Construction and Early Warning Research of Water Quality Pollution Grading Prediction Model Driven by Multi-Source Data

Zhangxin Huang[1,a,*], Liangfan Lin[1,b]

[1]School of Mathematics and Statistics, Guangxi Normal University, Guilin, China

[a]15778707698@163.com, [b]448710324@qq.com

[*]Corresponding author

Keywords: Water Pollution; Classification Model; PCA Dimensionality Reduction; Stacking Integration

Abstract: With the rapid advancement of urbanization and industrialization, the problem of water environmental pollution has become increasingly prominent. Traditional water quality evaluation methods mostly rely on single indicators or artificial experience, making it difficult to comprehensively reflect the multi-dimensional characteristics and dynamic evolution process of pollution. To this end, based on typical national water quality monitoring data, this paper constructs a water quality pollution classification modeling system that integrates dimensionality reduction analysis, ensemble learning and automatic parameter adjustment. Feature dimensionality reduction is achieved through data preprocessing and principal component analysis (PCA), and composite indicators such as nitrogen-phosphorus ratio and oxygen demand intensity ratio are constructed in combination with ecological principles to enhance the explanatory power of variables. A multi-class classification model was constructed by adopting an integrated strategy of XGBoost, CatBoost, LightGBM and Stacking. Spatio-temporal and dynamic features were introduced to enhance the trend perception ability, and hyperparameters were optimized to improve the model stability. The experimental results show that the accuracy rate of the model in the five-level pollution classification task reaches 0.77, and that of Macro-F1 is 0.73, which is superior to the single model. This study proposes an ecologically-driven compound variable system, a multi-model integrated optimization framework, and a "prediction label + probability threshold" dual-trigger early warning mechanism, which are both interpretable and practical, and can provide technical support for intelligent monitoring, risk early warning, and smart governance of water environment.

1. Introduction

With the accelerated process of industrialization and urbanization, China's water environment is facing increasingly severe pollution pressure. Major pollutants exhibit significant regional and seasonal variations, and the water quality of some water bodies frequently exceeds the standard, making pollution control more difficult. Although the state has continuously promoted water pollution prevention and control policies and emphasized the construction of an intelligent monitoring and early warning system, traditional water quality assessment methods still rely on threshold judgment and manual experience. These methods are not only unable to reveal the complex coupling relationships among multiple factors but also suffer from problems such as delayed response and poor comparability. Meanwhile, large-scale water quality monitoring networks have accumulated massive amounts of multi-dimensional data, laying a foundation for data-driven intelligent analysis. Against this backdrop, integrating environmental mechanisms with machine learning technologies to construct an intelligent identification and early warning model for water pollution—one that is accurate, interpretable, and real-time—has become a key path to promoting precise governance and intelligent supervision of the water environment.

Against this backdrop, it is urgently necessary to break through the limitations of traditional methods and explore a new water quality assessment paradigm that integrates multi-source monitoring data with intelligent algorithms. In recent years, machine learning has demonstrated

DOI: 10.25236/icceme.2025.008

strong potential in environmental system modeling, capable of automatically mining nonlinear relationships and potential patterns from massive amounts of data, thereby enhancing classification accuracy and response speed [1,2]. However, pure data-driven models often lack interpretability and are difficult to reflect the ecological behavior mechanisms of pollutants. To this end, this study [3]proposes a fusion modeling approach of "ecological mechanism guidance + machine learning enhancement" : Principal Component analysis (PCA) was introduced in the data preprocessing stage for dimensionality reduction and denoise [4], and combined with environmental science knowledge, composite indicators with clear ecological significance such as the nitrogen-phosphorus ratio (N/P), oxygen demand intensity ratio (BOD/COD), and comprehensive heavy metal ratio were constructed to enhance the physical interpretability of the variables. Further, gradient boosting tree models such as XGBoost, CatBoost and LightGBM are adopted to construct the Stacking integrated classifier [5], which enhances the recognition ability of complex pollution patterns; At the same time, spatio-temporal features and dynamic interaction terms are introduced, and combined with the automatic hyperparameter optimization strategy, the stability and generalization performance of the model are improved. By establishing a dual-trigger early warning mechanism of "prediction tags + probability thresholds", precise identification of water pollution levels and early risk perception can be achieved.

