Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9555790, 13 pages
https://doi.org/10.1155/2022/9555790

Research Article

Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification

Lei Xin[1] **and Tianyu Mou**[2]

[1]*College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China*
[2]*School of Marine Electrical Engineering, Dalian Maritime University, Dalian, China*

Correspondence should be addressed to Lei Xin; 9201310419@stu.njau.edu.cn

Received 30 May 2022; Revised 14 June 2022; Accepted 20 June 2022; Published 26 July 2022

Academic Editor: Chia-Huei Wu

With the development of society and the accelerated industrialization, the problem of water pollution has become increasingly prominent. In order to stop the gathering and diffusion of harmful substances in water bodies, leading to further deterioration of water quality and more serious environmental problems, environmental management departments have developed a series of pollutant discharge standards to prevent water pollution in real time. Common testing methods are the colorimetric method and TDS (total dissolved solids) value testing method, which are mostly through water bodies that contain acid, alkali, salt, and other indicators of the concentration test, to produce an assessment of water quality. However, the traditional methods of water quality testing, whether in the measurement time or in the accuracy of the test, are certain defects. In order to be able to quickly detect the concentration of water quality indicators in water bodies, timely response and treatment of highly polluted water bodies are urgently needed. In this paper, we propose a water quality detection classification model based on multimodal machine learning algorithm. Firstly, we preprocessed and analyzed the collected water quality dataset and determined the reasonable and perfect water quality classification influencing factors. Then, we successively built 15 kinds of classification models based on machine learning algorithms for water quality detection. At the same time, we evaluated the performance of each model. From the four evaluation indexes of precision, recall rate, F1 value, and accuracy, respectively, the real value is compared with the predicted value of each model. The experimental results show that sulfate, pH, solids, and hardness are the important influencing factors to perform water quality testing. And the three models XGBoost (Extreme Gradient Boosting), CatBoost (Categorical Boosting), and LGBM (Light Gradient Boosting Machine) have better performances in conducting water quality testing. Finally, we further optimized the classification models based on XGBoost, CatBoost, and LGBM by using two major tools: cross-validation and hyperparameter tuning.

1. Introduction

Water is an important environmental resource that has a variety of uses. For example, water is indispensable for irrigation of crops, and drinking water is indispensable for our daily life. However, with the global climate change, urbanization, and industrialization, a large amount of domestic sewage and industrial wastewater is discharged into natural water bodies, and the pollutants in natural water bodies are continuously enriched, and the available water resources are under serious threat. Therefore, how to solve the pollution of water bodies and use relevant technologies to control it has become an increasing concern of scholars and governmental and corporate sectors. To control the concentration and discharge of toxic and harmful pollutants, we need to first introduce various water quality influencing factors, and most of these factors are related to chemical substances. For example, according to the *Guidelines for Drinking-Water Quality* released by World Health Organization (WHO) in 2011, a certain range of sodium, iron, manganese, and other chemicals will lead to poor water quality, thus affecting the taste of drinking. The Food and Agriculture Organization (FAO) of the United Nations has also published data showing that many inorganic salts can affect the growth of crops when they exceed the prescribed levels.

The manual water quality testing process is labor-intensive and time-consuming. This traditional testing has many shortcomings, but the safety of human water and the balance of aquatic ecosystems are such important and urgent issues that we need to solve. Therefore, we propose the use of machine learning models to efficiently and rapidly detect and monitor water quality in real time.

In the traditional water quality analysis, regression analysis is often used to determine the correlation between variables. For example, some scholars have used regression analysis to process data on the relationship between activity ratios of 234 U and 238 U and TDS in Saudi groundwater [1]; Mamun and Kwang-Guk also used regression analysis to evaluate the water quality of Yeongsan River in Korea with pollution source analysis and found that the river can be influenced by other systems [2]. Due to the existence of uncertain and nonlinear factors in the data, neural network has become a very popular water quality prediction method with its own good nonlinear mapping ability. For example, in the research on water quality by Liu and Wang, they established a water quality prediction model based on LSTM (long short-term memory) network under big data in combination with in-depth learning method [3]; some scholars have also established a data-driven model based on BP neural network to predict and analyze water quality in time and space, providing a basis for decision-making and disposal of sudden water pollution events [4]. With the rapid development of computer vision-related theory and application research, some researchers have found that the important index of water turbidity that determines whether tap water can be drunk can be detected by using image processing technology. A convolution neural network turbidity detection system based on embedded platform is designed. Compared with the traditional turbidity identification method, this scheme not only has lower cost, but also has strong real-time performance, And it has high accuracy. Therefore, the new detection method has good application value [5].

However, traditional analysis methods have limitations in solving large-scale dynamic problems. And since the context of the problem is relatively simple, using deep learning techniques is too complex and does not offer significant advantages. This has led to the introduction of machine learning theory for improvement. Early researchers have developed automated machine learning (AutoML) models that use multivariate statistical analysis and water quality indices to automate the assessment of water body pollutant migration transformations over time and space. This saves time and reduces development costs [6]; similarly, it has been proposed that AutoML models can be used to define controlling factors in water bodies and explain changes in alpha and beta in the aggregate. It is also emphasized that AutoML models can be used to process datasets with different geological and hydrological conditions [7]. Varol et al. used cluster analysis, factor analysis, and principal component analysis to iteratively analyze data with less polluted, severely polluted, and seriously polluted water and found that soluble salts, organic pollution, and nutrients were the main factors responsible for water quality changes [8]. In addition, support vector machines (SVM) have also been proposed to have broad research prospects in monitoring water quality indicators. For example, Yu et al. used support vector machines (SVM) to perform multicomponent spectroscopy on four drugs with similar fluorescence properties that could not be distinguished and found that the error in predicting the concentration in the components using fluorescence spectroscopy was less than 0.1% [9]. When we encounter complex samples with high variance and low bias, the variance can be reduced by using a random forest model for drug effectiveness. For example, Fang et al. used the random forest method to construct a lake response model with anthropogenic nutrient inputs, hoping to identify the sources of factors that have a critical impact on water quality and respond to reduce their discharge [10]. In Al-Mukhtar's study of suspended matter in the Diglis River, random forest, support vector machine, and artificial neural network methods were applied simultaneously to predict sediment content, resulting in effective reduction of sediment content in the river water [11]. The integrated learning algorithm has a great advantage over other algorithms in terms of accuracy. In addition to the random forest algorithm mentioned above [12, 13], XGBoost algorithm [14], CatBoost algorithm, and other integrated learning algorithms are also widely used in studies about water bodies. For example, J. H. Lee et al. analyzed the short-term prediction of urban water quality by using decision tree, random forest, and XGBoost algorithms to model the variation of dissolved oxygen (DO) and found that electrical conductivity, cumulative precipitation, total nitrogen, and water temperature are the key factors limiting DO [15].

