Article

Prediction model of wastewater pollutant indicators based on combined normalized codec

Chun-Ming Xu [1], Jia-Shuai Zhang [2,3], Ling-Qiang Kong [1], Xue-Bo Jin [2,3,*], Jian-Lei Kong [2,3], Yu-Ting Bai [2,3], Ting-Li Su [2,3], Hui-Jun Ma [2,3], and Prasun Chakrabarti [4]

[1] School of Light Industry, Beijing Technology and Business University, Beijing 100048, China
[2] Artificial Intelligence College, Beijing Technology and Business University, Beijing 100048, China
[3] China Light Industry Key Laboratory of Industrial Internet and Big Data, Beijing Technology and Business University, Beijing 100048, China
[4] ITM (SLS) Baroda University, Vadodara, Gujarat, India
xucm@th.btbu.edu.cn (C.-M.X.); zhangjiashuai@btbu.edu.cn (J.-S.Z.); acekong@sina.com (L.-Q.K.); jinxuebo@btbu.edu.cn (X.-B.J.); kongjianlei@btbu.edu.cn (J.-L.K.); baiyuting@btbu.edu.cn (Y.-T.B.); sutingli@btbu.edu.cn (T.-L.S.); mahuijun@th.btbu.edu.cn(H.-J.M.); drprasun.cse@gmail.com(P.C.)
* Correspondence: jinxuebo@btbu.edu.cn (X.-B.J.)

Abstract: Effective prediction of wastewater treatment is beneficial for precise control of wastewater treatment processes. The nonlinearity of pollutant indicators such as COD and TP makes the model difficult to fit and has low prediction accuracy. The classical deep learning methods have been shown to perform nonlinear modeling. However, there are enormous numerical differences between multi-dimensional data in the prediction problem of wastewater treatment, such as COD above 3000 mg/L and TP around 30 mg/L. It will make current normalization methods challenging to handle effectively, leading to the training failing to converge and the gradient disappears or exploding. This paper proposes a multi-factor prediction model based on deep learning. The model consists of a combined normalization layer and a codec. The combined normalization layer combines the advantages of three normalization calculation methods: z-score, Interval, and Max, which can realize the adaptive processing of multi-factor data, fully retain the characteristics of the data, and finally cooperate with the codec to learn the data characteristics and output the prediction results. Experiments show that the proposed model can overcome data differences and complex nonlinearity in predicting industrial wastewater pollutant indicators and achieve better prediction accuracy than classical models.

Keywords: wastewater treatment; combinatorial normalization; codec; pollutant indicators; predict

MSC:

1. Introduction

In order to protect water resources and reduce the pollution of production and domestic wastewater to the environment, it is necessary to reduce the discharge of pollutants through the harmless treatment of wastewater [1]. Therefore, the effect of wastewater treatment has received extensive attention, and innovative technologies and management methods have become a current research focus.

Anaerobic biological treatment technology, also known as anaerobic digestion (AD), is widely used in the sewage treatment link of wastewater treatment plants (WWTPs) [2]. Its processing cost includes anaerobic granular sludge (AnGS) bed reactors, e.g., the upflow anaerobic sludge blanket (UASB) reactor, the expanded granular sludge bed (EGSB) reactor, and the internal circulation (IC) reactor [3], etc. Due to the complexity of sludge composition, its application has limitations, mainly in the inability to fully use functional anaerobic microorganisms, resulting in a slow hydrolysis rate and poor biodegradability [4]. Although ultrasonic irradiation and other methods can improve the efficiency of anaerobic treatment, improper use of parameters will inhibit sludge metabolism and affect the economy of wastewater treatment [5]. Moreover, the anaerobic biological action in the reactor is vulnerable to the impact of influent water, and thus the action is reduced. For example, during heavy rain, the anaerobic biological reactor is under a high hydraulic load to treat low-concentration sewage. It will lead to a period of famine.

In the case of industrial wastewater, the influent composition and flow rate are more prone to large fluctuations or even complete disruptions, affecting the microbial activity and the treatment capacity of wastewater treatment systems [2]. A large amount of surplus sludge will be discharged with the effluent, affecting the environment [6, 7]. Therefore, effectively removing sludge from anaerobic reactors or reducing sludge production has become an essential topic in recent years. Another disadvantage is that the removal rate of nitrogen and phosphorus is low. The enhancement of endogenous microbial metabolism will also promote the release of nutrients such as nitrogen and phosphorus in microbial cells, increasing the nutrients in the water and affecting the removal efficiency of nitrogen and phosphorus [7].

Anaerobic/aerobic conditions (A/O) biological nitrogen removal process is a biological sewage treatment system composed of anoxia and aerobic reaction. After the sewage enters the anoxic pool, it successively goes through the stages of anoxic denitrification, aerobic removal of organic matter, and nitrification. The advantages of the A/O process are lower operating costs, higher organic matter removal efficiency, less aerobic sludge, and no pH correction [8]. In the process of aerobic sludge treatment, the endogenous respiration rate is high, so the content of aerobic sludge in the effluent is small [7].

Aerobic/anoxic/anaerobic conditions(A/A/O), an anoxic tank was added to the A/O. Part of the mixed liquid from the aerobic tank was returned to the front of the anoxic tank to achieve the purpose of nitrification and denitrification. It can keep the function of nitrogen and phosphorus removal of activated sludge to the maximum extent. Moreover, the standby time is greatly improved, quickly recovering the activity when the wastewater is fed back [9]. This combination process combines the advantages of each of the three reactors. The combined process is more energy efficient. Although most chemical oxygen demand (COD) and suspended solids can be removed under anaerobic and anoxic conditions, the aerobic process can further reduce the concentration of pollutants in the wastewater [10].

China has strict discharge standards for wastewater pollutants and has limited water quality indicators such as COD and suspended solids (SS) in the treated wastewater. Take the beer industry pollutant discharge standard (GB19821-2005) [11] as an example: COD, SS, total nitrogen (TN), and total phosphorus (TP) should be lower than 80 mg/L, 70 mg/L, 15 mg/L, and 3 mg/L respectively. To ensure that the wastewater can be discharged up to the standard, some studies consider using the time series prediction method to model and predict the COD and other indicators at historical moments to provide a basis for adjusting treatment strategies.

The modeling methods commonly used in current research are divided into machine learning [12] and deep learning methods [13]. Machine learning methods, such as K-nearest neighbor (KNN), artificial neural network (ANN), etc., have the advantages of convenient modeling and few parameters and have specific applications in some simple prediction tasks. However, in the face of multi-factor and complex nonlinear data, its prediction accuracy is difficult to meet expectations. The deep learning methods are currently the most widely used methods, mainly including recurrent neural network (RNN), long short-term memory (LSTM) neural network, and so on. Deep learning relies on data-driven modeling and has a solid fitting ability. It usually obtains better results than machine learning methods in robust nonlinear and random modeling [14].

