IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2021, accepted November 29, 2021, date of publication December 6, 2021,
date of current version December 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3133107

An Improved-Bagging Model for Water Chemical Oxygen Demand Measurements Using UV-Vis Spectroscopy

JINGWEI LI, SISI PAN, JIE BIAN, AND WEI JIANG
College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou 225127, China
Corresponding author: Wei Jiang (jingwei_li@yzu.edu.cn)

This work was supported by the Yangzhou City—Yangzhou University Cooperation Foundation under Grant YZ2020169.

ABSTRACT The ultraviolet-visible (UV-Vis) spectroscopy measurement method of Chemical Oxygen Demand (COD) in water is a simple physical method that can measure water without secondary pollution from chemical reagents. To solve the problems of low accuracy and insufficient generalization capability of the COD prediction model, an improved Bagging algorithm is proposed and evaluated in this study. The Improved-Bagging algorithm can reduce model variance and bias concurrently, and improves the accuracy and stability of the traditional Bagging algorithm. Results show that the Improved-Bagging algorithm achieves a better prediction ability on different preprocessed data than the traditional Bagging algorithm. After ensemble empirical mode decomposition based (EEMD-Based) algorithm denoising and stability competitive adaptive reweighted sampling (SCARS) algorithm dimension reduction, Improved-Bagging model achieves the best prediction performance. Its coefficient of determination (R^2) on the prediction set reached 0.9317, its root mean square error of prediction (RMSEP) reached 5.39 mg/L, and its variance reached 5.53 mg^2. Results also show that the Improved-Bagging algorithm can accurately measure the COD concentration in water, which lays the foundation for the wide application of spectroscopy to measure water quality parameters.

INDEX TERMS COD measurements, improved-bagging model, UV-Vis spectroscopy, water.

I. INTRODUCTION

COD describes the pollution of water by reducing substances, is an important parameter for evaluating water, but also a required parameter in water quality measurements. Generally, there are two methods for measuring COD: chemical methods and physical methods (i.e., spectroscopy methods) [1], [2]. Chemical methods generally use strong oxidants, such as potassium permanganate and potassium dichromate, to oxidize water samples under strong acid conditions, and then calculate the COD in water by measuring the amount of oxidant consumed. Chemical methods have disadvantages such as secondary pollution and long measurement periods, which are not suitable for online and real-time measurements [3], [4]. The wavelength range for measuring COD via spectroscopy is generally in the ultraviolet-visible interval. After UV-Vis spectroscopy is transmitted through water, the corresponding COD value is obtained by measuring the absorbance of the water [5]. The water quality measurement method based on UV-Vis spectroscopy has received increasing attention in recent years, and its application prospects are good.

Essentially, using UV-Vis spectroscopy to measure COD in water allows a COD prediction model to be built based on UV-Vis spectra. By building a calibration model between the UV-Vis spectrum data of water and the COD standard values, the corresponding COD concentration in water can be predicted based on water's spectrum. Therefore, the prediction accuracy of the model depends on the pros and cons of the calibration modeling method [6]. In recent years, with the development and breakthrough of statistics, applied mathematics, chemometrics, artificial intelligence and other fields, some new modeling methods have been applied to UV-Vis spectroscopy to measure COD, providing a new idea to measure COD concentrations in water with a complex composition [7]–[9]. Modeling methods primarily include statistical methods and machine learning methods. Appropriate modeling methods should be used based on the specific application

The associate editor coordinating the review of this manuscript and approving it for publication was Wen-Sheng Zhao.

environments being considered. According to the morphological characteristics of different water spectra, statistical methods or machine learning methods are used to build a prediction model suitable for water, which has attracted substantial attention and led to useful results. In early studies, COD was measured via UV-Vis spectroscopy at one or several wavelengths, which is referred to as a single wavelength or multi-wavelength method. Linear regression (LR) and multiple linear regression (MLR) are primarily used for modeling. and is a simple and accurate method for water samples that are of uniform composition and remain relatively similar over time; however, these methods are unsuitable for water samples whose composition changes markedly with time [10]–[13]. However, the full spectrum contains more abundant information, which can effectively improve the accuracy of COD measurement. Therefore, the application of a combination of chemonetrics methods and a full spectrum is the current development trend. For example, partial least squares (PLS), support vector machine (SVM), random forest regression (RFR), and artificial neural networks (ANNs), have already been applied to UV-Vis spectroscopy to measure COD in water [14]–[22]. Machine learning algorithms developed with the rise of artificial intelligence have the characteristics of various types, flexible use, and fast optimization speed, and can be improved. Both measurement performance and the scope of application have made a qualitative leap. However, many machine learning algorithms have not yet been used in real-world water quality measurement applications using UV-Vis spectroscopy, such as ensemble learning methods, thus, there is still much work to be done. Ensemble learning has achieved good results in applications in many fields and has the advantages of improving prediction accuracy, and stability, and eliminating overfitting [23]–[27]. Therefore, this paper uses ensemble learning in COD measurements based on UV-Vis spectroscopy to build a COD calibration model. To improve model accuracy, the basic ensemble learning method is improved to be more suitable for the research demand of COD measurements.

This paper thus proposes an Improved-Bagging algorithm based on elastic net regression. Drawing on the idea of the two-phase learning of the Stacking algorithm, the ''learning method'' is used to replace the ''simple average'' ensemble strategy (regression problem) in the traditional Bagging algorithm. Combined with the characteristics of elastic net regression, which is simple and has a feature extraction function, elastic net regression is used as the combination (ensemble) strategy of base learners in Bagging algorithm. Thus, the influence of the high collinearity among the base learners in Bagging on the accuracy of the final ensemble model can be mitigated. Therefore, elastic net regression is used instead of simple averaging to improve the Bagging algorithm. The improved algorithm can retain the advantages of both the Bagging algorithm and the Stacking algorithm, which can reduce the variance and bias concurrently. Thus, the prediction accuracy of the Bagging model is improved. In addition, a variety of spectrum denoising and dimension reduction methods are used to preprocess the spectrum data that are input to the mode. Spectrum features that can comprehensively reflect the COD in water are extracted from high-dimensional spectrum data, to speed up model convergence and reduce computational complexity.

The remainder of this paper is organized as follows. Section 2 presents the materials required for the experiment and the proposed methods for spectrum modeling. Section 3 reports the results of COD predictions from different models and discusses the prediction results. Section 4 concludes the paper.

II. MATERIALS AND METHODS

A. INSTRUMENTS AND SAMPLES

1) EXPERIMENTAL INSTRUMENTS

The measurement of COD in water by UV-Vis spectroscopy was performed with a COD measurement instrument system. The system structure is shown in Figure 1 and is primarily composed of a light source, sample cell, spectrometer and a computer. The light source used in the experiment was a DH-2000-DUV deuterium-halogen-tungsten light source, which can provide 190-2500 nm light (Ocean Optics, USA). The optical path length of the sample cell for water is 10 mm. The UV-VIS spectrometer used in the experiment was USB2000+ and can measure light with wavelengths between 165 and 1200 nm with a resolution of 0.45 nm (Ocean Optics, USA). OceanView spectrum acquisition software was used to analyze all data, comes with a spectrometer, and stores the spectrum wavelength range from 193.91 to 1121.69 nm. The baseline was corrected based on deionized water, and the integration time of the spectrometer was 10 ms. Each water sample was successively scanned 10 times, and the average value was taken. A computer is used to save data, process data and build models with the corresponding software.

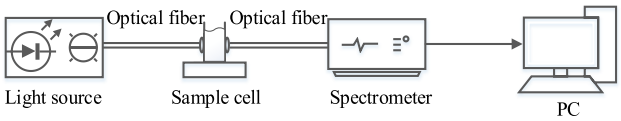


FIGURE 1. Structure diagram of the COD measurement instrument system based on the UV-Vis spectroscopy method.

