Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Sustainable Urban Water Systems—Article

Multimodal Machine Learning Guides Low Carbon Aeration Strategies in Urban Wastewater Treatment

Hong-Cheng Wang [a,#], Yu-Qi Wang [a,#], Xu Wang [a], Wan-Xin Yin [a], Ting-Chao Yu [b], Chen-Hao Xue [b], Ai-Jie Wang [a,*]

[a] *State Key Laboratory of Urban Water Resource and Environment, School of Civil and Environmental Engineering, Harbin Institute of Technology Shenzhen, Shenzhen 518055, China*
[b] *Key Laboratory of Drinking Water Safety and Distribution Technology of Zhejiang Province, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China*

ARTICLE INFO

Article history:
Received 21 July 2023
Revised 10 November 2023
Accepted 28 November 2023

Keywords:
Wastewater treatment
Multimodal machine learning
Deep learning
Aeration control
Interpretable machine learning

ABSTRACT

The potential for reducing greenhouse gas (GHG) emissions and energy consumption in wastewater treatment can be realized through intelligent control, with machine learning (ML) and multimodality emerging as a promising solution. Here, we introduce an ML technique based on multimodal strategies, focusing specifically on intelligent aeration control in wastewater treatment plants (WWTPs). The generalization of the multimodal strategy is demonstrated on eight ML models. The results demonstrate that this multimodal strategy significantly enhances model indicators for ML in environmental science and the efficiency of aeration control, exhibiting exceptional performance and interpretability. Integrating random forest with visual models achieves the highest accuracy in forecasting aeration quantity in multimodal models, with a mean absolute percentage error of 4.4% and a coefficient of determination of 0.948. Practical testing in a full-scale plant reveals that the multimodal model can reduce operation costs by 19.8% compared to traditional fuzzy control methods. The potential application of these strategies in critical water science domains is discussed. To foster accessibility and promote widespread adoption, the multimodal ML models are freely available on GitHub, thereby eliminating technical barriers and encouraging the application of artificial intelligence in urban wastewater treatment.

1. Introduction

To address the global energy crisis and mitigate climate change, a collaborative approach involving multiple industries is needed to reduce greenhouse gas (GHG) emissions [1–3]. Current reports suggest that the worldwide treatment of approximately $188.1 \times 10^9\ \mathrm{m}^3$ per year of wastewater equates to approximately 1% of total energy consumption and contributes 0.77×10^9 t CO_2-equivalent GHG emissions, or approximately 1.57% of global GHG emissions (49×10^9 t CO_2-equivalent) [4–6]. Notably, the wastewater treatment industry has gained increasing recognition for its potential for carbon reduction [7,8]. This industry is deeply intertwined with societal development and human life, as it purifies wastewater and replenishes valuable water resources. Notably, GHG emissions from wastewater treatment systems play a significant role in global GHG emissions across all countries. With the ongoing surge in energy consumption, individuals are increasingly recognizing the significance of energy conservation and GHG emissions reduction [9]. Consequently, a multitude of approaches have emerged, including different energy allocation methods and intelligent low-carbon technologies. These advancements offer valuable insights into reducing carbon emissions in wastewater treatment systems [10–12].

As a relatively large emitter of GHG in the world, China critically shapes global environmental trajectories with its carbon reduction and energy conservation efforts. In 2019, China's municipal wastewater treatment industry emitted a total of 5.3 million tons of CO_2-equivalent, with the national average wastewater GHG intensity escalating by 17.2% from 2009 to 2019 [13]. China's swift industrial and economic ascension presents formidable challenges in its quest for carbon neutrality [14]. Unexpected carbon emissions have imposed significant constraints on achieving carbon neutrality objectives for wastewater treatment plants (WWTPs). However, striving for carbon neutrality in wastewater systems also uncovers novel pathways for sustainable environmental development [15,16]. Globally, nations are embracing intelligent strategies

[*] Corresponding author.
E-mail address: waj0578@hit.edu.cn (A.-J. Wang).
[#] These authors contributed equally to this work.

https://doi.org/10.1016/j.eng.2023.11.020

to simultaneously tackle water pollution and target carbon neutrality in wastewater treatment systems [17]. Consequently, achieving low-carbon operations and implementing intelligent control mechanisms have emerged as pivotal areas of focus for the efficient operation of WWTPs.

Among the different processes of WWTPs, the biochemical aeration process significantly contributes to both energy consumption and carbon emissions [18], accounting for approximately 75% of energy consumption and subsequently affecting the carbon emissions of WWTPs [19,20]. However, achieving precise aeration control remains a considerable challenge. Implementing intelligent control in this process is difficult due to the complex, slow, and disorganized dynamics of the underlying biochemical reactions [21,22]. Although activated sludge models, the most prevalent traditional mechanistic models in WWTPs, can offer relatively reliable data, their practical application is limited. This limitation arises from their dependency on complex model parameters and computational constraints [23]. The current aeration control method, which is based on biochemical mechanisms, is hindered by complex parameter settings and cannot efficiently operate wastewater plants in accordance with intelligent control strategies. Consequently, accurate prediction and real-time control of the aeration rate are crucial for WWTPs to work toward achieving carbon neutrality.

The rapid advancement of artificial intelligence (AI) has ushered in a new era of simplified and effective solutions for tackling interdisciplinary scientific challenges [24–26]. The widespread adoption and assistance of AI have led to a surge in potential applications in wastewater treatment [27,28]. These applications span various areas, including water quality testing, identification of emerging contaminants, and resource recovery [29–31]. Previous research has demonstrated the effectiveness of AI technology for predicting and controlling aeration in biochemical wastewater treatment (Table 1) [32–46].

The potential for achieving significant reductions in GHG emissions and energy consumption through intelligent control of wastewater treatment is immense. However, previous studies have often relied on single-category machine learning (ML) models, including classic ML, deep learning, and reinforcement learning, to directly predict air demand or optimize parameters in the aeration process. Unfortunately, these studies often overlook aspects of interpretability and generalizability. Moreover, as the generation of environmental data continues to escalate in speed, quantity, and complexity, a wealth of information across diverse data formats such as tables, images, and videos remains unexploited. Progress in constructing multimodal models for intelligent control of the wastewater biochemical treatment process has been scant. In addition, previous research has largely focused on integrating mechanistic models with single-category ML models, with little attention to models from diverse categories, hindering model generalization. The persistent challenge of the "black box" effect further hampers model interpretability, limiting our understanding of the underlying environmental principles. Therefore, this study pioneers an approach that deploys multimodal ML, merging classical ML models with visual models, with a specific focus on the intelligent control of the aeration process in WWTPs. The foundational steps and principles of these multimodal ML models are introduced, and the generalizability and interpretability of this multimodal approach in WWTPs are discussed. Furthermore, the performance of the multimodal models in intelligently controlling air demand is validated through application in a full-scale WWTP, affirming their effectiveness in energy conservation and consumption reduction.

2. Materials and methods

2.1. Description of the data source

The raw data of this study are obtained from a full-scale WWTP in Shandong Province, China. This plant employs a modified anaerobic–anoxic–oxic (A^2O) process for biological nutrient removal. Real-time measurements of wastewater characteristics, including ammonium (NH_4^+-N, mg·L^{-1}), nitrate (NO_3^--N, mg·L^{-1}), chemical oxygen demand (COD, mg·L^{-1}), and dissolved oxygen (DO, mg·L^{-1}), were taken using an autosampler (Integrated Quality Sensor Net system, Wissenschaftlich-Technische Werkstätten; Xylem Inc., Germany). Temperature, flow rate, COD, NH_4^+-N and total nitrogen (TN) of the influent were obtained at the different locations of the WWTP (Table S1 in Appendix A). The data were automatically collected at 15-minute intervals using the supervisory control and data acquisition (SCADA) system, resulting in a total of 8832 sets of data over a 92-day period from July 1, 2022, to September 30, 2022. Simultaneously, video files documenting the aerobic tank surface aeration were captured and saved in Mobile Pentium 4 (MP4) format. These segmented video clips were further converted into Joint Picture Group (JPG) format images and seamlessly merged with the structured data based on the corresponding time index.

The dataset, comprising 8064 samples, is used for model development through training and validation, with the remaining 768 samples employed to test model performance. The dataset utilized in this study comprises 16 features, encompassing temperature, flow rate, water quality, and other variables. The air demand is designated as the feature that needs to be controlled. As a result of the continuous motion of the data acquisition unit, the picture data originate from various aeration locations within the aerobic tank, leading to significant variations in bubble location, size, number, and shape.

Table 1
Previous research on the role of AI techniques in predicting and controlling biochemical wastewater treatment.

Model	Advantage	Reference
Reinforcement learning	Optimize dissolved oxygen injection	[32]
Multivariate adaptive regression	Optimize air injection and water quality	[33]
Artificial neural network	Energy conservation and control of risk of non-compliance	[34]
Linear stochastic	Equipment performance upgrades	[35]
Long short-term memory	Reduced energy costs	[36]
Quantile regression neural network	Stable operation and control of water quality	[37]
Coupling convolutional neural network and recurrent neural network	Efficiency and stability	[38]
Artificial neural network	Adjustment of removal rate	[39]
Artificial neural network	Reduction of energy expenditures	[40]
Ensemble learning	Simulating aerobic granular sludge	[41]
K-Nearest Neighbor	Aeration efficiency	[42]
Machine learning-mechanistic transfer models	Significant energy savings	[43]
Dynamic supervised machine-learning	Cost savings and automation	[44]
Reinforcement learning control strategies	Data management and integration processes	[45]
Random forest	Without manual calibration	[46]

2.2. Data preprocessing

The performance and characteristics of the dataset are illustrated in Table S1, including minimum, maximum, mean, standard deviation, and 25%, 50%, and 75% quantile values for all indicators. The indicators reveal subtle but distinct time series patterns. Influent parameters exhibit broad variability, with influent COD ranging from 84 to 295 mg·L^{-1} (with an average of (162.8 ± 48.9) mg·L^{-1}), TN ranging from 17.57 to 49.70 mg·L^{-1} (with an average of (33.30 ± 8.11) mg·L^{-1}), and NH$_4^+$-N ranging from 11.55 to 40.78 mg·L^{-1} (with an average of (25.68 ± 6.48) mg·L^{-1}). In contrast, effluent parameters demonstrate a uniform distribution with slight average, variance, and offset variations. Overall, noticeable fluctuations in water quality indicators are observed, posing expected challenges during model training.

