IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 17, 2021, accepted October 22, 2021, date of publication October 26, 2021, date of current version November 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123225

Using LSTM Neural Network Based on Improved PSO and Attention Mechanism for Predicting the Effluent COD in a Wastewater Treatment Plant

XIN LIU, QIMING SHI, ZHEN LIU, AND JIA YUAN
School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China

Corresponding author: Jia Yuan (yuanjia@hebeu.edu.cn)

This work was supported in part by the Project of the University Science and Technology Research Youth Fund of Hebei Province, in 2021, under Grant QN2021034; in part by the Innovation Fund Project of Hebei University of Engineering under Grant SJ2101003143, Grant SJ2101003148, and Grant SJ2101003149; in part by the Natural Science Foundation of Hebei Province of China under Grant F2021402005; and in part by National Defense Key Laboratory Fund Project of the Equipment Development Department under Grant 614240319010.

ABSTRACT Enhancing the monitoring capabilities of wastewater treatment plant (WWTP) key features can accomplish accurate prediction to help WWTPs develop a plan, which is of great significance to control regional water environmental pollution. Chemical oxygen demand (COD) is one of the key features of wastewater treatment. Traditional monitoring methods are time consuming and have high costs making it difficult to meet the needs of rapid monitoring in practical applications. To address this issue, a method for optimizing a long short term memory (LSTM) neural network model based on adaptive hybrid mutation particle swarm optimization (AHMPSO) and an attention mechanism (AM) is proposed. As the hyperparameters of the LSTM are difficult to select, AHMPSO is used to optimize the LSTM. A nonlinear variable inertia weight with random factors is introduced to balance the global search ability and the local search ability and to improve the convergence speed of the PSO algorithm. In addition, the hybrid mutation strategy is added in the search process to reduce the risk of particles falling into local optimal solutions. Finally, an AM is added to the LSTM model to mine local water quality features to improve the effluent COD prediction accuracy. Compared with other models (LSTM, LSTM-AM, and PSO-LSTM-AM), the RMSE(the root mean square error) of the optimized model decreased by 7.803%-19.499%, the MAE(the mean absolute error) of the optimized model decreased by 9.669%-27.551%, the MAPE(the mean absolute error) of the optimized model decreased by 8.993%-25.996%, and the R^2 (the coefficient of determination) value of the optimized model increased by 3.313%-11.229%. The experimental results show that the optimized model has better performance and achieves a more accurate prediction of the effluent COD.

INDEX TERMS Attention mechanism, deep learning, particle swarm optimization, prediction, wastewater treatment process.

I. INTRODUCTION

With the increasing capacity of wastewater treatment, the problem of water pollution prevention and control has shifted from simply improving the quality of the water environment to an organic combination of water quality improvement, water resource protection and water ecological protection. The most effective approach to strengthen the protection of water resources and prevent further deterioration and pollution of water resources is to enhance the monitoring capacity of wastewater treatment's key features.

Due to the characteristics of uncertainty, nonlinearity, and time lag in wastewater treatment, the structure of the mechanism is difficult to describe with traditional mathematical models [1]. In addition, in the wastewater treatment process, there are many important key features that are not easy to directly measure, such as effluent COD [2], which especially hinders the effective monitoring and control of wastewater treatment quality. Although an accurate concentration can be obtained through traditional detection methods, such as the

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

dichromate and the determination of the permanganate index methods, it is inevitable that there may be a significant time delay that ranges from minutes to hours. This delay is too late for advanced wastewater treatment systems that require more precise and timely control [3]. Otherwise, traditional detection methods may cause secondary pollution [4]. Hence, to address the above problems, it is necessary to design a fast and accurate monitoring method that can estimate hard-to-measure key features from other existing key features (instrumental key features). These methods are of great significance for improving the monitoring ability of wastewater treatment features.

In the past, scholars used mechanism models, such as Activated Sludge Model No. 1 (ASM1) and Activated Sludge Model No. 2 (ASM2), to simply describe the complexity of wastewater treatment [5]. Then, to objectively evaluate the performance of the wastewater treatment control strategy, scholars and related organizations have jointly developed the activated sludge water treatment benchmark simulation model (BSM1), which can monitor the key features of wastewater [6]. While it may provide better experimental results, users need to know the expertise of the various systems in advance [7]. In addition, these models are designed according to specific circumstances. If these models need to be applied in other circumstances, many modifications and tests are required, which limits the generalization of the models [8],[9]. Unlike mechanism models, data-driven models can be made by da ta and algorithms, which means that data-driven models do not need to fully understand the mechanism of the process [10]. Furthermore, WWTPs monitor, store and accumulate a large amount of data in daily production, which makes data-driven models more practical in applications [11].

A neural network is a data-driven model that imitates the structure of biological neurons [12]. Based on its powerful fitting ability and adaptability, neural networks have been gradually introduced into the field of wastewater treatment for data-driven modeling [13]. Matheri *et al.* constructed an ANN model to predict the concentration of COD and trace metals in WWTPs with the goal of revealing the relationship between the two parameters. The results prove that neural networks can be used to build a simulation model of WWTPs [14]. Bakr *et al.* established an ANN model optimized based on the Levenberg-Marquardt algorithm to simulate the performance of auto-aerated immobilized biomass (AIB) reactor packed with sponge media, experiment showed that the R^2 (the coefficient of determination) value of the model was satisfactorily fit in training, verification and testing and that the model can reflect reality [15]. In [16], Facchini *et al.* developed a decision model supported by an ANN model to determine an economic sludge management strategy and experimentally showed that the model can identify appropriate sludge treatment options based on multiple characteristics, thus supporting decision-making. Bekkari *et al.* used the ANN model to predict the effluent COD of the Touggourt WWTP, and the results indicated that the ANN modeling approach can provide an effective tool for simulating, controlling and predicting the performance of WWTPs [17]. Nourani *et al.* used a variety of artificial intelligence models (SVM, FFNN, ANFIS, and MLR) to predict the performance of the Nicosia WWTP, and the results of those experiments illustrated that the AI models could satisfactorily predict the Nicosia WWTP effluent COD [18].

However, the above research ignores the time series characteristics of wastewater data [19], [20] and lacks effective treatment of the sequence dependence between input variables, which limits the model's ability to treat time series forecasting tasks. Moreover, with the increase in neural network layers, gradient vanishment and explosion conditions will arise [21]. The LSTM neural network was proposed as a way to address gradient vanishment and explosion by implementing gating [22]. It is an improved neural network based on a recurrent neural network (RNN) that can balance the temporal and nonlinear relationship of wastewater data [23].

At present, the LSTM neural network has been successfully used in speech recognition, natural language processing and other applications [24]–[27]. Based on the performance of the LSTM neural network in these fields, many scholars have tried to introduce the LSTM neural network into the field of wastewater treatment. Zhiwei *et al.* built the LSTM model to simulate the wastewater treatment process [28]. Yaqub *et al.* used an LSTM neural network to predict the nutrient removal efficiency of WWTPs [29]. Cheng *et al.* designed six kinds of neural networks based on the LSTM method and gated recurrent units (GRUs) to compare the prediction effects of WWTP features [30]. Pisa *et al.* verified the effectiveness of an LSTM-based internal model controller (IMC) applied to WWTPs [31]. Although the LSTM neural network can extract valid features from data, it lacks the ability to learn locally important features [32]. Recent studies suggest that LSTM neural networks based on attentional mechanism can be used for time-series task prediction [33]. He *et al.* showed that adding a self-attention mechanism to an LSTM neural network can not only capture local information but can also solve the long-term dependencies well [34]. Zang *et al.* compared the LSTM model based on an attention mechanism with a variety of neural network models and proved the effectiveness of the LSTM model in a time series prediction task [35]. The experimental results showed that adding an attention mechanism to the LSTM model can improve the practicability and accuracy [36]. Therefore, the introduction of an attention mechanism into an LSTM neural network can improve the ability of the neural network to mine locally important features from wastewater data can effectively improve the prediction accuracy and the stability of the model.

Although LSTM neural network has many advantages, it is difficult to select hyperparameters, similar to traditional neural networks [37]. Some scholars have attempted to use the particle swarm optimization (PSO) algorithm to optimize

the LSTM hyperparameters. In [38], the PSO algorithm was applied to optimize the LSTM hyperparameters, which makes up for the cumbersome and time-consuming shortcoming of manual selection. Zhang *et al.* found that the prediction accuracy of the LSTM neural network optimized by PSO was improved [39]. Song *et al.* showed that the performance of the LSTM neural network using PSO was better than that of other methods [40]. However, the standard PSO algorithm has the problem of slow convergence speed and easily falls into a local optimum [41]. In response to the above problem, an adaptive hybrid mutation particle swarm optimization algorithm (AHMPSO) is proposed in this research.

In view of the above problems, the main contributions of this research are as follows:

1) To improve the monitoring ability of WWTP features and to provide a new type of wastewater treatment indicator monitoring tool; a data-driven neural network model is proposed in this paper.
2) To improve the prediction accuracy of the model for the WWTP key features, an attention mechanism is introduced to improve the ability of the LSTM model to learn the importance of local wastewater features. In addition, to compensate for the cumbersome shortcomings of manually selecting hyperparameters and to more reasonably determine the model hyperparameters, the AHMPSO algorithm is used to optimize the number of neurons in the hidden layer and the learning rate of the LSTM-AM model.
3) To verify the prediction effect of the AHMPSO-LSTM-AM model, taking the effluent COD of WWTPs as an example, a comparative analysis of all the models proposed in this paper was carried out, and the efficiency and stability of the proposed hybrid model on a real dataset were evaluated.

The outline of this paper is as follows: Section II presents a model (AHMPSO-LSTM-AM) that uses the AHMPSO algorithm to optimize the LSTM-AM neural network for effluent COD prediction of WWTPs. Section III introduces related datasets and then compares and discusses the prediction effect of the model. Section IV summarizes the paper.

II. LSTM-AM EFFLUENT COD PREDICTION MODEL BASED ON AHMPSO OPTIMIZATION

The LSTM-AM effluent COD prediction model (AHMPSO-LSTM-AM) based on the AHMPSO optimization is composed of three parts: data preprocessing, the LSTM-AM unit and the AHMPSO unit. The details are as follows:

1) Data preprocessing: Given the missing wastewater data, the Lagrange interpolation method is used. The data with different orders of magnitude is normalized, and the training set, verification set, and test set are selected.
2) LSTM-AM unit: The LSTM-AM model is constructed, the key features of the treated wastewater data are input to the model, and the model output is processed by the fully connected layer to obtain the effluent COD prediction results.
3) AHMPSO unit: The number of hidden layer neural units and the learning rate of the LSTM-AM are optimized, and the optimal hyperparameters obtained are assigned to the LSTM-AM model.