As we can see, water quality testing has been a headache for public administrators due to the backwardness of traditional methods. Thus, in this paper, we have established an effective water quality testing model for real-time and repetitive testing tasks by introducing a machine learning approach. In terms of the cost of use, the labor-intensive process and time of manual labor are greatly reduced [16]. Specifically, our contributions are as follows.

(1) We built three efficient machine learning frameworks, namely XGBoost, CatBoost, and LGBM for water quality and obtained excellent classification accuracy

(2) We introduced 12 additional basic machine learning models as a benchmark for comparative evaluation, which objectively and realistically reflects the effectiveness of our adopted models

(3) We performed hyperparameter optimization for the existing model and used cross-validation to determine multiple sets of optimal parameters, which substantially improved the robustness and generalization ability of the model

The remaining chapters of this paper are organized as follows: Chapter two gives a detailed description of the dataset and discusses our model work. Chapter three is about our experiments and discussion section. The last chapter provides the corresponding summary.

2. Related Work

The quality of water bodies has a great impact on people's production and life. In the natural environment, water resources are easily affected due to the lack of supervision of sewage discharge, resulting in the overall quality of water bodies declining. Therefore, real-time and accurate testing of water quality is of great practical importance. In this chapter, we will first provide a brief introduction to the water quality detection dataset. Then we will focus on the explanation of the three major machine learning methods, XGBoost, CatBoost, and LGBM.

2.1. Dataset Introduction

2.1.1. Dataset Collection. In this paper, we used the open-source water quality testing dataset on the kappa platform. The data includes 3,276 water quality data collected under different water quality conditions. The attribute values, i.e., water quality influencing factors, are pH, hardness, total dissolved solids (TDS), chloramine, sulfate, electrical conductivity, organic carbon, trihalomethanes (THM), turbidity, and potability. Details of these ten influencing factors are as follows.

(1) pH is an important indicator of the acidity and alkalinity of a water body. It is also an indicator of the acidic or alkaline conditions of the water state. WHO recommends a maximum permissible limit of pH of 6.5 to 8.5. The current range of investigation is 6.52-6.83, which is in line with WHO standards

(2) Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from the geological deposits through which the water passes. The length of time that the water is in contact with the hardness-producing material helps to determine the hardness in the raw water. Hardness is initially defined as the total amount of calcium and magnesium ions in the water body

(3) Water is capable of dissolving a variety of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonate, chloride, magnesium, and sulfate. These minerals produce unwanted flavors and dilute the color in the appearance of the water. This is an important parameter for water use. High mineralization of water with TDS indicates high mineralization of water. In drinking water, the optimum concentration of TDS in water is 500 mg/L, and the maximum limit is 1000 mg/L

(4) Chloramines and chlorine are widely used in public disinfection due to their ability to produce reactive chlorine species, which are very powerful oxidizers. Chloramines are usually formed when ammonia is added to chlorine to treat drinking water. Chlorine levels of up to 4 milligrams per liter (mg/L or parts per million (ppm)) in drinking water are considered safe

(5) Sulfates are naturally occurring substances found in minerals, soils, and rocks. Groundwater, air, plants, and food all contain large amounts of sulfate. The main commercial use of sulfate is in the chemical industry. The concentration of sulfate in seawater is about 2700 milligrams per liter (mg/L). In most freshwater supplies, the concentration ranges from 3 to 30 mg/L, although much higher concentrations (1000 mg/L) are found in some geographic locations

(6) Pure water is not a good conductor of electric current but a good insulator. Usually, the amount of dissolved solids in the water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic processes in a solution that allow it to transmit current. According to WHO standards, the EC value should not exceed 400 μS/cm

(7) Total organic carbon (TOC) in source water is derived from decaying natural organic matter (NOM) and synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to the United States Environmental Protection Agency (US EPA), the TOC content in treated/drinking water should be <2 mg/L, and the TOC content in source water used for treatment should be <4 mg/L

(8) THMs are chemicals that may be found in chlorine-treated water. The concentration of THMs in drinking water depends on the amount of organic matter in the water, the amount of chlorine required to treat the water, and the temperature of the water being treated. THM levels of up to 80 ppm in drinking water are considered safe

(9) Turbidity of water depends on the amount of solid matter in suspension. It is a measure of the luminous properties of water, and the test is used to indicate the quality of the wastewater discharge in relation to colloidal substances. The average turbidity value for the Wondo Genet Campus (0.98 NTU) is below 5.00 NTU recommended by WHO

2.1.2. Dataset Description. We divided the 3,276 water quality data into high-quality water bodies and polluted water bodies according to drinking water standards. 0 indicates high-quality water bodies and 1 indicates polluted water bodies. As shown in Figure 1, 61.0% of the water bodies are of high quality, and 39.0% are of polluted water bodies, which is also in line with the good to bad ratio of water bodies in our actual scenario.

Then, we visualize the distribution of each feature for type 0 and type 1 separately. The results are shown in Figure 2.

We found that the factors affecting water quality are normally distributed. Then we performed outlier detection on the characteristic distributions of type 0 and type 1 data, respectively. The results are shown in Figure 3.

The data itself has obvious outlier points for all the influencing factors except for the distribution of three

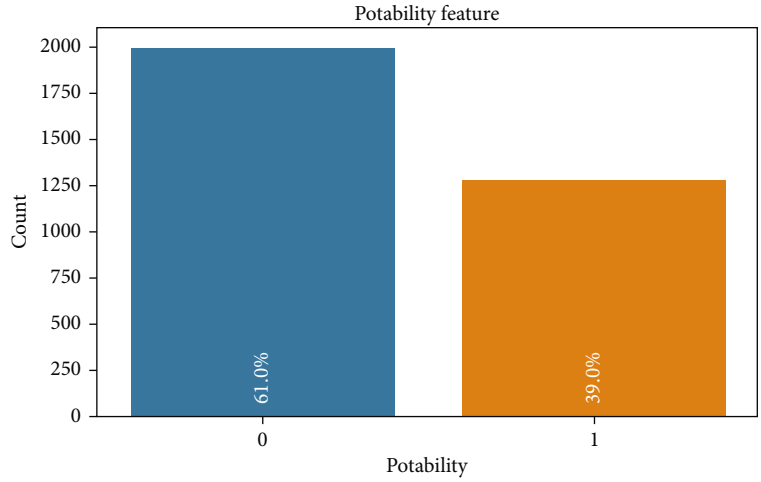

Figure 1: Data proportion.

indicators of electrical conductivity, organic carbon, and turbidity.

Finally, we conducted a correlation analysis of the statistical characteristics of the water quality data. The correlation matrix was obtained as in Figure 4.

From the results, it can be seen that the correlations among the 9 variables are small and relatively independent from each other. Therefore, all 9 feature data are used as the input of the machine learning model in the experiment part of this paper.

