Journal Pre-proof

Development of a wide-range soft sensor for predicting wastewater BOD_5 using an eXtreme gradient boosting (XGBoost) machine

P.M.L. Ching, X. Zou, Di Wu, R.H.Y. So, G.H. Chen

PII: S0013-9351(22)00280-8

DOI: https://doi.org/10.1016/j.envres.2022.112953

Reference: YENRS 112953

To appear in: *Environmental Research*

Received Date: 29 October 2021

Revised Date: 6 February 2022

Accepted Date: 10 February 2022

Please cite this article as: Ching, P.M.L., Zou, X., Wu, D., So, R.H.Y., Chen, G.H., Development of a wide-range soft sensor for predicting wastewater BOD_5 using an eXtreme gradient boosting (XGBoost) machine, *Environmental Research* (2022), doi: https://doi.org/10.1016/j.envres.2022.112953.

Development of a wide-range soft sensor for predicting wastewater BOD_5 using an eXtreme gradient boosting (XGBoost) machine

P.M.L. Ching [a], X. Zou [b], Di Wu [b,c,d*], R. H.Y. So [e], G.H. Chen [b]

a. Bioengineering Graduate Program, Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.
b. Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.
c. Center for Environmental and Energy Research, Ghent University Global Campus, Republic of Korea.
d. Department of Green Chemistry and Technology, Ghent University, Belgium.
e. Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

* Corresponding author:
Di WU, Center for Environmental and Energy Research, , Ghent University Global Campus, Republic of Korea. (Email: di.wu@ghent.ac.kr)

Abstract

In wastewater monitoring, detecting extremely high pollutant concentrations is necessary to properly calibrate the treatment process. However, existing hardware sensors have a limited linear range which may fail to measure extremely high levels of pollutants; and likewise, the conventional “soft” model sensors are not suitable for the highly-skewed data distributions either. This study developed a new soft sensor by using eXtreme Gradient Boosting (XGBoost) machine learning to ‘measure’ the wastewater organics (in terms of 5-day biochemical oxygen demand (BOD5)). The soft sensor was tested on influent and effluent BOD5 of two different wastewater treatment plants to validate the results. The model results showed that XGBoost can detect these extreme values better than conventional soft sensors. This new soft sensor can function using a sparse input matrix via XGBoost’s sparsity awareness algorithm - which can address the limitation of the conventional soft sensor with the fallibility of supporting hardware sensors even.

Keywords: Soft sensor, machine learning, XGBoost, real-time monitoring, biochemical oxygen demand (BOD)

1. Introduction

Online monitoring is an important prerequisite for advancements in wastewater treatment. Real-time information allows the plant to implement more cost-efficiently and gives evidence that the quality regulations are consistently being met. A conventional online monitoring system for relevant wastewater parameters (e.g. chemical oxygen demand (COD), ammonia concentration) would be composed of hardware sensors. However, the existing hardware sensors for these parameters have a limited useful lifespan due to the harsh conditions of wastewater. The accumulation of sludge and precipitates on the sensor lowers its accuracy over time, and necessitates frequent maintenance (Haimi et al., 2013). The sensor itself loses its functionality over time, such as the dissolution of Ag/AgCl layers observed in electrode-based sensors (Hill et al., 2020) or the degradation of the microorganism culture used in biosensors (Raud et al., 2012). Besides, there is no mature sensor product for measuring five-day biochemical oxygen demand (BOD_5) to reflect the biodegradable organics content in the wastewater. To solve this problem, one good option is to use a machine learning-based soft sensor model, which estimates parameter values from other hardware sensors using machine learning. In this way, soft sensors can facilitate real-time monitoring by avoiding the delays or missing data resulting from frequent maintenance; manual measurement. However, the accuracy of the soft sensor is still dependent on the (1) choice of hardware sensors used as its basis for estimation; (2) the volume and range of data, and (3) the appropriateness of the machine learning model used in estimation.

In choosing the hardware basis for the machine learning-based soft sensor, the ideal choice is to choose simple and stable sensors (e.g. pH, conductivity). Yet, the majority of soft-sensor studies add complex sensors (e.g. chemical oxygen demand (COD), NH_4) to enhance accuracy. When there is a large quantity of potential soft sensors, parameter selection techniques can be employed to reduce the number of model inputs. These techniques aim to identify the input parameters sharing the strongest relationship with the output parameter(s) (Zhu et al., 2017). In addition, the performance of the soft sensor may be improved by the removal of some inputs, as collinearity between the input variables may promote overfitting (Asante-Okyere et al., 2020).

It should note that datasets used in soft sensor development vary in size (Ye et al., 2020). While there is no defined minimum for the size of the dataset, a larger dataset is preferred for higher generalizability. The volume and range of wastewater datasets are limited by sensor degradation and infrequent sampling, resulting in missing sensor readings in the dataset. These missing values can be filled in using a statistic (e.g. mean, median), or using a statistical method to impute the missing values (Wu et al., 2008). While these methods can produce additional samples to the dataset, samples with missing parameters may increase the uncertainty in the model, and skew the estimations of the soft sensor (Li et al., 2020).

Although any mathematical model can be applied in soft sensor development, machine learning approaches are preferred in recent studies. One reason is that these utilize the existing wastewater treatment databases, and produce new insights without additional experimentation (Asami et al., 2021; Qiu et al., 2021). Using machine learning, mathematical relationships are automatically ‘learned’ instead of manually developed based on theoretical knowledge, and this may be more efficient in some cases. Some examples include applications in predicting the concentration of novel pollutants and pathogens of interest (Abdeldayem et al., 2022). It can also capture a broad range of operating conditions, whereas traditional mechanistic modelling is typically limited to steady state analysis (Wang et al., 2021).

Currently, the most popular machine learning models applied in wastewater treatment are artificial neural networks (ANN) and support vector machines (SVM) (Ye et al., 2020). The ANN model is composed of several layers of node equations, which form a highly nonlinear relationship. Its primary advantage is its ability to present complex underlying relationships between variables, and has improved the accuracy of predicting several key wastewater parameters (Matheri et al., 2021). However, the disadvantage of this complex nonlinear structure is that ANN models have a tendency to overfit to the dataset used for training, and thus require a large number of samples in order for the trained model to be generalizable (Ye et al., 2020). Some modifications of the classical neural network have been proposed: Zhu et al. (2017) integrated the radial basis function in an ANN model for predicting total phosphorus (TP), as this function is associated with enhanced generalizability even with smaller datasets. Cong and Yu (2018) used wavelet tranforms in an ANN model, to prevent it from overfitting to noise in the training set.

On the other hand, the advantage of SVM is its generalizability. Specifically, the objective function used in determining the optimal parameters of an SVM model seeks to maximize generalizability (Liu & Xie, 2020; Jiang et al., 2020). Because of this, SVM can be used even with relatively small datasets, which can be important when analysing novel processes and technologies (Hosseinzadeh et al., 2022; Moufid et al., 2021). The disadvantage of SVM is that its generalizability objective may lead the model to overfit to the dominant condition in the dataset (Jaramillo et al., 2018). A soft sensor based on SVM may thus fail in accurately measuring extreme values in the statistical distribution of a parameter, or in differentiating between normal and abnormal operating conditions.