2) EXPERIMENTAL SAMPLES

Water samples were collected from Qian Lake in the center of Nanjing city, which is important to local fisheries, bird habitats, water resources and the regional environment. The water quality of this lake is affected by domestic sewage from the growing urban population. Therefore, it is critical for local residents to effectively measure and monitor water quality, and to warn of water quality problems quickly. From June 2019 to June 2020, water samples from the lake were collected once per day (except holidays) for one year; a total of 249 samples were collected throughout the year. Some samples had many impurities, were not suitable for further

research, and were not considered in this study. Concurrently, considering the next step of sample set division, 240 samples were selected from all collected samples. Each water sample was divided into two parts: one was used to measure the standard value of COD, and the other was used to collect UV-Vis spectrum data. To retain the original characteristics of the lake water to the maximum extent, the collected water samples should be immediately measured for UV-Vis spectroscopy and COD standard values.

B. UV-VIS SPECTRUM COLLECTION AND COD STANDARD VALUE MEASUREMENT

1) UV-VIS SPECTRUM COLLECTION

Using the UV-Vis spectrum collection instrument system introduced in Figure 1, the collected water samples were immediately measured for UV-Vis spectrum data. Baseline correction was based on deionized water. The integration time of the spectrometer was set to 10 ms, each water sample was scanned 10 times, and the average value was taken. The original UV-Vis spectra of 240 water samples are shown in Figure 2.

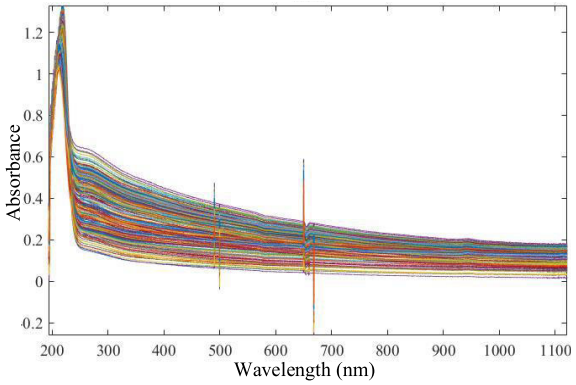


FIGURE 2. Original UV-Vis spectra of 240 samples.

Figure 2 shows the curve of the original UV-Vis spectra of the collected water samples. The curve trend of the original spectrum data of the water samples collected at different times is similar. A strong absorption peak is shown at approximately 235nm, which is due to absorption by COD. Jumps and pinnacles appear at approximately 500 nm and 680 nm, which are due to the presence of noise in the spectrum. Therefore, due to the deficiency of the collected spectra, it is necessary to preprocess the collected spectra before building a model to improve the accuracy of COD measurement based on UV-Vis spectroscopy. Jumps and pinnacles at approximately 500 nm and 680 nm are non-sensical and are removed before preprocessing.

2) COD STANDARD VALUE MEASUREMENT

According to the rapid digestion spectrophotometry method, the COD of the collected water samples was measured by a DRB200 digester and a DR3900 visible spectrophotometer (HACH, USA). The required chemical reagents and water samples to be tested were fully mixed and put into a DRB200 COD digester preheated to 165 °C in advance for 20 minutes. After digestion, each digestion tube was placed onto the cooling shelf to cool. After cooling to room temperature (25±1)°C, color reagent was added to each digestion tube. Finally, the COD value was measured with a DR3900 spectrophotometer.

C. SAMPLE SET DIVISION

A reasonable calibration set can improve the prediction ability of the built calibration model. Common sample selection methods include random sampling (RS), conventional selection (CS), Kennard stone (KS) and sample set portioning based on joint X-Y distance (SPXY) [28]–[31]. The RS method cannot guarantee the representativeness of the selected samples because it randomly selects the samples of the calibration set. The CS method selects the samples according to the chemical measurement values of the samples and selects the samples with the maximum or minimum chemical measurement values as the calibration set samples. RS and CS are subjective in sample selection. The KS method selects the two sample pairs with the farthest Mahalanobis distance for inclusion in the calibration set; calculates the distance from each remaining sample to each selected sample in the calibration set; determines the minimum distance sample and the maximum distance sample; and adds them to the calibration set. This step was repeated until the number of samples in the calibration set met the requirements. The SPXY algorithm is a new sample set division method based on the KS algorithm and considers the scientific division of the sample set by comprehensively considering the spectrum and chemical values of the samples. This algorithm has the advantages of covering multidimensional vector space and effectively improving the prediction performance of the calibration model. According to the SPXY algorithm, the calibration set and prediction set are divided according to the ratio of 2:1; therefore, among the 240 water samples, 160 samples are used as the calibration set, and 80 samples are used as the prediction set. The statistical characteristics of the samples are shown in Table 1.

D. UV-VIS SPECTRUM PREPROCESSING

Due to the influence of experimental conditions including spectrometer hardware and natural light, the original spectrum data collected will contain some noise, as shown in Figure 2. If the original spectra are used directly, the reliability and stability of the calibration model will be affected; thus, it is important to preprocess the original spectrum data properly in advance. By preprocessing the original spectrum data, we can effectively reduce the influence of external factors on the spectrum; improve the correlation between the spectrum and the component to be measured; and then build a robust and reliable prediction model. In this paper, Gaussian smoothing (SG), Fourier transform (FT), wavelet transform (WT) and EEMD-Based denoising algorithms are used to process the spectrum, and the effects of the four

TABLE 1. Statistical results of COD standard values of 240 water samples.

Sample set	Samples	Minimum (mg/L)	Maximum (mg/L)	Average value (mg/L)	Standard deviation (mg/L)
Calibration	160	14	116.2	59.7	31.2
Prediction	80	16.3	115.8	59.6	31.3
All	240	14	116.2	59.7	31.2

denoising algorithms are compared [32]–[36]. SG smoothing is a type of linear smoothing method that is suitable for eliminating Gaussian noise and is widely used in various data. FT has a good effect on the denoising of stationary signals, and WT denoising methods have been widely studied and have achieved good results in a variety of spectral denoising techniques. The EEMD-Based denoising method is a new denoising method that has strong adaptability and plays an important role in signal denoising.

E. SPECTRUM DIMENSION REDUCTION ALGORITHM

After denoising, the UV-Vis spectra collected by the instrument system still have serious nonlinear or linear overlap, and the spectrum data dimension is high. High-dimensional data contain a large amount of redundant information and hide important information. If the full UV-Vis spectrum is used as the input variable of the calibration model, it will lead to large computing resources and introduce the interference of unknown substances, thus reducing the measurement accuracy and generalization performance of the model. Therefore, it is necessary to reduce the dimension of the high-dimensional spectrum, extract its effective feature information, and improve the efficiency of model training and generalization ability. Data dimension reduction can be divided into two categories: feature transformation and feature selection. From the results of spectrum feature selection, feature selection can be divided into continuous feature selection (wavelength interval selection) and discontinuous feature selection (limited discontinuous wavelengths). According to the results of spectrum feature dimension reduction, this paper will select a representative algorithm from the feature transformation, continuous feature selection and discontinuous feature selection to analyze. PCA, interval partial least squares (iPLS), and SCARS were used to reduce the dimension of the full UV-Vis spectrum, and the calibration model performance of COD was analyzed [37]–[39].

