An Efficient and Highly Accurate Water Quality Detection Model Based on Machine Learning

Qingping Wang[1], Lishan Zhang[1], Zhiqiang Zhao[2], Yining Liu[3]*

[1] College of Life Sciences, Guilin University of Electronic Technology, China
[2] Information Engineering University, China
[3] School of Data Science and Artificial Intelligence, Wenzhou University of Technology, China
wqp0421@163.com, lishanzhang@guet.edu.cn, zhiqiang0909@126.com, lyn7311@sina.com

Abstract

Water quality monitoring is critical for human health and ecological protection. Traditional methods are time-consuming, costly, and complex. Machine learning algorithms can be used to increase automation and efficiency by analyzing large data sets. However, existing methods often suffer from low detection accuracy, and need to manage a large number of features. In this paper, a machine learning framework is proposed for water quality monitoring by combining mutual information feature selection and cost-sensitive learning. Mutual information is used to select the most relevant features, thus to improve the detection efficiency. Cost-sensitive learning is introduced during model training to improve the detection accuracy. An ablation study is conducted to assess the impact of mutual information feature selection and cost-sensitive learning, which in turn is evaluated, and comparative experiments are conducted with the state-of-the-art methods. The results show that the proposed framework significantly improves the model performance and enhances the accuracy and reliability of water quality monitoring. Feature importance analysis helps to gain insights into the contribution of individual features to support effective water protection.

Keywords: Machine Learning, Mutual information, Cost-sensitive learning, Water quality monitoring, Water quality detection model

1 Introduction

With the rapid advancement of industrialization and urbanization, water resource pollution has become increasingly severe, posing significant impacts on human life and ecosystems [1]. Water quality monitoring stands as a critical component in safeguarding water environment safety and maintaining ecological health. The accurate acquisition and high-efficiency, high-precision analysis of water quality data are essential for identifying pollution sources, assessing pollution levels, and formulating effective pollution control measures.

Traditional water quality monitoring primarily relies on chemical analysis methods, involving the detection of chemical components in water samples through spectroscopic, chromatographic, and mass spectrometric analyses conducted in laboratories [2]. However, these conventional methods are fraught with numerous drawbacks [3]. Firstly, they typically require specialized equipment and skilled operators, leading to higher costs [4]. Secondly, these methods often necessitate extended periods for testing, making them unsuitable for real-time monitoring [5]. Additionally, traditional methods exhibit limitations in detecting unknown chemicals or complex mixtures, complicating the accurate evaluation of overall water toxicity [6]. Due to constraints in sampling points and the quantity of monitoring equipment, traditional methods exhibit lower efficiency when addressing widely distributed water quality issues [5]. Therefore, the development of new-generation water quality monitoring technologies that are rapid, cost-effective, and user-friendly has emerged as a focal point of current research endeavors.

Machine Learning, as a critical subfield of Artificial Intelligence, has increasingly played a pivotal role in the environmental sector, particularly in addressing the intricate challenges of water quality regulation. Machine Learning exhibits significant advantages over traditional methods, notably in its ability to efficiently automate water quality assessment, thereby compensating for the time-consuming and labor-intensive nature of conventional analytical approaches [7]. This automation enhances the precision and efficiency of evaluating water quality parameters [8]. In a case study focused on municipal drinking water in Wuhan, China, Xia et al. utilized the Water Quality Index (WQI) and Artificial Neural Network models to assess the spatiotemporal distribution of water quality over an extended period, successfully forecasting water quality conditions for subsequent years [9]. Similarly, Samsudin et al. trained a predictive model based on machine learning for assessing water quality in the estuarine areas of West Malaysia [10]. Chou et al. proposed a Machine Learning approach employing the Carlson Trophic State Index to predict reservoir water quality [11]. Furthermore, Machine Learning methods have been applied to explore the impact of water quality on aquaculture in freshwater ponds. By analyzing changes in water quality variables such as dissolved oxygen, salinity, and temperature, researchers have investigated their

*Corresponding Author: Lishan Zhang; Email: lishanzhang@guet.edu.cn
DOI: https://doi.org/10.70003/160792642026112706002

effects on pond culture harvest performance [12]. These applications highlight the versatility and effectiveness of Machine Learning in advancing the domain of water resource management and environmental protection.

The water quality detection issue addressed in this paper falls under the category of anomaly detection, characterized by the challenge of imbalanced datasets where the number of samples belonging to the anomaly class is significantly fewer compared to those of the normal class. This imbalance leads to insufficient learning of the anomaly class by the model, thereby reducing the detection accuracy for anomalies. The existing research work performs poorly when exposed to the unbalanced dataset problem, in addition, a large number of features are required to perform water quality detection, which results in less efficient detection. In practical scenarios, water quality detection should be simple and efficient, meaning that high-precision detection can be achieved without requiring complex operations. In light of these challenges, the primary contributions of this study are as follows:

1. Addressing Class Imbalance: This study introduces innovative strategies to tackle the problem of class imbalance, ensuring that the model can effectively learn from the limited number of anomaly samples and improve the detection accuracy for anomalies.

2. Enhancing Anomaly Detection Accuracy: This study seeks to improve the model's sensitivity to anomalies by leveraging advanced machine learning methodologies, thereby augmenting the overall accuracy of anomaly detection within water quality monitoring.

