World Journal of Agriculture and Forestry Sciences
ISSN: 2960-0227
DOI: https://doi.org/10.61784/wjafs3019

AN INTELLIGENT ANALYSIS MODEL OF FACTORS INFLUENCING INDUSTRIAL WASTEWATER TREATMENT EFFICIENCY BY INCORPORATING XGBOOST

Yang Xu
Gongshu District Water and Air Pollution Control Office, Hangzhou 310015, Zhejiang, China.
Corresponding Email: xuy21102@gmail.com

Abstract: In order to improve the accuracy and system understanding of industrial wastewater treatment efficiency prediction, an intelligent analysis model integrating XGBoost is constructed with a typical A²/O process wastewater treatment system as an example. 14 high-frequency operational variables are collected and processed in the system, and a multi-dimensional input system containing ratio features and time-difference features is designed. Combining the PCA dimensionality reduction and the temporal sliding window mechanism, the model effectively compresses the redundant information and enhances the expression ability of the dynamic features. The model stability and generalization ability are improved by the joint tuning strategy of grid search and Bayesian optimization. Comparison of the SVR, RF and MLP models shows that XGBoost is better in terms of prediction accuracy, robustness and feature interpretation, and SHAP analysis further clarifies the dominant roles of COD, NH_4^+-N and other variables in the performance of the system, which verifies the potential and scalability of the constructed model in complex industrial scenarios. The constructed model is validated for its practical potential and extension value in complex industrial scenarios.
Keywords: Industrial wastewater treatment efficiency; XGBoost-based intelligent analysis; Feature engineering; Principal component analysis (PCA); Time-series sliding window

1 INTRODUCTION

With the continuous advancement of industrialization, wastewater discharge has been growing rapidly, and its treatment efficiency has become an important index to measure the level of resource utilization and the ability of environmental governance. Industrial wastewater has complex composition and large fluctuation of pollution load, and the traditional treatment system is easy to be disturbed under highly variable working conditions, which makes it difficult to realize stable and compliant discharge. At present, the influence of multivariate coupling relationship and dynamic characteristics on the treatment effect is becoming more and more prominent, and there is an urgent need to construct a high-precision analysis model with strong generalization ability to realize the accurate perception and process optimization of the operating state of wastewater treatment system. In this context, the modeling strategy based on the integrated learning method provides a new path for revealing the influence mechanism of treatment efficiency, which has important theoretical value and engineering significance for promoting the construction of intelligent water treatment system.

2 DATA ACQUISITION AND PRE-PROCESSING OF INDUSTRIAL WASTEWATER TREATMENT EFFICIENCY INFLUENCING FACTORS

2.1 Experimental Data Acquisition Program Design

In order to ensure the representativeness and engineering usability of the input data of the industrial wastewater treatment efficiency influencing factors analysis model, a systematic and reproducible experimental data collection program is designed based on the typical operation process of a municipal wastewater treatment plant. The A²/O biochemical treatment system is selected as the object, and high-frequency data collection points are deployed in the key units of pretreatment, anaerobic, aerobic and sedimentation, mainly collecting water quality indicators (pH, COD, BOD_5, NH_4^+-N, SS, TN, TP), operation parameters (influent flow rate, reflux ratio, sludge concentration, sludge concentration, etc.), and data collection data. NH_4^+-N, SS, TN, TP, operating parameters (influent flow rate, reflux ratio, sludge concentration, aeration intensity), environmental variables (air temperature, water temperature, humidity) and so on, a total of 14 core variables, the collection frequency is set to 15 minutes / times, the continuous operation of the collection of 30 days in order to ensure the timeliness and integrity of the data [1]. The collection equipment adopts online multi-parameter water quality analyzer (such as YSI 6600 V2) and PLC system linkage control, and through the RS485 interface to access the SCADA platform, real-time transmission and automatic recording. The data flow structure is shown in Figure 1, the system organizes the raw data according to the time series structure, and automatically marks the missing measurements and abnormal points. In order to follow up the structural consistency of feature construction and model input, the collection program defines the data field format and metadata standards simultaneously to ensure that the pre-processing stage can accurately identify and convert the indicators.