To ensure data reliability and appropriateness for analysis, preprocessing and cleaning procedures are undertaken. Missing values are handled using data interpolation techniques, with the column's average value used to fill in absent data. Several filling methods are examined, including the use of values preceding or following missing positions, a fixed value of 0, or a fixed value of the mean, with the most advantageous results achieved using mean filling. After substituting the missing values and analyzing the distribution of the numerical data, outliers that could affect model accuracy are identified. These outliers are detected and managed using appropriate scaling methods. Ultimately, the data undergo a logarithmic transformation to enhance the speed of training convergence.

2.3. Base ML model

To verify the generality of the multimodal approach and characterize different ML models, it is essential to select a diverse array of ML models for this investigation. As such, eight ML models are meticulously chosen as the base models for this study (Fig. 1). This broad selection aims to ensure robust and varied testing of the adaptability across different algorithmic structures. This diverse set of models facilitates a comprehensive evaluation of their performance and unique characteristics. These eight base models are categorized into three groups.

First, three classical ML models, linear regression (LIN), Huber *k*-nearest neighbors (KNN), and support vector machines (SVMs), are incorporated. Linear regression, a widely implemented ML algorithm, aims to identify the optimal function by minimizing the sum of squared errors (Fig. 1(a)) [47]. KNN classifies a point based on the categories of its *k* nearest neighbors, subsequently performing regression for the prediction (Fig. 1(b)). SVM seeks a hyperplane that separates categories, maximizing the margin between the hyperplane and the training samples (Fig. 1(c)). A more detailed explanation of these ML models can be found in Section S1 in Appendix A.

The study also includes two ensemble learning models: gradient random forest (RF) and light gradient boosting machine (LGBM). Ensemble learning models utilize diverse strategies to integrate submodels. RF utilizes the idea of bagging, where a subtraining set for each base model is formed by randomly sampling from the original training set (Fig. 1(e)). The training process of boosting-based ensemble learning models such as LGBM follows a ladder-shaped approach. When a specific data point is misclassified in one iteration, it is assigned a higher weight (Fig. 1(f)). A more detailed explanation of ensemble learning models can be found in Section S2 in Appendix A.

Finally, three kinds of deep learning are utilized: deep neural networks (DNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM). The DNN demonstrates extensive coverage and exceptional adaptability due to its networks comprising numerous layers of substantial width (Fig. 1(d)). RNN and LSTM are deep learning methods specifically designed for handling time series data. RNN excels in short-term memory (Fig. 1(g)), while LSTM is better for long-term memory (Fig. 1(h)). A more detailed explanation of deep learning models can be found in Section S3 in Appendix A.

2.4. Visual model

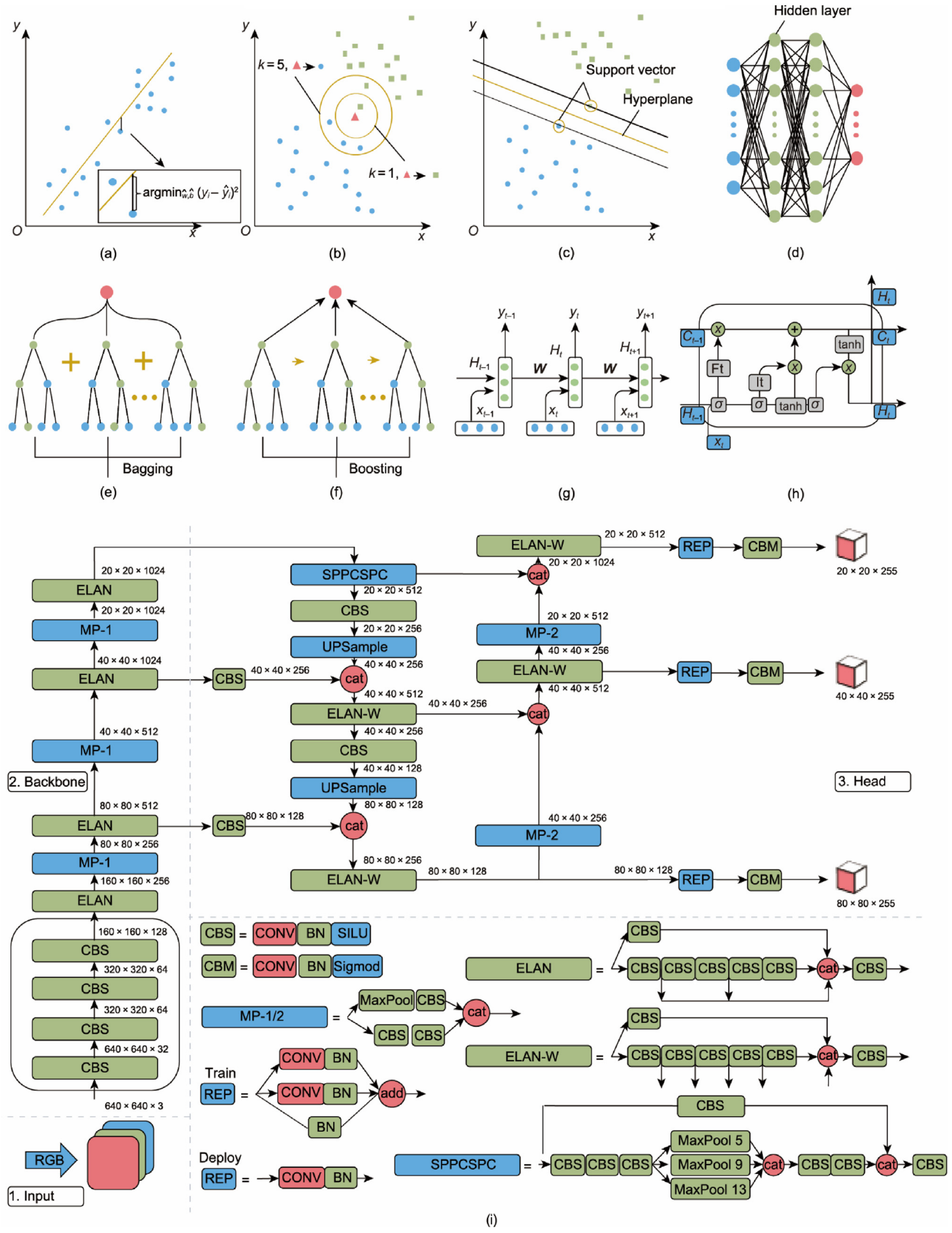
Visual models are widely used in the field of environmental science [48–50]. The emergence of deep learning, particularly convolutional neural networks (CNNs), has profoundly transformed traditional computer vision. Typically, a CNN comprises convolutional, pooling, and fully connected layers, which together form its basic components [51]. Convolutional layers distill feature maps encapsulating the multidimensional characteristics of an image, while pooling layers reduce the dimensionality of these feature maps by discarding extraneous information, thereby accelerating model convergence and minimizing the parameter count [52,53]. A prominent model based on CNNs is You Only Look Once (YOLO), which has become a leading model in computer vision due to its real-time training capabilities, fast training speed, and strong generalization feature learning ability [54,55]. The visual model framework utilized in this study extends the backbone network of YOLO, incorporating adaptations to meet the specific requirements of the wastewater treatment system.

The visual model consists of three distinct components. The input module first normalizes images of varying sizes into a 640 pixels × 640 pixels × 3 pixels format, encoding the RedGreenBlue (RGB) channels. Next, the backbone network extracts pivotal features from the input images, specifically focusing on bubble features associated with aeration. Finally, the detection head module facilitates the detection and classification of objects at three scales—small, medium, and large. This module differentiates bubbles from other solids, suspended matter, and so on. An accompanying diagram provides further insights into the individual modules (Fig. 1(i)). At its core lies the module which includes a CONV, a BN and a SILU (CBS), which consists of convolutional layers(CONV), batch normalization (BN), and the activation function (SILU). Additional modules are built or combined on the foundation provided by the CBS module. The numerical values indicating multiplication next to each module in the diagram represent the image size at each stage, illustrating the transformations from 640 pixels × 640 pixels × 3 pixels to subsequent stages.

2.5. Multimodal model

Multimodal learning integrates multiple types of data and facilitates the extraction of comprehensive features by leveraging the diverse information within each modality [56]. Data partitioning in multimodal learning involves a more fine-grained concept of modality, where different modalities can coexist within the same medium. Modalities encompass a range of elements, including ideographic symbols; semantic representations such as word vectors or knowledge graphs; structured and unstructured data units; and mathematical descriptions such as formulas, logic symbols, function diagrams, and explanatory texts.

