Front. Environ. Sci. Eng. 2023, 17(6): 67
https://doi.org/10.1007/s11783-023-1667-3

RESEARCH ARTICLE

Water quality soft-sensor prediction in anaerobic process using deep neural network optimized by Tree-structured Parzen Estimator

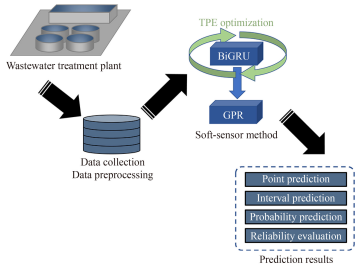
Junlang Li[1], Zhenguo Chen (✉)[1], Xiaoyong Li[1], Xiaohui Yi[1], Yingzhong Zhao[2], Xinzhong He[2], Zehua Huang[2], Mohamed A. Hassaan[3], Ahmed El Nemr[3], Mingzhi Huang (✉)[1,4]

1 SCNU Environmental Research Institute, Guangdong Provincial Key Laboratory of Chemical Pollution and Environmental Safety & MOE Key Laboratory of Theoretical Chemistry of Environment, School of Environment, South China Normal University, Guangzhou 510006, China
2 Fujian Environmental Protection Design Institute Co. Ltd, Fuzhou 350000, China
3 National Institute of Oceanography and Fisheries, NIOF, Alexandria 21556, Egypt
4 SCNU Qingyuan Institute of Science and Technology Innovation Co., Ltd., Qingyuan 511517, China

HIGHLIGHTS

- Hybrid deep-learning model is proposed for water quality prediction.
- Tree-structured Parzen Estimator is employed to optimize the neural network.
- Developed model performs well in accuracy and uncertainty.
- Usage of the proposed model can reduce carbon emission and energy consumption.

GRAPHIC ABSTRACT



ARTICLE INFO

Article history:
Received 7 March 2022
Revised 18 October 2022
Accepted 21 October 2022
Available online 18 December 2022

Keywords:
Water quality prediction
Soft-sensor
Anaerobic process
Tree-structured Parzen Estimator

ABSTRACT

Anaerobic process is regarded as a green and sustainable process due to low carbon emission and minimal energy consumption in wastewater treatment plants (WWTPs). However, some water quality metrics are not measurable in real time, thus influencing the judgment of the operators and may increase energy consumption and carbon emission. One of the solutions is using a soft-sensor prediction technique. This article introduces a water quality soft-sensor prediction method based on Bidirectional Gated Recurrent Unit (BiGRU) combined with Gaussian Progress Regression (GPR) optimized by Tree-structured Parzen Estimator (TPE). TPE automatically optimizes the hyperparameters of BiGRU, and BiGRU is trained to obtain the point prediction with GPR for the interval prediction. Then, a case study applying this prediction method for an actual anaerobic process (2500 m^3/d) is carried out. Results show that TPE effectively optimizes the hyperparameters of BiGRU. For point prediction of COD_{eff} and biogas yield, R^2 values of BiGRU, which are 0.973 and 0.939, respectively, are increased by 1.03%–7.61% and 1.28%–10.33%, compared with those of other models, and the valid prediction interval can be obtained. Besides, the proposed model is assessed as a reliable model for anaerobic process through the probability prediction and reliable evaluation. It is expected to provide high accuracy and reliable water quality prediction to offer basis for operators in WWTPs to control the reactor and minimize carbon emission and energy consumption.

1 Introduction

Energy consumption and carbon emission in wastewater treatment plants (WWTPs) have attracted attention. Various power equipment is used in WWTPs, causing a large amount of energy consumption (Di Maria and Micale, 2015). Besides, organic pollutants in wastewater

✉ Corresponding authors
E-mails: zhenguo.chen@m.scnu.edu.cn (Z. Chen);
mingzhi.huang@m.scnu.edu.cn (M. Huang)

Special Issue—Artificial Intelligence/Machine Learning on Environmental Science & Engineering (Responsible Editors: Yongsheng Chen, Xiaonan Wang, Joe F. Bozeman III & Shouliang Yi)

are mineralized into CO_2, thus increasing carbon emission and aggravating the greenhouse effect (Hauck et al., 2016). In WWTPs, energy should be recycled, and carbon emission should be reduced. Among many wastewater treatment processes, the anaerobic process has been recognized as a green and sustainable process due to the advantages of high efficiency, low carbon emission, low energy consumption, and considerable bioenergy reproduction by converting biogas into heat, electricity, or fuel (Kim et al., 2012; Wei et al., 2020; Heydari et al., 2021; Şenol, 2021).

Although the anaerobic process has many merits in wastewater treatment, some of the water quality metrics of this process are not measurable in real time to reflect the actual water quality due to lack of corresponding sensor (Newhart et al., 2019; Ching et al., 2021; Darvishi et al., 2021). Besides, the inner reaction is an environmentally sensitive process, which may cause acidification collapse because of misjudgment and poor control (Han et al., 2021; Wu et al., 2021). Moreover, poor control of the reaction system leads to poor water quality, increased carbon emission, and extra usage of the equipment with increased energy consumption (Wang et al., 2016). The water quality soft-sensor prediction is needed by operators in WWTPs to obtain all water quality metrics at the same time for improved control and minimize carbon emission and energy consumption. The soft-sensor technique is an emerging method applied in industrial process monitoring (Yaginuma et al., 2022). Traditional data collection in the industry relies on a large number of sensors, whereas the soft-sensor technique uses software to predict the uneasily measurable variables from real-time measurable variables (Kadlec et al., 2009). Soft-sensor technique can effectively save cost on equipment procurement and extra space on equipment installation and avoid the financial loss and serious safety issues due to possible hardware failure (Jiang et al., 2021).

In recent years, the machine learning-based soft-sensor method is increasingly becoming popular. The machine learning method, such as Support Vector Machine (Zeng et al., 2006), Random Forest (Szelag et al., 2017), and Gaussian Process Regression (GPR) (Samuelsson et al., 2017), uses a data-driven model to predict future data from historical data. This method is able to deal with data with high dimension, low stability, and nonlinearity. However, given the increase in the amount and dimension of data, high accuracy is difficult to obtain using the single traditional machine learning method.

Nowadays, deep learning has been rapidly developed to provide researchers with a new way to deal with proposed problems. The Recurrent Neural Network (RNN) (Chen et al., 2010) is a deep learning model aiming at the sequential problem and is suitable for predicting water quality. However, for data with large length, RNN exists a long-term dependency and easily causes gradient extinction and gradient explosion, thus limiting its application. The Long Short-Time Memory (LSTM) (Li et al., 2021) solves the mentioned defects by adding the input gates, forget gates, and output gates in RNN. LSTM has high converged rate and high prediction accuracy, and its applications have covered many fields. The Gated Recurrent Unit (GRU) (Li et al., 2021) is the variant of LSTM that has a simpler gate structure. GRU can reduce training time and obtain the prediction accuracy close to LSTM. Moreover, the Bidirectional Gated Recurrent Unit (BiGRU) (Wang et al., 2021), as an innovation of GRU by adding a bidirectional recurrent layer, is proven to overcome the forgetting problem and enhance the accuracy efficiently.

For water quality monitoring, prediction should provide the trend and give the scope of change, i.e., interval prediction, which is useful for the operators to know the fluctuation of water quality. Besides the point prediction, some machine learning methods, such as GPR, are skilled at obtaining the interval prediction to describe the uncertainty of the prediction result. GPR can achieve good performance in water quality prediction and produce high reliable prediction interval (Zhang et al., 2019).

