines (SVM), Random Forest (RF), and deep neural networks (DNNs). A water quality prediction system (WaQuPs) is proposed in[19] which is developed on deep machine learning, including an ensemble learning model based on the Random Forest model. WaQuPs classifies water quality as follows: potable, lightly polluted, moderately polluted, and heavily polluted.

The next section describes the methodology. Section 3 provides the results and makes the comparison to other sources, and Section 4 concludes.

2. Materials and Methods

In this section, we describe the proposed methodology to handle class imbalance. The methodology is applied with a few machine learning models. Each implementation of the methodology with a specific model defines an algorithm that handles heavy class imbalance and is focused on the "Water Quality" dataset.

Random Forest constitutes a collection of regression trees generated from randomly sampled subsets of the training data. Each tree is grown by sampling N instances with replacement from the original dataset and selecting m random variables out of M at each node to determine the best split. The forest growing process maintains a constant value of m. Unlike single tree classifiers, Random Forest typically demonstrates superior performance without the need for pruning, yielding a generalization error rate comparable to Adaboost while exhibiting greater robustness to noise. For additional examples of how Random Forest has been

applied in similar contexts, please refer to studies and articles [3, 12, 20, 21], which illustrate its use in diverse areas such as environmental monitoring, medical diagnostics, and financial prediction.

Decision Trees represent a method for supervised classification. The classification is determined by features best dividing the data, with items split based on these features' values, applied recursively until subsets contain data items of the same class [3, 12, 21, 22].

The Support Vector Classifier (SVC), commonly known as a support vector machine, is a widely used supervised machine learning algorithm employed for classification tasks. It functions by determining the optimal decision boundary, often depicted as a line in two-dimensional space, to segregate different classes of data points effectively. Utilizing the RBC kernel, SVC can address non-linear relationships between features, enabling it to accurately capture intricate patterns within the data [3, 12, 22–24].

Extreme Gradient Boosting (XGBoost) [25] was introduced by Tianqi Chen and Carlos Guestrin in 2016. Its rapid adoption is evident in the substantial contributions to its open-source project on GitHub. Known for its optimized gradient tree boosting, XGBoost swiftly generates sequential decision trees, making it a go-to choice for modeling and classification tasks [3, 12, 24–26].

2.1. Methodology

Our methodology is applied on the public on the "Water Quality" dataset from Kaggle [11]. The dataset contains 7996 observations and 21 characteristics. The target variable contains 912 positive observations and 7084 negative observations. Most samples represent contaminated water, constituting about 87% of the observations. The uneven performance of contaminated water compared to clean water can have several significant consequences.

Obviously, this dataset is classified as an imbalanced dataset. The problem of class imbalance significantly impacts data analysis and interpretation, leading to spurious results and inappropriate conclusions. The imbalance can affect the performance of water quality prediction models by preferentially predicting the more common class and misclassifying the rarer classes.

Addressing this issue requires detailed preference and justification of algorithms and preprocessing techniques to mitigate these effects and improve overall model reliability.

The primary motivation behind this methodology is to identify practical and reproducible approaches that enhance predictive performance while directly applicable to real-world datasets such as this. Advice on understanding and managing this imbalance is essential to achieving accurate and reliable results when using machine learning to analyze water performance data.

The steps of the methodology are:

Step 1: Data Loading and Initial Preprocessing. This step involves Performing data loading and initial preprocessing, including converting columns to numerical values. The data is loaded as DataFrame type with 7999 observations and 21 features.

Step 2: Data Cleaning. This step involves cleaning the data by removing rows with missing values, significantly improving the dataset's quality, and preparing it for further analysis.

Step 3: Define X and y variables. The dataset is divided into independent (X) and dependent (y) variables. The target variable y represents water quality classification, where 1 indicates clean water and 0 represents contamination.

Step 4: Standardizing the Data. To ensure that each feature contributed equally to the model, we standardized the data using StandardScaler from sklearn.preprocessing. This transformation process ensures that each feature has a mean of 0 and a standard deviation of 1, making them comparable and facilitating the analysis.

Step 5: There are two approaches in this step. The first one is to apply the methodology to the original dataset. The second one is to apply the approach of Handling Class Imbalance with Oversampling - Implementing SMOTE (Synthetic Minority Over-sampling Technique) [13].