3. Promoting Simplicity and Efficiency: The proposed methods are designed to be straightforward and efficient, enabling high-precision water quality detection with minimal operational complexity, which is crucial for practical application in real-world settings.

4. Supporting Environmental Protection: By improving the efficiency and accuracy of water quality detection, this research contributes to better environmental protection efforts, facilitating timely and effective responses to water pollution incidents.

2 The Proposed Methodology

This section provides a detailed account of the construction process of the water quality detection model, encompassing key steps such as data preprocessing, feature selection, dataset partitioning, model training, and model testing. The framework of the model is illustrated in Figure 1.

2.1 Descriptive Statistical Analysis

Descriptive statistical analysis constitutes the initial step in data preprocessing, serving to provide foundational insights necessary for subsequent data processing and model construction. By conducting statistical descriptions of the dataset, one can gain a comprehensive understanding of data distribution, central tendency such as mean and median, and dispersion such as standard deviation and variance. Moreover, descriptive statistical analysis aids in the identification of outliers and missing values within the dataset, offering guidance for data-cleansing activities.

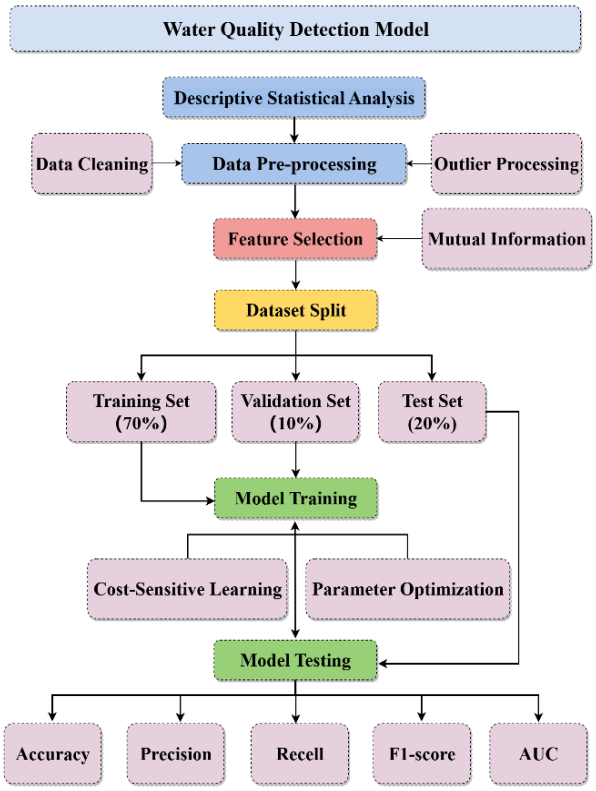


Figure 1. Flowchart of the proposed scheme
(The flowchart illustrates the proposed machine learning-based classification model. The model consists of several stages, including data preprocessing, feature selection, dataset splitting, model training, and model testing. To evaluate the performance of classification models, researchers used both the original imbalanced dataset and the balanced dataset generated through cost-sensitive learning methods. This approach ensures that the model's ability to handle class imbalance is thoroughly assessed. The proposed scheme uses mutual information to ensure that only the most relevant features are used in the classification process.)

2.2 Data Preprocessing

Data preprocessing constitutes a pivotal phase in guaranteeing data quality and mainly includes data cleaning and outlier handling:

Data Cleaning: Mainly checking for errors and inconsistencies in the dataset to improve the accuracy and reliability of the data. Data cleaning tasks include filling in missing values, correcting erroneous data, and removing duplicate records.

Outlier Handling: A large number of outliers can have a negative impact on the training and prediction of models. We use box plots to visualize each feature and identify anomalies, then take appropriate measures to handle them. It is important to note that while some outliers may deviate significantly from other data points, they might contain valuable information and should not be simply deleted [13].

2.3 Feature Selection

Feature selection is a critical step in enhancing model performance and reducing overfitting. Additionally, through feature selection engineering, unnecessary operations can be minimized. In this study, mutual information is employed to evaluate the correlation between each feature and the target variable, with the most relevant features being selected.

Mutual information is a metric used to measure the dependence between two variables, closely related to the concept of entropy in random variables. It quantifies the amount of information obtained about one variable through knowledge of another, indicating the difference between the joint distribution of random variables (X, Y) and the product of their marginal distributions. When two variables are identical, mutual information reaches its maximum value; thus, the greater the mutual information, the more correlation there is between the variables [14].

Compared to other correlation calculation methods, mutual information can capture a broader range of relationship types, including linear and non-linear relationships. Traditional correlation coefficients are limited to measuring linear relationships, whereas mutual information is not restricted by such limitations, making it more advantageous in analyzing complex systems. The mutual information between the feature and the target label is computed, the most relevant features can be identified, which improves model performance and interpretability, as well as simplifies the model and reduces computational costs [15].

The process of feature selection proceeds as follows:

1. Calculation of Mutual Information

For each feature, calculate the mutual information value with respect to the target variable. Mutual information reflects the information gain between features and labels; the greater its value, the stronger the feature-label correlation. The calculation process is as follows:

Let the joint distribution of two random variables (X, Y) be denoted as $p(x, y)$, and their marginal distributions be denoted as $p(x)$, $p(y)$ respectively. Mutual information $I(X, Y)$ is defined as the relative entropy (or Kullback-Leibler divergence) between the joint distribution and the product of the marginal distributions.