However, in the prediction of wastewater treatment indicators, classical deep learning methods also face some difficulties [15]. The first is the difficulty of data processing. Because wastewater treatment requires multi-factor forecasting, many forecasted indicators and the values of each indicator vary greatly, making it difficult for a single normalization method to achieve sound treatment effects for all indicators. The second is the high data complexity. In prediction tasks, it is often necessary to learn from long historical data, coupled with the nonlinearity and strong randomness of the data, which seriously affects the model's prediction accuracy.

The solution to this problem is to modify the normalization processing part of the model so that the data can be reasonably limited to a specific range, reducing the complexity of the data and speeding up the convergence of the model. The current research considers adaptive normalization layer, automatic selection of normalization layer, etc., and adopts a data-driven way to select a suitable normalization method adaptively. However, these improvements are primarily for univariate forecasting, and the final calculation method is still one. In the prediction task with multiple factors and significant data differences, it is effective to consider multiple normalization processing methods.

In summary, this paper considers a combined normalization codec (CNC) model for predicting water quality indicators in wastewater treatment. The model consists of a combined normalization layer, a denormalization layer, and a codec. The advantages of the processing method can be improved to improve the model's prediction accuracy.

Our main contributions are summarized as follows:

(1) A combined normalized encoder structure is proposed for the multi-factor prediction problem of wastewater pollutant indicators. This structure combines the advantages of three normalization methods, which can adaptively normalize and encode pollutant index data of different magnitudes, simplify complex index data processing processes, and improve the data processing capability in multi-factor prediction.

(2) A combined renormalized decoder structure is proposed for the prediction task. The structure uses three renormalization methods to adaptively renormalize the output value of the decoder and map to obtain the real prediction result. Its feature of adaptively adjusting parameters in model optimization can improve model prediction accuracy.

The rest of this paper is organized as follows: Section 2 introduces related research work in this area, Section 3 describes the proposed method in detail, Section 4 validates and analyzes the proposed model through experiments, and Section 5 summarizes the work and suggests future work. The abbreviations used in this paper are shown in Table 1.

Table 1. List of abbreviations.

Full name	**Abbreviation**
anaerobic digestion	AD
wastewater treatment plants	WWTPs
anaerobic granular sludge	AnGS
up-flow anaerobic sludge blanket	UASB
expanded granular sludge bed	EGSB
internal circulation	IC
anaerobic/aerobic conditions	A/O
aerobic/anoxic/anaerobic conditions	A/A/O
chemical oxygen demand	COD
suspended solids	SS
total nitrogen	TN
total phosphorus	TP
K-nearest neighbor	KNN
artificial neural network	ANN
recurrent neural network	RNN
long and short-term memory	LSTM
combined normalization codec	CNC
extreme learning machine	ELM
least squares support vector machine	LS-SVM
convolutional neural networks	CNN
shared weight long short-term memory	SWLSTM
Gaussian process regression	GPR
exponential weighted moving average	EWMA
deep neural network	DNN

gated recurrent unit	GRU
mean square error	MSE
mean absolute error	MAE
mean absolute percentage error	MAPE
Pearson correlation coefficient	R

2. Related Work

Currently, some studies use machine learning methods to predict the quality of wastewater treatment. Arismendy et al. [16] developed an intelligent system based on multilayer perceptrons. The system can predict the COD index to support the relevant decision-making of the sewage treatment plant. Hilal et al. [17] used the BKNN-ELM model combining KNN and extreme learning machine (ELM) to predict the SS index, and the prediction accuracy reached 93.56%. Liu et al. [18] used the least squares support vector machine (LS-SVM) to build a prediction model, which was validated in the COD prediction of an anaerobic wastewater treatment system. These models based on machine learning can complete the prediction of water quality indicators in practice but generally target a single factor. Because the models are relatively simple, the prediction accuracy still needs to be improved.

Therefore, there are studies considering prediction models based on deep learning. Han et al. [19] used an adaptive fuzzy neural network to achieve multi-objective predictive control. They dealt with conflicting control objectives by capturing the nonlinear behavior of the sewage treatment plant to improve its operational performance of the sewage treatment plant. Farhi et al. [20] used LSTM to build a wastewater prediction model, which showed better results than machine learning in predicting ammonia and nitrate concentrations in wastewater. Wan et al. [21] integrated the spatial feature of convolutional neural networks (CNN), the temporal feature of sharing-weight long short-term memory (SWLSTM), and the probabilistic reliability of Gaussian process regression (GPR) to construct a new water quality prediction CSWLSTM-GPR. And it is applied to high-precision point prediction and interval prediction monitoring of papermaking wastewater treatment systems.

These applications demonstrate the superiority of deep learning methods in wastewater treatment quality prediction. However, with the increase in pollutant index modeling needs and training data, deep learning methods also expose some problems. When faced with multiple factors and numerical differences, due to the enormous amount of training data, the existing data processing methods are complicated to operate and difficult to meet the processing requirements. Studies have shown improper normalization can significantly affect model performance, reducing model generalization and prediction accuracy [22]. Therefore, more efficient data processing methods must be adopted to cope with the growing demand for forecasting [23].

Passalis et al. [24] designed an adaptive normalization layer based on the z-score normalization method and applied it to the field of time series forecasting. The model adaptive optimization method can achieve better processing results than a fixed normalization scheme. Since this study only considers one basic normalization method, it is challenging to adapt widely to multiple forecasting scenarios. Jin et al. [25] combined z-score, Interval, decimal, and Min-Max normalization methods to design the normalization layer and renormalization layer and obtained the best predictions for a greenhouse weather dataset.

Based on the above analysis, this paper proposes the CNC model in combination with the actual characteristics of the deep learning state estimation method. In this paper, the combined normalization method is adopted, the advantages of various normalization methods are integrated, the data processing effect is improved, and the normalization layer and renormalization layer for the prediction task of wastewater treatment indicators are designed.

3. Combined normalized codec prediction model

The structure of the proposed combined normalized codec prediction model is shown in Figure 1. The model contains a variety of data normalization methods, which can adaptively integrate the advantages of multiple data processing methods through the end-to-end model optimization process. Thereby, the learning effect of the model on multi-dimensional data is improved, and the purpose of improving the prediction accuracy is finally achieved.