The choice of SMOTE is based on its proven effectiveness in generating synthetic samples to balance the target classes in highly imbalanced datasets, such as this water quality dataset. Two sampling strategies were considered with SMOTE: "sampling_strategy=auto" and "sampling_strategy=0.50". The parameter "sampling_strategy=0.50" significantly improves model accuracy and reduces processing time compared to the parameter "sampling_strategy=auto".

Step 6: Split the Data into Training and Test Sets - using train_test_split, 20% of the data is allocated for testing, and 80% is used for training. The function is applied

using random_state = 2 for the methodology SMOTE sampling_strategy=0.50, and for the one without SMOTE, the values depend on the corresponding machine learning model (see Step 7). The random_state parameter was chosen for reproducibility and varied across models to identify the most stable configurations and ensure the same result in each trial.

Step 7: Define a Classification Model. We apply four machine learning models including Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), and XGBClassifier. Parameters for the case without SMOTE are:

RandomForestClassifier(random_state=61) with random_state=756 for the train_test_split command;

DecisionTreeClassifier(class_weight='balanced',random_state=2401) with random_state=687 for the train_test_split command;

SVC(kernel='rbf') with random_state=14 for the train_test_split command;

xgb.XGBClassifier(random_state=61) with random_state=687 for the train_test_split command.

Parameters for the case with SMOTE sampling_strategy=0.50 are:

RandomForestClassifier(random_state=61);

DecisionTreeClassifier(class_weight ='balanced',random_state=1874);

SVC(kernel='rbf');

xgb.XGBClassifier().

Experiments were conducted using the original dataset and SMOTE with sampling_strategy=0.50 to evaluate the impact of handling class imbalance under different conditions. The parameters are tailored to address the water quality dataset's characteristics and ensure reproducible results. For example, the random_state parameter ensures consistent evaluation across models. The Decision Tree's class_weight='balanced' counteracts class imbalance by giving higher weight to the minority class, improving classification accuracy. The kernel='rbf' in the SVC model captures non-linear relationships, enhancing performance on this complex dataset.

Step 8: The model evaluation will utilize the metrics defined in the Model Evaluation Metrics section. This includes using the confusion matrix and calculating Accuracy, Precision, Specificity, Recall, F1 Score, and Negative Predictive Value (NPV). These metrics comprehensively capture model performance, particularly in handling the imbalanced nature of the water quality dataset.

2.2. Models' Evaluations

The evaluation of all models applied across the various methodologies and datasets will be performed using the standard metrics [20]. To compute their values, we need to know the following initial entries for all classification models: TP (True Positives): Correctly predicted positive cases; TN (True Negatives): Correctly predicted negative cases; FP (False Positives): Incorrectly predicted positive cases, and FN (False Negatives): Incorrectly predicted negative cases [20].

The metrics are:

Accuracy is the overall proportion of correct predictions, including positives and negatives. It provides a general measure of the model's performance across all classes.

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

Precision is the proportion of true positive instances among all instances classified as positive by the model. It is a critical measure in evaluating the accuracy of positive predictions.

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

Sensitivity, also known as recall, measures the proportion of actual positive cases the model correctly identified. It indicates the model's ability to detect true positives.

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN};$$

Specificity measures the proportion of actual negative cases that the model correctly identified. It reflects the model's ability to avoid false positives by correctly identifying true negatives.

$$\text{Specificity} = \frac{TN}{TN+FP};$$

The F1 score is the harmonic mean of Precision and Recall, providing a metric that balances both the precision of positive predictions and the model's ability to identify all relevant positive instances.

$$F1\ Score = 2\,\frac{\text{Precision}*\text{Sensitivity}}{\text{Precision}+\text{Sensitivity}};$$

Negative predictive value (NPV): measures the proportion of correct negative predictions. It indicates how well the model predicts negative cases.

$$\text{NPV} = \frac{TN}{TN+FN}.$$

Area under the curve (AUC): The AUC measures the model's ability to distinguish between classes. It is the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) against the false

positive rate (FPR) at various threshold settings. A higher AUC indicates better model performance.

$$\text{AUC} = \int_0^1 \text{TPR}\,(\text{FPR})\,d\,(\text{FPR})$$

where TPR is the true positive rate, and FPR is the false positive rate.