It should also be noted that, aside from the recurring problems in terms of missing sensor readings and noise, data on water treatment is characterized by skewed and non-normal distributions. This may render approaches that emphasize generalizability unsuitable for modeling. Ensemble models are a non-parametric modeling approach that makes estimations using the average of a large number of simple models (Sharafati et al., 2020). Each model within the ensemble may represent a characteristic of the distribution of the predicted parameter. This enhances the robustness of the model while allowing it to model non-normal variables.

In this study, extreme gradient boosting (XGBoost), a new ensemble method, is proposed in soft sensor development for BOD_5 analysis. This method was selected because of its robustness and ability to model non-normal variables. In addition, XGBoost includes a sparsity-awareness algorithm that allows it to train using samples with missing sensor readings. Operating as a soft sensor, XGBoost can also make inferences from inputs with missing parameters, which is faster compared to using a separate model to estimate the missing values. This study used two case studies of wastewater treatment plants to identify the dataset characteristics. Finally, the comparisons with other popular machine learning techniques were drew to verify the merits of XGBoost machine learning.

2. Materials and Methods

The framework of developing a new machine learning-based soft sensor is illustrated in Figure 1. The details were described as 1) the data source (two case studies for BOD5 soft sensor development); 2) the general steps involved in the development of the soft sensor; 3) the method of developing Modified Partial Least Squares used in selecting supporting sensors for the soft sensor; 4) the methods for missing sensor reading in the dataset; 5) the development approach for the proposed XGBoost soft sensor and other potential soft sensor development methods for comparison.

2.1. Data Source

The proposed soft sensor development approach was demonstrated through two case studies: Case 1, the public wastewater treatment dataset published by the UCI Machine Learning Repository (Dua & Graff, 2019); and Case 2, a dataset collected from a wastewater treatment plant in Hong Kong (see supplementary information Figure S1). The data used in this study came from manual measurements to allow easier comparatability of results, avoid variance resulting from the choice of sensor and sensor performance. Thus, we can assume that all model input data is accurate. Although in the context of real operation, input data collected from sensors would suffer from noise and interference, there is already existing work on mathematical models that address these problems (see Ba-Alawi et al., 2021; Fan et al., 2020; Wang et al., 2020).

The case study based on the dataset of the UCI Machine Learning Repository (Case 1) describes the treatment of urban wastewater in an unnamed plant. It contains 527 daily readings, with missing data found in 84 samples in the influent and 72 samples in the effluent. The Hong Kong dataset (Case-2) was collected from January 2013 to December 2018. It contains 2,189 daily samples, with missing data in 1,576 samples in the influent and 1,575 samples in the effluent. The supplementary information for this study contains more details on the statistical properties of this dataset, namely its range, skewness, and the number of missing readings for each parameter in the dataset (Table S1 and Table S2). Generally speaking, both datasets are highly skewed, with a higher level of skewness among effluent parameters. Skewness measures the tendency of samples in the dataset to cluster towards lower (positive skew) or higher values (negative skew). High levels of skewness indicate that the dataset is not normally distributed, which is a key assumption in most data-driven models. It is also notable that the distribution of effluent BOD (BOD_{eff}) is more skewed compared to influent BOD (BOD_{inf}), while BOD_{inf} has a higher variance compared to BOD_{eff}.

2.2. General Soft Sensor Development

This study developed soft sensors for BOD_{inf} and effluent BOD_{eff} for the two cases described in the previous section. For Case 1 (using data from the UCI repository), BOD_{inf} was modeled using other influent parameters as supporting sensors; and likewise, BOD_{eff} was modeled only using other effluent parameters as supporting sensors. Case 2 differs slightly as it includes ambient temperature (represented by temperature measured at the reactor, $Temp_{Reac}$) as a potential supporting sensor. This was included as a supporting sensor for BOD_{inf}, representing the potential for organic degradability before the treatment process.

There are multiple potential supporting sensors, and some information is redundant across the different hardware sensors (e.g., NH_3-N and NO_2-N). To identify the best-supporting sensors to use as the basis for the soft sensor, the study incrementally added supporting sensors as inputs to the soft sensor model and evaluated the change in performance as the output. The order of adding supporting sensors to the model was based on a modified Partial Least Squares approach (the details see next section). Limiting the number of input variables also limits the potential for the soft sensor to fail; its inputs are other supporting sensors, therefore its performance is dependent on its supporting sensors.

Given that a significant portion of the samples in both cases include missing parameters, the study considered three methods of handling the missing values: (i) removing the samples with missing values; (ii) using *k*-nearest neighbors (kNN) to fill in the missing values; and (iii) using the sparsity-awareness algorithm of XGBoost to train a model using samples with missing parameters. The disadvantage of removing the samples with missing values is that it significantly reduces the size of the dataset. Depending on the distribution and noisiness of the data, a smaller dataset could prevent the machine learning models from representing the complete and general behavior of BOD_5. Conversely, using a model to impute the missing values could also worsen soft sensor performance through the errors in the imputed values. For most estimation models (e.g. ANN and SVM), it is necessary to have a separate method such as *k*-nearest neighbors for handling the missing values. But XGBoost differs from these methods as it has a built-in algorithm to incorporate the samples with missing values in the training process. This is one of the key advantages of XGBoost and will be described further in the following section. It will be compared with the aforementioned methods (i) and (ii) of handling missing values.

Performance analysis was based on root mean square error (RMSE, see Eq. 1) in units of mg/L. This reflects the actual deviation of the soft sensor reading from the 'real' value, based on laboratory tests. It also reflects the effect of differences in dataset characteristics such as the minimum, maximum and kurtosis on the magnitude of error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{N}} \qquad \text{Eq. 1}$$

The model results were validated using 10-fold cross-validation. This approach divides the training set into 10 sets with no overlap. Each of the ten sets represents a test set for the model, where it will be trained using all the other samples not included in the test set (as shown in Figure 1b). The purpose of this method is to determine the general performance of each method using different datasets. This also allows for a comparison of the consistency of model performance.

2.3. Modified Partial Least Squares for Supporting Sensor Selection

PLS is a form of linear regression that maximizes the covariance between model inputs and the predicted output. It has been used in related studies for soft sensor development because it simultaneously maximizes the variance in the inputs, and the correlation between the model's inputs and outputs (Zhu et al., 2017; Qin et al., 2012). This means that the strongest supporting sensors with the least redundancy will be selected as inputs. Although there are various ways of interpreting the results of PLS for input variable selection, one of the most reliable and straightforward ways is by measuring the absolute value of the PLS regression coefficients (Mehmood et al., 2020). The greater the value of the regression coefficient, the more significant its corresponding input variable is based on PLS regression.

However, in the context of wastewater treatment, the effectiveness of the supporting sensor has to be weighed with respect to the practicality of selecting this particular soft sensor. Simpler sensors (i.e., pH, conductivity, temperature, flow rate) may be easier to maintain or replace. Using these sensors as supporting sensors would make the proposed soft sensor more reliable, although these variables may not have the strongest correlation with BOD_5. This study applied a modified PLS approach in selecting the supporting sensors. Several versions of the soft sensor were built using different sets of supporting sensors as inputs to the model. There lessen the number of combinations that would have to be tested, the study used the modified PLS approach to guide the selection process. This approach prioritizes the simpler sensors as inputs for the initial model. Then, sensors are added incrementally in the order of their PLS regression coefficients. The optimal soft sensor design was selected based on the model which resulted in the lowest and most consistent RMSE.