2.1.3. Dataset Splitting. This is an important step in running the model. Splitting the dataset into a training set and a test set are the key to start the model. Next, the machine learning model is trained, and the stability of the system is improved by testing the model. The data is split into training and test sets in a ratio of 3 : 1. Thus, out of the 3,276 data, 2,457 data were used as training set, and 819 data were used as test set.

2.2. Water Quality Classification Model. Classification algorithms and regression algorithms are the two common supervised learning algorithms; the difference between them is whether the type of variables output is continuous or not. Classification algorithms output continuous variables, while regression algorithms output discrete variables [17]. The water quality monitoring process is a classification process and can be handled using classification algorithms. In the subsections, we will focus on the algorithmic process of the three mainstream artifacts XGBoost, CatBoost, and LGBM as classification models in the framework of GBDT (gradient boosting decision tree) algorithm. All these methods are an improved implementation in the framework of GBDT algorithm.

2.2.1. XGBoost Algorithm. The XGBoost algorithm is transformed from the Boost algorithm framework and belongs to the integrated learning algorithm in machine learning. It is an efficient classification algorithm that can be applied to unbalanced datasets. In the algorithm, different classification trees are aggregated together, and the predicted values are obtained by summing each tree [18].

As shown in Figure 5, this algorithm is significantly superior in terms of speed and fault tolerance due to the clustering of multiple weak classifiers into one strong classifier in the algorithm, so that the weak classifiers can compensate each other's deficiencies. At the same time, to avoid data loss, we use a leftward growing strategy when growing the tree [19].

First, after extracting all the relevant variables related to water quality monitoring as feature vectors, we get the dataset D, which is defined as $\{(xi, yi), i = 1 \cdots n, xi \in Rm, yi \in R\}$, where n denotes the number of samples and m denotes the number of features for each sample, assuming that there are K regression trees in XGBoost, the model can be defined as

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F. \tag{1}$$

The role of each node of the tree in the XGBoost model is to do feature splitting, so the metric of the number of times a feature is selected as a split feature can be used as a criterion for judging the importance of the feature for the water quality detection classification task.

2.2.2. CatBoost Algorithm. Gradient boosting decision trees (GBTD) are often used in many different types of big data processing, and it has also been widely studied in recent years. In April 2017, Yandex developed a CatBoost model based on GBTD. The model mainly has the following advantages: it enriches the dimensionality of data processing by combining category-based features during data processing; the introduction of the leaf node method can reduce overfitting.

The main steps of the CatBoost algorithm are (1) binarization of the data, (2) conversion of categorical data into numerical data, and (3) splitting and combining of feature combinations according to the "greedy strategy." In addition, the tree is constructed by first determining the structure of the tree, then determining the leaf node values and splitting them, and selecting the best splitting scheme

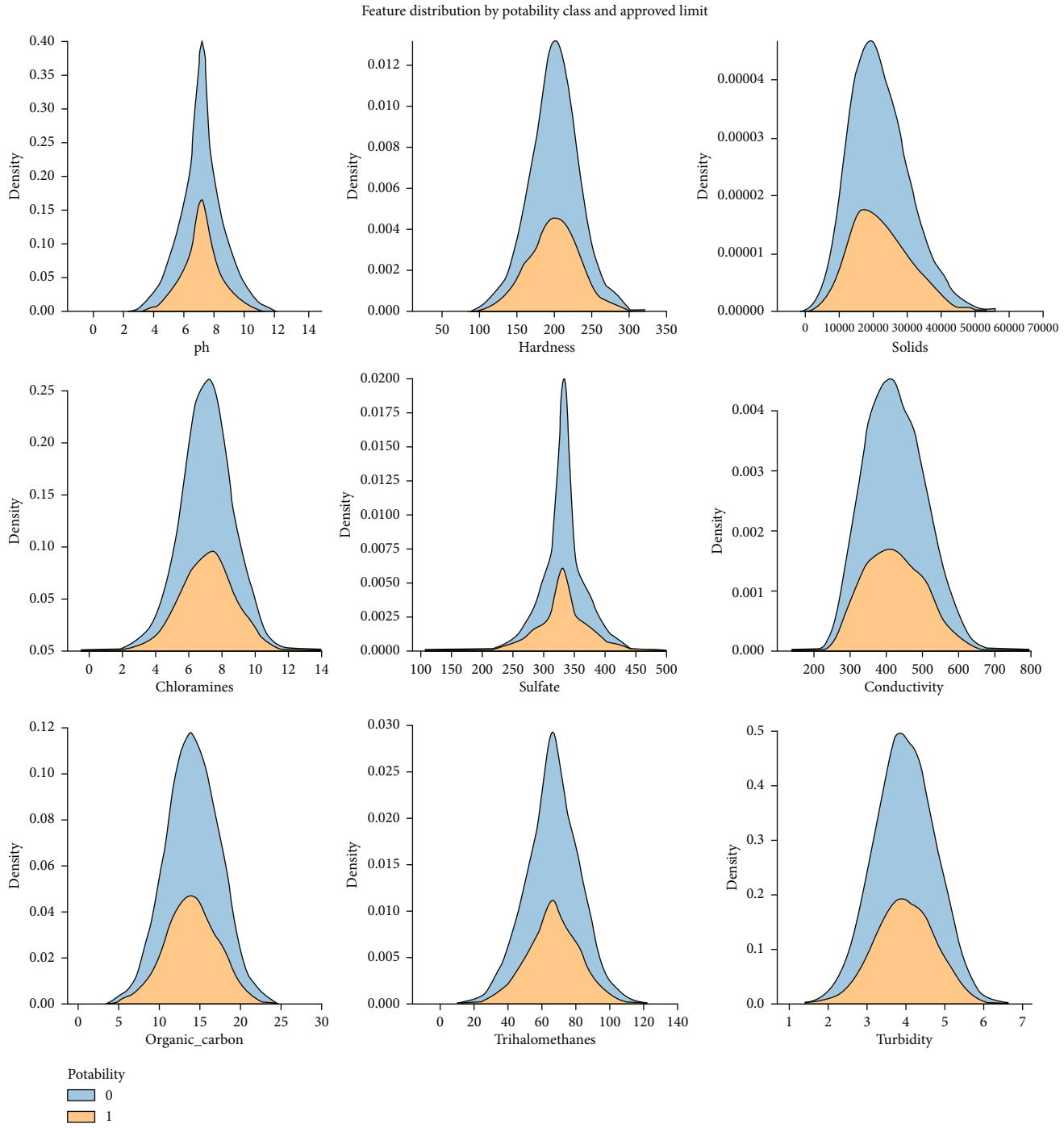

Figure 2: visualization type data scale.

according to the calculated leaf node values. The best splitting scheme is selected according to the calculated leaf node values [20].