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

Here, $p(x, y)$ represents the joint probability of occurrences of x and y, while $p(x)$ and $p(y)$ represent the probabilities of occurrence of x and y independently.

2. Feature Selection

Features are selected based on their mutual information values. Generally, a threshold is established, and only features with mutual information values above this threshold are chosen. Alternatively, the top features with the highest mutual information values can be selected.

Figure 2 illustrates the mutual information values between the target variable and each feature variable in the dataset. It is observed that the importance of the 'nitrates' and 'lead' attributes is 0. Consequently, in this study, features with mutual information values greater than 0 are selected as the characteristics to be used for detection.

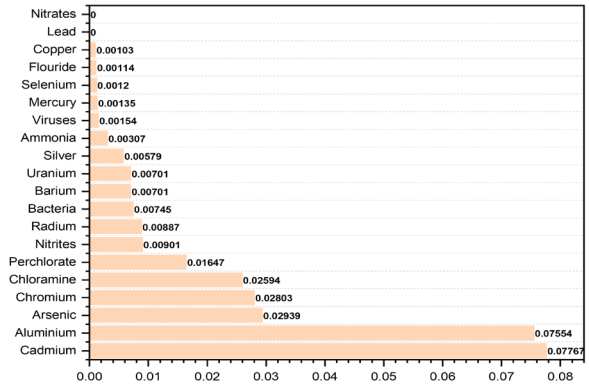


Figure 2. Mutual information value for each attribute (The mutual information values between the target variable and each attribute in the dataset are presented. According to the figure, the 'nitrates' and 'lead' attributes have mutual information values of 0, indicating no significant relationship with the target variable. Therefore, for the purpose of detection, this study selects attributes with mutual information values greater than 0, ensuring that only those features that contribute to the predictive power of the model are retained.)

2.4 Dataset Splitting

The dataset was partitioned into training, validation, and testing sets to rigorously evaluate model performance. The training set was used for model training, the validation set for hyperparameter tuning and model selection, and the test set, held aside until the end, for final performance assessment. This partitioning strategy ensured robust generalization to unseen data.

2.5 Machine Learning Models

Karthick et al. identified XGBoost, LightGBM, CatBoost, and Random Forest as high-performing machine learning algorithms in their study. In this study, we also conduct training and testing on these four models to further demonstrate the superiority of our approach [16].

XGBoost is a tool for large-scale parallel boosting trees, recognized as the fastest and most efficient open-source boosting tree toolkit, outperforming common toolkits by more than ten times in speed. In data science, a significant number of Kaggle competitors choose XGBoost for data mining competitions, making it a must-have weapon in various data science contests [17].

LightGBM is designed with two key goals: first, to minimize memory usage, enabling single machines to handle large datasets without speed compromises; and second, to reduce communication overhead, improving efficiency and achieving linear acceleration in multi-machine parallel processing [18].

CatBoost is a GBDT framework that implements fewer parameters, supports categorical variables, and offers high accuracy by using symmetric decision trees as

base learners. It efficiently processes categorical features, combining the concepts of Categorical and Boosting [19].

Random Forest is a classification algorithm that trains multiple trees on samples and predicts outcomes by aggregating the results. It is an algorithm based on ensemble learning, integrating multiple decision trees. During the training process, datasets are randomly sampled with replacement, and a subset of features is randomly selected as input for each tree.

2.6 Model Training

Model training is the core step in constructing a water quality detection model. During this phase, we adopt the following strategies to optimize the model:

1. Hyperparameter Tuning

By adjusting the hyperparameters of the model, such as the learning rate, we aim to improve model performance. Techniques like cross-validation and grid search are employed to find the optimal combination of parameters. Table 1 gives the parameter values for the four machine learning models.

Table 1. Parameters of the machine learning model

XGBoost	Max_depth	Learning rate	Objective	Eval_metric
	30	0.2	Binary	logloss
LightGBM	Num_leaves	Learning rate	Objective	Boosting_ type
	100	0.1	Binary	gbdt
CatBoost	Depth	Learning rate	Loss-function	Eval_metric
	6	0.5	MultiClass	MultiClass
Random Forest	n_estimators	bootstrap	max_depth	-
	300	True	None	-

2. Cost-Sensitive Learning:

Cost-sensitive learning is an effective method in machine learning for dealing with imbalanced datasets [20]. The underlying idea is to assign different costs to incorrect predictions for different classes during the model training process. By adjusting the costs associated with different classes, the model can pay more attention to the types of errors with higher costs.

In this study, we adopt adjusting class weights to implement cost-sensitive learning, and the specific steps are as follows:

Step 1: Count the number of samples for the safe and unsafe categories in the dataset.

Step 2: Calculate the sample weights using the following formula:

$$class_weight_i = \frac{\sum_{j=1}^{n} class_count_j}{n \times class_count_j} \tag{2}$$

Where $class_weight_i$ is the weight of the category i, n is the total number of categories, and $class_count_j$ is the number of samples in the first category. This calculation method can make the categories with fewer samples get higher weights in the model training process, thus improving the model's while learning ability for these categories.

