Original Research Article

Application of stacking-based ensemble learning model for water quality prediction

ABSTRACT

Water is the source of life, and the growth of animals and plants cannot leave the water source. The quality of water will directly affect the life and health of humans, animals and plants. In order to predict the concentration and changing trend of various pollutants in water bodies and promote the comprehensive management of water resources, this paper proposes a new integrated model based on the idea of Stacking integrated learning. The model is based on XGBoost, support vector regression, and multi-layer perceptron. The model is constructed with ridge regression as the meta-model. The model was applied to the pH and total nitrogen content of water quality, and the mean absolute percentage error was used to quantitatively evaluate the prediction results, and the results of the ensemble learning model were compared with the prediction results of a single base model. The results show that the stacking ensemble learning idea can effectively improve the prediction ability and generalization performance of the base model. The proposed ensemble learning model has very good prediction ability and generalization ability, and has great potential in other prediction fields such as water quality.

Keywords: Ensemble learning, water quality prediction, multi-step prediction, machine learning

1. INTRODUCTION

1.1 BACKGROUD

Water is the source of life and an essential resource for human survival. The origin and development of human civilization depend on the supply and nourishment of different water systems. However, due to the gradual intensification of human production and exponential population growth in recent years, many watersheds' ecological and environmental quality has been declining year by year, and the carrying capacity is heavy. As a result of environmental pollution, some watersheds' ecological resources and regulating functions have been massively damaged[1], affecting the healthy, sustainable, and coordinated development of the watershed areas. In recent years, water pollution incidents have been reported around the world[2], and it is evident that many water pollution incidents have occurred in the absence of information management. Water quality prediction is an essential task in water environment management; accurate water quality prediction can provide a basis for decision-makers, and offer suggestions to environmental management departments to prevent it before it happens, so the precise prediction value will effectively improve water resources management[3].

Water is an essential and vital resource for human life, and quality is a fundamental issue because it is directly related to human health [4]. Consumption of contaminated water can lead to health problems and increased morbidity and mortality. With the advancement of technology and industry, most industrial production activities pose a serious threat to

groundwater quality. While driving the accelerated technological development of civilization, the energy generated by crude oil and natural gas has also led to severe pollution of water bodies by oil derivatives, among others [5].

Monitoring and predicting the concentrations and trends of various pollutants in water bodies can promote decision-makers in the rational development, optimal allocation, comprehensive conservation, efficient use, adequate protection, and comprehensive management of water resources. They can guarantee the safety of the water supply in drinking water source areas.

1.2 RELATED WORK

Water quality prediction research has been a hot research area. With the development of artificial intelligence and big data and other fields in recent years, more and more scholars have introduced machine learning, deep learning, and other models into water quality prediction and achieved satisfactory results.

Rahim Barzegar et al. applied wavelets and extreme learning machines to the prediction of electrical conductivity (EC) of water[6]; Sakshi Khullar and Nanhey Singh used BiLSTM to predict water quality factors in Yamuna River, India [7], and the final chemical oxygen demand (COD)) analysis showed MAPE values of and 20.32% and biochemical oxygen demand (BOD) study showed MAPE values of 18.22% for the Pala region, respectively; Lu Hongfang and Ma Xin proposed two-hybrid decision tree-based water quality prediction models in 2020 and studied the most polluted Tuaratim River in Oregon, USA as an example [3]; Abubakr Saeed Abobakr Yahya et al. used support vector machine to predict the water quality of Leng Yue River in Malaysia [8]; Rahim Barzegar et al. proposed a hybrid CNN-LSTM model and applied it to short-term river water quality prediction [9]; AliNajah Ahmed et al. used wavelet denoising technique and adaptive fuzzy neural network to effectively improve the accuracy of water quality prediction [10]. Moges B. Wagena et al. used a process-based model (SWAT-VSA), stochastic model (artificial neural network-ANN), and autoregressive moving average (ARMA) model to evaluate flow prediction. They formed an integrated Bayesian model using the results of SWAT-VSA, ANN, and ARMA [11]; Jiayue Gu et al. proposed an integrated learning model with KNN, XGB, SVR, and ANN as the base models. They used this model to predict monthly rainfall and finally obtained the prediction results with high accuracy[12]; Longfeng Zhang et al. proposed a model combining the Harris Hawk optimization algorithm with the XGBoost model and predicted the short-term traffic flow, and the results showed that the accuracy and stability of the model were more significantly improved[13].