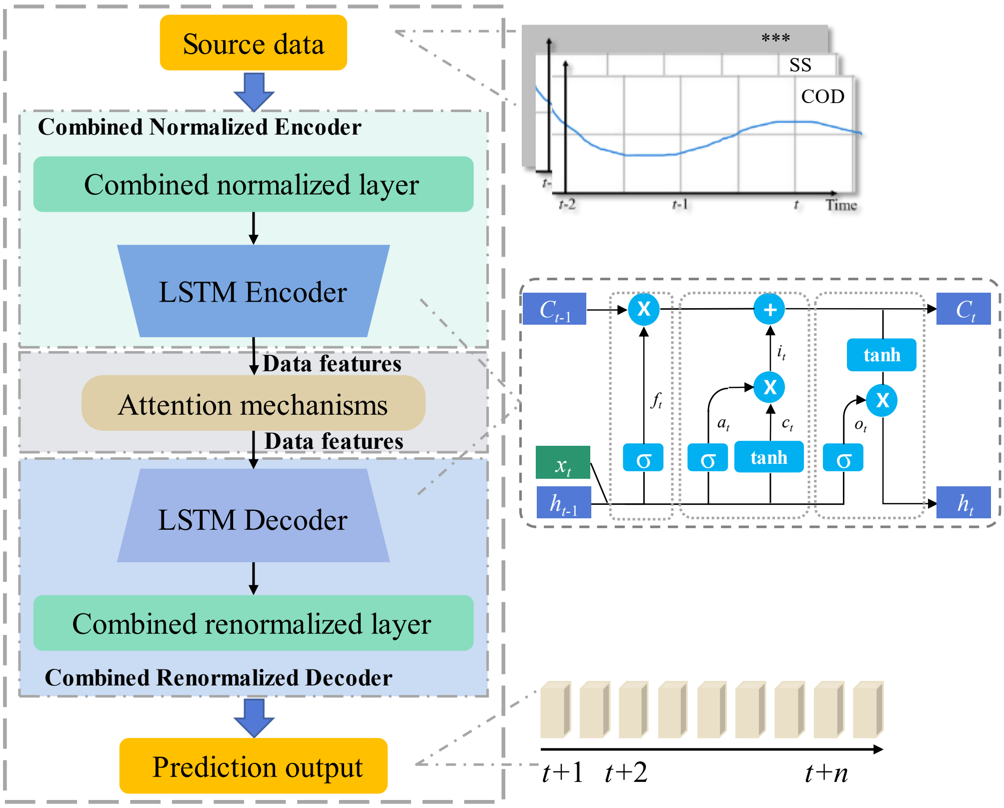


Figure 1. Schematic diagram of the model structure.

The CNC model is mainly composed of three parts: combined normalization encoder, attention mechanism, and combined renormalization decoder. The combined normalization encoder integrates an adaptive combined normalization layer containing three normalization calculation methods: z-score, Interval, and Max normalization. During the model training process, the unprocessed pollutant indicator data are directly input into the adaptive combined normalization layer in batches. Three normalized calculations are obtained by separately obtaining the batch data's mean, variance, and other statistics. In order to synthesize the advantages of the three calculation methods and get the optimal processing effect, the results of the three normalization calculations are weighted and selected based on the Softmax function. The weights are obtained from the model training to finally generate the weighted normalized processing results. These results are scaled and panned by the learnable parameter α β s that can be dynamically adjusted according to the current model training effect. The exponential weighted average method is used to fit the global distribution of the data, and the iterative estimation is performed according to the statistics of each batch of data. The optimal global statistics are retained, and the prediction accuracy of the data by the final training model is improved. The normalized data are encoded by a multilayer LSTM [26].

The attention mechanism [27] focuses on the encoded features, selecting the most favorable features for the model output values and ignoring the unimportant ones, thus reducing the model's internal parameters and learning more distant historical information. The features filtered by the attention mechanism are fed into the combined renormalization decoder.

The combined renormalization decoder decodes the data features. The decoding of features is mainly achieved by multilayer LSTMs containing sophisticated gating mechanisms that preserve and learn long-term information about the sequence. After decoding the prediction values, the final prediction values are output through the adaptive combined renormalization layer. Corresponding to the adaptive combined normalization layer, this layer contains three renormalization algorithms, which respectively perform renormalization calculation on the output features of the LSTM according to the statistics during data normalization. This layer also uses the Softmax function to weight the three sets of renormalized results and comprehensively considers the three sets of results through the trainable combined weights to obtain the best estimation results. Also, this layer adds similar trainable parameters λ and ν as the normalization layer to scale and translate the results, and the values of λ and ν can also be trained by backpropagation. The structure of the switched normalization encoder and the switched renormalization decoder will be described below.

3.1. Combined normalized encoder

The schematic structure of the combined normalized encoder is shown in Figure 2. The combined normalization encoder integrates the combined normalization layer on top of the conventional encoder. Therefore, it can automatically switch the normalization method of the input data and improve the effect of normalization processing, and finally improve the feature coding capability of the encoder. In the combined normalization layer, we use a normalization method that includes z-score, Interval and Max. It is calculated as respectively:

$$\hat{x} = \frac{x - mean}{\sqrt{\sigma^2 + \Delta}} \tag{1}$$

$$\hat{x} = a + \frac{(b - a)(x - min)}{max - min} \tag{2}$$

$$\hat{x} = \frac{x}{|x|_{max}} \tag{3}$$

where x represents the input data, $\hat{x}$ represents the normalized calculation result. min, max, $mean$, σ^2 represent the minimum, maximum, mean, and variance of the source data, respectively, and *a*, *b* represents the normalized interval. △ represents a fixed, smaller positive number.

Each of the three normalizations has its strengths and can process the input data to the standard normal distribution, (*a*, *b*) specific interval, and between (-1, 1), respectively, to exert different effects on the data. Among them, z-score processing can obtain data conforming to the standard normal distribution and reduce data distribution differences [28]; Interval method processing fixes the results in a specific interval to prevent gradient disappearance and gradient explosion problems; Max is scaling normalization scales down the input data without changing the scale characteristics of the input data.

In order to use the effect of the three normalization methods on the input data, this paper uses the adaptive combined normalization method to weigh the calculation results of normalization and determine the most suitable normalization calculation method. In the combined normalization layer, the Softmax function acts as a combined function and is calculated as follows:

$$\text{Sotfmax}(t_i) = \frac{e^{t_i}}{\sum_{i=1}^{n} e^{t_i}} \tag{4}$$

where *t* is the trainable parameter. It can optimize end-to-end by error backpropagation and is dynamically adjusted according to the model training effect. In this paper, three trainable parameters are set to output the combined weights for the results of the three normalization calculations to enhance the effectiveness of the combined normalization method. The calculation formula for combining using the Softmax function [29] is:

$$X\text{=Softmax}(t_1)\otimes x_1+\text{Softmax}(t_2)\otimes x_1+\text{Softmax}(t_3)\otimes x_3 \qquad (5)$$

where t_1, t_2 and t_3 denote the three selected trainable parameters, x_1, x_2 and x_3 denote the results obtained from the three normalization calculations, Softmax denotes the Softmax function, X represents the final output, and $\otimes$ denotes matrix multiplication.

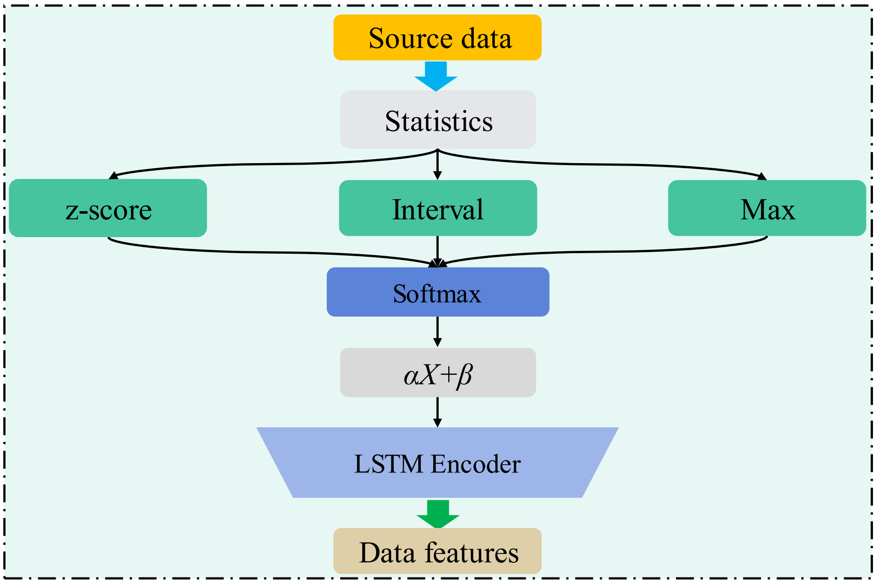


Figure 2. Combined normalized encoder structure.