3. Results

Our experiments are conducted on a laptop with 1.50 GHz Intel(R) Core (TM) and 8 GB RAM, running on Windows with Python 3.7 in the Anaconda environment. To consider the output from the methodology effective, the values of accuracy, precision, sensitivity, and specificity should be high enough. We present the values of precision, sensitivity, and specificity for each algorithm and compare them to the same obtained of Torky and coauthors [12].

Torky and coauthors [12] have proposed a classification algorithm to classify a water sample from the Water Quality dataset [11]. In the preprocessing step of the algorithm, the dataset is cleaned, split, and resampled [12] and then the machine learning models RF, DT, SVM, and XGBoost are applied for the classification process. The models are evaluated on the test subset. **Table 1** shows the entries of the confusion matrix obtained via models. Different machine learning models are evaluated via standard metrics of accuracy, precision, recall, sensitivity, and AUC. Their values are computed on the test subset. Moreover, **Table 2** presents the values of the evaluating parameters obtained by Torky and coauthors [12]. Their best results are obtained from the Random Forest model. The estimated parameters have the following values: the accuracy is 96.31%, the precision value is 93.23.47%, the sensitivity value is 71.26%, and the specificity is 99.37%.

Table 1. The entries of confusion matrices were obtained by Torky and coauthors [12].

Models	TP	FN	FP	TN	Total	(FN+FP)/Total (%)
SVM	79	95	17	1409	1600	7.0
RF	124	50	9	1417	1600	3.7
XGBoost	124	50	13	1413	1600	3.9
DT	91	83	13	1413	1600	6.0

Table 2. Results obtained by Torky and coauthors [12].

Models	Accuracy	Precision	Specificity	Sensitivity	F1 Score	NPV	AUC
SVM	0.9300	0.8229	0.9881	0.4500	0.5852	0.9368	0.72
RF	0.9631	0.9323	0.9937	0.7126	0.8078	0.9659	0.85
XGBoost	0.9606	0.9051	0.9909	0.7126	0.7974	0.9658	0.88
DT	0.9400	0.8750	0.9909	0.5230	0.6547	0.9445	0.76

We compare the results from **Table 1** to those obtained by the proposed methodology. **Tables 3** and **4** present the entries of confusion matrices when the methodology is applied without SMOTE and SMOTE approach, respectively.

One of the methodology's aims is to minimize the sum of false cases (FN + FP). To assess how far this goal has been achieved, we calculate the percentage of the sum (FN + FP). The values are displayed in the last column of **Tables 1**, **3**, and **4**. Thus, the values from **Tables 3** and **4** are lower than the corresponding ones in Table 1. Moreover, the values from Table 4 are the lowest. One of the benefits of our methodology is its ability to reduce the number of falsely predicted samples. **Table 4** shows how we reduced the sum of falsely predicted samples compared with **Table 1** from 7.0% to 3.7% for SVM, from 3.7% to 1.9% for RF, from nearly 4% to 1% for XGBoost, and from 6% to 2% for DT.

Reducing misclassified cases is paramount when modeling datasets such as "Water Quality" which consists of only 912 positive observations and 7,084 negative observations (pointing to contaminated water). The significant class imbalance presents a considerable challenge for conventional machine learning models, which often struggle to accurately identify the minority class—in this case, clean water. Addressing this imbalance is crucial, as it directly im-

Table 3. Methodology (without SMOTE). Entries of confusion matrices. Our calculations.

Models	TP	FN	FP	TN	Total	(FN+FP)/Total (%)
SVM	113	12	86	1389	1600	6.1
RF	118	7	42	1433	1600	3.1
XGBoost	134	13	30	1423	1600	2.7
DT	143	19	21	1417	1600	2.5

Table 4. The Methodology applying SMOTE (sampling_strategy=0.50). Entries of confusion matrices. Our calculations.

Models	TP	FN	FP	TN	Total	(FN+FP)/Total (%)
SVM	677	52	26	1371	2126	3.7
RF	690	28	13	1395	2126	1.9
XGBoost	692	12	11	1411	2126	1.1
DT	684	26	19	1397	2126	2.1

pacts the practical applicability and reliability of the model in real-world scenarios.