According to the results of the literature review, it can see that machine learning and deep learning models are widely used in the field of water quality prediction because of their powerful predictive ability and good stability. However, it is worth mentioning that most scholars are committed to using a single prediction model, but this approach depends heavily on the model's merit. Therefore, in this study, we are committed to proposing an integrated learning model based on the Stacking method, which combines the advantages of multiple models to improve the prediction accuracy and model stability effectively.

1.3 MOTIVATION & ARTICLE STRUCTURE

Scientific experiments have shown that changes in water pH can affect the ability of algae to take in oxygen and the sensitivity of animals to food intake and that unacceptable pH levels can disrupt the supply of phosphate and inorganic nitrogen compounds in the water column, thus seriously affecting the biological productivity of the water column. In addition, too high or too low PH can have a massive impact on the human gut. In addition to PH, total nitrogen content (TN) is also an essential criterion for water quality. Excessive water intake or food with high nitrate-nitrogen can cause digestive tract cancer or liver cancer. High levels of the

three nitrogen (ammonia nitrogen, nitrate-nitrogen, and nitrate-nitrogen) in water source water and drinking water can be toxic to both humans and aquatic organisms in water bodies. For example, ammonia nitrogen in the water of more than 1mg/L will reduce the ability of marine organisms to combine oxygen in the blood with more than 3mg/L, and fish will die. The presence of ammonia and nitrogen in the water source will cause a significant increase in chlorination in water treatment. High ammonia nitrogen will lead to the rise in other disinfection by-products, which can be harmful to human health.

Therefore, this study will analyze and predict the PH and TN of water quality. To avoid the limitations of a single model, in this paper, we use the Stacking integration method and propose an integrated learning model based on XGBoost, Support Vector Regression (SVR), Multilayer Perceptron (MLP), and Ridge Regression and apply it to the prediction of PH and TN in water quality. The remainder of this paper is as follows: Section 2 discusses the principles of the leading models in this paper, Section 3 shows the experimental design process of this paper, Section 4 shows the prediction results and analyzes and discusses the results, and Section 5 summarizes the whole essay and gives the main conclusions.

2. METHODOLOGY

2.1 Extreme Gradient Boosting

XGBoost is an optimization algorithm based on AdaBoost and GBDT [14], which is essentially a Boosting-based idea GBDT (Gradient Boosting Decision Tree) [15]. XGBoost seeks to maximize speed and efficiency based on GBDT. The core idea of the XGBoost algorithm is that each time a new function is used to fit the previous residuals of the prediction. Then the score of each node is calculated based on the characteristics of the sample, and the sum of the scores of all nodes is the predicted value of the final selection. The idea of this algorithm can be understood in Figure 1.