The complementarity and redundancy of multiple modalities enable the representation and summarization of complex information. Nevertheless, the heterogeneity of multimodal data presents challenges in constructing effective representations. Tabular data directly reflect information, but image and video data are represented as signals. Single-modal representation aims to convert information into numerical vectors that can be processed by computers or abstracted into higher-level feature vectors. In contrast, multimodal representation aims to acquire more informative and discriminative feature representations by leveraging


y
x
O
argmin$_{\hat{w},\hat{b}}$ $(y_i - \hat{y}_i)^2$
(a)
k = 5,
k = 1,
(b)
Support vector
Hyperplane
(c)
Hidden layer
(d)
Bagging
(e)
Boosting
(f)
y_{t-1}
y_t
y_{t+1}
H_{t-1}
H_t
H_{t+1}
W
x_{t-1}
x_t
x_{t+1}
(g)
C_{t-1}
C_t
H_{t-1}
H_t
x_t
Ft
It
tanh
σ
(h)
1. Input
RGB
2. Backbone
3. Head
CBS
ELAN
MP-1
640 × 640 × 3
640 × 640 × 32
320 × 320 × 64
160 × 160 × 128
160 × 160 × 256
80 × 80 × 256
80 × 80 × 512
40 × 40 × 512
40 × 40 × 1024
20 × 20 × 1024
SPPCSPC
UPSample
cat
ELAN-W
MP-2
REP
CBM
20 × 20 × 512
20 × 20 × 256
40 × 40 × 256
40 × 40 × 512
40 × 40 × 128
80 × 80 × 128
20 × 20 × 255
40 × 40 × 255
80 × 80 × 255
CBS = CONV BN SILU
CBM = CONV BN Sigmod
MP-1/2 = MaxPool CBS cat CBS CBS
ELAN = CBS CBS CBS CBS CBS CBS cat CBS
ELAN-W = CBS CBS CBS CBS CBS CBS cat CBS
Train
REP = CONV BN CONV BN BN add
Deploy
REP = CONV BN
SPPCSPC = CBS CBS CBS CBS MaxPool 5 MaxPool 9 MaxPool 13 cat CBS CBS cat CBS
(i)

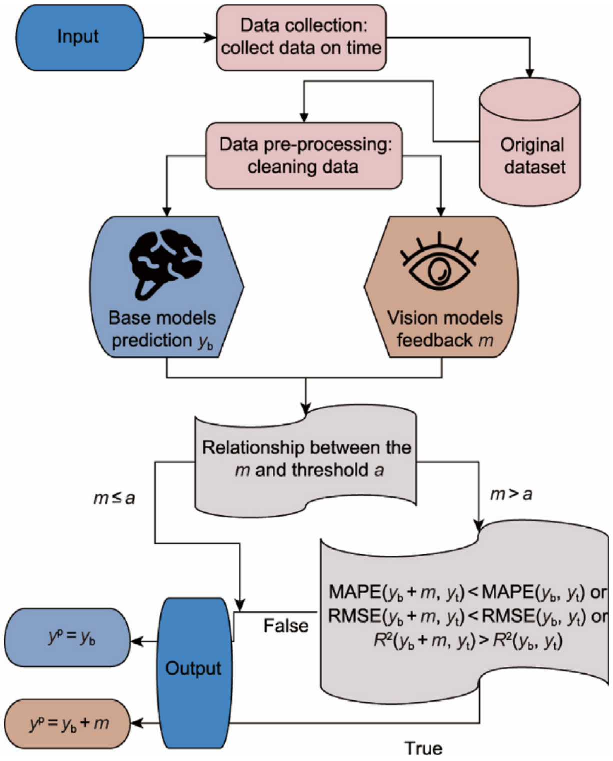

Input
Data collection: collect data on time
Original dataset
Data pre-processing: cleaning data
Base models prediction y_b
Vision models feedback m
Relationship between the m and threshold a
$m \leq a$
$m > a$
MAPE($y_b + m$, y_t) < MAPE(y_b, y_t) or RMSE($y_b + m$, y_t) < RMSE(y_b, y_t) or R^2($y_b + m$, y_t) > R^2(y_b, y_t)
False
True
Output
$y^p = y_b$
$y^p = y_b + m$

Fig. 2. Flowchart of multimodal strategy for intelligent aeration control of wastewater treatment. The data is categorized and preprocessed, and then enters under different judgments to choose whether to use the visual model feedback calibration or not, producing different prediction. MAPE: mean absolute percentage error; RMSE: relative root mean square error; y_t: true value ; y^p: predict value ; R^2: coefficient of determination.

the complementarity among different modalities and eliminating redundancy between them. The pursuit of improved feature representations in multimodal learning is motivated by the desire to effectively capture the rich and varied information encapsulated within diverse data types.

Fusion plays a crucial and challenging role in multimodal learning, where information from multiple modalities contributes to predictive tasks collaboratively. Each modality can have unique predictive capabilities, noise patterns, and the potential for data loss in at least one modality. The multimodal strategy proposed in this study employs fusion. As shown in Fig. 2, this study proposes an interrelated multimodal strategy that combines visual models with classical ML models.

Algorithm 1 shows the pseudocode for the multimodal strategy. In the data collection phase, after collecting data from multiple modalities, meticulous steps such as data cleaning, reduction and integration are performed. The processed data are simultaneously entered into the base ML model and the visual model to obtain the results. The prediction of the base ML model is y_b, and the prediction of the visual model is m, which implies the calibration parameters of feedback. This m is 0 within the threshold range, which means the calibration role of the visual model is not enabled and the results of the base ML model are directly used as the final prediction. The threshold a is designed to be a variable parameter during the code development phase. The threshold used is 91.5 $m^3 \cdot h^{-1}$, which is 5-fold the standard deviation of instantaneous oxic tank flow rate (F_{oxi}). When $m \leq a$, the predictions can be output directly without further reliance on the multimodal model. However, if $m > a$, the coupling parameters from the visual model need to be carefully evaluated. Model performance is evaluated by three indicators: relative root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) (Section S4 in Appendix A). If any metrics are optimized, the multimodal predicted value $y_m = y_b + m$ is output as the final value; otherwise, the final value is y_b.

Algorithm 1

Multimodal strategy pseudocode.

```
Input:   Base model → LIN, SVM, KNN, RF, LGBM, ANN, RNN, LSTM, and so forth
         Vision model → YOLO, SSD, DETR, FCOS, EfficientDet, Faster R-CNN, and so forth
         Multimodal environmental dataset
         Threshold $a$ → Customized or standard deviation about important feature
Output:  Predict value
         function Split (multimodal environmental dataset)
             Feature and true value $y^t$ ← Multimodal environmental dataset
             Structured-data and image ← feature
         function BaseTrain (base model, structured-data, $y^t$)
             Prediction $y_b$ ← Base model (structured feature, $y^t$)
         function VisionTrain (vision model, image, $y^t$)
             Feedback $m$ ← Vision model (image, $y^t$)
         function MulStrategy ($y_b$, $m$, $y^t$)
             if $m > a$ then
                 Evaluation $A$ ← MAPE($y_b + m$, $y^t$) < MAPE ($y_b$, $y^t$)
                 Evaluation $B$ ← RMSE($y_b + m$, $y^t$) < RMSE ($y_b$, $y^t$)
                 Evaluation $C$ ← $R^2$($y_b + m$, $y^t$) > $R^2$($y_b$, $y^t$)
                 if Evaluation $A$ or Evaluation $B$ or Evaluation $C$ then
                     Predict value ← $y_b + m$
                 else if Predict value ← $y_b$
                 end if
             if $m \leq a$ then Predict value ← $y_b$
             end if
             return Predict value
```

ANN: artificial neural network; SSD: Single Shot MultiBox Detector; DETR: DEtection TRansformer; FCOS: Fully Convolutional One-Stage Object Detection. SSD, DETR, FCOS, EfficientDet, and Faster R-CNN are some commonly used computer vision models.

Fig. 1. Schematic diagram of the principle of the base model and visual model. (a) Linear regression, x: feature; y: label; argmin: minimize; $\widehat{w}$: weights; $\widehat{b}$: bia; y_i: value; $\widehat{y}_i$: average value; (b) K-nearest neighbor, x: feature; y: label; k: number of recent values; (c) support vector machines, x, y: the axes of the Cartesian coordinate system; (d) deep artificial neural network; (e) random forest; (f) light gradient boosting machine; (g) recurrent neural network, y_{t-1}, y_t, y_{t+1}: label; H_{t-1}, H_t, H_{t+1}: hidden state; $\boldsymbol{W}$: weighting matrix; x_{t-1}, x_t, x_{t+1}: feature; (h) long short-term memory, C_{t-1}, C_t: memory cell; x, x_t: feature; It: input gate; H_{t-1}, H_t: hidden state; tanh: activation function; Ft: forget gate; σ: sigmoid function; (i) visual model framework, the visual model framework is structured into three distinct components. The input module in the lower left transforms images of varying sizes into a standardized format, encoding the RedGreenBlue (RGB) channels. The top left is the backbone module which assumes the responsibility of extracting essential features from the input images. The top right is the head module that facilitates object detection and classification across different scales. The bottom right provides specific construction for each submodule. All values are in pixel points. The numerical values denoting multiplication adjacent to each submodule within the diagram reflect the current image size. CONV: convolutional layer; BN: batch normalization; SILU: the activation function; Sigmod: sigmoid function; CBS: a CONV, a BN, and a SILU; CBM: a CONV, a BN, and a Sigmod; ELAN, MP-1, SPPCSPC, UPSample, ELAN-W, MP-2, REP, and MaxPool: the structure of these modules are shown on the diagram; cat: weight splicing; add: weight add up .

