Multi-grained cascade forest for effluent quality prediction of papermaking wastewater treatment processes

Chen Xin, Xueqing Shi, Dongsheng Wang, Chong Yang, Qian Li and Hongbin Liu

ABSTRACT

The real time estimation of effluent indices of papermaking wastewater is vital to environmental conservation. Ensemble methods have significant advantages over conventional single models in terms of prediction accuracy. As an ensemble method, multi-grained cascade forest (gcForest) is implemented for the prediction of wastewater indices. Compared with the conventional modeling methods including partial least squares, support vector regression, and artificial neural networks, the gcForest model shows prediction superiority for effluent suspended solid (SS_{eff}) and effluent chemical oxygen demand (COD_{eff}). In terms of SS_{eff}, gcForest achieves the highest correlation coefficient with a value of 0.86 and the lowest root-mean-square error (RMSE) value of 0.41. In comparison with the conventional models, the RMSE value using gcForest is reduced by approximately 46.05% to 50.60%. In terms of COD_{eff}, gcForest achieves the highest correlation coefficient with a value of 0.83 and the lowest root-mean-square error value of 4.05. In comparison with the conventional models, the RMSE value using gcForest is reduced by approximately 10.60% to 18.51%.

Key words | effluent indices, ensemble methods, multi-grained cascade forest, prediction accuracy, wastewater treatment processes

Chen Xin
Xueqing Shi
Chong Yang
Hongbin Liu (corresponding author)
Co-Innovation Center of Efficient Processing and Utilization of Forest Resources,
Nanjing Forestry University,
Nanjing 210037,
China
E-mail: *hongbinliu@njfu.edu.cn*

Dongsheng Wang
School of Automation,
Nanjing University of Posts and Telecommunication,
Nanjing 210023,
China

Qian Li
Department of Environmental Science and Engineering, College of Engineering,
Kyung Hee University,
Yongin 446701,
Korea

INTRODUCTION

With the increased public concern of environmental protection in recent years, it is noteworthy to improve technology capability and flexibility for online process monitoring in wastewater treatment plants (WWTPs) (Haimi *et al.* 2015). Especially in order to meet the discharge standard, several effluent indices on wastewater should be controlled within a reasonable range, such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), suspended solids (SS), daily sewage sludge, and daily flow rates (Qiao *et al.* 2016; Najafzadeh & Zeinolabedini 2018, 2019; Zeinolabedini & Najafzadeh 2019). In general, specific hardware sensors are often used for wastewater monitoring. However, hardware sensors have obvious disadvantages, such as high maintenance costs, short service life, and time consuming processes (Cecil & Kozlowska 2010). As an alternative method, modeling technology based on data-driven characteristics is gradually applied in the monitoring of wastewater treatment processes to predict the effluent indices (Liu *et al.* 2010, 2020; Jin *et al.* 2015; Najafzadeh & Zeinolabedini 2018). According to the prediction results, an early warning message can be generated when the effluent concentrations exceed the discharge standard, which can help the operators in WWTPs to carry out corresponding process adjustments (Liu & Ge 2018).

At present, partial least squares (PLS) (Wang *et al.* 2015), support vector regression (SVR) (Meng *et al.* 2015), and artificial neural networks (ANN) (Mjalli *et al.* 2007) are the main methods applied to the prediction of wastewater effluent indices. As a classical method, PLS was employed by MacGregor for analyzing the operations of a mineral processing plant (MacGregor & Kourti 1995). PLS takes both the variance structure and the correlation between the inputs and the outputs into consideration (Shi *et al.* 2016). Besides, PLS could avoid excessive time consumption

doi: 10.2166/wst.2020.206

in the training period by using less latent variables (LVs) instead of full input variables (Huang *et al.* 2018). However, PLS is a statistical method based on the linear correlation of process variables, which may not be able to handle the complex and nonlinear characteristics in papermaking wastewater treatment processes. Therefore, in order to solve the above problems, other advanced methods need to be embedded into PLS. In terms of the prediction for COD, a dynamic Gaussian process regression based PLS model was used to improve the prediction performance (Liu *et al.* 2019). Furthermore, the information loss of PLS will be inevitable even choosing the desirable latent variables in the modeling process (Singh *et al.* 2010). In this case, the emergence of ANN offers a new idea for predicting effluent indices (Ráduly *et al.* 2007). Similar to biological neurons, the structure of an ANN model has a strong nonlinear fitting ability (Gonzaga *et al.* 2009), which allows ANN to be successfully applied in predicting the daily flow rates, BOD, and SS concentrations in WWTPs (Hamed *et al.* 2004; Najafzadeh & Zeinolabedini 2019). However, the training process of ANN may end prematurely and the weights decline phenomenon might appear which will reduce the modeling accuracy. Additionally, the hyper-parameters and the fitting ability of ANN are often based on expertise and the sample size of training data, respectively. Considering the complexity and time-variability characteristics of WWTPs, it will be a time-consuming work to adjust the parameters for further industrial validation via ANN method. To overcome the above drawbacks, deep learning and wavelet functions were applied into ANN, and the prediction results indicated that the optimized model behaved more favorably (Hinton & Salakhutdinov 2006; Zeinolabedini & Najafzadeh 2019). Compared to ANN, SVR solves the unavoidable local extremum problem of ANN and selectively finds the finite vector in the input data, which is more efficient than ANN's iteration calculation for the whole sample (Yan *et al.* 2004). The concentration of total nitrogen from a wastewater treatment plant was successfully predicted by Guo (Guo *et al.* 2015). However, for large-scale training samples, SVR is difficult to effectively restore the sample's valid information and the prediction accuracy can be decreased significantly (Suykens *et al.* 2002).