The overall architecture of the AHMPSO-LSTM-AM model is shown in Fig. 1.

A. LSTM MODEL BASED ON ATTENTION MECHANISM

1) LSTM

The LSTM is an improved neural network based on RNN that can effectively address the problems of vanishing and exploding gradients seen in traditional neural networks. Through the additional memory unit, the long-term timing information is stored to capture the long-term dependencies in the data. Based on this feature, LSTM neural network is frequently used to deal with time series tasks. Because wastewater data have distinct time series features, LSTM neural network can be used to mine the time series variation rules existing in wastewater data.

As shown in Fig. 2, the memory unit of the LSTM neural network maintains three gates at each time step, including the forget gate, input gate and output gate. Due to the gating, the LSTM neural network can realize filtering and information storage functions [42].

The forget gate combines the hidden layer state h^{t-1} of the previous time step with the input wastewater instrumental key feature x^t of the current time step. Through the sigmoid activation function, all the input features are scaled within the interval [0,1], and the scaling value is used to control the forgetting degree of the cell state C^{t-1} of the previous time step. The formula of the forget gate is as follows:

$$F_t = sigmoid\left(\sum_{j=1}^{J} W_j^{Fx} x_j^t + \sum_{l=1}^{L} W_l^{Fh} h_l^{t-1} + b_F\right) \tag{1}$$

In this formula, J and L are the dimensions of the input feature vector x^t and the hidden layer state h^{t-1}, respectively, and W and b are the weight matrices and bias in each gated, respectively.

The input gate combines h^{t-1} and x^t to generate new candidate values $\tilde{C}_t$ via the tanh activation function. Then, similar to the forgetting gate, the scaled value is utilized to control the degree to which the candidate values are updated. The formula for the input gate is as follows:

$$\tilde{C}_t = tanh\left(\sum_{j=1}^{J} W_j^{\tilde{C}x} x_j^t + \sum_{l=1}^{L} W_l^{\tilde{C}h} h_l^{t-1} + b_{\tilde{C}}\right) \tag{2}$$

$$I_t = sigmoid\left(\sum_{j=1}^{J} W_j^{Ix} x_j^t + \sum_{l=1}^{L} W_l^{Ih} h_l^{t-1} + b_I\right) \tag{3}$$

After that, the current cell state C_t is updated by the following equation:

$$C_t = F_t * C_{t-1} + I_t * \tilde{C}_t \tag{4}$$

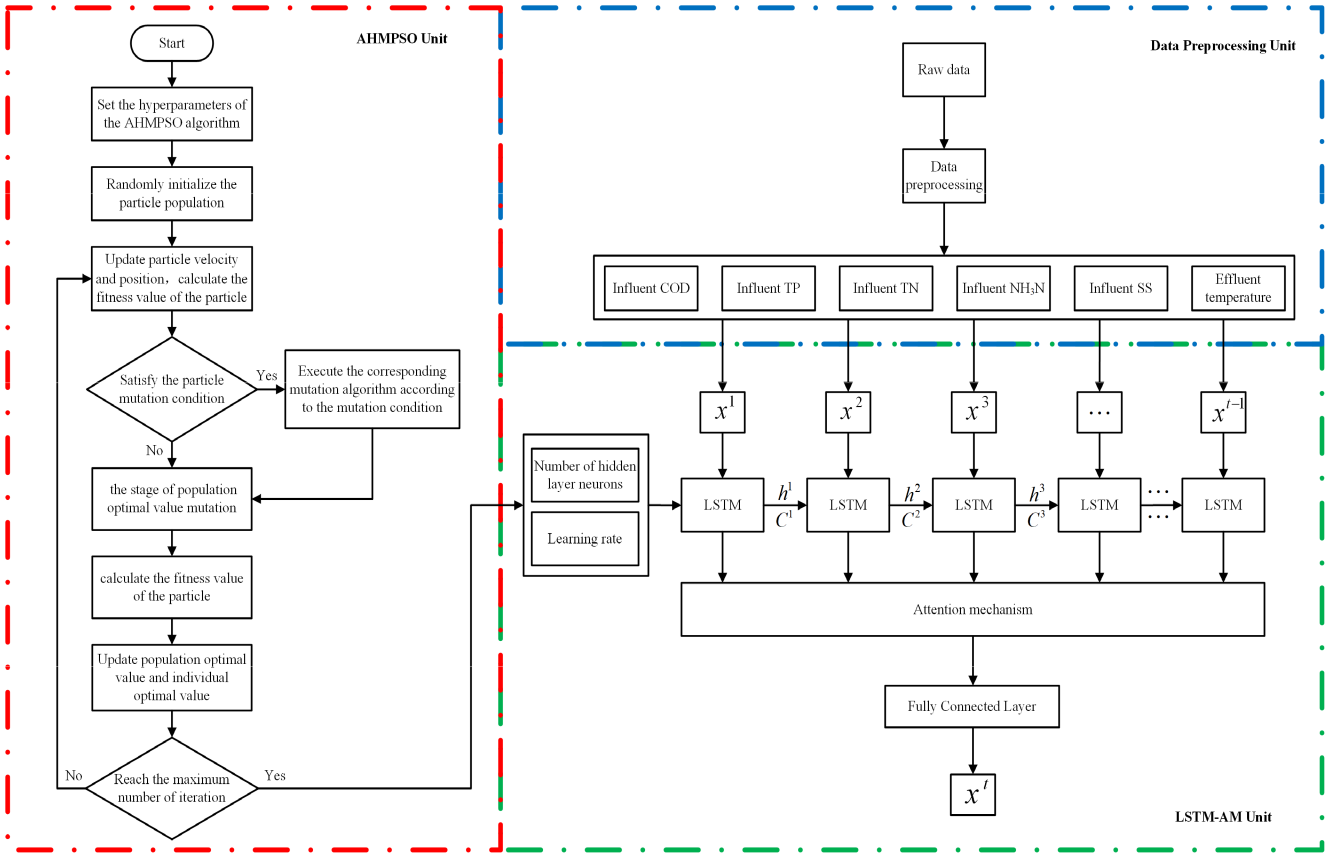

FIGURE 1. Overall architecture diagram of the AHMPSO-LSTM-ATT model. AHMPSO unit is used to obtain the optimal hyperparameters; Data preprocessing is used to process the key features of the input wastewater data; LSTM-AM unit is trained through input data to implement the prediction of effluent COD.

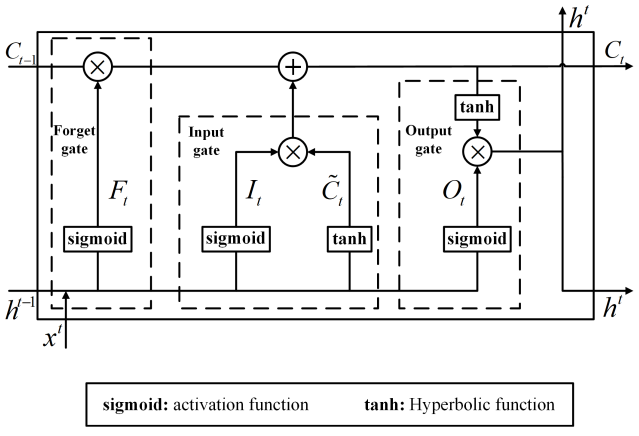

FIGURE 2. Structure diagram of LSTM unit.

The output gate determines what part of the information is exported for the current cell state C_t. The formula of the output gate is as follows:

$$O_t = sigmoid\left(\sum_{j=1}^{J} W_j^{Ox} x_j^t + \sum_{l=1}^{L} W_l^{Oh} h_l^{t-1} + b_O\right) \tag{5}$$

$$h^t = O_t * \tanh(C_t) \tag{6}$$

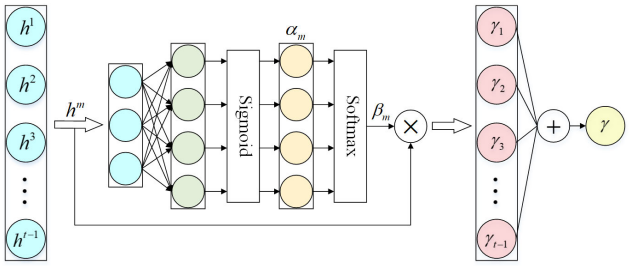

FIGURE 3. The structure of attention module.

2) ATTENTION MECHANISM

The essence of the attention mechanism is to imitate human visual mechanisms. For example, when people observe something, they tend to pay more attention to some information that can assist judgment and ignore irrelevant information [43]. The attention mechanism can be simply understood as a weighted summation that can allocate corresponding weights according to the importance of the input wastewater instrumental key feature. In this way, the ability of the LSTM neural network to learn the importance of local wastewater data features can be improved to achieve the purpose of more accurate WWTP effluent COD prediction. The structure of the attention mechanism is shown in Fig. 3.