For the traditional neural network constructing process, manually tuning the hyperparameters, the configurations of the network, is laborious and time-consuming for researchers. Automatic hyperparameters optimization methods, like Tree-structured Parzen Estimator (TPE) (Nguyen et al., 2020), provide a new way to construct the neural network efficiently, thus saving remarkable time for researchers.

In summary, the main contributions of the paper are as follows:

1) A hybrid model combined BiGRU and GPR, called BiGRU–GPR, is introduced in this paper to achieve high accuracy prediction and obtain its uncertain information.

2) An automatic hyperparameter optimization method called TPE is employed to tune the hyperparameters of BiGRU to save the neural network constructing time and avoid the local optimum.

3) Data from a full-scale anaerobic WWTP is employed to verify the performance of the proposed model. This study is the first to use BiGRU-GPR and the TPE optimization method to predict the anaerobic water quality.

4) The proposed method aims to obtain a high-precision and reliable model to predict the water quality and promote the reduction of carbon emission and energy consumption for anaerobic process.

2 Methodology

2.1 Tree-Structured Parzen Estimator (TPE)

In machine learning, hyperparameters define the model structure and decide the prediction accuracy and training

time, such as learning rate, batch size, iteration, and cell size of hidden layer. Hyperparameters need to be manually tuned, which is time-consuming and a potential local optimum dilemma. Automatic optimization provides a new method to tune the hyperparameters for researchers (Qiao et al., 2020; Xu et al., 2021). Random Search (RS) and Grid Search (GS) (Putatunda et al., 2018) are typical automatic hyperparameter optimization methods that select the hyperparameters by random combination and traversal combination, respectively. Using RS and GS can free researchers from manual tuning. However, large amounts of computing resources are consumed because of their randomness and traversal. TPE is an enlightening automatic hyperparameter optimization method based on Bayesian Optimization (Nguyen et al., 2020), which can improve the hyperparameter combination process compared with RS and GS.

TPE algorithm assumes a set of observations obeying the Gaussian distribution that takes $\{x^{(i)}, y^{(i)}, i = 1, 2, ..., k\}$, where $x^{(i)}$ denotes the hyperparameter set, $y^{(i)}$ denotes the corresponding training loss, and k is the number of iterations. At the initial iterations, RS is employed to select the hyperparameters in search space. Then, observations are split into D_l and D_g as follows:

$$p(x|y) = \begin{cases} l(x), & y < y^* \\ g(x), & y \geqslant y^*, \end{cases} \tag{1}$$

where $l(x)$ and $g(x)$ respectively denote the probability density function (PDF) of D_l and D_g. y^* is selected to be a γ-quantile of the observation results, satisfying $p(y^* > y) = \gamma$. Therefore, y^* can be defined as the splitting point between D_l and D_g.

An Expected Improvement (EI) function is used as an objective function as follows:

$$\begin{aligned} \mathrm{EI}(x) &= \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)\, p(y)}{p(x)} \mathrm{d}y \\ &\propto \left(\gamma + \frac{g(x)}{l(x)} (1 - \gamma) \right)^{-1}. \end{aligned} \tag{2}$$

Then, TPE selects the next hyperparameters $x^{(i+1)}$ by maximizing EI. The TPE repeats the above steps until the max iteration is reached.

2.2 Bidirectional Gated Recurrent Unit (BiGRU)

2.2.1 Gated Recurrent Unit (GRU)

RNN and its variants, i.e., LSTM and GRU, have been widely used in recent years to predict the nonlinear time series data. GRU consists of an update gate z_t that decides how much the past information is carried over to the future and a reset gate r_t that determines how much the previous information is passed into the current memory content. The formulas involved in GRU are shown as follows:

$$r_t = \sigma(W_r \cdot h_{t-1} + W_r \cdot x_t), \tag{3}$$

$$z_t = \sigma(W_z \cdot h_{t-1} + W_z \cdot x_t), \tag{4}$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot x_t + W_{\tilde{h}} \cdot (r_t \odot h_{t-1})), \tag{5}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}, \tag{6}$$

where W_r, W_z, and W_h represent weight matrices; x_t represents current inputs; $\tilde{h}_t$ is the current memory content; h_t is current outputs; and h_{t-1} is the past outputs. Besides, $\sigma(\cdot)$ denotes sigmoid function, $\tanh(\cdot)$ denotes hyperbolic tangent function, and $\odot$ denotes the Hadamard product. This structure enables GRU to connect the dependencies in long-term or short-term time scale. Thus, GRU can solve the time series problems. The gate structure of GRU is shown in Fig. 1(a).

Comparing LSTM involving input gate, forget gate, and output gate, GRU only mentions two gates, thus consuming less time on training and test while performing similarly (Hochreiter and Schmidhuber, 1997).

2.2.2 Bidirectional Gated Recurrent Unit (BiGRU)

The x_t is carried by forward propagation in the proposed model as a one-way neural network. However, some connections between the previous and future data for the actual time series information may be present. BiGRU is an advanced version aiming at the bidirectional propagation by adding a bidirectional recurrent layer in original GRU that can enhance prediction precision and solve the forget problem (Wang et al., 2021).

2.2.3 Dropout algorithm

Overfitting is a critical and common problem in the training process of deep neural network. When samples of the training set are few or the hyperparameters of the model are many, overfitting may appear quickly. Specifically, high accuracy in the training set but low accuracy in the test set may be observed. A method to avoid overfitting is using the hybrid model to exert the merits of every model to increase the accuracy through time-consuming training.

A dropout algorithm can effectively alleviate overfitting (Pham et al., 2014). The cells of the hidden layer are endowed with a random probability to be invalid during the training process. The proposed probability p, also called the dropout rate, is one of the hyperparameters of the dropout algorithm. As illustrated in Fig. 1(b), some of the cells of the hidden layer are removed when applying the dropout algorithm during the actual training process (Srivastava et al., 2014). Then, each hidden cell can create useful information independently without relying on other cells.

The whole network structure is shown in Fig. 1(c).