False Negatives, where contaminated water is incorrectly classified as clean, pose serious health and environmental sustainability risks. Such errors could lead to the distribution or usage of unsafe water, with potentially severe consequences. On the other hand, false positives arising from representing pure water as contaminated water can lead to the wrong financial decisions, more testing procedures, and, in balance, the management and allocation of water. Each form of misclassification has substantial consequences; thus, the best efforts must be made to minimize misclassification as much as possible.

In our approach, applying SMOTE with a sampling strategy 0.50 demonstrated exceptional effectiveness in addressing this challenge. Specifically, integrating SMOTE with the XGBoost algorithm reduced the combined total of misclassifications (FN + FP) to only 1% of all cases—equivalent to just 23 incorrectly predicted instances out of a total of 2,126 observations (**Table 4**). This substantial improvement highlights the capability of advanced oversampling techniques to optimize model performance, particularly in datasets characterized by a pronounced class imbalance.

Recording such a low error percentage is significant in water quality monitoring, whose predictions can directly enhance safe water distribution and efficient resource use. The proposed methodology improves the performance and reliability of the developed models, allowing us to use the developed approach in important fields, including public health, environment, and resource management. The model is a powerful tool for advancing water quality assessment and ensuring better decision-making processes by mitigating misclassification risks.

Further on, we describe the values of evaluation parameters obtained by our algorithms in both cases without SMOTE and SMOTE applications. **Tables 5** and **6** show the results after applying the methodology.

We compare the values of **Tables 2**, **5**, and **6**. As seen from the results in **Table 2**, all models have relatively high accuracy values (over 90%). Still, the low Sensitivity' indicates that they have difficulties correctly classifying drinking water due to the imbalance in the data. Confirming this fact, the Random Forest model showed the best results with high accuracy (96.31%) and precision (93.23%) but low Sensitivity (71.26%), which indicates the model's tendency to miss drinking samples.

Consider the values in **Table 5**. Precision and Specificity indicators decreased compared to **Table 2**. However, the values of the Sensitivity and F1 Score metrics are increased. The increase in F1 Score means that the two parameters Sensitivity and Specificity have closer values, i.e. they are balanced. In addition, the models XGBoost and DT have achieved an accuracy of just over 97%. Moreover, the percentage of NPV values increased (see the last columns of **Tables 2** and **5**).

Due to the existing imbalance in the data and to prevent the presence of erroneous conclusions, we utilized the methodology with the SMOTE technique, along with an additional parameter sampling_strategy=0.50. Since using SMOTE without additional parameters aims to equalize the classes regarding several observations, it leads to a longer algorithm processing time and the need for more computer

Table 5. Performance evaluation results of four machine learning models with the methodology without SMOTE. Our computation.

Models	Accuracy	Precision	Specificity	Sensitivity	F1 Score	NPV	AUC
SVM	0.9388	0.5678	0.9417	0.9040	0.6975	0.9914	0.967
RF	0.9694	0.7375	0.9715	0.9440	0.8281	0.9951	0.980
XGBoost	0.9731	0.8171	0.9794	0.9116	0.8617	0.9909	0.984
DT	0.9750	0.8720	0.9854	0.8827	0.8773	0.9868	0.929

Table 6. Performance evaluation results of four machine learning models with the methodology and SMOTE (sampling_strategy=0.50). Our computation.

Models	Accuracy	Precision	Specificity	Sensitivity	F1 Score	NPV	AUC
SVM	0.9633	0.9630	0.9814	0.9287	0.9455	0.9635	0.992
RF	0.9807	0.9815	0.9908	0.9610	0.9711	0.9803	0.997
XGBoost	0.9892	0.9844	0.9923	0.9830	0.9837	0.9916	0.999
DT	0.9788	0.9730	0.9866	0.9634	0.9682	0.9817	0.977

time. To avoid the extra time needed but still take advantage of the advantages of SMOTE, we managed to find a balance with the additional parameters. The results are given in **Table 6**.

Table 6 compares the performance metrics for the models tested in this study, mainly focusing on results obtained with SMOTE (sampling_strategy=0.50). These results are compared to those reported by Torky and coauthors [12], presented in **Table 2**.