2.6. Applicability and performance tests

The applicability of the developed multimodal framework model was assessed by evaluating the performance of aeration control strategies implemented at the same full-scale WWTPs (China). The secondary treatment processes at this WWTP were arranged in parallel pairs. Half of the processes employed the original aeration method, while the other half used the multimodal aeration method for control. This operational strategy was implemented for 29 days to comprehensively assess the practicality and effectiveness of the multimodal aeration approach. Throughout the 29-day data collection period, the influent COD exhibited fluctuations ranging from 103 to 227 mg•L^{-1}, and the TN levels varied between 15 and 56 mg•L^{-1}. The dissolved oxygen levels in the five regions of the modified A^{2}O process (Fig. S1 in Appendix A) were carefully regulated within specific ranges: (0.05 ± 0.02), (0.15 ± 0.05), (1.5 ± 0.5), (0.15 ± 0.05), and (1.5 ± 0.5) mg•L^{-1}.

3. Results and discussion

3.1. Preliminary analysis of variables of investigated WWTPs

For a better understanding of the distribution of the datasets, the original indicators are normalized and demonstrated in Fig. 3. Notably, distinct variations emerge in the distribution of these different normalized indicators of the WWTPs. The original feature value is y, the normalized feature value of $\frac{y-\min(\text{feature column})}{\max(\text{feature column})-\min(\text{feature column})}$. While most feature columns tend to cluster around 0.25–0.75, the mean and mode positions among these columns differ significantly. Intriguingly, no apparent temporal correspondence between the changing trends of feature and label columns is found. This raises a challenge in deciphering the environmental significance of the data distribution, thereby underscoring the need for robust model interpretability.

The performance metrics MAPE, RMSE, and R^2 for all eight algorithms under two scenarios (base ML models and multimodal ML models) are presented in Fig. 4. Notably, the simulated values of the base ML models fail to align satisfactorily with the actual values. All the base ML models consistently exhibit a MAPE ranging from approximately 11.6% to 15.2% and an RMSE ranging from 1908 to 2242, with R^2 values consistently below 0.301. These findings indicate a significant discrepancy between the predicted and actual results using the base ML model. Among the specific base models, the DNN achieves the lowest MAPE value of 11.6%, while the RNN attains the lowest RMSE of 1908 and the highest R^2 value of 0.301. These results indicate that the base ML model has limited generalizability and interpretability for air demand prediction, thereby indicating an avenue for improvement to enhance precision in aeration control. The visualization results of the training process show that RF, RNN, and LSTM have excellent learning ability, and the prediction results are very close to the true values (Fig. S2 in Appendix A). Nevertheless, the MAPE in the test set does not show a significant difference, with values between 11.7% and 12.7% (Fig. 4). Conversely, LIN and KNN underperform in the training process, with their respective MAPE values in the test set reaching only 13.3% and 13.7%, respectively. These phenomena substantiate the conjecture that base ML models predict aeration with poor interpretability.

3.2. Multimodal ML model development and performance

The evaluation of the performance of the multimodal ML models on the test dataset is presented in Figs. 4(a)–(h). Incorporating the visual model significantly enhances the performance of the multimodal model when compared to the base ML models. All multimodal prediction curves are closer to the true situation than their corresponding base model curves. Fig. 4(i) illustrates that LIN-M, KNN-M, SVM-M, DNN-M, and RNN-M achieve MAPE values below 12% and exhibit improved R^2 values of 0.090, 0.344, 0.484, 0.468, and 0.424, respectively. Moreover, RF-M, LGBM-M, and LSTM-M demonstrate further improvements in their R^2 values, achieving 0.894, 0.737, and 0.606, respectively. The notable improvement in R^2 values provides a reliable foundation for interpreting the models.

The key feature information regarding the training of the visual model is displayed in Fig. S3 in Appendix A. In both the training and validation sets, the box loss remains stable at 0.03 and 0.06, respectively. Likewise, the objectivity loss remains stable at 0.10 in the training set and 0.02 in the validation set. The mean average precision (mAP) represents the average precision calculated for all categories across various confidence levels. mAP@0.5 denotes the average mAP with a confidence threshold greater than 0.5, steadily converging at 0.8. mAP@0.5:0.95 represents the average mAP across different confidence levels from 0.5 to 0.95, illustrating stable convergence at 0.6. Due to the multimodal nature of real-world data, multimodal learning offers a more robust theoretical foundation, albeit with associated complex training challenges

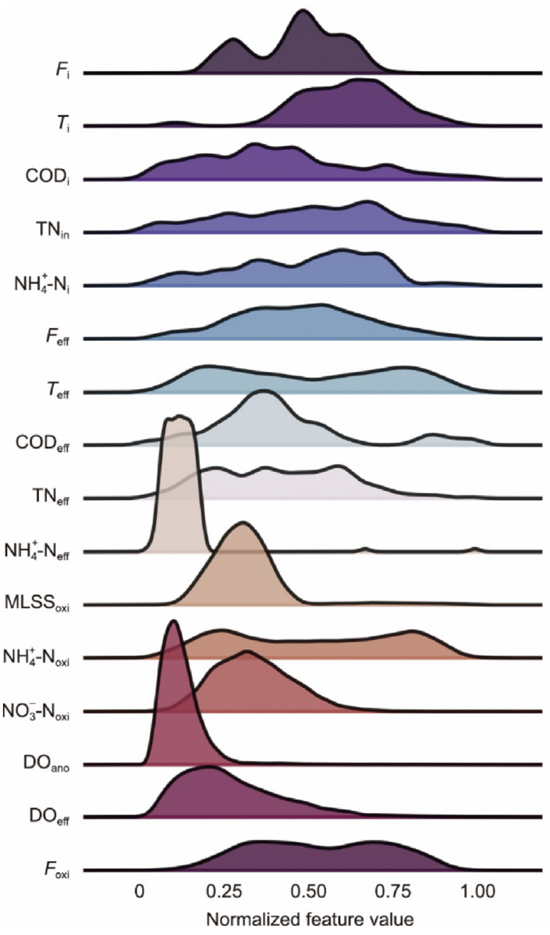


Fig. 3. Feature values of structured input parameters for the multi-modal ML in an actual WWTP. Each column of feature value is normalized. F_i: instantaneous influent flow rate; T_i: influent temperature; COD$_i$: influent chemical oxygen demand; TN$_i$: influent total nitrogen; NH$_4^+$-N$_i$: influent ammonia; F_{eff}: instantaneous effluent flow rate; T_{eff}: effluent temperature; COD$_{eff}$: effluent chemical oxygen demand ; TN$_{eff}$: effluent total nitrogen; NH$_4^+$-N$_{eff}$: effluent ammonia; MLSS$_{oxi}$: oxic tank mixed liquor suspended solids ; NH$_4^+$-N$_{oxi}$: oxic tank ammonia ; NO$_3^-$- N$_{oxi}$: oxic tank nitrate ; DO$_{ano}$: anaerobic tank dissolved oxygen ; DO$_{eff}$: effluent dissolved oxygen.

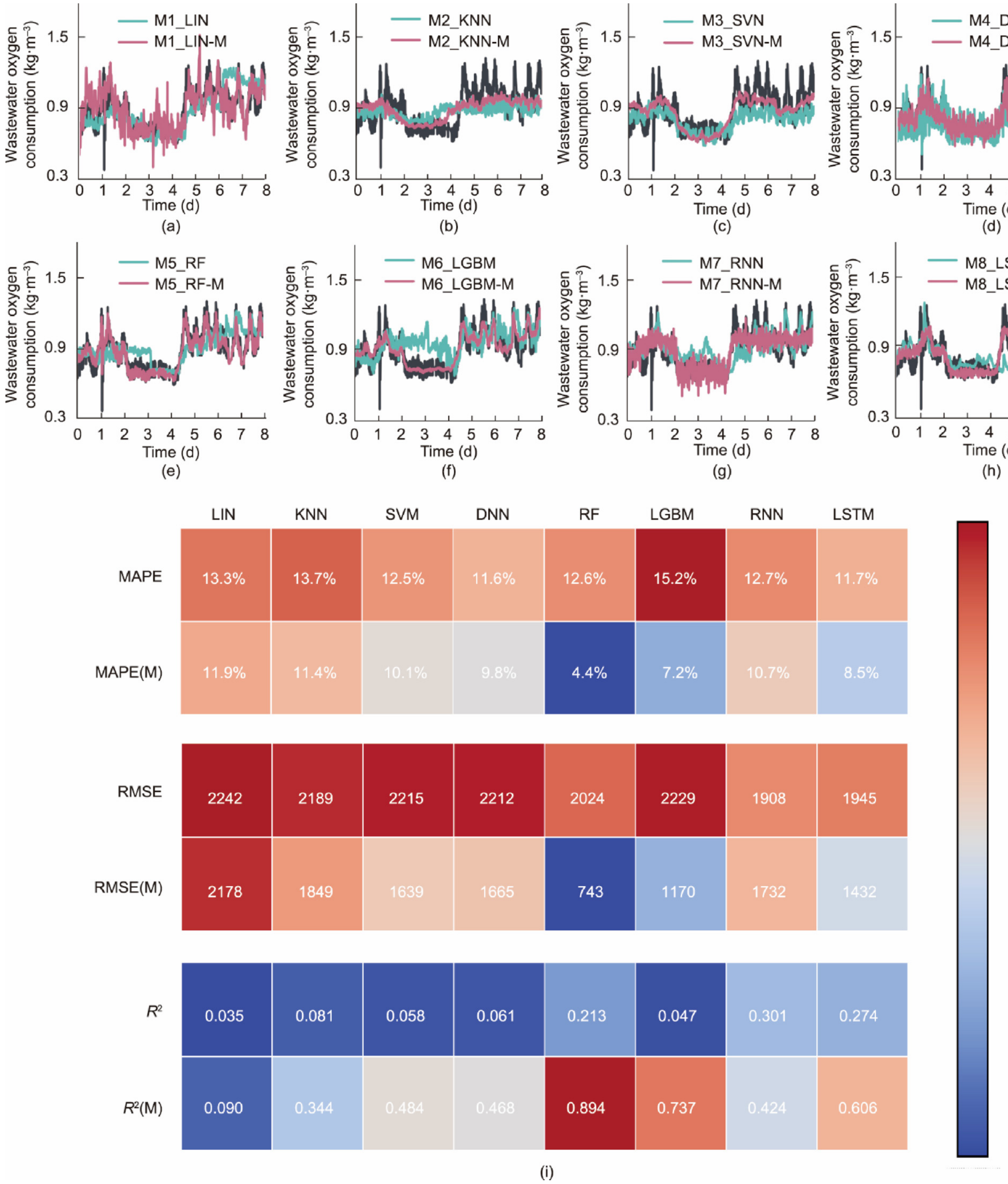


Fig. 4. (a–h) Comparison of base model and multimodal ML model verification results. Black line: the true value; green line: base ML model predict value; red line: multimodal ML model predict value. The panels a–h represent the differences between the eight base models. (i) Training performance comparison of multimodal and base ML models. The indicators for comparison are MAPE, RMSE, and R^2 from top to bottom. The color bar on the right indicates the relative size of the data, with blue to red indicating the size of the data from small to large. The smaller the MAPE and RMSE, the larger the R^2, and the stronger the performance of the model. M: multimodal.