Step 3: Balance the importance of positive and negative classes during training using the following:

$$scale_pos_weight = (\frac{class_weights_i}{class_weights_j}) \times \alpha, \tag{3}$$

which α is dynamic adjustment factor.

3 Data Preprocessing

3.1 Descriptive Statistics of the Dataset

The dataset used in this study is the water quality dataset obtained from an open-source digital library (Data Link 2021). This dataset comprises 7,999 records and is divided into 21 attributes. The last column, "is_ safe," represents the true class label for each sample, indicating whether the water sample is safe or not. The first 20 attributes represent different substances present in the water samples and serve as the feature variables. Additionally, a detailed analysis of the dataset reveals important characteristics. Figure 3 displays the histogram of the dataset, which provides insights into the basic features of the data and whether it follows any particular pattern (such as normal distribution or skewness). This visualization helps us understand and analyze the distribution characteristics of the data from multiple dimensions.

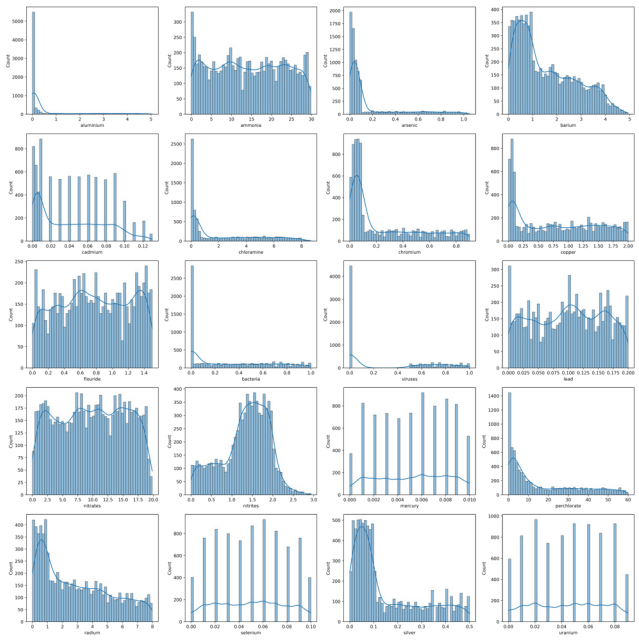


Figure 3. Histogram of sample data
(The histogram shows us clearly the basic characteristics and distribution of the data for the 20 attributes in the dataset.)

3.2 Data Cleaning

Data cleaning refers to the process of preparing and transforming data to make it suitable for analysis and modeling. This encompasses a range of activities, including the removal of duplicate entries, imputation of missing values, management of outliers, and conversion of data formats, all aimed at enhancing the accuracy and reliability of the dataset. As an essential step in the data processing pipeline, data cleaning helps eliminate errors and noise within the data, thereby improving the precision and robustness of subsequent analyses and models [21].

In this dataset, there are no duplicate values, but the attribute "ammonia" has 3 missing values. Following the removal of missing values, the resulting dataset (7,996 records, 21 attributes) exhibits a significant class imbalance, as illustrated in Figure 4. Specifically, 88.6% of instances are classified as "unsafe" (0), while only 11.4% are "safe" (1).

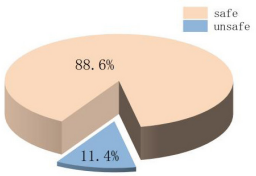


Figure 4. Visualizing the distribution of safe/unsafe labels (This dataset exhibits a significant class imbalance: 88.6% are labeled "unsafe" (0), and 11.4% "safe" (1).)

3.3 Outlier Handling

In previous studies, the presence of outliers is primarily attributed to data corruption and incorrect measurement results during the data collection process [22]. Outliers can distort results and affect the accuracy of predictions, making their detection and handling a critical step in improving the performance of machine learning models.

The box plot [23], also known as a box-and-whisker plot, is a classic statistical tool for detecting outliers by visually representing the spread and dispersion of a dataset. It highlights key elements: the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values. The box plot visually highlights outliers, the degree of spread, and the symmetry of the data distribution. Outliers are typically represented as individual points outside the whiskers of the box plot.

Figure 5 presents the box plots for the distribution of each feature in the dataset. The box plots provide a simplified format to display a large amount of information, helping to train the model on a more representative dataset, thereby improving accuracy and generalization capability. The primary advantage of box plots is their insensitivity to outliers, allowing for a precise and stable depiction of the data's dispersion, which also aids in data cleaning.

Through the analysis of box plots, it is evident that only the attributes "aluminum" and "arsenic" exhibit outliers. Although these outliers deviate from other data points, they may encapsulate valuable information. Given that the essence of water quality detection models lies in anomaly detection, it is advisable not to eliminate such outliers but rather to conduct further analysis into their causes, thereby potentially uncovering novel insights.