Ensemble methods improve prediction accuracy by combining different models (Krogh & Sollich 1997; Altman & Krzywinski 2017) and have already been applied in the field of effluent wastewater prediction (Granata & Marinis 2017). Numerous studies have proven that the prediction performance of an ensemble model is superior to the single prediction model (Sendi *et al.* 2019). A new fault diagnosis method based on multi-grained cascade forest (gcForest) showed higher fault classification accuracy than other popular deep learning algorithms (Hu *et al.* 2018). Combined with deep Boltzmann machine approach, the modified gcForest fault diagnosis method not only specializes in discovering the complex relationship between real industrial data and potential faults but also spends much less time compared with conventional methods. The gcForest was also applied to the classification of cloud/snow images from single-spectral satellites (Xia *et al.* 2019). With the capability of effectively extracting the satellite cloud/snow imagery features, gcForest improves the feature utilization rate and increases the accuracy of the classification. Additionally, the gene function of yeast–human polyprotein interaction networks was predicted based on gcForest (Zhang & Deng 2019), which shows that gcForest is not sensitive to the hyper-parameters, especially the number of dimensions for function prediction, and achieves competitive results on the data sets of yeast and human.

Aiming to obtain higher estimation accuracy and overcome the disadvantages of the conventional models listed in Table 1 (Ge 2018), gcForest is proposed in this paper for the prediction of effluent indices. This paper is organized as follows. First, the basic concept of gcForest is introduced in more detail. Second, gcForest and other conventional models are applied into a wastewater treatment process. Finally, the conclusions are given.

Table 1 | The disadvantages of PLS, SVR, and ANN for process modeling

Conventional modeling methods	Disadvantages
PLS	(1) Linear method which cannot handle nonlinear characteristics of WWTPs data (2) Information loss caused by choosing partial latent variables is inevitable
SVR	(1) High computational complexity when the scale of training samples is large (2) The choice of kernel function is lack of a uniform standard
ANN	(1) The training process tends to ending prematurely and the weights declining phenomenon often appears (2) Too many hyper-parameters need to be adjusted (3) Low convergence rate and easy to be trapped in local minima

METHODS

Multi-grained cascade forest (gcForest)

As an ensemble-based method, gcForest is composed of many decision trees which are much easier to train and work well for small-scale training data. When numerous 'trees' get together, they can be vividly called forest. The essence of gcForest lies in its powerful ensemble learning ability, which combines several weak classifiers into a strong classifier to improve the accuracy of the model. The gcForest can be divided into two parts: multi-grained scanning part and cascade forest part. The design of the former part can map the sample data from low dimensional space to high dimensional space with the tool of sliding windows, which aim to adequately extract the feature information from training samples. The latter part is inspired by the representation learning method in deep neural networks. Each cascade structure will receive the feature information processed by its preceding layer and transfer its result to the next layer so that the information of a training set will be fully learned.

Multi-grained scanning part

The purpose of multi-grained scanning is to transform the raw feature vectors into higher-dimensional feature vectors, so that the model can fully utilize the information contained in each sample and overcome the problem of scale variations, which enhances the model learning ability of representation. For example, each wastewater sample has six input features (Q, SS_{in}, pH, T, COD_{in}, and DO), which means every sample is a raw vector with six dimensions. As shown in Figure 1, four three-dimensional feature samples, three four-dimensional feature samples, and two five-dimensional feature samples are obtained by scanning the input raw features using sliding windows with sizes of three, four, and five, respectively. For n training samples, a three-dimensional window will generate $4n$ feature vectors. Similarly, for a four-dimensional window, $3n$ feature vectors will be generated and $2n$ feature vectors can be obtained from a five-dimensional window.