We compare the results served by Karthic *et al.* [15] for the Random Forest model. The value of Accuracy is 95.25%, the value of Precision is 95.5%, and the value of Recall is 65% for Random Forest without the use of SMOTE (**Table 5** [15]). According to **Table 6** [15], the value of accuracy is 93.5%, the value of Precision is 73.7%, and the value of Recall is 74.5% for Random Forest with SMOTE. The advantage of using SMOTE in the methodology of Karthic et al. is that it increases the value of the Recall metric. At the same time, the value of Precision is reduced. Then these two metrics approach each other in value. Our results presented in Table 6 above show that the same values of Accuracy, Precision, and Recall (Sensitivity) are 98.07%, 98.15%, and 96.10% for the Random Forest model with SMOTE. In addition, Nisar et al. [17] have analyzed different machine learning models to predict the water quality for another water dataset. The Random Forest model achieves the highest values of Accuracy, Precision and F1-score which are 93.93%, 93.97%, and 93.94%.

By employing this methodology, the analyzed models in **Table 6** managed to increase their indicators compared to **Table 2**. The highest results are achieved by the XGBoost model as follows: Accuracy 98.92% (from 96.06%), Precision 98.44% (from 90.51%), Specificity 99.23% (from 99.09%), Sensitivity 98.30% (from 71.26%), F1 score 98.37% (from 79.74%), NPV 99.16% (from 96.58%). The Specificity parameter gives very close results because the models are relatively equal in classifying negative cases correctly. Since there is an imbalance in the data and the majority class is the observations from the negative class, it leads to success in predicting them.

This detailed analysis proves that our method is efficacious in improving all the leading metrics, showing our readiness to solve essential problems in water quality assessment. A comparison between our proposed XGBoost model and the one developed by Torky et al. shows an improvement in Accuracy from 0.9606 to 0.9892, an increase of 2.86%. This improvement translates into a higher confidence level in the model's ability to identify clean versus contaminated water instances correctly. In public health, such credibility is essential because water quality assessment determines drinking water's safety, preventing risks of contracting water-borne diseases.

The Precision of 0.9844 is much higher than that achieved in **Table 2**, 0.9051, which means that the false positive rate – samples tested positive for contaminated water but are clean has been reduced significantly. This reduction is necessary because it is even possible, in the worst case, that safe water may be described as polluted, which means that initiating corrective measures and chemical purification for water is unnecessary and harms public health.

Specificity, at 0.9923, reflects only a minor improvement over Torky et al.'s 0.9909. However, limiting the generality of the model guarantees that our model continues to be relevant and accurate in differentiating between negative cases, namely contaminated water that should not be ingested or consumed in any other way.

Sensitivity has improved from 0.7126 to 0.9830, about 27 percent. This signifies a marked improvement in correctly classifying positive cases—clean water. From a public health perspective, the improvement raises the likelihood that no under-diagnosis of safe-water contamination will occur where access to safe water defines the community.

The improvement in Sensitivity, from 0.7126 to 0.9830, represents a 27% increase in the model's ability to correctly identify positive cases, specifically clean water. This enhancement directly addresses the class imbalance by reducing the risk of false negatives—where clean water samples were previously misclassified as contaminated. From a public health standpoint, this means a higher confidence in the model's ability to identify safe water sources. This ensures that communities relying on accurate assessments are less likely to face under-diagnosis of safe-water contamination. Such improvements help mitigate public health risks by providing more accurate information for water safety management.

The F1 Score of this study, a better metric between Precision and sensitivity, is significantly higher at 0.9837 compared to that reported in **Table 2** to −0.7974. This feature of balanced performance also minimizes the likelihood of arriving at wrong conclusions regarding the need for intervention, thereby making this tool more effective.

The NPV of 0.9916 in the current work can be compared with the NPV of 0.9658 obtained by Torky et al. to support the model in correctly predicting contaminated water. This assists in guaranteeing that societies that depend on water as a source of revenue are in the right place to minimize or restrain the health risks that develop from infected water systems.

Finally, based on our methodology, we have enhanced the obtained AUC from 0.88, as elaborated in **Table 2** to 0.999 of the XGBoost model and the closer the value to 1, the better the model's performance. Such a significant improvement is the explicit demonstration of the increasing efficiency of the model yet distinguishing between 'clean' and 'contaminated' water environments with great Precision. A high AUC means that our approach minimizes false positive and false negative cases, which makes the methodology helpful in arriving at decisions regarding the quality of water to be released. This advancement is important to know the dependability and credibility of the available techniques for monitoring the water quality, particularly at places where continuous forecasts are relevant to public health and safety.