Among them, in terms of processing deviations, the CatBoost algorithm completes the tree construction in two steps: first, the tree structure is selected, and the values of the leaf nodes are calculated after the tree structure is fixed; second, different splitting methods are enumerated, and the obtained tree is scored by calculating the values of the leaf nodes to select the best split.

CatBoost achieves simultaneous processing of the training dataset and the processing dataset, which effectively improves the shortage of feature processing efficiency of the GBTD model; at the same time, it generates a random arrangement of the training set and uses nonrepeated data to train the model, which reduces overfitting [21].

2.2.3. LGBM Model. Commonly used machine learning algorithms, such as neural networks, can be trained in a minibatch fashion, and the size of the training data is not

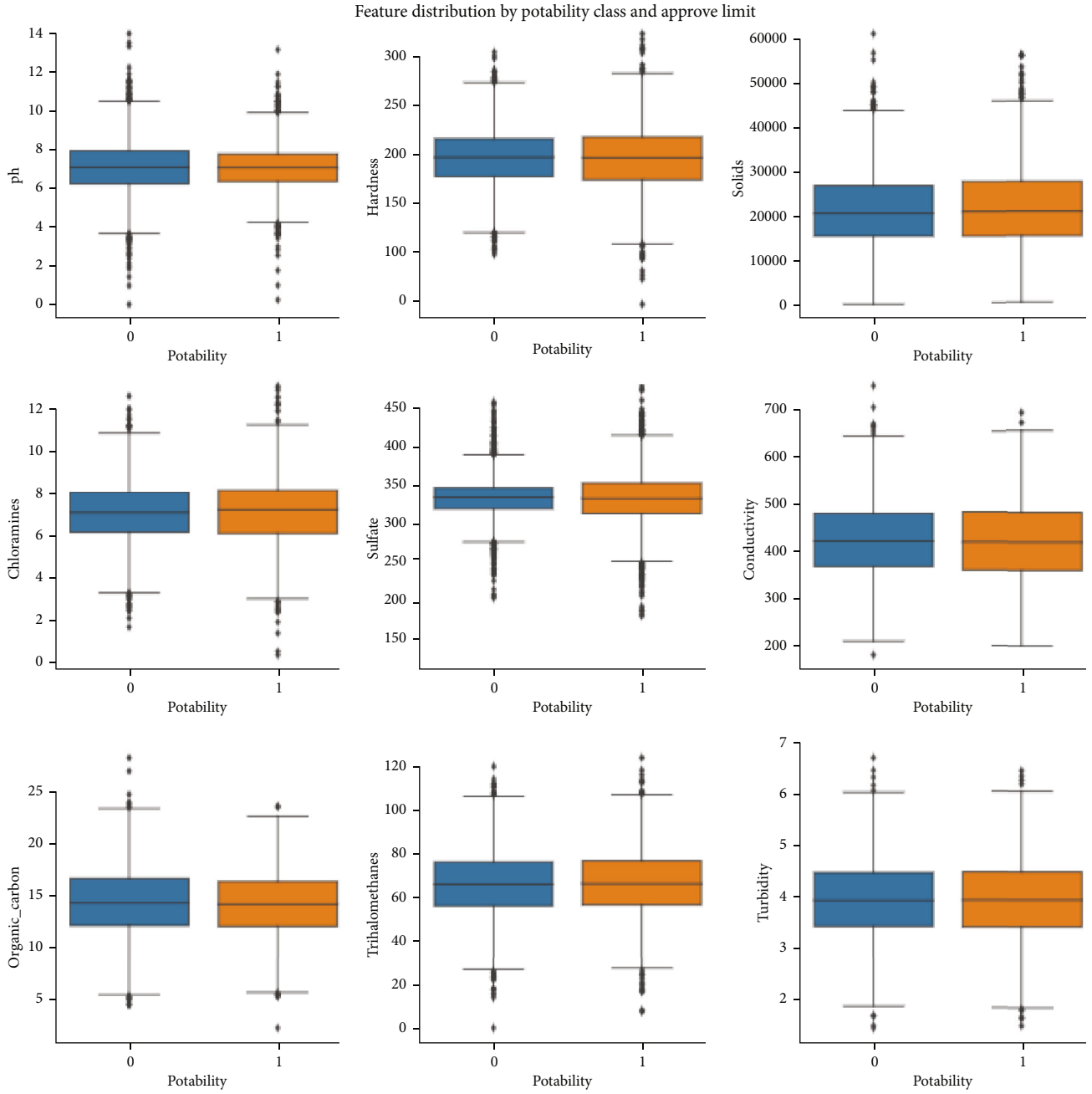

Figure 3: Data distribution box plot statistics.

limited by memory. GBDT, on the other hand, needs to traverse the entire training data several times during each iteration. If the entire training data is loaded into memory, it will limit the size of the training data; if it is not loaded into memory, repeatedly reading and writing the training data will consume a lot of time. Especially in the face of industrial-grade massive data, the ordinary GBDT algorithm cannot meet the demand. Therefore, in order to solve the problems encountered by GBDT in massive data and make GBDT better and faster to be used in industrial practice, LGBM was invented.

LGBM model is an optimization model based on decision tree, which mainly includes one-sided decision process and feature bundling process. The GOSS (Gradient-Based One-Side Sampling) algorithm is used to reduce the dimensionality of the sample data in the unilateral decision process; and the EFB algorithm is also introduced to reduce the high-dimensional feature data elements and reduce the complexity of data processing. Compared with the traditional boosting method, the LGBM model can show higher accuracy and computing speed when dealing with a large amount of high-dimensional data [22].

3. Experiments and Analysis

In the previous chapters, we have introduced the classification model for water quality detection based on machine

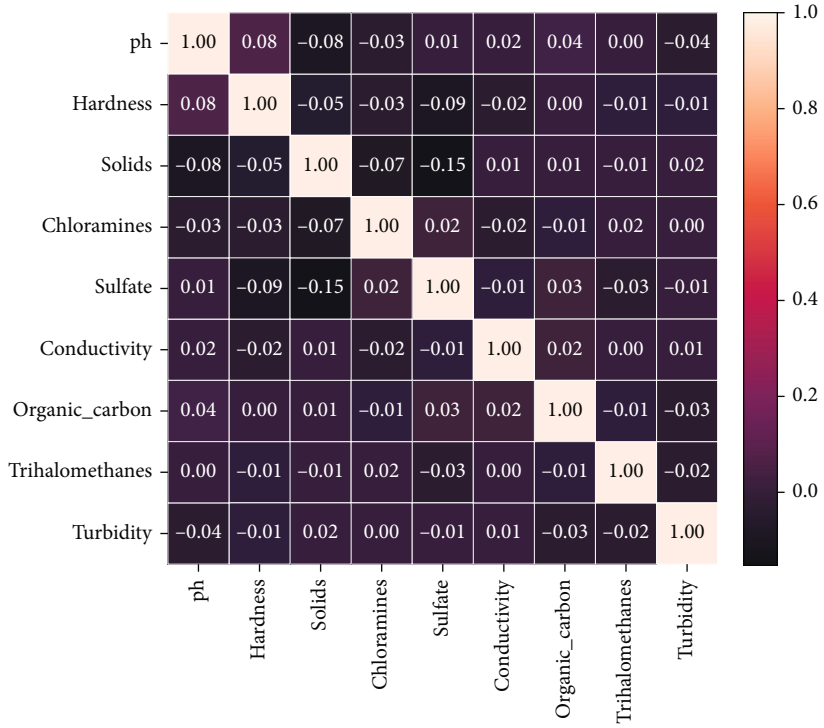

Figure 4: Data correlation matrix.