[57,58]. Multimodal learning facilitates a comprehensive understanding of data, emphasizing the algorithms learned from multimodal data [59,60]. Humans can perceive others or objects using both visual and auditory modalities. Multimodal deep learning aspires to imbue computers with similar capabilities, enabling models to process inputs from multiple modalities concurrently [61,62]. This elucidates the robust convergence and generalization performance of multimodal models on training datasets, as illustrated in Fig. 4(i).

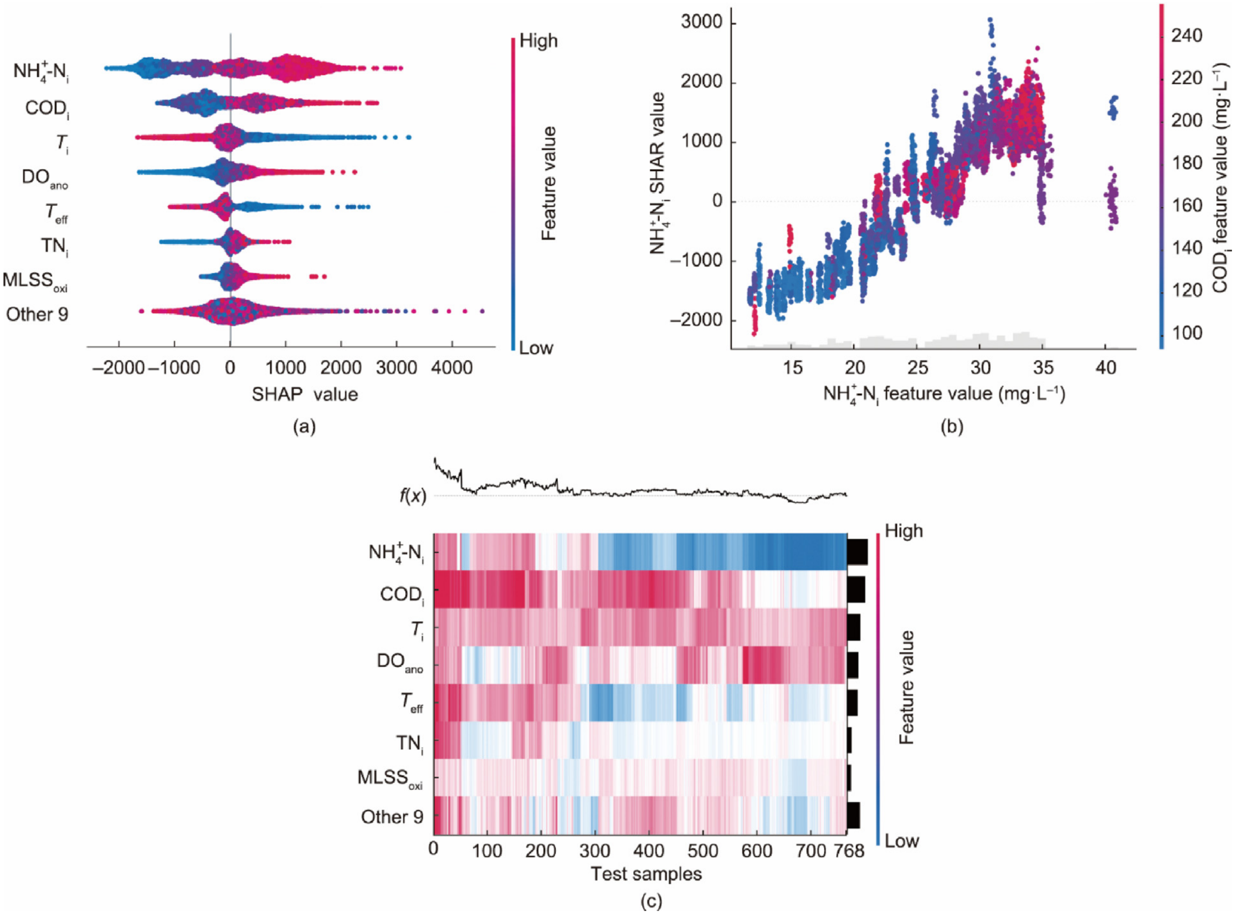
Regarding the individual algorithms, significant differences in improvement magnitude exist among the different base models. After incorporating the visual model, the improvement in RNN and LSTM is not substantial, possibly because RNN can learn from partial instantaneous information through its memory cells [63]. RNN excels in short-term memory, whereas LSTM excels in long-term memory [64,65]. This is supported by Fig. 4(i), which shows a notably higher improvement in LSTM compared to RNN. Conversely, multimodal models have a significant improvement effect on other algorithms. The two ensemble learning algorithms, RF and LGBM, perform exceptionally well, possibly due to the extraction of instantaneous information from the visual model, which provides a better path for subtree classification in tree models [66]. RF utilizes a sample retrieval method with recycling, facilitating parallel training of subtrees. In contrast, LGBM relies on altering sample weights, necessitating sequential training [67,68]. RF outperforms LGBM primarily because parallel training requires more instantaneous temporal information, as observed in Fig. 4(i). Additionally, the multimodal model significantly enhances the performance of the remaining four ML methods. RF-M exhibits remarkable predictive effects, achieving an MAPE of 4.4% and an RMSE of 743. Furthermore, R^2 reaches 0.894. RF demonstrates reliable accuracy and clear interpretability, rendering it suitable for intelligent control of wastewater plants. By visualizing the difference between y_b and y_m on RF-M, we find that there is a corrective effect of the multimodal strategy. Comparing Fig. 4(e) and Fig. S4 in Appendix A, the multimodal strategy reduces the jitter of the anomalies by visualizing the corrected values of the model, which enhances the model accuracy.

3.3. Interpretability analysis

The Shapley Additive Explanation (SHAP) method, an interpretative approach drawing inspiration from game theory, calculates the SHAP value as the average of a feature's marginal contribution across all feature permutations. Essentially, the SHAP method allocates the output value to each feature's SHAP value, thus quantifying the influence of different features on the final output value [69–71]. In RF, SHAP interprets the output of each tree and calculates the average interpretation across all trees to obtain the final interpretation result [72]. The SHAP values of all indicator features in the multimodal ML models are shown in Fig. S5 in Appendix A.

In multimodal ML models, the top three features are ammonia of the influent (NH_4^+-N_i), influent chemical oxygen demand (COD_i), and temperature of the influent (T_i) (Fig. 5(a)). The features were ordered on the vertical axis based on the aggregated SHAP values for all samples, while the horizontal axis illustrates the distribution of the impact of a single sample's SHAP value on the model output. Each point represents a sample, the sample size is stacked vertically, and the color represents the size of the feature values. The global importance of the second-ranked COD_i is 57.1% relative to that of NH_4^+-N_i, while the last-ranked feature variable has a mere importance of 1.4%. Compared to other feature indicators, NH_4^+-N_i exhibits tighter variation boundaries. Furthermore, within the multimodal framework, the coupled visual data display a larger variability, and the diverse visual picture states correlate more significantly with the subtle changes in NH_4^+-N_i. Additionally, it demonstrates that elevated levels of NH_4^+-N_i, COD_i, and T_i positively impact the system, whereas high T_i exert a negative influence. This insight substantiates that variations in influent water quality indicators substantially impact the water quality modeling of WWTPs, affirming the conjecture made during the data preprocessing stage.

The dependence scatter plot depicted in Fig. 5(b) elucidates the interplay between the two most consequential factors (NH_4^+-N_i and COD_i). Each point corresponds to a sample, with the horizontal axis



(a)

(b)

(c)

Fig. 5. Interpretability analysis of multimodal ML models by SHAP. Red corresponds to high values, and blue corresponds to low values in all figures. (a) Global interpretation of distinguishing feature values. (b) Feature dependency of NH_4^+-N_i and COD_i. (c) Supervised clustering of the validation set samples. Other 9: 9 other features not shown; $f(x)$: the sum of SHAP values.

representing the feature value of NH_4^+-N_i, the vertical axis representing the SHAP value of NH_4^+-N_i, and the color representing the feature value of COD_i. At lower concentrations of NH_4^+-N_i, the oxygen demand escalates concomitantly with the increase in NH_4^+-N_i concentration. Specifically, at NH_4^+-N_i concentrations of 15–20 mg·L^{-1}, a rise in COD_i concentrations accompanies the increase in NH_4^+-N_i concentrations. Notably, when the NH_4^+-N_i concentration reaches the 30 mg·L^{-1} threshold, the oxygen demand will level off. At this time, the COD_i concentration does not increase linearly with NH_4^+-N_i, rendering NH_4^+-N_i the determining factor. For samples with lower NH_4^+-N_i, COD_i corresponds to a lower SHAP value for NH_4^+-$N_{i,}$ and COD_i has a greater dependence and influence on NH_4^+-N_i. Crucially, the plot reveals that samples with identical feature values can exhibit disparate SHAP values, implying the presence of interactions between these features and other variables.

