Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Adv.*, 2026, **5**, 1174

Application of explainable artificial intelligence and machine learning in predicting wastewater treatment plant variables: a comparative study of small- and large-scale treatment plants

Fuad Bin Nasir * and Jin Li*

Received 9th November 2025
Accepted 27th February 2026

DOI: 10.1039/d5va00408j

rsc.li/esadvances

Explainable artificial intelligence (XAI) can play a significant role in the application of machine learning (ML) in wastewater treatment plants (WWTPs). The present research focuses on evaluating the performance and generalizability of widely used ML models for predicting key effluent quality variables at multiple WWTPs. Effluent variables including ammonia nitrogen (NH_3–N), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total phosphorus (TP), and total suspended solids (TSS) were predicted using eXtreme Gradient Boosting (XGBoost) and Random Forest (RF) models. Several feature selections (FS) and XAI tools were used to understand the influence of input variables on target variables and the impact of input variables on model performance. The study demonstrates that XAI can enhance the understanding of model behavior by identifying key input variables, thereby supporting more informed and transparent decision-making at WWTPs. The study finds that XAI methods are effective in capturing the influence of variables regardless of the choice of model for variable prediction. XAI tools, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), are successful in providing deeper understanding of the key factors influencing ML model performance. The findings of this research could facilitate WWTP operation with better decision-making in choosing ML models to optimize treatment performance and improve environmental sustainability.

Environmental significance

Wastewater treatment plants play a crucial role in safeguarding public health and the environment by effectively removing contaminants before their release. However, managing these plants efficiently can be challenging due to the complex nature of the treatment processes. Machine learning (ML) can predict water quality outcomes, enabling plant operators to optimize their management strategies. This study demonstrates that explainable artificial intelligence (XAI) can provide rationales behind ML model predictions, thereby enhancing operators' confidence in decision-making. In summary, this study promotes efficient and sustainable management of wastewater treatment processes to ensure better water quality and overall environmental health.

1 Introduction

Wastewater treatment plants (WWTPs) play a crucial role in sustainable water management by removing pollutants from wastewater before discharging to the natural environment. However, approximately 48% of the world's wastewater is discharged untreated, which leads to significant environmental challenges.[1] Traditionally, WWTPs have relied on empirical models and manual monitoring to manage their operation which failed to address the nonlinear, multivariate, and dynamic nature of wastewater treatment processes. It is crucial to predict influential effluent variables accurately to overcome the unpredictability and vulnerability of current WWTP operations, thus comply with effluent regulations as well as to reduce costs and enhance operational efficiency. Recently, machine learning (ML) has gained popularity in the WWTP sector, demonstrating significant accuracy in predicting key variables that can optimize treatment plant performance while minimizing costs and labor.[2–11]

While ML has proven successful in predicting variables with significant accuracy, operators of WWTPs cannot solely rely on black-box ML models, which lack transparency in their predictions. Recently, explainable artificial intelligence (XAI) has been studied with ML models to address the opaque nature of their predictions, providing insights into the reationales of specific outcomes.[12,13] This transparency enhances model acceptance among practitioners and identifies key factors for process optimization. However, existing studies have not accounted for the varying operational settings of various WWTPs. The validity of XAI in the context of WWTPs requires investigation that

Dept. of Civil and Environ Engr., Univ. of Wisconsin-Milwaukee, WI 53211, USA. E-mail: fnasir@uwm.edu; li@uwm.edu

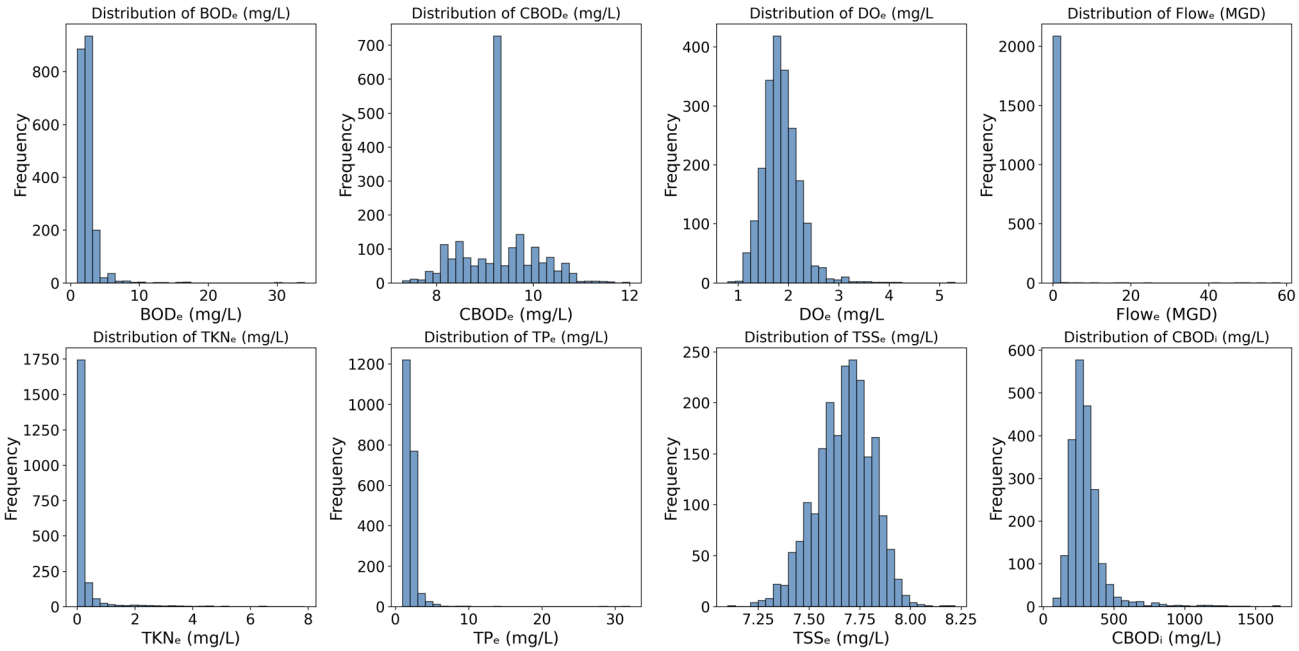

Fig. 1 Monroe WWTF variable distribution.

encompasses a diverse range of facilities. Small-scale plants often encounter resource constraints, limited data availability, and fluctuating influent loads, whereas large-scale facilities must manage high-throughput operations and maintain robust control systems. A comparative analysis of ML performance across small- and large-scale WWTPs will provide valuable insights into the scalability and adaptability of ML model prediction.

This study examines the application of ML and XAI in predicting critical variables at WWTPs, comparing their effectiveness in both small-scale and large-scale facilities. The research utilizes models such as eXtreme Gradient Boosting (XGBoost) and random forest (RF) to investigate the performance of ML algorithms for distinct WWTPs. Additionally, XAI techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are employed to

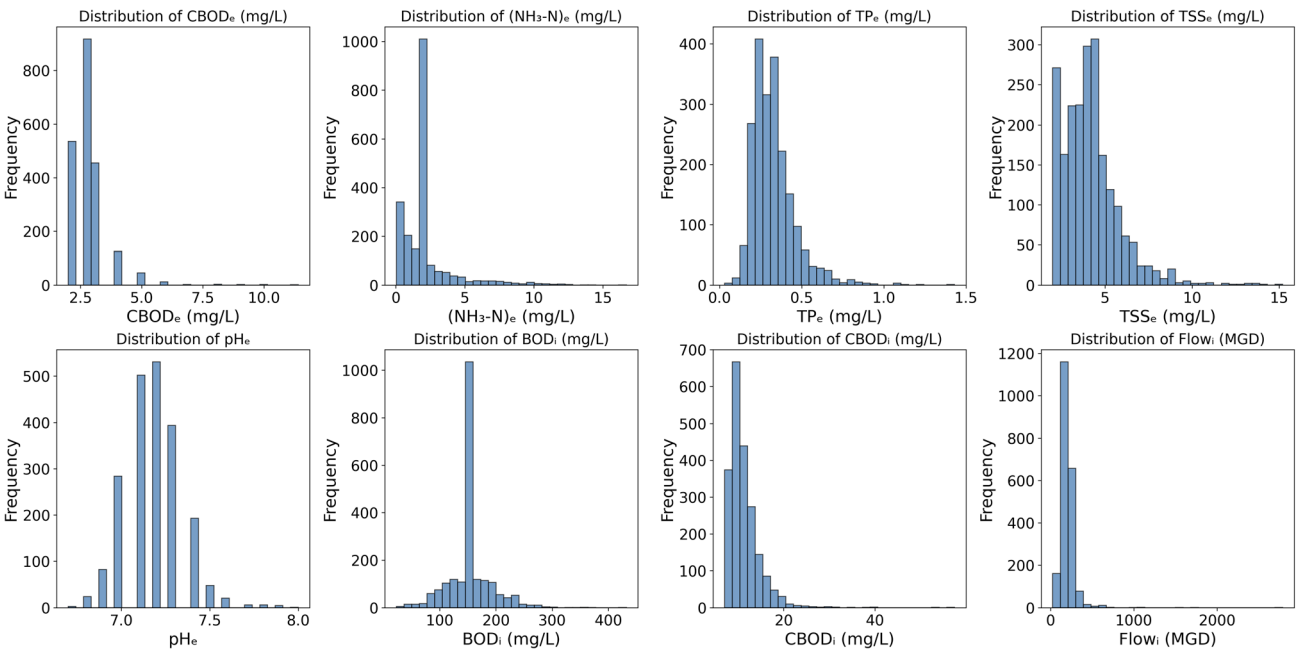

Fig. 2 Sheboygan WWTF variable distribution.