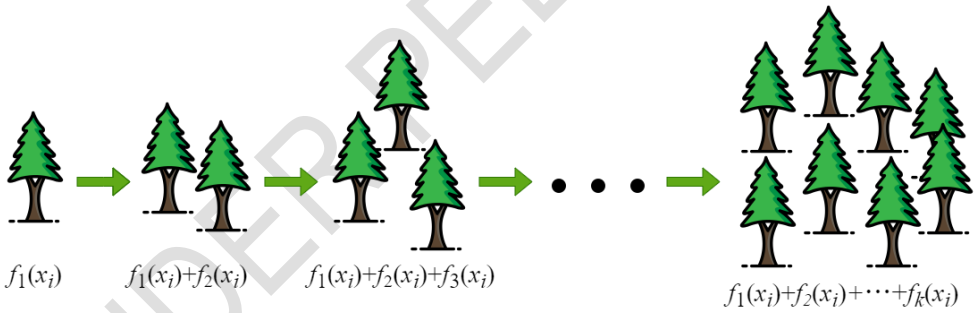


Figure 1 The process of gradient lifting integration model

The objective function of the XGBoost algorithm mainly consists of two parts, the loss function, and the regularization term, and the definition of $D = \{(x_i, y_i)\}, (|D| = n)$ is an $n \times m$ sample set, where n represents the number of samples and m represents the number of features. The mathematical model of the XGBoost algorithm is shown in Eq.1:

$$\hat{y}_i = \varphi(x_i) = \sum_{t=1}^{T} f_t(x_i) \tag{1}$$

Where $f_t(x_i)$ represents the regression tree space, where $f_t(x_i) = \omega_{q(x)}$, and q in $\omega_{q(x)}$ denotes the structure of each tree when the samples are mapped to the leaves, and each f_k

contains a separate q and $\omega_{q(x)}$; T represents the number of leaves. Then the objective function of the XGBoost algorithm can be expressed as:

$$\begin{aligned} obj &= \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ where\ \Omega(f) &= \gamma T + \frac{1}{2}\gamma \left|\left|\omega\right|\right|^2 \end{aligned} \tag{2}$$

Based on the core idea of the XGBoost algorithm, the iterative process of residual fitting can be derived as shown below:

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ &\vdots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \tag{3}$$

Each iteration aims to choose an optimal $f_t(x_i)$ that minimizes the objective function. The objective function is approximated using the second-order expansion of Taylor's formula and removing the constant term. Define I_j as the set of subscripts of samples on each leaf node j. Map each sample value x_i to a leaf node by the function $q(x_i)$ and expand ω. Rewrite the objective function again as in Eq.4:

$$\begin{aligned} obj^{(t)} &= \sum_{j=1}^{T}\left[G_j\omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma T \\ where\ G_j &= \sum_{i \in I_j} g_i\ , H_j = \sum_{i \in I_j} h_i \end{aligned} \tag{4}$$

Since $q(x)$ is deterministic, the optimal ω_j^* can be derived from the following equation.

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{5}$$

At this point, the corresponding objective function value can be calculated from equation .7. The result of the accurate function value can then be used to assess the quality of the tree results, and the smaller the matter, the better the tree's structure.

$$obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j + \lambda} + \gamma T \tag{6}$$

2.2 Multi-layer Perceptron

The multilayer perceptron (MLP) is an improvement and refinement of the feedforward neural network [16], which consists of three main parts: an input layer, an output layer, and a hidden layer, Figure 2 shows a schematic diagram of a three-layer perceptron. Each layer in the MLP has its fixed task, and the input layer is mainly responsible for receiving input

samples. The hidden layer is the core part of MLP, which is primarily responsible for processing and handling the input data. The output layer mainly performs activation processing and operates according to the tasks needed, such as prediction and classification.

Firstly, the input is input from the input layer and then passed by the hidden layer processing to the output layer. If the actual output of the output layer does not match the desired result, it will enter the backpropagation stage of the error. The error back propagation is to backpropagate the output error in some form through the hidden layer to the input layer and spread the error to all units in each layer to obtain the error signal of each team in each layer. This error signal will be used to correct the weights of individual units until the output error meets certain conditions or the number of iterations reaches a specific number.

Each neuron in the output and hidden layers performs the following computations.

$$O(x) = G(b(2) + W(x)h(x)) \tag{7}$$

$$h(x) = \Phi(x) = s(b(1) + W(1)x) \tag{8}$$

Where $b(1)$ and $b(2)$ denote the bias vectors, $W(1)$ and $W(2)$ mark the weight matrices, and G and S are the activation functions. It is worth mentioning that $W(1)$, $b(1)$, $W(2)$, and $b(2)$ are the parameters to be optimized. There are several choices of activation functions, as shown in Table 1.