Supervised clustering and a heatmap are employed to visualize the underlying substructure of the 768 test samples (Fig. 5(c)). The horizontal axis is the validation set sample instance, the vertical axis is the model feature input, and the color bar is the encoded SHAP value. The gray dotted line is the baseline, and the bar chart on the right is the global importance of each model feature input. It visualizes the distribution of features within each sample. Upon vertically inspecting the arrangement of all samples, the color block of the initial sample exhibits a prominent red hue. The preceding sample distribution substantially impacts the model's effectiveness, accounting for 91.8%. Furthermore, the sum of SHAP values ($f(x)$) surpasses the mean line in 95.8% of cases, indicating their classification as high-quality samples. This finding demonstrates the relative stability of the multimodal ML models during the prediction stage, successfully avoiding significant overfitting. Consequently, this attribute contributes to the superior performance of the multimodal ML models.

3.4. Feasibility and practical implications of multimodal learning

Upon validation of the 29 days of data obtained from the full-scale WWTPs, the multimodal model developed in this study demonstrates superior performance compared to traditional fuzzy aeration prediction and control. The utilization of multimodal effluent leads to a substantial decrease in effluent COD and TN compared to the fuzzy control. Fig. 6(a) shows that the average effluent COD decreases by 38.5%, and Fig. 6(b) demonstrates a 26.3% decrease in average effluent TN. The removal rates of COD and TN indicate the model's conversion ability, and the multimodal framework notably improves these rates. Fig. 6(c) illustrates the maximal COD removal rate of 96.3%, whereas Fig. 6(d) exhibits the peak TN removal rate of 95.6%. Moreover, the implementation of the multimodal framework results in a remarkable reduction in carbon emissions within the WWTPs.

The notable enhancement in crucial indicators can be attributed to alterations in dissolved oxygen levels in the effluent. Fig. 6(e) demonstrates an elevation in dissolved oxygen, signifying an enhanced efficiency of oxygen utilization in the initial stage. The

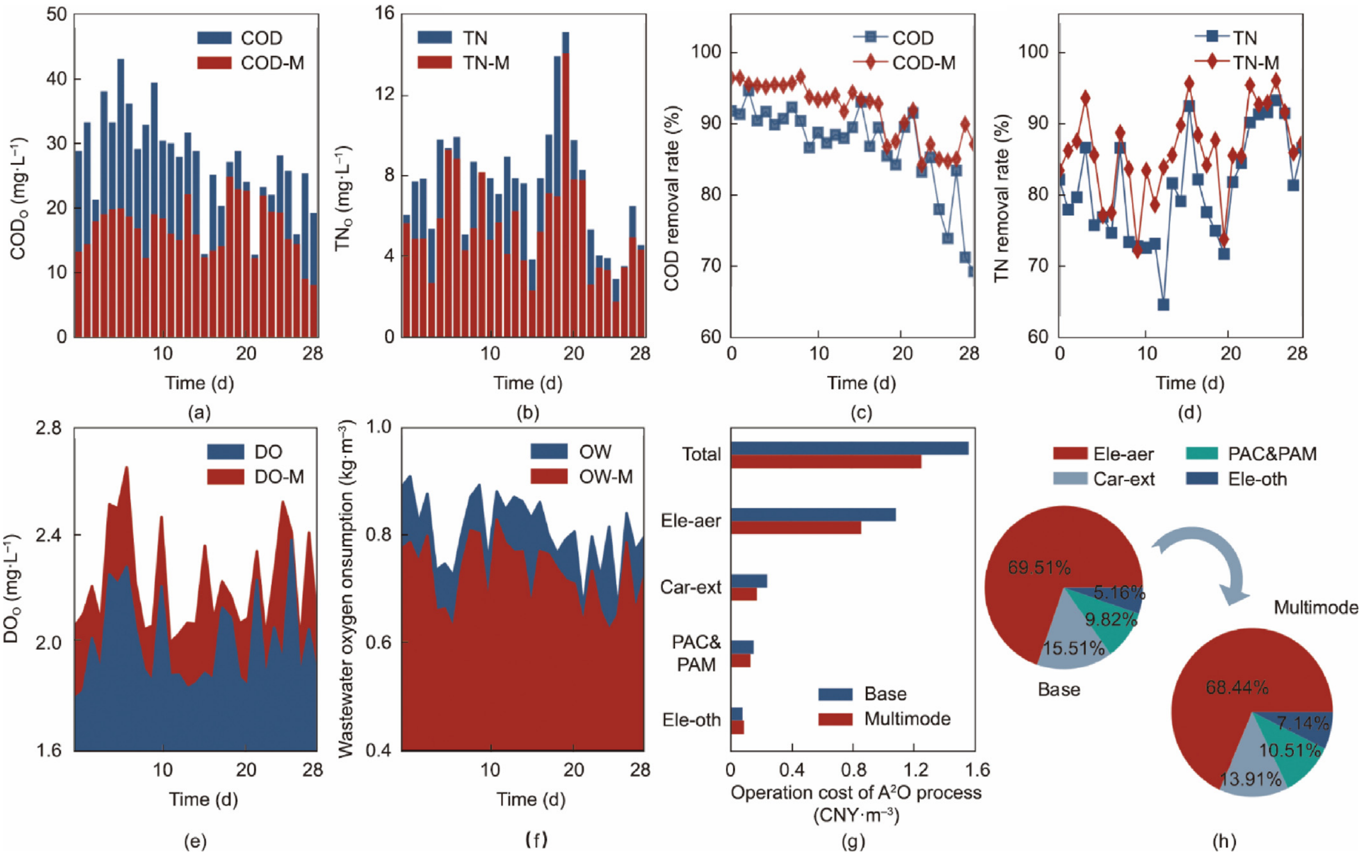


Fig. 6. Application example of the multimodal framework in a full-scale WWTP. The legend of the original method is shown without string-M, while the legend of using the multimodal framework is shown with string-M. (a,b) Evaluate the water quality and stability by the effluent of COD and TN. (c,d) Evaluate conversion ability capability by COD and TN removal rate. (e) Dynamic changes in dissolved oxygen concentration in effluent. (f) The amount of oxygen required to treat a unit volume of wastewater in an aerobic tank. (g) Total and unit cost of the WWTP when using traditional fuzzy control and multimodal strategy. (h) The percentage cost of the WWTP when using traditional fuzzy control and multimodal strategy. Ele-aer: electricity of aeration; Car-ext: methanol for carbon source; PAC&PAM: polyacrylamide and polyaluminum chloride; Ele-oth: other electricity consumption; COD_o: COD of outflow; DO_o: DO of outflow; OW: oxygen demand per unit of wastewater.

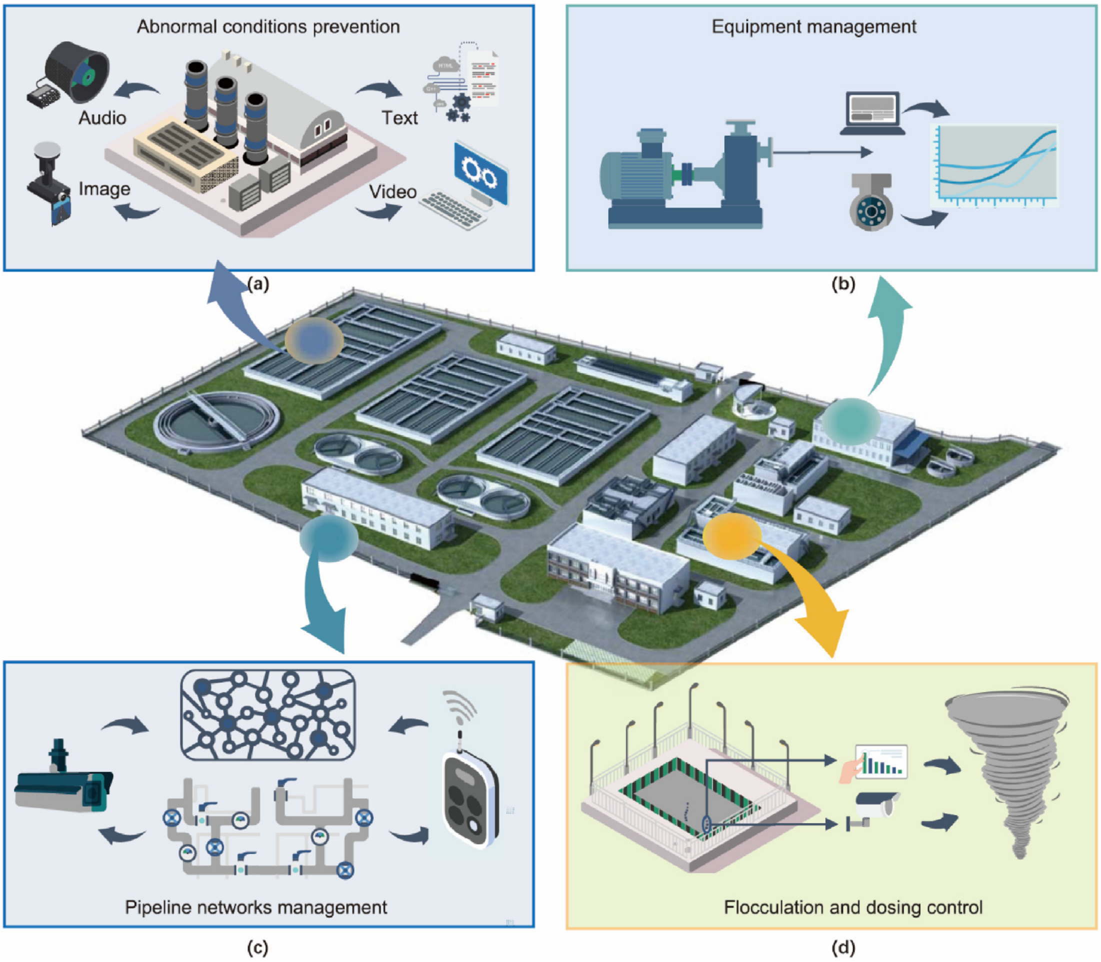

Abnormal conditions prevention
Audio
Text
Image
Video
Equipment management
(a)
(b)
Pipeline networks management
Flocculation and dosing control
(c)
(d)