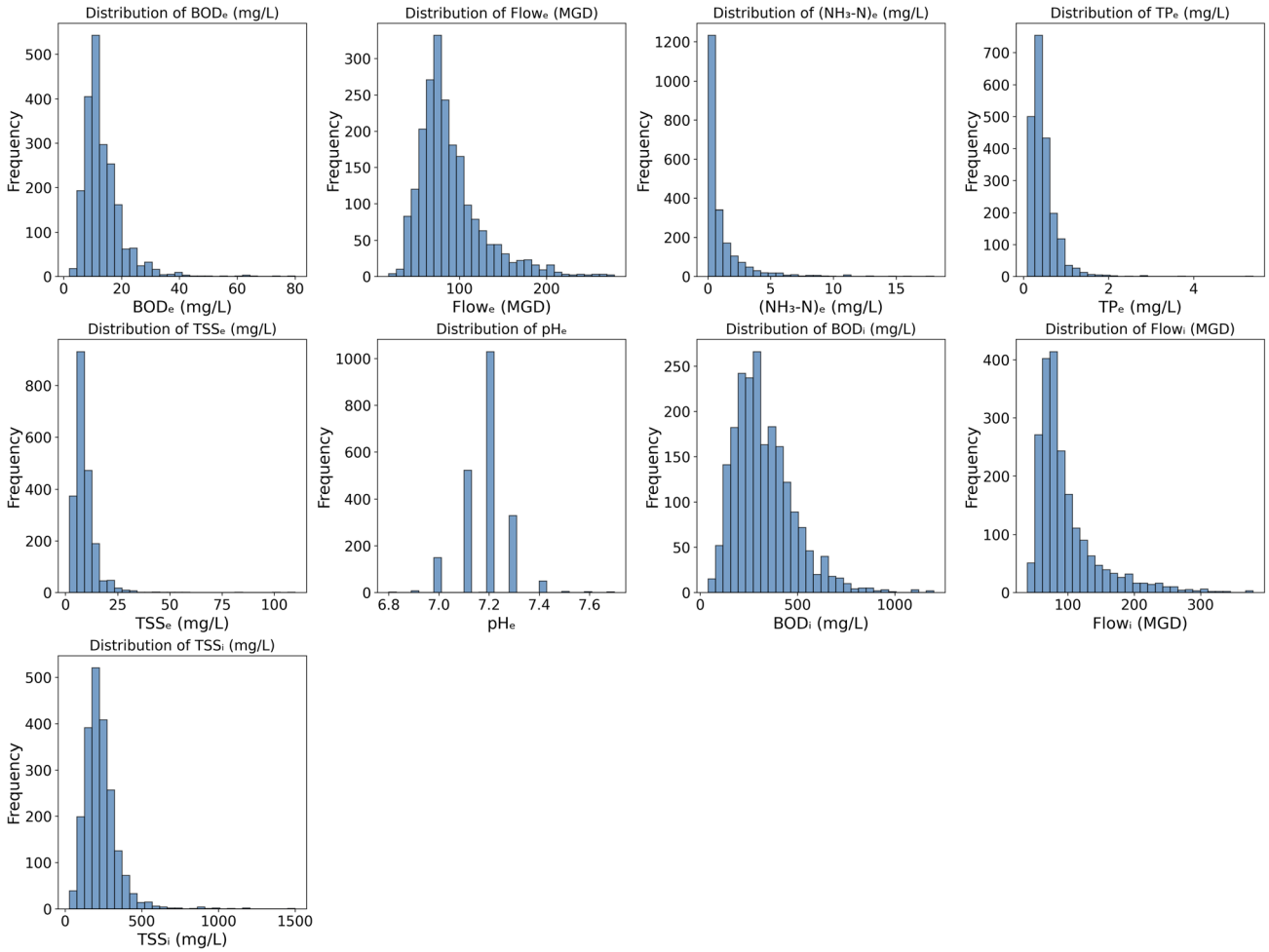

Fig. 3 MMSD SSWRF variable distribution.

interpret predictions, providing a deeper understanding of the key factors influencing ML model performance.

This study demonstrates that integrating ML with XAI creates a robust and transparent framework for predicting key effluent quality variables in varying scales of WWTPs. The comparative analysis between small- and large-scale WWTPs showed that advanced ML models, such as XGBoost and RF, can achieve desirable predictive performance despite differences in operational complexity, data availability, and influent variability. The use of XAI techniques, including SHAP and LIME, allowed for meaningful interpretation of model predictions by consistently identifying influential input variables and capturing their real-world relevance across different treatment plant settings. These insights enhance trust in ML-based decision support systems and provide practical guidance for process optimization and regulatory compliance.

2 Methods

2.1 Data collection and processing

Data were obtained from four treatment plants located in Wisconsin, USA. The City of Monroe Wastewater Treatment Facility (WWTF) has the capacity to treat 3.7 million gallons per day (MGD) of flow. The Sheboygan WWTF can treat an average of 18.4 MGD with a peak design capacity of 58.6 MGD. The Madison Metropolitan Sewerage District (MMSD), conveys approximately 37 MGD wastewater to the plant. The Milwaukee Metropolitan Sewerage District South Shore Water Reclamation Facility (MMSD SSWRF) has the capacity to treat a daily flow of 250 MGD and a peak hourly flow of 300 MGD. The intention for choosing this combination of small- and large-scale treatment plants was to capture the XAI performance from operational settings at WWTPs with various treatment capacities. The dataset consists of daily data from 1st January 2019 to 30th November 2024. The collected data were thoroughly checked for repetition. In the case of missing data, if a variable has more than 50% missing data, it was deleted to avoid excessive uncertainty introduced by imputation. Missing values of a variable (<50%) were replaced with mean values to preserve dataset continuity for baseline model development. To reduce multicollinearity, variables that have high correlation (greater than or equal to |0.9|) were removed. The WWTP variable distributions are shown in Fig. 1–4 and variable statistics are shown in Tables 1–4.

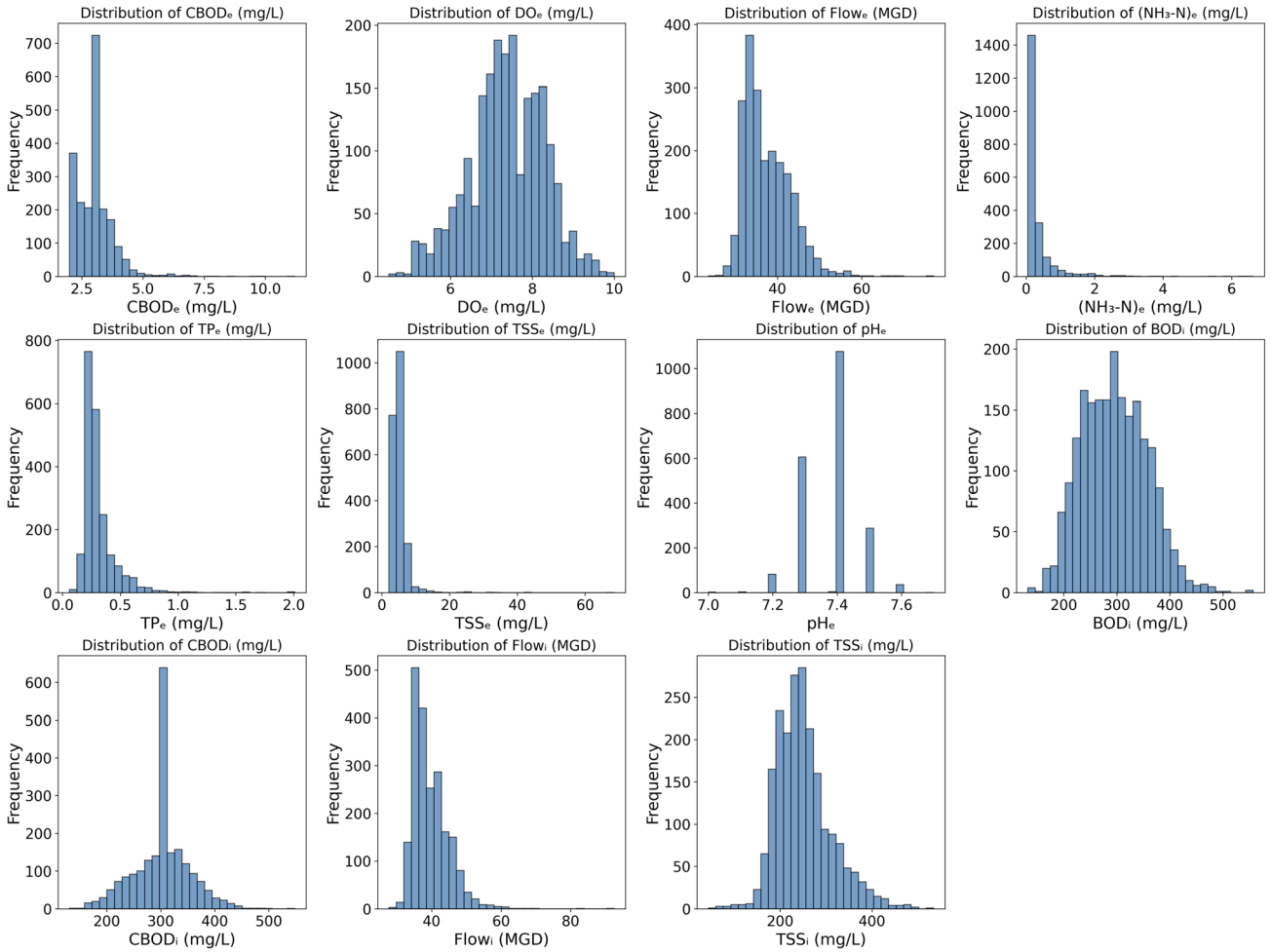

Fig. 4 MMSD variable distribution.

Table 1 Monroe WWTF variable statistics ("i" denotes influent and "e" denotes effluent)

Variable name	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
BOD_e (mg L^{-1})	1.00	34.00	2.80	1.50	0.54
$CBOD_e$ (mg L^{-1})	7.30	12.00	9.30	0.73	0.08
DO_e (mg L^{-1})	0.80	5.31	1.86	0.36	0.19
$Flow_e$ (MGD)	0.04	58.40	0.33	2.60	7.87
TKN_e (mg L^{-1})	0.01	7.88	0.30	0.71	2.39
TP_e (mg L^{-1})	1.00	32.00	2.06	1.25	0.61
TSS_e (mg L^{-1})	7.10	8.22	7.67	0.14	0.02
$CBOD_i$ (mg L^{-1})	70.00	1675.00	302.67	137.00	0.45

2.2 Feature selection (FS)

Some well-known FS approaches, *i.e.* least absolute shrinkage and selection operator (LASSO), mutual information (MI), random forest (RF), Pearson correlation (PC), and principal component analysis (PCA), were tested to compare their selected influential variables with XAI tool results. Since multiple FS methods were investigated in the study, a normalized importance scaling technique was used to standardize the range of scores to compare different methods regardless of their different scales or units. A score of 0 indicates that the feature has the lowest importance, while a score of 1 indicates that the feature has the highest importance.