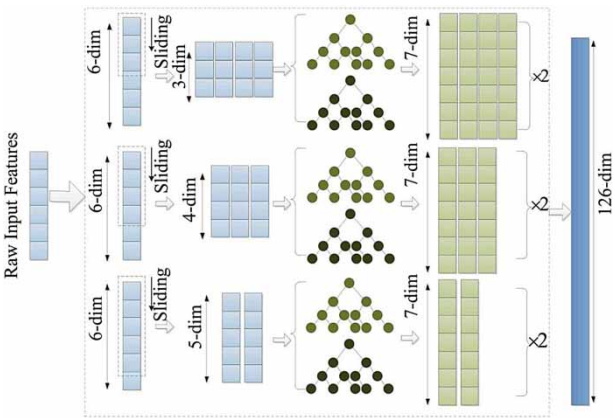

Figure 1 | Illustration of multi-grained scanning part.

Because the diversity of ensemble construction is of great importance in the ability to generalize models, all feature vectors will be utilized to train each completely random tree forest and random forest. Each leaf node of decision trees in a completely random tree forest is split by selecting features randomly in a feature space, while each leaf node of the decision trees in a random forest is split by the best Gini value in a feature subspace (Raileanu & Stoffel 2004). The index of Gini is defined by Equation (1),

$$Gini(p) = \sum_{i=1}^{N} P_i(1 - P_i) = 1 - \sum_{i=1}^{N} P_i^2 \quad (1)$$

where N is the class of predicted value and P_i is the proportion that a certain sample belongs to i class. According to the definition, if the Gini value is smaller, the probability that the sample belongs to a certain class is higher.

For example, suppose SS_{eff} is between 19 and 26, then each forest will output a seven-dimensional class vector, so a three-dimensional window produces a 28-dimensional (4×7) feature subsample. Similarly, a four-dimensional window produces a 21-dimensional (3×7) feature subsample and a five-dimensional window produces a 14-dimensional (2×7) feature subsample. The subsamples generated by each window are trained by two different forests through a combination of all feature subsamples. A 126-dimensional transformed feature vector is generated and the transformed feature vector as an output enters the cascade forest part.

Cascade forest part

The cascade forests adopt a layer by layer network structure which is similar to the structure of multi-layer artificial neural network. It can be regarded as an ensemble of random forests. The flow chart of the network structure is illustrated in Figure 2.

Each layer is composed of several random forests in Figure 2 and each decision tree in a certain forest generates a class distribution. Then, a class vector of the forest is produced by means of averaging the class distribution generated by all decision trees. For SS_{eff}, each forest produces a seven-dimensional class vector. Finally, all the class vectors are combined to produce a 42-dimensional (6×7) feature

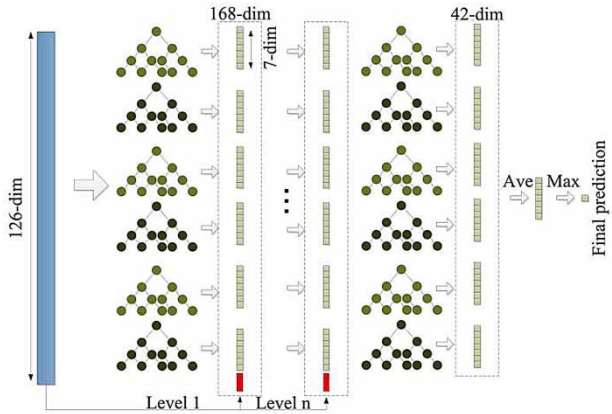

Figure 2 | Illustration of cascade forest part.

vector, which then concatenate with the vectors generated from the multi-grained scanning to produce a 168-dimensional (42 + 126) representation vector. The representation vector can be used as the input for the next cascade level. Each layer of cascade outputs its processing results to the next layer and the processing results pass on layer-by-layer until the prediction performance in the next level of cascade is not improved. Fortunately, the optimal number of cascades can be determined automatically, enabling the complexity of the model to be greatly reduced. After averaging all of the class vectors generated by the final output cascade, the predicted value of a specific sample can be obtained.

Modeling performance indices

The prediction performance of the gcForest and other conventional models are evaluated by using correlation coefficient (r), mean absolute percentage error (MAPE) and root mean square error (RMSE) as defined by Equations (2)–(4), respectively. High value of r and small value of MAPE and RMSE mean a better performance of the prediction.

$$r = \frac{\sum_{t=1,i=1}^{N} (y_t - \bar{y}_t)(\hat{y}_i - \bar{y}_i)}{\sqrt{\sum_{t=1}^{N} (y_t - \bar{y}_t)^2 \sum_{i=1}^{N} (\hat{y}_i - \bar{y}_i)^2}} \quad (2)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_t - \hat{y}_i}{y_t} \right| \times 100 \quad (3)$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 / N} \quad (4)$$

where y_t is measured values, $\hat{y}_i$ is predicted values, $\bar{y}_t = \frac{1}{N}\sum_{t=1}^{N} y_t$ and $\bar{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i$ are average values of y_t and y_i, respectively.