Table 1 Several common activation functions

Function	Formula	Derivative
ReLU (Rectified linear unit)	$f(x) = \max(x, 0)$	$f(x)' = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$
sigmoide	$sigmoid(x) = \frac{1}{1+e^{-x}}$	$sigmoid(x)' = \frac{e^{-x}}{(1+e^{-x})^2}$
tanh	$tanh(x) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$	$tanh'(x) = 1 - \left(\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}\right)^2$

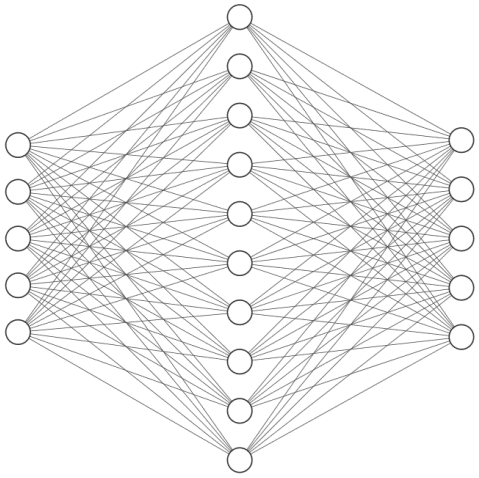


Figure 2 Schematic diagram of three-layer neural network

2.3 Support Vector Regression

Support Vector Machines (SVM) is a machine learning algorithm that constructs hyperplanes for separating different classes, usually for analyzing data with categorical output variables. In contrast, we use regression analysis instead of classification for continuous numerical output variables, called support vector regression (SVR) [17].The SVR model is used to obtain an approximation function g(x) from a given complex sample data $G = \{(x_i, y_i)\}\binom{N}{i=1}$ to obtain the approximate function $g(x)$. The main idea is to first map the nonlinearly separable data to a high-dimensional linearly separable feature space and then use this feature space for linear programming calculations.

$$f(x) = \sum_{i=1}^{D} w_i \phi_i(x) + b \tag{9}$$

In Eq.9, $\phi_i(x)$ is a feature with b and w_i as variables and is computable from the data. When the data are nonlinearly divisible, we need to map the data into a more decadent feature space where the data are separable. The minimum function coefficient w_i can be obtained from the following equation.

$$R[w] = \frac{1}{N}\sum_{i=1}^{N} |f(x_i) - y_i|_\epsilon + \lambda \left||w|\right|^2 \tag{10}$$

In Eq.10 λ is a normalization constant and the function $|f(x_i) - y_i|_\epsilon$ can be defined by Eq.11.

$$\left|f(x_i) - y_i\right|_\epsilon = \begin{cases} |f(x) - y| - \epsilon, \left|f(x_i) - y_i\right| \geq \epsilon \\ 0, \quad other \end{cases} \tag{11}$$

The minimization function can be expressed in the following form:

$$f(x, a, a^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{12}$$

Meanwhile, when $\alpha_i\alpha_i^* = 0$, $\alpha_i, \alpha_i^* \geq 0, i = 1, \cdots, N$, then the inner product kernel function can be expressed as:

$$k(x, y) = \sum_{j=1}^{D} \phi_j(x)\phi_j(y) \tag{13}$$

The following equation can calculate the parameters α_i and α_i^*:

$$\begin{aligned} R(\alpha_i^*, \alpha_i) = &-\varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) \\ &-\frac{1}{2}\sum_{i,j=1}^{N} (\alpha_i^* + \alpha_i)(\alpha_i^* - \alpha_i)k(x_i, x_j) \end{aligned} \tag{14}$$

There is a constraint: $\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0, \alpha_i \geq 0, \alpha_i^* \leq C$.