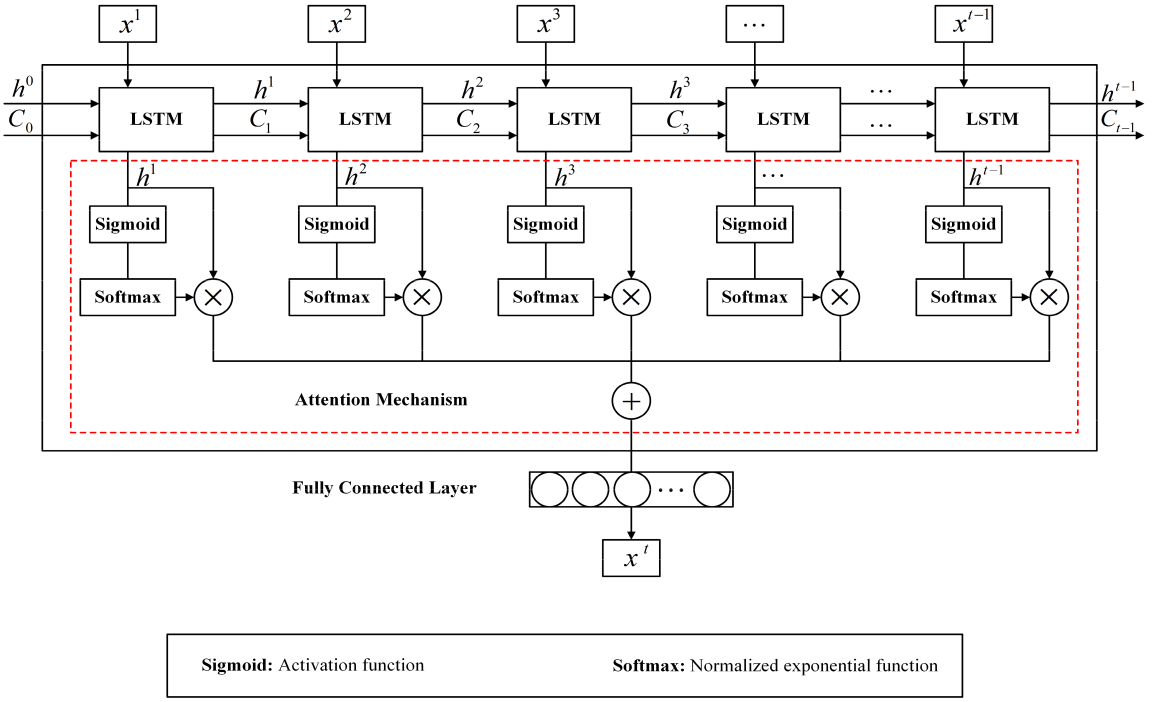

FIGURE 4. The structure of LSTM-AM model.

The calculation formula is as follows [44]:

$$\alpha_m = sigmoid\left(\sum_{l=1}^{L} W_l^{\alpha h} h_l^m + b_\alpha\right) \tag{7}$$

$$\beta_m = \frac{e^{\alpha_m}}{\sum_{q=1}^{T} e^{\alpha_q}} \tag{8}$$

$$\gamma = \sum_{m=1}^{T} \beta_m h^m \tag{9}$$

where T is the total time step; h^m is the output feature vector of the LSTM; α_m is the result of the first weighted calculation through the full connection layer; $W_p^{\alpha h}$ and b_α are the weight matrix and bias of the full connection layer, respectively; β_m is the final weight assigned to the corresponding h^m calculated by a softmax activation function; and vector γ is the key feature of extraction.

3) ARCHITECTURE OF THE LSTM-AM MODEL

The architecture of the LSTM-AM model is shown in Fig. 4. where the input vectors $x^1, x^2, \ldots\ldots, x^t$ are the wastewater instrumental key feature vectors before the time step to be forecasted. The LSTM model disposes of the input vectors in time steps and obtains several hidden layer states $h^1, h^2, \ldots\ldots, h^t$. The attention mechanism calculates the attention weights β_m by using a neural network with a softmax activation function. Then, the attention weights β_m are assigned to the corresponding hidden state h^m. Finally, the key feature γ is extracted by summation. The forecast result of the effluent COD in the next time step is obtained by inputting key feature γ into the fully connected layer.

B. ADAPTIVE HYBRID MUTATION PARTICLE SWARM OPTIMIZATION (AHMPSO)

In the process of constructing the LSTM-AM model, due to the different number of neurons and learning rate in the LSTM hidden layer, the prediction accuracy of the WWTP effluent COD is not the same. However, many experiments are needed to manually select the two hyperparameters. Therefore, to compensate for the cumbersome and time-consuming shortcomings of manual selection and to improve the prediction accuracy of the model, we designed an AHMPSO algorithm to optimize the two hyperparameters of the LSTM-AM model.

1) PRINCIPLE OF PSO ALGORITHM

The PSO algorithm is a swarm intelligence algorithm derived from research on the social behavior of birds. In the algorithm, a solution in the search space is called a particle, and all particles are described by three indices: position, velocity and fitness value determined by the objective function. The position is the solution of the search space, the velocity determines the direction and distance of the particle, and the fitness value determines the quality of the particle [45].

The velocity and position update formulas of the PSO algorithm are as follows:

$$\begin{cases} v_{i,d}^{k+1} = \omega v_{i,d}^k + c_1 r_1 \left(pbest_{i,d}^k - x_{i,d}^k\right) \\ +c_2 r_2 \left(gbest_d^k - x_{i,d}^k\right) \\ x_{i,d}^{k+1} = x_{i,d}^k + v_{i,d}^{k+1} \end{cases} \tag{10}$$

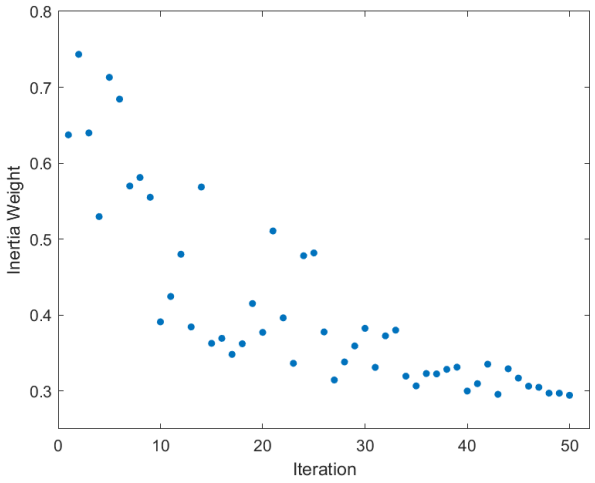

FIGURE 5. Changing curve of nonlinear inertia weight.

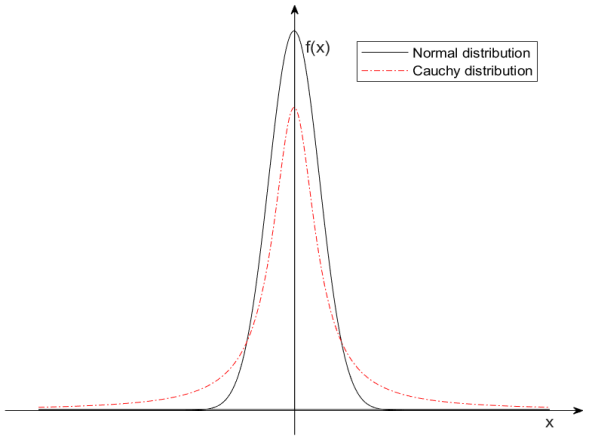

FIGURE 6. Cauchy distribution and Normal distribution.

where $v_{i,d}^k$ is the velocity component of particle i in the dth dimension at the kth iteration; ω is the inertia weight; c_1 and c_2 are the learning factors for particles and populations, respectively; r_1 and r_2 are two independent random numbers with uniform distribution on the interval [0,1]; $x_{i,d}^k$ is the position component of particle i in the dth dimension at the kth iteration; $pbest_{i,d}^k$ is the individual optimal value component of particle i in the dth dimension at the kth iteration; and $gbest_d^k$ is the component of the optimal population value in the dth dimension at the kth iteration.

2) PRINCIPLE OF THE AHMPSO ALGORITHM

The inertial weight ω can measure the ability of particles to maintain the motion state at the previous moment, which can be used to balance the global search ability and local search ability of the algorithm. In the standard PSO algorithm, ω is a fixed value, which will reduce the global search ability and convergence speed of the algorithm, making it easy for the algorithm to fall into local optima and premature convergence. Hence, this paper introduces a nonlinear variation inertia weight with a random factor to solve the above problems[46]. The optimized inertia weight is shown in Eq. (11):

$$\omega(k) = \omega_{fix} e^{-\left(\frac{k}{k_{max}}\right)^{\varepsilon}} \tag{11}$$

where ω_{fix} is a fixed constant that limits the maximum value of the inertia weight; k_{max} is the maximum iteration number of the PSO algorithm; and ε is a random number uniformly distributed within the interval [0,1].

As shown in Fig. 5, as the number of iterations k increases, $\omega(k)$ shows a decreasing trend overall and exhibits randomness among particles. That is, the inertia weight of the descendant particles is not necessarily less than the inertia weight of the previous generation particles. This can improve the randomness and diversity of particles and enhance the effect of inertia weight balance algorithm searchability. In addition, $\omega(k)$ is close to ω_{max} in the early search stage, which makes the algorithm have a strong global search ability. In the later search stage, $\omega(k)$ decreases with the increase in the iteration time, which makes the algorithm have good local search ability. Thus, the joint optimization of global search and local search can be achieved through the adaptive change of the inertia weight.

The improvement of the algorithm effect is limited only by the method of optimizing the inertia weight, which further reduces the risk of the algorithm falling into a local optimum and, at the same time, further balances the global and local search capabilities of the algorithm. According to the mutation principle of the genetic algorithm, we proposed an adaptive hybrid mutation method. The mutation method is divided into two stages: (1) particle mutation and (2) population optimal value mutation.

At the kth iteration, if the current particle i satisfies the condition that its fitness value f_i^k is greater than its individual optimal fitness value f_{pi}^k and the random number r_3 which is uniformly distributed within the interval [0,1] is greater than 0.9, the algorithm enters the particle mutation stage. In the early particle mutation process, the algorithm uses the adaptive mutation method based on the Cauchy distribution. In the later particle mutation process, an adaptive mutation method based on the Normal distribution is adopted [47]. The formulas for the two mutation methods are as follows:

$$\begin{aligned} &x_{i,d}^{k*} = \left| x_{i,d}^k + Cauchy * \left(pbest_{i,d}^k - x_{i,d}^k \right) \right|, \\ &\frac{k}{k_{max}} \le \frac{1}{2} \; and \; r_3 > 0.9 \end{aligned} \tag{12}$$

$$\begin{aligned} &x_{i,d}^{k*} = x_{i,d}^k * \left[1 + \frac{2}{5} - \frac{1}{5} tan\left(\frac{\pi}{4} \frac{k}{k_{max}} \right) * Normal \right], \\ &\frac{k}{k_{max}} > \frac{1}{2} \; and \; r_3 > 0.9 \end{aligned} \tag{13}$$

where $x_{i,d}^{k*}$ is the position of the particle after mutation; *Cauchy* is a random number satisfying the standard Cauchy distribution; and *Normal* is a random number that conforms to the standard Normal distribution.