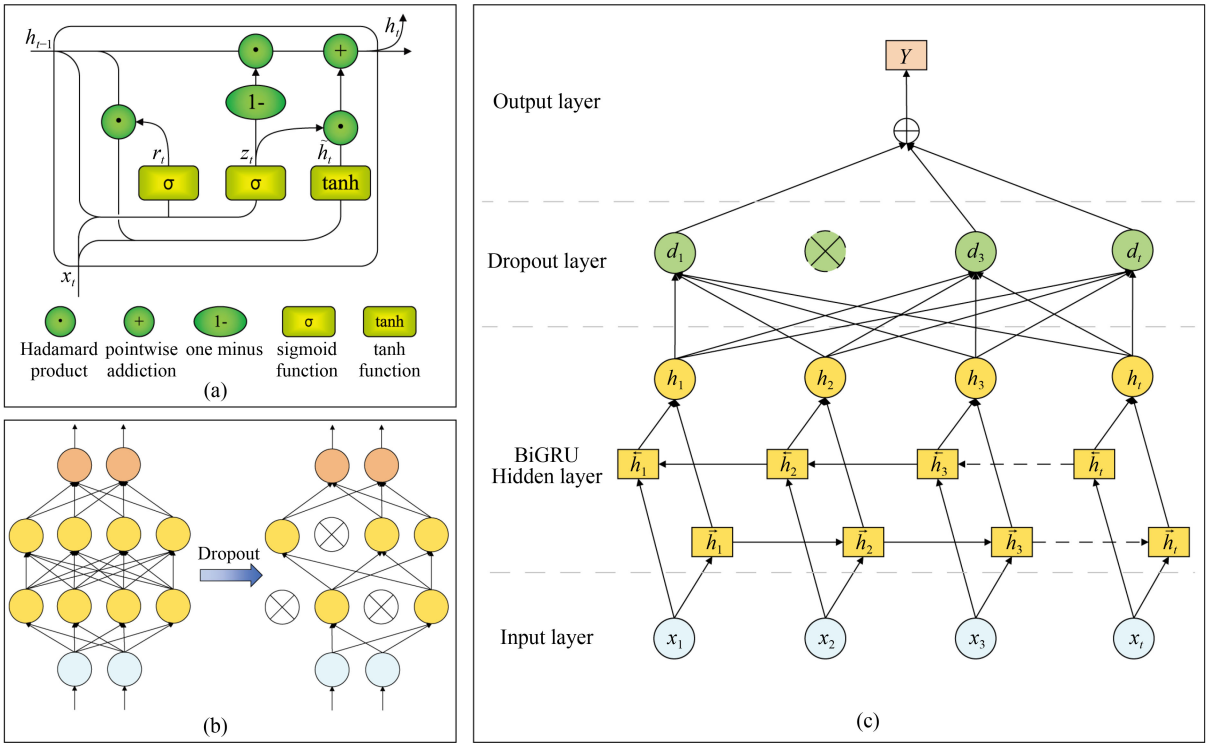


Fig. 1 Structure of proposed network and algorithm: (a) Structure of GRU; (b) A full connected network after using dropout algorithm; (c) The whole structure of BiGRU.

2.3 Gaussian Progress Regression (GPR)

GPR is a non-parametric machine learning approach based on Bayesian regression (Zhang et al., 2016). It skills at the regression of data with high-dimension, non-linear, and small samples. GPR is usually combined with neural network to obtain corresponding interval predictions to increase the accuracy and reliability of the prediction.

The construction process of GPR is as follows. The kernel function and its hyperparameters are first set to determine prior distribution. Then, the GPR model is trained to obtain the optimal hyperparameters by using the maximum likelihood estimation. The GPR model is uniquely determined once the training data, kernel function, and hyperparameters are confirmed. Lastly, the mean and variance of test data are calculated.

Suppose a training set $\{x_i, f(x_i), i=1, 2, \ldots, m\}$ and a test set $\{ x_{*j}, f(x_{*j}), j=1, 2, \ldots, n\}$ split from a time series dataset, when there is noise ε in data, GPR assumes that the noise obeys the Gaussian distribution whose mean is equal to 0 and variance is equal to σ_n^2.

$$y = f(x) + \varepsilon,\ \varepsilon \sim N(0, \sigma_n^2). \tag{7}$$

GPR assumes that observations obey the joint Gaussian distribution as follows.

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim N\left(0, \begin{bmatrix} K(x,x)+\sigma_n^2 I_N & K(x,x_*) \\ K(x_*,x) & K(x_*,x_*) \end{bmatrix}\right) = N\left(0, \begin{bmatrix} K_y & K_* \\ K_*^T & K_{**} \end{bmatrix}\right), \tag{8}$$

$$K_y = K + \sigma_n^2 I_N, \tag{9}$$

where K is the covariance matrix, i.e., kernel function, * and ** are the symbols distinguishing different covariance matrices, and I_N is the $N \times N$ unit matrix.

According to the Bayesian regression, the predictive distribution of y^* is calculated as follows:

$$P(y_*|y) = N\left(\bar{y}, \sigma_y^2\right), \tag{10}$$

$$\bar{y} = K_*^T K_y^{-1} y, \tag{11}$$

$$\sigma_y^2 = K_{**} - K_*^T K_y^{-1} K_*, \tag{12}$$

where $\bar{y}$ and σ_y^2 denote the mean and variance of the probability distribution function of the point prediction respectively.

Kernel functions include Radial Basis Function (RBF), Rational Quadratic (RQ), Square Exponential covariance function (SE) and Mattern function. The formulas are as follows:

$$K_{\mathrm{RBF}} = \exp\left(-\frac{\left\|x_i, x_j\right\|^2}{2\sigma^2}\right), \tag{13}$$

$$K_{\mathrm{RQ}} = \left(1 + \frac{\left\|x_i, x_j\right\|^2}{2\alpha l^2}\right)^{-\alpha}, \alpha, l > 0, \tag{14}$$

$$K_{\mathrm{SE}} = \exp\left(-\frac{\left\|x_i, x_j\right\|^2}{2l}\right), \tag{15}$$

$$K_{\mathrm{Mattern}} = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\left\|x_i, x_j\right\|}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}\left\|x_i, x_j\right\|}{l}\right), \tag{16}$$

where σ, α, l, and v are the hyperparameters of the kernel function.

The hyperparameters of the kernel function are optimized using the maximum likelihood estimation. The formula is shown below:

$$\begin{aligned} L(\theta) &= -\log N(0, K_{**}(\theta)) \\ &= \frac{1}{2} y^T K_{**}^{-1} y + \frac{1}{2}\log|K_{**}| + \frac{n}{2}\log(2\pi), \end{aligned} \tag{17}$$

where θ denotes the hyperparameter to be optimal.

Therefore, for GPR, the point prediction result is $\bar{y}$, and the corresponding 95% prediction interval is $[\bar{y} - 1.96\sigma_y, \bar{y} + 1.96\sigma_y]$.

2.4 Proposed model

Given the high point prediction accuracy of BiGRU and reliable interval prediction result of GPR, a hybrid model consisting of BiGRU and GPR is constructed in this study. The constructing process includes three steps. First, the original structure of the BiGRU model is built. Second, TPE is employed to optimize the hyperparameters by training BiGRU. Third, TPE selects the best trial and outputs the first prediction result, i.e., point prediction result. Then, on the basis of the first prediction data, GPR is built and outputs the second prediction results, including interval prediction and probability prediction. Finally, reliability evaluation is employed to test whether the proposed model is reliable. The complete construction process of the proposed model is shown in Fig. 2.

3 Case study

3.1 Data resources

The experimental dataset from the Internal Circulation process in a full-scale anaerobic WWTP (2500 m^3/d) located in Guangdong province, China is employed to train and test the proposed model. The features of the dataset include pH, alkalinity (ALK, mg/L), organic loading rate (OLR, kgCOD/(m^3·d)), hydraulic retention time (HRT, h), flow rate (Q_{inf}, m^3/d), and inflow chemical oxygen demand (COD_{inf}, mg/L). Labels include effluent chemical oxygen demand (COD_{eff}, mg/L) and biogas yield (m^3/d).

The pieces of equipment of the automatic monitoring system include probes (HACH® SC1000, USA) for pH, biogas yield, and ALK; flowmeters (OMEGA® FL-6101B, USA) for HRT and Q_{inf}; and COD online monitor (INESA® COD-583, China) for OLR, COD_{inf}, and COD_{eff}.