2.4 Ridge Regression

In the linear regression model, the parameter estimation formula is $\beta = (X^T X)^{-1} X^T Y$, but when $X^T X$ is not invertible, it is impossible to find β. In addition, if $|X^T X||$ tends to 0, the regression coefficient will tend to infinity, and the regression coefficient obtained at this time is meaningless. Ridge regression is an excellent solution to this problem.In the multiple linear regression, the objective function is:

$$J(\beta) = \sum (Y - X\beta)^2 \tag{15}$$

And to ensure that the regression coefficients β can be found in the ridge regression, an L_2 parametrization is introduced on the objective function as a penalty:

$$\begin{aligned} J(\beta) &= \sum (Y - X\beta)^2 + \lambda \left|\left|\beta\right|\right|_2^2 \\ &= \sum (Y - X\beta)^2 + \sum \lambda\beta^2 \end{aligned} \tag{16}$$

Where λ is a non-negative number, the larger λ is to make $J(\beta)$ minimum, the smaller the regression coefficient β will be at this time. For the above objective function values, the derivation process of ridge regression is given as follows.

$$\begin{aligned} J(\beta) &= \sum (Y - X\beta)^T (Y - X\beta) + \lambda\beta^T\beta \\ &= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta + \lambda\beta^T\beta \end{aligned} \tag{17}$$

We let $\frac{\partial J(\beta)}{\partial \beta} = 0$, then we can obtain.

$$\beta = (X^T X + \lambda I)^{-1} X^T Y \tag{18}$$

The addition of the L_2 parametric penalty term makes $(X^T X + \lambda I)$ total rank, but also the addition of the penalty term estimates the regression coefficients are no longer unbiased. Therefore, ridge regression is a method to solve multicollinearity and pathological matrices at the cost of giving up unbiasedness and reducing accuracy.

2.5 Construction of Ensemble Learning Model

The core idea of Stacking is to use the initial training data to learn several base learners and use the prediction results of these learners as a new training set to learn a new learner.

This study uses the Stacking fusion method to construct an integrated learning model based on Stacking with four models, XGBoost, MLP, SVR, and Ridge regression, as the base models. The base models of the first layer are: XGBoost, MLP, and SVR, and the base model of the second layer is: Ridge regression. The construction idea of the whole integrated learning model is shown in Figure 3.

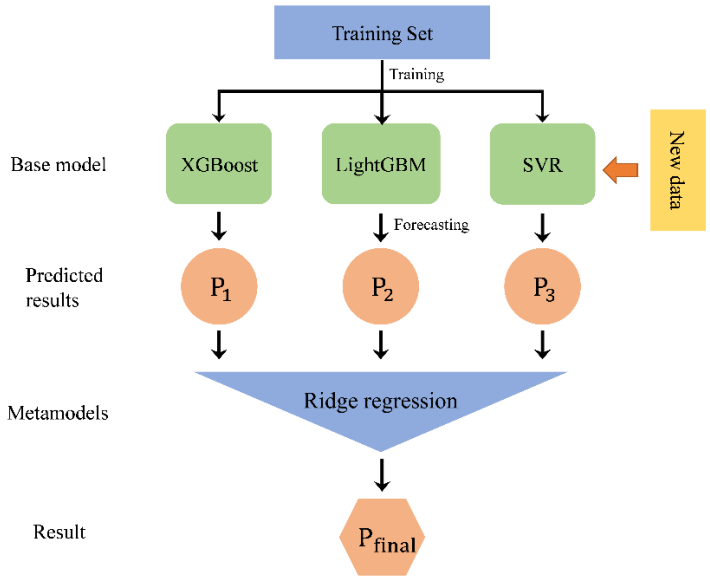


Figure 3 Stacking integration model construction

3. RESULTS AND DISCUSSION

3.1 Data Description

The data used in this study were the daily average monitoring data of PH and TN from 2021-10-01 to 2022-03-03 at Zhuhai Bridge Crossing, Guangdong Province. The data were derived from the Environmental Expertise Service System (http://envi.ckcest.cn/environment/). Figure.4 shows the correlation trends of the two different data sets. It can be seen that both data sets have extreme volatility and certain periodicity, and the more volatile data are often prone to underfitting, thus causing a decrease in inaccuracy.