2.4. Methods for Missing Sensor Readings in the Dataset

In general, having a larger dataset is preferred as it should help enhance the generalizability of the model. Several studies have attempted to fill in missing values in a dataset to enhance model performance. Among these studies, *k*-Nearest neighbors (kNN) has emerged as a standard for determining the missing values. To fill in the missing parameters of a sample, kNN uses the weighted average of samples with the highest similarity based on the available parameters for that sample (Qi et al., 2021). In this case, the similarity is based on a distance measure such as Euclidean distance (Alfeilat et al., 2019).

XGBoost has its algorithm for addressing the missing values. This algorithm is known as a sparsity awareness split-finding algorithm, referring to the dataset splitting involved in determining the optimal structure for the XGBoost model. The sparsity-awareness algorithm applies for any commonly recurring value (e.g., NaN, 0). In the context of wastewater treatment, this can apply to missing values and very low levels of effluent pollutants below the threshold for recording.

The sparsity awareness algorithm differs from kNN, as the former is a method that is integrated in model training, while the latter is completely independent of soft sensor development. This study sought to identify the best method for handling the missing sensor readings in Cases 1 and 2, according to the characteristics of these respective datasets. The study compared three methods of handling the missing values by developing models using (1) a dataset containing no samples with missing readings; (2) a dataset where the missing sensor readings were filled in using kNN; and (3) the sparsity-awareness split-finding algorithm to train using a dataset with missing values.

2.5. XGBoost and Comparison with other Soft Sensor Models

XGBoost is an ensemble method, meaning that it is a collection of weaker models, as opposed to being a single, highly complex model (i.e., ANN, SVM) (Chen & Guestrin, 2016). Specifically, it is composed of regression trees (f_k) (Eq. 2). The structure of the regression tree is represented by its leaves, which correspond to a numerical weight (*w*). Each sample is assigned to a set of leaves based on the values of its input variables. The model's estimated output for that sample is obtained by adding the sum of the leaves assigned to that sample for each regression tree (visualized in Figure 2-b).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad \text{Eq. 2}$$

These regression trees are introduced additively to the ensemble (as f_t for iteration *t*), such that each new regression tree minimizes the learning objective (eq. 3). This is different from singular models, which tend to have a pre-defined structure and are optimized in a Euclidean space.

$$\mathcal{L}^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{\,t-1}) + f_t(x_i)) + \Omega(f_t) \qquad \text{Eq. 3}$$

For benchmarking, XGBoost was compared with an ANN model and an SVM model. The ANN model was based on Zhu et al. (2017), which is composed of one hidden layer with 10 neurons, using the radial basis function as its activation function. The SVM model was based on Zaghloul et al. (2020), which used a Gaussian kernel function.

3. Results

3.1. Selection of Supporting Sensors

A modified PLS approach was used to identify the best-supporting sensors for the proposed BOD_5 soft sensors. In typical implementations of PLS for parameter selection, the PLS regression coefficients are used as the basis for selection. The regression coefficients obtained from building models for BOD_5 using data from Cases 1 and 2 are presented in Figure 2. In all cases, the supporting sensor with the highest PLS regression coefficient was a complex sensor, i.e. COD and total suspended solids (TSS). On the other hand, simpler sensors, i.e. $Temp_{Reac}$, flow rate (Q), conductivity (Cond), and pH, ranked lower in the order of recommended supporting sensors.

Through the modified PLS approach, the potential of using these simpler sensors to build the soft sensor was explored. The study compared the difference in RMSE resulting from using different sets of supporting sensors (see Figure 3). The initial soft sensor model for each case was built using only simple sensors. Specifically, for Case 1, these simple sensors were pH_{inf}, flowrate (Q_{inf}) and conductivity ($Cond_{inf}$) for BOD_{inf}; and pH_{eff} and $Cond_{eff}$ for BOD_{eff}. For Case 2, simple sensors refer to Q_{inf}, $Temp_{Reac}$ and pH_{inf} for BOD_{inf}, and Q_{eff} and pH_{eff} for BOD_{eff}. Supporting sensors were incrementally added based on the order of their PLS regression coefficients until all potential supporting sensors were exhausted. The sensitivity analysis showed that using a large number of complex supporting sensors did not improve accuracy. Based on these results, we found that: A soft sensor for BOD_5 could be built using simple sensors and one complex sensor.

The results for Case 1 showed that a soft sensor for BOD_{inf} could be developed based on simple sensors and COD_{inf}. There was a significant decrease in RMSE from the model using only simple sensors, to the model using simple sensors and COD_{inf}.

However, the improvement in RMSE became minimal for additional supporting sensors. As such, the proposed soft sensor for BOD_{inf} in Case 1 is based on the simple sensors and COD_{inf}. The soft sensor for BOD_{eff} was the only case where COD did not have the highest PLS regression coefficient. In this specific case, the coefficient for COD_{eff} is lower than TSS_{eff} and $Sediments_{eff}$, although it is notable that there was a slight decrease in RMSE when COD_{eff} is added as a soft sensor along with TSS, sediments and the simple sensors. It is also notable that soft sensor performance worsens both in terms of average performance and consistency between the soft sensor with 3 supporting sensors (i.e., simple sensors and TSS_{eff}), and that with 4 supporting sensors (i.e., simple sensors, TSS_{eff} and $Sediments_{eff}$). This suggests that having more supporting sensors may generally even worsen performance, potentially due to noise or multicollinearity. Thus, the proposed BOD_{eff} sensor for Case 1 takes the version of the model using 3 supporting sensors. Based on the same reasoning, the proposed supporting sensors for BOD_{inf} are the simple sensors and COD_{inf}. For BOD_{eff}, the proposed supporting sensors are the simple sensors, COD_{eff}, and othophosphates ($OP\text{-}P_{eff}$). This is the only case where further improvement was observed from adding more than one complex sensor as an input. This shows that a soft sensor can be developed with relatively few and accessible supporting sensors.

3.2. Comparing Methods for Missing Sensor Readings

The wastewater datasets contain a significant number of samples with missing readings, owing to sensor failure or manual measurement of the parameters. This is a common problem in data-driven modeling, particularly in water treatment (Ma et al., 2020). This study compared different approaches for the missing sensor readings in the dataset. Specifically, the study compared (i) the case where only samples without missing readings were used in training, (ii) the case where kNN was used to fill in the missing readings, and (iii) the case where a dataset with missing values was used to train the XGBoost model, to be processed by its sparsity awareness split-finding algorithm. For Case 1, including the missing sensor readings in training generally improved performance (see Table 1). The sparsity awareness split-finding algorithm of XGBoost resulted in the highest accuracy (i.e., lowest RMSE) for both BOD_{inf} and BOD_{eff}, although the XGBoost model using missing values was obtained from *kNN* had the highest consistency. Meanwhile, for Case 2, the results of the models were in favor of removing the samples with missing readings from the dataset, for both BOD_{inf} and BOD_{eff}.

