Front. Environ. Sci. Eng. 2024, 18(3): 31
https://doi.org/10.1007/s11783-024-1791-x

RESEARCH ARTICLE

Brain-inspired multimodal approach for effluent quality prediction using wastewater surface images and water quality data

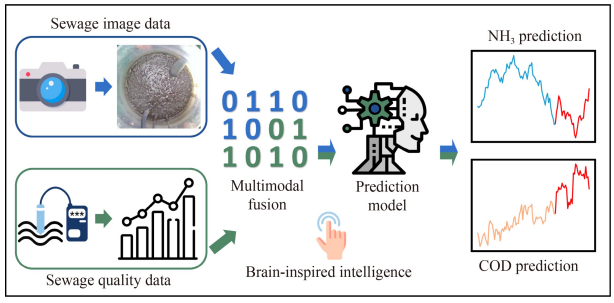
Junchen Li[1,2], Sijie Lin[2,3], Liang Zhang[4,5], Yuheng Liu[4,5], Yongzhen Peng (✉)[4,5], Qing Hu (✉)[2,3]

1 School of Environment, Harbin Institute of Technology, Harbin 150090, China
2 School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
3 Engineering Innovation Center of SUSTech (Beijing), Southern University of Science and Technology, Beijing 100083, China
4 Faculty of Environment and Life, Beijing University of Technology, Beijing 100124, China
5 Engineering Research Center of Intelligence Perception and Autonomous Control, Ministry of Education, Beijing 100124, China

HIGHLIGHTS

- A novel brain-inspired network accurately predicts sewage effluent quality.
- Sewage-surface images are utilized in data analysis by the model.
- The developed method outperforms traditional ones by reducing error by 23%.
- The model offers the potential for cost-effective monitoring.

GRAPHIC ABSTRACT



ARTICLE INFO

Article history:
Received 29 May 2023
Revised 17 October 2023
Accepted 17 October 2023
Available online 10 November 2023

Keywords:
Wastewater treatment system
Water quality prediction
Data driven analysis
Brain-inspired model
Multimodal data
Attention mechanism

ABSTRACT

Efficiently predicting effluent quality through data-driven analysis presents a significant advancement for consistent wastewater treatment operations. In this study, we aimed to develop an integrated method for predicting effluent COD and NH_3 levels. We employed a 200 L pilot-scale sequencing batch reactor (SBR) to gather multimodal data from urban sewage over 40 d. Then we collected data on critical parameters like COD, DO, pH, NH_3, EC, ORP, SS, and water temperature, alongside wastewater surface images, resulting in a data set of approximately 40246 points. Then we proposed a brain-inspired image and temporal fusion model integrated with a CNN-LSTM network (BITF-CL) using this data. This innovative model synergized sewage imagery with water quality data, enhancing prediction accuracy. As a result, the BITF-CL model reduced prediction error by over 23% compared to traditional methods and still performed comparably to conventional techniques even without using DO and SS sensor data. Consequently, this research presents a cost-effective and precise prediction system for sewage treatment, demonstrating the potential of brain-inspired models.

1 Introduction

Wastewater treatment plants (WWTPs) are essential components of sustainable and effective waste management schemes and play a crucial role in water environmental protection (Geerdink et al., 2017). The treatment of wastewater not only prevents the release of harmful pollutants into natural water bodies but also mitigates the risks to public health and aquatic ecosystems. Stable and high-quality effluent is indispensable for the effective operation of WWTPs. In determining the effluent quality, ammonia nitrogen (NH_3) and chemical oxygen demand (COD) are essential. As key control parameters in WWTPs, NH_3, and COD represent the level of nitrogenous waste and organic

✉ Corresponding authors
E-mails: pyz@bjut.edu.cn (Y. Peng); huq@sustech.edu.cn (Q. Hu)

pollutants in the effluent, respectively, that affect WWTP performance and compliance with regulatory standards (Wu et al., 2019). Therefore, the accurate prediction of NH_3 and COD effluent concentrations is vital for ensuring water safety, satisfying regulatory standards, and optimizing the management of WWTPs.

Data-driven methods for predicting effluent water quality have become increasingly popular due to the increasing volume of available data produced by WWTP operations. Several models have been developed to establish connections between available water quality data and effluent quality. Among these, time series analysis using linear and nonlinear approaches has emerged as a widely accepted method (Tealab, 2018). Early research primarily utilized linear techniques such as multiple linear regression (Poutiainen et al., 2010; Zare Abyaneh, 2014) and the autoregressive integrated moving average model (Al-Asheh et al., 2007). These models assume linear relationships among the water quality parameters and provide low computational costs and easily interpretable results. In comparison, nonlinear techniques are effective at enhancing prediction accuracy because of their ability to simulate brain-like computations, making them particularly successful at predicting intricate systems (Mehonic and Kenyon, 2022). Thus, brain-inspired models, such as artificial neural networks (Lee et al., 2011; Nasr et al., 2012; Bekkari and Zeddouri, 2019), support vector machines (Guo et al., 2015; Granata et al., 2017; Wang et al., 2018; Fu et al., 2023), and kernel function-based optimization methods (Bagheri et al., 2015; Fernandez de Canete et al., 2016; Zhu et al., 2017; Liu et al., 2019), have been increasingly employed and are preferred when managing the complex and nonlinear behaviors of wastewater treatment systems, especially when considering the high interdependence among the parameters involved (Zodrow et al., 2017). These models adopt a broader set of characteristics such as the ability of the brain to process large amounts of information, to improve their prediction accuracy. In recent years, deep learning methods such as Long Short-Term Memory networks (LSTM) and convolutional neural networks (CNN), have emerged as powerful tools for time-series analysis in water quality prediction (Wang et al., 2023; Zhang and Li, 2023). These methods differ significantly from traditional linear and nonlinear models because they can handle more extensive and complex data sets, leading to improved prediction accuracy. For example, the LSTM model, which is inspired by the temporal dependency of human memory, excels at processing sequential data. Meanwhile, a CNN emulating the hierarchical structure of the human visual system allows effective feature extraction from images. For instance, Wang et al. (2017) demonstrated that LSTM models are more reliable than back-propagation neural networks and extreme learning machine models in predicting dissolved oxygen (DO) and total phosphorus (TP) in water. Similarly, Ta and Wei (2018) presented a CNN model for predicting DO concentration, which provided better feature extraction results and improved accuracy compared with other models. Despite significant advancements in data-driven methods, remaining challenges must be addressed to fully harness their potential for effluent water quality prediction. First, the availability of high-quality data are crucial for the success of these methods; however, obtaining such data are often limited and costly. Second, these methods require further optimization and improvement to enhance their performance, particularly regarding prediction accuracy and computational efficiency.

To achieve enhanced effluent quality prediction, researchers have explored brain-inspired techniques, particularly hybrid models that combine CNN and LSTM (Barzegar et al., 2020). Hybrid approaches leverage the efficient feature extraction capabilities of CNN and the time correlation capture ability of LSTM, resulting in a more dependable water quality prediction performance than single models. Additionally, image recognition technology offers a promising avenue for water quality prediction and serves as a potential integration method for diverse models. By analyzing images of wastewater or sludge samples, researchers can extract valuable information related to water quality parameters, making relatively high-accuracy prediction possible (Litjens et al., 2017; Rawat and Wang, 2017). Studies analyzing wastewater surface images aim to capture the relationships between the visual characteristics that are perceivable by the human eye and the target water quality parameters. This method offers benefits for real-time prediction as it capitalizes on readily available visual characteristics that are closely linked to effluent quality (Liu et al., 2014; Mullins et al., 2018; Li et al., 2022b). Data-driven models offer advantages in terms of capturing temporal correlations, whereas image recognition technology provides valuable information related to water quality parameters. Recently, multimodal data fusion has emerged as a focal point of research, yielding significant advancements across various domains (Li et al., 2022a). This approach, which is reminiscent of the ability of the human brain to process diverse data types concurrently, has demonstrated multiple advantages in terms of efficient and accurate practical applications (Lahat et al., 2015). For instance, its implementation in the financial and medical sectors has significantly enhanced the precision of stock market analysis and disease diagnostics, respectively (Lee and Yoo, 2020; Muhammad et al., 2021). We hypothesize that a brain-inspired multimodal approach mirroring the human brain's proficiency in processing diverse data would offer a more comprehensive representation of complex water quality patterns. This method not only aims to achieve enhanced prediction accuracy but also strives to reduce

the dependency of traditional methods on expensive sensors, subsequently lowering costs. Although recent algorithms and computational resource advancements have bolstered the brain-inspired multimodal strategy, consequential challenges remain. These encompass the limited availability of multimodal data in the wastewater sector, the uncharted territories of data fusion, model robustness, and data preprocessing and compatibility problems. Practical experiments with real-world data are paramount to validate the efficacy of the approach. Despite these challenges, the proposed brain-inspired multimodal method enhances traditional monitoring techniques, offering a more encompassing view of wastewater dynamics.

