IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 24, 2020, accepted October 1, 2020, date of publication October 5, 2020, date of current version October 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028683

Effluent Quality Prediction of Papermaking Wastewater Treatment Processes Using Stacking Ensemble Learning

HONGBIN LIU[1,2], CHEN XIN[1], HAO ZHANG[1], FENGSHAN ZHANG[2], AND MINGZHI HUANG[3]

[1]Co-Innovation Center of Efficient Processing and Utilization of Forest Resources, Nanjing Forestry University, Nanjing 210037, China
[2]Laboratory for Comprehensive Utilization of Paper Waste of Shandong Province, Shandong Huatai Paper Company Ltd., Dongying 257335, China
[3]SCNU Environmental Research Institute, Guangdong Provincial Key Laboratory of Chemical Pollution and Environmental Safety and MOE Key Laboratory of Theoretical Chemistry of Environment, School of Environment, South China Normal University, Guangzhou 510006, China

Corresponding authors: Fengshan Zhang (htjszx@163.com) and Mingzhi Huang (mingzhi.huang@m.scnu.edu.cn)

This work was supported by the Guangdong Provincial Natural Science Foundation of China under Grant 2016A030306033.

ABSTRACT Advanced process modeling methods have been used for prediction and monitoring of key quality indices in wastewater treatment processes. However, single conventional models usually have limited precision accuracy when predicting the effluent indices in papermaking wastewater treatment processes. To achieve a better prediction accuracy and robustness, we propose a stacking ensemble learning (SEL) method which utilizes the advantages of the internal base-learning models. The method combines base-learning algorithms including partial least squares, support vector regression, and artificial neural networks with a meta-learning algorithm, which is a multiple-response linear regression in this work. To evaluate the model performance in practical applications, both real wastewater data and simulation wastewater data are used for modeling. The predicted effluent indices include effluent suspended solid (SS_{eff}), effluent chemical oxygen demand (COD_{eff}), effluent ammonia concentration (S_{NHeff}), and effluent nitrate concentration (S_{NOeff}). Compared with base-learning algorithms and other ensemble learning methods, the results demonstrate that SEL significantly improves the prediction accuracy and reduces the prediction errors, which provides a new way to achieve real-time monitoring of wastewater treatment processes.

INDEX TERMS Stacking ensemble learning, papermaking process modeling, effluent indices, prediction accuracy, wastewater treatment processes.

I. INTRODUCTION

The key to improving the quality management efficiency in wastewater treatment processes (WWTPs) heavily relies on the implementation of effective real-time monitoring of the effluent concentrations [1]. In recent years, hardware sensors in WWTPs have been exposed to a series of shortcomings in the monitoring process, such as significant time lags and high maintenance costs [2]. On the contrary, soft sensing methods can save measurement cost and improve monitoring quality in WWTPs, which is not only economical and reliable but also have a dynamic response [3]–[5]. For example, soft sensors can make real-time predictions for key WWTP variables including the concentrations of suspended solids (SS), chemical oxygen demand (COD), total phosphorus (TP), total nitrogen (TN), and daily sewage sludge. For some variables with periodic features, variables information can be divided into the periodic component and the residual component based on periodic analysis. Soft sensors combining with periodic analysis can achieve a better prediction performance [6]. Nowadays, increasing attention has been paid in advanced monitoring techniques in WWTPs [7], [8].

In recent years, the main conventional methods including partial least squares (PLS), support vector regression (SVR), and artificial neural networks (ANN) have been used for the prediction of wastewater effluent indices. However, the complex characteristics in wastewater treatment processes, such as nonlinearity, time-varying characteristic, and uncertainty make it difficult to obtain a satisfactory prediction result using these conventional methods. For example, the linear

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

method such as PLS usually shows bad modeling performance when nonlinear characteristic exists in the WWTP data [9]. Embedding kernel functions into PLS has been considered as an effective way for improving the prediction performance, in which the original data is transformed into a high-dimensional feature space by nonlinear mapping [10]. Compared with PLS, ANN has better nonlinear fitting performance and adaptive learning ability, which allows ANN to be successfully applied in WWTPs [11], [12]. However, there always exist low efficiency and local minimum problems in the ANN modeling process. To improve its modeling performance, the original ANN model needs to combine with other optimization methods such as fuzzy subtractive clustering and optimize fuzzy rule [13]. Although SVR has proven to work well under limited data sets, the computational cost is relatively large for large-scale data sets [14]. Aiming at curbing this limitation of the conventional SVR, the LSSVR algorithm has been proposed and it can provide a more effective solution by transforming the optimization problem into a set of linear equations problem [15]. Reducing the computing complexity, the improved LSSVR model can be successfully applied for predicting the wastewater effluent indices [16].

However, conventional models inevitably have some limitations. Without further optimization, none of the original models has the capability to interpret the complex characteristics of wastewater treatment processes. Moreover, there always exists a contradiction between model complexity and its generalization ability for limited samples. It is difficult and usually impossible for an over-optimized model to reach high prediction performance for all the data sets. Fortunately, it has been confirmed that ensemble learning methods could improve prediction accuracy without making the model too complicated [17]. Rather than transforming a single model and hoping the modified model to display its full potential, ensemble learning methods combine different types of models' advantages to achieve a better prediction performance. By considering various viewpoints of training data and multiple training principles, ensemble learning methods can be of great benefit for excavating the potential information between WWTP variables so the model's generalization ability is greatly increased. Ensemble learning methods have been an important direction of process modeling in future.

All of the conventional models are also called base-learners in ensemble learning methods which improve prediction ability by diversifying its base-learners [18]. At present, ensemble learning methods are generally divided into three types: Bagging, Boosting and stacking ensemble learning (SEL) [19]–[21]. For Bagging and Boosting, the emphasis is mainly placed on the data resampling technique [22], [23]. Thus, the diversity between all of the base-learners is focused on the multiformity of the training samples. Unlike Bagging and Boosting, SEL pays more attention to the diversity of training principles. More specifically, SEL integrates several distinct base-learning algorithms through a meta-learning algorithm, which aims to improve the prediction accuracy and generalization capability. From the viewpoint of diversity, SEL has a better prospect [24].

If the well-trained base-learning algorithms with higher prediction accuracy are prerequisites to SEL, the meta-learning algorithm determines the quality of SEL to a degree. Multi-response linear regression (MLR) has been confirmed as the most suitable meta-learning algorithm in SEL. Different from the voting or average methods in Bagging and Boosting, SEL uses MLR to further generalize the output values of the base-learning algorithms. Previous research has shown that the main superiority of MLR depends on its powerful function for reducing variance and bias of different base-learning algorithms [25].