In addition, we make calculations and statistics on the correlation properties of the datasets, as shown in Table 2 for some statistical characteristics of the two different datasets.

Table 2 Several features of the dataset

Data	Mean	Std	Max	Min	Mode	Numbers
PH	7.924	0.154	8.390	7.670	7.800	149
TN	1.910	0.239	2.890	1.480	1.900	149

The table shows that the two datasets are taught to differ significantly in their statistical characteristics, although they come from the same water source. In addition, the KS-test (Kolmogorov-Smirnov test) was used to test the distribution of the data, and the test P-value of the data set PH was obtained as 0.2247. The test P-value of the data set TN was 0.0006. It can be obtained that the test P-value of data set PH is more significant than 0.05 and therefore satisfies the normal distribution. The test P-value of the data set TN is less than 0.05 and does not satisfy the normal distribution.

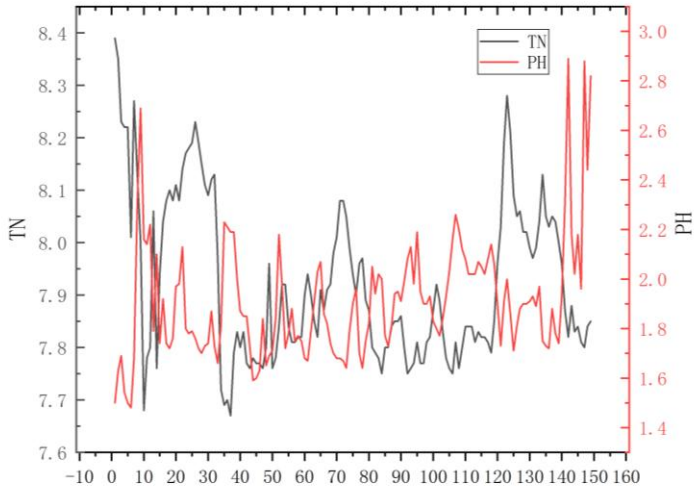


Figure 4 Data set trend chart

3.2 Evaluation Metrics

To quantitatively evaluate the model's prediction performance, the mean absolute percentage error (MAPE) is introduced in this study to assess the model's prediction performance. In addition, to highlight the enhancement effect of the Stacking method, we also compare the results of three base models, XGBoost, SVR, and MLP, with the Stacking integrated learning model for analysis.

3.2 Forecasting result

This section will show the prediction results and analyze and discuss the results. In this study, we used a multi-step prediction approach to predict PH and TN for the next five days based on historical data, and the consequences of each prediction step indicate a prediction for the next day's value. The prediction results of all models are presented in Table 3.

Table 3 Predicted MAPE values (%)

Data	Step / Model	1-step	2-step	3-step	4-step	5-step
TN	Stacking model	8.032	8.997	9.349	9.547	9.683
	XGBoost	8.843	9.714	10.502	10.562	10.561
	MLP	9.235	10.368	11.345	12.061	12.632