RESULTS AND DISCUSSION

Industrial WWTP data

In order to verify the superiority of gcForest method and observe the prediction results of conventional methods, 170 samples were used in the work which was obtained from a papermaking wastewater treatment plant in Dongguan, China. As shown in Figure 3, each sample contains eight variables (six input variables and two output variables), including wastewater flow rate (Q), influent suspended solid (SS_{in}), effluent suspended solid (SS_{eff}), pH, temperature (T), influent chemical oxygen demand (COD_{in}), effluent chemical oxygen demand (COD_{eff}), and dissolved oxygen (DO). Both COD_{eff} and SS_{eff} meet the Chinese national discharge standards, which are less than 100 mg/L and 30 mg/L, respectively. Probes (HACH) and cards (Advantech) were used for collecting the WWTP data. The signals, filtered in a transmitter, were captured by a data acquisition card (ADAM4017, Advantech, China). A power relay output board (ADAM4024, Advantech, China) was used for optimizing equipment functioning.

To improve the prediction accuracy, data pre-processing is an essential step. First, the outliers in the measured data were detected using the three-sigma rule and replaced by reasonable values. In this work, regression equations were established using the measured data, and the fitted values obtained from the regression equations were used to replace the outliers. Second, these treated data were normalized. Finally, the samples were divided into a training data set and a test data set. 170 samples were randomly divided into two data sets with the ratio of 8:2, 136 of which were used for the training data set and the remaining samples for the test data set. In consideration of the measuring error in the data-collection process, uncertainty analysis of all variables needs to be implemented (Sadeghi *et al.* 2020). As a data-driven model, the relationships between process variables are difficult to describe. In this work, 10-fold cross-validation was used to observe the impact of data uncertainty. Specifically, after the sequences of samples were rearranged, they were divided into 10 parts among which nine parts were selected as the training data set in

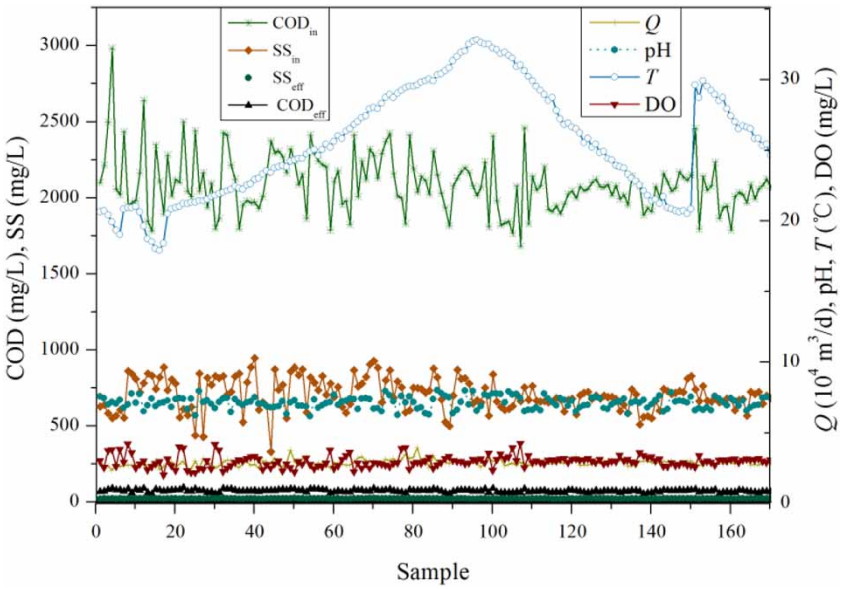

Figure 3 | Variables in the papermaking WWTP data.

turn, and the remaining part was used for validation. After 10 times of calculation, the average result was taken as the final result.

Conventional methods for the WWTP data

According to the cumulative variance proportion computed via PLS, three latent variables (LVs) were selected for modeling. It is worth noting that the cumulative variance proportion of three LVs was up to 78.39%, which means that these three LVs contain enough information for the process modeling. In terms of the SVR model, radial basis kernel function was used for process modeling. The ANN model is composed of three layers. Considering that the number of nodes and the network structure are critical to prediction accuracy, a grid search was used to select the optimal parameters. All the important parameters of the conventional methods are listed in Table 2.