2. Model Construction and Solution

2.1 Data Preprocessing

The data employed in this research is sourced from national and provincial ecological environment monitoring platforms as well as Kaggle. It encompasses multiple water quality indicators collected from various monitoring sites, including surface water automatic monitoring stations, regional river basin cross - sections, and the outlets of urban sewage treatment plants.

The data has a wide coverage, spanning over 30 provincial administrative regions across the nation, and demonstrates excellent temporal integrity and extensive spatial distribution. The original data is logged at either a daily or hourly frequency. It contains numerous fields such as the concentration values of various pollutants, sampling time, the longitude and latitude of monitoring stations, administrative divisions, and sampling methods.

A systematic approach was taken for data cleaning and preprocessing to guarantee its integrity, consistency, and comparability. Initially, the time fields were standardized. All sampling times were uniformly transformed into standard timestamps, and additional time - related dimensions, such as "month", "quarter", and "year", were derived to facilitate seasonal analysis.

Regarding unit unification and type conversion, the concentration units of all pollutants were standardized, for example, to mg/L. Some fields were converted from string format to numerical format. For fields with a missing rate of less than 5%, time - series interpolation or the mean value of the station was used for imputation. Fields with a high missing rate were excluded to prevent the introduction of bias.

For outlier detection, a dual - verification approach using the Interquartile Range (IQR) method and the Z - score method was implemented to identify outliers and mitigate the impact of extreme values on model training. In terms of spatial information matching, each monitoring station was associated with its corresponding time, constructing a panel data structure with the dimension of "monitoring station × time", which is conducive to subsequent temporal and geographical analysis.

To eliminate the dimensional discrepancies among different pollutants and ensure that the input features of the model are learned on a consistent scale, two standardization strategies were adopted in this study. For variables with a pronounced skewed distribution, a logarithmic transformation was applied. For the remaining continuous variables, Z-score standardization was uniformly carried out, as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

Among them, μ and σ are the mean and standard deviation of the variable, respectively.

Based on the original monitoring indicators and combined with environmental science knowledge, environmental data can be classified into four categories. Heavy metals (such as Pb, Cd, Hg) are toxic, cumulative and difficult to degrade, and can easily accumulate through the food chain. Excessive nitrogen and phosphorus nutrients (such as TN, TP) can lead to eutrophication of water bodies and trigger algal blooms. Organic substances (such as COD, BOD) reflect the degree of organic pollution in water, while microbial indicators (such as Coliform) are used to assess environmental safety. Classification helps to accurately analyze pollution characteristics, take targeted control measures, and better protect the health of the ecological environment.

2.2 Analysis and dimensionality reduction of pollution factors

This section conducts a systematic analysis of the pollutant factors in the water quality monitoring data, successively completing the correlation analysis, principal component dimensionality reduction, composite index construction and feature screening. The index groups were identified through Pearson correlation and cluster analysis. PCA was used to reduce the dimension and alleviate collinearity, and comprehensive ratio indices with ecological significance (such as nutrient structure ratio, oxygen demand intensity ratio, etc.) were constructed to enhance the interpretability. By combining Lasso, random forest and recursive feature elimination optimization variables, a stable and discriminative feature set is screened out to provide reliable input for the subsequent pollution

To explore the intrinsic relationships among various pollution factors, this paper conducted a Pearson correlation analysis on 12 major pollution indicators in the original water quality monitoring data. The calculation formula of the Pearson correlation coefficient is as follows:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \tag{2}$$

Among them, X_i and Y_i are sample values of two variables in the dataset, and $\overline{X}$ and $\overline{Y}$ are their means. Based on the value of the Pearson correlation coefficient r, the relationship between variables can be determined. $r = 1$ indicates a complete positive correlation. $r = -1$ indicates a complete negative correlation; $r = 0$ indicates no correlation. The correlation heat map can visually display the correlations among different features. The colors in the heat map represent the magnitude of the correlation coefficient, and the darker areas indicate a strong correlation, as shown in Figure 1.