In order to make the output of combined normalization better adaptable to complex data, in this paper, the trainable parameters α and β are used as scaling and translation factors, respectively. The two parameters can be back-propagated to be trained and updated end-to-end during training. The output of the combined normalization method is adjusted according to the training effect. The trainable parameters are calculated as:

$$Y=\alpha X+\beta \qquad (6)$$

where Y denotes the output of the normalized layer of the batch, X denotes the value of the batch after normalization calculation, α is the scaling factor, and β is the translation factor. Finally, the combined normalized output adjusted by trainable parameters is encoded by an encoding structure composed of LSTMs to obtain the encoded features.

In the model training, in order to grasp the global distribution of the data according to the batch data and ensure the fitting effect of the model to the input data at the end of the training, this paper uses the exponential weighted moving average (EWMA) method [30] to iteratively estimate the statistics of each batch and record the optimal statistical distribution. It is calculated as:

$$\begin{aligned}
&running\_min_t=m*running\_min_{t-1}+(1-m)*min_t\\
&running\_max_t=m*running\_\text{max}_{t-1}+(1-m)*\text{max}_t\\
&running\_mean_t=m*running\_mean_{t-1}+(1-m)mean_t\\
&running\_\sigma_t^2=m*running\_\sigma_{t-1}^2+(1-m)\sigma^2{}_t
\end{aligned} \qquad (7)$$

where min_t, max_t, $mean_t$ and σ_t^2 denote the minimum, maximum, mean, and variance statistics of the batch data at the moment t. $running\_min_t$ and $running\_min_{t-1}$ denote the estimates of the minimum value at the moment with t and t-1, $running\_max_t$ and $running\_max_{t-1}$ denote the estimates of the maximum value at the moment with t and t-1, $running\_mean_t$ and $running\_mean_{t-1}$ denote the estimates of the mean value at the moment with t and t-1, $running\_\sigma_t^2$ and $running\_\sigma_{t-1}^2$ denote the estimates of the variance at the moment with *t* and *t*-1, and m denotes the weight of retaining the information of the previous moment, respectively. In this paper, we set it to 0.6. The flow of the algorithm for combined normalization layer is shown in Algorithm 1.

Algorithm 1: Combined normalization Layer

```
Input: data: $R$={$x_1$,…,$x_m$}, Interval: $a,b$, Forgetting weight: $m$,
Learning parameters: $\alpha,\beta,t_1,t_2,t_3$
Output: $\{y_i = \mathrm{SNLayer}_{\alpha,\beta}(x_i)\}$
$\min \leftarrow x_{\min}$, $\max \leftarrow x_{\max}$, $\mu_R \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$, $\sigma_R^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i\text{-}\mu_R)^2$, $d = 10\,\hat{}\,\lceil \log_{10}|x|_{\max} \rceil$
$\mathrm{Softmax}(t_1) = \frac{e^{t_1}}{e^{t_1}+e^{t_2}+e^{t_3}}$, $\mathrm{Softmax}(t_2) = \frac{e^{t_2}}{e^{t_1}+e^{t_2}+e^{t_3}}$, $\mathrm{Softmax}(t_3) = \frac{e^{t_3}}{e^{t_1}+e^{t_2}+e^{t_3}}$
$running\_max_t \leftarrow m * running\_\mathrm{max}_{t-1} + (1-m) * \max$
$running\_min_t \leftarrow m * running\_min_{t-1} + (1-m) * min$
$running\_mean_t \leftarrow m * running\_mean_{t-1} + (1-m)\mu_R$
$running\_var_t \leftarrow m * running\_var_{t-1} + (1-m)\sigma_R^2$
$running\_d_t \leftarrow m * running\_d_{t-1} + (1-m) * d$
$output_1 \leftarrow \frac{x_i - running\_mean_t}{\sqrt{running\_\mathrm{var}_t + 1 \times 10^{-5}}}$
$output_2 \leftarrow a + \frac{(b\text{ - }a)(x_i\text{ - }running\_min)}{running\_max\text{ - }running\_min}$
$output_3 \leftarrow \frac{x_i}{running\_d}$
$output = \mathrm{Softmax}(t_1) \otimes output_1 + \mathrm{Softmax}(t_2) \otimes output_2 + \mathrm{Softmax}(t_3) \otimes output_3$
$y_i \leftarrow output_i * \alpha + \beta \equiv \mathrm{SNLayer}_{\alpha,\beta}(x_i)$
```

3.2. Attention mechanism

In this paper, the scaled dot product attention mechanism is used to pay attention to the input features of the combined normalization encoder. By adaptively selecting relevant feature information, highly relevant features are retained, and irrelevant features are ignored, thereby improving the renormalization encoding. The structure of the scaled dot product attention mechanism is shown in Figure 3.

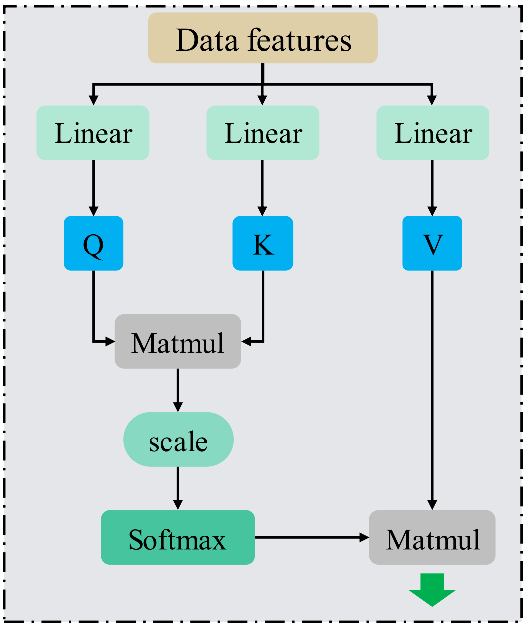


Figure 3. Attention mechanism.

We can see that, the feature vectors from the combined normalized coder are passed through three different linear layers to obtain the query vector *Q*, the key vector *K* and the value vector *V*. Firstly, the dot product calculation is performed on *Q* and *K* to obtain the similarity matrix of *Q* and *K*. Next, the similarity matrix is scaled. Then, the attention weights are obtained by normalizing the values of the similarity matrix using the Softmax function. The purpose of using the Softmax function is to ensure that the sum of the weights is 1. Then, the attention weights and *V* are computed as a dot product to obtain the final result. The calculation process is as follows:

$$Attention(Q,K,V) = \text{Softmax}(\frac{Q \bullet K^T}{\sqrt{d}}) \bullet V \tag{8}$$

where *d* denotes the scaling multiplier, *Q*, *K*, and *V* denote the query vector, key vector, and value vector, respectively, Softmax denotes the Softmax function, and *Attention* (*Q*, *K*, *V*) denotes the final result.