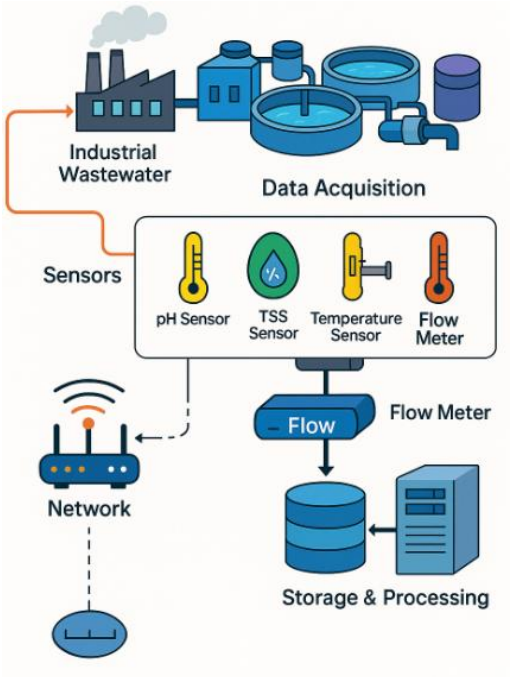

Figure 1 Industrial Wastewater Treatment Data Acquisition System Architecture Diagram

2.2 Data Preprocessing

The data preprocessing process includes key steps such as identifying and filling in missing values, detecting outliers, normalizing variables and aligning time series. Due to the existence of sensor signal drift, communication interruption and human intervention in the actual acquisition process, data integrity is difficult to ensure. First, the original timestamp sequence is sampled synchronously, the uniform interval $\Delta t = 15$ min is set, and the linear interpolation method is used to fill in the missing data with an interruption of no more than 1 hour, and the missing segments with an interruption of more than 1 hour are marked as unavailable. Second, anomaly detection was performed based on the IQR method for the distribution intervals of each variable and combined with engineering upper and lower limits to screen out physically unreasonable values (e.g., COD > 2000 mg/L or pH > 14). In terms of variable scale standardization, considering that XGBoost is insensitive to feature scales, but the subsequent feature construction involves similarity measures and PCA, the Z-score standardization method is uniformly adopted [2]:

$$x_i' = \frac{x_i - \mu}{\sigma} \tag{1}$$

Where μ and σ are the sample mean and standard deviation, respectively. To cope with the problem of multi-source data synchronization, a time window sliding mechanism is introduced to achieve asynchronous feature alignment and compress high-frequency noise. This preprocessing design not only improves the robustness of feature engineering, but also provides standardized, high-quality data input for the model training phase.

2.3 Feature Engineering

The feature engineering scheme for XGBoost modeling is designed to cover three aspects: variable selection, feature construction and feature dimensionality reduction. In the variable selection stage, basic indicators with significant influence were screened based on domain knowledge and statistical correlation analysis (Pearson's coefficient $|r| > 0.6$), and the Variance Threshold method (Variance Threshold ≥ 0.01) was introduced to exclude the redundant features with low information content [3]. In terms of feature construction, combined with the physicochemical coupling characteristics of water quality parameters, ratio class combination features (e.g., COD/TN, BOD_5/COD) and time-difference categorization dynamic features (e.g., ΔCODt= CODt- CODt-1)were designed to enhance the model's ability to perceive the system's nonlinearities and temporal variations. Considering the problem of dimensional catastrophe caused by feature dimension expansion, principal component analysis (PCA) is used for compression to retain the principal components with more than 95% of cumulative explained variance (Figure 2), and the final modeling input feature matrix $X \in Rn \times d$ is generated on this basis.