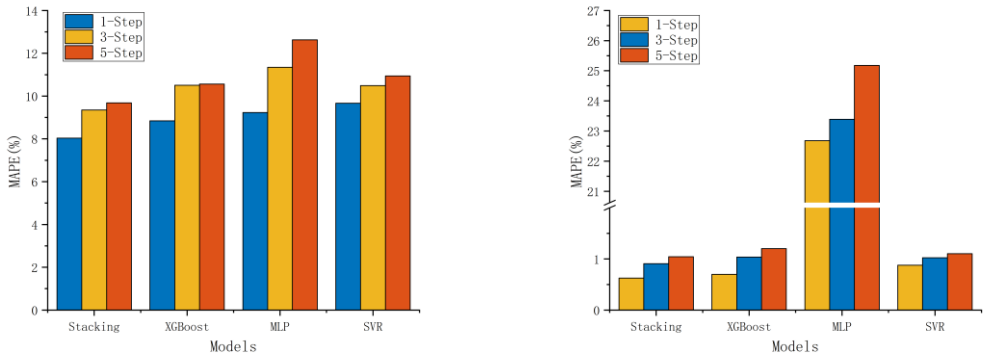
PH	SVR	9.669	10.218	10.490	10.733	10.945
	Stacking model	0.625	0.808	0.906	0.966	1.043
	XGBoost	0.700	0.914	1.036	1.123	1.202
	MLP	22.682	22.704	23.391	24.447	25.178
	SVR	0.879	0.961	1.024	1.069	1.103

According to the results in the above table, it can be learned that the Stacking integrated learning model has the smallest MAPE value from one-step to five-step prediction in the data TN, which indicates that the Stacking model does have a potent prediction ability. In the data PH, the MAPE values of the Stacking integrated learning model for one-step to five-step prediction are 0.625%, 0.808%, 0.906%, 0.956%, and 1.043%, respectively, which shows that the prediction ability of the Stacking prediction model is powerful from the MAPE. It is also worth noting that among all the prediction results, the prediction results of the Stacking integrated learning model are significantly better than the other three base models, which shows that the Stacking fusion method can effectively combine the points of each base model to improve the prediction ability of the model further.

It is also worth mentioning that the statistical characteristics of the two datasets and their distributions are described in Section 3.1. The two datasets differ significantly in both quantitative and statistical aspects, and the two datasets do not belong to the same distribution. However, from the prediction results, the Stacking integrated learning model has the best prediction performance in both datasets, which also shows the stability and generalization performance of the Stacking integrated learning model.

After discussing the overall prediction results, we next discuss the prediction step size and the three base models.

The change in MAPE for each model when the prediction step is increased on both datasets is shown in Figure 5.



(a) TN (b) PH

Figure 5 Error variation with prediction step

From the image, we can see that when the prediction step increases, the prediction error of the model becomes larger and larger, so it can be explained that the mistake of short-term prediction is much smaller than the error of long-term prognosis, the reason for this situation is the phenomenon of error accumulation, because there is an error in the accuracy of each prediction calculation, as the step length increases, the last prediction is based on the error of the previous forecast. Therefore, the longer the number of prediction steps, the larger the prediction error.

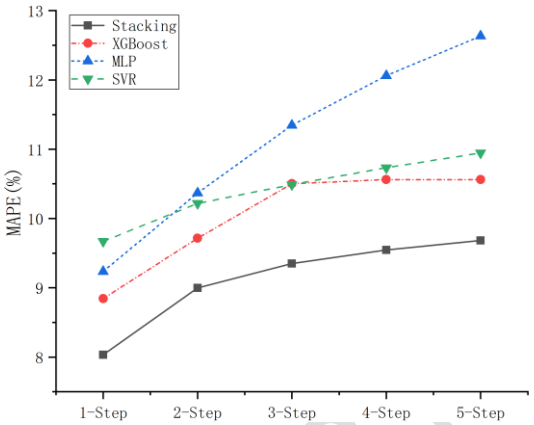


Figure 6 Prediction error of the base model

Finally, we discuss the prediction ability and generalization ability of the three base models separately, taking TN prediction as an example for the three base models. The prediction errors of the three base models on the dataset TN are shown in Figure 6. From the figure, we can see that, except for the Stacking integrated learning model, the prediction performance of the other base models on the dataset is similar. Still, XGBoost is the base model with the best prediction result. In addition, we can notice that in both the SVR and MLP models, the prediction result of MLP is due to SVR at one-step prediction. Still, with the increase of the prediction, step increases the prediction of the MLP model gradually increases. More noteworthy is that in the results of PH prediction in Table 3, the prediction error MAPE of MLP is higher than 20%, so it can indicate that the stability of MLP in water quality prediction is poor, and the prediction effect still needs to be improved compared with the other two base models.