F. MODELING ALGORITHM

1) BAGGING ALGORITHM

Bagging [40], [41] is an ensemble learning method, that uses the same base learner (different data subsets) to generate multiple different learners and combines these learners into an ensemble model. The core idea of Bagging is bootstrap. For a given dataset containing n samples, we first perform m random samplings with replacement to obtain a data subset containing m samples and use this data subset to train a base learner. The operation is repeated T times, and T base learners are generated. Due to the bootstrap sampling method, there

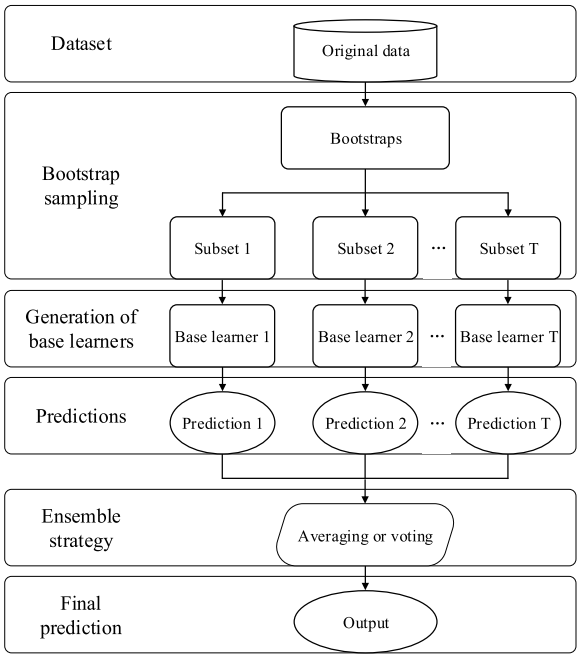


FIGURE 3. Schematic diagram of bagging algorithm.

are differences between the data subsets; thus, there are also marked differences between the T base learners. Finally, the T base learners are combined. For the regression problem, the average method is used to combine the base learners. For the classification problem, the voting method is used to combine the base learners to produce a final ensemble model. Among the T base learners, the accuracy of each base learner is not necessarily high, but the result of their ensemble is very high. A schematic diagram of Bagging algorithm is shown in Figure 3.

The Bagging algorithm changes the distribution of the original dataset by resampling and produces a number of data subsets with differences. The more unstable the base learner is to the data subset, the better the performance of Bagging. Currently, decision tree (DT) and artificial neural networks (ANNs) are commonly used as the base learners of Bagging because these two algorithms are sensitive to training data. Considering the limited number of samples used in this paper, it is not suitable to use ANNs as the base learner; thus, DT is used as the base learner. The primary advantage of the Bagging algorithm is to reduce model variance; the increase in performance by reducing bias is negligible. Therefore, this paper explores methods to reduce Bagging bias.

2) STACKING ALGORITHM

Stacking [42], [43] is also a famous ensemble learning method and is different from Bagging in base learner selection. The base learner of Bagging is usually the same algorithm (with different training data subsets), while the base learner of Stacking algorithm is a different learning algorithm. The Stacking algorithm trains base learners, takes their outputs as inputs of the second learner (meta-learner), and generates the final ensemble results through two-phase learning. The Stacking algorithm first trains the first phase learners from the original training dataset, and then uses the prediction results of the first phase learners to form a new dataset for training the meta-learner. A schematic diagram of a two-phase Stacking algorithm is shown in Figure 4.

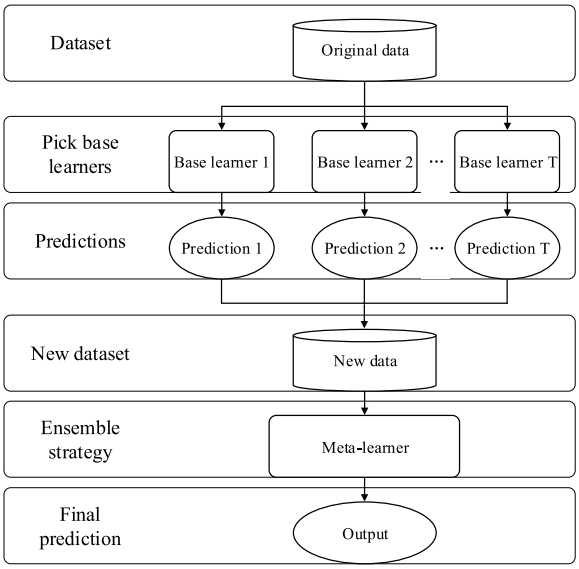


FIGURE 4. Schematic diagram of a two-phase Stacking algorithm.

Different from Bagging, Stacking divides the model into two phases. The first phase trains the sample set and predicts the results, and the meta-learner uses the results of the first phase for further learning to find and correct the bias in the first phase and improve the accuracy of the ensemble model. Stacking is a generalization of ensemble strategy and is an ensemble method based on the ''learning method'' that uses meta-learner to replace the average method (regression problems) in Bagging to reduce model bias. Therefore, Stacking can make full use of the advantages of two-phase learning, reduce model bias, and improve the accuracy of the ensemble model.

3) ELASTIC NET REGRESSION

In the standard linear regression model, the model relates y to x as follows:

$$y = \omega^T x_i + b \tag{1}$$

The regression coefficients in ω can be estimated by optimizing the following elastic net penalty function as Equation (2). When using the elastic net penalty, we obtain the elastic net regression [44]:

$$min \frac{1}{2} \sum_{i=1}^{N} \left(\omega^T x_i + b - y_i\right)^2 + r \|\omega\|_1 + \frac{1-r}{2} \|\omega\|_2^2 \tag{2}$$

where $\{x_i, y_i\}$ is the sample data, $x_i \in R^n$ is the independent variable, and y_i is the corresponding dependent variable. $\omega \in R^n$ is the feature weight vector, and $b \in R^n$ is the intercept. $0 \leq r \leq 1$ is the regularization parameter, which controls how much of the loss function is ridge regression and lasso regression. When $r = 0$, a complete ridge regression is performed. When $r = 1$, a complete lasso regression is performed.

As a typical linear regression technology, elastic net regression integrates the ridge regression and lasso regression algorithms. Elastic net regression can shrink regression coefficients while performing regularization to select characteristic variables such as lasso regression and obtain a simpler model. Elastic net regression can also select closely associated variables, such ridge regression, to select features, simplify the model, and ensure its stability. Therefore, elastic net regression combines the advantage of ridge regression and lasso regression, and performs feature extraction and regression analysis concurrently. Elastic net regression achieves better performance with data that contain many characteristic variables that are associated with each other.

4) IMPROVED-BAGGING ALGORITHM

When the Bagging algorithm is used to solve regression problems, the ensemble of base learners is typically reported as a simple average. However, there is a high collinearity between the base learners, and the simple average between the collinear variables is limited to improve model accuracy. To eliminate the effects of collinearity, the meta-learner of Stacking was referenced, and the simple average was replaced by the ''learning method''. The meta-learner selected in this study should overcome the problem of high collinearity between the base learners in Bagging and should avoid the risk that the meta-learner is too complex to lead to overfitting of the ensemble model. Therefore, the elastic net regression algorithm is simple and performs feature extraction, which can reduce the influence of the high collinearity between the base learners in the Bagging algorithm. Therefore, elastic net regression as a meta-learner is introduced to replace the simple average to improve the Bagging algorithm. Elastic net regression can make the improved Bagging (Improved-Bagging) algorithm retain the advantages of both Bagging algorithm and Stacking algorithm, which can reduce model variance and bias concurrently. The Improved-Bagging algorithm is shown in Figure 5.

The workflow of the Improved-Bagging algorithm is as follows:

Input: Original data $S = \{(x_p, y_p), p = 1, \cdots, N\}$;
Base learner f;
Meta-learner h;
Number of base learners T.

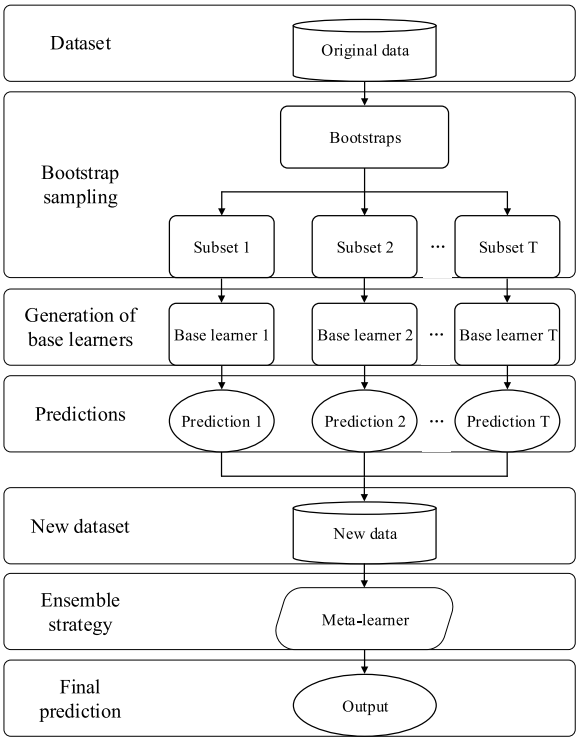


FIGURE 5. Schematic diagram of improved-bagging algorithm.