As shown in Fig. 6, compared with the Normal distribution, the Cauchy distribution has a larger range of values on the x-axis. That is, the mutation method based on the Cauchy distribution can help the algorithm expand the search range of particles and obtain more feasible solutions with better fitness values. Therefore, the mutation method based on the Cauchy distribution is more suitable for improving the global searchability in the early stage of the algorithm. The Normal distribution has a larger range of values on the y-axis. In other words, the mutation method based on a Normal distribution can improve the convergence accuracy of the algorithm and carry out a local search with a small change near the optimal solution. Thus, the mutation method based on the Normal distribution is more suitable for local searchability in the later stage of the developed algorithm.

In the PSO algorithm, the function of the population optimal value $gbest_d^k$ is to guide the particles to iterate toward the optimal solution of the problem. As the particles are updated, new population optimal values will continue to be generated, making the algorithm gradually approach the optimal solution of the problem. If the algorithm is trapped at a local optimum, it may be difficult to jump out of the local optimum within a certain number of iterations. Consequently, in the mutation stage of the population optimal value, the adaptive mutation method based on the elite substitution strategy is used to mutate the population optimal value to help the algorithm jump out of the local optimal value[48]. The formula of the mutation method is as follows:

$$gbest_d^{k*} = gbest_d^k * \left[\mu + \tan\left(e^{-coh}\right)\right] \tag{14}$$

where $gbest_d^{k*}$ is the optimal value of the population after mutation and μ is the scaling factor, which is uniformly distributed in the interval [0,1], $\tan\left(e^{-coh}\right)$ is the perturbation function, and $coh(d)$ can measure the cohesion of particles. The formula is as follows:

$$coh(d) = \frac{1}{I}\sum_{i=1}^{I}\left|\frac{x_{i,d}^k - \bar{x}_d^k}{\bar{x}_d^k}\right| \tag{15}$$

In the formula, I is the number of particles; $\bar{x}_d^k$ is the average value of the particles in the dth dimension at the kth iteration. A smaller $coh(d)$ results in a greater degree of particle agglomeration, and more particles agglomerate, which requires a greater amount of disturbance. A larger $coh(d)$ results in smaller degree of particle agglomeration, more dispersed particles, and a smaller amount of disturbance is required. Before updating the optimal value of the population in each iteration, the algorithm enters the optimal population value mutation stage.

The analysis of Eq. (14) shows that the population optimal value variation method first scales $gbest_d^k$ by the scaling factor μ and then dynamically adjusts the various sizes of the population optimal value through the perturbation function. If the fitness value of $gbest_d^{k*}$ is better than the fitness value of $gbest_d^k$, then the optimal value of the population is updated, $gbest_d^k$ is replaced with $gbest_d^{k*}$, and $gbest_d^{k*}$ participates in the subsequent algorithm calculations.

In essence, the population optimal value mutation stage is a supplement to the particle mutation stage. In the early stage of the algorithm iteration, the particle mutation method based on the Cauchy distribution can produce larger mutations, and in conjunction with the larger inertia weight in the early stage, the diversity and randomness of the particles can be improved, and the search space of the algorithm can be expanded. Due to the sufficient population search of the algorithm and small particle cohesion, the perturbation function will generate a small perturbation. By reducing the variation value to avoid the oscillation of the optimal population value, the algorithm can smoothly converge at a faster rate. In later algorithm iteration stage, the particle mutation method based on the Normal distribution can produce smaller mutations, and in conjunction with the smaller inertia weight in the later stage, this particle mutation method can improve the local search ability of particles. Now, the particle cohesion is high, and the perturbation function will generate a large perturbation. By increasing the variation value, the risk of the population optimal value falling into a local optimal value is reduced so that the algorithm can jump out of the local optimal value with a certain probability.

C. OPTIMIZATION OF THE LSTM-AM MODEL USING AHMPSO

In this paper, two hyperparameters, the number of neurons in the hidden layer of LSTM and the learning rate, which can have a significant impact on the performance of the LSTM-AM model, were selected for optimization.

The process of AHMPSO to optimize the parameters of the LSTM-AM model is as follows:

Step 1: Preprocess the acquired dataset of WWTP key features.

Step 2: Set the hyperparameters of the AHMPSO algorithm, including the number of particles, the number of iterations, the initial inertia weight, the learning factor, and the limits of the particle speed and position.

Step 3: Determine the fitness function of the algorithm. The formula is as follows:

$$f = \sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(y_p - \hat{y}_p\right)^2} \tag{16}$$

where P is the number of samples in the verification set; y_p is the true value of the verification sample; $\hat{y}_p$ is the predicted value.

Step 4: The number of neurons in the hidden layer of the LSTM-AM model and the learning rate are used as the optimization features of the AHMPSO algorithm. Determine the optimization range of the feature to be optimized, initialize the population randomly, and determine the initial population optimal value and the individual optimal value.

Step 5: The position and velocity of particles are updated according to Eq. (10) and Eq. (11), and the LSTM-AM model is constructed according to the corresponding

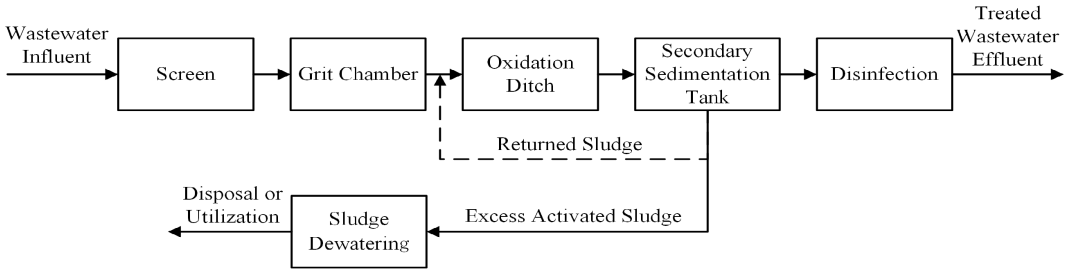

FIGURE 7. The Schematic of wastewater treatment process.

hyperparameters of each particle. The fitness value of each particle is calculated according to Eq. (16).

Step 6: Determine whether to enter the mutation stage. If the algorithm enters the particle mutation stage, the corresponding particle mutation method will be executed according to Eq. (12) and Eq. (13) depending on the corresponding conditions. If the algorithm enters the stage of population optimal value variation, then the population optimal value variation method based on the elite substitution strategy is implemented according to Eq. (14) and Eq. (15).

Step 7: In light of the fitness value of the particles, the optimal value of the population and the optimal value of the individual are determined.

Step 8: Determine whether the algorithm meets the conditions for terminating iteration. If the number of iterations reaches the maximum, the optimal hyperparameter of the model is returned; otherwise, Step 5 is repeated to continue execution until the algorithm meets the termination condition.

Step 9: Use the optimal hyperparameters to construct and train the LSTM-AM model and output the predicted effluent COD value.

III. EXPERIMENT AND ANALYSIS

A. DATA PREPARATION AND PREPROCESSING

The data used in this research are derived from the operation report of a WWTP in Rizhao City, Shandong Province, China. The treatment plant adopts the activated sludge method to treat urban wastewater, and the annual effluent index reaches the standard, which can stably meet level A of China's ''Urban Water Pollution Discharge Standard'' (GB18918). The specific process is shown in Fig. 7. First, the larger suspended solids and inorganic particles in the wastewater were removed through a screen and grit chamber. Second, the wastewater circulated in the oxidation ditch and secondary sedimentation tank to remove the pollutants and separate the sludge from the treated wastewater. The sludge is returned to the oxidation ditch under the action of the return pump. Finally, the treated wastewater is discharged or reused after disinfection, and the fully reacted remaining sludge is dewatered for disposal or utilization.

This paper selects the operation report of the treatment plant from January 1, 2019, to January 31, 2020, with a total of 396 data points. Taking the first 255 data points in 2019 as the training set, the remaining 110 data points as the validation set, and a total of 31 data points in January 2020 as the test set. The influent COD, influent total phosphorus (TP), influent NH_3N, influent total nitrogen (TN), influent suspended solids (SS), and effluent temperature (TE) were selected as auxiliary variables of the AHMPSO-LSTM-AM model to predict the WWTP effluent COD. Part of the data is shown in Table 1.

Due to the difference in the order of magnitude of wastewater data, the LSTM model has a high sensitivity to the data scale. To avoid problems, such as slowing down the convergence speed of the model due to the influence of large data changes, the wastewater data need to be normalized according to Eq. (17).

$$x^{t*} = \frac{x^t - x^t_{min}}{x^t_{max} - x^t_{min}} \tag{17}$$

where x^{t*} is the normalized wastewater data value; x^t is any value in the wastewater dataset; and x^t_{max} and x^t_{min} are the maximum and minimum values in the wastewater dataset, respectively.

B. ENVIRONMENT CONFIGURATION

All experiments in this article are performed on a computer with an Intel(R) Core(TM)i7-9750H with a 2.60 GHz and a GTX1660Ti GPU. All models are built on the Keras 2.2.4 framework using Juypter Notebook based on Python 3.6 to edit the code.

C. SETTING THE MODEL PARAMETER

The parameters of the PSO algorithm and the AHMPSO algorithm are shown in Table 2. The number of hidden layer neurons and the learning rate in the LSTM-AM model are selected as the optimized objects. The range of hidden layer neurons is [10,200], and the range of learning rates is [0.001,0.01]. To reduce the influence of other parameter changes on the experimental results, except for changing the (nonlinear) inertia weight, the rest of the parameter settings are identical.

The fitness curves of the PSO and AHMPSO algorithms are shown in Fig. 8. First, compared to the PSO algorithm, the AHMPSO algorithm falls into the local optimal value fewer times. Second, the convergence speed and accuracy of the AHMPSO algorithm are better than those of the PSO algorithm. The reason is that the nonlinear inertia weight

TABLE 1. Partial sample data.