2.3 SHapley Additive exPlanations (SHAP)

SHAP analysis is a recently developed method in XAI based on game theory. It provides insights into the behavior of ML models by explaining how predictions are made, particularly highlighting the relative impact of input variables on model performance.[10,14] In this study, SHAP values based on game

Table 2 Sheboygan WWTF variable statistics

Variable name	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
$CBOD_e$ (mg L^{-1})	2.00	11.40	2.81	0.77	0.28
$(NH_3–N)_e$ (mg L^{-1})	0.04	16.70	2.23	1.89	0.85
TP_e (mg L^{-1})	0.03	1.43	0.32	0.13	0.40
TSS_e (mg L^{-1})	2.00	15.20	4.16	1.62	0.39
pH_e	6.70	8.00	7.19	0.16	0.02
BOD_i (mg L^{-1})	23.00	432.00	155.09	38.45	0.25
$CBOD_i$ (mg L^{-1})	7.06	57.44	11.33	3.61	0.32
$Flow_i$ (MGD)	25.00	2790.00	204.05	107.17	0.53

Table 3 MMSD SSWRF variable statistics

Variable name	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
BOD_e (mg L^{-1})	2.00	80.00	13.42	6.86	0.51
$Flow_e$ (MGD)	19.00	278.00	89.52	36.90	0.41
$(NH_3–N)_e$ (mg L^{-1})	0.02	18.00	1.08	1.67	1.55
TP_e (mg L^{-1})	0.09	5.40	0.46	0.31	0.68
TSS_e (mg L^{-1})	1.90	110.00	9.32	6.06	0.65
pH_e	6.80	7.70	7.18	0.09	0.01
BOD_i (mg L^{-1})	40.00	1200.00	326.89	152.79	0.47
$Flow_i$ (MGD)	39.00	379.00	97.41	48.04	0.49
TSS_i (mg L^{-1})	28.00	1500.00	232.17	113.13	0.49

Table 4 MMSD variable statistics

Variable name	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
$CBOD_e$ (mg L^{-1})	2.00	11.20	3.04	0.77	0.25
DO_e (mg L^{-1})	4.50	10.00	7.37	0.92	0.12
$Flow_e$ (MGD)	23.57	77.12	37.51	5.53	0.15
$(NH_3–N)_e$ (mg L^{-1})	0.05	6.63	0.32	0.52	1.62
TP_e (mg L^{-1})	0.06	2.01	0.31	0.16	0.52
TSS_e (mg L^{-1})	2.10	68.60	5.19	2.83	0.55
pH_e	7.00	7.70	7.38	0.08	0.01
BOD_i (mg L^{-1})	132.00	557.00	295.64	61.60	0.21
$CBOD_i$ (mg L^{-1})	131.00	548.00	301.76	51.73	0.17
$Flow_i$ (MGD)	27.71	92.55	39.67	5.25	0.13
TSS_i (mg L^{-1})	44.20	534.00	251.51	60.53	0.24

theory[10,15,16] were utilized to understand how input variables affect the model's performance.

2.4 Local Interpretable Model-Agnostic Explanations (LIME)

LIME is an XAI tool designed to interpret black-box ML models by creating a local, interpretable model that provides explanations for each prediction made by the original model.[13] It identifies the critical features influencing a specific prediction and assigns a weight to each feature, reflecting its respective impact. Features with positive weights contribute positively to the prediction, whereas those with negative weights exert a detrimental influence. The magnitude of each weight highlights the strength of the feature's effect. Subsequently, the features were ranked according to their significance, with the most influential features prominently positioned at the top. The explanations provided by LIME follow the methodology outlined by Ribeiro.[17]

2.5 ML models

RF is an ensemble learning technique and one of the most influential methods in ML. It employs multiple decision trees to generate predictions.[18–21] An important feature of this algorithm is its ability to provide insights into the importance of each variable, which is essential for further analysis.[4] XGBoost is an optimized and highly efficient version of the gradient boosting algorithm. It has become a popular choice for both regression and classification tasks because of its superior performance and scalability. This technique enhances traditional gradient boosting by incorporating regularization techniques and

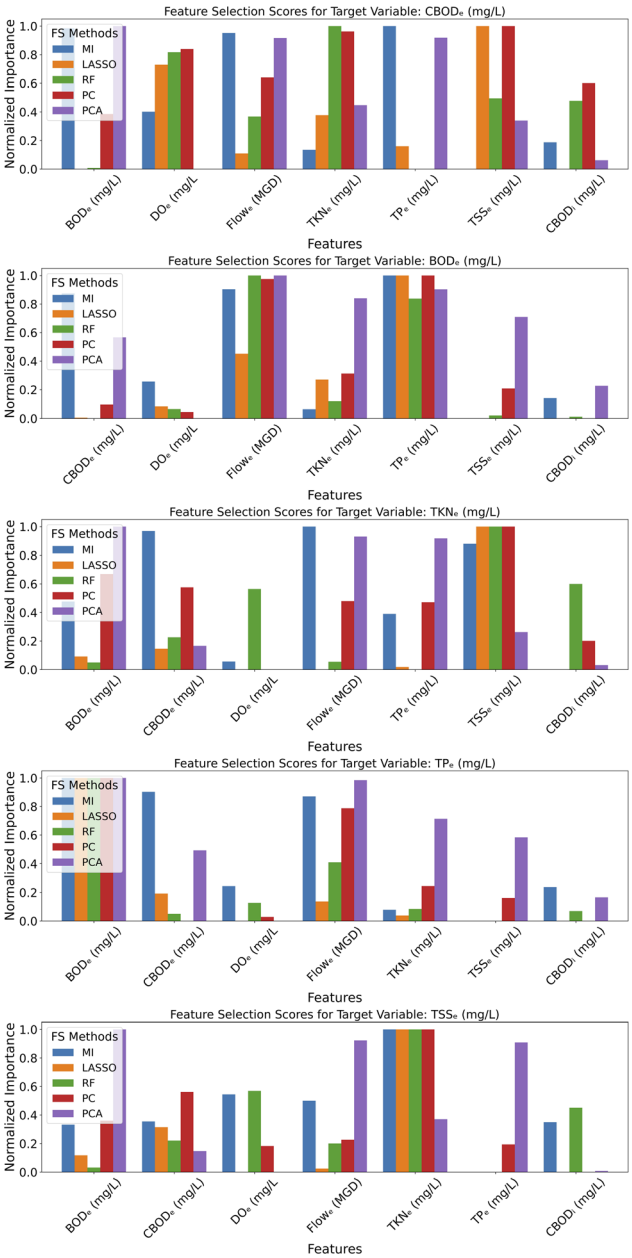

Fig. 5 Feature selection scores of Monroe WWTP target variable.

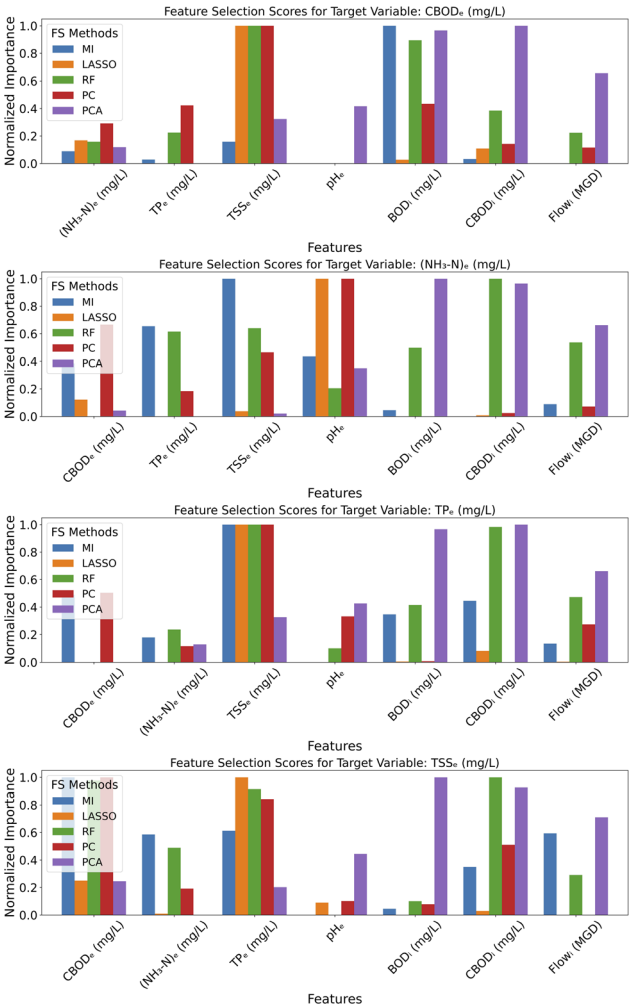

Fig. 6 Feature selection of Sheboygan WWTP target variable.

parallel processing capabilities. RF and XGBoost were chosen because of their widespread application in the water sector. For the RF model we imported the RandomForestRegressor class from the sklearn ensemble module. GridSearch method was used to find the best hyperparameters for building the RF model. The tuning process involved varying the number of trees (n_estimators: 50, 100, 150, 200, 250, 300), the maximum depth of each tree (max_depth: none, 5, 10, 15, 20), the minimum number of samples required to split an internal node (min_samples_split: 2, 5, 10, 15, 20), and the minimum number of samples required in a leaf node (min_samples_leaf: 1, 2, 4, 6, 8). Additionally, the number of features considered for each split (max_features: 'auto', 'sqrt', 'log 2') was fine-tuned to balance model accuracy and computational efficiency.