In recent years, SEL has been successfully applied to the industrial field as a real-time prediction method. Divina developed an approach for short-term electricity consumption forecasting based on SEL. Compared with other conventional methods, the proposed method realized an efficient and promising way for solving the forecasting accuracy problem [26]. Khairalla proposed a modified SEL method to predict the average growth rate of total oil demand, which was superior to other benchmark methods in the aspects of error rate and directional accuracy [27]. Sun successfully applied SEL to the river ice forecasting field and obtained a better prediction result with higher accuracy [28]. In this work, a novel SEL algorithm is proposed to predict the wastewater effluent indices. This article is organized in the following manner. In Section 2, the training and testing processes are illustrated in more details, then the modeling principles of base-learning algorithms and meta-learning algorithm are briefly introduced. In Section 3, data processing and parameter optimization are illustrated first, and then other ensemble learning methods are introduced for comparison. To evaluate the prediction performance of SEL, both real wastewater data and simulation wastewater data are used for modeling. Finally, the conclusions are given in Section 4.

II. METHODS

SEL can improve estimation ability by combining the advantages of several different algorithms. In this work, SEL can be divided into two parts. The first part contains base-learning algorithms and the second part is the meta-learning algorithm. The base-learning algorithms should be efficient, diversiform and simple. As a prerequisite for building an ensemble model, strong learning ability of base-learning algorithm is helpful to improve the predicted performance of SEL. In terms of training principle, the diversity between each base learner should be as large as possible, which enables SEL to interpret data characteristics from multiple perspectives. Moreover, lower computational complexity will be beneficial to further improvement and optimization. Based on the above criteria, PLS, SVR, and ANN are chosen as the candidate base-learning algorithm. Compared with the simple average and voting strategy, MLR has access to a further generalization result, which is also the most commonly used meta-learning algorithm at present.

A. TRAINING AND TESTING PROCESS

For the training process, three base-learning algorithms including PLS, SVR, and ANN are defined as ζ_1, ζ_2, and ζ_3, respectively. As the meta-learning algorithm, MLR is defined as ξ. In the training process, five-fold cross validation approach is adopted.

Original training set $D= \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$ is randomly divided into five data sets with the same size as D_1, D_2, D_3, D_4, D_5 and $\mathbf{x}_i$ represents sample feature, y_i represents the target value. Among which, four data sets are used for training base-learners, and the remaining data set is used to verify the predicted performance, then optimizing the specific parameters of base-learner $h_i^{(j)}$ from the i-th learning algorithm ζ_i. The above-mentioned process needs to be carried out five times. Then, the well-trained base-learner $h_i^{(j)}$ is used to predict the samples which not participate in modeling, and the prediction result can be expressed as $h_i^{(j)}(\mathbf{x}_i)$. The secondary training set produced by the three base-learning algorithms can be expressed as $D' = \{(\mathbf{x}_i', \ y_i)\}_{i=1}^m$. Among them, $\mathbf{x}_i' = (h_1^{(j)}(\mathbf{x}_i), \ h_2^{(j)}(\mathbf{x}_i), \ h_3^{(j)}(\mathbf{x}_i))$, and y_i is still the target value in the original training set D. The D' is used for training the meta-learner $h' = \xi(D')$ by the meta-learning algorithm ξ.

For the testing process, five base-learners $h_i^{(j)}$ get corresponding prediction results for original test data set D_{test}, which can be described as $h_i^{(1)}(\mathbf{x}_t)$, $h_i^{(2)}(\mathbf{x}_t)$, $h_i^{(3)}(\mathbf{x}_t)$, $h_i^{(4)}(\mathbf{x}_t)$, and $h_i^{(5)}(\mathbf{x}_t)$. By averaging all of the prediction results, a prediction vector is obtained as follows:

$$\overline{h_i}(\mathbf{x}_t) = \sum_{j=1}^{n=5} h_i^{(j)}(\mathbf{x}_t)/5 \tag{1}$$

Three base-learning algorithms produce three prediction vectors, which can be described as $\mathbf{x}_i'' = (\overline{h_1}(\mathbf{x}_t), \ \overline{h_2}(\mathbf{x}_t), \ \overline{h_3}(\mathbf{x}_t))$. Because y_i' in D_{test} has not changed, the secondary test set can be expressed as $D'' = \{(\mathbf{x}_i'', \ y_i')\}_{i=1}^n$. By applying the D'' into the meta-learner h', the final prediction result $H(x) = h'(D'')$ is obtained. The implementation process of SEL and the formation process of the secondary data set is shown in Figures 1 and 2, respectively.

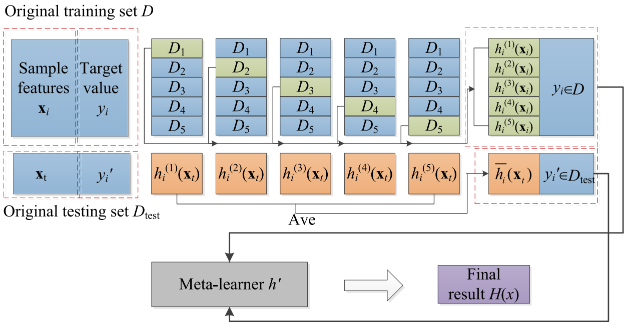

Original training set D
Sample features x_i
Target value y_i
D1 D2 D3 D4 D5
h_i^(1)(x_i) h_i^(2)(x_i) h_i^(3)(x_i) h_i^(4)(x_i) h_i^(5)(x_i)
y_i∈D
x_t
y_i'
h_i^(1)(x_t) h_i^(2)(x_t) h_i^(3)(x_t) h_i^(4)(x_t) h_i^(5)(x_t)
h̄_i(x_t)
y_i'∈D_test
Original testing set D_test
Ave
Meta-learner h'
Final result H(x)

FIGURE 1. Scheme of stacking ensemble learning.

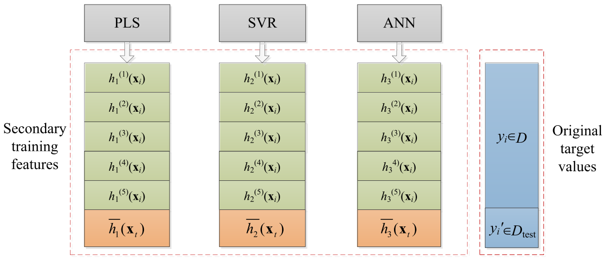

PLS
SVR
ANN
h_1^(1)(x_i) h_1^(2)(x_i) h_1^(3)(x_i) h_1^(4)(x_i) h_1^(5)(x_i) h̄_1(x_t)
h_2^(1)(x_i) h_2^(2)(x_i) h_2^(3)(x_i) h_2^(4)(x_i) h_2^(5)(x_i) h̄_2(x_t)
h_3^(1)(x_i) h_3^(2)(x_i) h_3^(3)(x_i) h_3^(4)(x_i) h_3^(5)(x_i) h̄_3(x_t)
Secondary training features
y_i∈D
y_i'∈D_test
Original target values

FIGURE 2. Formation process of secondary data set.