3.3. Combined renormalized decoder

The combined renormalization decoder consists of an LSTM model and an adaptive combined renormalization layer. Figure 4 shows the schematic structure of the combined renormalization decoder layer. The output features of the attention mechanism first goes through a decoder consisting of multiple layers of LSTMs, which decode the features into normalized predicted values. In order to get the actual predicted value, this value needs to be processed using a combined renormalization layer. Corresponding to the normalization calculation, the adaptive merging and renormalization layer includes three renormalization calculations, which are calculated as follows:

$$x = \hat{x} * \sqrt{\sigma^2 + \triangle} + mean \tag{9}$$

$$x = \frac{(max - min)(\hat{x} - a)}{b - a} + min \tag{10}$$

$$x = \hat{x} * |x|_{max} \tag{11}$$

where x represents the data after renormalization, $\hat{x}$ represents the data without renormalization, and min, max, $mean$, and σ^2 represent the maximum value, minimum value, mean value, and variance of the input data, respectively, which all share the statistics from the normalization calculation and are updated with different batches of values. *a* and *b*, on the other hand, represents the interval set by the renormalization method and $\triangle$ represents a fixed smaller positive number.

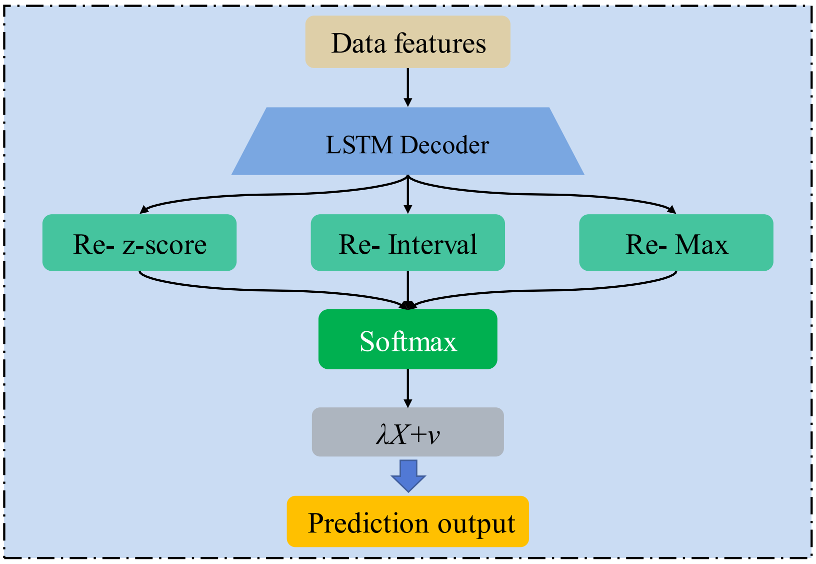


Figure 4. Combined denormalization decoder.

In order to combine the results of the three renormalization calculations and improve the overall data processing, the Softmax combining function is also added to the combined renormalization layer to select the results. This function is used as a combining function to calculate three trainable parameters and output the combined weights for the results of the three renormalization calculations. Three trainable parameters can be optimized by error backpropagation to improve the effectiveness of the renormalization combination. The Softmax function for combining is calculated as follows:

$$H\text{=Softmax}(c_1)\otimes h_1+\text{Softmax}(c_2)\otimes h_1+\text{Softmax}(c_3)\otimes h_3 \tag{12}$$

$$\text{Softmax}(c)=\frac{1}{1+e^{-c}} \tag{13}$$

where c_1, c_2 and c_3 denote the three selected trainable parameters, h_1, h_2 and h_3 denote the results obtained from the three renormalization calculations, Softmax denotes the Softmax function, H denotes the final output, and $\otimes$ denotes the matrix multiplication.

Similarly, the combined renormalization layer incorporates the learnable parameters λ and ν as the scaling and translation factors, respectively. In the output of the renormalization layer, the two parameters can scale the fixed renormalization output and dynamically adjust the output values according to the model training effect. The expression at the output of the renormalization layer can be expressed as:

$$O=\lambda H+\nu \tag{14}$$

where O denotes the actual value output by the inverse normalization layer, H denotes the value after the renormalization calculation, λ is the scaling factor, and ν is the translation factor. Finally, O is output as the state estimate of the model. The flow of the algorithm for combined renormalization layer is shown in Algorithm 2.

```
Algorithm 2: Combined renormalized layer
Input: data: $\hat{R}=\{\hat{x}_1,\ldots,\hat{x}_m\}$, Interval: $a,b$, Forgetting weight: $m$,
Learning parameters: $\lambda,\nu,c_1,c_2,c_3$
Output: $\{\hat{y}_i=\mathrm{SRNLayer}_{\lambda,\nu}(\hat{x}_i)\}$
$\mathrm{Softmax}(c_1)=\frac{e^{c_1}}{e^{c_1}+e^{c_2}+e^{c_3}}$, $\mathrm{Softmax}(c_2)=\frac{e^{c_2}}{e^{c_1}+e^{c_2}+e^{c_3}}$, $\mathrm{Softmax}(c_3)=\frac{e^{c_3}}{e^{c_1}+e^{c_2}+e^{c_3}}$
$output_1 \leftarrow \hat{x} * \sqrt{running\_var+1\times10^{-5}}+running\_mean$
$output_2 \leftarrow \frac{(running\_max - running\_min)(\hat{x}_i - a)}{b - a}+running\_min$
$output_3 \leftarrow \hat{x} * running\_d$
$output=\mathrm{Softmax}(c_1)\otimes output_{21}+\mathrm{Softmax}(c_2)\otimes output_{22}+\mathrm{Softmax}(c_3)\otimes output_{23}$
$\hat{y}_i \leftarrow output*\lambda+\nu\equiv\mathrm{SRNLayer}_{\lambda,\nu}(\hat{x}_i)$
```

4. Experiment

The experiments in this paper use actual data from beer production wastewater treatment. Beer is an alcoholic beverage brewed with malt grain, hops, and water as the primary raw materials, through liquid gelatinization and saccharification and then through liquid fermentation [31]. Beer is the fifth largest consumer beverage globally, second only to tea, carbonated beverages, milk, and coffee, with an average consumption of 23 liters per person per year [32]. The beer uses a lot of water to produce; for each cubic meter of beer produced, the water consumed in general is 10-20 m^3, of which more than 90% will be discharged into a sewer system, and wastewater is produced at all stages of production [33]. Moreover, beer wastewater has a high concentration of soluble organic pollutants and SS [34], and the COD of the wastewater produced in the production process is high because most organic matter in the water is made up of sugars, starches, and proteins [35]. The biological methods commonly used for beer wastewater treatment include aerobic sequential batch reactor, cross-flow ultrafiltration membrane anaerobic reactor, and UASB [36]. Beer wastewater produces methane [35], and better wastewater treatment strategies could lead to better economic benefits while protecting the environment.