This study aims to enhance water quality prediction using a brain-inspired approach, integrating wastewater surface image features with time series data, and imitating how the human brain processes complex multimodal information. We propose a novel brain-inspired image and temporal fusion (BITF) with a CNN-LSTM network (CL) model for effective prediction which includes 1) an image feature extraction module that captures high-dimensional visual information from wastewater surface images; 2) an adaptive feature fusion method that assigns different weights to image and time series data; and 3) a CNN-LSTM architecture that effectively leverages both short and long-term data for accurate predictions. To evaluate the performance of the BITF-CL model, we compared it with other widely used models using real-world data sets and appropriate evaluation metrics to demonstrate the effectiveness and advantages of our proposed approach.

2 Experimental platform and data collection

2.1 Experimental platform

To investigate the dynamics of essential parameters in sewage treatment processes, we developed an experimental platform utilizing a 200 L sequencing batch reactor (SBR) made of acrylic, as depicted in Fig. 1(a). The influent of the reactor was collected from an actual urban sewage treatment plant to ensure accurate and reliable data for developing the water quality prediction model. The reactor operates in a 180-min cycle, with phases of synchronous filling and discharging (20 min), anaerobic treatment (25 min), aeration (120 min), and settling (15 min). During the aeration phase, the air pump typically provides an aeration velocity of 15 L/min under regular operation. This rate is adjustable and is continuously monitored using a flow gauge. An inlet tank was used to temporarily store the influent wastewater before it was fed into the SBR during the filling phase, while an outlet tank temporarily stored the treated water before being discharged. The adaptable design of the SBR allows adjustments to operational parameters as needed, thereby enhancing efficiency and control and facilitating the optimization of sewage treatment.

2.2 Image and sensor data collection

To acquire multimodal wastewater data, a high-resolution overhead camera (WSD-2133-V01, Weishida, China) with a resolution of 1920 × 1080 was used to capture images of the sewage surface every minute. These images provided insights into the physical and chemical changes at the liquid surface during treatment. As shown in Fig. 1 (b), water quality sensor arrays (LH-G8820, Lohand, China) were strategically placed. Influent water quality sensors were placed in the inlet tank, while those for effluent readings were positioned near the outlet of the SBR. These sensors continuously monitored essential indicators such as COD, DO, pH, NH_3, electrical conductivity (EC), oxidation-reduction potential (ORP), suspended solids (SS), and temperature at consistent 1-min intervals. The study monitored 16 indicators using 8 sensors in influent and effluent streams. However, when predicting specific effluent parameters, such as NH_3 and

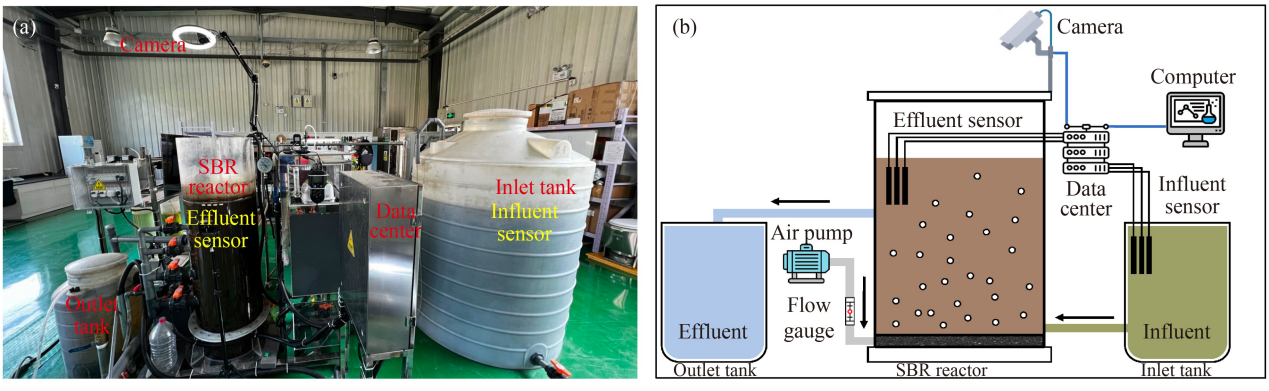


Fig. 1 Schematic of the (a) experimental platform and (b) platform structure.

COD, we excluded their effluent values from the input features to prevent data leakage and ensure robust model training. As a result, the maximum number of parameters used in the predictions is 15.

All the monitored indicators were chosen based on their known impact on microbial activity and the overall effectiveness of wastewater biological treatments (Mulkerrins et al., 2004; Guštin and Marinšek-Logar, 2011). Furthermore, they are essential tools for effectively monitoring the treatment process (Yu et al., 2013; Alattabi et al., 2017). We used a registration script to synchronize image capture timestamps and sensor readings to ensure consistent data collection. This precise alignment was essential, especially considering the 1-min data collection intervals. Maintaining data quality was of identical importance. The Outliers were removed, and data normalization was performed to ensure consistent scaling of all water quality metrics, preventing any excessive impact on the model.

2.3 Experiment settings

To rigorously evaluate the efficiency of the BITF-CL model in predicting effluent COD and NH_3 concentrations, we collected 40 d of multimodal data consisting of 40246 data points, each including corresponding water quality and image data. The data set was divided into 36 d under normal conditions and 4 d under anomalous conditions. These anomalous conditions were created by manually modifying the aeration velocity. Further information about these operational shifts will be provided in later sections.

For model assessment, we adopted two distinct partitioning methodologies. In the regular conditions, data was split in a 6:3:1 ratio for training, validation, and testing. Results from this test set formed our primary evaluation metrics, revealing model performance in typical scenarios. Conversely, regular conditions data was used for training and validation under anomalous conditions, while anomalous data was reserved for testing, allowing us to evaluate the model's adaptability to operational changes. The previous 150 data points were employed for each prediction since this period covers necessary processing steps throughout the entire 180-min SBR cycle. This approach ensures accurate short- and long-term forecasts, and we based our predictions on historical data at 1-min and 1-h intervals.

We contrasted the model against LSTM and CNN-LSTM in performance benchmarks to highlight its specific advantage in multimodal prediction accuracy. The development of these models was facilitated using Pytorch 1.8.0 and TensorFlow 2.2.0 (Paszke et al., 2019; Pang et al., 2020) frameworks. Our experimental environment featured an Intel Xeon E5-2679 v4 CPU and an RTX 2080Ti GPU. We used Ubuntu 18.04 as the operating system and Python 3.6 as the development language.

3 Model establishment and evaluation

3.1 Development of the BITF module

Inspired by the brain's ability to integrate multimodal information, we developed a BITF module. This module aims to combine sewage surface images and water quality parameters, offering valuable insights for wastewater treatment. These images, rich in visual indicators, reveal the physicochemical of sewage attributes through variations in texture and color. For instance, color variations can reflect changes in chemical concentrations, while changes in texture can indicate differences in sediment granularity or distribution. To accurately interpret these visual cues and to further the utility of the BITF module, a robust image feature extraction method was necessary. We incorporated the VGG11 network into the BITF due to its proven efficiency in extracting intricate image features with a relatively simpler architecture than deeper networks, ensuring accuracy and computational efficiency. As illustrated in Fig. 2, VGG11 comprises stacked convolutional (conv), max-pooling (pooling), and fully connected (fc) layers. These layers transform simple textures into more complex visual concepts (Sengupta et al., 2019).