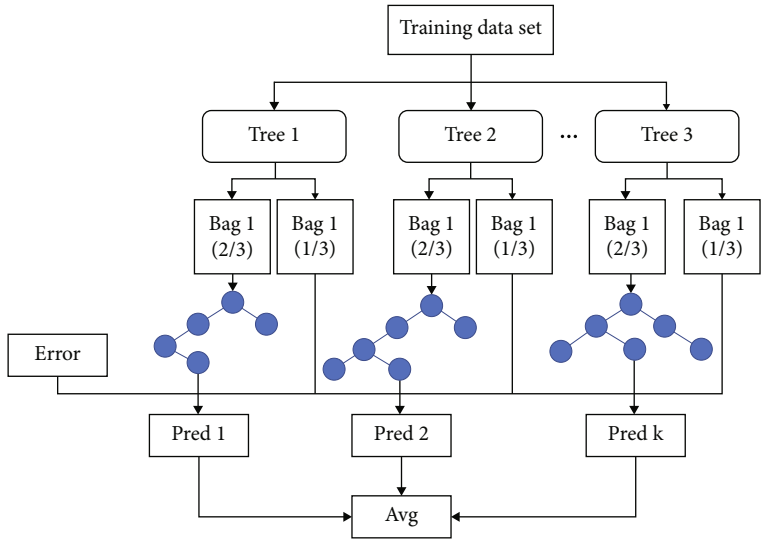

Figure 5: Flow chart of integrated learning algorithm analysis.

learning. Next, in this chapter, we will use the collected water quality detection dataset to test the identification performance of each model. Here, in addition to our previous introduction of the GBDT algorithm framework under the three mainstream artifacts of XGBoost, CatBoost, and LGBM as classification models, we also built a 12 additional common machine learning water quality detection classification models such as Bayesian and SVM to assist in the model evaluation.

3.1. Experiment Environment. The experiment environment for this paper is configured with an i5-8300H 2.30GHz processor and NVIDIA GTX 1050Ti graphics card, and the experiments were conducted under Python 3.6 and PyCaret 1.0.0.

3.2. Evaluation Indicators of Experiments. In order to comprehensively evaluate the performance of the classification model for water quality testing, four metrics of precision, recall rate, F1 value, and accuracy were chosen as the judging criteria in this paper. Table 1 shows the confusion matrix composed of the labels of the true and predicted results.

Table 1: Confusion matrix.

True results	Forecast results	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Among them, the four classification criteria indicators of water quality testing are calculated as shown in equations (2)–(5).

The precision rate is calculated by the formula:

$$P = \frac{TP}{(TP + FP)}. \tag{2}$$

The accuracy rate is calculated by the formula:

$$Acc = \left(\frac{TP + TN}{TP + TN + FP + FN}\right). \tag{3}$$

The recall rate is calculated by the formula:

$$R = \frac{TP}{(TP + FN)}. \tag{4}$$

The F1 score is the summed average of the precision and recall rates:

$$F1 = 2 \times \frac{P \times R}{P + R}. \tag{5}$$

Among them, ACC can be understood as the probability of correct prediction. Its defect is that when the proportion of positive and negative samples is very uneven, the category with a large proportion will affect acc. If 99% of the abnormal points are nonabnormal points, we will regard all samples as nonabnormal points, and the ACC will be very high; P can be understood as how many predicted things are of interest to users; R can be understood as how many things users are interested in are predicted. Generally speaking, P and R are a pair of contradictory measures. In order to better characterize the performance measurement of the learner in P and R, we can also introduce F1 value. Further, this paper compares the accuracy differences between water quality detection classification models by the magnitude of the area under the ROC curve. The area under the ROC curve is denoted as AUC, which can be used to evaluate the performance of the classifier model, and is calculated as

$$AUC = \frac{\sum_{i=1}^{n_0} r - n_0 \times (n_0 + 1)/2}{n_0 \times n_1}. \tag{6}$$

3.3. Experimental Comparison of Data Balancing. In order to deal with unbalanced data, either upsampling or downsampling methods are generally used. Since, the downsampling method is prone to data loss and the upsampling method is prone to overfitting, the SMOTE (Synthetic Minority Oversampling Technique) algorithm is chosen in this paper to accomplish the task of expanding the dataset.

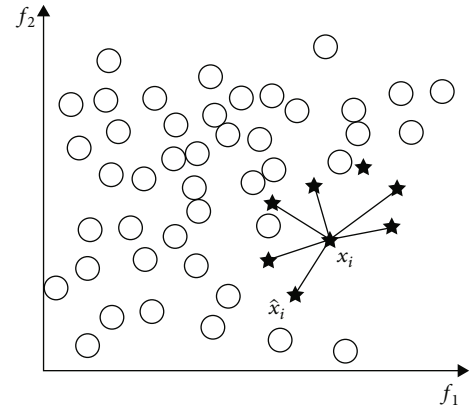

Figure 6: The principle of SMOTE algorithm.

Table 2: Classification results before data balancing.

	Precision	Recall	F1 score	Support
0	0.84	0.74	0.79	510
1	0.65	0.77	0.70	309
Accuracy			0.75	819
Macro avg	0.75	0.76	0.75	819
Weighted avg	0.77	0.75	0.76	819

Table 3: Comparison results after data balancing.

	Precision	Recall	F1 score	Support
0	0.78	0.90	0.84	510
1	0.78	0.59	0.67	309
Accuracy			0.78	819
Macro avg	0.78	0.75	0.76	819
Weighted avg	0.78	0.78	0.78	819

This is an upsampling method, and the main idea of the algorithm is as follows:

(1) Compute all negative class samples x_i to all samples K-nearest neighbors of similar samples

(2) For each minority class samples, the nearest neighbor sample $\hat{x}_i$ is selected

(3) For $\hat{x}_i$ and negative class samples x_i, take random distances and construct new minority class samples. The main idea is shown in Figure 1

The coordinates f_1 and f_2 in Figure 6 represent the sample space, x_i represents the ith minority class sample value selected, and $\hat{x}_i$ represents the nearest neighbor sample selected. The SMOTE algorithm is beneficial to expand the data for better machine learning training but has some drawbacks of its own. During the calculation of the algorithm, the number of nearest neighbor K values needs to be defined manually by the trainer, and the upper limit of

Table 4: Performance analysis of classification models for water quality detection based on common machine learning.