The concentration of pollutants such as COD, SS, TN, and TP detected in the wastewater treatment process is an essential indicator of wastewater treatment, and whether it meets the national discharge standards is the determining factor for judging the effect of wastewater treatment. Predicting the future treatment effect according to the pollutant concentration index of the input wastewater at a historical time to assist in decision-making is a hot issue in current research. However, due to the multi-factor, complex, and nonlinear characteristics of forecasting tasks, higher requirements are placed on forecasting models' data processing and modeling capabilities. Therefore, this study uses COD, SS, TN, and TP data before and after brewery wastewater treatment to verify the model's prediction accuracy.

4.1. Experimental procedure and evaluation index

Based on the data of pollutant concentration indicators in the actual brewery wastewater treatment process, the prediction accuracy of the proposed model and seven classical prediction models, including ANN [37], deep neural network (DNN) [38], LSTM [26], gated recurrent unit (GRU) [39], Attention_LSTM [40], Attention_GRU [41] and Codec [42] are compared.

We build predictive models based on the open-source Tensorflow deep learning framework. In comparative experiments, we set the hyperparameters of the model. Specifically, all prediction models were optimized using the Adam hyperparameter optimization algorithm, and the optimized learning rate was set to 0.0001; the batch size of the data input network was set to 10; and the number of iterations per training was 300. In order to avoid the influence of random errors of the model on the prediction results, all comparative experiments were repeated 10 times independently, and the average value was taken as the final result.

In this paper, four evaluation indicators are used to evaluate the experimental results: root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and Pearson correlation coefficient (R). All four evaluation indicators can measure the difference between the prediction value given by the model and the actual value and evaluate the model's performance. The smaller RMSE, MAE, and MAPE values represent the minor difference between the prediction value given by the model and the actual value. In comparison, the larger R values represent the model's better fitting ability. The calculation equations (15)-(18) for the four evaluation indicators are as follows.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{15}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|(y_i-\hat{y}_i)\right| \tag{16}$$

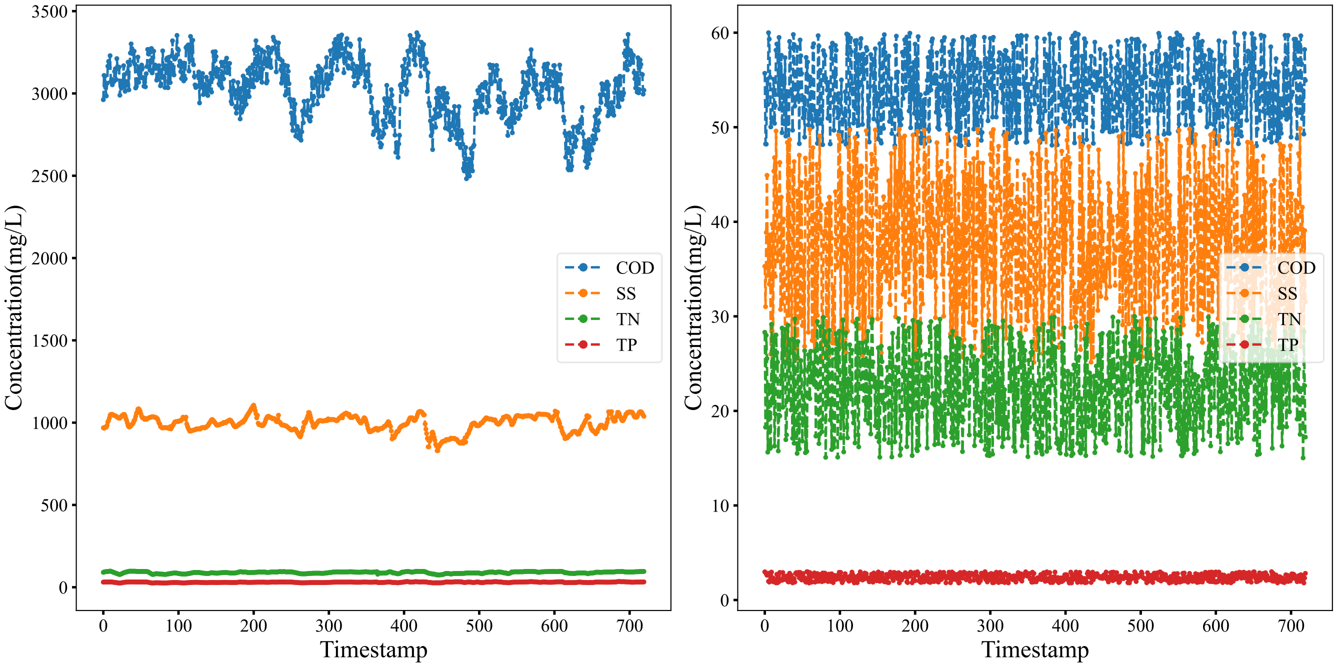
$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{\hat{y}_i}\right| \tag{17}$$

$$R=\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)(\hat{y}_i-\bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}}_i)^2}} \tag{18}$$

where n is the total amount of data, $\hat{y}$ denotes the actual value of the data, y is the state estimate given by the model, $\bar{\hat{y}}$ is the mean of the actual values, and $\bar{y}$ denotes the mean of the prediction values.

4.2. Validation results

The dataset consists of four pollutant concentration indicators of COD, SS, TN, and TP detected during the brewery wastewater treatment. The data set was collected from a wastewater treatment station. 720 sets were collected from June 11 to July 11, 2022. The data sampling interval was 1 h. Each data set includes four pollutant concentration indicators at the inlet and outlet. The structure of the dataset used is shown in Figure 5.



(**a**) (**b**)

Figure 5. Data comparison of water inlet and outlet. (a) COD, SS, TN, and TP detected at the water inlet. (b) COD, SS, TN, and TP detected at the outlet.

This experiment compares the CNC model with other classical prediction models and verifies the superiority of the improvement proposed in this paper compared with other models in applying actual wastewater treatment effect prediction. The comparison models include: ANN [37], DNN [38], LSTM [26], GRU [39], Attention_LSTM [40], Attention_GRU [41] and Codec model [42]. The experiment uses the pollutant concentration index of the water inlet from time *t*-30 to *t* to predict the pollutant concentration index of the water outlet at time *t*+1, and the data set is divided into 90% training set and 10% test set.

The prediction accuracy evaluation indexes of each comparative model are shown in Table 2. Figure 6 compares the predicted and actual values of each model. The comparison results show that the model proposed in this paper has better performance indicators, and the prediction results are closer to the actual situation. RMSE, MAE, and MAPE of the proposed model are reduced by 1.5%, 3.2%, and 0.5%, respectively, and the R indicator is increased by 0.1%, which verifies the improvement proposed in this paper.

Table 2. Comparison of results of evaluation indicators.