The conv layer is pivotal in this transformation process, which allows the network to identify and highlight intricate image patterns by the following equation (Eq. (1)):

$$I_{\text{out}} = \sum_i \sum_j I_{\text{in}}(i,j) \times K(x-i, y-j). \tag{1}$$

First, in the conv layer, the x and y denote the coordinates in the output image. Meanwhile, the i and j refer to the positions in the input image multiplied with a specified filter. I_{out} and I_{in} are the output and input images, respectively. Lastly, K is the conv kernel or filter

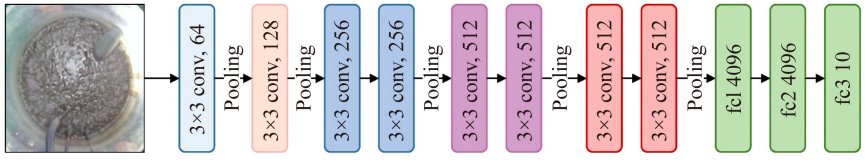


Fig. 2 Image feature extraction process.

that detects specific features. Subsequently, the pooling layer is used to reduce the size of the feature map while preserving essential features (Peng et al., 2019). This methodology enhances computational performance and augments the network's resilience to overfitting. Finally, after the pooling layer, the extracted features are integrated into feature vectors through stacked fc layers to capture patterns and relationships in the image data for practical analysis and prediction. With this foundational understanding, it is crucial to consider the practical adaptations made for our specific study. We resized the original 1920 × 1080 wastewater images to 224 × 224 to comply with the network input requirements. This allowed us to efficiently extract visual features using the VGG11 architecture tailored to our needs. We modified the last fully connected layer to output a 10-dimensional vision feature vector c instead of the conventional 1000-dimensional feature vector, ensuring a more accurate representation of the visual features inherent to wastewater during treatment.

After image feature extraction, we focused on integrating image and temporal data within the BITF module, utilizing a self-attention mechanism. Inspired by the brain's capacity to prioritize specific sensory inputs, this mechanism quantifies the significance of diverse elements in the input sequence, and subsequently modulates their impact on the output (Niu et al., 2021). In the fusion process, the 10-dimensional image feature vector c was concatenated with the water quality feature vector s, which has a dimensionality of up to 15, depending on the sensors used. This results in an input feature vector m with a maximum dimensionality of 25. The Fig. 3 visualizes this self-attention-based fusion process.

The input features from m were first transformed into three distinct feature spaces, $f(m_i)$, $g(m_j)$, and $h(m_i)$, by their respective linear transformations W_f, W_g, and W_h. As showcased in Fig. 3, the attention weights, denoted as $A_{i,j}$, where i and j are indices of the input feature map, were derived using the formula (Eq. (2)):

$$A_{i,j} = \mathrm{softmax}\left(f(m_i)^{\top} g(m_j)\right). \tag{2}$$

The $f(m_i)$ and $g(m_j)$ are the query and key vectors. After being processed by the softmax function, their dot product yields $A_{i,j}$. This weight quantifies the importance of each data input, indicating which features should be emphasized more during the subsequent fusion, whether from imagery or water quality metrics.

By using the attention weights in Eq. (2), the value vector $h(m_i)$ informs the weighted sum of features to generate the final multimodal feature vector O_j. This vector reflects the joint expression of sewage image and water quality features, as given by (Eq. (3)):

$$O_j = \sum_{i=1}^{N} A_{i,j} h(m_i). \tag{3}$$

Significantly, the multimodal feature vector O_j, which assimilates insights from image and water quality data, preserves its dimensionality, matching that of m at 25 dimensions. The distinctiveness lies in its construction: rather than relying on predefined weights, it is molded by the model's adaptively learned weights. The BITF proficiency in feature extraction and fusion exemplifies a brain-inspired design that emulates the ability of the human brain to synthesize sensory inputs into a coherent interpretation. The multimodal feature vector is critical for achieving accurate water quality predictions because of its encompassing nature.

3.2 Construction of the BITF-CL model

Expanding on the BITF module introduced in Section 3.1, we further integrated it with the CNN-LSTM network to formulate the comprehensive BITF-CL model. As depicted in Fig. 4, the multimodal feature vectors O_j, generated by processing sewage images and water quality data through the BITF module, serve as a vital input to the subsequent CNN-LSTM network. This approach

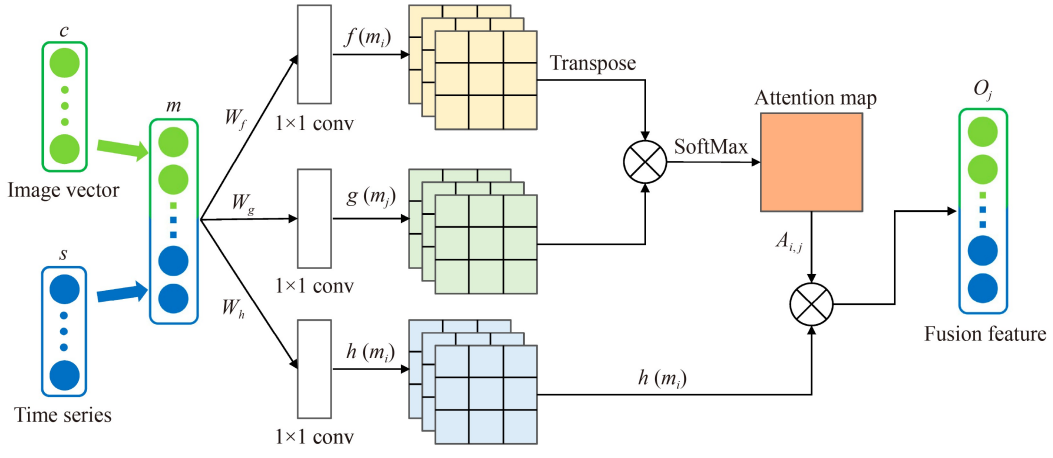


Fig. 3 Image and time series feature fusion based on self-attention.

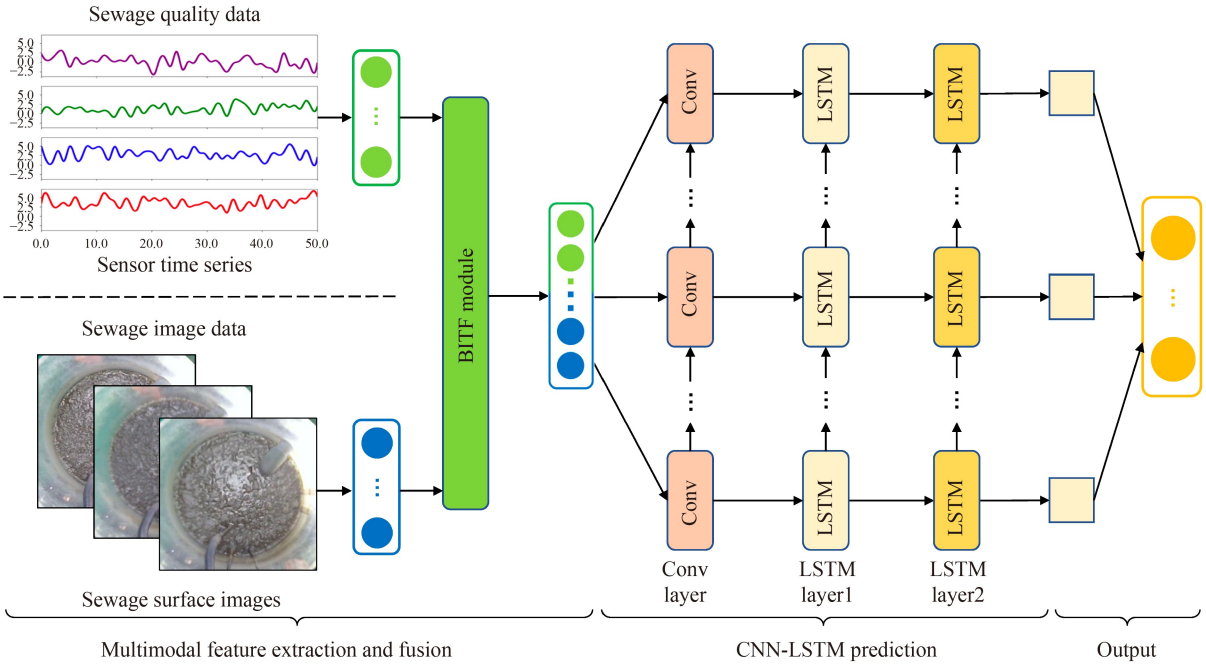


Fig. 4 Model architecture of BITF-CL.

leverages the advantages of multimodal data, enhancing prediction accuracy and robustness. The CNN-LSTM architecture, consisting of an initial CNN layer followed by two LSTM layers, is designed for intricate temporal pattern extraction. It excels in extracting both local temporal patterns and long-range dependencies from O_j, achieving superior water quality forecasting compared to traditional LSTM networks (Yang et al., 2021).