	Model	Accuracy	AUC	Recall	Prec.	F1
0	CatBoost classifier	0.78730	0.8869	0.76710	0.78100	0.77360
1	Light gradient boosting machine	0.77790	0.8789	0.76160	0.76810	0.76450
2	Extreme gradient boosting	0.75570	0.8604	0.76950	0.72940	0.74850
3	Gradient boosting classifier	0.75110	0.8581	0.76240	0.72610	0.74340
4	Random Forest classifier	0.74700	0.8266	0.66780	0.76870	0.71420
5	Extra trees classifier	0.71910	0.7927	0.62560	0.74060	0.67760
6	Decision trees classifier	0.70560	0.7052	0.69720	0.68790	0.69190
7	Ada boosting classifier	0.69840	0.7954	0.69310	0.67860	0.68510
8	Quadratic discriminant classifier	0.64640	0.6998	0.50560	0.66890	0.57500
9	K neighbors classifier	0.63670	0.6791	0.64300	0.61040	0.62560
10	Naive Bayes	0.56400	0.5905	0.38860	0.55850	0.45670
11	Ridge classifier	0.53840	0.0000	0.29090	0.52550	0.37320
12	Logistic regression	0.53800	0.5249	0.29090	0.52470	0.37300
13	Linear discriminant analysis	0.53800	0.5249	0.29090	0.52470	0.37300
14	SVM—linear kernel	0.50490	0.0000	0.35290	0.46800	0.38630

Table 5: CatBoost 10-fold cross-validation.

	Accuracy	AUC	Recall	Prec.	F1
0	0.8158	0.9004	0.7857	0.8182	0.8016
1	0.7895	0.8798	0.8095	0.7612	0.7846
2	0.7669	0.8806	0.7460	0.7581	0.7520
3	0.7970	0.8783	0.7460	0.8103	0.7769
4	0.7331	0.8468	0.7143	0.7200	0.7171
5	0.8120	0.9080	0.8175	0.7923	0.8047
6	0.8075	0.8894	0.8080	0.7891	0.7984
7	0.7774	0.8759	0.7840	0.7538	0.7686
8	0.7396	0.8609	0.7222	0.7280	0.7251
9	0.7698	0.8771	0.7460	0.7642	0.7550
Mean	0.7809	0.8797	0.7679	0.7695	0.7684
SD	0.0275	0.0167	0.0357	0.0310	0.0293

K values can only be trained and tested according to the training value characteristics.

We used Gradient boosting algorithm as an example of water quality detection model to classify the original water quality detection data and got the results as shown in Table 2.

After we use the SMOTE algorithm to upsample the water quality detection data, the task of balancing the data and data expansion of the dataset is completed. Next, the classification results obtained by the water quality detection classification model using the equalized data are shown in the following Table 3.

When we can compare the results, we can see that the data can achieve a high accuracy for both type 0 and type 1 after data balancing, while the model has a large accuracy difference for the classification of the two types of water quality before data balancing.

3.4. Performance Comparison of Different Algorithms. In this paper, we propose to use the AutoML technique and have conducted a related study using the PyCaret third library, and the experimental results are shown in Table 4.

We obtained the results of 15 different machine learning models on balanced training data, and comparing the results, we can find that three models, CatBoost, LGBM, and XGBoost, have the highest accuracy rates. It also further validates the reasonableness of our model selection. In the later chapter, we will optimize the hyperparameters of these three models to obtain the optimal model and use 10-fold cross-validation to demonstrate the applicability and model robustness for this task.

3.5. Fine-Tuning and Optimization of Model Parameters. The performance of a water quality detection classification model will also depend on our choice of hyperparameters.

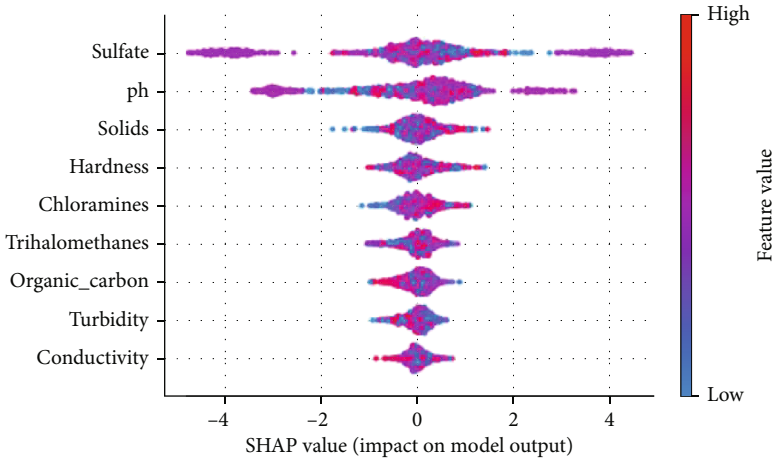

Figure 7: Visualization of important features.

Table 6: LGBM 10-fold cross-validation.

	Accuracy	AUC	Recall	Prec.	F1
0	0.8083	0.8902	0.7937	0.8000	0.7968
1	0.7632	0.8754	0.7778	0.7368	0.7568
2	0.7857	0.8724	0.7778	0.7717	0.7747
3	0.7970	0.8773	0.7381	0.8158	0.7750
4	0.7444	0.8455	0.6984	0.7458	0.7213
5	0.8045	0.8973	0.8095	0.7846	0.7969
6	0.8038	0.8877	0.8160	0.7786	0.7969
7	0.7774	0.8753	0.7600	0.7661	0.7631
8	0.7509	0.8731	0.7143	0.7500	0.7317
9	0.7585	0.8642	0.7460	0.7460	0.7460
Mean	0.7794	0.8758	0.7632	0.7695	0.7659
SD	0.0227	0.0137	0.0372	0.0244	0.0258

However, in general, due to the large search space of hyperparameters, we usually have little time and computational cost to try every possible case, and even with automated hyperparameter tuning methods such as grid search and Bayesian optimization, their practicality is still not strong. Therefore, we combine practical empirical methods for fine-tuning parameters, both in terms of reducing the time spent on finding parameters and in terms of computational cost overhead. In this subsection, we expect to find the optimal hyperparameters of the machine learning-based water quality detection model by fine-tuning and optimizing the parameters of the model.