3.2 Data preprocessing

3.2.1 Feature selection

The proposed features with poor correlation between the labels are unnecessary to the training process and may cause interference, which may decrease the robustness of the model (Newhart et al., 2019; Xu et al., 2021). Thus, the unnecessary variable should be removed. A correlation analysis is conducted for all variables. The Spearman correlation coefficient (ρ) of two variables is calculated using the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{18}$$

where d_i denotes the rank difference i-th sample of two variables, and n is the number of samples. The ρ value close to 1 or −1 indicates correlation between the two variables, and ρ value higher than 0 represents positive correlation. After calculating all ρ values, a correlation coefficient matrix is generated. The result is shown in Fig. 3 with a heatmap. The factors affecting COD_{eff} and biogas yield most are COD_{inf} and OLR, respectively. The ρ value between Q_{inf} and COD_{eff} is −0.0016, showing a very low correlation. Besides, the ρ between Q_{inf} and HRT is closer to −1, demonstrating that they exist a high linear correlation and are not independent. Therefore, Q_{inf} is suggested as an unnecessary variable for training process and removed in the latter prediction.

3.2.2 Dataset split

Generally, for small scale dataset, it is split into training set and test set in a ratio of 8:2. In this study, the dataset includes 150 samples and is split into 120 (80%) samples for training and 30 (20%) samples for test. The plots for the dataset are shown in Fig. S1.

3.2.3 Normalization

Given the huge differences in magnitude between different variables, a normalization process is needed to eliminate the dimensional effects and increase the coverage rate before training the model. In this study, a typical normalization method is employed to normalize the training data to [0,1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{19}$$

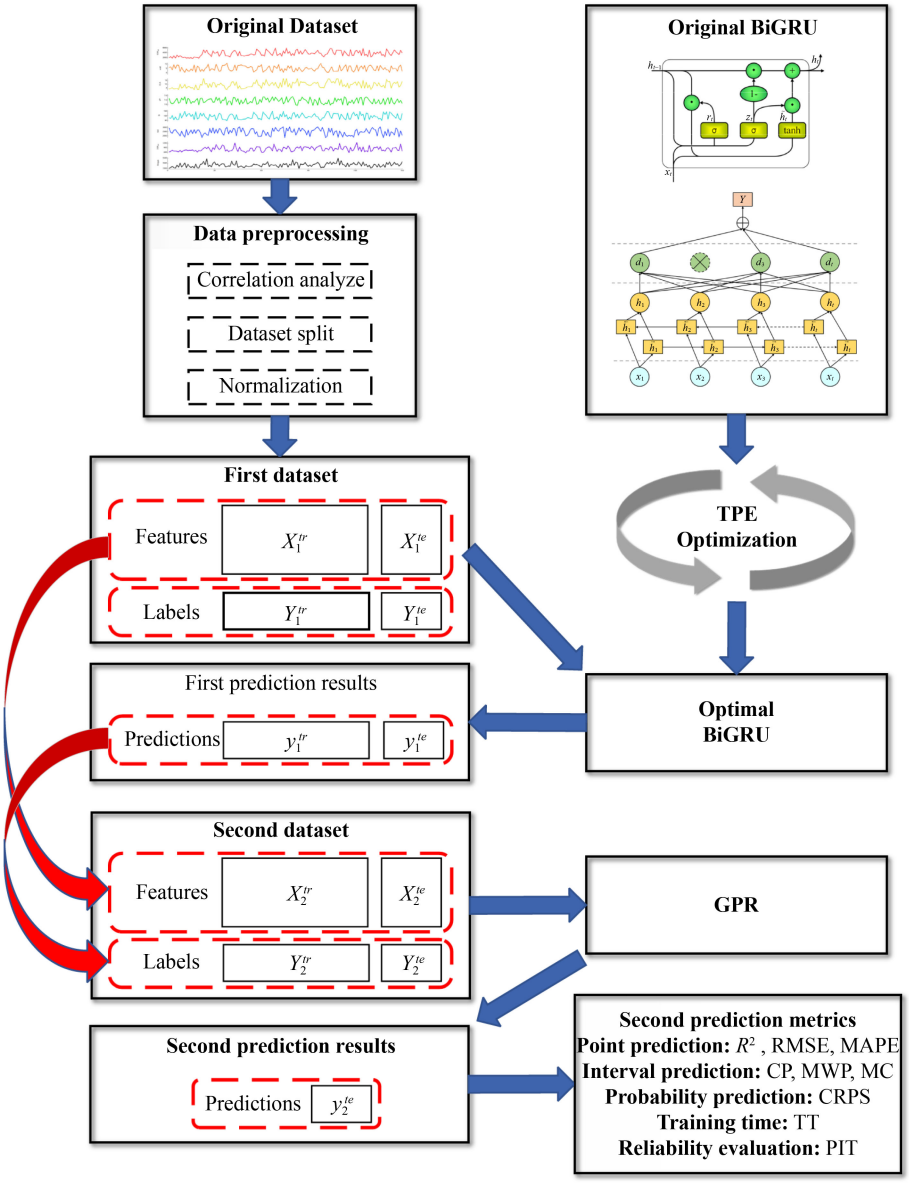


Fig. 2 Flowchart of construction process of proposed model. The blue arrows represent the construction process, and the red arrows represent dataset combination. X denotes the features, Y denotes the labels. The superscript "tr" and "te" respectively denote training set and test set. The subscript "1" and "2" respectively denote the first and the second.

where x is original value of any feature; max(x) and min(x) denote the max and min values of the corresponding features, respectively; and x' is the normalized value.

3.3 Evaluation metrics

3.3.1 Evaluation metrics of point prediction

1) Coefficient of determination (R^2)

R^2 characterizes the deviation degree between predictions and observations. It ranges from 0 to 1, and it value close to 1 indicates high point prediction accuracy. The formula is as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{20}$$

where $\hat{y}_i$ is the i-th observation, y_i is the i-th observation, $\bar{y}$ is the mean of observations, and n is the number of samples.

2) Root mean square error (RMSE)

RMSE is defined as the root of the mean of squared error between predictions and observations. RMSE close to 0 indicates high degree of congruence between predictions and observations.

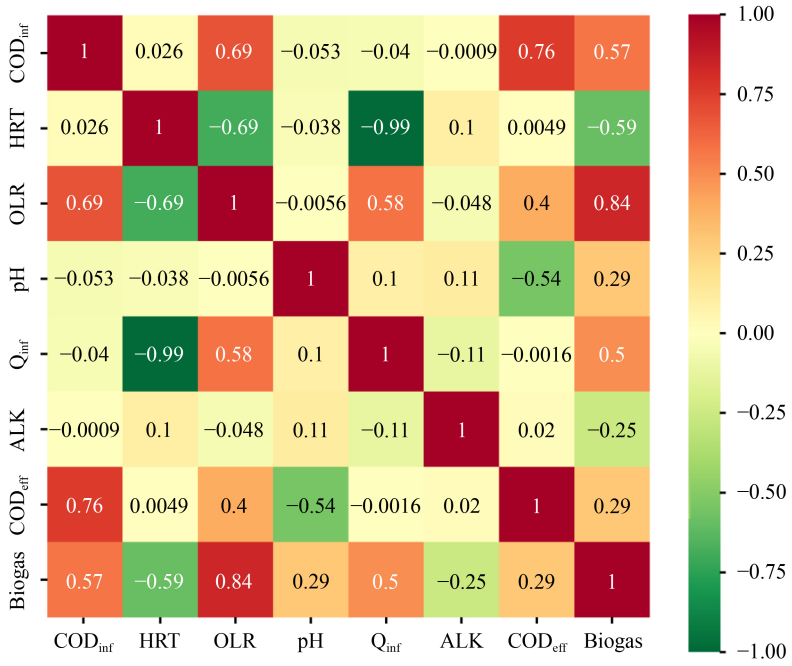


Fig. 3 A heatmap with correlation analyze.