Date	COD_{eff} mg/L	COD_{inf} mg/L	TN_{inf} mg/L	SS_{inf} mg/L	TP_{inf} mg/L	NH_3N_{inf} mg/L	TE_{eff} ℃
2019.1.1	13.9	194	46.8	125	4.07	34.6	10.1
2019.1.2	13.9	231	33.4	156	8.17	15.3	10.1
2019.1.3	15.6	215	34.2	127	4.07	30.3	10.1
…	…	…	…	…	…	…	…
2020.1.29	21.1	264	48.8	124	5.24	36.7	9.3
2020.1.30	20.6	125	41.5	105	3.6	36.5	9.6
2020.1.31	21.4	102	48.1	89	3.2	22.3	9.8

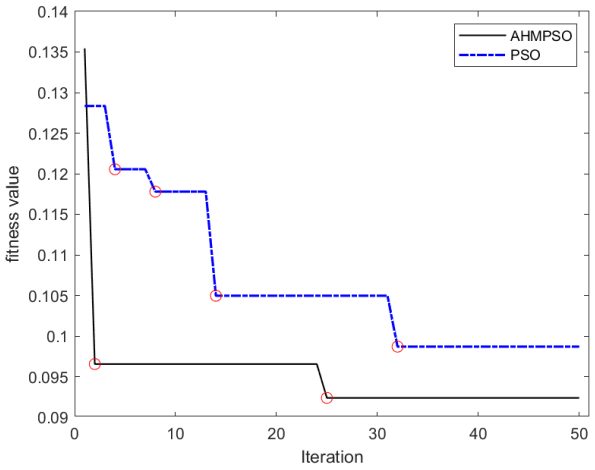

FIGURE 8. Fitness curve.

TABLE 2. The parameter setting of PSO and AHMPSO.

parameter	PSO	AHMPSO
(nonlinear)Inertia weight	0.8	[0.3,0.8]
Hidden layer neuron value range	[10,200]	[10,200]
Learning rate range	[0.001,0.01]	[0.001,0.01]
Number of particles	20	20
The number of iterations	50	50
Learning factor C1,C2	2,2	2,2
Particle velocity range	[-2,10]	[-2,10]

with the random factor can enhance the randomness and diversity of the particles. At the same time, the adaptive mutation method based on the Cauchy distribution adopted in the early iteration can expand the algorithm search range so that the algorithm can jump out of the local optimum in time and improve the convergence speed of the algorithm. In the later algorithm iteration stage, thanks to the adaptive mutation method based on Normal distribution, the local search ability of the algorithm has been strengthened, and the convergence accuracy of the algorithm is improved. In addition, as a supplement to the particle mutation methods, the population optimal value mutation method can generate different disturbances through particle cohesion, balance the global and local search capabilities of the algorithm, and guide the particles to iterate toward the optimal solution of the problem.

In summary, the AHMPSO algorithm is better than the PSO algorithm in terms of convergence accuracy, speed, and global optimization. The optimal model hyperparameters

TABLE 3. Hyperparameter settings for each model.

parameter	LSTM	LSTM-AM	PSO-LSTM-AM	AHMPSO-LSTM-AM
Number of hidden layers neurons	70	70	142	114
Learning rate	0.002	0.002	0.0056	0.0066
Time-step	3	3	3	3
Batch size	50	50	50	50
Epoch	150	150	150	150
Optimizer	Adam	Adam	Adam	Adam
Loss function	RMSE	RMSE	RMSE	RMSE
Dropout rate	0.2	0.2	0.2	0.2

obtained by the AHMPSO algorithm are as follows: the number of hidden layer neurons is 114, and the learning rate is 0.0066. However, algorithm performance improvement comes at a price. For the case of using the wastewater dataset collected in this paper, the average training time of the AHMPSO algorithm is approximately 3 hours, and the average training time of the PSO algorithm is approximately 2.5 hours. The reason is that the AHMPSO algorithm introduces a hybrid mutation strategy, which increases the calculation time of the algorithm, making its training time slightly higher than that of the standard PSO algorithm. The training speed of AHMPSO-LSTM-AM needs to be further optimized.

The hyperparameter settings of the four models proposed in this paper, including LSTM, LSTM-AM, PSO-LSTM-AM, and AHMPSO-LSTM-AM, are shown in Table 3. The Adam optimizer[49] was used to train all the models by minimizing the root mean square error (RMSE), the batch size was set to 50, the time-step was set to 3, and the epoch was set to 100. The probability of model overfitting is reduced by adding a dropout layer. In the selection of the number of hidden layer neurons and the learning rate, unlike the PSO-LSTM-AM model and the AHMPSO-LSTM-AM model to obtain the optimal parameters through optimization, the parameters of the LSTM model and the LSTM-AM model are randomly selected and are the same. The aim is to make a comparison and to verify the performance of those models.

TABLE 4. Performance indicators of all proposed model.

Model	R^2 Mean ± std	RMSE: Mean ± std	MAPE: Mean ± std	MAE: Mean ± std
AHMPSO-LSTM-AM	0.842 ± 0.013	1.028 ± 0.045	3.046 ± 0.134	0.710 ± 0.030
PSO-LSTM-AM	0.815 ± 0.011	1.115 ± 0.031	3.347 ± 0.084	0.786 ± 0.015
LSTM-AM	0.792 ± 0.005	1.182 ± 0.014	3.709 ± 0.097	0.878 ± 0.025
LSTM	0.757 ± 0.015	1.277 ± 0.037	4.116 ± 0.185	0.980 ± 0.046

D. EVALUATION CRITERION

In this paper, the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the coefficient of determination (R^2) are selected as evaluation metrics to measure the prediction efficacy of the model.

Smaller values of the above metrics indicate a better model prediction effect. The relevant formulae are as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{y}_n\right)^2} \tag{18}$$

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|y_n - \hat{y}_n\right| \tag{19}$$

$$MAPE = \frac{1}{N}\sum_{n=1}^{N}\frac{\left|y_n - \hat{y}_n\right|}{y_n} \tag{20}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}\left(y_n - \hat{y}_n\right)^2}{\sum_{n=1}^{N}\left(y_n - \bar{y}\right)^2} \tag{21}$$

In the formulae, N is the number of samples, y_n is the true value, $\bar{y}$ is the mean of the true value, and $\hat{y}_n$ is the predicted value of the model.

E. MODELING RESULTS AND ANALYSIS

This paper conducted 20 random experiments on the proposed model, and the evaluation criteria of all models are shown in Table 4 and in Fig. 9-Fig. 16. As seen from Table 4, the average values of the evaluation criteria of the LSTM model are as follows: R^2 is 0.757 ± 0.015, RMSE is 1.277 ± 0.037, MAPE is 4.116% ± 0.185, and MAE is 0.980 ± 0.046. The average values of the evaluation indicators of the LSTM-AM model are as follows: R^2 is 0.792 ± 0.005, RMSE is 1.182 ± 0.014, MAPE is 3.709% ± 0.097, and MAE is 0.878 ± 0.025; The average values of the evaluation indicators of the PSO-LSTM-AM model are as follows: R^2 is 0.815 ± 0.011, RMSE is 1.115 ± 0.031, MAPE is 3.347% ± 0.084, and MAE is 0.786 ± 0.015; The average values of the evaluation indicators of the AHMPSO-LSTM-AM model are as follows: R^2 is 0.842 ± 0.013, RMSE is 1.028 ± 0.045, MAPE is 3.046% ± 0.134, and MAE is 0.710 ± 0.030.

From the analysis of Fig. 10, Fig. 12, Fig. 14 and Fig. 16, we can see that compared with the model that introduces the attention mechanism, the LSTM model has more outliers,

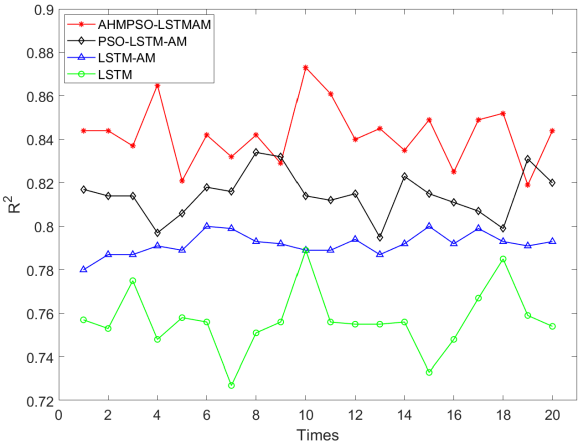

FIGURE 9. Variation curve of R^2 of each model in 20 experiments.

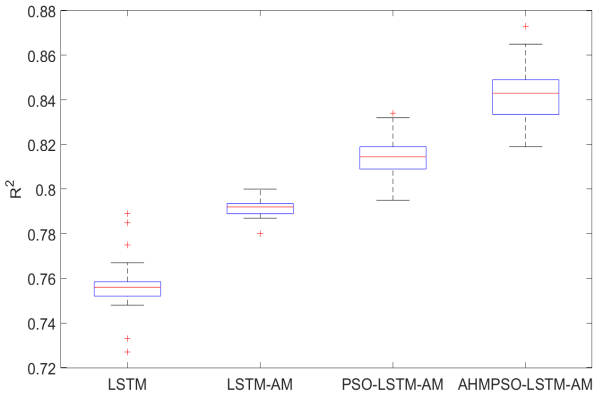

FIGURE 10. Box plot of R^2 obtained from each model.

the average value of the evaluation criteria is larger, the accuracy and stability of the model are poorer. As seen from Table 4, the accuracy and stability of the LSTM model are improved after the introduction of the attention mechanism. Using the same parameters as the LSTM model, the R^2 of the LSTM-AM model is increased by 4.624%, the RMSE is reduced by 7.439%, the MAPE is reduced by 9.888%, and the MAE is reduced by 10.408%. The reason for the above phenomenon is that the attention mechanism allocated corresponding weights to the input features according to the different importance degrees of the input features. Thus, the attention mechanism reduces the error of the model,

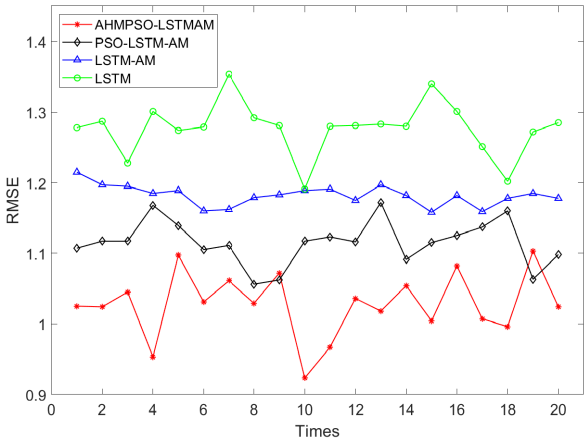

FIGURE 11. Variation curve of RMSE of each model in 20 experiments.