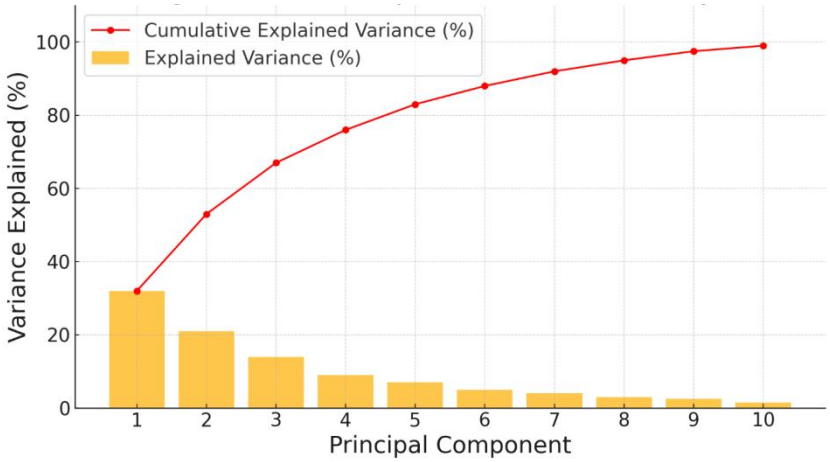

Figure 2 Plot of PCA Principal Component Contribution Analysis

2.4 Data Set Division

After completing the preprocessing and feature construction of the data related to industrial wastewater treatment, in order to ensure the scientific and generalization ability of the model evaluation, the training set, validation set and test set division scheme based on the temporal integrity and sample balance is designed. Considering the dynamic characteristics of the wastewater treatment process and the temporal correlation of the data, the overall dataset is initially segmented according to the chronological order and the sample units are constructed with a non-overlapping sliding window (window length of 60 minutes, step length of 15 minutes) [4]. In order to avoid the information leakage problem, a forward rolling segmentation strategy is used to ensure that the training set is completely independent of the test set time interval, thus realistically simulating the model deployment scenario. Eventually, the dataset is divided into training set (70%), validation set (15%) and test set (15%), each of which covers the smooth period, high load period and abnormal fluctuation period respectively to ensure the diversity and representativeness of sample distribution. The structure of the dataset is summarized in Table 1, which takes into account the operating cycle of the industrial system and the typical load variation characteristics, and helps to improve the adaptability and stability of the subsequent XGBoost model in the real engineering environment.

Table 1 Summary of Dataset Division Scheme

Data set	Time interval coverage	Percentage	Number of windows with time series	Characterization
Training set	Day 1-21	70%	2016	Covering both smooth and varying working conditions with sufficient samples
Validation set	Day 22-25	15% of the total number of samples	432	For parameter tuning
Test Sets	Day 26-30	15 percent	432	No data at all, simulating deployment effects

3 INTELLIGENT ANALYTICS MODEL CONSTRUCTION BY INCORPORATING XGBOOST

3.1 Model Selection and Problem Definition

Considering the high degree of nonlinear coupling and complex dynamic disturbances among multiple variables in the industrial wastewater treatment system, the wastewater treatment efficiency prediction problem is modeled as a typical multi-input regression problem, with the treatment efficiency index (e.g., COD removal rate) as the target variable, and the construction of a multidimensional feature input vector $X \in Rn \times d$, corresponding to the output vector $y \in Rn$. To maintain good generalization performance while dealing with high-dimensional features, nonlinear mapping and strong feature interactions, eXtreme Gradient Boosting (XGBoost) is preferred as the core model in this paper. eXtreme Gradient Boosting (XGBoost) is an enhanced integrated learning algorithm, which introduces second-order derivatives, regularization control, and parallel processing mechanisms based on the traditional gradient boosting tree (GBDT), and is equipped with highly efficient training and strong generalization capabilities. It is especially suitable for modeling situations with missing values, noise perturbations and variable covariance in industrial data. The model loss function is designed as follows [5]:

$$L = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2}$$

Where $l(\cdot)$ denotes the squared error loss, $\Omega(f)$ is the regularity term, T is the number of leaf nodes, and λ controls the weight penalty. The model selection process is shown in Figure 3. Considering the data size, variable characteristics and

application requirements, XGBoost achieves a good balance between accuracy and computational efficiency compared with SVM, RF and ANN, which provides a theoretical guarantee and algorithmic basis for modeling training and parameter tuning.