Table 2 | Modeling parameters of the conventional methods

Methods	COD	SS
PLS	Number of latent variables: 3	Number of latent variables: 3
SVR	Kernel function: radial basis function Kernel parameter $\gamma = 56$ Regularization parameter C = 118	Kernel function: radial basis function Kernel parameter $\gamma = 2.8$ Regularization parameter C = 60
ANN	Number of neural network layers: 3 Number of hidden layer: 1 Input layer nodes: 6 Hidden layer nodes: 2 Output layer nodes: 1 Activation function of hidden layer: tansig Activation function of output layer: purelin Max training time: 1000	Number of neural network layers: 3 Number of hidden layer: 1 Input layer nodes: 6 Hidden layer nodes: 3 Output layer nodes: 1 Activation function of hidden layer: tansig Activation function of output layer: purelin Max training time: 1000

Multi-grained scanning forest for the WWTP data

The hyper-parameters need to be determined in advance, including the shape of the sample (dimension of the sample), the number of forests in multi-grained scanning and cascade forest, the growth rules of decision trees, the number and size of decision trees in each forest, the sliding windows, and the sliding step. The shape or dimension of the data, determined by the number of features contained in the sample, contains six input features. Generally speaking, the larger the number of forests and decision trees in each forest, the better chance of prediction accuracy tends to be increased. According to this phenomenon, three completely random forests and three random forests were applied in the multi-grained scanning part and the cascade forest part, and each of the forests contains 500 decision trees. It is worth noting that the model is not sensitive to the change of the number of decision trees. Once the number of decision trees reaches a certain value, the model will show a promising prediction performance.

Compared with the process of parameter optimization in ANN, the training time is shorter. Similarly, the number and the size of sliding windows were gradually increased to the desired value and it is confirmed that the sliding windows with the size of 3, 4, and 5 were the best selection for multi-grained scanning, respectively. Considering that the number of samples in this work is small, the number of sliding steps was selected as one step to obtain as many subsamples as possible. Each tree cannot be pruned back until its maximum growth was achieved.

Modeling results of gcForest and conventional methods

The prediction results of SS_{eff} and COD_{eff} are shown in Figure 4. The prediction performance of each model is evaluated by the index values of r, RMSE, and MAPE for SS_{eff} and COD_{eff} (Table 3). In terms of COD_{eff}, the best prediction results come from gcForest, with $r = 0.83$, RMSE = 4.05 and MAPE = 4.26. Compared with PLS, ANN, and SVR, the values of RMSE and MAPE using gcForest are reduced by 18.02% and 26.93%, 18.51% and 28.04%, 10.60% and 18.55%, respectively. In terms of SS_{eff}, the gcForest shows the highest prediction accuracy ($r = 0.86$). On the other hand, gcForest has the smallest RMSE (0.41) and MAPE (2.24). Compared with PLS, ANN, and SVR, the values of RMSE and MAPE using gcForest are reduced by 50.60% and 23.55%, 48.10% and 19.42%, 46.05% and 16.10%, respectively. In addition, taking the uncertainty factor of the WWTP data into account, we implemented 10-fold cross-validation whose results are listed in Table 3, which shows that there is no significant change in the prediction accuracy. Specifically, the change of RMSE and MAPE is no more than 4.88%, and the change of r is no more than 2.41%, which can prove that gcForest is a robust modeling method used for the prediction of the papermaking wastewater treatment process.

The results demonstrate that gcForest is more suitable for predicting the SS_{eff} and COD_{eff} than the other