For XGBoost, we imported the XGBRegressor class from the XGBoost library. The tuning process involved varying the number of trees from 50 to 300 to determine the optimal number of boosting rounds and learning rate from 0.01 to 0.3. To prevent overfitting and control tree complexity, the maximum depth varied from 3 to 10, and the minimum sum of instance weights required in a child node was varied from 1 to 5. Further optimization included tuning the subsample from 0.6 to 1.0 to regulate the fraction of samples used per boosting round and the subsample ratio of columns from 0.6 to 1.0 to manage the number of features considered in each tree. The optimal hyperparameter combination, identified through systematic tuning, enhanced model generalization while maintaining computational efficiency, ensuring reliable predictions in wastewater treatment applications.

2.6 Model training and evaluation

In the analysis, the dataset was divided into a training set, a validation set and a test set. The first 85% of data (from 1st January 2019 to 20th November 2023) were used as the training and validation set and last 15% (from 21st November 2023 to

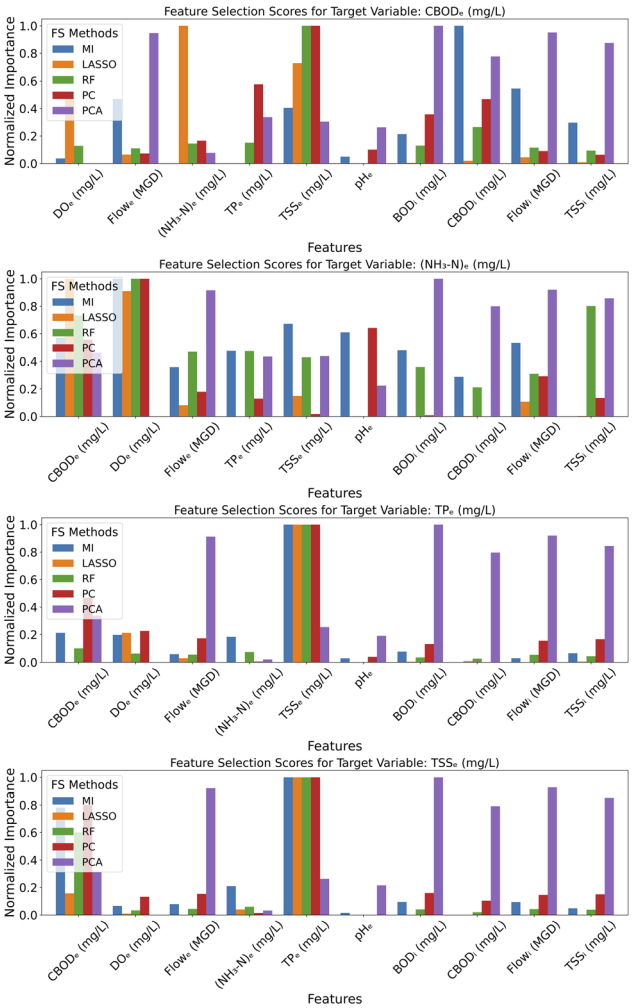

Fig. 7 Feature selection scores of Madison WWTP target variable.

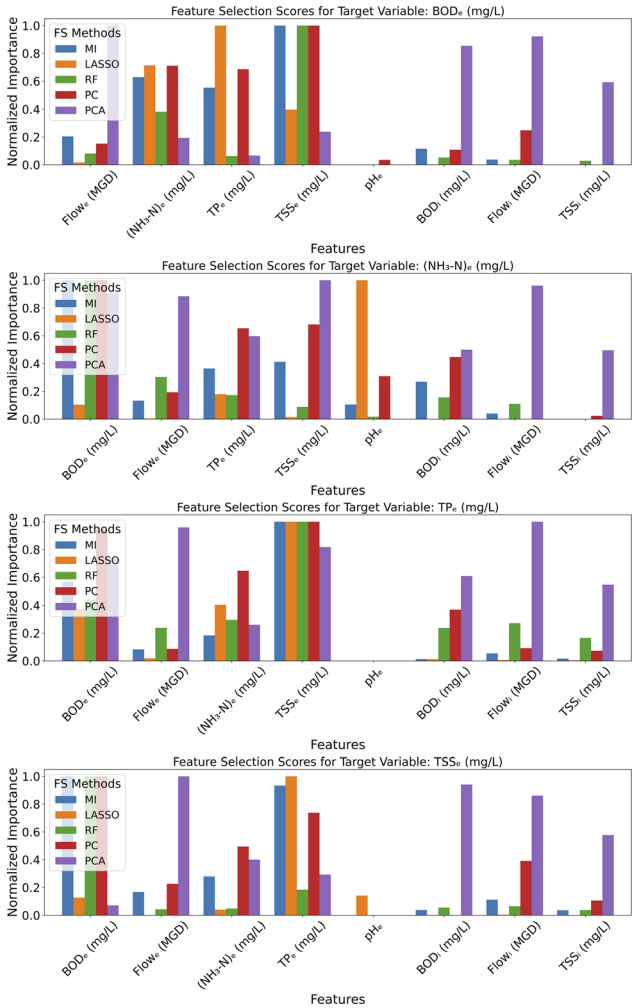

Fig. 8 Feature selection scores of Milwaukee WWTP target variable.

30th November 2024) of data were used as the test set, representing future unseen data. This approach represented how models would be applied into real settings, where predictions are made for future data based on past observations. Out of the first 85% data, 70% of data were randomly chosen for training and 15% for validation. From the training set, the model learned patterns, relationships, and features from this subset of the data that allowed the model to optimize its parameters and learn from the data. The validation set allowed for systematic adjustment of model settings without affecting the final unbiased evaluation. By using a separate validation set for tuning, the test set remained untouched until final evaluation, ensuring a reliable assessment. The test set provided an independent assessment of model performance on unseen data. Since the test set was not seen during training or tuning, it mimics how the model would perform in real-world scenarios with new data.

The k-fold cross-validation was also performed to lower the risk of overfitting[22] by splitting the entire dataset into five equal-sized sections. Thus, the 70% training set was further divided into 5 folds for cross-validation, where 4 folds are used for training and 1 fold was used for validation. All cross-validation and tuning procedures were conducted only during the pre-test period, and the final model performance was assessed specially on the test set that was held out chronologically. Using a well-known hyperparameter tuning method, GridSearch, various hyperparameter combinations for RF and XGBoost models were tested to find the optimum set that yielded the best performance on certain validation metrics.

To evaluate the model's performance, four widely used assessment metrics – Mean Absolute Error (MAE), MSE (Mean Squared Error), R-squared (R^2), and Root Mean Squared Error (RMSE) – were used to evaluate the performance of the ML models. MAE is a valuable metric that assesses the average magnitude of errors between predicted and actual values, offering insights into how closely predictions align with real outcomes. MSE complements this by calculating the average squared difference between predicted and actual values, helping to illustrate the potential magnitude of deviations in predictions. R^2 is particularly useful as it quantifies the percentage of variance in the data that the model can explain, with values ranging from 0 to 1. A value of 1 indicates a perfect explanation of variability, while a value of 0 suggests that the

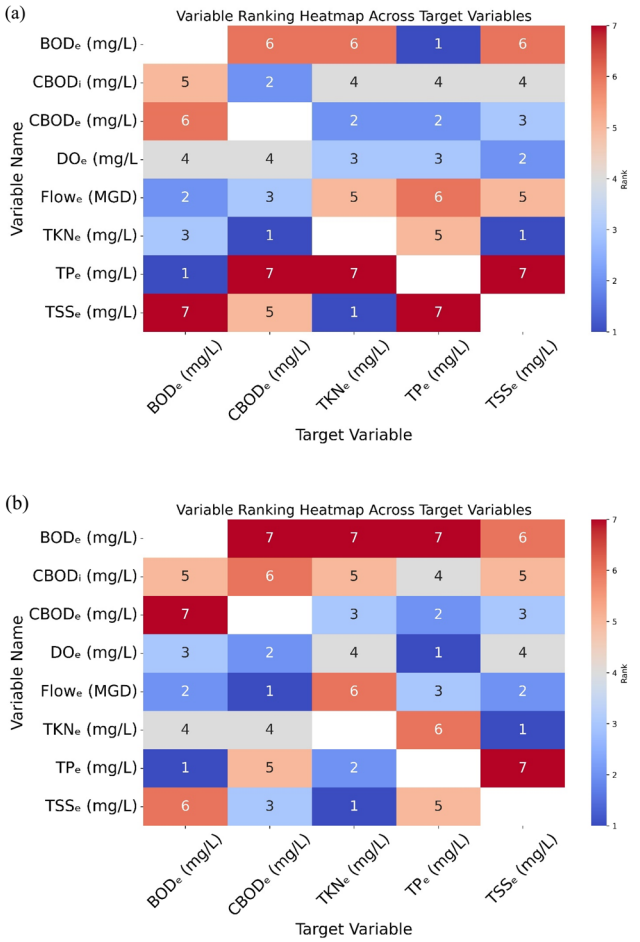

Fig. 9 (a) SHAP Monroe for the RF model. (b) LIME Monroe for the RF model.