Phase 1: %Training base learners

for t = 1, . . . ,T:

$S_t = Bootstrap(t)$ %Generating different data subsets S_t from the original dataset S by the bootstrap method

$F_t = f\,(S_t)$ %Using data subset S_t for training the base learner to obtain the prediction result F_t

end;

for n = 1, . . . , T: %Building a new dataset D_h used for meta-learner training

$X'_n = \{F_1\,(x_n)\,, \cdots, F_T\,(x_n)\}$

$D_h = \left\{\left(X'_n, y_n\right), n = 1, \cdots, T\right\}$ % New dataset D_h

end;

Phase 2: %Training meta-learner

$\mathrm{H} = \mathrm{h}\,(D_h)$

Output: Ensemble mode H

Based on the two phases of the Improved-Bagging algorithm, the Bagging algorithm first trains the base learner f repeatedly using different data subset to obtain T base prediction models $F_1, F_2, \cdots, F_T$ and outputs the prediction results. Then, according to the outputs of the base learners, a new dataset D_h for meta-learner training is constructed. Finally, the new dataset D_h is used to train the meta-learner h to obtain the ensemble model H.

G. MODEL TUNING

When the data set and model remain unchanged, tuning model hyper-parameters is an effective method to reduce model complexity and improve model accuracy. Grid Search (GS) is a model parameter tuning method based on traversal. With the improvement of computer hardware, the computing power and speed of the computer have been greatly improved. Therefore, more search levels and smaller search steps can be set during GS to improve the accuracy of the model. The hyper-parameters that the Improved-Bagging algorithm needs to tune include the parameters of Improved-Bagging and the parameters of the base learner (DT), forming a two-level grid search. The root mean square error of calibration (RMSEC) is used as a fitness function to assess the pros and cons of each group of parameters. The smaller the fitness function value, the higher the model accuracy.

H. PERFORMANCE INDICES

Machine learning methods must be used to build a spectrum data model, but different types of modeling methods have different advantages and disadvantages. Therefore, the comparison of the prediction performance of different models must use quantitative model performance indices. The evaluation of the prediction performance of the model was based on several performance indices, including R^2, root mean square error of calibration (RMSEC) and RMSEP, variance (s^2). The larger R^2 and the smaller RMSEC/RMSEP are, the better the model. The smaller RMSEC/RMSEP, the smaller model bias. The smaller s^2, the smaller model variance. The equations of these performance indices are shown as follows:

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\hat{y}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y}_{\mathrm{i}} - \bar{y}\right)^2} \tag{3}$$

$$\mathrm{RMSEC} = \sqrt{\frac{1}{\mathrm{n}^{\mathrm{c}} - 1} \sum_{\mathrm{i}=1}^{\mathrm{n}^{\mathrm{c}}} \left(\hat{y}_{\mathrm{i}}^{\mathrm{c}} - \mathrm{y}_{\mathrm{i}}^{\mathrm{c}}\right)^2} \tag{4}$$

$$\mathrm{RMSEP} = \sqrt{\frac{1}{\mathrm{n}^{\mathrm{p}}} \sum_{\mathrm{i}=1}^{\mathrm{n}^{\mathrm{p}}} \left(\hat{y}_{\mathrm{i}}^{\mathrm{p}} - \mathrm{y}_{\mathrm{i}}^{\mathrm{p}}\right)^2} \tag{5}$$

$$s^2 = \frac{\sum_{\mathrm{i}=1}^{n^m} \left(\hat{y}_{\mathrm{m-i}}^{\mathrm{p}} - \mu\right)^2}{n^m} \tag{6}$$

where y_i is the measured value based on the standard method; $\bar{y}$ is the average value of y_i; $\hat{y}_i$ is the predicted value based on spectroscopy method; n is the number of samples; y_i^c is the measured value based the standard method of calibration set; $\hat{y}_i^c$ is the predicted value based on spectroscopy method of calibration set; n^c is the number of samples of calibration set; y_i^p is the measured value based the standard method of prediction set; $\hat{y}_i^p$ is the predicted value based on the spectroscopy method with the prediction set; n^p is the number of samples of prediction set; μ is the average value of $\left(\hat{y}_{\mathrm{m-1}}^{\mathrm{p}}, \hat{y}_{m-2}^{\mathrm{p}}, \cdots, \hat{y}_{\mathrm{m-n}}^{\mathrm{p}}\right)$; and n^m is the number of predictions for the same sample. The model variance (s^2) represents the variance of the predicted value for the same sample under different calibration sets. In this paper, the sample with the median value of COD in the calibration set and prediction set (both COD = 46.3 mg/L) was used to calculate the variance. Tenfold cross-validation modeling was used to predict the sample (COD = 46.3 mg/L) 10 times, and the variance of

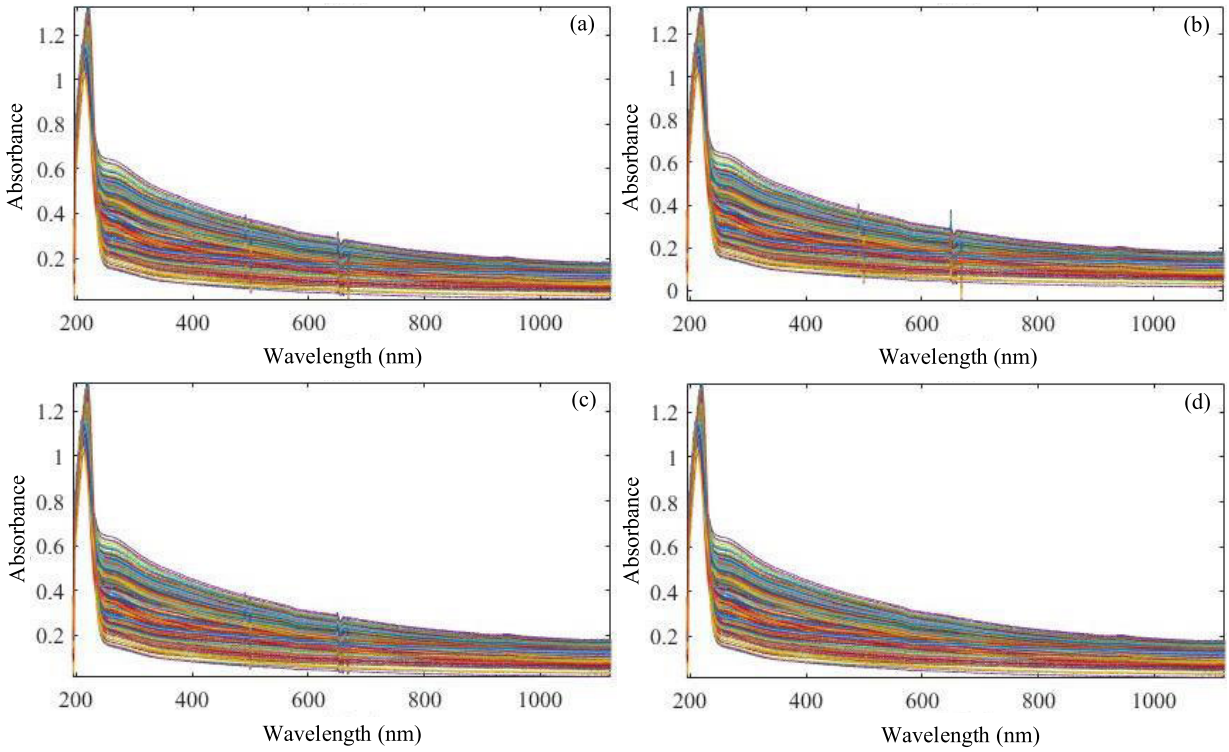


FIGURE 6. UV-Vis spectra denoised by different methods. (a) GS denoising. (b) FT denoising. (c) WT denoising. (d) EEMD-based denoising.

the 10 prediction results was used to represent the variance of the model.