B. BASE-LEARNING ALGORITHMS

1) PARTIAL LEAST SQUARES

Partial least squares algorithm has been widely used in regression, mainly by finding the reasonable latent variables. Assume that input matrix is $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_q]_{n\times q}$ and output matrix is $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdot\cdot\cdot, \mathbf{y}_p]_{n\times p}$. Among them, n is the number of the samples, q and p represents the number of input variables and output variables, respectively. To better exploit variance structures of process, $\mathbf{X}$ and $\mathbf{Y}$ are projected into a lower dimensional space as follows:

$$\begin{cases} \mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \end{cases} \tag{2}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are loading matrices, $\mathbf{T}$ and $\mathbf{U}$ are latent matrices which carry enough variation information in $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{E}$ and $\mathbf{F}$ represent residual matrices.

2) SUPPORT VECTOR REGRESSION

Support vector regression algorithm is the modified version of support vector machine (SVM) which can be used to solve linear and non-linear regression tasks. It performs better for small-scale data set by using the structural risk minimization principle. The key to SVR lies in finding suitable mapping function $\varphi(x)$ between input vectors and output vectors. With the help of mapping function $\varphi(x)$, the training data $T = \{(\mathbf{x}_1, \ y_1), (\mathbf{x}_2, \ y_2), \cdot\cdot\cdot, (\mathbf{x}_i, \ y_i), \cdots, (\mathbf{x}_n, \ y_n)\}$ is mapped into a high dimensional feature space. Then, an optimized regression function can be constructed in the new feature space as follows:

$$f(x) = \mathbf{w}^{\mathrm{T}}\varphi(x) + b \tag{3}$$

where $\mathbf{w}$ is the weight vector and b represents the bias term. By introducing insensitive loss function ε, the errors within a specified tolerance range can be neglected. Then, the slack variables ξ_i and ξ_i^* are added into Equation (3). The regression task can be transformed into an optimization problem as follows:

$$\begin{cases} y_i - \mathbf{w}^{\mathrm{T}}\varphi(\mathbf{x}) - b \le \xi_i + \varepsilon, & i = 1, 2, \cdot\cdot\cdot, n \\ \mathbf{w}^{\mathrm{T}}\varphi(\mathbf{x}) + b - y_i \le \xi_i^* + \varepsilon, & i = 1, 2, \cdots, n \\ \xi_i\xi_i^* \ge 0, & i = 1, 2, \cdots, n \end{cases} \tag{4}$$

By adding Lagrangian coefficients β_i and β_i^*, the weight vector **w** can be expressed as follows:

$$\mathbf{w} = \sum_{i=1}^{n} (\beta_i - \beta_i^*)\varphi(\mathbf{x}) \tag{5}$$

Finally, the SVR can be defined as the following regression function:

$$f(x) = \sum_{i=1}^{n} (\beta_i - \beta_i^*)K(\mathbf{x}, \mathbf{x}') + b \tag{6}$$

where $K(\mathbf{x},\ \mathbf{x}')$ corresponds to the kernel function. The radial basis function (RBF) was used in this work, which can be expressed as follows:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|^2) \tag{7}$$

where γ is kernel parameter which is used to control the radial range of the kernel function.

3) ARTIFICIAL NEURAL NETWORKS

Artificial neural networks algorithm has a powerful self-learning and self-adaptive ability to reduce prediction error. Figure 3 shows the classic topology structure of artificial neural networks with three layers including input layer, hidden layer, and output layer.

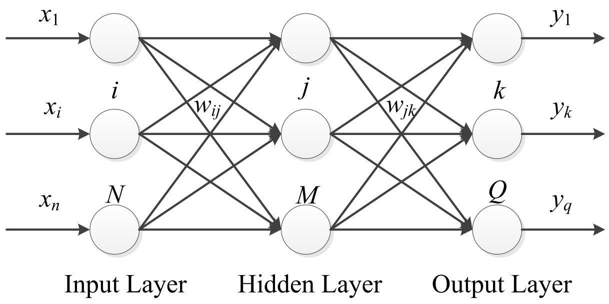


FIGURE 3. **Topology structure of artificial neural networks.**

Firstly, it is assumed that the number of input layer nodes, hidden layer nodes, and output layer nodes is N, M, and Q, respectively. Moreover, x_i represents the i-th input value in the input layer, w_{ij} represents the weight value from the input layer to the hidden layer, w_{jk} represents the weight value from the hidden layer to the output layer and θ_j corresponds to the threshold value. The input information is firstly propagated forward from the input layer to the hidden layer then the error is propagated backward according to weight values and transfer function $f(x)$. Through repeated correction of weight values, the predicted values are gradually closer to the actual values. During the whole process, x_j' represents the output value of the j-th neuron in the hidden layer which can be defined as follows:

$$x_j' = f(\sum_{i=1}^{N} w_{ij}x_i - \theta_j), \quad j = 1, 2, \cdots M \tag{8}$$

The output value of the k-th neuron in the output layer is written as y_k, which can be defined as follows:

$$y_k = f(\sum_{j=1}^{M} w_{jk}x_j' - \theta_k), \quad k = 1, 2, \cdots Q \tag{9}$$

The network weights and the thresholds need to be adjusted depending on the minimum mean square error (MSE) which is defined as follows:

$$\text{MSE} = \frac{1}{l}\sum_{n=1}^{l} (t_n - y_n)^2 \tag{10}$$

where l is the number of training samples, t_n corresponds to the expected value of the neural node and y_n denotes the predicted value. The MSE will be reduced to certain extent by repeating the back-propagation mechanism. Another stopping criterion of the algorithm is that the training time reaches its maximum.