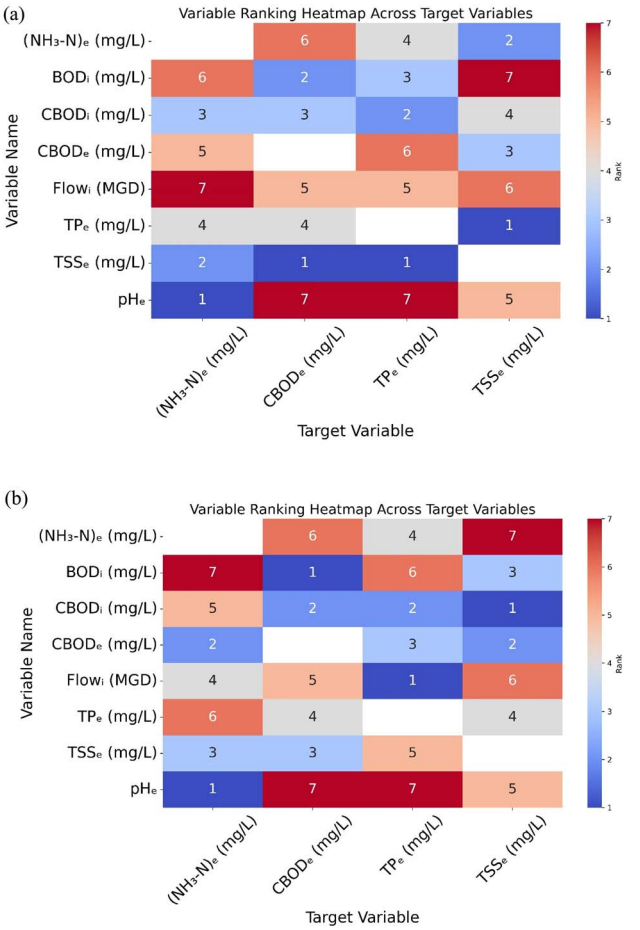

Fig. 10 (a) SHAP Sheboygan for the RF model. (b) LIME Sheboygan for the RF model.

model does not capture any variance. Meanwhile, RMSE provides an average size of residuals and is always non-negative; lower values indicate a stronger fit to the data. Together, these metrics offer constructive feedback on the precision, goodness-of-fit, and accuracy of ML models.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (1)$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \quad (2)$$

$$R^2 = 1 - \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}} \quad (3)$$

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

where $\hat{y}_i$ is the predicted value, y_i is the experimental data and n is the number of test observations.

3 Results and discussion

3.1 FS selection

FS methods demonstrated consistent variable importance across varying scales of WWTPs, as shown in Fig. 5–8. For the Monroe WWTP, $CBOD_e$, TKN_e, and $Flow_e$ were key predictors across all methods, with BOD_e and TSS_e also playing significant roles in various predictions. In the Sheboygan WWTP, $CBOD_e$, $(NH_3–N)_e$, TSS_e, and BOD_i were influential in multiple predictions. For the Milwaukee WWTP, $Flow_e$, $(NH_3–N)_e$, TP_e, and TSS_e were consistently important across different models, with TP_e showing dependencies on both influent and effluent parameters. In the Madison WWTP, $CBOD_e$, $Flow_e$, TSS_e, and influent variables such as $CBOD_i$ and TSS_i were crucial for multiple predictions. Overall, key variables such as BOD_e, TSS_e, TP_e, and $Flow_e$ were frequently identified as significant predictors across WWTPs, highlighting their critical role in effluent quality modeling.

3.2 SHAP and LIME

The SHAP- and LIME-based analysis of all target variables across the four WWTPs, as shown in Fig. 9–16, provided valuable insights into the factors influencing effluent quality predictions. The variability in significant predictors among

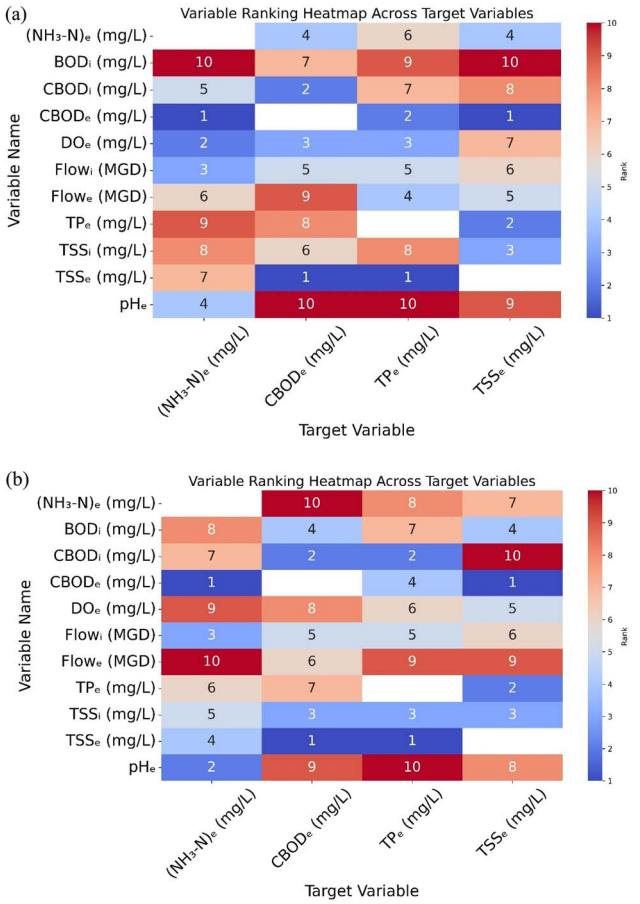

Fig. 11 (a) SHAP Madison for the RF model. (b) LIME Madison for the RF model.

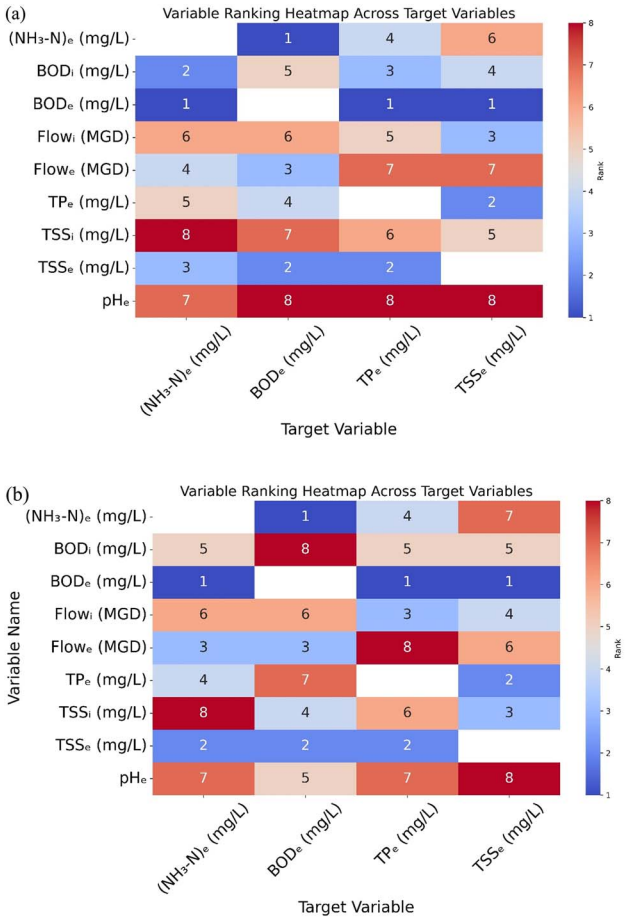

Fig. 12 (a) SHAP Milwaukee for the RF model. (b) LIME Milwaukee for the RF model.

target variables reflected the complex interplay of physical, chemical, and biological processes within WWTPs, underscoring the need for designed monitoring and management strategies.

For BOD_e predictions, both SHAP and LIME consistently identified $Flow_e$, $(NH_3–N)_e$, TP_e, and TSS_e as the dominant contributors in both the RF and XGBoost models. These variables represent major factors such as organic load, nutrient concentration, solid content, and hydraulic conditions, all of which significantly affect BOD_e levels. For $CBOD_e$, both SHAP and LIME analysis indicated that BOD_i, $Flow_e$, TKN_e, and TSS_e were among the top predictors across all WWTPs. These results imply that both the incoming organic load and nitrogen, as well as the solid content, are crucial for determining $CBOD_e$.

TSS_e predictions were strongly driven by other effluent quality variables according to SHAP. Both SHAP and LIME values indicated that BOD_e, $CBOD_e$, TKN_e, and TP_e had substantial impacts on the TSS_e levels. This is insightful because higher organic loads and nutrient concentrations were associated with changes in biomass and particulate matter during the treatment process. LIME, focusing on individual instances, sometimes ranked influential factors such as $CBOD_i$ and $Flow_i$ as influential for TSS_e. For TP_e, both SHAP and LIME were largely in agreement. They consistently highlighted the predictive value of BOD_e, $CBOD_e$, $(NH_3–N)_e$, and TSS_e.

SHAP and LIME results for TKN_e indicated that variables related to organic matter and solids were influential, as BOD_e and $CBOD_e$ were frequently top contributors. This suggests that the effluent organic nitrogen and particulate-associated nitrogen might be linked to the overall organic load and solid retention in the system. LIME outputs for TKN_e also pointed to TP_e. For $(NH_3–N)_e$, both XAI methods identified a mix of carbon, solids, and phosphorus indicators as key explanatory variables. SHAP and LIME analyses emphasized the roles of BOD_i, BOD_e, $CBOD_e$, TSS_e, and TP_e in the ammonia predictions, indicating that higher organic carbon and solids, as well as higher effluent phosphorus, were associated with changes in ammonia removal or production. This alignment between SHAP and LIME suggests strong coupling between carbon usage, solids, and phosphorus in the transformation of ammonia during the treatment process.

In summary, the SHAP and LIME results were largely complementary. SHAP was well suited for identifying consistent, globally important drivers of model predictions across the entire dataset and for each target variable. LIME provides insight into instance-specific nuances, capturing how the

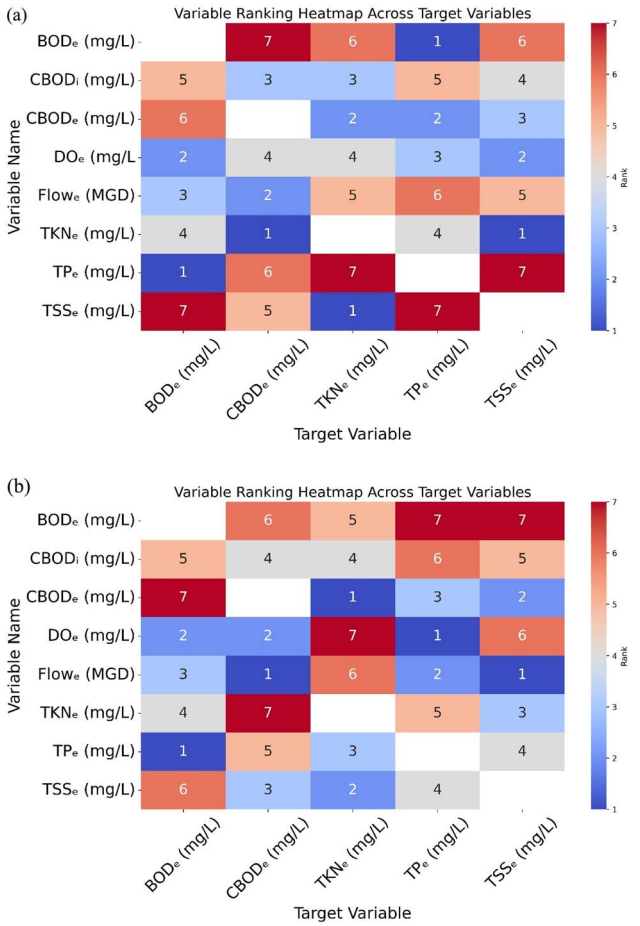

Fig. 13 (a) SHAP Monroe for the XGBoost model. (b) LIME Monroe for the XGBoost model.