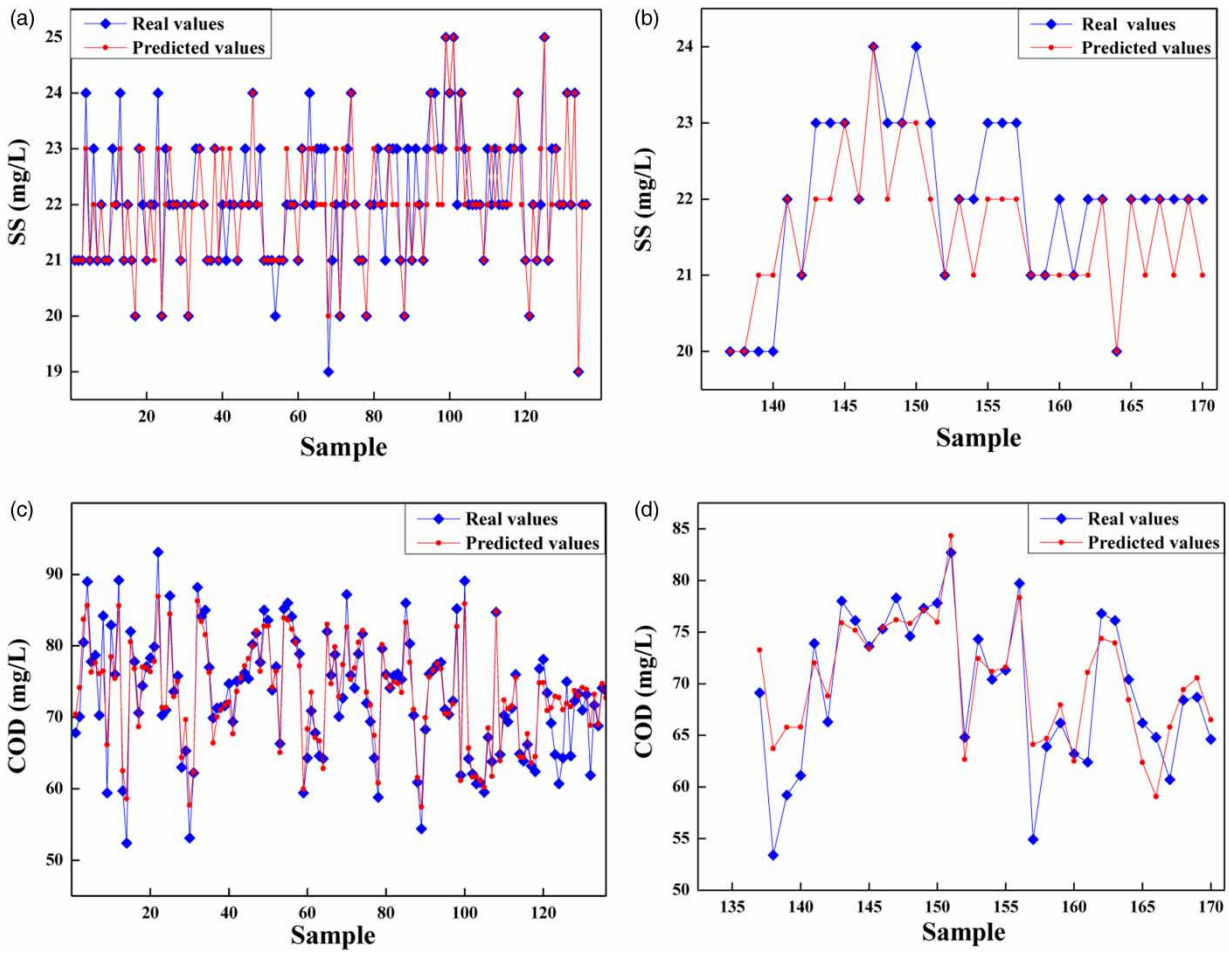

Figure 4 | Prediction results of SS_{eff} and COD_{eff} using gcForest. (a) Training data. (b) Test data. (c) Training data. (d) Test data.

Table 3 | Comparison of different methods for prediction of SS_{eff} and COD_{eff}

Models		COD_{eff}			SS_{eff}		
		RMSE	MAPE (%)	*r*	RMSE	MAPE (%)	*r*
PLS	Training data	4.32	4.56	0.80	0.79	2.41	0.78
	Test data	4.94	5.83	0.71	0.83	2.93	0.70
ANN	Training data	4.56	5.86	0.84	0.68	2.13	0.82
	Test data	4.97	5.92	0.71	0.79	2.78	0.76
SVR	Training data	5.11	5.32	0.83	0.72	2.58	0.83
	Test data	4.53	5.23	0.78	0.76	2.67	0.72
gcForest	Training data	2.34	2.45	0.92	0.52	1.70	0.85
	Test data	4.05	4.26	0.83	0.41	2.24	0.86
gcForest (cross validation)	Training data	2.33	2.43	0.90	0.51	1.73	0.82
	Test data	4.12	4.27	0.81	0.43	2.23	0.84

conventional models. Due to the complex and nonlinear characteristics of wastewater treatment processes, it is difficult for a linear model like PLS to extract the related information between input variables and output variables. Moreover, PLS usually utilizes partial information of latent variables in modeling process, so it cannot make full use of the information contained in the samples. Compared with PLS, ANN has stronger nonlinear fitting ability. However, there are many hyper-parameters needing to be determined when ANN is used for modeling. Meanwhile, there is no uniform and standardized optimization method to determine the hyper-parameters such as network structure, momentum factor, the number of nodes. Therefore, it is time consuming to find the optimal parameters through optimization which usually needs expertise and skills. Additionally, off-line samples are often obtained by laboratory analysis, so the sample size might be small, which is likely to be the reason for poor generalization of ANN. Furthermore, ANN still has the limitation of getting trapped in local minima. Compared with ANN, SVR shows lower prediction accuracy and the introduction of kernel functions also increases the difficulty for modeling. By contrast, gcForest has achieved better prediction performance in this work. Unlike PLS, the multi-grained scanning technique in gcForest can make full use of the feature's information in the training process without reducing the dimension of input variables. Through utilizing the adaptive function of the cascade structure, gcForest greatly reduces the time cost of hyper-parameter optimization in comparison to ANN. Besides, because the step of kernel optimization used in SVR is omitted, the gcForest modeling is easier to implement. Like other data-driven models, gcForest does not require expertise in the modeling process. In other words, the complex biochemical reaction mechanisms in WWTPs need not be fully understood.