The difference between the performance of the methods in Cases 1 and 2 was attributed to the volume of missing values in each case. Specifically, only 15.9% of influent samples and 13.7% of effluent samples contained missing sensor readings. This is small compared to Case 2, with 72.0% of influent samples and 71.9% of effluent samples containing missing values. In addition, it was notable that data in Case 1 tended to contain fewer parameters with missing readings in each sample. In comparison, there samples in Case 2 with missing readings tended to contain several missing sensor readings (see supplementary information Table S3). Because of this, kNN and the sparsity awareness algorithm had less inputs for handling the missing values, resulting in poorer estimations.

These characteristics of Case 1 make it more viable to include the samples with missing readings in training. This illustrates that there is a threshold for uncertainty in the samples included in the training set. While including some of these samples with missing readings can improve performance, adding a large number of the samples, or using samples with too many missing parameter values, worsen performance. Related studies concerning unlabelled datasets have also encountered this problem, necessitating the selective inclusion of samples for model development (Li et al., 2020).

As a method for handling missing values, the results demonstrated that the sparsity awareness algorithm of XGBoost was at least equal to kNN. This makes the estimation process of the soft sensor model more efficient, as the algorithm can directly process the samples with missing readings, whereas kNN results in a two-step approach of imputation and estimation. The significance of the sparsity awareness algorithm method is that it assigns a direction for any sparsely occurring value, whereas other regression tree ensembles would either not be able to use a missing value as an input, or would treat the recurring value as any continuous value. This method is helpful both in training and operating the soft sensor, as the algorithm may allow the soft sensor to continue functioning even if some of the supporting sensors fail.

3.3. Comparison of Soft Sensor Models

XGBoost differs from other implementations of regression tree ensembles as its learning objective is penalized with the term $\Omega(f_k)$. This limits the complexity of the regression trees, preventing overfitting. The learning objective is used to determine the optimal structure of regression trees, the assignment of leaves for each sample, and the weighted value of the leaves. The performance of XGBoost was compared with more popular methods in soft sensor development, i.e. ANN and SVM. First, a comparison of observed (laboratory-tested) and estimated (soft sensor) values was conducted to identify the source of error in the models in relation to RMSE. Results to demonstrate this analysis in Case 1 is shown in Figure 4. For BOD_{inf}, the RMSE of XGBoost was inferior to both ANN and SVM; and for BOD_{eff}, the RMSE of XGBoost was superior to both models. The stark difference in performance indicates the dataset characteristics where each model would be more appropriate. Specifically, a continuous regression approach seems to be more effective for the high-variance case of BOD_{inf}, while the ensemble learning approach is compatible with the high skewness BOD_{eff}.

Figure 5 shows the results for Case 2. In this case, XGBoost ranks second to ANN in terms of performance for BOD_{inf}. This supports the notion that continuous regression is more appropriate for BOD_{inf}. However, more cases would be needed to identify the difference between Cases 1 and 2 that allowed XGBoost to have an advantage over SVM. On the other hand, the results for BOD_{eff} show that XGBoost had the lowest RMSE in this case, which supports the conclusions drawn from Case 1 on the effectiveness of XGBoost on skewed and non-normal distributions. In spite of this, it was found that all three models were challenged when it came to estimating extremely high and extremely low values. In particular, given the high skewness of BOD_{eff}, there were significantly fewer samples to represent extremely high values of BOD_{eff} in the dataset, which can account for poor performance. The visual comparison of observed and estimated values shows that XGBoost is superior in estimating some of these low-frequency cases.

The findings based on Cases 1 and 2 analysis were validated with 10-fold cross-validation. This means that each model was tested using 10 different test sets, in cases where these samples were not included in training the model. The results of cross-validation for Case 1 are presented in Table 2a. For BOD_{inf}, the results confirmed that continuous regression was superior for this case; and likewise, the results for BOD_{eff} confirmed that XGBoost was advantageous for skewed distributions. In addition, it can also be observed that while XGBoost did not always have the lowest RMSE, it consistently had the lowest standard deviation, which supports the notion that the residual errors were not higher for extreme values.

The results of cross-validation for Case 2 (shown in Table 2b) confirm that both ANN and SVM are superior to XGBoost for BOD_{inf}. Previously, the results of a single test set presented in Figure 5 showed that XGBoost was more accurate than SVM for at least one case. Although the average RMSE from cross validation seemed to converge (between 67.49 – 67.79 mg/L), some variation on a case-to-case basis can be expected given that the SVM model had the highest standard deviation based on cross validation. A model can achieve the highest accuracy for a certain fold if it is the most suitable model for the characteristics of the data in that fold. This was demonstrated by the results of BOD_{eff} , which supported the appropriateness of XGBoost for skewed datasets. Specifically, XGBoost had the highest accuracy and consistency among the three models.

In general, XGBoost has some advantages over singular models in terms of robustness and scalability. The characteristic of being an ensemble of models is intended to allow each model to capture some aspect of the data structure. Being composed of several weaker (less complex) models prevents the likelihood of overfitting, even for smaller datasets. Together, the ensemble characteristic and the additive model development process prevent convergence to local minima, a tendency of singular models. These characteristics can also help XGBoost to cover a larger space of potential solutions, resulting in a higher potential for good potential.

4. Discussion

This study developed soft sensors for BOD_5 for two different wastewater treatment plants. In both cases, the supporting sensors used were a combination of relatively simple sensors (e.g., pH, temperature, flow rate) and minimal complex sensors (i.e., COD and/or nutrients). This is a common approach in most soft sensor development studies, as simpler sensors may be more stable or easily replaced, while the complex sensors may share stronger correlations with BOD_5. In a literature study, Xiao et al. (2019) predicted effluent BOD_5 from sensors for pH, effluent ammonia, influent TSS, and influent COD using multivariate regression models. The soft sensor designed by Ebrahimi et al. (2017) predicted effluent BOD_5 from influent TSS, influent total phosphorus (TP) and influent total nitrogen (TN), specifically using the interactions between these parameters in the soft sensor model. Similarly, the supporting sensors for the soft sensor developed by Liu (2017) include influent TSS, effluent ammonia, and simpler sensors such as dissolved oxygen, oxidation-reduction potential, and flow rate.

Notably, most studies used sensors for nutrients (e.g., TN, TP, ammonia), COD and TSS as supporting sensors. In this study, both cases showed significant accuracy improvement when COD was included as a supporting sensor. Case 2 also demonstrated the potential improvement from using sensors for nutrients (i.e., OP-P) in predicting effluent BOD_5. However, this study was able to keep the complex sensors to a minimum by using the modified PLS approach for prioritization and performing a sensitivity analysis of soft sensor performance using different supporting sensors.