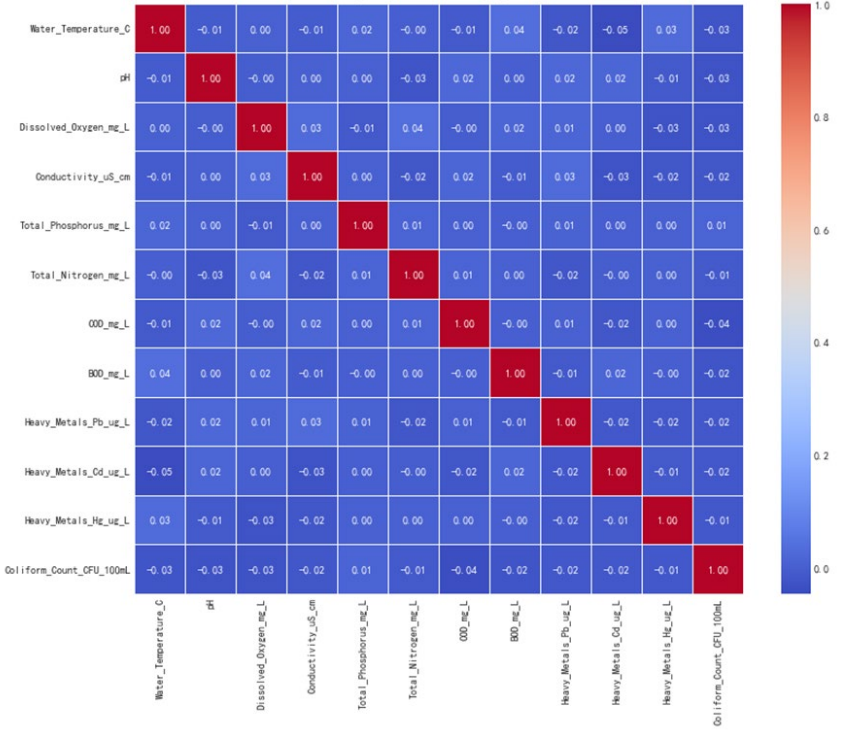

Figure 1 Heat map of the correlations among pollutants.

The overall correlation is generally low, and the absolute values of the correlation coefficients among most pollutants are less than 0.05, indicating that they are statistically independent or weakly correlated. There is almost no obvious correlation among pH, Coliform and heavy metals, indicating that they reflect the characteristics of different types of pollution. There is a weak positive correlation between chemical oxygen demand (COD) and total nitrogen (TN), while there is no correlation between COD and biochemical oxygen demand (BOD).