Fig. 7. The application feasibility and prospect of the multimodal method are provided in WWTPs. (a) Preventing abnormal working conditions. (b) Dispatching of pump stations for intelligent sludge discharge in equipment. (c) Real-time detection of pipeline networks. (d) Identification of flocculation effect and intelligent dosing.

minimum dissolved oxygen level in the effluent increases from 1.79 to 1.98 mg·L^{-1}, while the maximum value increases from 2.38 to 2.65 mg·L^{-1}. The average dissolved oxygen in the effluent increases by 11.2%, while the maximum increase is 31.3%, equating to 0.59 mg·L^{-1}. Supporting evidence is provided in Fig. 6(f), illustrating a reduction in the necessary oxygen content for treating unit wastewater in aerobic tanks. By implementing the multimodal aeration framework, intelligent control over oxygen supply and utilization is achieved, thereby optimizing the provision of substrates for carbon and nitrogen removal processes in wastewater biological treatment. As a result of this optimization, stable effluent quality and enhanced removal rates are achieved.

The operational expenses of industrial park WWTPs principally encompass electricity of aeration (Ele-aer), chemical costs, and other electricity consumption (Ele-oth), such as sludge return pumps, internal return pumps, and sludge discharge pumps. The primary chemicals utilized comprise methanol for carbon source (Car-ext) supplementation and polyacrylamide and polyaluminum chloride (PAC&PAM) for chemical precipitation. Table S2 in Appendix A delineates the primary cost components of WWTPs and their corresponding unit costs. Overall, for half of the process employing the traditional fuzzy control strategy, the treatment cost measures 1.57 CNY for every cubic meter of wastewater. In contrast, for half of the process units utilizing a multimodal ML control strategy, the processing cost is significantly curtailed by 19.7% to 1.26 CNY for every cubic meter of wastewater, as shown in Fig. 6(g). Specifically, a significant reduction in aeration electricity consumption is achieved, with the multimodal ML model demonstrating a substantial decrease of 21.1% compared to base ML model operating conditions. The percentage of wastewater treatment costs also changes (Fig. 6(h)).

3.5. Perspective

Within the domain of intelligent wastewater treatment control, the utilization of multimodal learning frameworks extends beyond aeration, revealing significant environmental implications that remain unexplored (Fig. 7). The potential applications of multimodal methods in industrial WWTPs and broader water systems are immense. Multimodal ML incorporates diverse forms of environmental data, including text, audio, images, and videos, enabling comprehensive analysis. Importantly, multimodal frameworks exhibit significant potential and practical feasibility within critical water science domains.

The efficacy of multimodal frameworks' optimization solutions in effectively tackling the identification of abnormal operational conditions is demonstrated in Fig. 7(a). WWTPs and urban water systems feature complex parallel configurations of pipeline networks. The detection of anomalies in pipeline networks can be further improved through the optimization offered by multimodal frameworks (Fig. 7(c)). Similarly, in scenarios that encompass sludge treatment, water supply distribution, and pump utilization, real-time images, and operational data can be utilized to intelligently schedule pump stations (Fig. 7(b)). Furthermore, in advanced wastewater treatment and water supply treatment processes, intelligent chemical dosing can be accomplished by integrating preset data parameters and employing morphological recognition of alum flowers (Fig. 7(d)).

Several ideas are proposed for future optimization purposes. First, it is necessary to validate the practical application of the multimodal approach in other areas of industrial water systems. Second, encapsulating these methods into callable third-party software libraries will simplify the complexity related to multimodal modeling. Finally, a number of variables, including hydraulic retention time (HRT), pH, nitrate cycling ratio, and carbon to nitrogen ratio (C/N), were eliminated during the pretraining phase. Some of these variables could be addressed by other indicators, and some had insignificant effects on the model. At the same time, some indicators should be verified by researchers in the future, such as influent and effluent phosphorus and microbial fractions. However, it is crucial to acknowledge that data acquisition and reliability verification pose significant challenges that can be overcome through federated learning approaches.

4. Conclusion

This study introduces a novel ML modeling approach that utilizes multimodal learning, applied for the first time in the aeration process of wastewater treatment. By leveraging the multimodal framework, the performance and interpretability of eight ML models are significantly enhanced, leading to a notable reduction in operational costs and carbon emissions within WWTPs. Among the diverse categories of models explored, RF-M emerges as the top-performing model. Moreover, this study thoroughly examines the feasibility and potential of multimodal methods in tackling diverse challenges within the field of water science. The primary source code for replication purposes can be accessed via Section S5 in the Appendix A.

Acknowledgment

We gratefully acknowledge the financial support by the National Natural Science Foundation of China (52230004 and 52293445), the Key Research and Development Project of Shandong Province (2020CXGC011202-005), and the Shenzhen Science and Technology Program (KCXFZ20211020163404007, Grant KQTD20190929172630447).

Compliance with ethics guidelines

Hong-Cheng Wang, Yu-Qi Wang, Xu Wang, Wan-Xin Yin, Ting-Chao Yu, Chen-Hao Xue, and Ai-Jie Wang declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2023.11.020.

References

[1] McNicol G, Jeliazovski J, François JJ, Kramer S, Ryals R. Climate change mitigation potential in sanitation via off-site composting of human waste. Nat Clim Change 2020;10(6):545–9.

[2] Nerini FF, Sovacool B, Hughes N, Cozzi L, Cosgrave E, Howells M, et al. Connecting climate action with other sustainable development goals. Nat Sustainability 2019;2(8):674–80.

[3] Zhou N, Khanna N, Feng W, Ke J, Levine M. Scenarios of energy efficiency and CO_2 emissions reduction potential in the buildings sector in China to year 2050. Nat Energy 2018;3(11):978–84.

[4] Sabia G, Petta L, Avolio F, Caporossi E. Energy saving in wastewater treatment plants: a methodology based on common key performance indicators for the evaluation of plant energy performance, classification and benchmarking. Energy Convers Manage 2020;220:113067.

[5] IPCC. Climate change 2014: mitigation of climate change. Cambridge: Cambridge University Press; 2014.

[6] Jones ER, van Vliet MTH, Qadir M, Bierkens MFP. Country-level and gridded estimates of wastewater production, collection, treatment and reuse. Earth Syst Sci Data 2021;13(2):237–54.

[7] Van Loosdrecht MCM, Brdjanovic D. Water treatment. Anticipating the next century of wastewater treatment. Science 2014;344(6191):1452–3.

[8] Shao L, Chen GQ. Water footprint assessment for wastewater treatment: method, indicator, and application. Environ Sci Technol 2013;47(14):7787–94.

[9] Rothausen SGSA, Conway D. Greenhouse-gas emissions from energy use in the water sector. Nat Clim Change 2011;1(4):210–9.

[10] Shindell D, Smith CJ. Climate and air-quality benefits of a realistic phase-out of fossil fuels. Nature 2019;573(7774):408–11.

[11] Nguyen TKL, Ngo HH, Guo W, Chang SW, Nguyen DD, Nghiem LD, et al. Insight into greenhouse gases emissions from the two popular treatment technologies in municipal wastewater treatment processes. Sci Total Environ 2019;671:1302–13.

[12] Wang YQ, Wang HC, Song YP, Zhou SQ, Li QN, Liang B, et al. Machine learning framework for intelligent aeration control in wastewater treatment plants: automatic feature engineering based on variation sliding layer. Water Res 2023;246:120676.

[13] Du WJ, Lu JY, Hu YR, Xiao J, Yang C, Wu J, et al. Spatiotemporal pattern of greenhouse gas emissions in China's wastewater sector and pathways towards carbon neutrality. Nat Water 2023;1(2):166–75.

[14] Wang H, Lu X, Deng Y, Sun Y, Nielsen CP, Liu Y, et al. China's CO_2 peak before 2030 implied from characteristics and growth of cities. Nat Sustainability 2019;2(8):748–54.

[15] Ramaswami A, Tong K, Fang A, Lal RM, Nagpure AS, Li Y, et al. Urban cross-sector actions for carbon mitigation with local health co-benefits in China. Nat Clim Change 2017;7(10):736–42.

[16] Liang X, Zhang S, Wu Y, Xing J, He X, Zhang KM, et al. Air quality and health benefits from fleet electrification in China. Nat Sustainability 2019;2(10):962–71.

[17] Hao X, Liu R, Huang X. Evaluation of the potential for operating carbon neutral WWTPs in China. Water Res 2015;87:424–31.