III. RESULTS AND DISCUSSION

A. UV-VIS SPECTRUM PREPROCESSING

Figure 2 shows the original UV-Vis spectrum data curve of water samples collected by the experimental instrument. The trends of the original spectra of different water samples are relatively similar but contain considerable noise. Figures 6(a) $\sim$ 6(d) show the spectra denoised by the SG, FT, WT and EEMD-Based methods.

By comparing the denoised spectra in Figure 6 with the original spectra in Figure 2, the denoised spectrum is shown to retain the basic absorption characteristics of the original spectrum. After processing by different denoising methods, the jump and pinnacle noise in the original spectrum has been reduced to some extent, and the EEMD-Based denoising method achieves the best result. However, the influence of different denoising methods on the COD prediction results still must be studied in more detail via modeling.

B. MODEL PERFORMANCE OF FULL UV-VIS SPECTRUM

To study the processing results of different spectrum denoising methods on the original UV-Vis spectra, and study the fitting effect of the proposed modeling method on the spectrum data, this section compares the performance of the combination of different spectrum denoising and modeling methods on COD prediction. Data denoising methods include raw UV-Vis spectra without any processing, GS, FT, WT and EEMD-Based denoising. The modeling methods include DT, Bagging, random forest (RF) and Improved-Bagging. The COD prediction performance of the combination of different spectrum denoising and modeling methods is shown in Table 2.

Comparing the results shown in Table 2, if the modeling method is the same, the model built by UV-Vis spectrum denoised by EEMD-Based method is better than other denoising methods. The prediction set R^2 is the largest, and the RMSEP and variance are the smallest, which indicates that the EEMD-Based denoising method is more effective; thus, additional research is performed with EEMD-Based denoising used during preprocessing. If the denoising method is the same, the Improved-Bagging model is better than the other two modeling methods, the prediction set R^2 is the largest, and the RMSEP and variance are the smallest, which demonstrates the improved performance of the proposed Improved-Bagging modeling method. Among the full spectrum prediction models of COD in water shown in Table 2, the optimal prediction model is the UV-Vis spectrum processed by the EEMD-Based denoising method and modeled by Improved-Bagging algorithm. The R^2 of the prediction set was 0.9054, and the RMSEP and variance of the prediction set were 7.11 mg/L and 6.73 mg^2, respectively.

C. DIMENSION REDUCTION OF UV-VIS SPECTRUM

In this study, the original spectrum wavelength range is 193.91-1121.69 nm, and the spectrum resolution is 0.45 nm, including 2048 wavelength features. Full spectrum modeling

TABLE 2. COD prediction performance of combination of different denoising and modeling methods.

Preprocessing methods	Data dimension	Models	Calibration set			Prediction set		
			R^2	RMSEC (mg/L)	Variance (mg^2)	R^2	RMSEP (mg/L)	Variance (mg^2)
Raw spcetrum	2048	DT	0.7712	15.90	16.84	0.7356	18.23	19.23
	2048	Bagging	0.8197	12.72	13.42	0.7886	14.76	15.72
	2048	RF	0.8213	12.62	12.17	0.7924	14.51	14.36
	2048	Improved-Bagging	0.8221	12.56	8.04	0.8003	13.99	9.33
GS	2048	DT	0.8055	13.65	15.08	0.7628	16.45	17.29
	2048	Bagging	0.8491	10.80	11.37	0.8237	12.46	13.24
	2048	RF	0.8573	10.26	8.71	0.8214	12.62	10.13
	2048	Improved-Bagging	0.8616	9.98	7.35	0.8231	12.50	8.71
FT	2048	DT	0.7671	16.17	18.27	0.7369	18.15	19.82
	2048	Bagging	0.8253	12.35	13.55	0.8005	13.98	14.89
	2048	RF	0.8479	10.87	9.73	0.8267	12.26	11.57
	2048	Improved-Bagging	0.8520	10.60	7.28	0.8434	11.17	8.55
WT	2048	DT	0.7947	14.36	15.12	0.7684	16.08	17.53
	2048	Bagging	0.8539	10.48	11.67	0.8433	11.18	12.81
	2048	RF	0.8755	9.07	9.61	0.8549	10.42	11.76
	2048	Improved-Bagging	0.8874	8.29	6.95	0.8703	9.41	7.81
EEMD-Based	2048	DT	0.8744	9.14	10.34	0.8481	10.86	11.18
	2048	Bagging	0.8979	7.60	9.22	0.8749	9.11	10.23
	2048	RF	0.9122	6.66	6.47	0.8918	7.80	8.19
	2048	Improved-Bagging	0.9177	6.30	5.82	0.9054	7.11	6.73

TABLE 3. Principal component contribution rate of PCA dimension reduction.

Order number	1	2	3	4	5
Principal components	8.2621	0.7473	0.1059	0.0072	0.0041
contribution rate (%)	90.4987	8.1855	1.1602	0.0008	0.0005
cumulative contribution rate (%)	90.4987	98.6842	99.8444	99.8452	99.8457

will lead to the input variable dimension being too high, and the number of samples is far lower than the spectrum feature dimension. These differences increase the complexity of the calibration model, allowing the model to easily exhibit overfitting; thus, feature dimension reduction is necessary. Through feature dimension reduction, a COD prediction model with better generalization ability is constructed.

1) PCA DIMENSION REDUCTION OF UV-VIS SPECTRUM

In this study, 240 water samples were collected, of which 160 water samples were used for the calibration set and 80 water samples were used for the prediction set. Before PCA dimension reduction, EEMD-Based denoising preprocessing is performed on the collected spectra, and the denoising results are shown in Figure 6(d).

The PCA algorithm is used to reduce redundant information for the input UV-Vis spectrum matrix (160 × 2048), and results are shown in Table 3. Due to the limited space, only the first five principal components and their contribution rate and cumulative contribution rate are listed in the table. The row of cumulative contribution rates in Table 3 shows the first principal component contributed 90.4987% of the contribution rate, and the cumulative contribution rate of the first three principal components has reached 99.8444%, which is sufficient to replace the information for the original full spectrum. Adding the fourth principal component does not significantly improve the cumulative contribution rate.

TABLE 4. Feature matrix by PCA dimension reduction.

Feature	1	2	3
Sample 1	-1.7343	-1.0740	-0.2131
Sample 2	5.2718	-0.0434	-0.6094
Sample 3	3.9525	-0.3305	0.1262
Sample 4∼157	…	…	…
Sample 158	-0.6330	-1.3865	-0.3302
Sample 159	1.8585	1.2719	-0.0021
Sample 160	0.9746	1.6353	0.2016

Therefore, the original UV-Vis spectrum matrix can be simplified to a 160 × 3 matrix after PCA dimension reduction, as shown in Table 4. Due to space constraints, only part of the sample dimension reduction results are shown. For the UV-Vis spectrum of 80 water samples in the prediction set, PCA dimension reduction was performed using the same transformation strategy as the calibration set.

2) UV-VIS SPECTRUM WAVELENGTH INTERVAL SELECTION BY IPLS

The spectrum preprocessed by the EEMD-Based method is divided into n = 10, 20, 30, 40 and 50 subintervals. The PLS regression model is constructed for each subinterval, and the root mean square error of cross validation (RMSECV) of the model is compared. A regression model with minimum RMSECV under different subinterval divisions is built. The correlation coefficient (R) and RMSECV/RMSEP were

TABLE 5. Results of the best iPLS model with different interval divisions.