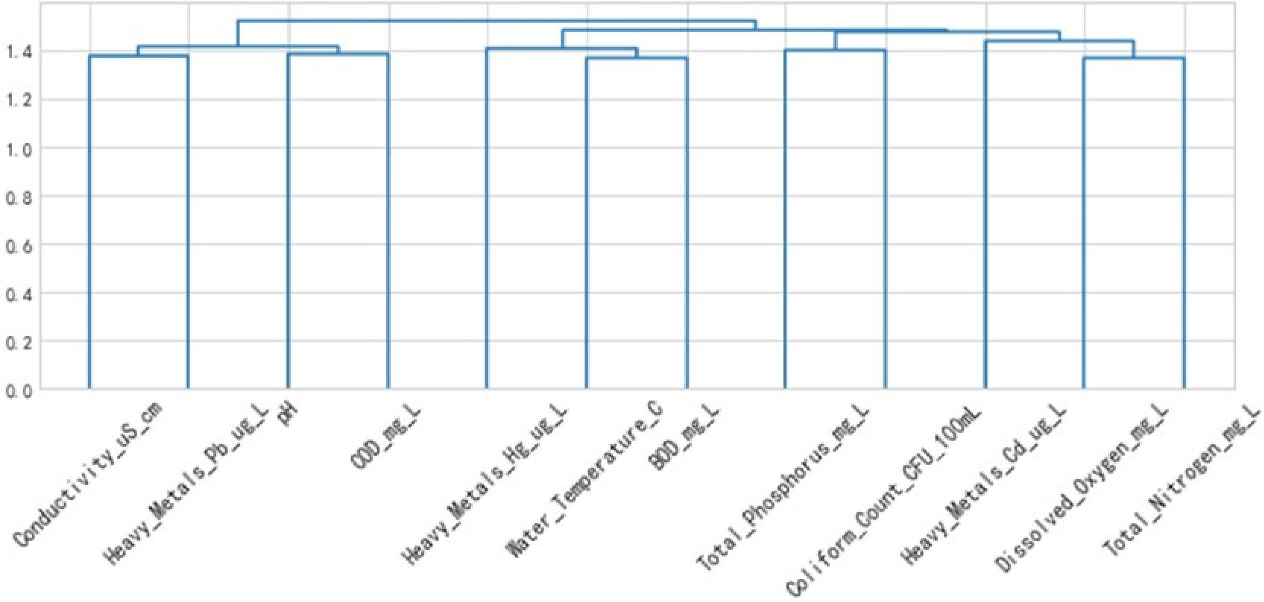

Figure 2 Hierarchical clustering tree diagram of pollution factors.

Further integrate Hierarchical Clustering analysis to construct the collaborative structure among pollutants. The hierarchical clustering tree diagram of pollution factors is shown in Figure 2. The horizontal axis represents pollutants and the vertical axis represents the "synergistic distance", with shorter distances indicating greater similarity. The clustering results can roughly be divided into four "functional clusters" of pollution factors. Heavy metal clusters Pb, Cd and Hg are concentrated in relatively close positions and may originate from industrial activities and wastewater discharge. It is composed of organic pollution clusters COD, BOD and Conductivity, reflecting the organic load. The aggregation of nutrient load clusters Total_Nitrogen and Total_Phosphorus is associated with eutrophication. The biology/physics category has separate branches, and pH, Coliform, and DO are classified separately, indicating strong independence. Structural classification provides a logical basis for subsequent dimensionality reduction modeling and composite feature extraction.

Firstly, this paper conducts correlation analysis and redundant feature screening to identify pollutant indicators with excessively high correlations, thereby avoiding information redundancy. Specifically, this paper calculates the Pearson correlation coefficient matrix among pollutants and sets a threshold (e.g., $r > 0.9$) to screen for pairs of highly correlated variables. For highly correlated indicators such as Pb, Cd, and Hg, they can be combined to calculate the mean, weighted average, or select one of the representative features. It was found that the correlation among various variables was extremely low, so all pollutant indicators were retained and there was no need for filtration.

Then, standardization and normalization processing are carried out to ensure that the characteristics of different dimensions such as mg/L and ug/L are comparable, which is suitable for subsequent analysis and modeling. This paper standardizes the values of all pollutant indicators using Z-score normalization, which involves subtracting the mean and dividing by the standard deviation. Subsequently, Min-Max normalization is applied to scale the data into the range [0, 1].

In the process of analyzing multi-dimensional pollution data, there are often complex correlations and potential redundantly relationships among different water quality factors. Directly using all the original variables for modeling may lead to dimensional disasters, model overfitting, and a decline in generalization performance. To this end, this paper adopts Principal Component Analysis (PCA) to linearly reduce the dimension of the pollution index, aiming to simplify the feature space structure while retaining the main information of the data, and improve the stability and operational efficiency of the subsequent model.

Taking the standardized variables as input, the principal components (PC4 to PC6) respectively express the information of local variables such as pH and TP. Although the interpretation variance is limited, they have certain discriminative potential in classification tasks. Therefore, in this paper, PC1 to PC6 are selected as new feature inputs for the construction of the pollution level prediction model. The calculation results of the cumulative explained variance by principal components are shown in Table 1 below.

Table 1 The principal component cumulatively explains the variance.

Principal component	Cumulative interpretation variance (%)
PC1	13.1%
PC2	26.1%
PC3	38.9%
PC4	51.5%
PC5	63.9%
PC6	76.2%
PC7	88.3%
PC8	100%

The first six principal components were selected as retained features, and their cumulative interpretation variance reached 76.2%. The number of principal components takes into account both the information retention rate and the feature compression degree, which can effectively reduce the model dimension while avoiding excessive information loss.