The initial CNN layer, equipped with 128 filters of 3 × 3 kernel size, refined the O_j vectors, producing feature maps represented as O_t for each time instance t. Following this, the feature map O_t *is* relayed to the two LSTM layers, each containing 64 units and incorporating a 0.2 dropout rate to mitigate overfitting. These LSTM layers are integral in analyzing time-series data, extracting essential insights for water quality prediction, and utilizing the forget, input, and output gate mechanisms, further illustrated by the subsequent equations (Eqs. (4)–(6)):

$$F_t = \sigma(W_f \cdot [h_{t-1}, O_t] + b_f), \tag{4}$$

$$I_t = \sigma(W_i \cdot [h_{t-1}, O_t] + b_i), \tag{5}$$

$$C_t = \tanh(W_c \cdot [h_{t-1}, O_t] + b_c), \tag{6}$$

where F_t represents the forget gate, which determines the information to be discarded from the previous cell state, improving the model's ability to focus on relevant data. I_t is the input gate, which controls the degree to which new information contributes to the updated cell state. C_t is a candidate value that helps incorporate new information into the cell state. The weight matrices of the corresponding gates are denoted by W_*, and the biases are represented by b_*. Here, * acts as a wildcard. The σ in equations is the sigmoid activation function, which helps control the flow of information through the gates. The tanh is the hyperbolic tangent function aiding in regulating the cell state.

The two LSTM layers work in tandem to capture short- and long-term correlations in the time-series data, which are essential for extracting crucial information for water quality prediction. By integrating these modules, the BITF-CL model can analyze multimodal features, providing a more comprehensive understanding of water quality dynamics. This ability can potentially enhance the predictive performance of the model in water quality prediction tasks.

3.3 Effluent quality prediction evaluation

In this study, we evaluate the effectiveness of the model using three evaluation indexes: root mean square error (RMSE), coefficient of determination (R^2), and mean absolute percentage error (MAPE) (Chicco et al., 2021). These indexes help measure the difference between predicted and observed values.

The RMSE calculates the square root of the mean square deviation between the predicted values and the target observation values to evaluate the model. The calculation formula is as follows (Eq. (7)):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{7}$$

where y_i is the actual value of the i^{th} target, $\hat{y}_i$ is the

model's predicted value for the i^{th} point, and n is the total number of data points. A lower RMSE value indicates a minor difference between the predicted and actual values, meaning that the model has a higher prediction accuracy.

The R^2 is a measure of the proportion of the variance in the target variable that is predictable from the independent variables. It provides an indication of how well the model's predictions fit the actual values. The calculation formula is as follows (Eq. (8)):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n} y_i - \bar{y}_i^{\,2}}, \tag{8}$$

where y_i is the actual value of the i^{th} target, $\widehat{y_i}$ is the model's predicted value for the i^{th} point, $\bar{y}_i$ is the mean of the target values, and n is the total number of data points. The R^2 value ranges from 0 to 1, with higher values indicating better model performance.

The MAPE is used to evaluate the mean absolute percentage deviation between the predicted values and the target values. The calculation formula is as follows (Eq.(9)):

$$\mathrm{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\widehat{y_i} - y_i}{y_i} \right|, \tag{9}$$

where y_i is the actual value of the i^{th} target, $\widehat{y_i}$ is the model's predicted value for the i^{th} point, and n is the total number of data points. The smaller MAPE is, the smaller the average percentage difference between the predicted and actual values, which means that the model's prediction accuracy is higher.

4 Results and discussion

4.1 BITF-CL and other unimodal models

In this study, we developed the BITF-CL model for predicting COD and NH_3 levels. Parallel to our brains processing multi-sensory information, our model integrated the sewage surface image features with water quality data. To evaluate the performance of the BITF-CL model, we employed data collected under regular conditions detailed in Section 2.3 to establish a standard benchmark. We evaluated the model's capability at 1-min and 1-h intervals, comparing it with conventional unimodal methods, specifically LSTM and CNN-LSTM, which rely solely on water quality data.

Tables 1 and 2 compare the prediction performance of different models for COD and NH_3 levels at 1-min and 1-h intervals, respectively. The BITF-CL model consistently outperformed the other models, with an R^2 improvement of 4.40% over CNN-LSTM and 9.20% over LSTM for COD at 1-min, and a gain of 5.43% over CNN-LSTM and 8.99% over LSTM for NH_3 at the same interval. Similar improvements were observed for predictions at the 1-h intervals. This improved performance can be attributed to the unique image feature extraction and fusion mechanism inspired by the brain processing visual information, as employed by the BITF-CL model. The mechanism effectively includes other visual features into the unimodal time series, enhancing the performance of model when handling multimodal data input.

Figure 5 compares the predicted NH_3 and COD values for BITF-CL, LSTM, and CNN-LSTM against the actual measurements at 1-min and 1-h intervals, which witness sudden COD value fluctuations ranging from 40 mg/L to 200 mg/L. These fluctuations mainly owe to the SBR process entering the aeration stage, where the sludge-water mixture disrupts the fluorescence-based COD sensor. Such interference does not compromise the predictive capacity of our model. Focusing on the 1-min interval (Figs. 5(a) and 5(b)), the BITF-CL model impressively captures the actual values, recording an RMSE of 20.75 for COD prediction and 0.30 for NH_3, demonstrating its precision. For MAPE, it exhibits superiority with the lowest recorded values of 0.18 and 0.04 for COD and NH_3, respectively. In contrast, the unimodal models tended to overestimate NH_3 levels, particularly during the 100–300 and 800–900 min, as suggested by their higher RMSE and MAPE values.

For the 1-h interval predictions (Figs. 5(c) and 5(d)), the BITF-CL model continues to demonstrate its exceptional performance. Despite the rise in RMSE by 17.59% and 16.67% for COD and NH_3, the BITF-CL model remains robust. The R^2 declined modestly by 4.21% and 4.12%, while the MAPE exhibited minor increases of 16.67% and 25.00%, underscoring the model's capability in longer-period prediction. In contrast, LSTM and CNN-LSTM models struggle with abrupt changes in actual COD values, reflected by their heightened RMSE and MAPE values. Significantly, the

Table 1 Performance comparison among different models (1-min)

Index	Model	COD			NH_3		
		RMSE	MAPE	R^2	RMSE	MAPE	R^2
Model 1	BITF-CL	20.75	0.18	0.95	0.30	0.04	0.97
Model 2	CNN-LSTM	26.94	0.21	0.91	0.52	0.05	0.92
Model 3	LSTM	29.69	0.25	0.87	0.56	0.06	0.89

Table 2 Performance comparison among different models (1-h)

Index	Model	COD			NH_3		
		RMSE	MAPE	R^2	RMSE	MAPE	R^2
Model 1	BITF-CL	24.40	0.21	0.91	0.35	0.05	0.93
Model 2	CNN-LSTM	32.12	0.23	0.86	0.61	0.07	0.89
Model 3	LSTM	35.41	0.27	0.81	0.67	0.10	0.86