Intervals	Optimum wavelength range (nm)	Latent variables	Feature numbers	Calibration set		Prediction set	
				R	RMSECV (mg/L)	R	RMSEP (mg/L)
10	193.91-291.39	3	205	0.8772	5.73	0.8623	5.91
20	193.91-242.78	3	103	0.8995	5.46	0.8811	5.68
30	193.91-226.52	3	69	0.8887	5.59	0.8739	5.77
40	218.86-243.26	4	52	0.8483	6.07	0.8227	6.38
50	213.58-232.74	4	41	0.8388	6.19	0.8259	6.34

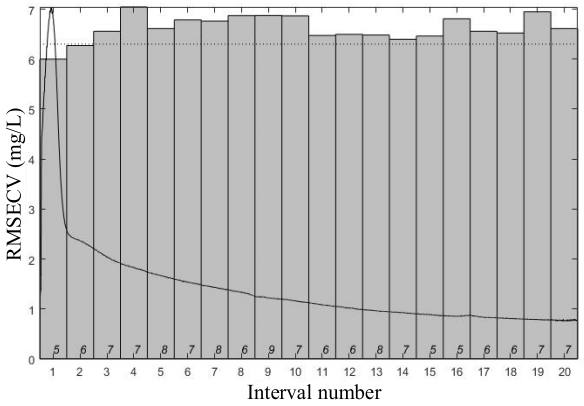


FIGURE 7. RMSECV value of each interval when the full spectrum is divided into 20 subintervals by iPLS.

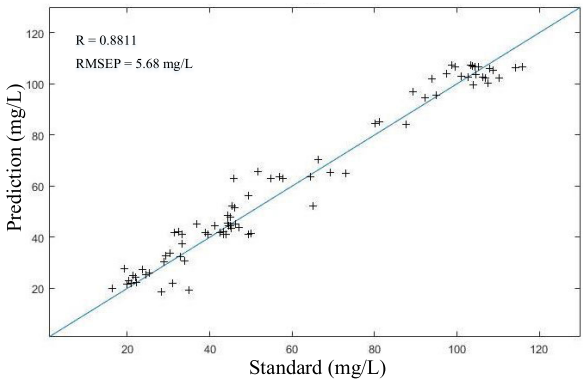


FIGURE 8. Scatter plot of COD prediction results in optimal subinterval of iPLS model and standard values of prediction set.

compared. Table 5 shows the best iPLS model results for the full spectrum with different interval divisions.

Table 5 shows the UV-Vis spectrum is divided into different subintervals, and the subintervals selected by the iPLS method and the prediction results of the PLS model built in the selected subinterval are different. When the full spectrum is divided into 20 subintervals, the RMSECV value of the iPLS model established in the first subinterval (wavelength range 193.91-242.78 nm) is the smallest, reaches 5.46 mg/L, and the number of optimal latent variables is 3, as shown in Figure 7. In this case, the model accuracy in the wavelength range is better than that in any other subinterval. Figure 8 shows the scatter plot of the COD value predicted by the optimal iPLS model and the COD standard value. The prediction set R is 0.8811, and RMSEP is 5.68 mg/L. Therefore, the wavelength range of 193.91-242.78 nm is the optimal wavelength subinterval after dimension reduction by the iPLS wavelength interval selection method.

3) SCARS WAVELENGTH SELECTION OF UV-VIS SPECTRUM

When using the SCARS algorithm to select feature wavelengths, it is necessary to first determine the optimal number of principal components (latent variables) in the PLS model. Initially, the maximum number of latent variables for the PLS model is set to 15, and the Monte Carlo sampling times is set to 3000. Figure 9 shows the RMSECV of the PLS model with different latent variables. Figure 9 shows that when the number of latent variables is 9, the minimum RMSECV is 6.2604 mg/L; thus, the optimal number of latent variables for the PLS model is 9.

Through many attempts to select a group of more appropriate SCARS parameters, this paper sets the Monte Carlo sampling times to 200, the number of latent variables to 9, and the number of cross validation groups to 10. Figure 10 shows that as the number of samplings increases, the number of optimized wavelength variables gradually decreases. The RMSECV value decreased continuously between 1 and 142 samplings, indicating that the variables removed in the screening process did not affect COD prediction. After 142 sampling, RMSECV began to rise, indicating that COD-related variables began to be removed, resulting in the increase in RMSECV. When the number of samplings reached 142, the RMSECV was the smallest (6.15 mg/L), and the corresponding feature wavelength subset was optimal. The subset contained 14 feature wavelengths: 195.83 nm, 197.27 nm, 202.07 nm, 204.95 nm, 205.43 nm, 206.39 nm, 210.23 nm, 212.15 nm, 215.98 nm, 216.46 nm, 216.94 nm, 220.29 nm, 238.48 nm, 241.35 nm. These wavelengths are the optimal wavelength features after dimension reduction of the SCARS wavelength selection method.

D. MODELS PERFORMANCE OF UV-VIS SPECTRUM AFTER FEATURE DIMENSION REDUCTION

This section compares the performance of different spectrum dimension reduction and modeling methods on COD prediction. Data dimension reduction methods include the PCA dimension reduction algorithm, iPLS wavelength interval selection algorithm and SCARS feature wavelength selection

TABLE 6. COD prediction performance of the combination of different spectrum dimension reduction and modeling methods.

Preprocessing methods	Data dimension	Models	Calibration set			Prediction set		
			R^2	RMSEC (mg/L)	Variance (mg^2)	R^2	RMSEP (mg/L)	Variance (mg^2)
EEMD-Based+PCA	3	DT	0.7899	14.67	14.18	0.7625	16.64	15.39
	3	Bagging	0.7996	14.04	11.76	0.7753	15.63	12.72
	3	RF	0.8276	12.20	8.76	0.8091	13.42	10.07
	3	Improved-Bagging	0.8450	11.06	7.81	0.8205	12.67	9.02
EEMD-Based+iPLS	69	DT	0.8020	13.88	12.41	0.7796	15.35	11.60
	69	Bagging	0.8753	9.08	8.07	0.8450	11.06	9.84
	69	RF	0.8814	8.68	6.72	0.8679	9.56	8.13
	69	Improved-Bagging	0.8906	8.08	5.75	0.8726	9.26	6.94
EEMD-Based+SCARS	14	DT	0.8441	11.12	8.90	0.8170	12.90	9.87
	14	Bagging	0.9138	6.56	7.74	0.8943	7.83	9.24
	14	RF	0.9315	5.40	5.79	0.9183	6.26	7.52
	14	Improved-Bagging	0.9459	4.46	4.97	0.9317	5.39	5.53

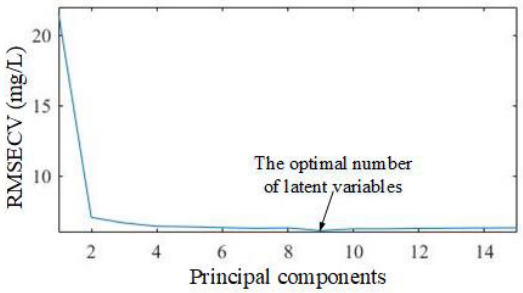


FIGURE 9. RMSECV with the number of principal components.

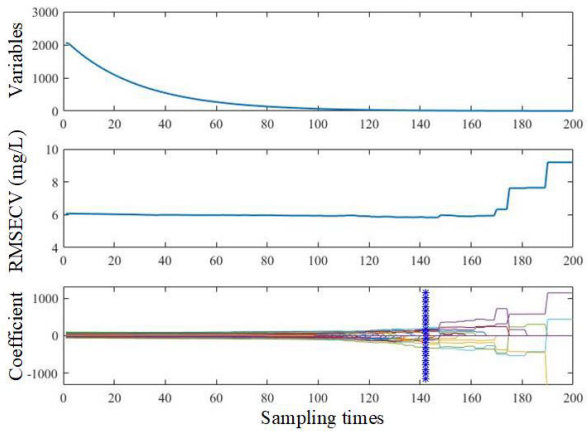


FIGURE 10. Feature optimization process of SCARS method.

algorithm. The modeling methods include DT, Bagging, RF and Improved-Bagging. The COD prediction performance of the combination of different spectrum dimension reduction and modeling methods is shown in Table 6.