The running times of different models are provided in Table 4, the training time and prediction time of PLS and SVR are shorter than 1 s. The training times of ANN and gcForest are 23.078 s and 11.254 s, respectively, which are significantly longer than those of PLS and SVR. The reason is that ANN uses a back-propagation algorithm for optimizing the hyper-parameters such as the weights and threshold of each neuron. The hyper-parameters need to be revised several times to reach the precision requirement. In terms of gcForest, the reason mainly lies in the fact that processing results of each layer need to pass on layer-by-layer until the prediction performance in the next level of cascade is not improved. This optimization step is a time-consuming process. Nevertheless, when the model is further applied to WWTPs with larger data volume, the training time of the model may be greatly increased. Therefore, to improve the execution efficiency, the parameter and structure of gcForest needs to be further optimized in the future.

Overall, gcForest offers several attractive properties, such as simple algorithm, fewer hyper-parameters during optimization, and powerful ability in extracting feature

Table 4 | Comparison of the execution time of different models

Models	Training time (s)	Prediction time (s)
PLS	0.080	0.003
SVR	0.120	0.011
ANN	23.078	1.677
gcForest	11.254	0.469

relationships. Moreover, some other advantages of gcForest are provided as follows: (1) higher prediction accuracy can be achieved when the scale of training samples is small; (2) the model has an adaptive function, and its complexity can be automatically set according to the actual situation; (3) the model is robust without over-fitting.

The following aspects are supposed to be explored in the future: the relatively high training time cost and limited guiding significance for production operation. With the development of the control strategies in WWTPs, the explanation of prediction results and the ability to introduce process operation knowledge to the gcForest have become more and more necessary.

CONCLUSIONS

In this paper, gcForest is used for the prediction of effluent indices of a papermaking wastewater treatment process. To illustrate the gcForest's superiority in terms of prediction performance, conventional models, including PLS, SVR, and ANN, are implemented for comparison. The results indicate that the prediction performance of the gcForest is significantly improved compared with the conventional models.

In terms of COD_{eff}, gcForest shows the highest prediction accuracy, with $r = 0.83$ and the smallest RMSE and MAPE, with values of 4.05 and 4.26, respectively. Compared with the conventional models, the values of RMSE and MAPE are reduced by approximately 10.60% to 28.04%. On the other hand, in terms of SS_{eff}, gcForest shows the highest prediction accuracy, with $r = 0.86$ and the smallest RMSE and MAPE, with values of 0.41 and 2.24, respectively. Compared with conventional models, the values of RMSE and MAPE are reduced by approximately 16.10% to 50.60%. However, taking the drawback of gcForest into account, there is still much research to be done. Future work will be focused on the improvement in the execution efficiency. Meanwhile, the number of training samples and input variables will be increased to further validate the prediction accuracy and robustness of gcForest in a new wastewater treatment plant.

CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

ACKNOWLEDGEMENTS

This study was supported by the Foundation of Nanjing Forestry University (No. GXL029) and the National Natural Science Foundation of China (No. 51708299).

REFERENCES

Altman, N. & Krzywinski, M. 2017 Points of significance: ensemble methods: bagging and random forests. *Nature Methods* **14**, 933–934.

Cecil, D. & Kozlowska, M. 2010 Software sensors are a real alternative to true sensors. *Environmental Modelling & Software* **25** (5), 622–625.

Ge, Z. 2018 Process data analytics via probabilistic latent variable models: a tutorial review. *Industrial & Engineering Chemistry Research* **57** (38), 12646–12661.

Gonzaga, J. C. B., Meleiro, L. A. C., Kiang, C. & Filho, R. M. 2009 ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Computers & Chemical Engineering* **33** (1), 43–49.

Granata, F. & Marinis, G. D. 2017 Machine learning methods for wastewater hydraulics. *Flow Measurement & Instrumentation* **57**, 1–9.

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J.-P., Kim, J. H. & Cho, K. H. 2015 Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences* **32**, 90–101.

Haimi, H., Corona, F., Mulas, M., Sundell, L., Heinonen, M. & Vahala, R. 2015 Shall we use hardware sensor measurements or soft-sensor estimates? case study in a full-scale WWTP. *Environmental Modelling & Software* **72**, 215–229.

Hamed, M. M., Khalafallah, M. G. & Hassanien, E. A. 2004 Prediction of wastewater treatment plant performance using artificial neural networks. *Environmental Modelling & Software* **19** (10), 919–928.

Hinton, G. E. & Salakhutdinov, R. R. 2006 Reducing the dimensionality of data with neural networks. *Science* **313** (5786), 504–507.

Hu, G., Li, H., Xia, Y. & Luo, L. 2018 A deep Boltzmann machine and multi-grained scanning forest ensemble collaborative method and its application to industrial fault diagnosis. *Computers in Industry* **100**, 287–296.

Huang, X., Luo, Y.-P., Xu, Q.-S. & Liang, Y.-Z. 2018 Incorporating variable importance into kernel PLS for modeling the structure–activity relationship. *Journal of Mathematical Chemistry* **56** (3), 713–727.