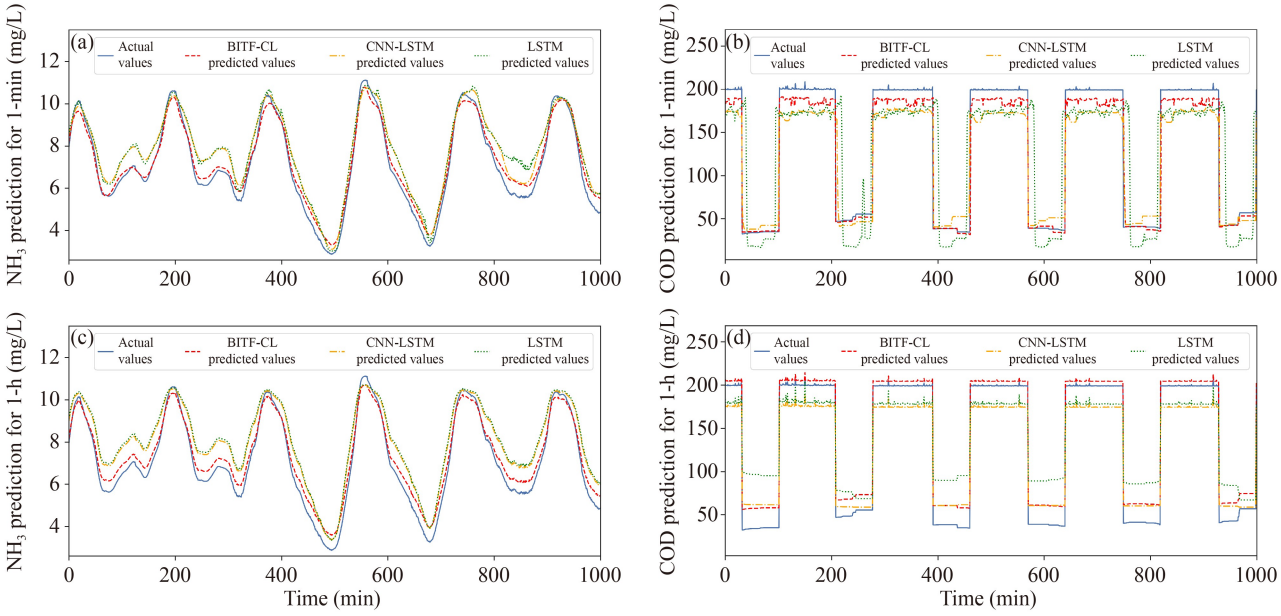


Fig. 5 Comparison of NH_3 and COD predictions and actual values for 1-min and 1-h intervals. (a) NH_3 and (b) COD prediction for 1-min; (c) NH_3 and (d) COD prediction for 1-h.

BITF-CL model uses multimodal inputs, which combine different data types such as images and time series data. This feature contributes to its consistently superior prediction performance across 1-min and 1-h intervals. Although 1-h predictions were slightly less precise than 1-min predictions, the model still produced results within an acceptable range. The lessened precision over the 1-h is attributed to the increased variability inherent to longer-duration wastewater quality data. Despite intrinsic challenges, the BITF-CL model provides accurate predictions, further validating our method's potential. Future analyses and experiments continue to use the 1-min predictions as the foundation, given their higher precision and lesser variability. This decision is driven by the increased accuracy of shorter-duration predictions, which offer a more reliable basis for further research.

The experimental results underscore the importance of the image feature extraction and fusion module in the BITF-CL model, which enhances unimodal time series data by incorporating visual elements from images, such as texture patterns, color shifts, and contour differences. These visual components represent physical attributes and chemical reactions pertinent to water quality parameters (Boztoprak et al., 2016; Tomperi et al., 2017). Previous studies have confirmed the correlation between the image particle size, shape, sludge distribution, and target water quality parameters (Khan et al., 2018; Costa et al., 2022). Conventional time series data fail to independently capture these properties and reactions. By integrating images and time series data, the BITF-CL model achieved superior COD and NH_3 level predictions compared with the LSTM and CNN-LSTM models. This enhanced performance can be attributed to the comprehensive representation of complex water quality patterns, addressing the limitations of unimodal models that rely on complex feature engineering, numerous water quality sensors, or manual variable selection.

4.2 Camera and sensor combination analysis

This section aims to ascertain the minimal sensor group required for stable prediction. To achieve this, we utilized data collected under regular conditions as outlined in Section 4.1, guaranteeing that the evaluation reflects the model's performance during the system's stable operation. This section focuses on model performance and examines the impact of various water-quality sensor and camera combinations. We performed ten replications of the test set to ensure a robust evaluation. Such repeated evaluations help capture the variability and ensure our findings' reliability based on RMSE, MAPE, and R^2 results.

We used eight setups of cameras and water quality sensors: CALL (using Camera, DO, SS, pH, EC, ORP, and temperature sensors), ALL (CALL without camera), CNS (CALL excluding SS sensor), NS (CNS excluding camera), CND (CALL excluding DO sensor), ND (CND without camera), CNDS (CALL without DO and SS sensors), and NDS (CNDS without camera). We use these abbreviations throughout the study to refer to their respective setups.

Figure 6 illustrates the predictive performance of different sensor configurations. The CALL configuration notably outperformed in predicting NH_3 and COD levels, evidenced by its average RMSE (mg/L) of 26.4, MAPE of 0.21, and R^2 of 0.457 for COD. Meanwhile, the CNS,

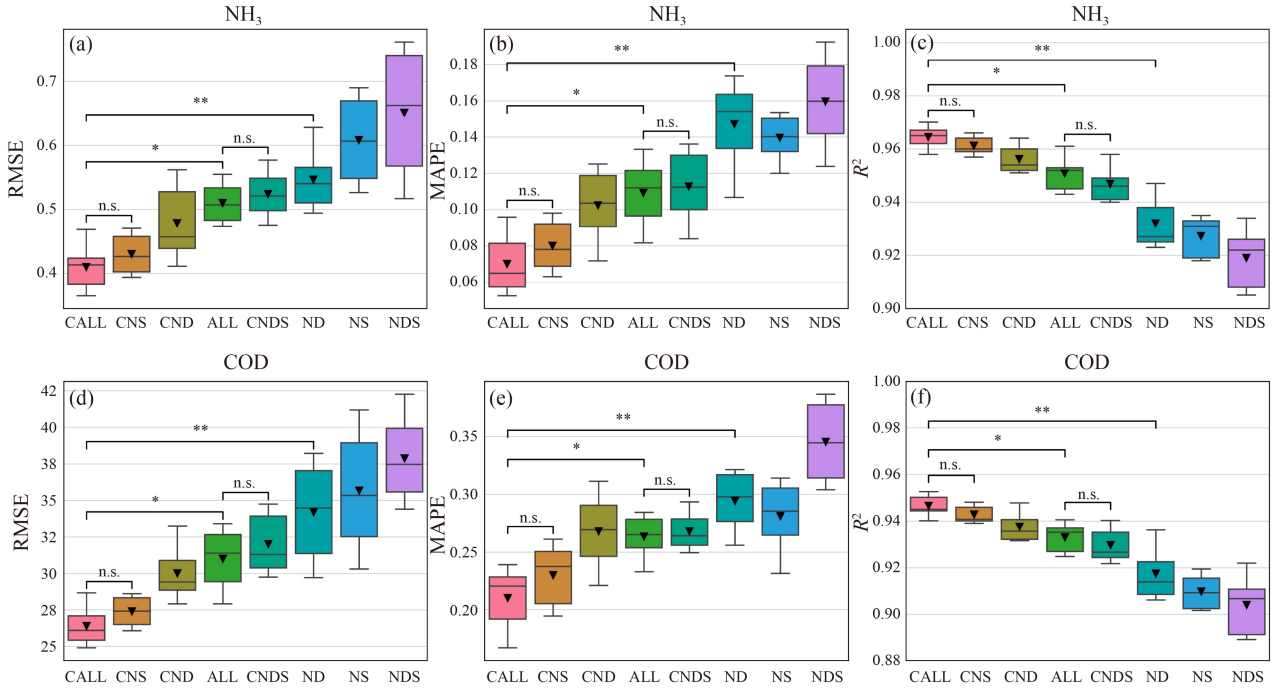


Fig. 6 Predictive performance for NH_3 and COD levels using different sensor configurations. (a) RMSE, (b) MAPE, and (c) R^2 for NH_3; (d) RMSE, (e) MAPE, and (f) R^2 for COD.