The two cases used in this study varied in terms of statistical properties (see supplementary information Table S4). This affected the model's performance based on RMSE, where data with a higher range (Case 1) also resulted in higher RMSE. Because of this, it is difficult to compare the reported performance of soft sensors developed using different datasets. It should note that most studies use a private dataset, which further limits the potential for comparison. These datasets may have unique characteristics which will influence the conclusions of the study. The size of the dataset alone is an influential factor, affecting the generalizability of the soft sensor. Mjalli et al. (2007) used a relatively small dataset of 73 samples from Doha West Wastewater Treatment Plant. In comparison, the dataset used by Ebrahimi et al.

(2017) was composed of 9,180 samples from Floyds Forks Water Quality Treatment Center.

This study aimed to make a comprehensive summary of the characteristics of the two datasets used in its analysis. This was intended to allow for comparison between the results of this study on XGBoost, as well as past and future efforts in soft sensor development for wastewater parameters. Aside from summarizing the characteristics of the datasets used, the study used a public dataset in Case 1 (Dua & Graff, 2019), allowing future studies to have the opportunity to make a direct comparison using the same dataset.

It was also observed that the majority of studies tended to focus on effluent prediction. In most cases, effluent parameters were predicted using influent parameters. The availability of influent parameters as supporting sensors may be one reason for the majority of soft sensor studies being concerned with the effluent. Previously, some studies were cited which used measures such as influent TSS and influent COD to predict BOD_5. Aside from this, influent parameters such as ammonia and flow rate have been used to predict effluent COD (Cong & Yu, 2018; Grieu et al., 2005). Effluent TP has been predicted using TP and TSS in the influent (Wang et al., 2021; Bagheri et al., 2015). Conversely, it makes no logical reason to predict the influent parameters using effluent data, which may be one reason that there are significantly more soft sensors that have been developed for the effluent, compared to the influent (Ye et al., 2020). In comparison, relatively few soft sensors have been developed for influent parameters. These include models for influent COD and TP developed by Wang et al. (2019); and the model for influent TP of Zhu et al. (2017). So far, this study is one of the only a few studies to develop a soft sensor for the influent BOD_5; the XGBoost-based machine learning model provided good opportunity for achieving this objective.

5. Conclusion

This study developed soft sensors for predicting BOD_5 using XGBoost machine learning. This new method was applied to two cases to evaluate its robustness. In both cases, XGBoost estimated a wide range of BOD_5 values, showing consistent performance across different test sets. Although the average performance of machine learning models tended to converge, XGBoost has an innate method of handling missing values; is less prone to overfitting; and was observed to be more effective in measuring higher values of pollutant concentration. XGBoost was particularly effective in estimating effluent BOD_5 which is characterized by important outliers, as cases of high pollutant concentration rarely occur. The soft sensor developed in this study was validated through 10-fold cross validation; however, in future work, we expect to validate the soft sensor in lab-scale or full-scale operation.

Acknowledgement

The authors give their warmest thanks to the Drainage Services Department of Hong Kong for supporting the research and providing the dataset used in this study. This study was also partially supported by the Hong Kong Innovation and Technology Commission (grant no ITC- CNERC14EG03), Hong Kong Research Grant Council (grant no T21-604/19-R), and Ghent University (BOF/STA/202109/022), Belgium.

References

1. Abdeldayem, O. M., Dabbish, A. M., Habashy, M. M., Mostafa, M. K., Elhefnawy, M., Amin, L., ... Rene, E. R., 2022. Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. Science of The Total Environment 803, 149834.
2. Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., Prasath, V. S., 2019. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. Big Data 7, 221-248.
3. Asami, H., Golabi, M., Albaji, M., 2021. Simulation of the biochemical and chemical oxygen demand and total suspended solids in wastewater treatment plants: Data-mining approach. Journal of Cleaner Production 296, 126533.
4. Asante-Okyere, S., Shen, C., Ziggah, Y. Y., Rulegeya, M. M., Zhu, X., 2020. Principal component analysis (PCA) based hybrid models for the accurate estimation of reservoir water saturation. Computers & Geosciences 145, 104555.

5. Ba-Alawi, A. H., Vilela, P., Loy-Benitez, J., Heo, S., Yoo, C., 2021. Intelligent sensor validation for sustainable influent quality monitoring in wastewater treatment plants using stacked denoising autoencoders. Journal of Water Process Engineering 43, 102206.
6. Bagheri, M., Mirbagheri, S. A., Ehteshami, M., Bagheri, Z., 2015. Modeling of a sequencing batch reactor treating municipal wastewater using multi-layer perceptron and radial basis function artificial neural networks. Process Saf. Environ. 93, 111-123.
7. Chen, T. Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 784-794.
8. Cong, Q., Yu, W., 2018. Integrated soft sensor with wavelet neural network and adaptive weighted fusion for water quality estimation in wastewater treatment process. Measurement 124, 436-446.
9. [dataset] Dua, D., Graff, C., 2019. Water treatment dataset. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
10. Ebrahimi, M., Gerber, E. L., Rockaway, T. D., 2017. Temporal performance assessment of wastewater treatment plants by using multivariate statistical analysis. J. Environ. Manag. 193, 234-246.
11. Fan, Y., Xu, Z., Huang, Y., Wang, T., Zheng, S., DePasquale, A., ... Li, B., 2020. Long-term continuous and real-time in situ monitoring of Pb (II) toxic contaminants in wastewater using solid-state ion selective membrane (S-ISM) Pb and pH auto-correction assembly. Journal of Hazardous Materials 400, 123299.
12. Grieu, S., Traoré, A., Polit, M., Colprim, J., 2005. Prediction of parameters characterizing the state of a pollution removal biologic process. Eng. Appl. Artif. Intell. 18, 559-573.
13. Haimi, H., Mulas, M., Corona, F., & Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: An overview. Environmental Modelling & Software 47, 88-107.
14. Hill, A., Tait, S., Baillie, C., Virdis, B., & McCabe, B., 2020. Microbial electrochemical sensors for volatile fatty acid measurement in high strength wastewaters: A review. Biosensors and Bioelectronics, 112409.
15. Hosseinzadeh, A., Zhou, J. L., Altaee, A., Li, D., 2022. Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process. Bioresource Technology 343, 126111.
16. Jaramillo, F., Orchard, M., Muñoz, C., Antileo, C., Sáez, D., Espinoza, P., 2018. On-line estimation of the aerobic phase length for partial nitrification processes in SBR based on features extraction and SVM classification. Chemical Engineering Journal 331, 114-123.
17. Jiang, H., Zou, B., Xu, C., Xu, J., Tang, Y. Y., 2020. SVM-Boosting based on Markov resampling: Theory and algorithm. Neural Netw. 131, 276-290.
18. Li, D., Liu, Y., Huang, D., 2020. Development of semi-supervised multiple-output soft-sensors with Co-training and tri-training MPLS and MRVM. Chemom. Intell. Lab. Syst. 199, 103970.
19. Liu, Y., Xie, M., 2020. Rebooting data-driven soft-sensors in process industries: A review of kernel methods. J. Process Control 89, 58-73.
20. Liu, Y., 2017. Adaptive just-in-time and relevant vector machine based soft-sensors with adaptive differential evolution algorithms for parameter optimization. Chem. Eng. Sci. 172, 571-584.
21. Ma, J., Ding, Y., Cheng, J. C., Jiang, F., Xu, Z., 2020. Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. Water Res. 170, 115350.
22. Matheri, A. N., Ntuli, F., Ngila, J. C., Seodigeng, T., Zvinowanda, C., 2021. Performance prediction of trace metals and cod in wastewater treatment using artificial neural network. Computers & Chemical Engineering 149, 107308.
23. Mehmood, T., Sæbø, S., Liland, K. H., 2020. Comparison of variable selection methods in partial least squares regression. J. Chemom. 34, e3226.
24. Mjalli, F. S., Al-Asheh, S., Alfadala, H. E., 2007. Use of artificial neural network black-box modelling for the prediction of wastewater treatment plants performance. J. Envi. Manag. 83, 329-338.