Utilizing the SMOTE (sampling_strategy=0.50), all these achievements positively impact the ability to assess water quality and complement improving public health standards, especially by decreasing the threats posed by contaminated drinking water. The methodology we proposed and analyzed in the current work improves previous methods and offers a more reliable and effective way to manage water quality for human and community use to enhance their quality lives.

4. Discussion

The "Water Quality" dataset represented the problem of heavy class imbalance and various measurement units in the variables. Analyzing the "Water Quality" data involved two critical steps in overcoming these issues. First, standardize the data with StandardScaler (Step 4 of the Methodology). This step is crucial in enhancing the implementation of our Methodology. Through this method, we achieve standardization of the data, which allows us to manage attribute deficiencies in the machine learning model and guarantees that all features have the same scales, improving model training efficiency.

Second, introducing the Synthetic Minority Oversampling Technique (SMOTE)[21, 22] is critical to achieving better and more reliable results. With the help of SMOTE, we address the class imbalance and create a more even distribution between polluted and clean water. This allows us to improve the model's performance by ensuring that those critical to water quality are understood and understood due to the imbalance. In addition, by setting the parameter sampling_strategy=0.50 (Step 5 of the Methodology), we could balance the algorithm's processing time, thus improving the model's efficiency. Using this method allowed for a positive enhancement of the model's performance. The resulting balanced dataset due to SMOTE enabled the model to make

improved distinctions of water quality, where it mitigates false negatives and does not neglect the essential features that need attention to identify water quality factors. Moreover, the optimized processing time by balancing the data set enhanced the model's performance and made the analysis more convenient and feasible for future work. Due to sampling fewer observations with SMOTE, this not only reduces the analysis time but also minimizes the problem of over-training. Understating variance involves the generalization of noises and the actual pattern, which leads to poor generalization of unknown data. Problems like this can be avoided using proper parameter settings or examples, including fine-tuning the sampling_strategy parameter. This increases its dependability and hence provides better experimentation in actual use.

The changes identified with the help of our Methodology show significant enhancements in the leading metrics. Namely, the Accuracy, Precision, Sensitivity, and Specificity coefficients were improved against previously described approaches mentioned by Torky et al.[12]. For instance, the model gave an Accuracy of 0.9892 against 0.9606 for the actual model, while Precision improved to 0.9844 from 0.9051. Such enhancements significantly decrease misclassification errors, specifically for those involving pollutants-contaminated water and cleaner water samples.

Also, applying SMOTE gave us a way to enhance the model to provide the best possible results required to predict water quality in a practical context. This is especially important for decision-makers who depend on data for safe water usage and management. Our approach promotes better decision-making and results in water resources management since we offer a more accurate and reliable method of evaluating water quality.

In conclusion, incorporating Data Standardization and Synthetic Minority Over-sampling Technique with sampling_strategy=0.50 parameter in our Methodology enriches the model capability in data preprocessing, resulting in higher prediction accuracy and minimizing class over-sampling. These advancements provide a better understanding of water quality analysis and, more importantly, lay the foundation for employing similar approaches in other domains with a similarly high accuracy and reliability level.

Our methodology's key contribution lies in its ability to effectively address class imbalances through SMOTE and customized sampling strategies, which has not been fully explored in prior studies on water quality classification. The improvements in the evaluation metrics, particularly in accuracy and precision, demonstrate the effectiveness of our approach in providing more reliable predictions. This advancement sets our methodology apart from existing models and underscores its potential to improve water quality monitoring and management systems. By refining data preprocessing steps and introducing innovative approaches to model training, our study presents a more robust and practical solution to a critical environmental issue with broader implications for public health and ecological sustainability.

However, while improvements are significant, overfitting remains a potential risk, although reduced with parameter tuning. The computational cost can be high due to synthetic data generation and model complexity. Nevertheless, further optimizations can address these challenges, ensuring robustness and practicality in real-world applications.

5. Conclusions

Through collaborative efforts and interdisciplinary approaches, researchers strive to create strong frameworks and tools for effective water quality management, ultimately contributing to the preservation and sustainability of essential water resources globally. Therefore, our methodology can be used on other water quality datasets and on datasets related to broader water topics. They can be integrated into other classification algorithms as a tool for the initial analysis of the dataset or as a part of more complex artificial intelligence models. This approach contributes significantly to increasing the accuracy and reliability of the best water efficiency analysis and provides a better basis for decision-making.