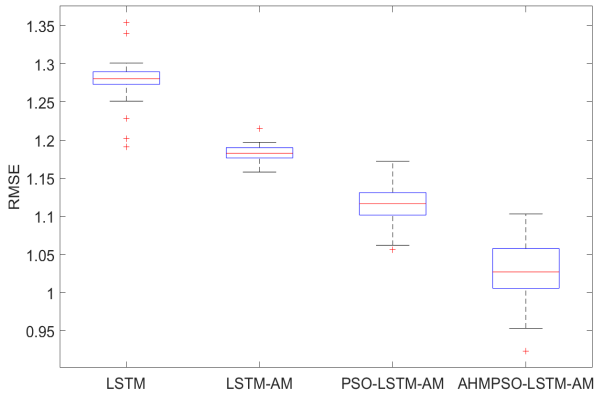

FIGURE 12. Box plot of RMSE obtained from each model.

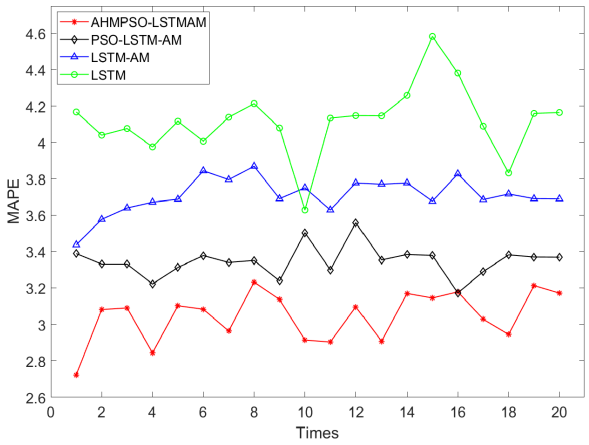

FIGURE 13. Variation curve of MAPE of each model in 20 experiments.

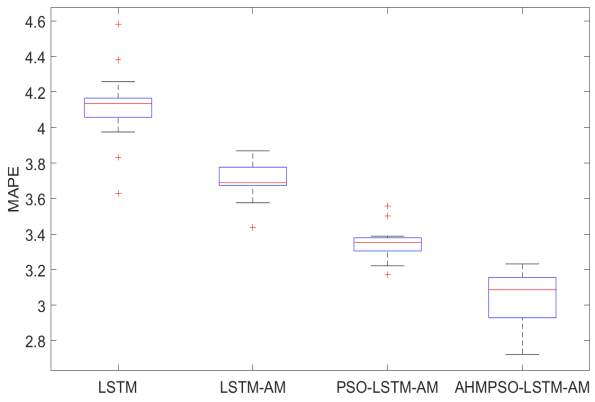

FIGURE 14. Box plot of MAPE obtained from each model.

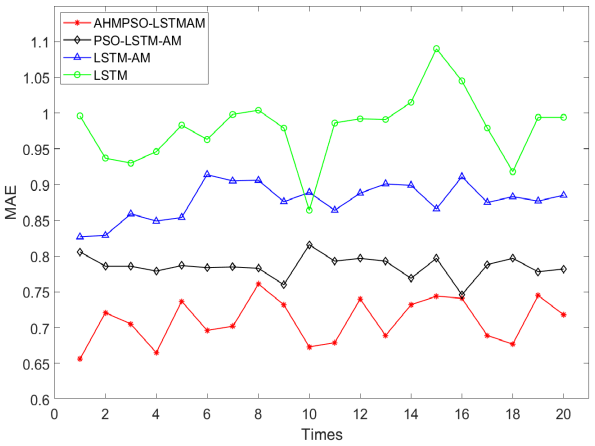

FIGURE 15. Variation curve of MAE of each model in 20 experiments.

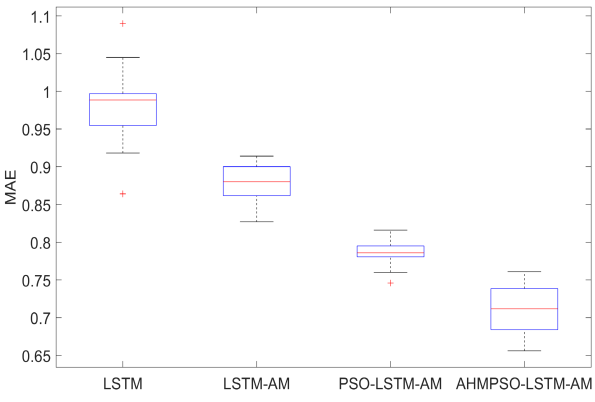

FIGURE 16. Box plot of MAE obtained from each model.

the ability of the LSTM model to fit the true value of effluent COD is enhanced, and the accuracy and stability of the model are further improved.

The above experimental results illustrate that the attention mechanism can improve the performance of the LSTM model. We will discuss the influence of the PSO algorithm and AHMPSO algorithm on model performance based on four evaluation criteria. The specific analysis is shown below.

1) MODEL GOODNESS OF FIT EVALUATION

R^2 can measure the goodness of fit for the model. A larger R^2 indicates that the model fits the dataset better. A smaller R^2 indicates that the model fits the dataset poorer. Compared with the LSTM-AM model, the R^2 of the PSO-LSTM-AM model is increased by 2.904%. Using the PSO algorithm to optimize the model hyperparameters can improve the model's goodness of fit, but this improvement is limited by the PSO algorithm's optimization accuracy. Compared with

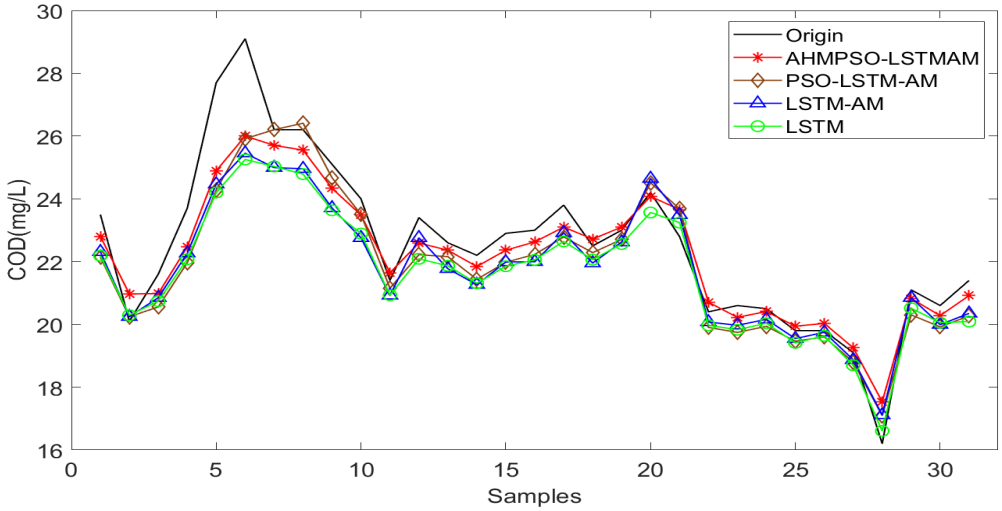

FIGURE 17. The prediction curve of the model proposed in the paper.

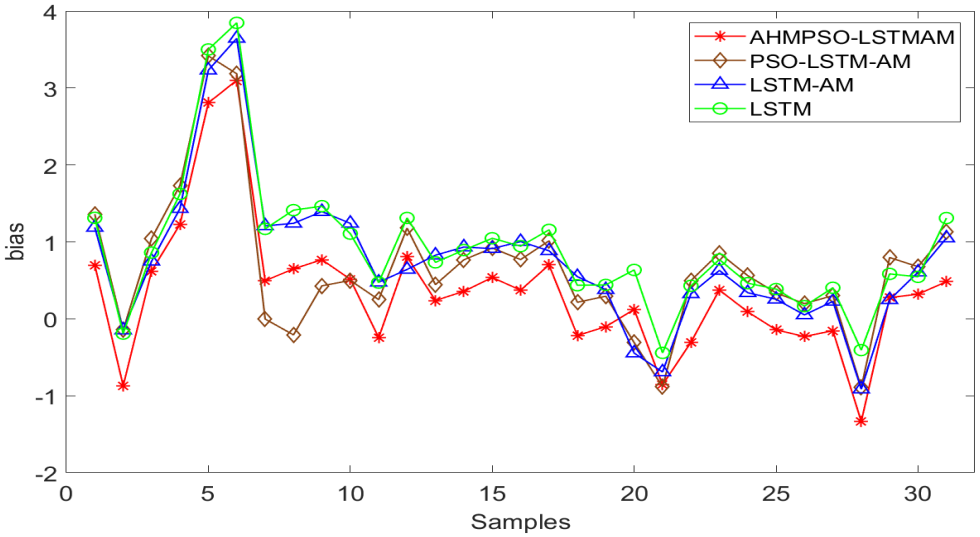

FIGURE 18. The bias curve of the model proposed in the paper.

PSO-LSTM-AM, the R^2 of the AHMPSO-LSTM-AM model is increased by 3.313%, which proves that the optimization accuracy of the AHMPSO algorithm is better than that of the PSO algorithm. Using the nonlinear inertia weight of the AHMPSO algorithm can expand the optimization space of the algorithm, and the hybrid mutation strategy improves the optimization accuracy of the AHMPSO algorithm by balancing the global and local search abilities.

As shown in Fig. 9, although the AHMPSO-LSTM-AM model achieves better goodness of fit than the other models, its goodness of fit curve varies greatly. The reason for this result is that the AHMPSO algorithm expands the optimization space, and the model obtains one of the optimal solutions in each optimization process. However, different optimal solutions have different influences on the goodness of fit of the model, which leads to a larger variation in the goodness of fit curve of the AHMPSO-LSTM-AM model.