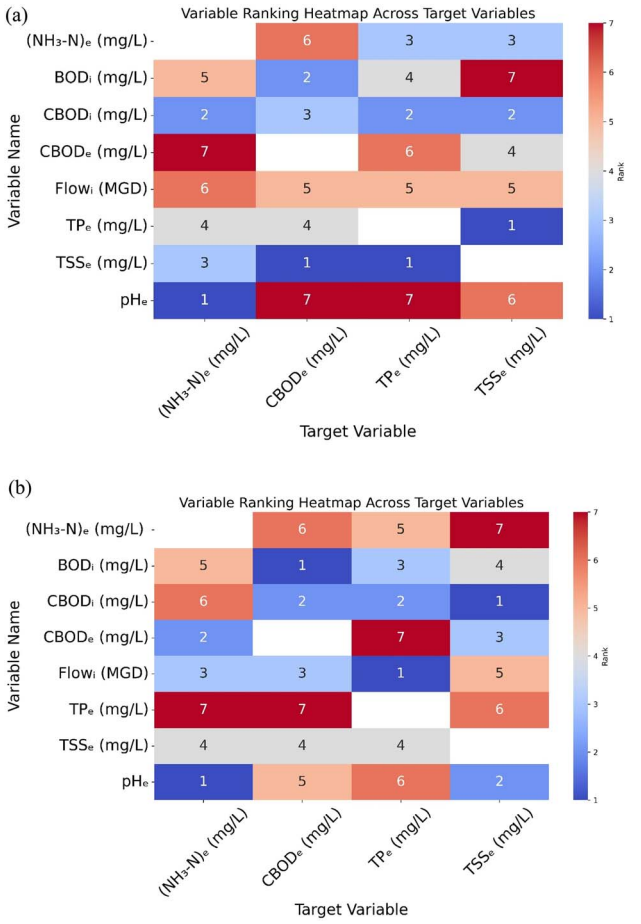

Fig. 14 (a) SHAP Sheboygan for the XGBoost model. (b) LIME Sheboygan for the XGBoost model.

influence of certain features can vary under various conditions or in different plants. The combination of both techniques confirms that the models are learning relationships that make physical and chemical sense and underscores the value of local explanations to detect when and why a model might rely on an unexpected factor for a particular prediction. These XAI insights can help WWTP operators and engineers trust ML model outputs and provide guidance on which variables are most critical for monitoring and controlling effluent quality.

3.3 Model performance

The performance of RF and XGBoost models for various WWTPs and target variables was evaluated using metrics including MAE, MSE, R^2, and RMSE. The analysis provides insights into the models' predictive accuracy for effluent variables across different datasets (training, validation, and test).

3.3.1 Monroe WWTP. For the Monroe WWTP, both RF and XGBoost displayed superior training performance but test results showed negative R^2 values, indicating that the models lacked generalization for all target variables except for total phosphorus (TP_e), as shown in Table 5. This suggests that none of the models achieved the desired level of prediction accuracy for the target variables, indicating that the underlying process dynamics at Monroe were not adequately captured by the modeling framework. These findings emphasize the unique challenges of complex wastewater systems and highlight that the proposed models are not universally applicable across all WWTPs, underscoring the need for site-adaptive modeling strategies.

3.3.2 Sheboygan WWTP. For $(NH_3\text{–}N)_e$, in both RF and XGBoost, the training R^2 dropped in validation and in the test, indicating poor generalization (Table 6). For TP_e and TSS_e, both models achieved comparable test scores.

3.3.3 Madison WWTP. $CBOD_e$ predictions by RF and XGBoost showed reasonable performance in training but comparable performance in test sets. For $(NH_3\text{–}N)_e$, both models struggled in validation and test sets, with negative R^2 values, indicating poor generalization. For TSS_e and TP_e, both RF and XGBoost showed strong training performance compared to test performance (Table 7).

3.3.4 Milwaukee WWTP. For BOD_e, RF demonstrated high training performance ($R^2 = 0.97$) and reasonably good test performance ($R^2 = 0.62$), whereas XGBoost fell behind slightly in training ($R^2 = 0.89$) but achieved comparable test set results ($R^2 = 0.67$). TSS_e predictions showed high accuracy in training for both models ($R^2 > 0.95$). However, RF achieved better test

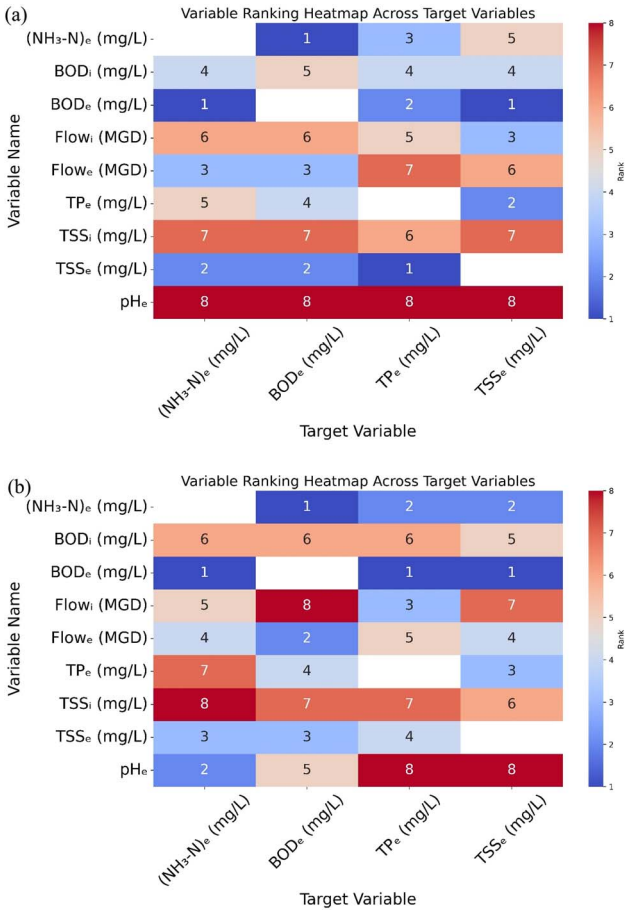

Fig. 15 (a) SHAP Milwaukee for the XGBoost model. (b) LIME Milwaukee for the XGBoost model.

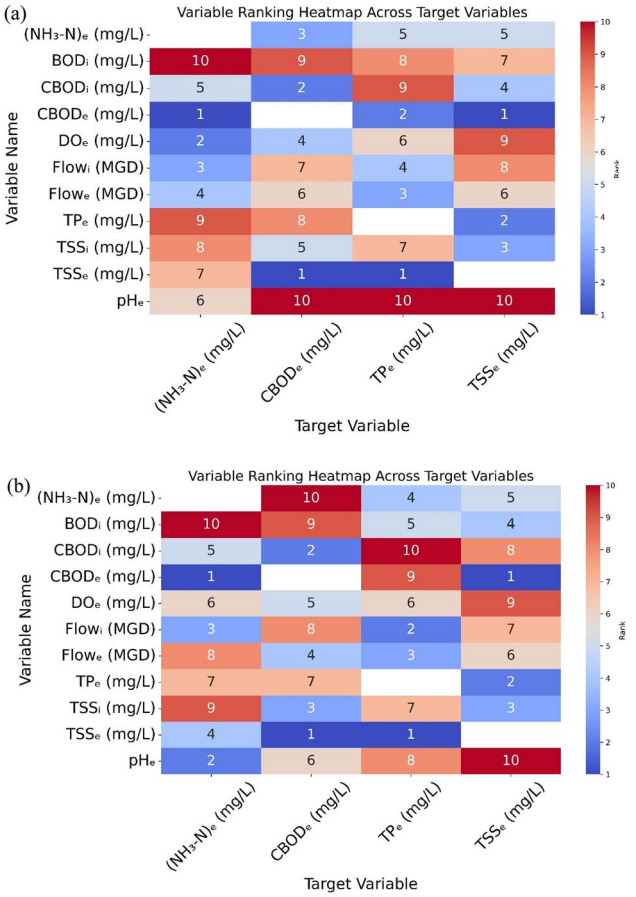

Fig. 16 (a) SHAP Madison for the XGBoost model. (b) LIME Madison for the XGBoost model.

performance ($R^2 = 0.69$) compared to XGBoost ($R^2 = 0.64$). Both RF and XGBoost performed moderately well for $(NH_3–N)_e$ on the validation set and test set. For TSS_e and TP_e, both RF and XGBoost showed comparable training and test performance (Table 8).