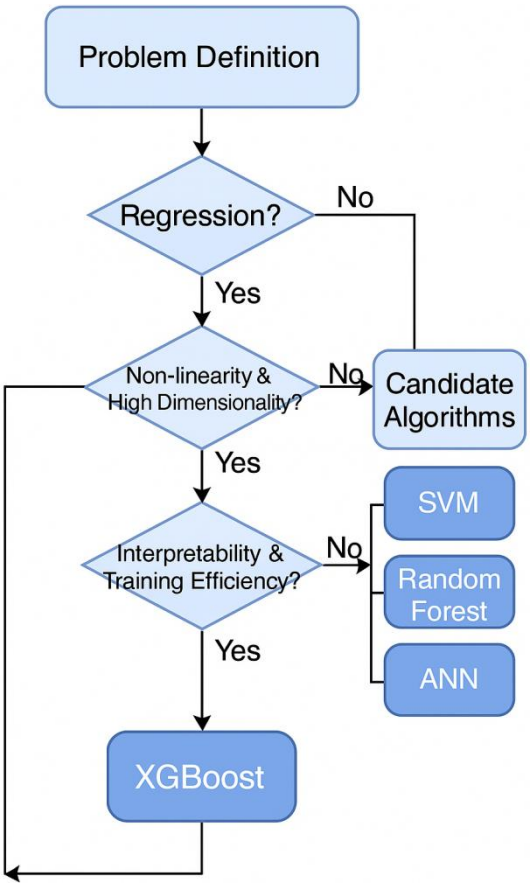

Figure 3 Model Selection Process and Decision Logic Diagram

3.2 Model Construction Process

After clarifying the task attributes and modeling objectives, this study systematically constructs an intelligent analysis model process integrating XGBoost based on the temporal characteristics and feature dimension design of industrial wastewater treatment data. The whole process starts from the standardized feature matrix $X \in Rn \times d$ input, flows into the gradient boosting framework at the core of the model via the training set data, generates multiple weighted regression trees by minimizing the regularized loss function in each iteration, and finally outputs the prediction result y^[6]. The input and output formats, data flow paths and processing order of each module in the model structure are shown in Figure 4. In order to maintain the versatility and scalability of the modeling process, the feature selection module is decoupled from the model engine, allowing subsequent replacement of the feature construction method or the model structure; for the dynamic perturbations that may exist in the wastewater data, this process introduces a time-sliding window mechanism and an incremental model updating mechanism to provide support for online optimization in the deployment environment. In addition, the training-validation-testing data channels are strictly independent, and the processing logic of data flow, model parameters and output indexes are standardized through a unified pipeline module (Pipeline) [7]. The whole modeling process emphasizes efficiency, traceability and engineering deployment friendliness, aiming to provide a structured foundation for parameter tuning and model performance evaluation.