C. META-LEARNING ALGORITHM

As the most efficient meta-learning algorithm, MLR combines weight values with output values coming from base-learning algorithms to obtain a further generalization result. The specific form can be expressed as follows:

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_kx_k \tag{11}$$

where y denotes the final prediction result, w_k corresponds to the weight value, and x_k corresponds to the output value from the base-learning algorithm. All of the weight values are obtained in the training process. By choosing appropriate weight values, MLR makes the square sum of the difference between the predicted values and the real values as small as possible. The formula is shown as follows:

$$e = \sum_{i=1}^{n} (y^{(i)} - \sum_{j=1}^{k} w_jx_j^{(i)})^2 \tag{12}$$

where n is the number of samples, k is the number of base-learning algorithms, $y^{(i)}$ represents the real value of the i-th sample, and $\sum_{j=1}^{k} w_jx_j^{(i)}$ represents the predicted value of the i-th sample.

D. MODELING PERFORMANCE INDICES

To determine whether the final results have a better prediction accuracy, three evaluation indices including determinate coefficient (R^2), mean absolute percentage error (MAPE), and root mean square error (RMSE) are used in this work, which is calculated from Equations (13), (14), and (15), respectively.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} \tag{13}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{14}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \Big/ N} \tag{15}$$

where y_i is measured value, $\hat{y}_i$ is predicted value, and $\bar{y}_i$ is the mean value of y_i.

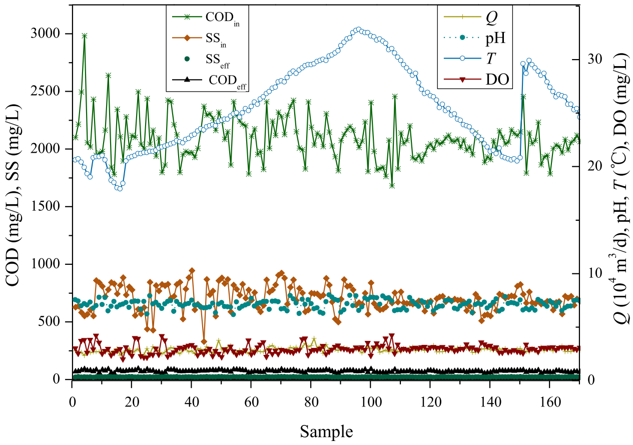


FIGURE 4. Papermaking WWTP data.

TABLE 1. Mean and standard deviation values of the papermaking WWTP data.

Variables	Mean	Standard Deviation	Unit
Q	2.75	0.23	10^4m^3/d
SS$_{\mathrm{in}}$	705.32	105.47	mg/L
SS$_{\mathrm{eff}}$	22.12	1.20	mg/L
pH	7.01	0.41	/
T	25.40	3.94	℃
DO	2.84	0.41	mg/L
COD$_{\mathrm{in}}$	2090.76	187.62	mg/L
COD$_{\mathrm{eff}}$	72.31	8.24	mg/L

III. RESULTS AND DISCUSSION

A. DESCRIPTION OF TWO DATA SETS

The modeling data used in this work include real wastewater data and simulation wastewater data. The actual wastewater data were collected from a papermaking WWTP in China. The data collection system includes various probes, signal acquisition card and power relay output board. As shown in Figure 4, the data include the following variables: wastewater flow rate (Q), influent suspended solid (SS$_{\mathrm{in}}$), effluent suspended solid (SS$_{\mathrm{eff}}$), pH, temperature (T), influent chemical oxygen demand (COD$_{\mathrm{in}}$), effluent chemical oxygen demand (COD$_{\mathrm{eff}}$), and dissolved oxygen (DO). The statistics are listed in Table 1. Among the eight variables, SS$_{\mathrm{eff}}$ and COD$_{\mathrm{eff}}$ are response variables (target values) which need to be controlled. Both COD$_{\mathrm{eff}}$ and SS$_{\mathrm{eff}}$ variables met the Chinese national effluent release standards. The rest of variables (sample features) are explanatory variables which are closely related to the response variables. The specific correlation coefficients between response variables and explanatory variables are presented in Table 2. These variables determine the degradation efficiency of pollutants by affecting the microbial activity of active sludge then indirectly influence the trend of SS$_{\mathrm{eff}}$ and COD$_{\mathrm{eff}}$.

TABLE 2. Correlation coefficients between response variables and explanatory variables.

Response variables	Explanatory variables					
	COD$_{\mathrm{in}}$	SS$_{\mathrm{in}}$	DO	T	pH	Q
COD$_{\mathrm{eff}}$	0.80	0.28	-0.41	-0.14	-0.15	0.10
SS$_{\mathrm{eff}}$	0.43	0.67	-0.39	0.09	0.08	0.03

The simulation wastewater data were generated from benchmark simulation model no. 1 (BSM1) which is designed to simulate a wastewater treatment system. As shown in Figure 5, BSM1 consists of five biological reactors and one secondary sedimentation tank. BSM1 can provide diverse control strategies under three weather conditions including dry weather, rainy weather and storm weather. In this article, the dry weather data were used for simulation. The sampling period is 14 days and the sampling interval is 15 minutes. Finally, 1345 samples were generated and each of the samples includes ten variables. Table 3 displays the specific process variables among which effluent ammonia concentration (S$_{\mathrm{NHeff}}$) and effluent nitrate concentration (S$_{\mathrm{NOeff}}$) are response variables and the rest are explanatory variables.

At the beginning of the modeling process, Jolliffe's three parameters method was used to detect the outliers [29]. Then, the processed data were normalized to ensure the values of different features have the same dimension. The corresponding transformation formula is as follows:

$$\mathbf{X}' = \frac{\mathbf{X} - \mu}{\sigma} \tag{16}$$

where $\mathbf{X}$ corresponds to the original sample vector, μ represents the mean value of the original sample vector, and σ is the standard deviation of the original sample data. Then, all the data were divided into training data and test data. For the real wastewater data, the first 120 samples were used as training data and the rest 50 samples were used as test data. For the simulation wastewater data, the first 672 samples were used for training, and the remaining 673 samples were used for testing.

B. PARAMETER OPTIMIZATION

Although PLS has been successfully applied in many industrial processes, the calculation process for obtaining latent variables is still lack of a uniform standard. In this work, with the help of the index of variable importance in the projection scores [30], the scores of latent variables higher than 1 are considered as valuable latent variables for modeling.