4 Discussion

Efficiency in predicting the variables of WWTPs depends on various factors, such as the quality and quantity of generated data and the complexity involved in the process.[23] The predictive performance of the RF and XGBoost models was evaluated for each WWTP using the metrics for training, validation, and test sets. Performance of the models in the study is comparable to those found in similar studies.[24,25]

Overall, both models showed good fits for the training data, with high R^2 and low error metrics. However, their ability to generalize the test data differed significantly between smaller and larger WWTPs. At the Monroe WWTP (small-scale), the test set R^2 values for most target variables were unsatisfactory, indicating that the models struggled to make accurate predictions beyond the training period. The poor test performance at the Monroe WWTP suggests possible overfitting or significant shifts in the data during the test period, which the models failed to capture. A similar trend was observed at the Sheboygan WWTP. In the Madison WWTP (large-scale), the models showed much more robust performance. For example, the predictions of $CBOD_e$ maintained reasonably good accuracy from training to testing, indicating that the models captured stable relationships for CBOD in that plant. Among the four plants, the Milwaukee WWTP (the largest facility) had the strongest predictive results. For BOD_e in Milwaukee, the RF model achieved an R^2 of about 0.97 on training and 0.62 on the test set, while the XGBoost model achieved a comparable test R^2 of 0.67. The smaller difference between the training and testing performances at the Milwaukee plant indicates better generalizations. These results suggest that the models trained on the large-scale data from the Milwaukee plant were able to capture the underlying patterns more reliably, likely due to the larger volume of data and the consistent operation of the facility. The study used basic techniques for handling missing data, such as variable exclusion and mean imputation. These methods do not adequately account for the temporal dependencies, seasonal patterns, or extreme values that are characteristic of environmental time-series data. Future research should adopt time-aware imputation strategies and conduct sensitivity analyses to more effectively capture temporal structures and enhance robustness for regulatory and operational applications.

Table 5 Metrices for the Monroe WWTP

WWTP name	ML model	Target variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
MONROE	RF	BOD_e	Training	0.18	0.12	0.96	0.35	708.65
MONROE	RF	BOD_e	Validation	0.47	0.78	0.58	0.88	0.01
MONROE	RF	BOD_e	Test	0.54	0.63	−0.35	0.79	0.01
MONROE	RF	$CBOD_e$	Training	0.23	0.12	0.78	0.34	674.73
MONROE	RF	$CBOD_e$	Validation	0.40	0.34	0.40	0.58	0.01
MONROE	RF	$CBOD_e$	Test	0.55	0.63	−0.36	0.79	0.01
MONROE	RF	TP_e	Training	0.42	0.44	0.76	0.66	735.80
MONROE	RF	TP_e	Validation	0.46	0.61	0.39	0.78	0.00
MONROE	RF	TP_e	Test	0.53	0.90	0.09	0.95	0.02
MONROE	RF	TSS_e	Training	0.07	0.01	0.63	0.09	698.20
MONROE	RF	TSS_e	Validation	0.10	0.01	0.23	0.12	0.03
MONROE	RF	TSS_e	Test	0.07	0.01	−0.05	0.10	0.02
MONROE	XGBoost	BOD_e	Training	0.31	0.23	0.91	0.48	445.11
MONROE	XGBoost	BOD_e	Validation	0.47	0.71	0.61	0.84	0.00
MONROE	XGBoost	BOD_e	Test	0.52	0.57	−0.23	0.76	0.00
MONROE	XGBoost	$CBOD_e$	Training	0.27	0.15	0.72	0.39	581.58
MONROE	XGBoost	$CBOD_e$	Validation	0.38	0.31	0.45	0.56	0.00
MONROE	XGBoost	$CBOD_e$	Test	0.57	0.67	−0.45	0.82	0.02
MONROE	XGBoost	TP_e	Training	0.41	0.38	0.79	0.61	442.73
MONROE	XGBoost	TP_e	Validation	0.47	0.62	0.39	0.79	0.00
MONROE	XGBoost	TP_e	Test	0.52	0.88	0.11	0.94	0.00
MONROE	XGBoost	TSS_e	Training	0.07	0.01	0.60	0.09	557.57
MONROE	XGBoost	TSS_e	Validation	0.10	0.01	0.22	0.12	0.00
MONROE	XGBoost	TSS_e	Test	0.07	0.01	−0.06	0.10	0.00

Table 6 Metrices for the Sheboygan WWTP

WWTP name	ML model	Target variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
SHEBOYGAN	RF	$(NH_3–N)_e$	Training	0.72	1.22	0.70	1.10	699.83
SHEBOYGAN	RF	$(NH_3–N)_e$	Validation	1.08	2.82	0.25	1.68	0.04
SHEBOYGAN	RF	$(NH_3–N)_e$	Test	0.92	1.50	−0.74	1.22	0.02
SHEBOYGAN	RF	TP_e	Training	0.06	0.01	0.52	0.09	685.73
SHEBOYGAN	RF	TP_e	Validation	0.08	0.02	0.23	0.13	0.02
SHEBOYGAN	RF	TP_e	Test	0.09	0.02	0.18	0.13	0.00
SHEBOYGAN	RF	TSS_e	Training	0.41	0.28	0.90	0.53	702.59
SHEBOYGAN	RF	TSS_e	Validation	0.82	1.22	0.49	1.11	0.02
SHEBOYGAN	RF	TSS_e	Test	1.05	1.90	0.19	1.38	0.01
SHEBOYGAN	XGBoost	$(NH_3–N)_e$	Training	0.55	0.67	0.83	0.82	539.04
SHEBOYGAN	XGBoost	$(NH_3–N)_e$	Validation	1.06	2.68	0.29	1.64	0.00
SHEBOYGAN	XGBoost	$(NH_3–N)_e$	Test	0.92	1.56	−0.82	1.25	0.02
SHEBOYGAN	XGBoost	TP_e	Training	0.07	0.01	0.44	0.09	523.68
SHEBOYGAN	XGBoost	TP_e	Validation	0.08	0.02	0.26	0.13	0.00
SHEBOYGAN	XGBoost	TP_e	Test	0.09	0.02	0.17	0.13	0.02
SHEBOYGAN	XGBoost	TSS_e	Training	0.72	0.86	0.69	0.93	529.73
SHEBOYGAN	XGBoost	TSS_e	Validation	0.84	1.27	0.47	1.13	0.00
SHEBOYGAN	XGBoost	TSS_e	Test	1.05	1.89	0.19	1.38	0.00

5 Conclusion

This study assessed the performance of two ML models, RF and XGBoost, in predicting key effluent quality variables at WWTPs with varying capacities. FS and XAI tools (SHAP and LIME) identified and interpreted the influence of the input variables on the model predictions. The results demonstrated that both the FS and XAI tools provided consistent and interpretable insights into variable importance across varying scales of WWTPs. In particular, the models and XAI analyses consistently highlighted crucial factors such as organic load, nutrient levels, and solids as primary drivers of effluent quality. These tools have proven to be valuable for enhancing the transparency of ML models and identifying key operational parameters that affect effluent outcomes. However, the predictive accuracy of ML models varied significantly among facilities depending on their scale. Large-scale WWTPs (*e.g.*, Madison and Milwaukee) exhibited more stable and reliable model performance, likely owing to their more consistent operations and the availability of larger, less noisy datasets. In contrast, the small-scale WWTPs (Monroe and Sheboygan) experienced

Table 7 Metrices for the Madison WWTP

WWTP name	ML model	Target variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
MADISON	RF	$CBOD_e$	Training	0.13	0.05	0.91	0.22	893.30
MADISON	RF	$CBOD_e$	Validation	0.31	0.27	0.56	0.52	0.03
MADISON	RF	$CBOD_e$	Test	0.53	0.60	0.28	0.78	0.04
MADISON	RF	$(NH_3–N)_e$	Training	0.09	0.03	0.90	0.18	929.43
MADISON	RF	$(NH_3–N)_e$	Validation	0.22	0.14	−0.06	0.37	0.01
MADISON	RF	$(NH_3–N)_e$	Test	0.32	0.24	−0.25	0.49	0.00
MADISON	RF	TP_e	Training	0.03	0.00	0.89	0.05	931.93
MADISON	RF	TP_e	Validation	0.07	0.01	0.71	0.10	0.03
MADISON	RF	TP_e	Test	0.08	0.01	0.32	0.10	0.02
MADISON	RF	TSS_e	Training	0.33	0.41	0.95	0.64	903.66
MADISON	RF	TSS_e	Validation	0.83	2.06	0.80	1.43	0.02
MADISON	RF	TSS_e	Test	0.95	3.65	−0.47	1.91	0.02
MADISON	XGBoost	$CBOD_e$	Training	0.18	0.06	0.88	0.25	779.92
MADISON	XGBoost	$CBOD_e$	Validation	0.33	0.29	0.52	0.54	0.02
MADISON	XGBoost	$CBOD_e$	Test	0.56	0.64	0.23	0.80	0.00
MADISON	XGBoost	$(NH_3–N)_e$	Training	0.16	0.08	0.73	0.29	734.60
MADISON	XGBoost	$(NH_3–N)_e$	Validation	0.22	0.16	−0.25	0.41	0.00
MADISON	XGBoost	$(NH_3–N)_e$	Test	0.29	0.23	−0.21	0.48	0.00
MADISON	XGBoost	TP_e	Training	0.05	0.01	0.79	0.07	718.03
MADISON	XGBoost	TP_e	Validation	0.07	0.01	0.66	0.11	0.02
MADISON	XGBoost	TP_e	Test	0.08	0.01	0.33	0.10	0.00
MADISON	XGBoost	TSS_e	Training	0.31	0.17	0.98	0.41	686.05
MADISON	XGBoost	TSS_e	Validation	0.81	1.82	0.82	1.35	0.01
MADISON	XGBoost	TSS_e	Test	0.95	5.43	−1.19	2.33	0.00