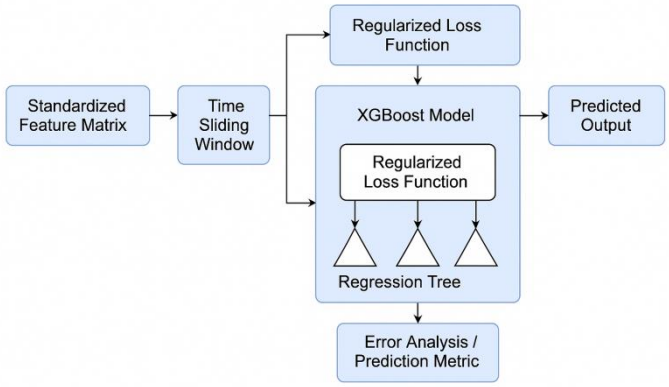

Figure 4 Flowchart of XGBoost Intelligent Analysis Model Construction

3.3 Hyper-Parameter Tuning

Based on the two-stage parameter tuning strategy jointly driven by Grid Search and Bayesian Optimization, the initial stage conducts coarse screening in a predefined discrete grid by Grid Search to determine the response trend of the objective function to the main control parameters (e.g., the learning rate η, the maximum tree depth max_depth, the proportion of subsamples, etc.), and to determine the response trend of the objective function to the main control parameters (e.g., the learning rate η, the maximum tree depth max_depth, the proportion of subsamples). subsample); subsequently, a Bayesian optimization method based on Gaussian process is introduced to iteratively approximate the optimal parameter combinations in the local optimal region. The whole tuning process aims at minimizing the validation set error, and adopts five-fold cross-validation to control the risk of overfitting, with the loss function as [8]:

$$L_{cv} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}(y_i^{(k)} - \hat{y}_i^{(k)})^2\right) \tag{3}$$

Where K=5, nk denotes the number of samples in the kth fold. The tuning parameter space design is shown in Table 2.

Table 2 XGBoost Key Hyperparameters and Search Interval Settings

Parameter name	Description	Search range
learning_rate (η)	Learning rate	[0.01, 0.05, 0.1, 0.2]
max_depth	Maximum depth of each tree	[3, 5, 7, 9]
subsample	Sample sampling ratio	[0.6, 0.8, 1.0]
colsample_bytree	Sampling ratio of feature columns	[0.5, 0.7, 0.9]
n_estimators	Upper limit on the number of weak learners	[100, 300, 500]
gamma	Split minimum loss function gain threshold	[0, 0.1, 0.2]

3.4 Model Training and Validation

After completing the hyper-parameter tuning, the XGBoost training and validation process based on structured pipelines is constructed to realize the robust training and scientific validation of the model in the task of industrial wastewater treatment prediction. The training stage takes the normalized feature matrix Xtrain ∈ Rn×d and the corresponding target value ytrain as inputs, adopts the parallel tree construction strategy, minimizes the squared error loss function with regular terms as the optimization objective, and uses the parameter combination after a priori tuning to control the training process. To avoid overfitting, an early stopping mechanism (early stopping) is introduced, where training is interrupted when the loss does not decrease in several consecutive rounds on the validation set Xval. Both training error and validation error are evaluated by Root Mean Square Error (RMSE) metric, defined as follows [9]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

The model training iteration process and error convergence trend are shown in Figure 5, which can be used to assess the model learning stability and generalization ability. In addition, in order to enhance the robustness of the assessment, the whole training process is nested with a five-fold cross-validation structure, with the average RMSE as the performance metric, based on which the model structure with the best validation performance is further preserved for the test phase deployment.