Model	RMSE	MAE	MAPE	R
ANN [37]	4.5633	3.3221	1.0059	0.9722
DNN [38]	4.5525	3.3194	0.9983	0.9723
LSTM [26]	4.4786	3.2571	1.0215	0.9733
GRU [39]	4.4888	3.2808	1.0135	0.9731
Attention_LSTM [40]	4.4478	3.2330	1.0086	0.9735
Attention_GRU [41]	4.4221	3.2171	1.0121	0.9738
Codec [42]	4.4221	3.2171	1.0121	0.9738
The proposed CNC	4.3547	3.1126	1.0071	0.9749

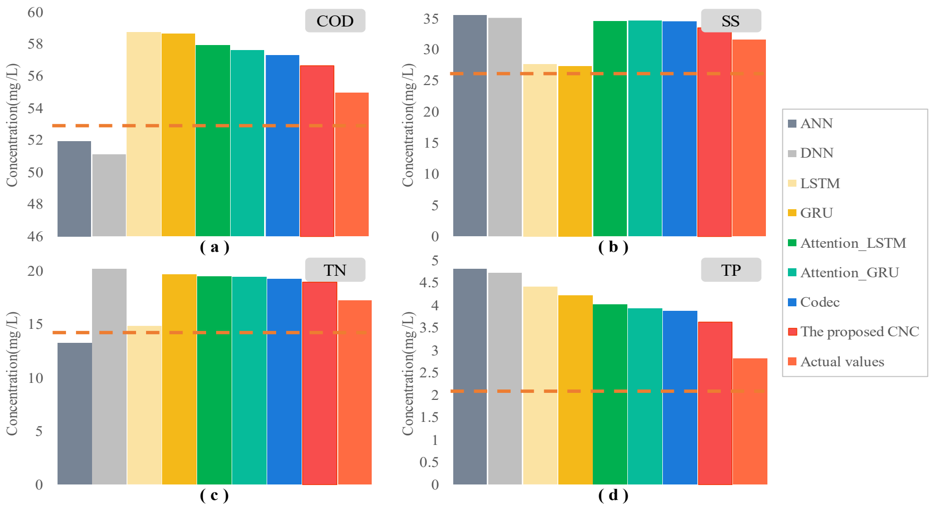


Figure 6. Comparison of predicted and actual values given by the model, based on four pollutant indicators. (a) COD, (b) SS, (c) TN, (d) TP. The last orange-red band is the actual ground-truth value, and the prediction results of all methods are compared using dashed lines. It can be seen that the red band (the method proposed in this paper) is the closest to the actual value.

5. Conclusions

The harmless treatment of wastewater is related to environmental protection and health. However, due to the volatility and nonlinear characteristics of wastewater treatment, it is difficult to carry out predictive modeling and guide early regulation, which seriously affects treatment efficiency.

Considering the prediction of pollutant indicators in brewery wastewater treatment to assist management, we propose an improved deep learning prediction model. The model is based on a combined normalized codec prediction for multi-factor and strongly nonlinear scenarios prediction tasks. In this model, the multi-factor pollutant index data such as COD and SS are first input into the combined normalization encoder, and the data is adaptively processed by combining the advantages of the three normalization methods. The encoder extracts the features of the data. Then, the decoder performs feature decoding after the features are paid attention to by the attention mechanism. Finally, a combined renormalization layer adaptively renormalizes the data and outputs the prediction results. The constructed CNC model was used to predict the four pollutant indicators of COD, SS, TN, and TP in brewery wastewater treatment and compared with the classical prediction model. The proposed model's RMSE, MAE, and MAPE indicators were 4.355, 3.113, and 1.007, and the R index reached 0.975, which is better than the comparison model. The experimental results show that the model is more suitable for managing and applying wastewater treatment.

In future work, we will continue to improve the model to enhance the accuracy of data predictions. Meanwhile, we will apply the model to more scenarios to verify the method's applicability.

Author Contributions: Conceptualization, C.-M.X. and J.-S.Z.; Formal analysis, J.-L.K.; Funding acquisition, X.-B.J.; Investigation, Y.-T.B.; Methodology, J.-S.Z. and X.-B.J.; Resources, C.-M.X.; Software, J.-S.Z. and L.-Q.K.; Supervision, J.-L.K., Y.-T.B., T.-L.S. and P.C.; Validation, X.-B.J.; Visualization, J.-S.Z. and H.-J.M.; Writing – original draft, J.-S.Z. and L.-Q.K.; Writing – review & editing, J.-S.Z. and X.-B.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China, No. 62173007, 62006008, 61903009.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