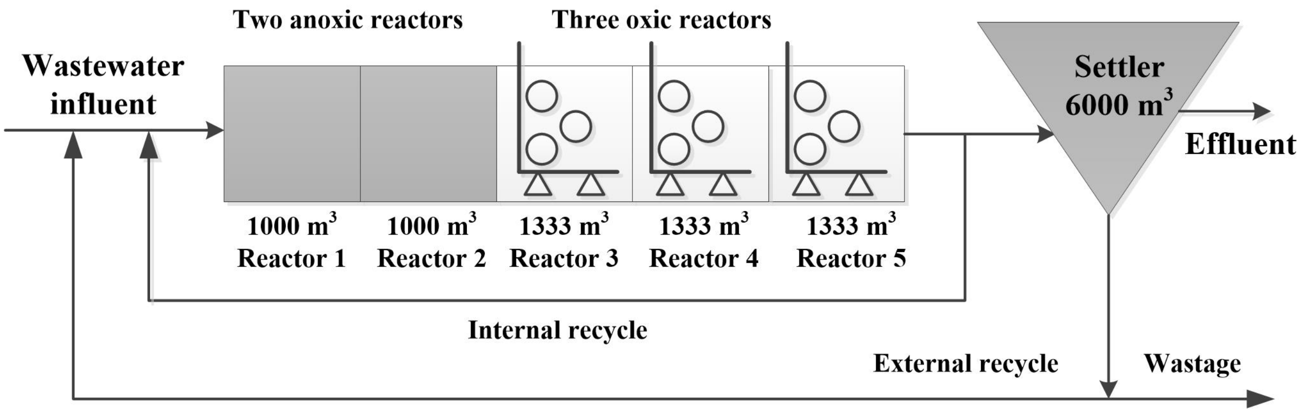


FIGURE 5. BSM1 layout.

TABLE 3. Process variables in BSM1.

Variables	Description	Unit
Q_{in}	Influent flow rate	m^3/d
S_{NHin}	Influent ammonia concentration	mg N/m^3
S_{NO2}	Nitrate concentration of the second reactor	mg N/m^3
S_{O3}	Dissolved oxygen concentration of third reactor	mg COD/m^3
S_{O4}	Dissolved oxygen concentration of fourth reactor	mg COD/m^3
T_{SS4}	Total suspended solid concentration of the fourth reactor	mg SS/m^3
KLa_5	Oxygen transfer coefficient of the fifth reactor	1/d
Q_{intr}	Internal recycle rate	m^3/d
S_{NHeff}	Effluent ammonia concentration	mg N/m^3
S_{NOeff}	Effluent nitrate concentration	mg N/m^3

The Kernel function is of key importance in SVR. Among all of the kernel functions, RBF is the most commonly used one. Its main advantage is that even if prior knowledge of the data is absent, the prediction effect is still robust. Therefore, RBF was adopted in this work. The kernel parameter γ directly affects the complexity of data distribution in the higher dimensional space, and hence needs to be confirmed first. If γ is too small, it will perform like linear kernel function. On the contrary, it will perform like polynomial kernel function. Another important parameter of SVR is the regularization parameter C also known as penalty factor, which is used to achieve a compromise between empirical risk and confidence level. If C value is too high, the SVR tends to result in over-fitting phenomenon. On the contrary, smaller C value is often accompanied with under-fitting problems. The grid searching method [31] was adopted to search for a suitable combination of C and γ in this work.

The key to constructing a well-performed ANN lies in finding a suitable network structure and activation function. Considering the data size is not large, the three-layer network structure was chosen in this work. The number of hidden layer nodes, the activation function of hidden layer, and output layer were selected according to the actual prediction performance. The specific parameters of the base-learning algorithms are shown in Table 4.

C. ENSEMBLE LEARNING MODELS COMPARISON

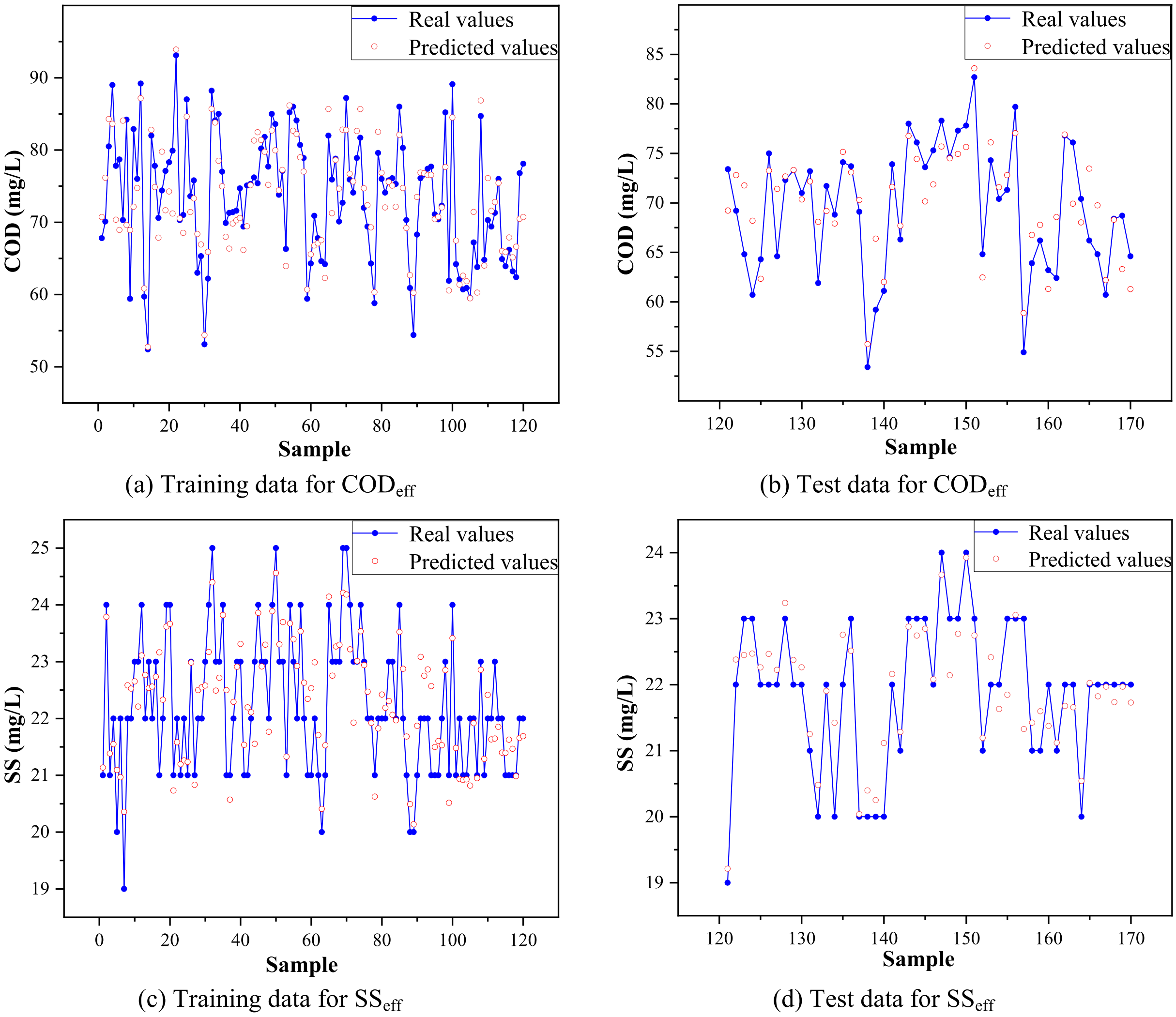
To compare SEL with other types of ensemble learning methods in terms of prediction accuracy, random forest (RF) model based on bagging algorithm and adaptive boosting (AdaBoost) model based on boosting algorithm were constructed by using the same wastewater data. The base-learners of RF and AdaBoost are both decision trees. Considering that prediction of wastewater effluent indices is essentially a regression task, the CART (classification and regression tree) is selected as the decision tree which is not pruned during the modeling process. The main difference between RF and AdaBoost is that the former integrates decision trees in a parallel manner, while the latter integrates decision trees in a serial manner. In terms of RF, the main tuning parameters include the number of decision trees n_{tree} and random variables used at each split m_{try}. In terms of AdaBoost, the main tuning parameters include the number of iterations n_T and learning rate of base-learners ν. All of above-mentioned tuning parameters were obtained by grid searching method.