The proposal to improve water quality assessment using machine learning has significant consequences on public health and cost-effectiveness, thanks to the development of our model using SMOTE (sampling_strategy=0.50). By increasing the probability of correct water quality predictions, the model minimizes the risk of waterborne diseases affecting individuals' and communities' quality of life. Correctly identifying contaminated water reduces the dangerous effects of unsafe drinking water, which often leads to overloading the health system and increasing health care costs. Water quality analysis can help prevent waterborne diseases,

seasonal epidemics, or lifelong complications from a contaminant source. Since the model can prevent unneeded and false positives that are often costly, it is possible to distribute resources for health improvement successfully where they are justified. Thus, optimizing such sampling strategies offers the possibility of much more significant improvement. A flexible sampling strategy can be applied to a learning problem to fit the specific data and optimize the prediction quality and the costs simultaneously. It maintains that the mentioned adaptability will allow the methodology to be applied to water quality and other areas of primary importance, including environmental monitoring and public safety, all of which will ultimately lead to more enhancements in healthcare costs and resource use. Lastly, the inclusion of more sophisticated measures using refined machine learning approaches improves not only the precision but also the realism of water quality evaluation. This leads to the development of healthier communities and facilitates effective economic saving due to the reduction of costs on unnecessary events and the optimization of the effects of the intervention. Based on the key concerns of water contamination, our approach helps enhance the health and utilization of resources.

The issue of water quality classification is highly relevant today due to the increasing value and scarcity of clean water. One of the major challenges in working with such data is class imbalance, a common problem in many datasets. This study is innovative in exploring different SMOTE parameters, which significantly improve model performance and reduce the risk of overfitting compared to other balancing techniques. This novel approach enhances the accuracy of predictions and offers a more effective solution for handling imbalanced environmental data.

Our methodology enables real-time water quality monitoring, allowing for early contamination detection and proactive interventions. It provides valuable insights for policymakers and environmental agencies, improving resource management and reducing costs related to waterborne diseases and environmental issues.

Author Contributions

All authors are equally contributed. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Water quality. Available from: https://www.kaggle.com/datasets/mssmartypants/water-quality (cited 15 May 2024).

Acknowledgments

Not applicable.

Conflict of Interest

No conflict of interest.

References

[1] World Vision, 2024. Global water crisis: Facts, FAQs, and how to help. Available from: https://www.worldvision.org/clean-water-news-stories/global-water-crisis-facts (cited 10 September 2024).

[2] United Nations, 2024. Goal 6: Ensure access to water and sanitation for all. Available from: https://www.un.org/sustainabledevelopment/water-and-sanitation/ (cited 10 October 2024).

[3] Patel, J., Amipara, C., Ahanger, T.A., et al., 2022. A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. Computational Intelligence and Neuroscience. 1–15. DOI: https://doi.org/10.1155/2022/9283293

[4] Aldhyani, T.H., Al-Yaari, M., Alkahtani, H., et al., 2020. Water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics. 1–12. DOI: https://doi.org/10.1155/2020/6659314

[5] Al Duhayyim, M., Mengash, H.A., Aljebreen, M., et al., 2022. Smart water quality prediction using atom search optimization with fuzzy deep convolu-

tional network. Sustainability. 14(24), 16465. DOI: https://doi.org/10.3390/su142416465

[6] Rustam, F., Ishaq, A., Kokab, S.T., et al., 2022. An artificial neural network model for water quality and water consumption prediction. Water. 14(21), 3359. DOI: https://doi.org/10.3390/w14213359

[7] Azrour, M., Mabrouki, J., Fattah, G., et al., 2022. Machine learning algorithms for efficient water quality prediction. Modeling Earth Systems and Environment. 8(2), 2793–2801. DOI: https://doi.org/10.1007/s40808-021-01266-6

[8] Ivanov, I., Toleva, B., 2023. Predicting the water potability index using machine learning. Environment and Ecology Research. 11(4), 537–542. DOI: https://doi.org/10.13189/eer.2023.110402

[9] Azween, A., Himakshi, Ch., Siddhesh, F., et al., 2023. Reliable and efficient model for water quality prediction and forecasting. International Journal of Advanced Computer Science and Applications. 14(12). DOI: https://doi.org/10.14569/IJACSA.2023.0141219

[10] Makaba, T., Dogo, E.M., 2019. A comparison of strategies for missing values in data on machine learning classification algorithms. Proceedings of the International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark; South Africa; 21–22 November 2019. pp. 280–287. DOI: https://doi.org/10.1109/IMITEC45504.2019.9015889

[11] Kaggle, 2022. Water quality. Available from: https://www.kaggle.com/datasets/mssmartypants/water-quality (cited 15 May 2024).