CND, ALL, and CNDS configurations also exhibited commendable outcomes with RMSE values ranging from 27.2 to 32.5 mg/L, MAPEs between 0.23 and 0.26, and R^2 values from 0.930 to 0.943. For NH_3, the CALL configuration achieved an average RMSE of 0.41 mg/L, MAPE of 0.07, and R^2 of 0.964, capturing 96.4% of the variance. These configurations produced consistent results, characterized by RMSE values between 0.43 and 0.52 mg/L, MAPEs from 0.08 to 0.11, and R^2 values from 0.946 to 0.961.

Moreover, to evaluate the significance of the differences among the configurations, we performed a one-way analysis of variance and Tukey's Honestly Significant Difference test on the RMSE, MAPE, and R^2 values of each configuration. The results are shown in Fig. 6, where n.s. indicates no significant difference, * indicates a difference of $p < 0.05$, and ** represents a more significant difference at $p < 0.01$. We found that CALL, which included all sensors, performed best in predicting COD and NH_3 levels. ALL showed a significant performance decline ($p < 0.05$), indicating the importance of image data for water quality prediction. Compared with ALL, CNDS lacked SS and DO sensors but achieved similar results after adding image features ($p > 0.05$). This demonstrates that benefiting from the multimodal feature fusion mechanism of the BITF-CL model, the camera can partially replace SS and DO sensors to achieve accurate sewage quality prediction. The finding is noteworthy because a camera is more cost-effective and easier to maintain than the aforementioned sensors for practical applications.

The NS, ND, and NDS configurations without image data significantly underperformed compared with CALL ($p < 0.01$). This underperformance suggests that the model can not capture the essential visual characteristics of water quality parameters such as chromaticity, transparency, and bubble shape. Evaluating the performance across setups, CALL exhibited exceptional performance, while CNS and CNDS achieved results comparable to ALL despite having fewer sensors. These results indicate that we can occasionally use fewer sensors without compromising predictive accuracy. Therefore, we focused on three configurations for further study: CALL for its top-tier performance, CNS for its robustness without a sensor, and CNDS for achieving comparable results to ALL with fewer sensors.

4.3 Predictive performance under varying aeration velocity

To further evaluate the robustness of the BITF-CL model under anomalous conditions, as detailed in section 2.3, we focused on three sensor configurations identified in the last section: CALL, CNS, and CNDS. We introduced a real-world disturbance by adjusting the aeration velocity of the SBR from 15 to 5 L/min. Such an intentional modification simulates deviations of crucial control parameters, reflecting common incidents like the block of aeration discs that typically lead to significant decreases in aeration velocity, impacting the effluent quality. Although this modified period persisted for 4 d, our analysis focuses on the most pivotal 1500 min marked by pronounced NH_3 and COD fluctuations.

We used this selected data as a test set to evaluate the model's performance, which was pre-trained under regular conditions. Our goal was to assess its ability to predict and adapt to unfamiliar anomalies accurately. Figure 7 shows a comparison between the values predicted by the model using the three sensor configurations and actual values after a sudden reduction in aeration. Error curves are plotted below the line graphs to intuitively reveal the differences among these configurations. The smaller the area between the error curve and *x*-axis, the more accurate its prediction.

After the reduction in aeration velocity, the DO amount decreased in the reactor. This decrease inhibited the activity of nitrifying bacteria and prevented the efficient conversion of NH_3 to nitrite and nitrate, leading to a continuous rise in NH_3 levels. As demonstrated in Figs. 7 (a) and 7(c), the CALL configuration outperformed the CNS and CNDS configurations in predicting the NH_3 level. However, all three configurations achieved close predictions of the actual values. The error plots reveal that during the initial 800 min, the model underestimated the actual data peak and overestimated valley values. The model underestimated the concentration during the subsequent 800 to 1500 min interval. This problem may be due to the model allocating more weight to track its variation trend, thereby compromising its fitting capability for local parameters. Nevertheless, the maximum error values for all configurations were lower than 1.5 mg/L. Meanwhile, CALL, CNS, and CNDS achieved excellent prediction performance with RMSEs of 0.37, 0.48, and 0.65 mg/L, respectively. These experiments demonstrated that an accurate prediction of NH_3 level can be realized using the BITF-CL model with the assistance of image data and fewer water quality sensors.

The predicted and measured results for the COD level after reducing the aeration velocity are shown in Figs. 7 (b) and 7(d). The actual value had a noise of approximately 200 mg/L owing to sensor interference by the mixed liquid of sludge and water during aeration. The COD measurement returned to normal after aeration was stopped. Reducing the aeration velocity decreased aerobic bacterial activity and organic matter transformation, thereby continuously increasing COD levels in the valley area.

The CALL configuration exhibits the best prediction performance throughout the error curve, with a maximum error of < 6.5 mg/L in the trough section. Compared with CNDS, the CNS configuration had an approximately 43% lower error in the trough section on average. Overall, CALL, CNS, and CNDS exhibited excellent performances in the valley interval of interest, with RMSEs of 25.94, 27.42, and 33.67 mg/L, respectively. However, this model is unsuitable for predicting the step peaks caused by sensor errors.

Figure 8 presents scatter plots for different sensor combinations predicting COD and NH_3 levels. The dotted line indicates perfect prediction, where the predicted value is equal to the observed value. The colors indicate the point density and data frequency across various intervals. The results revealed that the CALL, CNS, and CNDS sensor combinations closely followed the actual trends for predicting the concentrations of both pollutants. For NH_3 prediction, the respective R^2 values were 0.96, 0.94, and 0.91. For COD prediction, the overall fitting accuracy of the three sensor combinations was acceptable, except for a slight underperformance in fitting high abnormal peaks. The R^2 values for these combinations were 0.94, 0.91, and 0.90, respectively. Ultimately, while the CALL combination achieved the highest accuracy, the BITF-CL model captured the main characteristics and trends of the COD and NH_3 data when

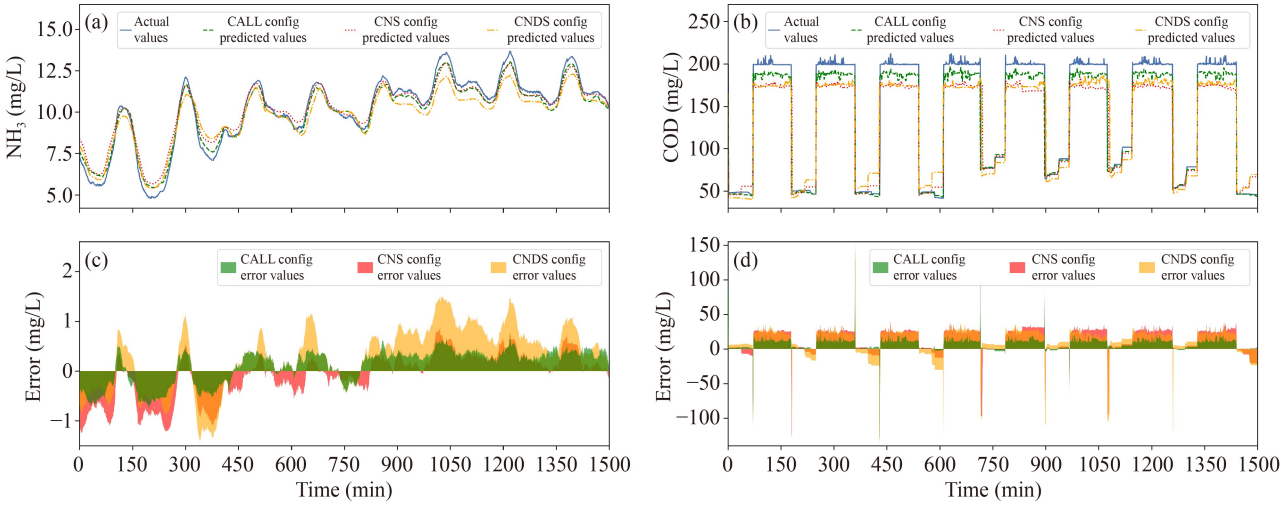


Fig. 7 Comparison of predicted and actual values of NH_3 and COD concentration after aeration velocity reduction. Predicted vs actual (a) NH_3 and (b) COD values; Error curve for (c) NH_3 and (d) COD values.