D. RESULTS AND DISCUSSION

To directly observe the prediction performance of SEL, we provide prediction figures for real wastewater data and simulation wastewater data. As shown in Figures 6 and 7, the data points of the two cases are both well modeled

TABLE 4. Specific parameters of base-learning algorithms.

Methods	Parameters	COD_{eff}	SS_{eff}	S_{NHeff}	S_{NOeff}
PLS	Number of latent variables	3	3	2	2
SVR	Kernel function	RBF	RBF	RBF	RBF
	Kernel parameter γ	56	2.8	17	21
	Regularization parameter C	118	60	11	13
ANN	Number of neural network layers	3	3	3	3
	Number of hidden layers	1	1	1	1
	Input layer nodes	6	6	8	8
	Hidden layer nodes	2	3	2	2
	Output layer nodes	1	1	1	1
	Activation function of hidden layer	tansig	tansig	tansig	tansig
	Activation function of output layer	purelin	purelin	purelin	purelin
	Max training time	1000	1000	500	500



(a) Training data for COD_{eff}

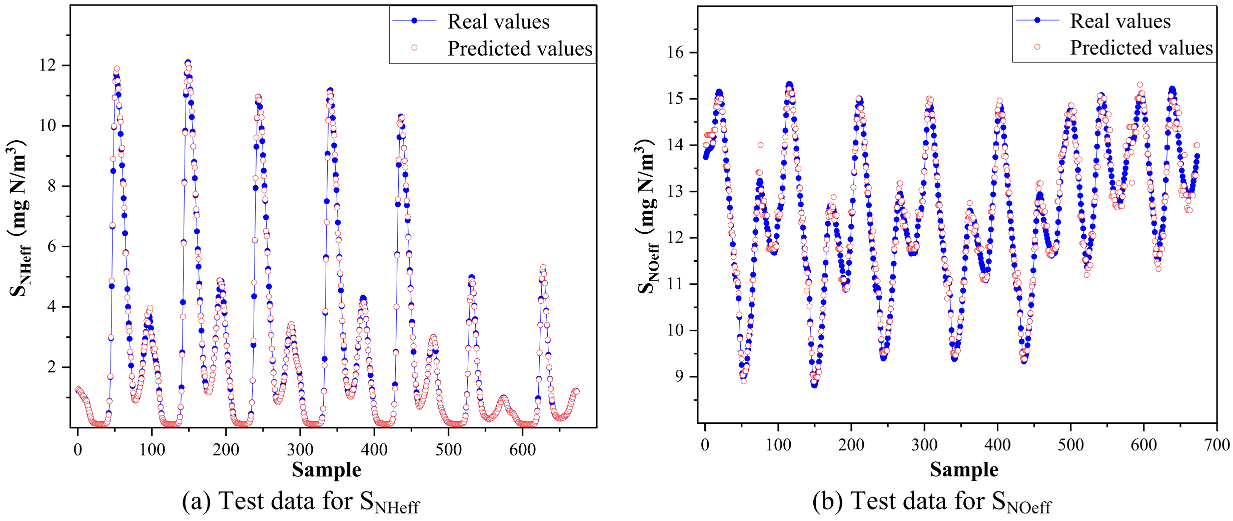
(b) Test data for COD_{eff}

(c) Training data for SS_{eff}

(d) Test data for SS_{eff}

FIGURE 6. Prediction results of SS $_{eff}$ and COD $_{eff}$ using stacking ensemble learning.

using SEL. The quantitative evaluation results including RMSE, MAPE and R^2 for PLS, SVR, ANN, RF, AdaBoost and SEL are listed in Tables 5 and 6. It can be seen that SEL achieves the highest prediction accuracy for two cases. For COD_{eff}, compared with the base-learning algorithms of PLS, SVR, and ANN, the RMSE of SEL is reduced by



(a) Test data for S_{NHeff}

(b) Test data for S_{NOeff}

FIGURE 7. Prediction results of S_{NHeff} and S_{NOeff} using stacking ensemble learning.

TABLE 5. Comparison of modeling results for COD_{eff} and SS_{eff}.

Methods		COD_{eff}			SS_{eff}		
		RMSE	MAPE (%)	R^2	RMSE	MAPE (%)	R^2
PLS	Training	4.32	4.56	0.67	0.79	2.41	0.66
	Testing	4.94	5.83	0.58	0.83	2.93	0.56
SVR	Training	5.11	5.32	0.69	0.72	2.58	0.70
	Testing	4.53	5.23	0.68	0.76	2.67	0.58
ANN	Training	4.56	5.86	0.71	0.68	2.13	0.70
	Testing	4.97	5.92	0.57	0.79	2.78	0.64
RF	Training	3.98	5.18	0.78	0.66	1.92	0.67
	Testing	4.76	5.45	0.67	0.78	2.69	0.65
AdaBoost	Training	4.33	5.37	0.72	0.72	2.03	0.67
	Testing	4.84	5.59	0.68	0.80	2.77	0.63
SEL	Training	4.14	5.06	0.74	0.76	2.17	0.69
	Testing	4.25	5.03	0.72	0.71	2.37	0.68

TABLE 6. Comparison of modeling results for S_{NHeff} and S_{NOeff}.