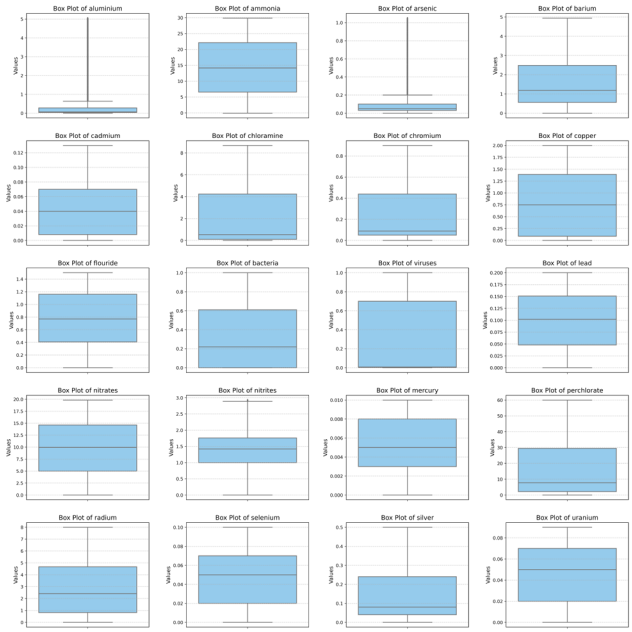


Figure 5. Box plots of feature distributions in the dataset

3.4 Dataset Splitting

The dataset was randomly partitioned into training (70%), validation (10%), and testing (20%) sets, using random_state=42 to guarantee reproducibility of the split. The training set is designated for model training, the validation set serves for optimizing the model's hyperparameters, while the testing set is reserved exclusively for evaluating the final model post-training. It is crucial to note that the testing set neither participates in the learning process nor influences the selection of hyperparameters; its sole purpose is to assess the model's performance. The division of the dataset into training, validation, and testing subsets plays a pivotal role in evaluating model performance, assessing generalization capabilities, and guiding the iterative refinement of the model [24]. After removing the three samples containing missing values, the resulting dataset consists of 7,996 samples, which were divided into training, validation, and testing sets at a ratio of approximately 70%, 10%, and 20%, respectively.

4 Experiments and Analysis

This section presents the experimental setup and analysis conducted to validate the superiority of the proposed water quality detection model. To achieve this, we employ both ablation studies and comparative experiments. Ablation studies are designed to evaluate the contribution of individual components to the overall performance of the model by systematically removing certain parts of the model. To assess the impact of each component, we will conduct ablation studies by removing

the feature selection module and the cost-penalized learning module, one at a time, and comparing the performance to the full model. Additionally, comparative experiments against the latest research are carried out to further substantiate the advantages of our method.

4.1 Evaluation Metrics

To comprehensively assess the performance of the model, four key performance metrics have been selected for comparative analysis. These metrics include: Confusion Matrix, Accuracy, Recall, Precision, F1-score.

The confusion matrix presents the relationship between the model's predicted outcomes and the actual labels in tabular form, based on the information provided in Table 2.

Table 2. Confusion matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positive (TP)	False Positive (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)

Accuracy measures the proportion of correctly predicted samples out of the total number of samples, and it is defined by the formula:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

Precision measures the proportion of true positive samples among those predicted as positive. When a model exhibits high precision, it implies a low false positive rate in the model's positive predictions, and it is defined by the formula:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

Recall quantifies the proportion of correctly identified positive instances out of all actual positive instances. A high recall indicates that the model is capable of identifying the majority of positive cases, and it is defined by the formula:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

The F1 score is the harmonic mean of precision and recall, designed to provide a comprehensive and balanced evaluation of these two key performance indicators. It offers a comprehensive evaluation, especially in cases of class imbalance, and is defined by the formula:

$$F1 - score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{7}$$

The Receiver Operating Characteristic (ROC) curve is a widely employed tool for assessing the performance of binary classification models. It demonstrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different threshold settings, helping to understand the model's behavior at varying levels of sensitivity. The area under the ROC curve (AUC) offers a single numerical metric that summarizes overall model performance, and an AUC value closer to 1 indicates better classification performance.

4.2 Ablation Study

To validate the effectiveness of the mutual information feature selection module and assess the effectiveness of the cost-penalized learning module introduced in the proposed model, a series of ablation experiments were performed based on the XGBoost model. The experimental results are summarized in Table 3.

Model 1 (Full Model): This model includes all components, namely the mutual information feature selection module, the cost-penalized learning module, and the original model components. This model aims to assess the performance improvement resulting from the addition of the two new modules.

Model 2 (Without Mutual Information Feature Selection Module): This model excludes the mutual information feature selection module while retaining all other components. By comparing the performance of this model with that of the full model, we can assess the contribution of the mutual information feature selection module to the overall model performance.

Model 3 (Without Cost-Penalized Learning Module): This model excludes the cost-penalized learning module while retaining all other components. By comparing the performance of this model with that of the full model, we can evaluate the contribution of the cost-penalized learning module to the overall model performance.

Model 4 (Baseline Model): This model excludes both the mutual information feature selection module and the cost-penalized learning module, retaining only the original model components. The purpose of this model is to serve as a baseline for comparison with the other models.