25. Moufid, M., Hofmann, M., El Bari, N., Tiebe, C., Bartholmai, M., Bouchikhi, B., 2021. Wastewater monitoring by means of e-nose, VE-tongue, TD-GC-MS, and SPME-GC-MS. Talanta 221, 121450.
26. Raud, M., Tenno, T., Jõgi, E., Kikas, T., 2012. Comparative study of semi-specific Aeromonas hydrophila and universal Pseudomonas fluorescens biosensors for BOD measurements in meat industry wastewaters. Enzyme and Microbial Technology, 50(4-5), 221-226.
27. Sharafati, A., Asadollah, S. B. H. S., Hosseinzadeh, M., 2020. The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. Process Saf. Environ. 140, 68-78.
28. Qi, X., Guo, H., Wang, W., 2021. A reliable KNN filling approach for incomplete interval-valued data. Eng. Appl. Artif. Intell. 100, 104175.
29. Qin, X., Gao, F., Chen, G., 2012. Wastewater quality monitoring system using sensor fusion and machine learning techniques. Water Res. 46, 1133-1144.
30. Qiu, J., Lü, F., Zhang, H., Shao, L., He, P., 2021. Data mining strategies of molecular information for inspecting wastewater treatment by using UHRMS. Trends in Environmental Analytical Chemistry, e00134.
31. Wang, G., Jia, Q. S., Zhou, M., Bi, J., Qiao, J., 2021. Soft-sensing of wastewater treatment process via deep belief network with event-triggered learning. Neurocomputing 436, 103-113.
32. Wang, T., Xu, Z., Huang, Y., Dai, Z., Wang, X., Lee, M., ... Li, B., 2020. Real-time in situ auto-correction of K+ interference for continuous and long-term NH4+ monitoring in wastewater using solid-state ion selective membrane (S-ISM) sensor assembly. Environmental Research 189, 109891.
33. Wang, X., Kvaal, K., Ratnaweera, H., 2019. Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. J.Process Control 77, 1-6.
34. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H. ... Zhou, Z.H., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14, 1-37.
35. Xiao, H., Bai, B., Li, X., Liu, J., Liu, Y., Huang, D., 2019. Interval multiple-output soft sensors development with capacity control for wastewater treatment applications: A comparative study. Chemom. Intell. Lab. Syst. 184, 82-93.
36. Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J. Wang, J., 2020. Tackling environmental challenges in pollution controls using artificial intelligence: A review. Sci. Total Environ. 699, 134279.
37. Zaghloul, M. S., Hamza, R. A., Iorhemen, O. T., Tay, J. H., 2020. Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors. J. Environ. Chem. Eng. 8, 103742.
38. Zhu, S., Han, H., Guo, M., Qiao, J., 2017. A data-derived soft-sensor method for monitoring effluent total phosphorus. Chin. J. Chem. Eng. 25, 1791-1797.

Tables

Table 1 RMSE (mg/L) of the model trained on the (a) UCI Machine Learning Repository dataset and (b) Hong Kong dataset (using different methods of handling missing values)

	(a) Case 1: UCI Machine Learning Repository				**(b) Case 2: Hong Kong Dataset**			
	Influent BOD		**Effluent BOD**		**Influent BOD**		**Effluent BOD**	
	Ave	Std. Dev.	Ave	Std. Dev.	Ave	Std. Dev.	Ave	Std. Dev.
Samples without missing values	52.41	9.06	10.59	7.96	67.79	17.52	0.47	0.20
Missing values filled in with kNN	52.07	8.96	10.59	8.01	70.64	16.64	0.77	0.86
Missing values processed by XGBoost	51.93	9.31	10.55	7.98	68.60	19.97	1.17	1.48

Table 2 RMSE (mg/L) of 10-fold cross-validation for models developed using the (a) UCI Machine Learning Repository dataset and (b) Hong Kong dataset.

	(a) Case 1: UCI Machine Learning Repository		**(b) Case 2: Hong Kong Dataset**	
	Influent BOD			
	Average	Std. Dev.	Average	Std. Dev.
XGBoost*	51.93	9.31	67.79	17.52
ANN with kNN	50.51	11.04	67.58	19.60
SVM with kNN	50.51	11.04	67.49	23.58
	Effluent BOD			
	Average	Std. Dev.	Average	Std. Dev.
XGBoost *	10.55	7.98	0.47	0.20
ANN	10.80	9.95	0.48	0.30
SVM	11.98	12.87	0.51	0.41