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}. \tag{21}$$

3) Mean absolute percentage error (MAPE)

MAPE characterizes the mean deviation degree between predictions and observations. It is a perfect prediction model with MAPE equal to 0, and a poor model with MAPE higher than 100%.

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|, \tag{22}$$

4) Test time (TT)

Since the GRU has a simpler structure than LSTM but can obtain a similar prediction accuracy, TT is chosen to evaluate the time consumed of GRU. TT evaluation begins with the start of TPE and ends with the test process of BiGRU.

3.3.2 Evaluation metrics of interval prediction

1) Coverage percentage (CP)

CP is defined as the percentage of the observation that falls within the prediction interval and describes the accuracy of the interval prediction. The formula is as follows:

$$\mathrm{CP}=n_j/n, \tag{23}$$

where n_j denotes the amount of the observations that fall within the interval.

2) Mean width percentage (MWP)

MWP reflects the mean width of the prediction interval. A low MWP means high reliability of the interval prediction. The formula is as follows:

$$\mathrm{MWP}=\sum_{i=1}^{n}\frac{\left|y_{\mathrm{upper},i}-y_{\mathrm{lower},i}\right|}{y_i}, \tag{24}$$

where $y_{\mathrm{upper},i}$ and $y_{\mathrm{lower},i}$ denote the upper and lower bounds of the i-th example, respectively.

3) MC value

The interval prediction cannot be evaluated only according to CP or MWP. On the one hand, a high MWP certainly obtains a high CP value, but such interval lacks reliability and fails to offer effective uncertain information for the prediction. On the other hand, an excessively low MWP may cause a low CP (Zhang et al., 2019). Therefore, MC is employed to evaluate the interval prediction synthetically. Its definition is given as follows:

$$\mathrm{MC}=\mathrm{MWP}/\mathrm{CP}. \tag{25}$$

3.3.3 Evaluation metrics of probability prediction

Continuous ranked probability score (CRPS) is a widely used probability prediction function. It evaluates the difference in distribution between prediction and observation. A low CRPS presents small differences between the above distributions (Ferro, 2014). Its formulas are as follows:

$$\mathrm{CRPS}=\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\left[F(y_i)-H(y_i-\hat{y}_i)\right]^2\mathrm{d}y_i, \tag{26}$$

$$F(y_i)=\int_{-\infty}^{y_i}p(x)\mathrm{d}x, \tag{27}$$

$$H(\hat{y}_i - y_i) = \begin{cases} 0 & \hat{y}_i < y_i \\ 1 & others, \end{cases} \quad (28)$$

where $F(y_i)$ denotes cumulative distribution function of prediction, and $H(\hat{y}_i - y_i)$ is a step function.

3.3.4 Evaluation metrics of reliability

Probability Integral Transform (PIT) means that any continuous random variable is transformed into a uniform distribution. Its application in this study is estimating the robust reliability of the model (Jupp and Kume, 2020). If PIT values follow a uniform $U(0, 1)$ distribution, the prediction of the proposed model is accurate (Safari et al., 2020). Furthermore, a Kolmogorov significance band is used to provide a formal evaluation of the uniformity (Laio and Tamea, 2007). The Kolmogorov significance band is an area between two straight lines parallel to the diagonal and $q(\alpha)/\sqrt{n}$ away from it, where $q(\alpha)$ is a coefficient determined by the significance level α, and n is the samples of features. In this study, $\alpha = 5\%$ is used.

3.4 Study environment

The models in this study are based on the TensorFlow in Python, which contain a large amount of standard libraries of neural networks. Deep learning needs huge computing resources, and the TensorFlow provides a platform to invoke the NVIDIA GPU to accelerate the calculation by using parallel computing. Thus, the proposed models are run by the NVIDIA GPU. The configures of study environment are shown in Table S1.

4 Results and discussion

4.1 Automatic hyperparameter optimization

In this study, five hyperparameters of BiGRU, including cell size of hidden layer, dropout rate, learning rate, iteration, and batch size, are set to be optimized. The search space of the mentioned hyperparameters is shown in Table S2.

For comparison, RS is employed to optimize the same model with the same search space. The optimization trials are set to 50 for two algorithms. The optimization effect is evaluated by training accuracy in each trial and the highest R^2 of all trials.

Figs. 4(a)–4(e) show the selection of hyperparameters in each trial of TPE and RS on the BiGRU training process. Besides, R^2, the corresponding training accuracy, is illustrated in Fig. 4(f). As mentioned in Section 2.1, TPE uses RS to select the hyperparameters and split the search space into D_l and D_g in accordance with training accuracy. Thus, it can be seen in Fig. 4 that the hyperparameter selection of TPE over the first 20 trials are similar to those of RS and the R^2 is approximated. After the initialization progress, the training accuracy of RS is still unstable, while TPE quickly increases and

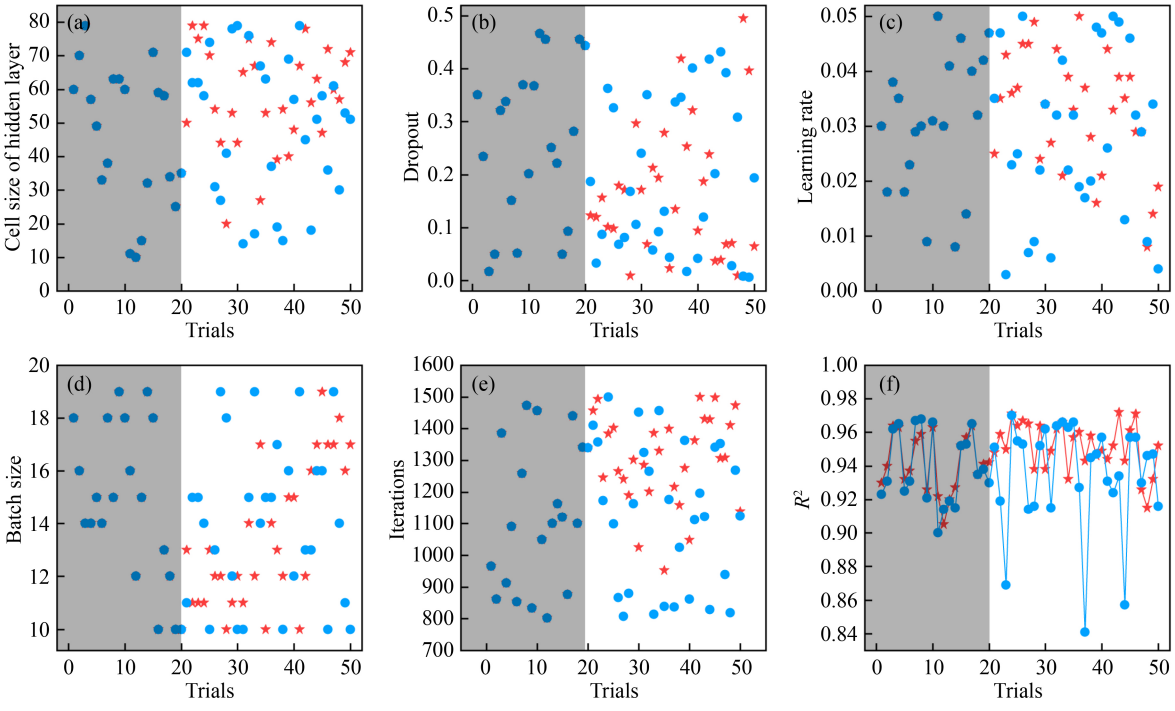


Fig. 4 Comparison over 50 trials by TPE and RS: (a)–(e) Hyperparameters selection; (f) Training accuracy. Grey area is the initial iterations. Red denotes TPE and blue denotes RS.