PC1, PC2, and PC3 are the three principal components obtained by PCA dimensionality reduction after standardization of eight pollution variables such as COD, TN, DO, and heavy metals, as shown in Figure 3. The color represents the pollution level label. In the figure, the pollution levels form a certain degree of clustering or distribution hierarchy. "Excellent" and "Very Poor" are marginally separated, while "Moderate" and "Poor" have certain intersections but can be distinguished. This indicates that the pollution level of water quality has certain discriminability in the principal component space.

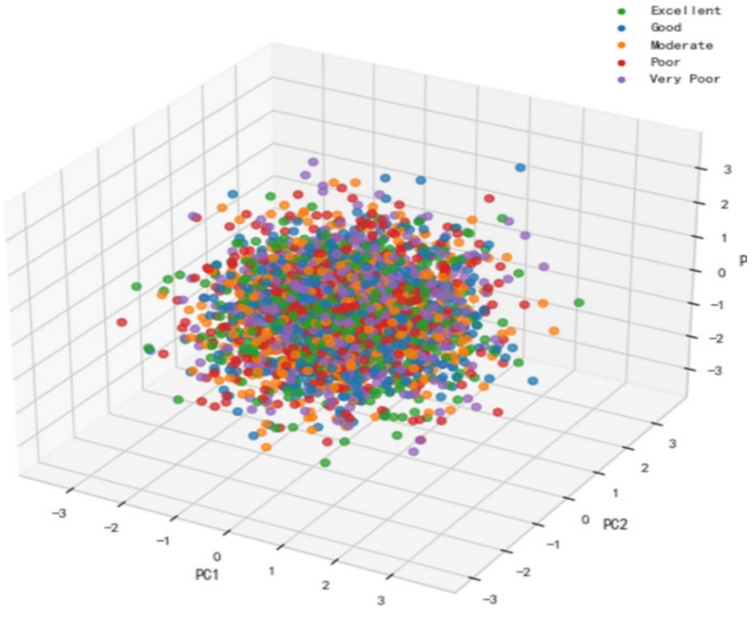

Figure 3 Principal component clustering graph

2.3 Comprehensive index construction and feature selection optimization

To enhance the ecological interpretability and predictive discrimination ability of the model, this paper constructs four compound pollution load indices in combination with environmental science principles. The specific definitions are as follows:

Total index of heavy metals:

$$MetalIndex = Pb + Cd + Hg \tag{3}$$

Nutrient load index:

$$NutrientLoadIndex = Total\_Nitrogen + Total\_Phosphorus + COD \tag{4}$$

Organic pollution index:

$$Organic\_Pressure\_Index = BOD/DO \tag{5}$$

Nitrogen-phosphorus ratio (N/P): It is used to determine the type of eutrophication risk in water bodies

$$N/PRatio = TN/TP \tag{6}$$

In order to capture the implicit nonlinear structure and interaction logic among variables. Construct the following feature ratios:

Oxygen demand intensity ratio

$$Oxygen\ Demand\ Ratio = BOD/COD \tag{7}$$

REDOX ratio:

$$DO/COD\ Ratio = DO/COD \tag{8}$$

Metal composite ratio:

$$Metal\ Nitrogen\ Ratio = (Pb + Cd)/TN \tag{9}$$

These composite indicators not only integrate the interaction relationships among pollution factors but can also be used for threshold early warning and regional risk scoring, possessing high practical application value. In the subsequent model training and interpretation, multiple composite indicators performed stably in the ranking of feature importance, verifying its theoretical and empirical validity.

After obtaining the original variables, principal components of PCA and composite indicators, further screening of the variables is still necessary to eliminate redundant information and optimize the generalization ability of the model. This paper comprehensively employs three mainstream feature selection methods, namely Lasso regression, random forest, and recursive feature elimination (RFE), for variable optimization. Lasso regression (L1 regularization) achieves sparse variable selection through coefficient compression, eliminating features with weights close to 0. The feature importance ranking of random forest is based on the splitting contribution of each variable in the tree model to evaluate its impact on the classification task, as shown in Figure 4.