Models	S_{NHeff}			S_{NOeff}		
	RMSE	MAPE (%)	R^2	RMSE	MAPE (%)	R^2
PLS	0.85	2.97	0.76	0.70	2.83	0.78
SVR	0.61	2.54	0.83	0.52	2.40	0.82
ANN	0.60	2.59	0.84	0.57	2.66	0.81
RF	0.47	2.37	0.83	0.44	2.19	0.82
AdaBoost	0.43	2.35	0.85	0.45	2.24	0.81
SEL	0.31	1.96	0.88	0.30	1.87	0.86

13.97%, 6.18%, and 14.49%, respectively. For SS_{eff}, the RMSE of SEL is reduced by 14.46%, 6.58%, and 10.13%, respectively. In terms of R^2, the prediction accuracy of SEL is also improved significantly range from 5.88%-26.32%, 6.25%-21.43% for COD_{eff} and SS_{eff}. Meanwhile, SEL also demonstrated its superiority of ensemble learning compared with RF and AdaBoost, specifically with the minimum RMSE value (4.25) and the maximum R^2 (0.72) for COD_{eff}, the minimum RMSE value (0.71) and the maximum R^2 (0.68) for SS_{eff}.

For S_{NHeff}, compared with the base-learning algorithms of PLS, SVR, and ANN, the RMSE of SEL is reduced by 63.53%, 49.18%, 48.33%, respectively. For S_{NOeff}, the RMSE of SEL is reduced by 57.14%, 42.31%, 47.37%, respectively. In terms of R^2, the prediction accuracy of SEL is also improved significantly range from 4.76%-15.79%, 4.88%-10.26% for S_{NHeff} and S_{NOeff}. Compared with RF and AdaBoost, SEL has the best prediction results, specifically with the minimum RMSE value (0.31) and the maximum R^2 (0.88) for S_{NHeff}, the minimum RMSE value (0.30) and the maximum R^2 (0.86) for S_{NOeff}.

Considering the fact that base-learning algorithms have their algorithm learning preference, their predictive capability may be limited for the papermaking wastewater effluent indices. Depending on different circumstances, SEL firstly uses the original training set to train base-learning algorithms, then uses the secondary training set generated by base-learning algorithms to train the meta-learning algorithm. In other words, the output values of base-learning algorithms are the input features of meta-learning algorithm. Based on the prediction results of the base-learning algorithms, SEL achieves a better generalization for the effluent indices. In addition, compared with other ensemble learning methods, SEL has higher prediction accuracy. From the viewpoint of theoretical analysis, the superior prediction performance of SEL can be mainly attributed to the following points:

(1) Through exploring the feature space from different perspectives, SEL can provide a more comprehensive analysis for complex characteristics in the wastewater data;

(2) The SEL makes use of the advantages of different base-learning algorithms while gets rid of their relatively worse prediction drawback, so as to reduce the risk of trapping into local minima;

(3) Compared with base-learning algorithms, SEL expends the hypothesis space in modeling process, which may be closer to the real hypothesis of the wastewater treatment process.

(4) From the viewpoint of the diversity and integration of base-learning algorithms, SEL can combine different types of base-learners corresponding to various base-learning algorithms compared with other ensemble learning methods, and constructing a meta-learning algorithm is a more reasonable way than adopting statistically averaging, which makes SEL a better capacity of generalization.

Although SEL has tremendous potential for improving the prediction accuracy in wastewater effluent data, the step of parameter optimization will be a time-consuming work. In this work, there are three groups of parameters need to be determined, which results in an increasing running time of SEL as shown in Table 7. The running time of SEL mainly spends on the ANN base-learning algorithm. Because ANN uses the back-propagation mechanism for optimizing the hyper-parameters, the network weights and the thresholds in each neuron need to be revised several times to finally reach the precision requirement, which immediately causes the increment of the total running time. In general, ensemble learning methods take more time than base-learning algorithms for training and prediction. Among ensemble learning methods, the running time of RF and AdaBoost is relatively shorter compared with SEL. This is mainly because the base-learners of RF and AdaBoost are relatively simple which makes RF and AdaBoost have low computational complexity.

TABLE 7. Comparison of running time for COD_{eff} and S_{NHeff}.

Methods	COD_{eff}		S_{NHeff}	
	Training time (s)	Prediction time (s)	Training time (s)	Prediction time (s)
PLS	0.080	0.003	0.101	0.013
SVR	0.120	0.011	0.183	0.024
ANN	23.078	1.677	26.454	1.895
RF	12.696	1.334	13.327	1.482
AdaBoost	17.463	1.485	19.522	1.656
SEL	24.132	1.856	28.621	2.031

With the ever-increasing amounts of data in WWTPs, the complexity and running time will directly affect its practical application and further development. In future work, besides optimizing the execution efficiency of SEL, more combinations of base-learning algorithms will be studied. Various machine learning algorithms, such as Gaussian process regression, relevance vector machine, gene expression programming, and evolutionary polynomial regression, can be integrated with the existing SEL. Meanwhile, the data set should be expended to further validate the prediction accuracy of SEL.

IV. CONCLUSION

In this work, the stacking ensemble learning is used for modeling the papermaking wastewater treatment process. By predicting the COD_{eff} and SS_{eff} from real wastewater data as well as S_{NHeff} and S_{NOeff} from wastewater simulation data, the proposed SEL method has successfully interpreted the complex characteristics of wastewater data. The prediction hypothesis space of SEL is more comprehensive for the wastewater treatment process. During the prediction process, the meta-learning algorithm enables SEL to obtain a further generalization result, which makes use of the advantages of each base-learning algorithm while avoiding the risk of trapping into local minima. The simulation results show that SEL has a better prediction ability compared with base-learning algorithms including PLS, SVR, ANN and other ensemble learning methods including RF and AdaBoost. Therefore, applying SEL into the real-time monitoring of wastewater treatment processes has practical reference value and realistic signification. Future work will be focused on the improvement in execution efficiency. Besides, more base-learning algorithms should be tried for stacking ensemble learning and sample size will be further expanded to verify the applicability of the proposed method.

ACKNOWLEDGMENT

(Hongbin Liu and Chen Xin contributed equally to this work.)

REFERENCES

[1] H. Liu, K.-H. Chang, and C. Yoo, "Multi-objective optimization of cascade controller in combined biological nitrogen and phosphorus removal wastewater treatment plant," *Desalination Water Treatment*, vol. 43, nos. 1–3, pp. 138–148, Apr. 2012.

[2] H. Haimi, F. Corona, M. Mulas, L. Sundell, M. Heinonen, and R. Vahala, "Shall we use hardware sensor measurements or soft-sensor estimates? Case study in a full-scale WWTP," *Environ. Model. Softw.*, vol. 72, pp. 215–229, Oct. 2015.