Table 3. Comparison of ablation experiment results

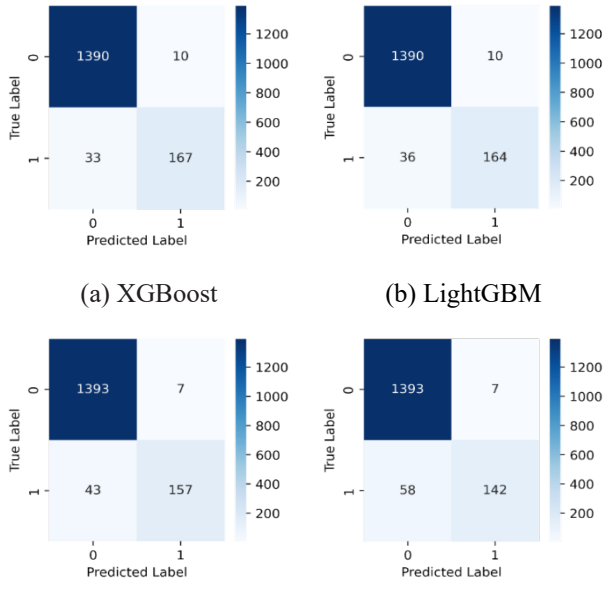
Model configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Model 1	97.31	94.35	83.50	88.60
Model 2	97.12	91.40	85.00	88.08
Model 3	97.19	95.86	81.00	87.80
Model 4	96.94	94.15	80.50	86.90

In the ablation study, we tested the XGBoost model under different configurations to evaluate the impact of the mutual information feature selection module and the cost-penalized learning module on model performance. The experimental results show that the full model (Model 1) performs the best in terms of overall performance, with an accuracy of 97.31%, precision of 94.35%, recall of 83.50%, and F1 score of 88.60%. In contrast, Model 2, which removes the Mutual Information Feature Selection

module, shows a decrease in Upper F1-score, Precision and Accuracy compared to the full model, and Model 3, which removes the Cost Penalty Learning module, shows a decrease in Accuracy, Recall and F1-score compared to the full model. In addition, Model 4, which only retains the original model components, is lower than the full model in all metrics, which further demonstrates the importance of mutual information feature selection and cost-penalty learning module for improving model performance. These results indicate that the two added modules in Model 1 play a key role in improving the predictive ability and robustness of the model.

4.3 Comparison Experiments

In this study, we compare the performance of the method proposed by Karthick et al. with our proposed method on four different machine learning models, which include LightGBM, XGBoost, CatBoost, and Random Forest. The experimental results are shown in Table 4, we performed a comprehensive evaluation in terms of four metrics, namely Accuracy, Precision, Recall, and F1-score. Figure 6 shows the confusion matrix using the four models, which clearly shows the classification results of the models on the test set. By analyzing the confusion matrix, we can calculate the performance metrics such as accuracy, precision, recall, and F1-score of the models.



(a) XGBoost (b) LightGBM

(c) Catboost (d) RandomForest

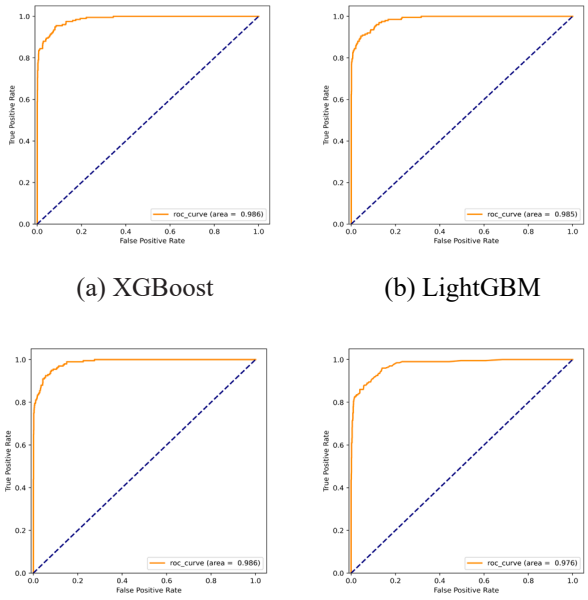
Figure 6. Confusion matrix of the proposed method under four models

For Karthick et al.'s method, the LightGBM model performed the best in terms of accuracy at 96.25 %, while the Random Forest model was slightly less accurate at 95.25 %. In terms of precision rate, the Random Forest model leads with 95.50 %, while the XGBoost model has the lowest precision rate among the four models at 93.30 %. In relation to recall, the XGBoost model performs the best with 76.00 %, while the Random Forest model has the lowest recall with 65.00 %. In terms of F1 score, the XGBoost model leads with 83.70 %, while the Random Forest model has the lowest F1 score with 77.40 %.

In comparison, our proposed method outperforms Karthick et al.'s method on all four metrics. In relation to accuracy, the XGBoost model performs the best with 97.31%, and the Random Forest model achieves 95.94%. In terms of precision, our method exceeds 94% on all models, with the CatBoost model having the highest precision at 95.73%. In terms of recall, our method also performs well, with the XGBoost model achieving 83.50% and the Random Forest model improving to 71.00%. In terms of F1 scores, the XGBoost model leads the way with 88.60%, and the Random Forest model has an F1 score of 81.37%, which is still higher than the method of Karthick et al.