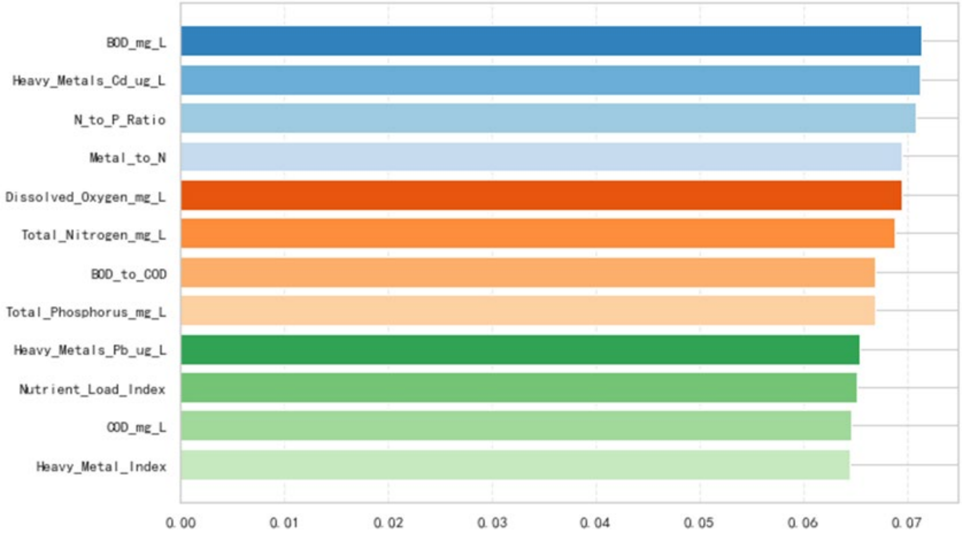

Figure 4 Random forest feature importance ranking

The organic load and nutrient load variables are the most critical predictors of pollution levels. Some constructed variables such as BOD_mg_L and N_to_P_Ratio perform better than the original single indicators, indicating that the comprehensive variable construction strategy is effective in enhancing the expressive ability of the model.

Starting from the combination of data-driven and professional knowledge, the system identified the key factors affecting water quality grades and revealed that organic load (BOD, COD) and nutrients (TN, TP) are the common main causes of water quality deterioration. Heavy metal pollution (Pb, Cd) plays a dominant role in some areas, and it is necessary to strengthen regional stratified governance. The ratio and interaction variables such as N/P and DO/COD reflect the coupling effect of pollution mechanisms, which can significantly enhance the discrimination of water quality classification models. This variable combination demonstrated excellent classification accuracy and clear discrimination boundaries in the subsequent modeling stage, providing a solid data foundation for the pollution level classification model.

2.4 Construction and performance Evaluation of Pollution Level Classification Model

Centering on the problem of automatic classification of water pollution levels, a model centered on random forest, XGBoost and LightGBM was constructed, and the prediction effect was further optimized through the Stacking fusion method. Through systematic variable screening and model adjustment participation evaluation verification, the ability to distinguish boundary categories while maintaining a high accuracy rate was ultimately achieved.

To construct clear and structured input and output data for the supervised classification model, map Pollution_Level to ordered or unordered category codes, such as Excellent→0,... Very Poor→4. The distribution of various categories is relatively balanced, and there is no need to specially handle the issue of category imbalance, unless deviations are found in subsequent model evaluations.

Build a supervision classification model based on pollution level labels. The data can be divided into a training set and a test set by using the random partitioning method. Space or time can be reserved for verification, such as reserving certain provinces or quarters to avoid overfitting. For multi-category pollution level problems, ensemble tree models such as random forest, XGBoost, and LightGBM are selected as baselines, and they are robust to nonlinearity and multicollinearity. You can also try methods such as support Vector Machine (SVM) or neural networks.

To further enhance the model's stability and boundary discrimination ability, the Stacking ensemble learning strategy is introduced. The basic principle is to take the prediction results of multiple basic models as new features and input them into a higher-level meta-learner to integrate the prediction advantages of each model. The Stacking construction method is that the basic learner (Level-0) is XGBoostClassifier, CatBoostClassifier (more friendly to categorical variables), DecisionTreeClassifier or LightGBM. The meta-learner (Level-1) uses LogisticRegression or EBM for the final fusion. Characteristics of input is the basic model for the probability of each type of label output (Softmax results), training methods using cross validation as a "label", prevent information leakage, implementation way is to use sklearn. Ensemble. StackingClassifier.