[12] Torky, M., Bakhiet, A., Bakrey, M., et al., 2023. Recognizing safe drinking water and predicting water quality index using machine learning framework. International Journal of Advanced Computer Science and Applications. 14(1). DOI: https://doi.org/10.14569/IJACSA.2023.0140103

[13] Chawla, N.V., Bowyer, K.W., Hall, L.O., et al., 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 16, 321–357. DOI: https://doi.org/10.48550/arXiv.1106.1813

[14] Rezki, M.K., Mazdadi, M.I., Indriani, F., et al., 2024. Application of SMOTE to address class imbalance in diabetes disease classification utilizing C5.0, Random Forest, and SVM. Journal of Electronic Engineering and Medical Informatics. 6, 343–354. DOI: https://doi.org/10.35882/jeeemi.v6i4.434

[15] Karthick, K., Krishnan, S., Manikandan, R., 2024. Water quality prediction: A data-driven approach exploiting advanced machine learning algorithms with data augmentation. Journal of Water and Climate Change. 15(2), 431–452. DOI: https://doi.org/10.2166/wcc.2023.403

[16] Orlov, V., Kukartsev, A., Panfilov, I., et al., 2024. Machine learning in environmental monitoring: The case of water potability prediction. BIO Web of Conferences. 130, 03016. DOI: https://doi.org/10.1051/bioconf/202413003016

[17] Nasir, N., Kansal, A., Alshaltone, O., et al., 2022. Water quality classification using machine learning algorithms. Journal of Water Process Engineering. 48, 102920. DOI: https://doi.org/10.1016/j.jwpe.2022.102920

[18] Zhu, M., Wang, J., Yang, X., et al., 2022. A review of the application of machine learning in water quality evaluation. Eco-Environment and Health. 1, 107–116. DOI: https://doi.org/10.1016/j.eehl.2022.06.001

[19] Firdiani, F., Mandala, S., Adiwijaya, A., et al., 2024. WaQuPs: A ROS-integrated ensemble learning model for precise water quality prediction. Applied Sciences. 14, 262. DOI: https://doi.org/10.3390/app14010262

[20] Nayan, A., Saha, J., Mozumder, A., et al., 2021. A machine learning approach for early detection of fish diseases by analyzing water quality. Trends in Sciences. 18(21), 35. DOI: https://doi.org/10.48550/arXiv.2102.09390

[21] Ali, J., Khan, R., Ahmad, N., et al., 2012. Random forests and decision trees. International Journal of Computer Science. 9(5), 1694–0814. Available from: https://www.uetpeshawar.edu.pk/TRP-G/Dr.Nasir-Ahmad-TRP/Journals/2012/Random%20Forests%20and%20Decision%20Trees.pdf (cited 15 November 2024).

[22] Juna, A., Umer, M., Sadiq, S., et al., 2022. Water quality prediction using KNN imputer and multilayer perceptron. Water. 14, 2592. DOI: https://doi.org/10.3390/w14172592

[23] Wien, M., Schwarz, H., Oelbaum, T., 2007. Performance analysis of SVC. IEEE Transactions on Circuits and Systems for Video Technology. 17, 1194–1203. DOI: https://doi.org/10.1109/TCSVT.2007.905530

[24] Zeravan, A., Abduljabbar, T.H., Sallow, A., et al., 2023. Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review. Academic Journal of Nawroz University. 12(2). DOI: https://doi.org/10.25007/ajnu.v12n2a1612

[25] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17 August 2016. pp. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

[26] Shams, M.Y., Elshewey, A.M., El-kenawy, E.S.M., et al., 2024. Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications. 83, 35307–35334. DOI: https://doi.org/10.1007/s11042-023-16737-4