[18] Skouteris G, Rodriguez-Garcia G, Reinecke SF, Hampel U. The use of pure oxygen for aeration in aerobic wastewater treatment: a review of its potential and limitations. Bioresour Technol 2020;312:123595.

[19] Pittoors E, Guo Y, van Hulle SWH. Modeling dissolved oxygen concentration for optimizing aeration systems and reducing oxygen consumption in activated sludge processes: a review. Chem Eng Commun 2014;201(8):983–1002.

[20] Åmand L, Olsson G, Carlsson B. Aeration control—a review. Water Sci Technol 2013;67(11):2374–98.

[21] Crini G, Lichtfouse E. Advantages and disadvantages of techniques used for wastewater treatment. Environ Chem Lett 2019;17(1):145–55.

[22] Sun Y, Guan Y, Pan M, Zhan X, Hu Z, Wu G. Enhanced biological nitrogen removal and N_2O emission characteristics of the intermittent aeration activated sludge process. Rev Environ Sci Bio Technol 2017;16(4):761–80.

[23] Fenu A, Guglielmi G, Jimenez J, Spèrandio M, Saroj D, Lesjean B, et al. Activated sludge model (ASM) based modelling of membrane bioreactor (MBR) processes: a critical review with special regard to MBR specificities. Water Res 2010;44(15):4272–94.

[24] Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. Nat Commun 2020;11(1):4428.

[25] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555(7698):604–10.

[26] Jones DT, Thornton JM. The impact of AlphaFold2 one year on. Nat Methods 2022;19(1):15–20.

[27] Eggimann S, Mutzner L, Wani O, Schneider MY, Spuhler D, de Vitry MM, et al. The potential of knowing more: a review of data-driven urban water management. Environ Sci Technol 2017;51(5):2538–53.

[28] Newhart KB, Holloway RW, Hering AS, Cath TY. Data-driven performance analyses of wastewater treatment plants: a review. Water Res 2019;157:498–513.

[29] Rodriguez-Perez J, Leigh C, Liquet B, Kermorvant C, Peterson E, Sous D, et al. Detecting technical anomalies in high-frequency water-quality data using artificial neural networks. Environ Sci Technol 2020;54(21):13719–30.

[30] Miller TH, Gallidabino MD, MacRae JI, Hogstrand C, Bury NR, Barron LP, et al. Machine learning for environmental toxicology: a call for integration and innovation. Environ Sci Technol 2018;52(22):12953–5.

[31] Garrido-Baserba M, Vinardell S, Molinos-Senante M, Rosso D, Poch M. The economics of wastewater treatment decentralization: a techno-economic evaluation. Environ Sci Technol 2018;52(15):8965–76.

[32] Hernández-del-Olmo F, Gaudioso E, Dormido R, Duro N. Energy and environmental efficiency for the N-ammonia removal process in wastewater treatment plants by means of reinforcement learning. Energies 2016;9(9):755.

[33] Asadi A, Verma A, Yang K, Mejabi B. Wastewater treatment aeration process optimization: a data mining approach. J Environ Manage 2017;203(Pt 2):630–9.

[34] Zhu JJ, Kang L, Anderson PR. Predicting influent biochemical oxygen demand: balancing energy demand and risk management. Water Res 2018;128:304–13.

[35] Lotfi K, Bonakdari H, Ebtehaj I, Mjalli FS, Zeynoddin M, Delatolla R, et al. Predicting wastewater treatment plant quality parameters using a novel hybrid linear–nonlinear methodology. J Environ Manage 2019;240:463–74.

[36] Wang J, Wan K, Gao X, Cheng X, Shen Y, Wen Z, et al. Energy and materials-saving management via deep learning for wastewater treatment plants. IEEE. Access 2020;8:191694–705.

[37] Icke O, van Es DM, de Koning MF, Wuister JJG, Ng J, Phua KM, et al. Performance improvement of wastewater treatment processes by application of machine learning. Water Sci Technol 2020;82(12):2671–80.

[38] Guo Z, Du B, Wang J, Shen Y, Li Q, Feng D, et al. Data-driven prediction and control of wastewater treatment process through the combination of convolutional neural network and recurrent neural network. RSC Adv 2020;10(23):13410–9.

[39] Khatri N, Khatri KK, Sharma A. Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant. J Water Process Eng 2020;37:101477.

[40] Newhart KB, Marks CA, Rauch-Williams T, Cath TY, Hering AS. Hybrid statistical-machine learning ammonia forecasting in continuous activated sludge treatment for improved process control. J Water Process Eng 2020;37:101389.

[41] Zaghloul MS, Iorhemen OT, Hamza RA, Tay JH, Achari G. Development of an ensemble of machine learning algorithms to model aerobic granular sludge reactors. Water Res 2021;189:116657.

[42] Sangeeta AS, Sharafati A, Sihag P, Al-Ansari N, Chau KW. Machine learning model development for predicting aeration efficiency through Parshall flume. Eng Appl Comput Fluid Mech 2021;15(1):889–901.

[43] Pan Y, Dagnew M. A new approach to estimating oxygen off-gas fraction and dynamic alpha factor in aeration systems using hybrid machine learning and mechanistic models. J Water Process Eng 2022;48:102924.

[44] Qambar AS, Al Khalidy MM. Optimizing dissolved oxygen requirement and energy consumption in wastewater treatment plant aeration tanks using machine learning. J Water Process Eng 2022;50:103237.

[45] Croll HC, Ikuma K, Ong SK, Sarkar S. Reinforcement learning applied to wastewater treatment process control optimization: approaches, challenges, and path forward. Crit Rev Environ Sci Technol 2023;53(20):1775–94.

[46] Schwarz M, Trippel J, Engelhart M, Wagner M. Dynamic alpha factor prediction with operating data—a machine learning approach to model oxygen transfer dynamics in activated sludge. Water Res 2023;231:119650.

[47] Visser H, Evers N, Bontsema A, Rost J, de Niet A, Vethman P, et al. What drives the ecological quality of surface waters? A review of 11 predictive modeling tools. Water Res 2022;208:117851.

[48] Jia T, Kapelan Z, de Vries R, Vriend P, Peereboom EC, Okkerman I, et al. Deep learning for detecting macroplastic litter in water bodies: a review. Water Res 2023;231:119632.

[49] Gnann N, Baschek B, Ternes TA. Close-range remote sensing-based detection and identification of macroplastics on water assisted by artificial intelligence: a review. Water Res 2022;222:118902.

[50] Zhou X, Tang Z, Xu W, Meng F, Chu X, Xin K, et al. Deep learning identifies accurate burst locations in water distribution networks. Water Res 2019;166:115058.

[51] Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018;2018:7068349.

[52] Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Networks Learn Syst 2022;33(12):6999–7019.

[53] Mallat S. Understanding deep convolutional networks. Philos Trans R Soc A 2016;374(2065):20150203.

[54] Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022. arXiv:2207.02696.

[55] Chen C, Zheng Z, Xu T, Guo S, Feng S, Yao W, et al. YOLO-based UAV technology: a review of the research and its applications. Drones 2023;7(3):190.

[56] Holler J, Levinson SC. Multimodal language processing in human communication. Trends Cognit Sci 2019;23(8):639–52.

[57] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Briefings Bioinf 2022;23(2):bbab569.

[58] Li M, Zareian A, Zeng Q, Whitehead S, Lu D, Ji H, et al. Cross-media structured common space for multimedia event extraction. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Stroudsburg, PA, USA. Kerrville: Association for Computational Linguistics; 2020. p. 2557–68.

[59] Gao J, Li P, Chen Z, Zhang J. A survey on deep learning for multimodal data fusion. Neural Comput 2020;32(5):829–64.

[60] Zhang Y, Chen M, Shen J, Wang C. Tailor versatile multi-modal learning for multi-label emotion recognition. 2022. arXiv:2201.05834.

[61] Azam MA, Khan KB, Salahuddin S, Rehman E, Khan SA, Khan MA, et al. A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Comput Biol Med 2022;144:105253.

[62] Huo Y, Zhang M, Liu G, Lu H, Gao Y, Yang G, et al. WenLan: bridging vision and language by large-scale multi-modal pre-training. 2021. arXiv:2103.06561.

[63] Yang B, Xiao Z, Meng Q, Yuan Y, Wang W, Wang H, et al. Deep learning-based prediction of effluent quality of a constructed wetland. Environ Sci Ecotechnol 2022;13:100207.

[64] Zhong S, Zhang K, Bagheri M, Burken JG, Gu A, Li B, et al. Machine learning: new ideas and tools in environmental science and engineering. Environ Sci Technol 2021;55(19):12741–54.

[65] Gupta S, Aga D, Pruden A, Zhang L, Vikesland P. Data analytics for environmental science and engineering research. Environ Sci Technol 2021;55(16):10895–907.

[66] Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res 2020;171:115454.

[67] Wang G, Jia QS, Zhou M, Bi J, Qiao J, Abusorrah A. Artificial neural networks for water quality soft-sensing in wastewater treatment: a review. Artif Intell Rev 2022;55(1):565–87.

[68] Zhu S, Lu H, Ptak M, Dai J, Ji Q. Lake water-level fluctuation forecasting using machine learning models: a systematic review. Environ Sci Pollut Res 2020;27(36):44807–19.

[69] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook: Curran Associates Inc.; 2017. p. 4768–77.

[70] Meng Y, Yang N, Qian Z, Zhang G. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. J Theor Appl Electron Commer Res 2021;16(3):466–90.

[71] Zhang J, Ma X, Zhang J, Sun D, Zhou X, Mi C, et al. Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model. J Environ Manage 2023;332:117357.

[72] Futagami K, Fukazawa Y, Kapoor N, Kito T. Pairwise acquisition prediction with SHAP value interpretation. J Finance Data Sci 2021;7:22–44.