3.5.1. CatBoost. The results of hyperparameter optimization and 10-fold cross-validation of the CatBoost model are shown in Table 5.

From the table, it can be seen that the average accuracy of the model can reach more than 78%, and the standard deviation of accuracy is small, which has a good classification performance.

Then we analyzed the factors that affect the model output and find out the factors that are more correlated with the output, as shown in Figure 7.

We experimentally found that Sulfate, pH, Solids and Hardness are important influencing factors for conducting water quality tests.

3.5.2. LGBM. Similarly, the results of our hyperparametric optimization and 10-fold cross-validation of the LGBM model are shown in Table 6.

The analysis results show that the accuracy of the model varies became less in ten cross-validations and the final accuracy reaches 78%.

Then we analyzed the factors that affect the output of the model to find out the factors that are more correlated with the output, as shown in Figure 8.

Similarly, we obtained the conclusion that sulfate, pH, solids, and hardness are important influencing factors for conducting water quality tests.

We plot the ROC curve of model classification. The closer the curve is to the upper left corner, the higher the prediction accuracy. The plotted results are shown in Figure 9.

From this, we can find that the area under each curve is larger, indicating that the prediction accuracy is also higher.

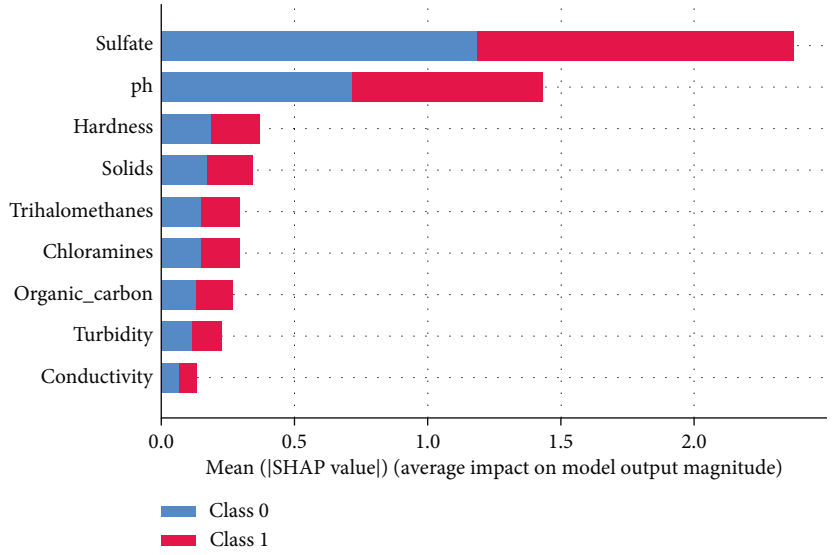

Figure 8: Visualization of important features.

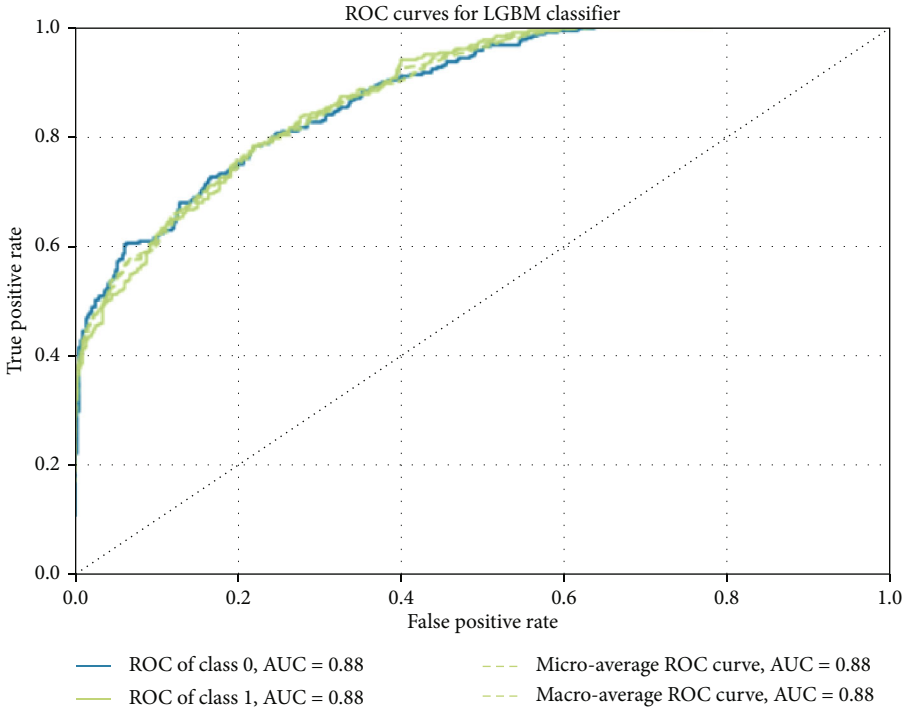

Figure 9: LGBM-ROC curve.

3.5.3. XGBoost. Finally, we perform hyperparameter optimization and 10-fold cross-validation of the XGBoost model, and the results are shown in Table 7.

From the table, it can be seen that the hyperparametric optimized XGBOOST model achieves 3% higher accuracy than the preoptimized model, which verifies the effectiveness of hyperparametric optimization. From the above results, it can be seen that the optimized XGBoost has the highest classification accuracy and the smallest standard deviation of accuracy, and the model has the best stability.

Next, we visualized the important features, and the results are shown in Figure 10.

We obtained that sulfate, pH, solids, and organic_carbon are the important basis for the classification of water quality testing. From the figure, we found that organic_carbon is more similar to hardness. Therefore, in the results of the analysis of the important characteristics of the three types of models, we can conclude that sulfate, pH, solids, and hardness are important influencing factors for water quality testing. Therefore, in the subsequent research work, we

Table 7: XGBoost 10-fold cross-validation.