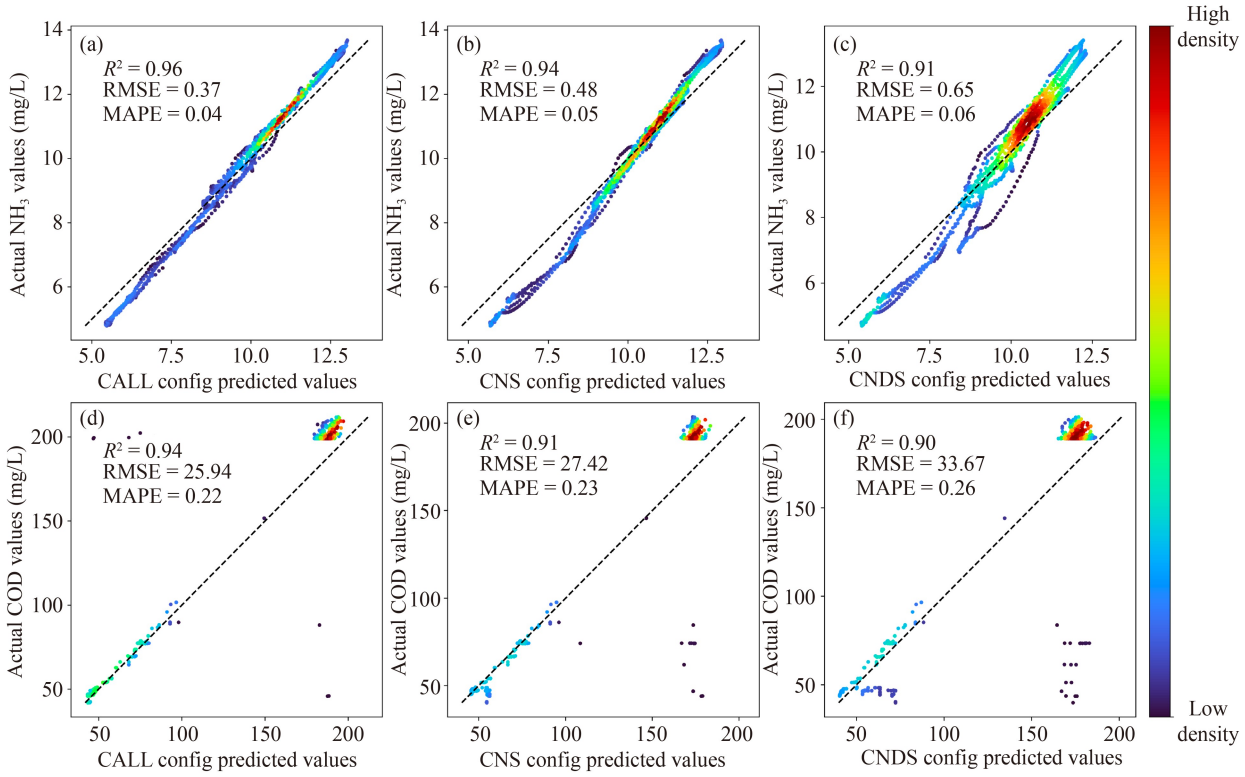


Fig. 8 Scatter plot of predicted and actual values of NH_3 and COD concentration. NH_3 prediction for (a) CALL, (b) CNS, and (c) CNDS; COD prediction for (d) CALL, (e) CNS, and (f) CNDS.

employing fewer sensors, such as the CNS and CNDS combinations. By eliminating costly DO and SS sensors, these alternative configurations still delivered comparatively excellent predictive performance at reduced expense.

4.4 Practical implications

Using the BITF-CL model, we applied a brain-inspired method to jointly analyze wastewater surface images and water quality sensor data, achieving high-precision effluent water quality predictions. When all water quality sensors and image data are accounted for, this model's predictive capability significantly surpasses traditional unimodal models, especially over forecasting intervals of 1-min and 1-h. Remarkably, even without critical sensors (DO and SS), the model's performance rivals traditional methods when image information is integrated. This multimodal fusion strategy enhances data dimensionality, ensuring sustained prediction stability even during sensor malfunctions or abnormalities. Economically speaking, the cost of obtaining image data are considerably lower than that of traditional sensors. The BITF-CL model effectively reduces the overall expenditure for WWTP monitoring systems. Additionally, it introduces tangible efficiency improvements for everyday operations. Leveraging non-contact image data as auxiliary prediction variables dramatically streamlines system maintenance and substantially reduces maintenance costs associated with traditional submerged sensors (Storey et al., 2011). This model allows wastewater treatment plants to swiftly adjust treatment processes over 1-min and 1-h forecasting scales, effectively mitigating risks.

Given the advantages outlined, the model adaptively extracts beneficial information from multiple data sources for water quality predictions. This multifaceted fusion strategy presents a novel solution for current wastewater treatment and holds promise for future applications, such as industrial wastewater analysis and lake and river quality forecasting. Of course, there remains room for optimization in the BITF-CL model, particularly in areas like image acquisition, data fusion algorithms, and computational capabilities. Nonetheless, its current iteration has showcased impressive performance and extensive application potential. Considering its cost-effective deployment, it holds immense value for both urban wastewater treatment facilities and decentralized rural wastewater treatment structures.

5 Conclusions

This study aimed to improve the prediction of COD and NH_3 levels in wastewater treatment facilities using the BITF-CL model, which mimics human cognitive abilities to integrate multimodal data, including image and time

series data. We compared the performance of our proposed model with that of conventional methods, such as LSTM and CNN-LSTM, and explored the potential for reducing the dependency on sensor data using our approach. In summary, the main findings are: 1) The BITF-CL model demonstrated superior prediction accuracy compared with LSTM and CNN-LSTM, effectively incorporating additional visual features into the unimodal time series through its unique image feature extraction and fusion mechanism. 2) This study highlighted the significance of sewage surface image data in improving water quality predictions, emphasizing the potential to reduce monitoring costs and maintenance requirements when specific DO and SS sensors are missing. 3) The BITF-CL model remained stable and accurate even when the aeration speed in the SBR suddenly changed, indicating its potential for practical application in wastewater treatment facility management. Future work should focus on refining the model, exploring more advanced feature extraction and fusion techniques, and evaluating its performance in different scenarios and applications. We aim to enhance the efficiency and cost-effectiveness of monitoring and managing water quality in wastewater treatment facilities and beyond by further optimizing the model and expanding its capabilities.

Acknowledgements This research was supported by the National Key R&D Program of China (No. 2021YFC1809001).

Conflict of Interests The author Yongzhen Peng is Editorial Board Member of *Frontiers of Environmental Science & Engineering*. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Accessibility Statement Due to the sensitive nature of the data and software copyright restrictions, the data used in this study cannot be made publicly available.