According to the comparative analysis of Table 6, if the model is the same, the model built by the UV-Vis spectrum after SCARS dimension reduction achieves better performance than the other two dimension reduction methods. The prediction set R^2 is the largest, and the RMSEP and variance are the smallest, indicating that the SCARS dimension reduction method is more effective. In terms of the same dimension reduction method, the model built by Improved-Bagging is better than the other two modeling methods. Its prediction set R^2 is the largest, and the RMSEP and variance are the smallest, which also demonstrates the superiority of the proposed Improved-Bagging modeling method. Among all the built COD prediction models, the best COD prediction model is the UV-Vis spectroscopy model through EEMD-Based denoising, SCARS dimension reduction, and Improved-Bagging modeling. The prediction set R^2 of the model is 0.9317, and the RMSEP and variance are 5.39 mg/L and 5.53 mg^2, respectively.

E. COMPARISON OF PREDICTION PERFORMANCE OF ALL MODELS

The COD optimal prediction model obtained by each modeling method is compared in Tables 2 and 6. The R^2 and RMSEC/RMSEP of each model's calibration set and prediction set show good consistency. In terms of the same spectrum preprocessing method, the Improved-Bagging algorithm proposed in this paper has better prediction performance than the RF, Bagging algorithm and DT, which fully validates the superiority of the proposed Improved-Bagging algorithm. After spectrum preprocessing (denoising), the accuracy of the COD prediction model can be improved to some extent. Compared with full spectrum modeling, through feature dimension reduction, the prediction accuracy of the COD prediction model can be improved further. For the water COD prediction model of the experimental water samples, the full spectrum DT model of the raw spectrum has the worst prediction performance, and the Improved-Bagging model, which is denoised by the EEMD-Based algorithm and dimension reduction by SCARS, has the best prediction performance. The optimal model's (EEMD-Based+SCARS+Improved-Bagging model) prediction set R^2 is 0.9317, and the RMSEP and variance are 5.39 mg/L and 5.53 mg^2, respectively. The COD prediction values and standard value scatter plots of the optimal model on the prediction set are shown in Figure 11. Figure 11 shows that the model performs well on the prediction set, and the prediction values and the standard values

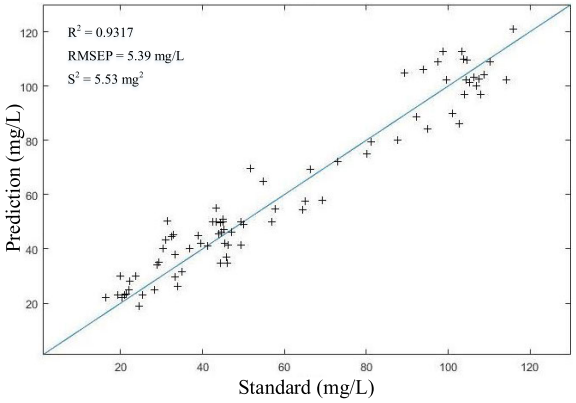


FIGURE 11. Scatter plot of COD prediction values and standard values based on EEMD-Based+SCARS+improved-bagging.

are similar, indicating that the prediction model built by this research has good robustness and adaptability, and can complete COD measurements in water accurately.

IV. CONCLUSION

Using the UV-Vis spectroscopy method to measure COD in water, an effective prediction model can be built using the UV-Vis spectrum of the water and the COD values. This paper proposed a model optimization method that used elastic net regression to improve the Bagging algorithm. Also, the input UV-Vis spectrum of the model was processed by spectrum preprocessing algorithms to further improve model performance. Results show that the prediction performance of Improved-Bagging algorithm is better than that of the traditional Bagging algorithm, and its prediction accuracy and generalization ability have been markedly improved. Appropriate denoising and feature dimension reduction methods can effectively reduce non-informative features, extract important features, and create a more accurate COD prediction model. Research shows that UV-Vis spectroscopy combined with the Improved-Bagging modeling method can perform COD measurements in water accurately. UV-Vis spectroscopy can thus be a new method for COD measurement in water.

REFERENCES

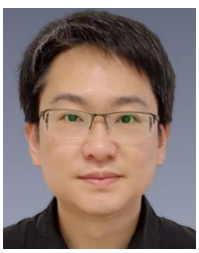
[1] J. Li, G. Luo, L. He, J. Xu, and J. Lyu, "Analytical approaches for determining chemical oxygen demand in water bodies: A review," *Crit. Rev. Anal. Chem.*, vol. 48, no. 1, pp. 47–65, Jan. 2018.
[2] A. Qasaimeh and Z. Al-Ghazawi, "Regression modeling for rapid prediction of wastewater BOD5," *Desalination Water Treat.*, vol. 201, pp. 165–172, Oct. 2020.
[3] B. Fleet, A. Y. W. Ho, and J. Tenygl, "A fully automated method for the determination of chemical oxygen demand," *Analyst*, vol. 97, no. 1154, pp. 321–333, May 1972.
[4] S. Ye, X. Chen, J. Wang, X. Wang, F. Wang, and D. Dong, "Rapid determination of water COD using laser-induced breakdown spectroscopy coupled with partial least-squares and random forest," *Anal. Methods*, vol. 10, no. 40, pp. 4847–4960, Oct. 2018.
[5] H. Cao, W. Qu, and X. Yang, "A rapid determination method for chemical oxygen demand in aquaculture wastewater using the ultraviolet absorbance spectrum and chemometrics," *Anal. Methods*, vol. 6, no. 11, pp. 3799–3803, Jun. 2014.
[6] B. L. Dinesha, S. Hiregoudar, U. Nidoni, K. T. Ramappa, A. Dandekar, and M. V. Ravi, "Comparison of chitosan based nano-adsorbents for dairy industry wastewater treatment through response surface methodology and artificial neural network models," *Water Sci. Technol.*, vol. 83, no. 5, pp. 1250–1264, Mar. 2021.
[7] S. Kim, M. Alizamir, K. Zounemat-Kermani, O. Kisi, and V. Singh, "Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea," *J. Environ. Manage.*, vol. 270, pp. 1–11, Sep. 2020.
[8] X. Chen, G. Yin, N, Zhao, T. Gan, R. Yang, M. Xia, C. Feng, Y. Chen, and Y. Huang, "Simultaneous determination of nitrate, chemical oxygen demand and turbidity in water based on UV-Vis absorption spectrometry combined with interval analysis," *Spectrochim. Acta, A*, vol. 244, pp. 1–14, Jan. 2021.
[9] H. Guo, J. J. Huang, B. Chen, X. Guo, and V. P. Singh, "A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery," *Int. J. Remote Sens.*, vol. 42, no. 5, pp. 1841–1866, Mar. 2021.
[10] M. Mrkva, "Evaluation of correlations between absorbance at 254 nm and COD of river waters," *Water Res.*, vol. 17, no. 2, pp. 231–235, Jan. 1983.
[11] M. P. Fernández, T. Knutz, and M. Barjenbruch, "Multi-parameter calibration of a UV/Vis spectrometer for online monitoring of sewer systems," *Water Sci. Technol.*, vol. 82, no. 5, pp. 927–939, Sep. 2020.
[12] C. P. Daniel, T. G. Juan, C. C. Fernando, and S. M. Juan, "Wastewater quality estimation through spectrophotometry-based statistical models," *Sensors*, vol. 20, no. 19, pp. 1–29, Oct. 2020.
[13] E. Kilic and N. Yucel, "Determination of spatial and temporal changes in water quality at Asi River using multivariate statistical techniques," *Turkish J. Fisheries Aquat. Sci.*, vol. 19, no. 9, pp. 727–737, Sep. 2019.
[14] M. Lepot, A. Torres, T. Hofer, N. Caradot, G. Gruber, J.-B. Aubin, and J.-L. Bertrand-Krajewski, "Calibration of UV/Vis spectrophotometers: A review and comparison of different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers," *Water Res.*, vol. 101, pp. 519–534, Sep. 2016.
[15] G. Langergraber, N. Fleischmann, and F. Hofstädter, "A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater," *Water Sci. Technol.*, vol. 47, no. 2, pp. 63–71, Jan. 2003.
[16] M. C. Sarraguça, A. Paulo, M. M. Alves, A. M. A. Dias, J. A. Lopes, and E. C. Ferreira, "Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy," *Anal. Bioanal. Chem.*, vol. 395, no. 4, pp. 1159–1166, Oct. 2009.
[17] J. Agustsson, O. Akermann, D. A. Barry, and L. Rossi, "Non-contact assessment of COD and turbidity concentrations in water using diffuse reflectance UV-Vis spectroscopy," *Environ. Sci., Processes Impacts*, vol. 16, no. 8, pp. 1897–1902, 2014.
[18] X. Chen, G. Yin, N. Zhao, R. Yang, M. Xia, C. Feng, Y. Chen, M. Dong, and W. Zhu, "Turbidity compensation method based on mie scattering theory for water chemical oxygen demand determination by UV-Vis spectrometry," *Anal. Bioanal. Chem.*, vol. 413, no. 3, pp. 877–883, Nov. 2020.
[19] Y. Hu, Y. Wen, and X. Wang, "Novel method of turbidity compensation for chemical oxygen demand measurements by using UV-Vis spectrometry," *Sens. Actuators B, Chem.*, vol. 227, pp. 393–398, May 2016.
[20] S. Fogelman, M. Blumenstein, and H. Zhao, "Estimation of chemical oxygen demand by ultraviolet spectroscopic profiling and artificial neural networks," *Neural Comput. Appl.*, vol. 15, nos. 3–4, pp. 197–203, Jun. 2006.
[21] E. M. Alves, R. J. Rodrigues, C. dos Santos Corrêa, T. Fidemann, J. C. Rocha, J. L. L. Buzzo, P. de Oliva Neto, and E. G. F. Núñez, "Use of ultraviolet–visible spectrophotometry associated with artificial neural networks as an alternative for determining the water quality index," *Environ. Monitor. Assessment*, vol. 190, no. 6, pp. 1–15, Jun. 2018.
[22] 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," *Anal. Methods*, vol. 10, no. 40, pp. 4879–4885, Oct. 2018.
[23] K. Wang, T. Chen, and R. Lau, "Bagging for robust non-linear multivariate calibration of spectroscopy," *Chemometric Intell. Lab. Syst.*, vol. 105, no. 1, pp. 1–6, Jan. 2011.
[24] D. Ghimire and J. Lee, "Extreme learning machine ensemble using bagging for facial expression recognition," *J. Inf. Process. Syst.*, vol. 10, no. 3, pp. 443–458, Sep. 2014.