	Accuracy	AUC	Recall	Prec.	F1
0	0.7970	0.8863	0.8095	0.7727	0.7907
1	0.7970	0.9034	0.7937	0.7812	0.7874
2	0.7707	0.8806	0.6984	0.7928	0.7426
3	0.8045	0.9095	0.8571	0.7606	0.8060
4	0.7707	0.8528	0.8571	0.7559	0.7589
5	0.7932	0.8787	0.7619	0.7886	0.7791
6	0.8113	0.9032	0.7698	0.8016	0.8016
7	0.8151	0.8991	0.8016	0.7984	0.8078
8	0.7887	0.8837	0.8175	0.7851	0.7724
9	0.7774	0.8542	0.7600	0.7538	0.7686
Mean	0.7926	0.8852	0.7840	0.7791	0.7815
SD	0.0150	0.0187		0.0166	0.0202

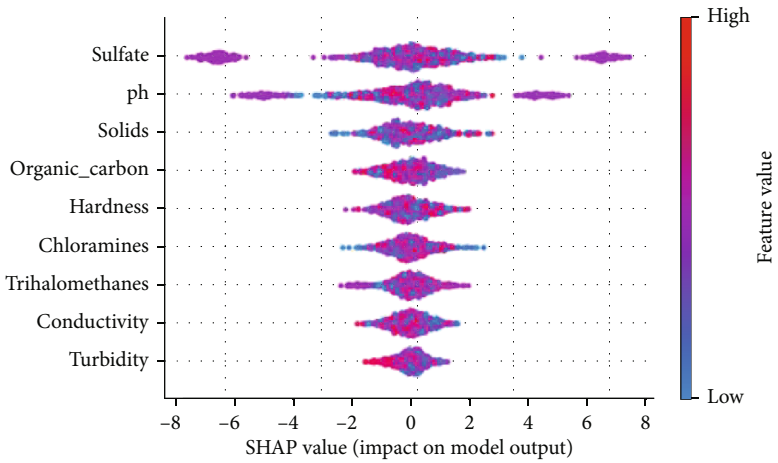

Figure 10: Visualization of important features.

should mainly collect the data of the above features and further explore the depth information of the features.

4. Conclusion

In this paper, the open-source water quality test dataset with the attributes of pH value, hardness, total dissolved solids (TDS), chloramine content, sulfate content, conductivity, organic carbon, trihalomethane (THM), turbidity, and drinkability is studied. 15 mainstream machine learning water quality detection algorithms based on XGBoost, CatBoost, RF, Naive Bayes, and LGBM are designed, respectively, and the real values are compared with the predicted values of various models from the four evaluation indexes of accuracy, recall, F1 value, and accuracy. The experimental results show that sulfite, pH, solids, and hardness are important influencing factors for water quality detection. And XGBoost, CatBoost, and LGBM have good performance in water quality detection. Finally, we further optimize the classification model of water quality detection based on XGBoost, CatBoost, and LGBM by means of cross-validation and super parameter optimization. The widely used models have proved that XGBoost, CatBoost, and LGBM models have good treatment effects on water quality indicators and can be applied to water quality detection on a large scale.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] A. M. El-Sharkawy, "$^{234}U/^{238}U$ activity ratios in groundwaters from two aquifers in Saudi Arabia, and correlation with water chemistry," *Journal of Radiation Research and Applied Sciences*, vol. 11, pp. 368–372, 2018.

[2] M. Mamun and A. Kwang-Guk, "Application of multivariate statistical techniques and water quality index for the assessment of water quality and apportionment of pollution sources in the Yeongsan River, South Korea," *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, p. 8268, 2021.

[3] P. Liu and J. Wang, "Analysis and prediction of water quality using LSTM deep neural networks in IoT environment," *Sustainability*, vol. 43, no. 2, pp. 294–299, 2021.

[4] Z. Rongjie, *Research on Spatio-Temporal Prediction and Pollution Warning Model of River Water Quality Based on Improved BP Neural Network*, Harbin Institute of technology, 2021.

[5] L. Ruiping, *Research on Water Turbidity Detection Based on Image Recognition*, Anhui University of technology, 2020.

[6] K. Zhong, *Fault Detection and Diagnosis Based on Multivariate Statistical Analysis*, Dalian University of Technology, 2020.

[7] O. Fallatah, M. Ahmed, B. Gyawali, and A. Alhawsawi, "Factors controlling groundwater radioactivity in arid environments: an automated machine learning approach," *Science of The Total Environment*, vol. 830, article 154707, 2022.

[8] M. Varol, B. Gökot, A. Bekleyen, and B. Şen, "Water quality assessment and apportionment of pollution sources of Tigris River (Turkey) using multivariate statistical techniques—a case study," *River Research and Applications*, vol. 28, no. 9, pp. 1428–1438, 2012.

[9] Y. Yu, M. Shao, L. Jiang et al., "Quantitative analysis of multiple components based on support vector machine (SVM)," *Optik*, vol. 237, p. 166759, 2021.

[10] X. Fang, X. Li, Y. Zhang et al., "Random forest-based understanding and predicting of the impacts of anthropogenic nutrient inputs on the water quality of a tropical lagoon," *Environmental Research Letters*, vol. 16, no. 5, 2021.

[11] M. Al-Mukhtar, "Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad," *Environmental monitoring and assessment*, vol. 191, no. 11, p. 673, 2019.

[12] S. Ye, X. Chen, D. Dong, J. Wang, X. Wang, and F. Wang, "Rapid determination of water COD using laser-induced breakdown spectroscopy coupled with partial least-squares and random forest," *Analytical Methods*, vol. 10, no. 40, pp. 4879–4885, 2018.

[13] Y. Liu and H. Wu, "Water bloom warning model based on random forest," in *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Okinawa, Japan, 2017.

[14] K. Joslyn and J. Lipor, "A supervised learning approach to water quality parameter prediction and fault detection," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2511–2514, Seattle, WA, USA, 2018.

[15] J. H. Lee, J. Y. Lee, M. H. Lee et al., "Development of a short-term water quality prediction model for urban rivers using real-time water quality data," *Water Supply*, vol. 22, no. 4, pp. 4082–4097, 2022.

[16] D. V. Prasad, P. S. Kumar, L. Y. Venkataramana et al., "Automating water quality analysis using ML and auto ML techniques," *Environmental Research*, vol. 202, article 111720, 2021.

[17] Y. He, "Research on urban road traffic flow prediction based on LGBM model," *Electronic Technology and Software Engineering*, vol. 3, pp. 259–262, 2022.

[18] W. P. Jiang, Z. C. Jiang, and Z. X. Dong, "XGBoost-based death risk assessment model and application of heart failure," *Modern Electronics Technique*, vol. 8, pp. 155–158, 2022.

[19] L. Minh Dang, S. J. Kyeong, Y. Li, H. W. T. N. Nguyen, and H. Moon, "Deep learning-based sewer defect classification for highly imbalanced dataset," *Computers & Industrial Engineering*, vol. 161, p. 107630, 2021.

[20] B. Asvija, R. Eswari, and M. B. Bijoy, "Security threat modelling with Bayesian networks and sensitivity analysis for IAAS virtualization stack," *Journal of Organizational and End User Computing (JOEUC)*, vol. 33, no. 4, pp. 44–69, 2021.

[21] F. J. Martínez-López, Y. Li, C. Feng, and D. López-López, "Buying through social platforms: perceived risks and trust," *Journal of Organizational and End User Computing (JOEUC)*, vol. 33, no. 4, pp. 70–93, 2021.

[22] B. Hewitt and G. White, "Factors influencing security incidents on personal computing devices," *Journal of Organizational and End User Computing (JOEUC)*, vol. 33, no. 4, pp. 185–208, 2021.