* Note: For Case 1, the XGBoost were analyzed with Sparsity Awareness Algorithm

Figures

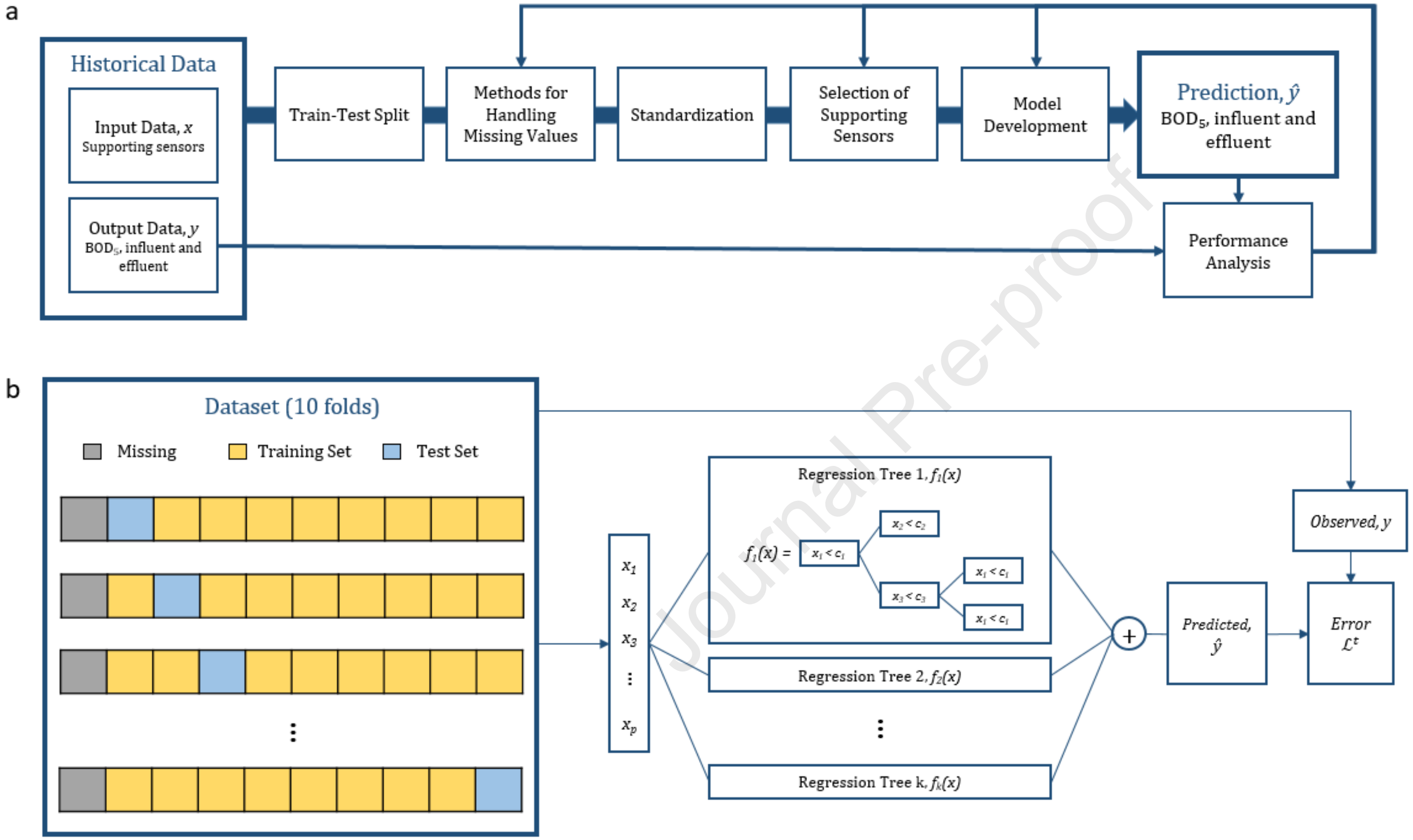

Figure 1 Soft sensor development frameworks: (a) methods applied, and (b) XGBoost model structure.

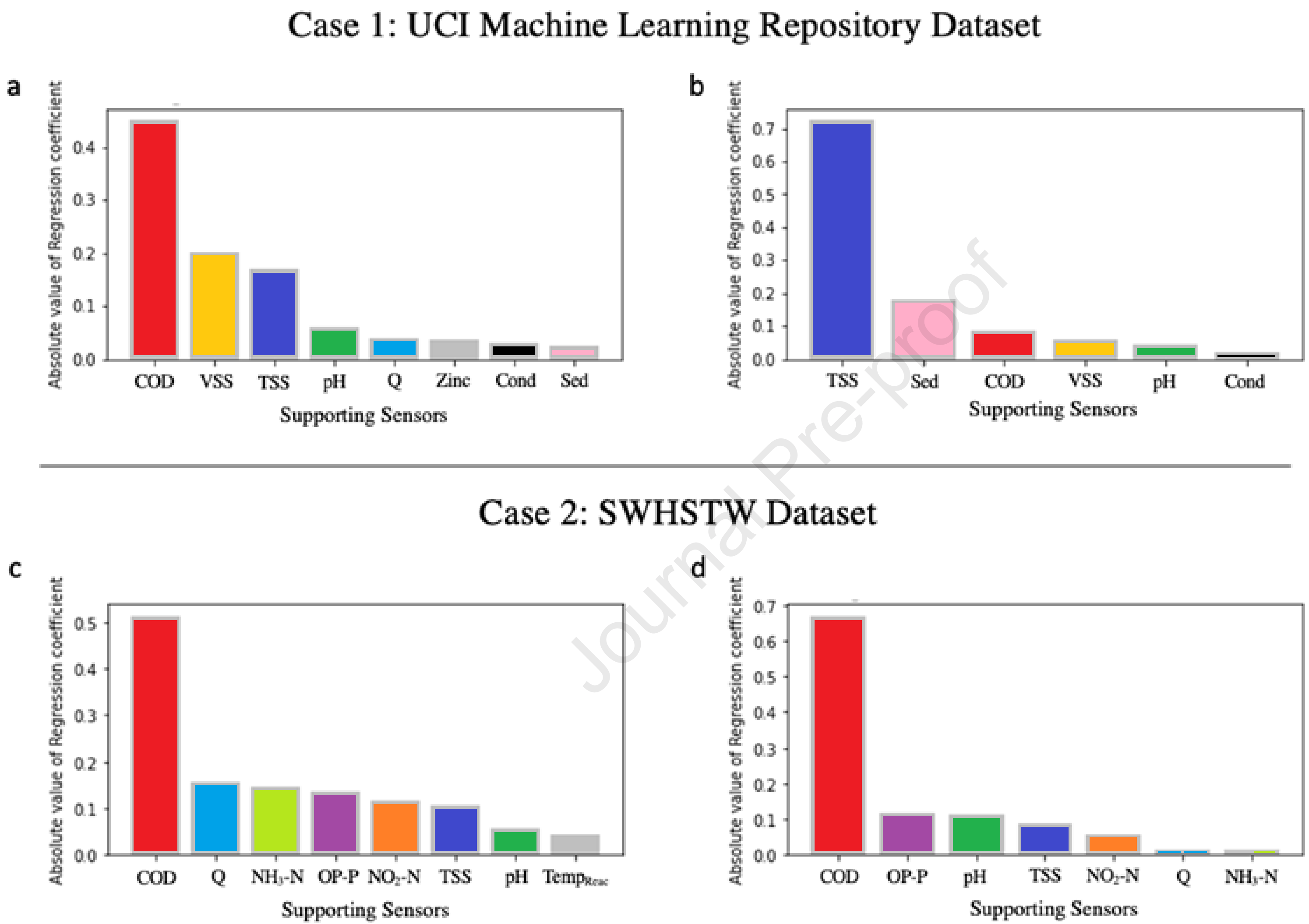

Figure 2 PLS regression coefficients for (a and c) influent BOD and (b and d) effluent BOD, with common parameters indicated by color.