Table 4. Comparative experimental results

		LightGBM	XGBoost	CatBoost	Random Forest
Karthick et al. methodology	Accuracy (%)	96.25	96.31	95.44	95.25
	Precision (%)	94.90	93.30	93.80	95.50
	Recall (%)	74.00	76.00	68.00	65.00
	F1-score (%)	83.10	83.70	78.80	77.40
Our proposed methodology	Accuracy (%)	**97.12**	**97.31**	**96.88**	**95.94**
	Precision (%)	**94.25**	**94.35**	**95.73**	**95.30**
	Recall (%)	**82.00**	**83.50**	**78.50**	**71.00**
	F1-score (%)	**87.70**	**88.60**	**86.26**	**81.37**



(a) XGBoost (b) LightGBM

(c) Catboost (d) RandomForest

Figure 7. ROC curves of the proposed method under four models

Figure 7 demonstrates the ROC curves using the four models, which shows the variation of TPR and FPR of the

models under different classification thresholds, and it can be seen that the proposed method has a better performance under all four models. Taken together, our proposed method shows better performance on all tested machine learning models, which is significantly better than the method of Karthick et al. This suggests that our method is able to perform water quality testing more accurately. These results not only validate the effectiveness of our method, but also provide new directions and references for future research.

4.4 Feature Importance Analysis

In this experiment, we constructed a water quality detection model and trained and evaluated four different machine learning models. As shown in Figure 8, both the XGBoost model and the LightGBM model, which have better detection performance, obtained the top five feature importance levels of aluminium, perchlorate, ammonia, radium, and chloramine in descending order. By analyzing the feature importance of these models after training, we were able to gain insights into the contribution and impact of individual feature's contribution and impact in water quality detection. The feature importance analysis not only helped us to filter out the indicators that are crucial for water quality judgment, but also provided guidance for model optimization and improvement. Through this experimental analysis, we hope to improve the accuracy and reliability of water quality testing and provide strong support for water environmental protection.

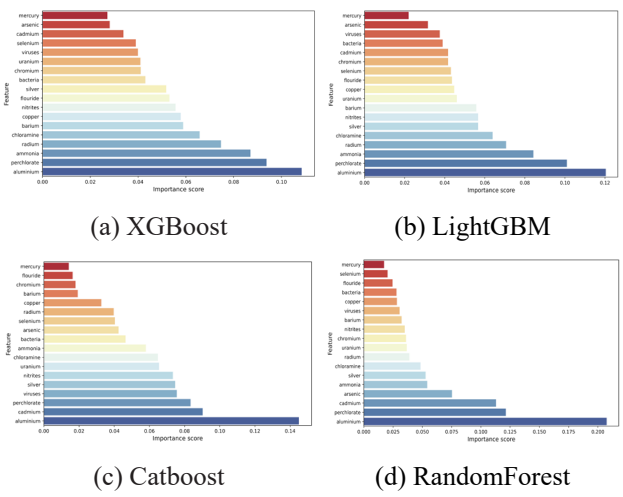


Figure 8. Feature importance analysis

5 Conclusion

This paper introduces a machine learning model for water quality detection based on mutual information feature selection and cost-sensitive learning, aiming to overcome the limitations inherent in traditional water quality detection methods. By enhancing the feature selection and model training processes, the proposed model significantly improves the efficiency and accuracy of water quality detection. Experimental results indicate that incorporating mutual information feature selection effectively reduces redundant features and enhances model performance, while the introduction of cost-sensitive learning contributes to increased detection accuracy and reduced risk of misclassification. Compared to the latest existing research, the method proposed in this paper demonstrates significant advantages in terms of accuracy and reliability. Furthermore, the analysis of feature importance reveals the distinct contributions of various features in water quality detection, providing valuable guidance for the formulation of strategies for water environmental protection. In summary, the application of machine learning-based water quality detection methods holds broad prospects. Future research could focus on further optimizing feature selection and model training mechanisms to advance the development of intelligent water quality monitoring technologies.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under grant No. 62072133, in part by Wenzhou Major Scientific and Technological Innovation Projects (ZG2023028).

References

[1] J. Aryal, B. Gautam, N. Sapkota, Drinking water quality assessment, *Journal of Nepal Health Research Council*, Vol. 10, No. 22, pp. 192-196, September-December, 2012.

[2] Y. Q. Zhou, L. L. Chen, L. Zhou, Y. L. Zhang, K. Peng, Z. J. Gong, K. S. Jang, R. G. M. Spencer, E. Jeppesen, J. D. Brookes, D. N. Kothawala, F. C. Wu, Key factors driving dissolved organic matter composition and bioavailability in lakes situated along the Eastern Route of the South-to-North Water Diversion Project, China, *Water Research*, Vol. 233, Article No. 119782, April, 2023.

[3] Z. Y. Su, T. Liu, Y. Men, S. Li, N. Graham, W. Z. Yu, Understanding point-of-use tap water quality: From instrument measurement to intelligent analysis using sample filtration, *Water Research*, Vol. 225, Article No. 119205, October, 2022.

[4] N. Mandal, S. Mitra, D. Bandyopadhyay, Paper-Sensors for Point-of-Care Monitoring of Drinking Water Quality, *Ieee Sensors Journal*, Vol. 19, No. 18, pp. 7936-7941, September, 2019.

[5] X. Fu, R. Wu, H. Qi, H. Yin, Long-term trends in surface water quality of China's seven major basins based on water quality identification index and big data analysis, *Environmental Impact Assessment Review*, Vol. 100, pp. 1-10, May, 2023.