[25] H.-C. Chan, A. Chattopadhyay, E. Y. Chuang, and T.-P. Lu, "Development of a gene-based prediction model for recurrence of colorectal cancer using an ensemble learning algorithm," *Frontiers Oncol.*, vol. 11, pp. 1–9, Feb. 2021.
[26] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electr. Power Syst. Res.*, vol. 192, pp. 1–14, Mar. 2021.
[27] M. Pirizadeh, N. Alemohammad, M. Manthouri, and M. Pirizadeh, "A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods," *J. Petroleum Sci. Eng.*, vol. 198, pp. 1–22, Mar. 2021.
[28] S. Tian, J. Zhang, L. Chen, H. Liu, and Y. Wang, "Random sampling-arithmetic mean: A simple method of meteorological data quality control based on random observation thought," *IEEE Access*, vol. 8, pp. 226999–227013, 2020.
[29] J. Lee and D.-W. Kim, "SCLS: Multi-label feature selection based on scalable criterion for large label set," *Pattern Recognit.*, vol. 66, pp. 342–352, Jun. 2017.
[30] Z. He, Z. Ma, M. Li, and Y. Zhou, "Selection of a calibration sample subset by a semi-supervised method," *J. Near Infr. Spectrosc.*, vol. 26, no. 2, pp. 87–94, Apr. 2018.
[31] Z. Yang, H. Xiao, L. Zhang, D. Feng, F. Zhang, M. Jiang, Q. Sui, and L. Jia, "Fast determination of oxide content in cement raw meal using NIR spectroscopy with the SPXY algorithm," *Anal. Methods*, vol. 11, no. 31, pp. 3936–3942, Aug. 2019.
[32] S. Minu and A. Shetty, "Prediction accuracy of soil organic carbon from ground based visible near-infrared reflectance spectroscopy," *J. Indian Soc. Remote Sens.*, vol. 46, no. 5, pp. 697–703, May 2018.
[33] W. Guo, Y. Huang, C. Han, and L. Yu, "Measurement of gain spectrum for Fabry–Pérot semiconductor lasers by the Fourier transform method with a deconvolution process," *IEEE J. Quantum Electron.*, vol. 39, no. 6, pp. 716–721, Jun. 2003.
[34] J. Zhang and Z. Zhao, "A method for determination of thiamethoxam in tea infusion by wavelet transform of self-enhanced absorption spectrum," *Food Anal. Methods*, vol. 10, no. 3, pp. 659–665, Mar. 2017.
[35] F. Zhou, C. Li, and Z. Hongqiu, "Development of a miniature spectrometer based on ultraviolet-visible spectroscopy for quantitative analysis of copper and cobalt," *IEEE Access*, vol. 8, pp. 131239–131247, 2020.
[36] J. Li, Y. Tong, L. Guan, S. Wu, and D. Li, "A UV-visible absorption spectrum denoising method based on EEMD and an improved universal threshold filter," *RSC Adv.*, vol. 8, no. 16, pp. 8558–8568, Feb. 2018.
[37] J. Melendez and G. Guarnizo, "Fast quantification of air pollutants by mid-infrared hyperspectral imaging and principal component analysis," *Sensors*, vol. 21, no. 6, pp. 1–14, Feb. 2021.
[38] W. Ju, C. Lu, Y. Zhang, W. Jiang, J. Wang, Y. Lu, and F. Hong, "Characteristic wavelength selection of volatile organic compounds infrared spectra based on improved interval partial least squares," *J. Innov. Opt. Health Sci.*, vol. 12, no. 2, pp. 1–19, Mar. 2019.
[39] J. Chen, C. Yang, H. Zhu, Y. Li, and W. Gui, "A novel variable selection method based on stability and variable permutation for multivariate calibration," *Chemometrics Intell. Lab. Syst.*, vol. 182, pp. 188–201, Nov. 2018.
[40] H. Pham and S. Olafsson, "Bagged ensembles with tunable parameters," *Comput. Intell.*, vol. 35, no. 1, pp. 184–203, Nov. 2019.
[41] A. Kadiyala and A. Kumar, "Applications of Python to evaluate the performance of bagging methods," *Environ. Prog. Sustain. Energy*, vol. 37, no. 5, pp. 1555–1559, Sep. 2018.
[42] S. Agarwal and C. R. Chowdary, "A-stacking and A-bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection," *Expert Syst. Appl.*, vol. 146, pp. 1–17, May 2019.
[43] D. Cho, C. Yoo, J. Im, Y. Lee, and J. Lee, "Improvement of spatial interpolation accuracy of daily maximum air temperature in urban areas using a stacking ensemble technique," *GISci. Remote Sens.*, vol. 57, no. 5, pp. 633–649, Jul. 2020.
[44] Q. Xu, X. Ding, C. Jiang, K. Yu, and L. Shi, "An elastic-net penalized expectile regression with applications," *J. Appl. Statist.*, vol. 48, pp. 1–26, Jul. 2020.

JINGWEI LI received the B.E. degree from the Zhengzhou University of Aeronautics, Zhengzhou, China, in 2013, and the M.E. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2017 and 2020, respectively. Since 2021, he has been a Lecturer with the College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou, China. His research interests include opto-mechatronics and artificial intelligence.

SISI PAN received the B.E. degree from the College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou, China, in 2019, where she is currently pursuing the M.E. degree.

JIE BIAN received the B.E. degree from the College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou, China, in 2020, where he is currently pursuing the M.E. degree.



WEI JIANG received the B.E. degree from Southwest Jiaotong University, Chengdu, China, in 2003, and the M.E. and Ph.D. degrees from The University of Texas at Arlington, TX, USA, in 2006 and 2009, respectively. Since 2020, he has been a Professor with the College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou, China. His research interests include power electronics and electromechanical energy conversion.