doi:10.20944/preprints202209.0486.v1

References

1. Ai, H.; Wang, Q.; Fan, X.; Xie, W.; Shinohara, R. Effect of hydraulic retention time on the efficiency of vertical tubular anaerobic sludge digester treating waste activated sludge. *Environ. Technol.* **2005**, 26, 725-731.
2. Wang, Y.; Geng, J.; Peng, Y.; Wang, C.; Guo, G.; Liu, S. A comparison of endogenous processes during anaerobic starvation in anaerobic end sludge and aerobic end sludge from an anaerobic/anoxic/oxic sequencing batch reactor performing denitrifying phosphorus removal. *Bioresour. Technol.* **2012**, 104, 19-27.
3. Chen, W.; Yu, T.; Xu, D.; Li, W.; Pan, C.; Li, Y. Performance of DOuble Circulation Anaerobic Sludge bed reactor: Biomass self-balance. *Bioresour. Technol.* **2021**,320, 124407.
4. Xu, Y.; Dai, X. Integrating multi-state and multi-phase treatment for anaerobic sludge digestion to enhance recovery of bio-energy. *Sci. Total. Environ.* **2020**, 698, 134196.
5. Zhu, Y.; Li, X.; Du, M.; Liu, Z.; Luo, H.; and Zhang, T. Improve bio-activity of anaerobic sludge by low energy ultrasound. *Water. Sci. Technol.* **2015**, 72, 2221-2228.
6. Chon, D.H.; Rome, M.; Kim, Y.M.; Park, K.Y.; Park, C. Investigation of the sludge reduction mechanism in the anaerobic side-stream reactor process using several control biological wastewater treatment processes. *Water. Res.* **2011**, 45, 6021-6029.
7. Li, Y.J.; Sun, L.P.; Ji, M. The nitrogen and phosphorus removal and sludge yield comparative study under anaerobic/aerobic and anaerobic/anoxic conditions. *Adv. Mat. Res.* **2012**, 610-613, 2068-2073.
8. Chan, Y.J.; Chong, M.F.; Law, C.L.; Hassell, D.G. A review on anaerobic–aerobic treatment of industrial and municipal wastewater. *Chem. Eng. J.* **2019**, 155, 1-18.
9. Yilmaz, G.; Lemaire, R.; Keller, J.; Yuan, Z. Effectiveness of an alternating aerobic, anoxic/anaerobic strategy for maintaining biomass activity of BNR sludge during long-term starvation. *Water. Res.* **2007**, 41, 2590-2598.
10. Lv, L.; Li, W.; Wu, C.; Meng, L.; Qin, W. Microbial community composition and function in a pilot-scale anaerobic-anoxic-aerobic combined process for the treatment of traditional Chinese medicine wastewater. *Bioresour. Technol.* **2017**, 240, 84-93.
11. Ministry of Ecology and Environment of the People's Republic of China. GB19821-2005: Discharge standard of pollutants for beer industry, 2005.07.18.
12. Ly, Q.V.; Truong, V.H.; Ji, B.; Nguyen, X.C.; Cho, K.H.; Ngo, H.H.; Zhang, Z. Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. *Sci. Total Environ.* **2022**, 832, 154930.
13. Wang, Z.; Man, Y.; Hu, Y.; Li, J.; Hong, M.; Cui, P. A deep learning based dynamic COD prediction model for urban sewage. *Environ. Sci. Water Res. Technol.* **2019**, 5, 2210-2218.
14. Dargan, S.; Kumar, M.; Ayyagari, M.R.; Kumar, G. A survey of deep learning and its applications: a new paradigm to machine learning. *Arch. Comput.* **2020**, 27, 1071-1092.
15. Li, X.; Yi, X.; Liu, Z.; Liu, H.; Chen, T.; Niu, G.; Ying, G. Application of novel hybrid deep leaning model for cleaner production in a paper industrial wastewater treatment system. *J. Clean. Prod.* **2021**, 294, 126343.
16. Arismendy, L.; Cárdenas, C.; Gómez, D.; Maturana, A.; Mejía, R.; Quintero M.C.G. Intelligent system for the predictive analysis of an industrial wastewater treatment process. *Sustainability* **2020**, 12, 6348.
17. Hilal, A.M.; Althobaiti, M.M.; Eisa, T.A.E.; Alabdan, R.; Hamza, M.A.; Motwakel, A.; Negm, N. An Intelligent Carbon-Based Prediction of Wastewater Treatment Plants Using Machine Learning Algorithms. *Adsorpt. Sci. Technol.* **2022**, 8448489.
18. Liu, G.; He, T.; Liu, Y.; Chen, Z.; Li, L.; Huang, Q.; Liu, J. Study on the purification effect of aeration-enhanced horizontal sub-surface-flow constructed wetland on polluted urban river water. *Environ. Sci. Pollut. R.* **2019**, 26, 12867-12880.
19. Han, H.; Liu, Z.; Hou, Y.; Qiao, J. Data-driven multi-objective predictive control for wastewater treatment process. *IEEE Trans. Industr. Inform.* **2019**, 16, 2767-2775.
20. Farhi, N.; Kohen, E.; Mamane, H.; Shavitt, Y. Prediction of wastewater treatment quality using LSTM neural network. *Environ. Technol. Innov.* **2021**, 23, 101632.
21. Wan, X.; Li, X.; Wang, X.; Yi, X.; Zhao, Y.; He, X.; Huang, M. Water quality prediction model using Gaussian process regression based on deep learning for carbon neutrality in papermaking wastewater treatment system. *Environ. Res.* **2022**, 211, 112942.
22. Jain, S.; Shukla, S.; Wadhvani, R. Dynamic selection of normalization techniques using data complexity measures. *Expert Syst. Appl.* **2018**, 106, 252-262.
23. Yang, H.; Ding, K.; Qiu, R.C.; Mi, T. Remaining useful life prediction based on normalizing flow embedded sequence-to-sequence learning. *IEEE T. Reliab.* **2020**, 70, 1342-1354.
24. Passalis, N.; Tefas, A.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Deep adaptive input normalization for time series forecasting. *IEEE Trans. Neural. Netw. Learn. Syst.* **2019**, 31, 3760-3765.
25. Jin, X.; Zhang, J.; Kong, J.; Su, T.; Bai, Y. A reversible automatic selection normalization (RASN) deep network for predicting in the smart agriculture system. *Agronomy* **2022**, 12, 591.
26. Van Houdt, G.; Mosquera, C.; Napoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, 53, 5929-5955.
27. Wang, Q.; Hao, Y. ALSTM: An attention-based long short-term memory framework for knowledge base reasoning. *Neurocomputing* **2020**, 399, 342-351.
28. Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.H.; Choo, J. Reversible instance normalization for accurate time-series forecasting against distribution shift. *ICLR* **2021**, 1-25.
29. Totaro, S.; Hussain, A.; Scardapane, S. A non-parametric softmax for improving neural attention in time-series forecasting. *Neurocomputing* **2020**, 381, 177-185.

30. Charles, C.H. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting* **2004**, 20, 5-10.
31. Karlovic, A.; Juric, A.; Coric, N.; Habschied, K.; Krstanovic, V.; Mastanjevic, K. By-products in the malting and brewing industries—re-usage possibilities. *Fermentation* **2020**, 6, 82.
32. Fillaudeau, L.; Blanpain-Avet, P.; Daufin, G. Water, wastewater and waste management in brewing industries. *J. Clean. Prod.* **2006**, 14, 463-471.
33. Mielcarek, A.; Janczukowicz, W.; Ostrowska, K.; Jóźwiak, T.; Kłodowska, I.; Rodziewicz, J. Biodegradability evaluation of wastewaters from malt and beer production. *J. Inst. Brew.* **2013**, 119, 242-250.
34. Shao, X.; Peng, D.; Teng, Z.; Ju, X. Treatment of brewery wastewater using anaerobic sequencing batch reactor (ASBR). *Bioresour. Technol.* **2008**, 99, 3182-3186.
35. Sangeetha, T.; Guo, Z.; Liu, W.; Cui, M.; Yang, C.; Wang, L. Cathode material as an influencing factor on beer wastewater treatment and methane production in a novel integrated upflow microbial electrolysis cell (Upflow-MEC). *Int. J. Hydrogen. Energ.* **2016**, 41, 2189-2196.
36. Feng, Y.; Wang, X.; Logan, B.E.; Lee, H. Brewery wastewater treatment using air-cathode microbial fuel cells. *Appl. Microbiol. Biotechnol*. **2008**, 78, 873-880.
37. Kang, J.H.; Song, J.; Yoo, S.S.; Lee, B.J.; Ji, H.W. Prediction of odor concentration emitted from wastewater treatment plant using an artificial neural network (ANN). *Atmosphere* **2020**, 11, 784.
38. Poznyak, A.; Chairez, I.; Poznyak, T. A survey on artificial neural networks application for identification and control in environmental engineering: Biological and chemical systems with uncertain models. *Annu. Rev. Control.* **2019**, 48, 250-272.
39. Oliveira, P.; Fernandes, B.; Analide, C.; Novais, P. Forecasting energy consumption of wastewater treatment plants with a transfer learning approach for sustainable cities. *Electronics* **2021**, 10, 1149.
40. Abbasimehr, H.; Paki, R. Improving time series forecasting using LSTM and attention models. *J. Amb. Intel. Hum. Comp.* **2022**, 13, 673-691.
41. Jung, S.; Moon, J.; Park, S.; Hwang, E. An attention-based multilayer GRU model for multistep-ahead short-term load forecasting. *Sensors* **2021**, 21, 1639.
42. Liu, L.; Song, X.; Chen, K.; Hou, B.; Chai, X.; Ning, H. An enhanced encoder–decoder framework for bearing remaining useful life prediction. *Measurement* **2021**, 170, 108753.