[6] L. Zhang, Y. Zhang, M. Zhu, L. Chen, B. Wu, A critical review on quantitative evaluation of aqueous toxicity in water quality assessment, *Chemosphere*, Vol. 342, Article No. 140159, November, 2023.

[7] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, T. Zimmermann, Software engineering for machine learning: A case study, *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Montreal, QC, Canada, 2019, pp. 291-300.

[8] S. Kaddoura, Evaluation of Machine Learning Algorithm

on Drinking Water Quality for Better Sustainability, *Sustainability*, Vol. 14, No. 18, pp. 1-18, September, 2022.

[9] F. Y. Xia, H. T. Liu, J. Zhang, D. Q. Wang, Migration characteristics of microplastics based on source-sink investigation in a typical urban wetland, *Water Research*, Vol. 213, pp. 1-12, April, 2022.

[10] M. S. Samsudin, A. Azid, S. I. Khalit, M. S. A. Sani, F. Lananan, Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones, *Marine Pollution Bulletin*, Vol. 141, pp. 472-481, April, 2019.

[11] J. S. Chou, C. C. Ho, H. S. Hoang, Determining quality of water in reservoir using machine learning, *Ecological Informatics*, Vol. 44, pp. 57-75, March, 2018.

[12] M. Rana, A. Rahman, J. Dabrowski, S. Arnold, J. McCulloch, B. Pais, Machine learning approach to investigate the influence of water quality on aquatic livestock in freshwater ponds, *Biosystems Engineering*, Vol. 208, pp. 164-175, August, 2021.

[13] H. Aguinis, R. K. Gottfredson, H. Joo, Best-Practice Recommendations for Defining, Identifying, and Handling Outliers, *Organizational Research Methods*, Vol. 16, No. 2, pp. 270-301, April, 2013.

[14] M. P. Uddin, Y. Xiang, X. Q. Lu, J. Yearwood, L. X. Gao, Mutual Information Driven Federated Learning, *Ieee Transactions on Parallel and Distributed Systems*, Vol. 32, No. 7, pp. 1526-1538, July, 2021.

[15] K. Rajab, F. Kamalov, Finite sample based mutual information, *Ieee Access*, Vol. 9, pp. 118871-118879, 2021.

[16] K. Karthick, S. Krishnan, R. Manikandan, Water quality prediction: A data-driven approach exploiting advanced machine learning algorithms with data augmentation, *Journal of Water and Climate Change*, Vol. 15, No. 2, pp. 431-452, February, 2024.

[17] J. T. Li, X. Q. An, Q. Y. Li, C. Wang, H. M. Yu, X. Y. Zhou, Y. A. Geng, Application of XGBoost algorithm in the optimization of pollutant concentration, *Atmospheric Research*, Vol. 276, pp. 1-16, October, 2022.

[18] D. Mahayana, Data-Driven LightGBM Controller for Robotic Manipulator, *Ieee Access*, Vol. 12, pp. 40883-40893, 2024.

[19] J. T. Hancock, T. M. Khoshgoftaar, CatBoost for big data: an interdisciplinary review, *Journal of Big Data*, Vol. 7, No. 1, pp. 1-45, November, 2020.

[20] A. Fernández, V. López, M. Galar, M. J. del Jesus, F. Herrera, Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches, *Knowledge-Based Systems*, Vol. 42, pp. 97-110, April, 2013.

[21] X. Chu, I. F. Ilyas, S. Krishnan, J. Wang, Data Cleaning: Overview and Emerging Challenges, *Proceedings of the 2016 International Conference on Management of Data (SIGMOD)*, San Francisco, CA, USA, 2016, pp. 2201-2206.

[22] A. Idini, An introduction to computational complexity and statistical learning theory applied to nuclear models, *Journal of Physics: Conference Series*, Vol. 2586, No. 1, Article No. 012155, 2023.

[23] S. Demir, E. K. Sahin, The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB: a study on undrained shear strength prediction, *Stochastic Environmental Research and Risk Assessment*, Vol. 38, No. 8, pp. 3273-3290, August, 2024.

[24] Yun Xu, R. Goodacre, 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*, Vol. 2, No. 3, pp. 249-262, July, 2018.

Biographies



Qingping Wang received the B.S. degree in Civil Engineering from Zhengzhou University in 2022. She is currently a master student majoring in Environmental Engineering at Guilin University of Electronic Technology. Her research interests lie in intelligent environmental protection and the construction of environmental ecological data models.



Lishan Zhang, an Australian citizen, received the Ph.D. degree of Engineering in Environmental Engineering from the University of Queensland in 2010. He is currently a professor at the School of Life and Environmental Sciences, Guilin University of Electronic Technology. His research focuses on ecological engineering treatment technologies for water environments.



Zhiqiang Zhao received the M.S. degree in computer science and technology from Guilin University of Electronic Technology, China, in 2024. He is currently a Ph.D. candidate of Information Engineering University, Zhengzhou, China. His research interests include data security and machine learning.



Yining Liu (Senior Member, IEEE) received the Ph.D. degree in mathematics from Hubei University, Wuhan, in 2007. He is currently a Professor with the School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, China. His research interests include AI and privacy.