Jin, H., Chen, X., Wang, L., Yang, K. & Wu, L. 2015 Adaptive soft sensor development based on online ensemble Gaussian process regression for nonlinear time-varying batch processes. *Industrial & Engineering Chemistry Research* **54** (30), 7320–7345.

Krogh, A. & Sollich, P. 1997 Statistical mechanics of ensemble learning. *Physical Review E* **55** (1), 811–825.

Liu, Y. & Ge, Z. 2018 Weighted random forests for fault classification in industrial processes with hierarchical clustering model selection. *Journal of Process Control* **64**, 62–70.

Liu, J., Chen, D.-S. & Shen, J.-F. 2010 Development of self-validating soft sensors using fast moving window partial least squares. *Industrial & Engineering Chemistry Research* **49** (22), 11530–11546.

Liu, H., Yang, C., Carlsson, B., Qin, S. & Yoo, C. 2019 Dynamic nonlinear PLS modeling using Gaussian process regression. *Industrial & Engineering Chemistry Research* **58** (36), 16676–16686.

Liu, H., Yang, C., Huang, M. & Yoo, C. 2020 Soft sensor modeling of industrial process data using kernel latent variables-based relevance vector machine. *Applied Soft Computing* **90**, 1–10.

MacGregor, J. F. & Kourti, T. 1995 Statistical process control of multivariate processes. *Control Engineering Practice* **3** (3), 403–414.

Meng, H., Gao, H. & Zhao, B. Y. 2015 Soft measurement of the cell concentration based on SVM. *Applied Mechanics and Materials* **742**, 239–243.

Mjalli, F. S., Al-Asheh, S. & Alfadala, H. 2007 Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. *Journal of Environmental Management* **83** (3), 329–338.

Najafzadeh, M. & Zeinolabedini, M. 2018 Derivation of optimal equations for prediction of sewage sludge quantity using wavelet conjunction models: an environmental assessment. *Environmental Science and Pollution Research* **25** (23), 22931–22943.

Najafzadeh, M. & Zeinolabedini, M. 2019 Prognostication of waste water treatment plant performance using efficient soft computing models: an environmental evaluation. *Measurement* **138**, 690–701.

Qiao, J., Hu, Z. & Li, W. 2016 Soft measurement modeling based on chaos theory for biochemical oxygen demand (BOD). *Water* **8** (12), 581–601.

Ráduly, B., Gernaey, K. V., Capodaglio, A. G., Mikkelsen, P. S. & Henze, M. 2007 Artificial neural networks for rapid WWTP performance evaluation: methodology and case study. *Environmental Modelling & Software* **22** (8), 1208–1216.

Raileanu, L. E. & Stoffel, K. 2004 Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence* **41** (1), 77–93.

Sadeghi, G., Najafzadeh, M. & Ameri, M. 2020 Thermal characteristics of evacuated tube solar collectors with coil inside: an experimental study and evolutionary algorithms. *Renewable Energy* **151**, 575–588.

Sendi, N., Abchiche-Mimouni, N. & Zehraoui, F. 2019 A new transparent ensemble method based on deep learning. *Procedia Computer Science* **159**, 271–280.

Shi, H., Kim, M., Liu, H. & Yoo, C. 2016 Process modeling based on nonlinear PLS models using a prior knowledge-driven time difference method. *Journal of the Taiwan Institute of Chemical Engineers* **69**, 93–105.

Singh, K. P., Basant, N., Malik, A. & Jain, G. 2010 Modeling the performance of 'up-flow anaerobic sludge blanket' reactor based wastewater treatment plant using linear and nonlinear approaches – a case study. *Analytica Chimica Acta* **658** (1), 1–11.

Suykens, J. A., De Brabanter, J., Lukas, L. & Vandewalle, J. 2002 Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing* **48** (1–4), 85–105.

Wang, Y., Hui, C., Yan, Z. & Zhang, Y. 2015 Nonlinear partial least squares regressions for spectral quantitative analysis. *Chemometrics & Intelligent Laboratory Systems* **148**, 32–50.

Xia, M., Liu, W. a., Shi, B., Weng, L. & Liu, J. 2019 Cloud/snow recognition for multispectral satellite imagery based on a multidimensional deep residual network. *International Journal of Remote Sensing* **40** (1), 156–170.

Yan, W., Shao, H. & Wang, X. 2004 Soft sensing modeling based on support vector machine and Bayesian model selection. *Computers & Chemical Engineering* **28** (8), 1489–1498.

Zeinolabedini, M. & Najafzadeh, M. 2019 Comparative study of different wavelet-based neural network models to predict sewage sludge quantity in wastewater treatment plant. *Environmental Monitoring and Assessment* **191** (3), 163–187.

Zhang, J. & Deng, L. 2019 Integrating multiple interaction networks for gene function inference. *Molecules* **24** (1), 30–45.

First received 20 February 2020; accepted in revised form 17 April 2020. Available online 28 April 2020