[3] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, Apr. 2009.

[4] C. Shang, F. Yang, D. Huang, and W. Lyu, "Data-driven soft sensor development based on deep learning technique," *J. Process Control*, vol. 24, no. 3, pp. 223–233, Mar. 2014.

[5] H. Liu, C. Yang, B. Carlsson, S. J. Qin, and C. Yoo, "Dynamic nonlinear partial least squares modeling using Gaussian process regression," *Ind. Eng. Chem. Res.*, vol. 58, no. 36, pp. 16676–16686, Aug. 2019.

[6] X. Yang, Y. Zou, J. Tang, J. Liang, and M. Ijaz, "Evaluation of short-term freeway speed prediction based on periodic analysis using statistical models and machine learning models," *J. Adv. Transp.*, vol. 2020, pp. 1–16, Jan. 2020.

[7] A. Shokry, P. Vicente, G. Escudero, M. Pérez-Moya, M. Graells, and A. Espuña, "Data-driven soft-sensors for online monitoring of batch processes with different initial conditions," *Comput. Chem. Eng.*, vol. 118, pp. 159–179, Oct. 2018.

[8] F. A. A. Souza, R. Araújo, and J. Mendes, "Review of soft sensor methods for regression applications," *Chemometric Intell. Lab. Syst.*, vol. 152, pp. 69–79, Mar. 2016.

[9] Y. Wang, H. Cao, Y. Zhou, and Y. Zhang, "Nonlinear partial least squares regressions for spectral quantitative analysis," *Chemometric Intell. Lab. Syst.*, vol. 148, pp. 32–50, Nov. 2015.

[10] H. Liu, C. Yang, M. Huang, and C. Yoo, "Soft sensor modeling of industrial process data using kernel latent variables-based relevance vector machine," *Appl. Soft Comput.*, vol. 90, pp. 1–10, May 2020.

[11] H. Mingzhi, Y. Ma, W. Jinquan, and W. Yan, "Simulation of a paper mill wastewater treatment using a fuzzy neural network," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5064–5070, Apr. 2009.

[12] S. H. Hong, M. W. Lee, D. S. Lee, and J. M. Park, "Monitoring of sequencing batch reactor for nitrogen and phosphorus removal using neural networks," *Biochem. Eng. J.*, vol. 35, no. 3, pp. 365–370, Aug. 2007.

[13] J. Wan, M. Huang, Y. Ma, W. Guo, Y. Wang, H. Zhang, W. Li, and X. Sun, "Prediction of effluent quality of a paper mill wastewater treatment using an adaptive network-based fuzzy inference system," *Appl. Soft Comput.*, vol. 11, no. 3, pp. 3238–3246, Apr. 2011.

[14] Z. Ge and Z. Song, "Nonlinear soft sensor development based on relevance vector machine," *Ind. Eng. Chem. Res.*, vol. 49, no. 18, pp. 8685–8693, Sep. 2010.

[15] J. Qu and M. J. Zuo, "An LSSVR-based algorithm for online system condition prognostics," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 6089–6102, Apr. 2012.

[16] Z.-J. Liu, J.-Q. Wan, Y.-W. Ma, and Y. Wang, "Online prediction of effluent COD in the anaerobic wastewater treatment system based on PCA-LSSVM algorithm," *Environ. Sci. Pollut. Res.*, vol. 26, no. 13, pp. 12828–12841, May 2019.

[17] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Mar. 2006.

[18] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, Mar. 2004.

[19] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[20] Y. Freund, R. E. Schapire, and N. Abe, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, 1999.

[21] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[22] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[23] L. Breiman, "Using iterated bagging to debias regressions," *Mach. Learn.*, vol. 45, no. 3, pp. 261–277, 2001.

[24] H. Parvin and H. Alizadeh, "Classifier ensemble based class weightening," *Amer. J. Sci. Res.*, vol. 19, pp. 84–90, May 2011.

[25] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *J. Artif. Intell. Res.*, vol. 10, pp. 271–289, May 1999.

[26] F. Divina, A. Gilson, F. Gómez-Vela, M. G. Torres, and J. Torres, "Stacking ensemble learning for short-term electricity consumption forecasting," *Energies*, vol. 11, pp. 949–979, Apr. 2018.

[27] M. A. Khairalla, X. Ning, N. T. Al-Jallad, and M. O. El-Faroug, "Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model," *Energies*, vol. 11, no. 6, pp. 1605–1625, Jun. 2018.

[28] W. Sun and B. Trevor, "A stacking ensemble learning framework for annual river ice breakup dates," *J. Hydrol.*, vol. 561, pp. 636–650, Jun. 2018.

[29] H. Liu and C. Yoo, "A robust localized soft sensor for particulate matter modeling in Seoul metro systems," *J. Hazardous Mater.*, vol. 305, pp. 209–218, Mar. 2016.

[30] I.-G. Chong and C.-H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometric Intell. Lab. Syst.*, vol. 78, nos. 1–2, pp. 103–112, Jul. 2005.

[31] H. Yasin, R. E. Caraka, and A. Hoyyi, "Prediction of crude oil prices using support vector regression (SVR) with grid search-cross validation algorithm," *Global J. Pure Appl. Math.*, vol. 12, no. 4, pp. 3009–3020, Apr. 2016.

HONGBIN LIU received the M.S. degree from the South China University of Technology, Guangzhou, China, in 2009, and the Ph.D. degree from Kyung Hee University, Seoul, South Korea, in 2013. He has been an Associate Professor with Nanjing Forestry University, Nanjing, China, since 2015. He is the author of more than 60 journal articles. His research interests include process modeling, process monitoring, fault detection and diagnosis, and wastewater treatment.

CHEN XIN received the B.S. degree from Nanjing Forestry University, Nanjing, China, in 2018, where he is currently pursuing the M.S. degree in process modeling and process monitoring direction.

HAO ZHANG received the B.S. degree from Nanjing Forestry University, Nanjing, China, in 2018, where he is currently pursuing the M.S. degree in process modeling and monitoring for wastewater treatment processes.

FENGSHAN ZHANG received the Ph.D. degree from Nanjing Forestry University, Nanjing, China, in 2010. He is currently the Chief Engineer with Shandong Huatai Paper Company Ltd., Dongying, China. His research interests include process modeling and engineering.

MINGZHI HUANG received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2010. He has been a Professor with South China Normal University, China, since 2017. He is the author of more than 100 journal articles. His research interests include process modeling, process monitoring, fault detection and diagnosis, and wastewater treatment.