2) MODEL ACCURACY EVALUATION

The RMSE can measure the accuracy of the model. A smaller RMSE indicates a higher model accuracy; conversely, a higher RMSE indicates worse model accuracy. Compared with the LSTM-AM model and the PSO-LSTM-AM model, the average RMSE of the AHMPSO-LSTM-AM model is reduced by 13.029% and 7.803%, respectively. The experimental results illustrate that the prediction accuracy of the AHMPSO-LSTM-AM model is better than that of the PSO-LSTM-AM model. The reason for this result is that the number of neurons in the hidden layer of the LSTM-AM model is 70, and the network structure is relatively simple. Although the pattern of the dataset can be learned quickly during the training process, the accuracy of the model is also limited. The numbers of hidden layer neurons in the PSO-LSTM-AM model and the AHMPSO-LSTM-AM model are 142 and 114, respectively. The network structure

is relatively complex, and the dataset pattern can be learned more fully during the training process. The overly complex network structure easily overfits the model and reduces the prediction accuracy of the model, which is also the reason why the accuracy of the PSO-LSTM-AM model is slightly lower than that of the AHMPSO-LSTM-AM model.

In addition, the Adam optimizer can adjust the learning rate adaptively. When the learning rate is set to 0.002, the weight and bias update amplitude of the LSTM-AM model is small, which limits the further improvement of the model accuracy. However, a larger learning rate will make the weight and bias of the model update within a larger amplitude, resulting in a larger RMSE variation amplitude. As shown in Fig. 11, the RMSE of the AHMPSO-LSTM-AM model reached a minimum value of 0.923 in the 10th experiment and a maximum value of 1.103 in the 19th experiment. In summary, although the RMSE of the AHMPSO-LSTM-AM model varies greatly, the hyperparameter obtained through the optimization of the AHMPSO algorithm ensures that its average RMSE is the smallest among the models proposed.

3) EVALUATION OF MODEL STRENGTHS AND WEAKNESSES

The MAPE can measure the strengths and weaknesses of models. A MAPE value that is closer to 0% indicates a model that is of higher quality; a MAPE value closer to 100% indicates that the model is inferior. As shown in Table 4, the average MAPE of the AHMPSO-LSTM-AM model (3.046) is the smallest. This indicates that compared with other models proposed in this paper, the model has better performance. The reason for this is that a more complex network structure enables the model to better handle nonlinear datasets, thereby making it easier for the model to identify patterns in the dataset to improve the performance of the model. In addition, the larger learning rate enables the model to quickly find the direction of gradient descent in the early stage of training, and the Adam optimizer can adaptively adjust the learning rate of the model so that the model can converge faster.

However, the more complex network structure and higher learning rate will also bring certain negative effects. From Fig. 13, we can see that the MAPE of the AHMPSO-LSTM-AM model varies greatly because the complexity of the network structure has caused parameter (weights, biases) complexity, and the higher learning rate will make the model weights and biases update more extensively, which leads to the MAPE of the model have a large amplitude change over several experiments.

4) MODEL ERROR EVALUATION

The MAE is used to describe the average value of the absolute value error between the predicted value and the true value. It is the average value of the error in a more general form and can better reflect the error of the predicted value. In Fig. 15, it can be seen that the MAE of the AHMPSO-LSTM-AM model in 20 historical experiments is lower than that of the other models. The reason for this result is that the attention mechanism can improve the accuracy of the model, thus reducing the error between the predicted value and the true value. In addition, the hyperparameters obtained by the AHMPSO algorithm can also guarantee a smaller MAE value for the AHMPSO-LSTM-AM model.

TABLE 5. Evaluation indicators for the average prediction effect of each model under twenty runs.

Model	R^2	RMSE	MAPE	MAE
LSTM	0.756	1.274	4.116	0.983
LSTM-AM	0.792	1.182	3.777	0.899
PSO-LSTM-AM	0.814	1.117	3.503	0.816
AHMPSO-LSTM-AM	0.869	0.953	2.843	0.665

F. ANALYSIS OF THE PREDICTION RESULTS

Fig. 17 shows the average prediction results for all models from 20 random experiments. All the models proposed in this paper can predict the effluent COD trend of the WWTP more accurately.

Fig. 18 shows the bias curve of the model. We can see from it that among all the models proposed in this paper, the LSTM model has the largest bias, and the goodness of fit for the effluent COD is poor. It can be seen from Table 5 that after the introduction of the attention mechanism, the goodness of fit of the LSTM-AM model is improved, R^2 is increased from 0.756 to 0.792, and the predictive ability is improved. In addition, other evaluation criteria have also been improved: RMSE decreased by 7.783%, MAPE decreased by 8.236%, and MAE decreased by 8.545%. This explains that the attention mechanism improves the ability of the LSTM-AM model to mine local important features of wastewater data under the same parameters as the LSTM model, thus increasing the prediction effect and accuracy of the model.

Compared with the LSTM-AM model, after the PSO algorithm was used to optimize the hyperparameters of the PSO-LSTM-AM model, R^2 was improved from 0.792 to 0.814, RMSE was reduced by 5.499%, MAPE was reduced by 7.254%, and MAE was reduced by 9.232%. In addition, Fig. 17 and Fig. 18 show that the bias of the PSO-LSTM-AM model was further reduced, and the prediction effect of some abnormal points was also improved. This illustrates that using the PSO algorithm to optimize the number of hidden layer neurons and the learning rate can enhance the prediction effect and accuracy of the model. As seen from Table 5, compared with the PSO-LSTM-AM model, the R^2 of the AHMPSO-LSTM-AM model increased from 0.814 to 0.869, RMSE decreased by 14.682%, MAPE decreased by 18.841% and MAE decreased by 18.505%. The bias curve is relatively small compared with other models and is closer to zero. This illustrates that the ability of the AHMPSO-LSTM-AM model to fit abnormal points has been further improved. At the same time, the ability of the model to fit the true value of the effluent COD trend has been improved, and the prediction accuracy of the model has been significantly improved compared with the LSTM model. The reason for this result is that the AHMPSO algorithm has better optimization capabilities than the standard PSO algorithm

and can help the model obtain a better network structure and learning rate.

IV. CONCLUSION

In this paper, an LSTM model based on the AHMPSO algorithm and attention mechanism was proposed to monitor the key features of sewage treatment in sewage treatment plants. First, the AHMPSO algorithm is used to optimize the hyperparameters of the LSTM-AM model, including the number of hidden layer neurons and the learning rate. Second, the hyperparameters obtained by the AHMPSO algorithm are used to establish an LSTM-AM model and train the model on the wastewater data training set. Finally, the wastewater data test set is input into the trained model to obtain the predicted value of the WWTP effluent COD. The simulation results show the following:

1) Compared with the PSO algorithm, the introduction of nonlinear inertial weights with random factors and adaptive hybrid mutation methods enables the AHMPSO algorithm to better balance the global search ability and the local search ability, reduce the probability of the algorithm falling into local optima, and improve the convergence speed and accuracy of the algorithm.
2) With the attention mechanism, the ability of the LSTM neural network to learn the importance of wastewater local features has been strengthened. The AHMPSO algorithm is used to optimize the hyperparameters of the LSTM-AM model, which offsets the cumbersome and time-consuming shortcomings of manual selection and effectively improves the prediction accuracy of the model.
3) Compared with the LSTM model, the LSTM-AM model, and the PSO-LSTM-AM model, the AHMPSO-LSTM-AM model achieves better prediction accuracy and stability in terms of predicting effluent COD. The AHMPSO-LSTM-AM model can enhance the monitoring ability of wastewater treatment key indicator features, which is beneficial to WWTPs to obtain more stable and accurate prediction results of wastewater features.

In summary, the AHMPSO-LSTM-AM model can provide a stable and effective tool for monitoring wastewater's key features.

The focus of this research is to optimize the structure of the effluent COD prediction model and improve the prediction accuracy of the model for effluent COD. However, the presence of some noise in the wastewater dataset will have some influence on the model's ability to extract key features. Therefore, in future research, we will consider further optimization of the model in terms of data noise reduction and feature construction.

REFERENCES

[1] J. Wang, K. Wan, X. Gao, X. Cheng, Y. Shen, Z. Wen, U. Tariq, and M. J. Piran, "Energy and materials-saving management via deep learning for wastewater treatment plants," *IEEE Access*, vol. 8, pp. 191694–191705, 2020.

[2] H. Han, S. Zhu, J. Qiao, and M. Guo, "Data-driven intelligent monitoring system for key variables in wastewater treatment process," *Chin. J. Chem. Eng.*, vol. 26, no. 10, pp. 2093–2101, Oct. 2018.

[3] M. Huang, Y. Ma, J. Wan, and X. Chen, "A sensor-software based on a genetic algorithm-based neural fuzzy system for modeling and simulating a wastewater treatment process," *Appl. Soft Comput.*, vol. 27, pp. 1–10, Feb. 2015.

[4] H. Jin, X. Chen, L. Wang, K. Yang, and L. Wu, "Adaptive soft sensor development based on online ensemble Gaussian process regression for nonlinear time-varying batch processes," *Ind. Eng. Chem. Res.*, vol. 54, no. 30, pp. 7320–7345, Oct. 2015.

[5] E. B. Hassen and A. M. Asmare, "Predictive performance modeling of Habesha brewery wastewater treatment plant using artificial neural networks," *Chem. Int.*, vol. 5, no. 1, p. 87, Jan. 2019.

[6] I. Pisa, I. Santin, A. Morell, J. L. Vicario, and R. Vilanova, "LSTM-based wastewater treatment plants operation strategies for effluent quality improvement," *IEEE Access*, vol. 7, pp. 159773–159786, 2019.

[7] P. A. Paraskevas, I. S. Pantelakis, and T. D. Lekkas, "An advanced integrated expert system for wastewater treatment plants control," *Knowl.-Based Syst.*, vol. 12, no. 7, pp. 355–361, Nov. 1999.

[8] H. Hauduc, I. Takács, S. Smith, A. Szabo, S. Murthy, G. T. Daigger, and M. Spérandio, "A dynamic physicochemical model for chemical phosphorus removal," *Water Res.*, vol. 73, pp. 157–170, Apr. 2015.

[9] C. Sicard, C. Glen, B. Aubie, D. Wallace, S. Jahanshahi-Anbuhi, K. Pennings, G. T. Daigger, R. Pelton, J. D. Brennan, and C. D. M. Filipe, "Tools for water quality monitoring and mapping using paper-based sensors and cell phones," *Water Res.* vol. 70, pp. 360–369, Mar. 2015.

[10] K.-J. Wang, P.-S. Wang, and H.-P. Nguyen, "A data-driven optimization model for coagulant dosage decision in industrial wastewater treatment," *Comput. Chem. Eng.*, vol. 152, Sep. 2021, Art. no. 107383.