4. CONCLUSION

To effectively predict the concentration and change trend of various pollutants in water bodies and promote decision-makers in rational development, optimal allocation, comprehensive conservation, efficient utilization, adequate protection, and comprehensive management of water resources work and have a basis. This study is based on a machine-learning algorithm to predict TN and PH in water quality. Since the generalization ability of individual models is often thin, an integrated learning model based on the Stacking integration idea and four models of XGBoost, MLP, SVR, and Ridge regression is established in this study and applied to water quality prediction.

Summarizing this paper and this study, we can get the following three conclusions: 1) Stacking integrated learning model has strong prediction ability and generalization performance, can be widely used in the field of water quality prediction, and also has strong potential in other prediction fields. 2) Stacking integrated idea can effectively combine the

advantages of each base model, based on the base model to effectively improve the prediction. 3) in the multi-step prediction, with the increase of the prediction step prediction accuracy will gradually decline, in the three base models compared to the MLP other two base models have better stability and generalization performance.

REFERENCES

1. Yin, Su, et al. "Contribution of the upper river, the estuarine region, and the adjacent sea to the heavy metal pollution in the Yangtze Estuary." Chemosphere 155 (2016): 564-572.
2. Hounslow, Arthur W. Water quality data: analysis and interpretation. CRC press, 2018.
3. Lu, Hongfang, and Xin Ma. "Hybrid decision tree-based machine learning models for short-term water quality prediction." Chemosphere 249 (2020): 126169.
4. Panaskar, D. B., et al. "Evaluating groundwater suitability for the domestic, irrigation, and industrial purposes in Nanded Tehsil, Maharashtra, India, using GIS and statistics." Arabian Journal of Geosciences 9.13 (2016): 1-16.
5. Kulisz, Monika, et al. "Forecasting water quality index in groundwater using artificial neural network." *Energies* 14.18 (2021): 5875.
6. Barzegar, Rahim, et al. "Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model." *Stochastic environmental research and risk assessment* 32.3 (2018): 799-813.
7. Khullar, Sakshi, and Nanhey Singh. "Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation." *Environmental Science and Pollution Research* (2021): 1-15.
8. Abobakr Yahya, Abobakr Saeed, et al. "Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios." *Water* 11.6 (2019): 1231.
9. Barzegar, Rahim, Mohammad Taghi Aalami, and Jan Adamowski. "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model." *Stochastic Environmental Research and Risk Assessment* 34.2 (2020): 415-433.
10. Ahmed, Ali Najah, et al. "Machine learning methods for better water quality prediction." *Journal of Hydrology* 578 (2019): 124084.
11. Wagena, Moges B., et al. "Comparison of short-term streamflow forecasting using stochastic time series, neural networks, process-based, and Bayesian models." *Environmental Modelling & Software* 126 (2020): 104669.
12. Gu, J.; Liu, S.; Zhou, Z.; Chalov, S.R.; Zhuang, Q. A Stacking Ensemble Learning Model for Monthly Rainfall Prediction in the Taihu Basin, China. Water 2022, 14, 492. https://doi.org/10.3390/w14030492.
13. Zhang, L., Yang, Y., Deng, Y., & Kang, H. (2022). Forecasting of Road Traffic Flow Based on Harris Hawk Optimization and XGBoost. Journal of Advances in Mathematics and Computer Science, 37(2), 21-29. https://doi.org/10.9734/jamcs/2022/v37i230433
14. CHEN, T., and C. GUESTRIN. "Xgboost: A scalable tree boosting system; proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, F, 2016 [C]."
15. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.
16. Raj, Pethuru, and Preetha Evangeline David. *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*. Academic Press, 2020.
17. Liu, Bing-Chun, et al. "Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang." *PloS one* 12.7 (2017): e0179763.