maintains a high level. It is because the enlightening search strategy of TPE has already reacted and the hyperparameter search focuses on D_g, whereas RS still uses the random strategy to select the hyperparameters. The accuracy metrics are given in Table 1. Compared with RS, TPE obtains 0.975, 0.954, and 0.000297 for the higher max R^2, average R^2, and lower variance of R^2, respectively. Besides, a t-test using the two-tail and two-sample tests of heteroscedasticity is employed to test whether the R^2 values of TPE and RS are significantly different. As shown in Table 1, p-value is 0.0127, smaller than 0.05, reflecting a significant difference for the R^2 between TPE and RS. In conclusion, TPE obtains a high and stable effect in hyperparameters optimization.

4.2 Point prediction result

The cell size of input layer and output layer are respectively set to five and two. The input layer is corresponding to five features and output layer is corresponding to two labels. For the optimizer of the neural network, compared with the typical gradient descent method, the adaptive moment estimation (ADAM) sets the self-adaptive learning rate by calculating the first-moment estimation and the second-moment estimation, rather than the steady learning rate. Thus, ADAM is chosen to be the optimizer of BiGRU. The activation function is set to RELU because it can maintain the convergence rate in a stationary state and avoid the gradient extinction or gradient explosion problem in tanh and sigmoid functions. Finally, mean squared error (MSE) is selected as the loss function of the model.

RNNs, including RNN, LSTM, GRU, and their variants, are designed to solve time series problems. Thus, BiRNN and BiLSTM are employed for comparison, whose dataset, hyperparameters to be optimized, and search space are the same as those of BiGRU.

Fig. 5 and Table 2 show the point prediction results and metrics of COD_{eff} and biogas yield among BiGRU, BiRNN, and BiLSTM. For COD_{eff}, the R^2, RMSE, and MAPE of BiGRU are 0.973, 24.94 mg/L, and 2.38%, respectively, showing the best performance among the three algorithms. Besides, for biogas yield, the R^2, RMSE, and MAPE of BiGRU are 0.939, 169.32 m^3/d, and 4.66%, respectively, presenting the lowest loss of the three models. It is because BiGRU can effectively avoid the long-term dependency, gradient explosion, and gradient extinction. Given that RNN has the simplest structure of the three models, its TT is the lowest. BiGRU saves 11% and 12% test time than BiLSTM because of the simpler structure. Therefore, BiGRU is certainly less time-consuming than BiLSTM for similar prediction accuracy. For point prediction, BiGRU has evident

Table 1 Comparison of R^2 over the last 30 trials between TPE and RS

Optimization methods	Max R^2	Average R^2	Variance of R^2	p value of the t-test
TPE	0.975	0.954	0.000297	0.0127
RS	0.970	0.935	0.000732	

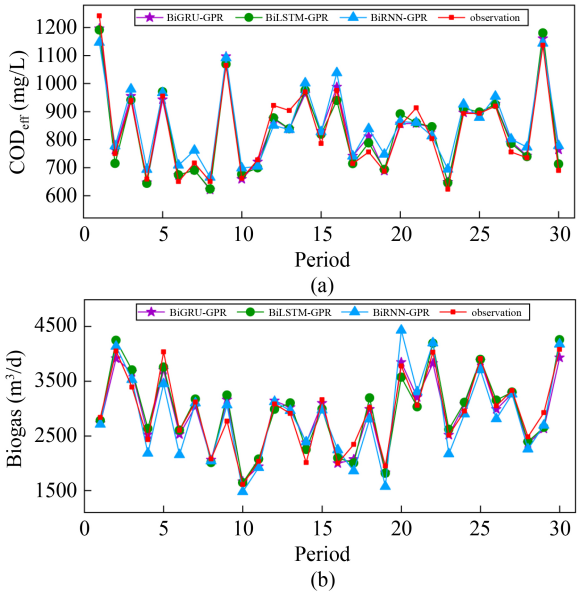


Fig. 5 Point prediction results of three models. (a) COD_{eff}; (b) Biogas.

Table 2 Metrics of three models.

Feature	Model	Point prediction			Interval prediction			Probability prediction	Test time (s)
		R^2	RMSE*	MAPE (%)	CP	MWP	MC	CRPS	
COD_{eff}	BiGRU-GPR	**0.973**	**24.94**	**2.38**	0.93	**0.12**	**0.13**	**0.0329**	2093
	BiLSTM-GPR	0.963	30.31	2.84	0.97	0.15	0.15	0.0412	2356
	BiRNN-GPR	0.899	48.54	5.22	**1**	0.24	0.24	0.0504	**1769**
Biogas	BiGRU-GPR	**0.939**	**169.32**	**4.66**	0.9	**0.24**	**0.27**	**0.0988**	1953
	BiLSTM-GPR	0.927	184.89	5.26	**0.97**	0.31	0.31	0.110	2238
	BiRNN-GPR	0.842	275.37	8.06	0.93	0.39	0.42	0.186	**1641**

Notes: *, The unit for RMSE of COD_{eff} and biogas production is mg/L and m^3/d, respectively.

dominance among the three comparison models in COD_{eff} and biogas yield prediction.

4.3 Interval prediction result

In this study, GPR is employed to obtain the prediction interval and calculate the metrics mentioned above. In GPR, different kernel functions can be selected (Kang et al., 2019). For flexibility, RBF is chosen because it is suitable for high-dimension and nonlinear data (Ozcan et al., 2016). BiGRU, BiLSTM, and BiRNN are combined with GPR for comparison. The interval prediction results of COD_{eff} and biogas yield between BiGRU, BiRNN, and BiLSTM are given in Fig. 6 and Table 2. For CP, BiLSTM-GPR simultaneously obtains 0.97, which is the highest value in COD_{eff} and biogas yield predictions. However, for MWP, BiGRU-GPR has narrower width of COD_{eff} and biogas yield prediction of 0.12 and 0.24, respectively. For the synthetical metric MC, BiGRU-GPR is 0.12 and 0.27, respectively, indicating that BiGRU-GPR achieves the best performance in interval prediction. This is because interval prediction with good performance is based on point prediction with high accuracy. The prediction interval becomes narrower and cover observation values when point prediction obtains high accuracy, thus the MC value will be higher. Therefore, the proposed BiGRU-GPR model is remarkably suitable for COD_{eff} and biogas yield prediction.

4.4 Probability prediction result

In probability prediction, the GPR model is the same as that in interval prediction. The probability prediction aims to evaluate the distribution differences between prediction and observation. Similarly, BiLSTM-GPR and BiRNN-GPR are employed for comparison with BiGRU-GPR. The probability prediction result is shown in Table 2. BiGRU-GPR obtains the lowest CRPS for COD_{eff} prediction by 0.0329 and biogas yield prediction by 0.0988. This finding is because BiGRU-GPR obtains the best performance in point prediction and interval prediction. Thus, its distributions of two predictions of BiGRU-GPR are approximated with the distributions of two observations.