Table 8 Metrices for the Milwaukee WWTP

WWTP name	ML model	Target variable	Set	MAE	MSE	R^2	RMSE	Runtime (s)
MILWAUKEE	RF	BOD_e	Training	0.92	1.60	0.97	1.26	786.96
MILWAUKEE	RF	BOD_e	Validation	2.22	9.93	0.69	3.15	0.04
MILWAUKEE	RF	BOD_e	Test	3.40	18.94	0.62	4.35	0.03
MILWAUKEE	RF	$(NH_3–N)_e$	Training	0.36	0.53	0.80	0.73	791.94
MILWAUKEE	RF	$(NH_3–N)_e$	Validation	0.52	0.80	0.40	0.90	0.01
MILWAUKEE	RF	$(NH_3–N)_e$	Test	1.00	3.08	0.25	1.76	0.02
MILWAUKEE	RF	TP_e	Training	0.11	0.04	0.63	0.20	793.77
MILWAUKEE	RF	TP_e	Validation	0.14	0.04	0.38	0.20	0.00
MILWAUKEE	RF	TP_e	Test	0.16	0.04	0.41	0.21	0.02
MILWAUKEE	RF	TSS_e	Training	0.87	1.96	0.95	1.40	789.63
MILWAUKEE	RF	TSS_e	Validation	1.90	7.85	0.61	2.80	0.01
MILWAUKEE	RF	TSS_e	Test	1.94	6.91	0.69	2.63	0.00
MILWAUKEE	XGBoost	BOD_e	Training	1.76	5.35	0.89	2.31	572.61
MILWAUKEE	XGBoost	BOD_e	Validation	2.17	9.79	0.70	3.13	0.00
MILWAUKEE	XGBoost	BOD_e	Test	3.20	16.40	0.67	4.05	0.02
MILWAUKEE	XGBoost	$(NH_3–N)_e$	Training	0.46	0.63	0.76	0.80	489.31
MILWAUKEE	XGBoost	$(NH_3–N)_e$	Validation	0.50	0.75	0.43	0.87	0.01
MILWAUKEE	XGBoost	$(NH_3–N)_e$	Test	1.03	3.20	0.22	1.79	0.00
MILWAUKEE	XGBoost	TP_e	Training	0.12	0.04	0.65	0.20	561.53
MILWAUKEE	XGBoost	TP_e	Validation	0.14	0.04	0.35	0.20	0.01
MILWAUKEE	XGBoost	TP_e	Test	0.16	0.04	0.41	0.21	0.00
MILWAUKEE	XGBoost	TSS_e	Training	0.78	1.11	0.97	1.06	506.23
MILWAUKEE	XGBoost	TSS_e	Validation	1.85	7.08	0.65	2.66	0.02
MILWAUKEE	XGBoost	TSS_e	Test	2.09	7.89	0.64	2.81	0.00

reduced prediction accuracy, which could be attributed to limited data, higher relative noise, greater variability in influent characteristics, and less consistent operational practices. Future research should focus on developing scalable, data-efficient ML frameworks that can be adapted to different plant sizes. Overall, such initiatives are important for advancing the practical integration of interpretable AI into real-world wastewater

management, ultimately contributing to more efficient and sustainable WWTP operation.

Conflicts of interest

The authors declare that they have no competing interests that could influence the work reported in this paper.

List of abbreviations

NH_3–N	Ammonia nitrogen
BOD	Biochemical oxygen demand
COD	Chemical oxygen demand
DO	Dissolved oxygen
XAI	Explainable artificial intelligence
XGBoost	eXtreme gradient boosting
FS	Feature selection
LASSO	Least absolute shrinkage and selection operator
LIME	Local interpretable model-agnostic explanations
MI	Mutual information
MGD	Million gallons per day
ML	Machine learning
PC	Pearson correlation
PCA	Principal component analysis
RF	Random forest
SHAP	SHapley additive exPlanations
TKN	Total Kjeldahl nitrogen
TP	Total phosphorus
TSS	Total suspended solids
TPD	Tons per day
WWTP	Wastewater treatment plant
WWTF	Wastewater treatment facility

Data availability

Data information available in the article. More data will be available from the author upon reasonable request.

Acknowledgements

The authors acknowledge the Wisconsin Department of Natural Resources for providing data of wastewater treatment plants for the study.

References

1 R. Li, K. Feng, T. An, P. Cheng, L. Wei, Z. Zhao, X. Xu and L. Zhu, Enhanced insights into effluent prediction in wastewater treatment plants: Comprehensive deep learning model explanation based on shap, *ACS ES&T Water*, 2024, **4**(4), 1904–1915, DOI: 10.1021/acsestwater.4c00040.

2 M. El-Rawy, M. K. Abd-Ellah, H. Fathi and A. K. Ahmed, Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques, *J. Water Proc. Eng.*, 2021, **44**, 102380, DOI: 10.1016/j.jwpe.2021.102380.

3 D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind and N. Souihi, A machine learning framework to improve effluent quality control in wastewater treatment plants, *Sci. Total Environ.*, 2021, **784**, 147138, DOI: 10.1016/j.scitotenv.2021.147138.

4 X. Zhang and C. A. Liu, Model averaging prediction by K-fold cross-validation, *J. Econom.*, 2023, **235**(1), 280–301, DOI: 10.1016/j.jeconom.2022.04.

5 J. J. Zhu, S. Borzooei, J. Sun and Z. J. Ren, Deep learning optimization for soft sensing of hard-to-measure wastewater key variables, *ACS ES&T Eng.*, 2022, **2**(7), 1341–1355, DOI: 10.1021/acsestengg.1c00469.

6 E. Aghdam, S. R. Mohandes, P. Manu, C. Cheung, A. Yunusa-Kaltungo and T. Zayed, Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques, *J. Clean. Prod.*, 2023, **405**, 137019, DOI: 10.1016/j.jclepro.2023.137019.

7 H. Y. Shyu, C. J. Castro, R. A. Bair, Q. Lu and D. H. Yeh, Development of a soft sensor using machine learning algorithms for predicting the water quality of an onsite wastewater treatment system, *ACS Environ. Au*, 2023, **3**(5), 308–318, DOI: 10.1021/acsenvironau.2c00072.

8 X. Wei, J. Yu, Y. Tian, Y. Ben, Z. Cai and C. Zheng, Comparative performance of three machine learning models in predicting influent flow rates and nutrient loads at wastewater treatment plants, *ACS ES&T Water*, 2023, **4**(3), 1024–1035, DOI: 10.1021/acsestwater.3c00155.

9 J. Yu, Y. Tian, H. Jing, T. Sun, X. Wang, C. B. Andrews and C. Zheng, Predicting regional wastewater treatment plant discharges using machine learning and population migration big data, *ACS ES&T Water*, 2023, **3**(5), 1314–1328, DOI: 10.1021/acsestwater.2c00639.

10 Y. Xu, Z. Wang, S. Nairat, J. Zhou and Z. He, Artificial intelligence-assisted prediction of effluent phosphorus in a full-scale wastewater treatment plant with missing phosphorus input and removal data, *ACS ES&T Water*, 2023, **4**(3), 880–889, DOI: 10.1021/acsestwater.2c00517.

11 M. A. Cechinel, J. Neves, J. V. Fuck, R. C. de Andrade, N. Spogis, H. G. Riella, N. Padoin and C. Soares, Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis, *J. Water Proc. Eng.*, 2024, **58**, 104758, DOI: 10.1016/j.jwpe.2023.104758.

12 J. Park, W. H. Lee, K. T. Kim, C. Y. Park, S. Lee and T. Y. Heo, Interpretation of ensemble learning to predict water quality using explainable artificial intelligence, *Sci. Total Environ.*, 2022, **832**, 155070, DOI: 10.1016/j.scitotenv.2022.155070.

13 Y. Hu, R. Wei, K. Yu, Z. Liu, Q. Zhou, M. Zhang, C. Wang, L. Zhang, G. Liu and S. Qu, Exploring sludge yield patterns through interpretable machine learning models in China's municipal wastewater treatment plants, *Resour. Conserv. Recycl.*, 2024, **204**, 107467, DOI: 10.1016/j.resconrec.2024.107467.

14 S. Shao, D. Fu, T. Yang, H. Mu, Q. Gao and Y. Zhang, Analysis of machine learning models for wastewater treatment plant sludge output prediction, *Sustainability*, 2023, **15**(18), 13380, DOI: 10.3390/su151813380.

15 S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, DOI: 10.48550/arXiv.1705.07874.
16 S. M. Lundberg, G. G. Erion, and S. I. Lee, Consistent individualized feature attribution for tree ensembles, *arXiv*, 2018 Feb 12, preprint, arXiv:1802.03888, DOI: 10.48550/arXiv.1802.03888.
17 M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016 Aug 13 pp. 1135–1144. DOI: 10.48550/arXiv.1602.04938.
18 H. Tyralis, G. Papacharalampous and A. Langousis, A brief review of random forests for water scientists and practitioners and their recent history in water resources, *Water*, 2019, **11**(5), 910, DOI: 10.3390/w11050910.
19 M. Jiang, J. Wang, L. Hu and Z. He, Random forest clustering for discrete sequences, *Pattern Recognit. Lett.*, 2023, **174**, 145–151, DOI: 10.1016/j.patrec.2023.09.001.
20 O. Szomolányi and A. Clement, Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters, *GEM. Int. J. Geomath.*, 2023, **14**(1), 20, DOI: 10.1007/s13137-023-00229-6.
21 Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang and X. Liang, An improved random forest based on the classification accuracy and correlation measurement of decision trees, *Expert Syst. Appl.*, 2024, **237**, 121549, DOI: 10.1016/j.eswa.2023.121549.
22 S. Zhang, H. Wang and A. A. Keller, Novel machine learning-based energy consumption model of wastewater treatment plants, *ACS ES&T Water*, 2021, **1**(12), 2531–2540, DOI: 10.1021/acsestwater.1c00283.
23 A. G. Sheik, A. Kumar, C. S. Srungavarapu, M. Azari, S. R. Ambati, F. Bux and A. K. Patan, Insights into the application of explainable artificial intelligence for biological wastewater treatment plants: Updates and perspectives, *Eng. Appl. Artif. Intell.*, 2025, **144**, 110132, DOI: 10.1016/j.engappai.2025.110132.
24 L. Bo-Qi, Z. Ding-Jie, Z. Yang and S. Long-Yu, Comparative analysis of supervised learning models for effluent quality prediction in wastewater treatment plants, *PLoS One*, 2025, **20**(6), e0325234, DOI: 10.1371/journal.pone.0325234.
25 F. B. Nasir and J. Li, Comparative analysis of machine learning models and explainable artificial intelligence for predicting wastewater treatment plant variables, *Adv. Environ. Eng. Res.*, 2024, **5**(4), 1–23, DOI: 10.21926/aeer.2404020.