References

Al-Asheh S, Mjalli F S, Alfadala H E (2007). Forecasting influent-effluent wastewater treatment plant using time series analysis and artificial neural network techniques. Chemical Product and Process Modeling, 2(3): 55–80

Alattabi A W, Harris C, Alkhaddar R, Alzeyadi A, Abdulredha M J P E (2017). Online monitoring of a sequencing batch reactor treating domestic wastewater. Procedia Engineering, 196: 800–807

Bagheri M, Mirbagheri S A, Ehteshami M, Bagheri Z (2015). Modeling of a sequencing batch reactor treating municipal wastewater using multi-layer perceptron and radial basis function artificial neural networks. Process Safety and Environmental Protection, 93: 111–123

Barzegar R, Aalami M T, Adamowski J (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stochastic Environmental Research and Risk Assessment, 34(2): 415–433

Bekkari N, Zeddouri A (2019). Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant. Management of Environmental Quality, 30(3): 593–608

Boztoprak H, Özbay Y, Güçlü D, Küçükhemek M (2016). Prediction of sludge volume index bulking using image analysis and neural network at a full-scale activated sludge plant. Desalination and Water Treatment, 57(37): 17195–17205

Chicco D, Warrens M J, Jurman G (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Computer Science, 7: e623

Costa J G, Paulo A M S, Amorim C L, Amaral A L, Castro P M L, Ferreira E C, Mesquita D P (2022). Quantitative image analysis as a robust tool to assess effluent quality from an aerobic granular sludge system treating industrial wastewater. Chemosphere, 291(Pt 2): 132773

Fernandez de Canete J, Del Saz-Orozco P, Baratti R, Mulas M, Ruano A, Garcia-Cerezo A (2016). Soft-sensing estimation of plant effluent concentrations in a biological wastewater treatment plant using an optimal neural network. Expert Systems with Applications, 63: 8–19

Fu X, Zheng Q, Jiang G, Roy K, Huang L, Liu C, Li K, Chen H, Song X, Chen J (2023). Water quality prediction of copper-molybdenum mining-beneficiation wastewater based on the PSO-SVR model. Frontiers of Environmental Science & Engineering, 17(8): 98

Geerdink R B, Sebastiaan Van Den Hurk R, Epema O J (2017). Chemical oxygen demand: historical perspectives and future challenges. Analytica Chimica Acta, 961: 1–11

Granata F, Papirio S, Esposito G, Gargano R, De Marinis G (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. Water, 9(2): 105–117

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

Guštin S, Marinšek-Logar R (2011). Effect of pH, temperature and air flow rate on the continuous ammonia stripping of the anaerobic digestion effluent. Process Safety and Environmental Protection, 89(1): 61–66

Khan M B, Nisar H, Ng C A (2018). Image processing and analysis of phase-contrast microscopic images of activated sludge to monitor the wastewater treatment plants. IEEE Access: Practical

Innovations, Open Solutions, 6: 1778–1791

Lahat D, Adali T, Jutten C (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE, 103(9): 1449–1477

Lee J W, Suh C, Hong Y S, Shin H S (2011). Sequential modelling of a full-scale wastewater treatment plant using an artificial neural network. Bioprocess and Biosystems Engineering, 34(8): 963–973

Lee S I, Yoo S J (2020). Multimodal deep learning for finance: integrating and forecasting international stock markets. Journal of Supercomputing, 76(10): 8294–8312

Li J, Hong D, Gao L, Yao J, Zheng K, Zhang B, Chanussot J (2022a). Deep learning in multimodal remote sensing data fusion: a comprehensive review. International Journal of Applied Earth Observation and Geoinformation, 112: 102926

Li J, Liu Y, Jiang H, Yang M, Lin S, Hu Q (2022b). A multi-view image feature fusion network applied in analysis of aeration velocity for WWTP. Water, 14(3): 345–357

Litjens G, Kooi T, Ehteshami Bejnordi B, Setio A A A, Ciompi F, Ghafoorian M, van der Laak J A, van Ginneken B, Sánchez C I (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42: 60–88

Liu L, Sheng S J, Yin J T, Na L (2014). Prediction and realization of DO in sewage treatment based on machine vision and BP neural network. Telecommunication Computing Electronics and Control, 12(4): 890–896

Liu Z J, Wan J Q, Ma Y W, Wang Y (2019). Online prediction of effluent COD in the anaerobic wastewater treatment system based on PCA-LSSVM algorithm. Environmental Science and Pollution Research International, 26(13): 12828–12841

Mehonic A, Kenyon A J (2022). Brain-inspired computing needs a master plan. Nature, 604(7905): 255–260

Muhammad G, Alshehri F, Karray F, Saddik A E, Alsulaiman M, Falk T H (2021). A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Information Fusion, 76: 355–375

Mulkerrins D, Dobson A D, Colleran E (2004). Parameters affecting biological phosphate removal from wastewaters. Environment International, 30(2): 249–259

Mullins D, Coburn D, Hannon L, Jones E, Clifford E, Glavin M (2018). Using image processing for determination of settled sludge volume. Water Science and Technology, 78(2): 390–401

Nasr M S, Moustafa M A, Seif H A, El Kobrosy G (2012). Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. Alexandria Engineering Journal, 51(1): 37–43

Niu Z, Zhong G, Yu H (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452: 48–62

Pang B, Nijkamp E, Wu Y N (2020). Deep learning with TensorFlow: a review. Journal of Educational and Behavioral Statistics, 45(2): 227–248

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L (2019). Pytorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32: 8026–8037

Peng C, Li Y, Jiao L, Chen Y, Shang R (2019). Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(8): 2612–2626

Poutiainen H, Niska H, Heinonen-Tanski H, Kolehmainen M (2010). Use of sewer on-line total solids data in wastewater treatment plant modelling. Water Science and Technology, 62(4): 743–750

Rawat W, Wang Z (2017). Deep convolutional neural networks for image classification: a comprehensive review. Neural Computation, 29(9): 2352–2449

Sengupta A, Ye Y, Wang R, Liu C, Roy K (2019). Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in Neuroscience, 13: 95–105

Storey M V, Van Der Gaag B, Burns B P (2011). Advances in on-line drinking water quality monitoring and early warning systems. Water Research, 45(2): 741–747

Ta X, Wei Y (2018). Research on a dissolved oxygen prediction method for recirculating aquaculture systems based on a convolution neural network. Computers and Electronics in Agriculture, 145: 302–310

Tealab A (2018). Time series forecasting using artificial neural networks methodologies: a systematic review. Future Computing and Informatics Journal, 3(2): 334–340

Tomperi J, Koivuranta E, Leiviskä K (2017). Predicting the effluent quality of an industrial wastewater treatment plant by way of optical monitoring. Journal of Water Process Engineering, 16: 283–289

Wang K, Wen X, Hou D, Tu D, Zhu N, Huang P, Zhang G, Zhang H (2018). Application of least-squares support vector machines for quantitative evaluation of known contaminant in water distribution system using online water quality parameters. Sensors, 18(4): 938–956

Wang Y, Zhou J, Chen K, Wang Y, Liu L (2017). Water quality prediction method based on LSTM neural network. In: International Conference on Intelligent Systems and Knowledge Engineering 2017, Nanjing. Beijing: IEEE, 1–5

Wang Z, Wang Q, Wu T (2023). A novel hybrid model for water quality prediction based on VMD and IGOA optimized for LSTM. Frontiers of Environmental Science & Engineering, 17(7): 88

Wu G, Hong J, Li D, Wu Z (2019). Efficiency assessment of pollutants discharged in urban wastewater treatment: evidence from 68 key cities in China. Journal of Cleaner Production, 233: 1437–1450

Yang Y, Xiong Q, Wu C, Zou Q, Yu Y, Yi H, Gao M (2021). A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environmental Science and Pollution Research International, 28(39): 55129–55139

Yu R F, Lin C H, Chen H W, Cheng W P, Kao M C J C E J (2013). Possible control approaches of the Electro-Fenton process for textile wastewater treatment using on-line monitoring of DO and ORP. Chemical Engineering Journal, 218: 341–349

Zare Abyaneh H (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science & Engineering, 12(1): 40–48

Zhang X, Li D (2023). Multi-input multi-output temporal convolutional network for predicting the long-term water quality of ocean ranches. Environmental Science and Pollution Research, 30(3): 7914–7929

Zhu S, Han H, Guo M, Qiao J (2017). A data-derived soft-sensor method for monitoring effluent total phosphorus. Chinese Journal of Chemical Engineering, 25(12): 1791–1797

Zodrow K R, Li Q, Buono R M, Chen W, Daigger G, Duenas-Osorio L, Elimelech M, Huang X, Jiang G, Kim J H, et al. (2017). Advanced materials, technologies, and complex systems analyses: emerging opportunities to enhance urban water security. Environmental Science & Technology, 51(18): 10274–10281