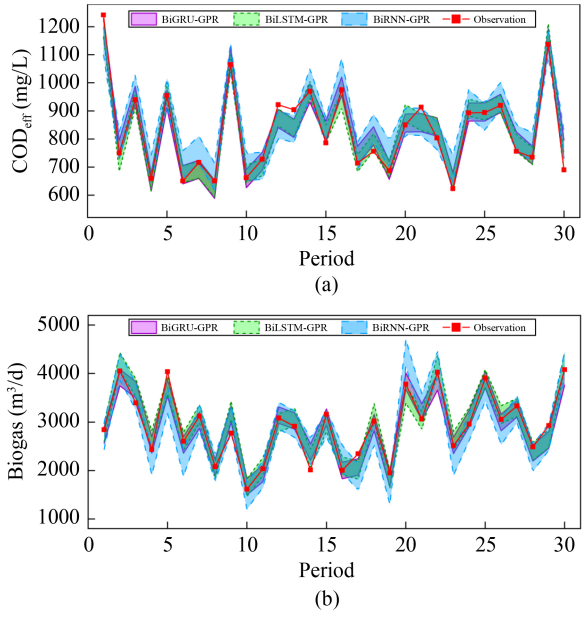


Fig. 6 Interval prediction results of three models. (a) COD_{eff}; (b) Biogas.

4.5 Reliability evaluation of BiGRU-GPR

Through the above three evaluations, BiGRU-GPR is the best model for COD_{eff} and biogas yield predictions. Then, a reliability test only for BiGRU-GPR is processed to ensure that the results are persuasive. In this test, PIT values are calculated to evaluate the reliability of the model. As shown in Fig. 7, PIT values are drawn in the plots. The points in the plots are similarly arranged in straight lines and fall into the Kolmogorov 5% significance band for COD_{eff} and biogas yield. This result demonstrates that PIT values are distributed quite uniformly and that the probability distributions are not too high or low and not too wide or narrow. Thus, the prediction of the proposed model is convincing and reliable.

4.6 Significance of the proposed model

The proposed soft-sensor model has obtained R^2 of 0.973 and 0.939 of point prediction. Also, the model provides a valid prediction interval, reflecting the fluctuation range of water quality. In addition, the model is assessed as a convinced model through probability prediction and reliability evaluation. The proposed model has performed well and can completely meet the demand of water quality prediction.

For full-scale wastewater treatment applications, the proposed model overcomes the shortcomings that some water quality metrics are unmeasurable online, thus achieving the sync output of all metrics. Besides, the neural network is used to evade the calculation of complex mechanism of anaerobic process and can provide a reliable basis for operators to control the reactor system well and avoid abnormal water quality, extra power consumption, increased carbon emission, and collapse of reaction system due to misjudgment and misoperation.

The proposed model is suitable for water quality prediction and is also skilled at solving the time series problems. A quantitative relationship also exists between water quality and carbon emission, which can be calculated in real time. Besides, energy consumption,

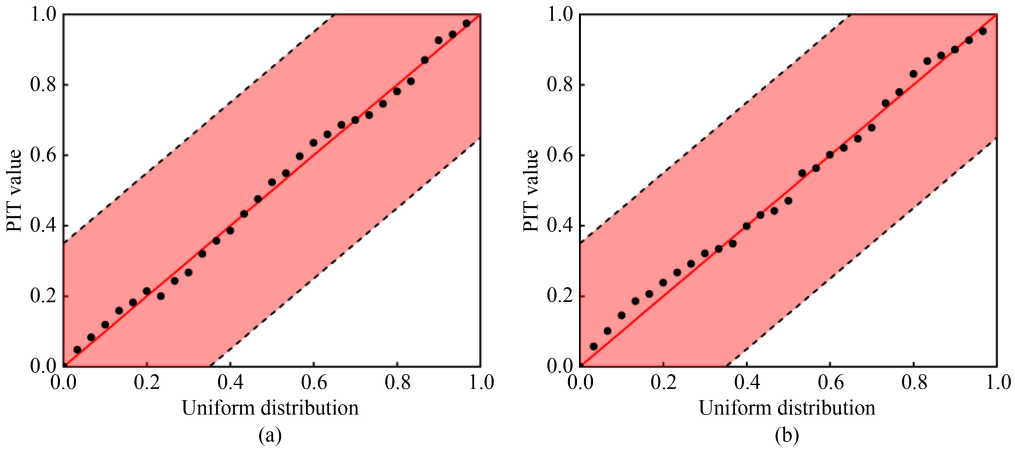


Fig. 7 Reliability evaluation of BiGRU-GPR. The red area denotes Kolmogorov 5% significance band, red diagonal line is the theoretical uniform distribution and the points are PIT values. (a) COD_{eff}; (b) Biogas.

including electricity and heat, in WWTPs can be measured or calculated in real time. Carbon emission and energy consumption data are also time series. An intuitive and precise basis for operators can be provided if the data of carbon emission and energy consumption are available, and the two can be directly predicted by the proposed model.

5 Conclusions

Water quality prediction based on deep learning, a new and developing technology for WWTPs, can precisely forecast the operation status and alleviate the adverse effect when manual management errors occur. A new water quality prediction method, BiGRU-GPR based on TPE optimization, is introduced in this article. In this method, TPE effectively avoids the time-consuming process of manual hyperparameter optimization and achieves good performance. BiGRU is adopted to obtain point prediction, and GPR obtains the uncertainty information and prediction interval. Nine evaluation metrics, including R^2, RMSE, MAPE, TT, CP, MWP, MC, CRPS, and PIT, are employed to verify the model. Research indicates the following results. 1) BiGRU has the point prediction result with high accuracy under the automatic hyperparameter optimization of TPE. 2) GPR calculates the PDF of point prediction and proves that BiGRU-GPR model has predominance in interval prediction and probability prediction in three comparison models. 3) BiGRU has good robustness because its PIT value distributes uniformly. 4) The soft-sensor model provides a reliable basis for the operator in WWTPs to control the water quality precisely for reducing carbon emission and energy consumption.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Nos. 41977300 and 41907297), the Science and Technology Program of Guangzhou (China) (No. 202002020055) and the Fujian Provincial Natural Science Foundation (China) (No. 2020I1001).

CRediT Author Contribution Statement
Junlang Li: Algorithm design, Writing – Original draft. **Zhenguo Chen**: Code debugging, Writing – Review & Editing, Investigation. **Xiaoyong Li**: Algorithm optimization, Writing – Review & Editing. **Xiaohui Yi**: Conceptualization. **Yingzhong Zhao**: Experimental data visualization. **Xinzhong He**: Dataset resources. **Zehua Huang**: Formal analysis. **Mohamed A. Hassaan**: Writing – Review & Editing. **Ahmed El Nemr:** Writing – Review & Editing. **Mingzhi Huang**: Project administration, Funding acquisition.

Data Accessibility Statement Data not available due to commercial restrictions.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11783-023-1667-3 and is accessible for authorized users.