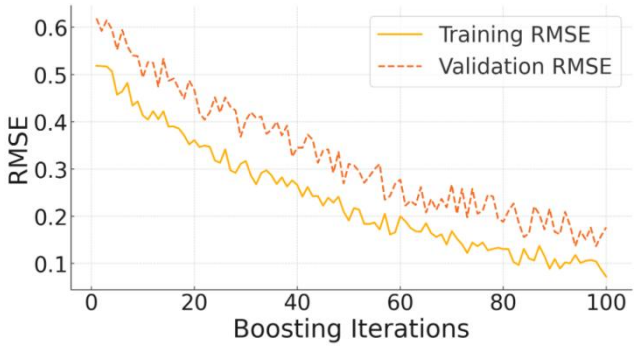

Figure 5 Convergence Curve of Model Training Error and Validation Error

4 EXPERIMENTAL VALIDATION AND RESULT ANALYSIS

4.1 Experimental Design

In order to systematically evaluate the performance of the intelligent analytical model integrating XGBoost in industrial wastewater treatment efficiency prediction, the experiment relies on the previously constructed dataset division structure, with the training set to complete the model training, the validation set for parameter tuning, and the test set for the final performance verification to ensure the independence of the whole process data. The experiments are categorized into three types of evaluation dimensions [10]: prediction accuracy, robustness and feature sensitivity, and the RMSE, MAE, R^2 and residual distribution structure are used as the main evaluation indexes, respectively. To test the advantages of XGBoost in nonlinear multivariate industrial data scenarios, comparison experiments with mainstream models such as Support Vector Regression (SVR), Random Forest (RF), and Multi-Layer Perceptron (MLP) are designed, and all the comparison models are uniformly configured on the same training and validation test set, and the dimensionality of the input features is controlled to be consistent to avoid the influence of bias. In addition, in order to analyze the degree of model dependence on key features, SHAP (SHapley Additive exPlanations) based method is introduced for explanatory experimental design, and feature importance mapping diagrams are constructed to provide mechanism-level support for the result analysis part. The parameters and evaluation indexes of each experimental setup are summarized in Table 3, and the experimental environment configuration and software version are supplemented in the Appendix to ensure the reproducibility and engineering promotion value of the experiments.

Table 3 Configuration Table of Experimental Program Design and Evaluation Indexes

Experiment type	Comparison model	Data source	Evaluation index	Feature interpretation method
Accuracy Validation	XGBoost, SVR, RF, MLP	Test Sets	RMSE, MAE, R^2	N/A
Robustness Analysis	XGBoost	Noise sample injection	RMSE, ΔR^2	N/A
Feature Sensitivity Analysis	XGBoost	Test Set	-SHAP	SHAP

4.2 Comparative Analysis of Model Performance

In order to systematically assess the comprehensive performance of the fused XGBoost model in predicting the efficiency of industrial wastewater treatment, this section constructs a design of experiments for comparison with three mainstream algorithms, namely SVR (Support Vector Regression), RF (Random Forest) and MLP (Multi-Layer Perceptron), focusing on three major dimensions, namely, prediction accuracy, model robustness and feature expression ability. All models uniformly use divided datasets with consistent input variables and are optimized separately in the tuning space to ensure the fairness of comparison and the principle of control variables. The performance evaluation metrics include RMSE (root mean square error), MAE (mean absolute error) and R^2 (coefficient of determination), and the residual distributions of the output fluctuations of the different models under stability test conditions are analyzed to check their disturbance resistance. The structural arrangement of the assessment results is shown in Figure 6, in which the results of each index are normalized to the [0,1] interval to uniformly compare the relative advantages of the performance of each model. In addition, in order to enhance the interpretability, the SHAP value attribution is visualized for the input responses of each model, and the feature importance ranking is listed in Table 4, which provides a structural explanation basis for the model performance differences. The overall analysis process not only reflects the horizontal comparison of prediction accuracy, but also combines the error composition, feature dependence and nonlinear fitting ability to carry out a multi-dimensional comprehensive analysis, forming a closed-loop performance evaluation system from data-driven to mechanism cognition.

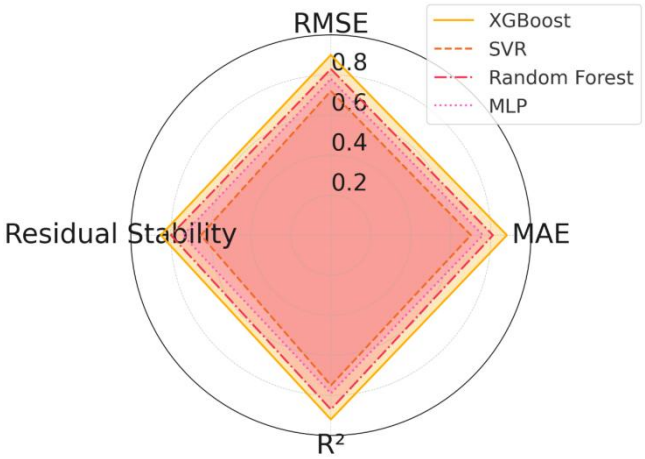

Figure 6 Radar Chart of Standardized Comparison of Multi-Model Performance Indicators