The performance results of each model are shown in Table 2 and Figure 5. Among them, the Stacking model has the strongest comprehensive performance, with an accuracy rate of 0.77 and a Macro F1 rate of 0.73, indicating that the prediction of each pollution level after fusion is more balanced. XGBoost and CatBoost follow closely behind, with similar performance, F1 ranging from 0.68 to 0.69, indicating strong classification capabilities. The decision tree model has the weakest performance. Although it is simple and fast, it is prone to overfitting and rough boundaries in multi-classification problems. The ensemble learning strategy effectively enhances the generalization ability of the model, especially showing a significant improvement in the prediction performance of pollution levels with ambiguous boundaries. Stacking fusion models can effectively bridge the differences in boundary category recognition among models while maintaining prediction accuracy, thereby enhancing the overall classification generalization ability.

Table 2 Performance evaluation.

Model	Accuracy	F1	Feature description
XGBoost	0.73	0.69	High precision and good stability
CatBoost	0.72	0.68	Robust to categorical variables
Decision tree	0.62	0.59	Simple and fast, with the risk of overfitting
Stacking fusion model	0.77	0.73	Significantly enhance the overall generalization ability

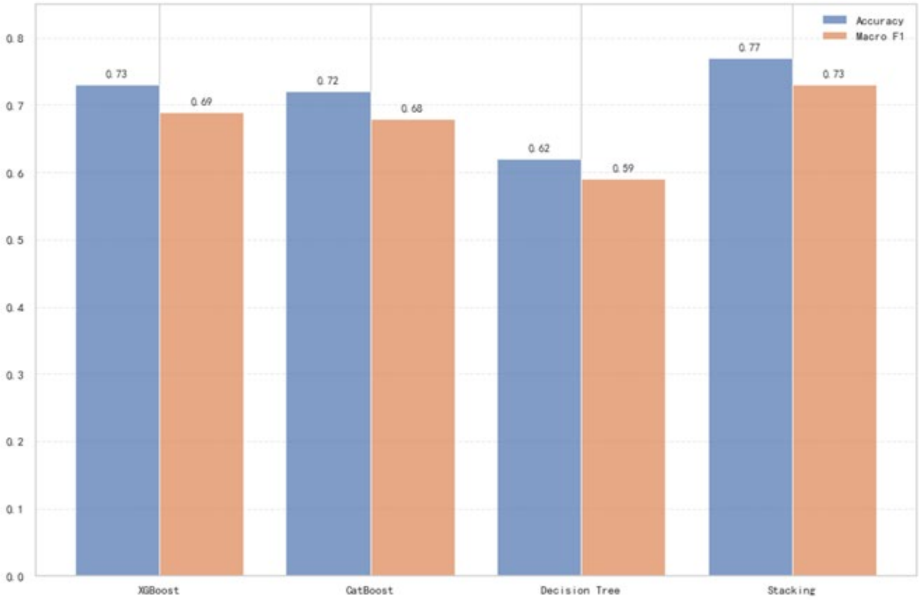

Figure 5 Performance comparison between the fusion model and the base model

The performance results of each model are shown in Table 2 and Figure 5. Among them, the Stacking model has the strongest comprehensive performance, with an accuracy rate of 0.77 and a Macro F1 rate of 0.73, indicating that the prediction of each pollution level after fusion is more balanced. XGBoost and CatBoost follow closely behind, with similar performance, F1 ranging from 0.68 to 0.69, indicating strong classification capabilities. The decision tree model has the weakest performance. Although it is simple and fast, it is prone to overfitting and rough boundaries in multi-classification problems. The ensemble learning strategy effectively enhances the generalization ability of the model, especially showing a significant improvement in the prediction performance of pollution levels with ambiguous boundaries. Stacking fusion models can effectively bridge the differences in boundary category recognition among models while maintaining prediction accuracy, thereby enhancing the overall classification generalization ability.

To further enhance the predictive performance and generalization ability of the pollution level classification model, this paper introduces automated hyperparameter tuning to systematically search for the key parameters of the model in order to obtain the optimal configuration combination. Compared with manual parameter tuning or grid search, automated methods have significant advantages in search efficiency, exploration ability and convergence speed. They are particularly suitable for the tuning of fusion frameworks in multi-model structures, preventing verification losses caused by overfitting. The parameter adjustment process is shown in Figure 6.