Abbreviations

WWTPs	Wastewater Treatment Plants
SVM	Support Vector Machine
RF	Random Forest
GPR	Gaussian Process Regression
ANN	Artificial Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Time Memory
GRU	Gated Recurrent Unit
BiGRU	Bidirectional Gated Recurrent Unit
TPE	Tree-structured Parzen Estimator
RS	Random Search
GS	Grid Search
RBF	Radial Basis Function
SE	Square Exponential Covariance function
RQ	Rational Quadratic

ALK	Alkalinity
OLR	Organic Loading Rate
HRT	Hydraulic Retention Time
COD	Chemical Oxygen Demand
MSE	Mean Square Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
CP	Coverage Percentage
MWP	Mean Width Percentage
CRPS	Continuous Ranked Probability Score
CDF	Cumulative Distribution Function
PDF	Probability Density Function
PIT	Probability Integral Transform

References

Chen Q L, Chai W, Qiao J F, IEEE (2010). Modeling of Wastewater Treatment Process Using Recurrent Neural Network. Jinan: IEEE, 5872–5876

Ching P M L, So R H Y, Morck T (2021). Advances in soft sensors for wastewater treatment plants: a systematic review. Journal of Water Process Engineering, 44: 102367

Darvishi H, Ciuonzo D, Eide E R, Rossi P S (2021). Sensor-fault detection, isolation and accommodation for digital twins via modular data-driven architecture. IEEE Sensors Journal, 21(4): 4827–4838

Di Maria F, Micale C (2015). The contribution to energy production of the aerobic bioconversion of organic waste by an organic Rankine cycle in an integrated anaerobic-aerobic facility. Renewable Energy, 81: 770–778

Ferro C A T (2014). Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 140(683): 1917–1923

Han H G, Zhang J C, Du S L, Sun H Y, Qiao J F (2021). Robust optimal control for anaerobic-anoxic-oxic reactors. Science China. Technological Sciences, 64(7): 1485–1499

Hauck M, Maalcke-Luesken F A, Jetten M S M, Huijbregts M A J (2016). Removing nitrogen from wastewater with side stream anammox: What are the trade-offs between environmental impacts? Resources, Conservation and Recycling, 107: 212–219

Heydari B, Sharghi E A, Rafiee S, Mohtasebi S S (2021). Use of artificial neural network and adaptive neuro-fuzzy inference system for prediction of biogas production from spearmint essential oil wastewater treatment in up-flow anaerobic sludge blanket reactor. Fuel, 306: 121734

Hochreiter S, Schmidhuber J (1997). Long short-term memory. Neural Computation, 9(8): 1735–1780

Jiang Y, Yin S, Dong J, Kaynak O (2021). A review on soft sensors for monitoring, control, and optimization of industrial processes, control, and optimization of industrial processes. IEEE Sensors Journal, 21(11): 12868–12881

Jupp P E, Kume A (2020). Measures of goodness of fit obtained by almost-canonical transformations on Riemannian manifolds. Journal of Multivariate Analysis, 176: 104579

Kadlec P, Gabrys B, Strandt S (2009). Data-driven Soft Sensors in the process industry. Computers & Chemical Engineering, 33(4): 795–814

Kang L, Chen R S, Xiong N, Chen Y C, Hu Y X, Chen C M (2019). Selecting hyper-parameters of gaussian process regression based on non-inertial particle swarm optimization in Internet of things. IEEE Access: Practical Innovations, Open Solutions, 7: 59504–59513

Kim M, Yang Y N, Morikawa-Sakura M S, Wang Q H, Lee M V, Lee D Y, Feng C P, Zhou Y L, Zhang Z Y (2012). Hydrogen production by anaerobic co-digestion of rice straw and sewage sludge. International Journal of Hydrogen Energy, 37(4): 3142–3149

Laio F, Tamea S (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrology and Earth System Sciences, 11(4): 1267–1277

Li X Y, Yi X H, Liu Z H, Liu H B, Chen T, Niu G Q, Yan B, Chen C, Huang M Z, Ying G G (2021). Application of novel hybrid deep leaning model for cleaner production in a paper industrial wastewater treatment system. Journal of Cleaner Production, 294: 126343

Newhart K B, Holloway R W, Hering A S, Cath T Y (2019). Data-driven performance analyses of wastewater treatment plants: a review. Water Research, 157: 498–513

Nguyen H P, Liu J, Zio E (2020). A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators. Applied Soft Computing, 89: 106116

Ozcan G, Pajovic M, Sahinoglu Z, Wang Y B, Orlik P V, Wada T, IEEE (2016). Online State of Charge Estimation for Lithium-Ion Batteries Using Gaussian Process Regression. Florence: IEEE, 998–1003

Pham V, Bluche T, Kermorvant C, Louradour J (2014). Dropout Improves Recurrent Neural Networks for Handwriting Recognition. Hersonissos, Greece: IEEE, 285–290

Putatunda S, Rama K, Acm (2018). A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. Shanghai: ACM

Qiao S, Wang Q, Zhang J, Pei Z (2020). Detection and classification of early decay on blueberry based on improved deep residual 3D convolutional neural network in hyperspectral images. Scientific Programming, 2020: 1–12

Safari M a M, Masseran N, Majid M H A (2020). Robust reliability estimation for lindley distribution: a probability integral transform statistical approach. Mathematics, 8(9): 1634

Samuelsson O, Björk A, Zambrano J, Carlsson B (2017). Gaussian process regression for monitoring and fault detection of wastewater treatment processes. Water Science and Technology, 75(12): 2952–2963

Şenol H (2021). Methane yield prediction of ultrasonic pretreated sewage sludge by means of an artificial neural network. Energy, 215: 119173

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15: 1929–1958

Szelag B, Gawdzik A, Gawdzik A (2017). Application of selected

methods of black box for modelling the settleability process in wastewater treatment plant. Ecological Chemistry and Engineering S-Chemia I Inzynieria Ekologiczna S, 24(1): 119–127

Wang H T, Yang Y, Keller A A, Li X, Feng S J, Dong Y N, Li F T (2016). Comparative analysis of energy intensity and carbon emissions in wastewater treatment in USA, Germany, China and South Africa. Applied Energy, 184: 873–881

Wang J, Cui Q, Sun X (2021). A novel framework for carbon price prediction using comprehensive feature screening, bidirectional gate recurrent unit and Gaussian process regression. Journal of Cleaner Production, 314: 128024

Wei J P, Liang G F, Alex J, Zhang T C, Ma C B (2020). Research progress of energy utilization of agricultural waste in China: Bibliometric analysis by citespace. Sustainability (Basel), 12(3): 812

Wu X, Wang Y, Wang C, Wang W, Dong F (2021). Moving average convergence and divergence indexes based online intelligent expert diagnosis system for anaerobic wastewater treatment process. Bioresource Technology, 324: 124662

Xu Y, Gao W, Qian F, Li Y (2021). Potential analysis of the attention-based LSTM model in ultra-short-term forecasting of building HVAC energy consumption. Frontiers in Energy Research, 9: 730640

Yaginuma K, Tanabe S, Kano M (2022). Gray-box soft sensor for water content monitoring in fluidized bed granulation. Chemical & Pharmaceutical Bulletin, 70(1): 74–81

Zeng G M, Li X D, Jiang R, Li J B, Huang G H (2006). Fault diagnosis of WWTP based on improved support vector machine. Environmental Engineering Science, 23(6): 1044–1054

Zhang C, Wei H, Zhao X, Liu T, Zhang K (2016). A Gaussian process regression based hybrid approach for short-term wind speed prediction. Energy Conversion and Management, 126: 1084–1092

Zhang Z, Ye L, Qin H, Liu Y, Wang C, Yu X, Yin X, Li J (2019). Wind speed prediction method using shared weight long short-term memory network and gaussian process regression. Applied Energy, 247: 270–284