Table 4 Summary of SHAP Importance Ranking of Main Input Features of Each Model

Feature Name	XGBoost Ranking	RF Ranking	SVR Ranking	MLP Ranking
COD	1	1	2	1
NH_4^+-N	2	3	1	3
BOD_5	3	2	4	2
TN	4	4	3	4

5 CONCLUSION

In this study, we constructed an intelligent analytical model of factors influencing industrial wastewater treatment efficiency by integrating XGBoost, which realized the organic integration of systematic acquisition of high-frequency time-series data, feature construction and nonlinear modeling, and enhanced the model's generalization ability and engineering adaptability while accurately predicting the treatment efficiency. Through feature selection and SHAP interpretive analysis, the dominant role of key water quality indicators on system performance is revealed, providing a quantitative basis for mechanism understanding. The model demonstrates superior accuracy and stability in multi-metric and multi-model comparison experiments, demonstrating the potential of integrated learning methods for a wide range of applications in modeling complex industrial processes. However, the sample source is limited to a single scenario, and the model mobility and real-time feedback ability still need to be improved. In the future, a multi-source heterogeneous data fusion mechanism can be further introduced to expand the cross-scene adaptability of the model, and at the same time, the online incremental learning strategy can be combined to explore new paths for intelligent optimization and decision support of wastewater treatment systems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Lv J, Du L, Lin H, et al. Enhancing effluent quality prediction in wastewater treatment plants through the integration of factor analysis and machine learning. Bioresource Technology, 2024, 393: 130008.

[2] Li J, Du Z, Liu J, et al. Analysis of factors influencing the energy efficiency in Chinese wastewater treatment plants through machine learning and SHapley Additive exPlanations. Science of The Total Environment, 2024, 920: 171033.

[3] Aparna K G, Swarnalatha R. Optimizing wastewater treatment plant operational efficiency through integrating machine learning predictive models and advanced control strategies. Process Safety and Environmental Protection, 2024, 188: 995-1008.

[4] Al-Jamimi H A, BinMakhashen G M, Saleh T A. From data to clean water: XGBoost and Bayesian optimization for advanced wastewater treatment with ultrafiltration. Neural Computing and Applications, 2024, 36(30): 18863-18877.

[5] Shao S, Fu D, Yang T, et al. Analysis of machine learning models for wastewater treatment plant sludge output prediction. Sustainability, 2023, 15(18): 13380.

[6] Nasir F B, Li J. Comparative Analysis of Machine Learning Models and Explainable Artificial Intelligence for Predicting Wastewater Treatment Plant Variables. Advances in Environmental and Engineering Research, 2024, 5(4): 1-23.

[7] Nguyen D V, Park J, Lee H, et al. Assessing industrial wastewater effluent toxicity using boosting algorithms in machine learning: A case study on ecotoxicity prediction and control strategy development. Environmental Pollution, 2024, 341: 123017.

[8] Shu Y, Kong F, Lin X, et al. Leveraging ionic information for machine learning-enhanced source identification in integrated wastewater treatment plant. Journal of Water Process Engineering, 2025, 74: 107784.
[9] Saddiqi H A, Javed Z, Ali Q M, et al. Optimization and predictive modeling of membrane based produced water treatment using machine learning models. Chemical Engineering Research and Design, 2024, 207: 65-76.
[10] Tenneti S, Divya P D, Tejaswini E S S, et al. Interpretability and performance assessment of advanced machine learning models for α-factor prediction in wastewater treatment plants. Journal of Water Process Engineering, 2025, 72: 107637.