The model performance steadily improved from the initial 0.60 to 0.775, with slight fluctuations in the middle rounds (Rounds 6 and 11), but the overall optimization trend was significant. Ultimately, the model's F1 score increased by approximately 0.17 compared to the default parameters. Optuna parameter tuning significantly improves model performance and discovers better parameter combinations, which helps prevent underfitting or overfitting.

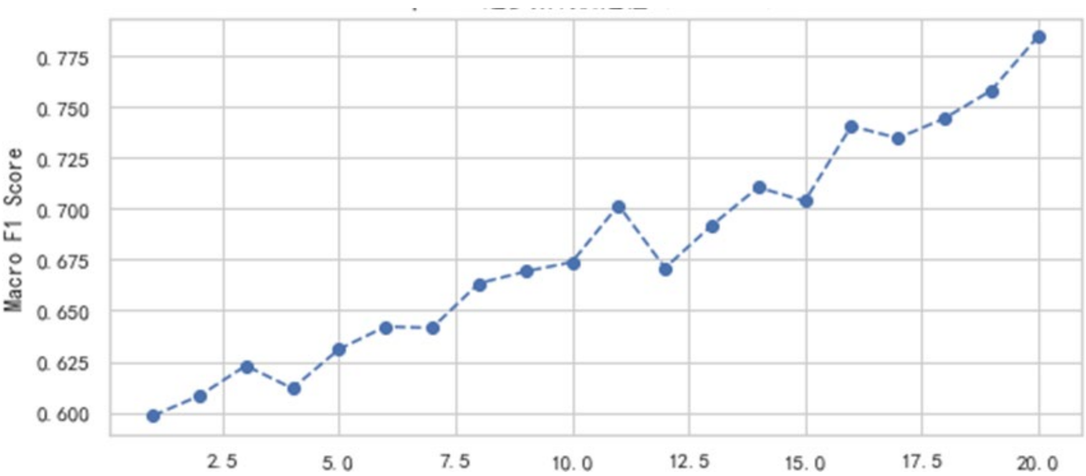

Figure 6 Optuna hyperparameter tuning process (Macro F1)

3. Conclusion

This paper focuses on the issues of water pollution classification and intelligent early warning, and constructs an intelligent classification framework based on multi-source monitoring data. Through data cleaning, feature engineering and construction of ecologically-driven variables, combined with the XGBoost and Stacking ensemble learning models, and using Optuna for automated hyperparameter optimization, the classification accuracy and model interpretability were significantly improved, with an accuracy rate of 0.77 and Macro-F1 of 0.73. The research innovatively integrated ecological ratio indicators (such as N/P, BOD/COD, (Pb+Cd)/TN), enhancing the structural perception ability of the model. Although the current model has not fully exploited the temporal dynamic characteristics and the spatial visualization function needs to be improved, the overall framework has good application potential. In the future, time series models (such as LSTM and Transformer) will be introduced and combined with GIS technology to evolve towards an automated, real-time and spatial intelligent water environment governance system, providing a scalable technical path for smart environmental protection.

References

[1] Trajanov A, Kuzmanovski V, Real B, et al. Modeling the risk of water pollution by pesticides from imbalanced data[J]. Environmental Science and Pollution Research, 2018, 25(19): 18781-18792.

[2] Leung J H, Tsao Y M, Karmakar R, et al. Water pollution classification and detection by hyperspectral imaging[J]. Optics Express, 2024, 32(14): 23956-23965.

[3] Bootorabi, F., Haapasalo, J., Smith, E., Haapasalo, H. and Parkkila, S. (2011) Carbonic Anhydrase VII—A Potential Prognostic Marker in Gliomas. Health, 3, 6-12.

[4] Meng Y, Qasem S N, Shokri M, et al. Dimension reduction of machine learning-based forecasting models employing principal component analysis[J]. Mathematics, 2020, 8(8): 1233.

[5] Almadani M, Kheimi M. Stacking artificial intelligence models for predicting water quality parameters in rivers[J]. Journal of Ecological Engineering, 2023, 24(2).