[11] L. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, and M. Poch, "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," *Environ. Model. Softw.*, vol. 106, pp. 89–103, Aug. 2018.

[12] D. Kumar, K. Karwasra, and G. Soni, "Bibliometric analysis of artificial neural network applications in materials and engineering," *Mater. Today, Proc.*, vol. 28, Jan. 2020, pp. 1629–1634.

[13] A. K. Maurya, B. S. Reddy, J. Theerthagiri, P. L. Narayana, C. H. Park, J. K. Hong, J.-T. Yeom, K. K. Cho, and N. S. Reddy, "Modeling and optimization of process parameters of biofilm reactor for wastewater treatment," *Sci. Total Environ.*, vol. 787, Sep. 2021, Art. no. 147624.

[14] A. N. Matheri, F. Ntuli, J. C. Ngila, T. Seodigeng, and C. Zvinowanda, "Performance prediction of trace metals and cod in wastewater treatment using artificial neural network," *Comput. Chem. Eng.*, vol. 149, Jun. 2021, Art. no. 107308.

[15] M. H. Bakr, M. Nasr, M. Ashmawy, and A. Tawfik, "Predictive performance of auto-aerated immobilized biomass reactor treating anaerobic effluent of cardboard wastewater enriched with bronopol (2-bromo-2-nitropropan-1,3-diol) via artificial neural network," *Environ. Technol. Innov.*, vol. 21, Feb. 2021, Art. no. 101327.

[16] F. Facchini, L. Ranieri, and M. Vitti, "A neural network model for decision-making with application in sewage sludge management," *Appl. Sci.*, vol. 11, no. 12, p. 5434, Jun. 2021.

[17] N. Bekkari and A. Zeddouri, "Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant," *Manage. Environ. Qual., Int. J.*, vol. 30, no. 3, pp. 593–608, Apr. 2019.

[18] V. Nourani, G. Elkiran, and S. I. Abba, "Wastewater treatment plant performance analysis using artificial intelligence—An ensemble approach," *Water Sci. Technol.*, vol. 78, no. 10, pp. 2064–2076, Dec. 2018.

[19] D. Zhang, E. S. Hølland, G. Lindholm, and H. Ratnaweera, "Hydraulic modeling and deep learning based flow forecasting for optimizing inter catchment wastewater transfer," *J. Hydrol.*, vol. 567, pp. 792–802, Dec. 2018.

[20] D. Zhang, G. Lindholm, and H. Ratnaweera, "Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring," *J. Hydrol.*, vol. 556, pp. 409–418, Jan. 2018.

[21] X. Wang, Y. Qin, Y. Wang, S. Xiang, and H. Chen, "ReLTanh: An activation function with vanishing gradient resistance for SAE-based DNNs and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 363, pp. 88–98, Oct. 2019.

[22] E. Balaji, D. Brindha, V. K. Elumalai, and R. Vikrama, "Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM network," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107463.

[23] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Syst. Appl.*, vol. 115, pp. 465–473, Jan. 2019.
[24] Z. Peng, J. Dang, M. Unoki, and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech," *Neural Netw.*, vol. 140, pp. 261–273, Aug. 2021.
[25] K. Shuang, Y. Tan, Z. Cai, and Y. Sun, "Natural language modeling with syntactic structure dependency," *Inf. Sci.*, vol. 523, pp. 220–233, Jun. 2020.
[26] C. Jörges, C. Berkenbrink, and B. Stumpe, "Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks," *Ocean Eng.*, vol. 232, Jul. 2021, Art. no. 109046.
[27] H. S. Gill and B. S. Khehra, "An integrated approach using CNN-RNN-LSTM for classification of fruit images," *Mater. Today, Proc.*, to be published, doi: 10.1016/j.matpr.2021.06.016.
[28] Z. Zhuang, Z. Sun, Y. Cheng, R. Yao, and W. Zhang, "Modeling and optimization of paper-making wastewater treatment based on reinforcement learning," in *Proc. 37th Chin. Control Conf. (CCC)*, Wuhan, China, Jul. 2018, pp. 8342–8346.
[29] M. Yaqub, H. Asif, S. Kim, and W. Lee, "Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network," *J. Water Process Eng.*, vol. 37, Oct. 2020, Art. no. 101388.
[30] T. Cheng, F. Harrou, F. Kadri, Y. Sun, and T. Leiknes, "Forecasting of wastewater treatment plant key features using deep learning-based models: A case study," *IEEE Access*, vol. 8, pp. 184475–184485, 2020.
[31] I. Pisa, A. Morell, J. L. Vicario, and R. Vilanova, "LSTM-based IMC approach applied in wastewater treatment plants: Performance and stability analysis," *IFAC-Papers OnLine*, vol. 53, no. 2, pp. 16569–16574, 2020.
[32] X. Wang, S. Chen, and J. Su, "Real network traffic collection and deep learning for mobile app identification," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–14, Feb. 2020.
[33] N. Yoon, J. Kim, J.-L. Lim, A. Abbas, K. Jeong, and K. H. Cho, "Dual-stage attention-based LSTM for simulating performance of brackish water treatment plant," *Desalination*, vol. 512, Sep. 2021, Art. no. 115107.
[34] H. Qun, L. Wenjing, and C. Zhangli, "B&ANet: Combining bidirectional LSTM and self-attention for end-to-end learning of task-oriented dialogue system," *Speech Commun.*, vol. 125, pp. 15–23, Dec. 2020.
[35] H. Zang, R. Xu, L. Cheng, T. Ding, L. Liu, Z. Wei, and G. Sun, "Residential load forecasting based on LSTM fusing self-attention mechanism with pooling," *Energy*, vol. 229, Aug. 2021, Art. no. 120682.
[36] Y. Han, C. Fan, M. Xu, Z. Geng, and Y. Zhong, "Production capacity analysis and energy saving of complex chemical processes using LSTM based on attention mechanism," *Appl. Thermal Eng.*, vol. 160, Sep. 2019, Art. no. 114072.
[37] F. Shahid, A. Zameer, and M. Muneeb, "A novel genetic LSTM model for wind power forecast," *Energy*, vol. 223, May 2021, Art. no. 120069.
[38] N. Ding, H. Li, Z. Yin, N. Zhong, and L. Zhang, "Journal bearing seizure degradation assessment and remaining useful life prediction based on long short-term memory neural network," *Measurement*, vol. 166, Dec. 2020, Art. no. 108215.
[39] Y. Zhang and S. Yang, "Prediction on the highest price of the stock based on PSO-LSTM neural network," in *Proc. 3rd Int. Conf. Electron. Inf. Technol. Comput. Eng. (EITCE)*, Xiamen, China, Oct. 2019, pp. 1565–1569.
[40] X. Song, Y. Liu, L. Xue, J. Wang, J. Zhang, J. Wang, L. Jiang, and Z. Cheng, "Time-series well performance prediction based on long short-term memory (LSTM) neural network model," *J. Petroleum Sci. Eng.*, vol. 186, Mar. 2020, Art. no. 106682.
[41] H. Feng, W. Ma, C. Yin, and D. Cao, "Trajectory control of electro-hydraulic position servo system using improved PSO-PID controller," *Autom. Construct.*, vol. 127, Jul. 2021, Art. no. 103722.
[42] H. Tian, P. Wang, K. Tansey, J. Zhang, S. Zhang, and H. Li, "An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong plain, PR China," *Agricult. Forest Meteorol.*, vol. 310, Nov. 2021, Art. no. 108629.
[43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 2048–2057.
[44] W. Huang, Y. Li, and Y. Huang, "Prediction of chaotic time series using hybrid neural network and attention mechanism," *Acta Phys. Sinica*, vol. 70, no. 1, pp. 235–243, Jan. 2021.
[45] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw. (ICNN)*, vol. 4, Perth, WA, Australia, Dec. 1995, pp. 1942–1948.
[46] J. Shi, Z. Yang, and P. Liu, "Hybrid mutation particle swarm optimization algorithm and its application," *Math. Pract. Theory*, vol. 51, no. 1, pp. 150–161, Jan. 2021.
[47] T. Wei and T. Pan, "Short-term power load forecasting based on LSTM neural network optimized by improved PSO," *J. Syst. Simul.*, vol. 33, no. 8, pp. 1866–1874, Sep. 2021.
[48] W. Ji, H. Xu, and P. Lin, "Adaptive mutation particle swarm optimization and its application in predicting the COVID-19 epidemic transmission," *J. Chin. Comput. Syst.*, vol. 42, no. 3, pp. 472–477, Mar. 2021.
[49] Z. Chang, Y. Zhang, and W. Chen, "Electricity price prediction based on hybrid model of Adam optimized LSTM neural network and wavelet transform," *Energy*, vol. 187, Nov. 2019, Art. no. 115804.

XIN LIU received the B.Sc. degree in telecommunication engineering from Jilin University, Jilin, China, in 2001, and the Ph.D. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010. He is currently a Professor with the Hebei University of Engineering, Handan, Hebei, China. From 2001 to 2003, he was a Technical Support Engineer with Huawei Technologies Company Ltd., Shenzhen, China. He was a recipient of the (2010–2011) Québec-China Postdoctoral Fellowship Scholarships and was a Postdoctoral Research Fellow with the Optical Zeitgeist Laboratory, INRS-EMT, Montreal, QC, Canada, from September 2010 to September 2011. From 2014 to 2015, he was a Visiting Scientist with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA. His research interests include the area of machine learning, intelligent water conservancy green, networks, and network reliability.

QIMING SHI was born in Shandong, China, in 1996. He received the bachelor's degree from the Chengdu University of Technology, in 2018. He is currently pursuing the master's degree with the School of Information and Electrical Engineering, Hebei University of Engineering. His research interests include machine learning and intelligent water conservancy.

ZHEN LIU received the B.S. degree in communication engineering and the M.S. degree in computer science and technology from the Hebei University of Engineering, Handan, Hebei, China, in 2013 and 2016, respectively, and the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2020. From 2016 to 2021, she was a Lecturer with the Hebei University of Engineering. Her research interests include optical networks, edge computing, and machine learning.

JIA YUAN received the Ph.D. degree in control science and engineering from Harbin Engineering University, Harbin, China, in 2019. She is currently a Lecturer with the School of Information and Electrical Engineering, Hebei University of Engineering. Her research interests include motion and attitude control of vehicle, high-performance ship control, and intelligent water conservancy systems.