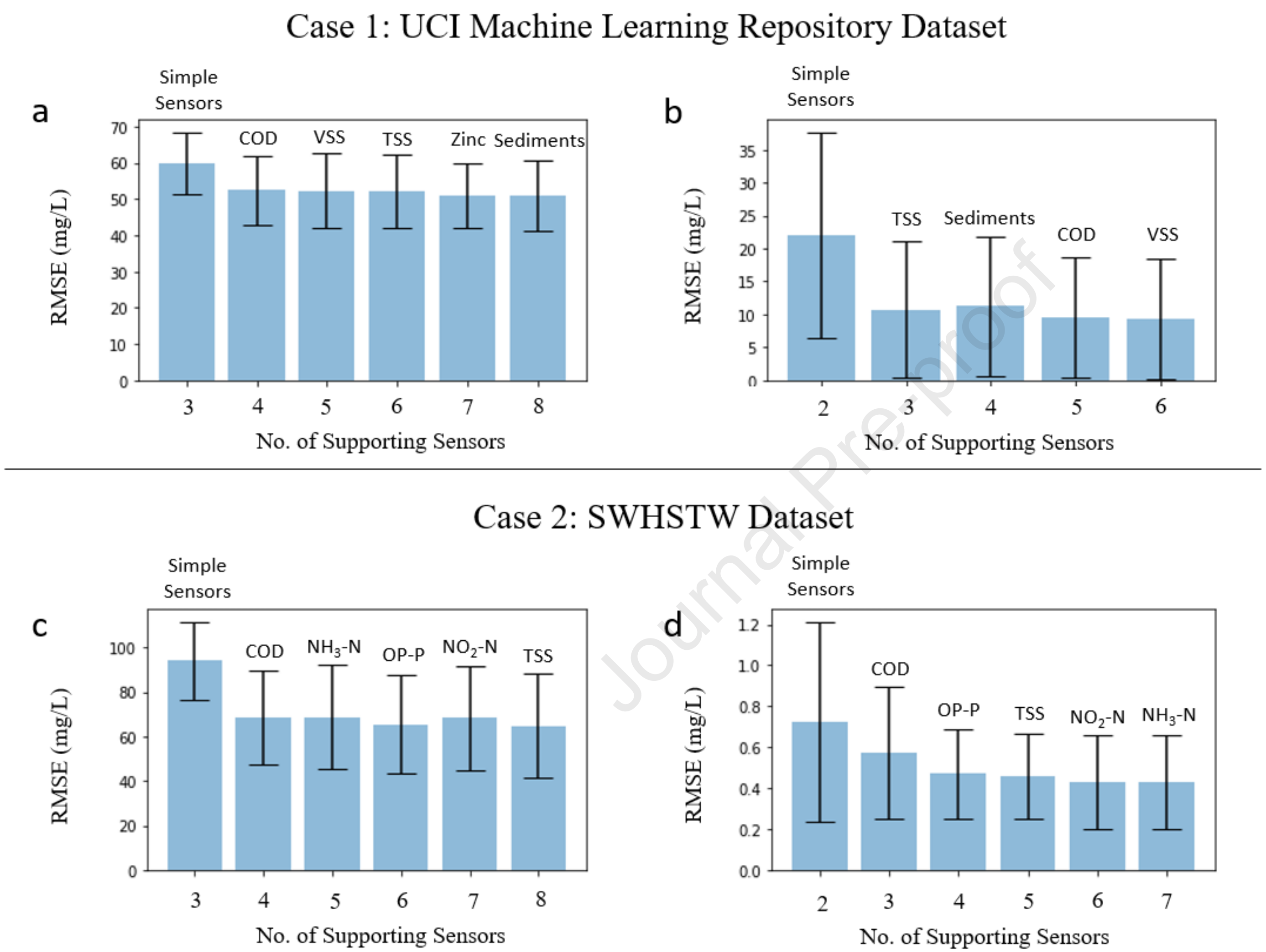

Figure 3 Change in RMSE (mg/L) supporting sensors are incrementally added to the soft sensor for (a, c) influent BOD and (b, d) effluent BOD.

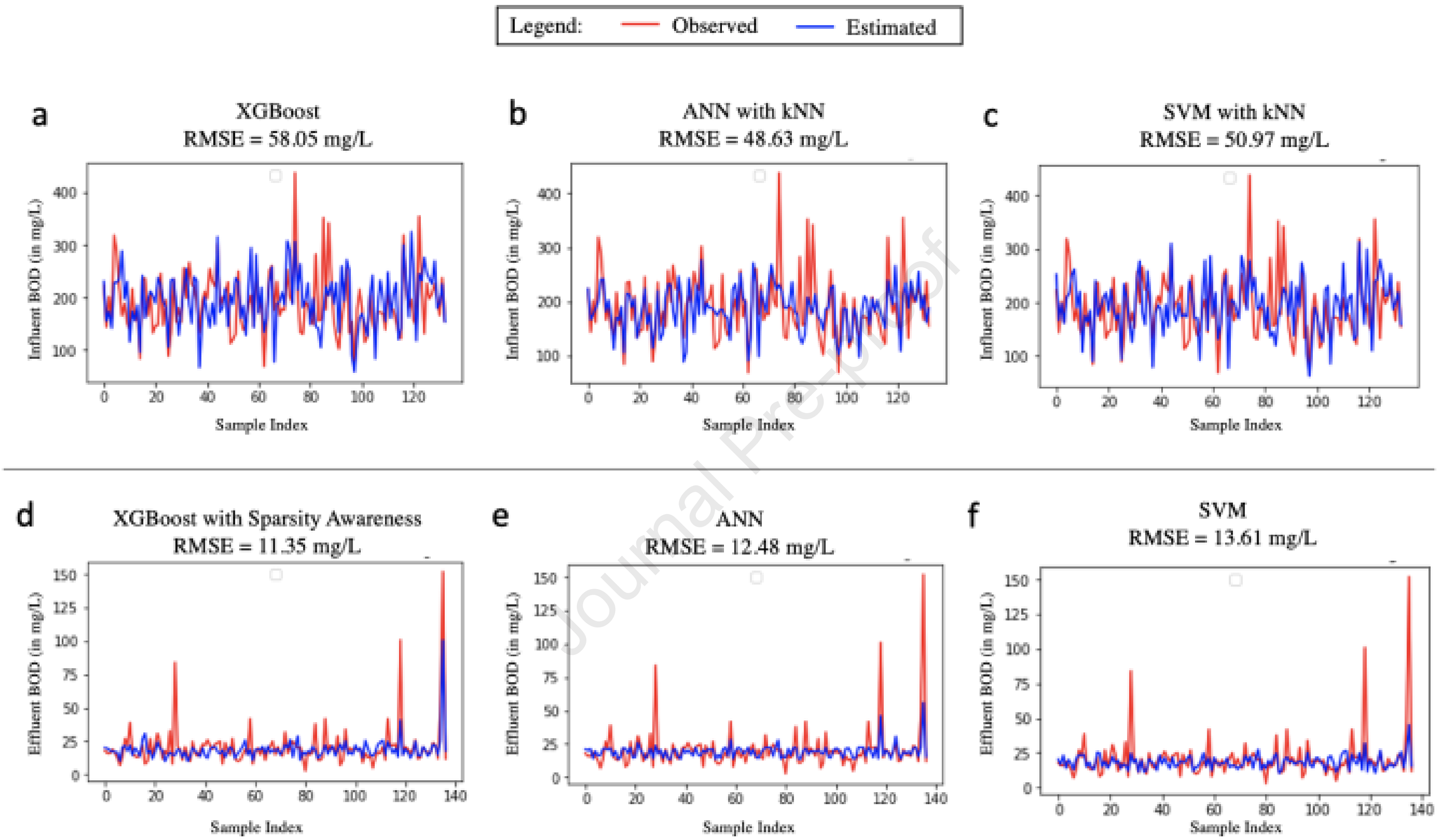

Figure 4 Visual comparison and RMSE (mg/L) of (a) BOD_{inf} estimated by XGBoost; (b) BOD_{inf} estimated by ANN with kNN; (c) BOD_{eff} estimated by SVM with kNN; (d) BOD_{eff} estimated by XGBoost; (e) BOD_{eff} estimated by ANN with kNN; and (e) BOD_{eff} estimated by SVM with kNN, modeled using the UCI Machine Learning Repository dataset.

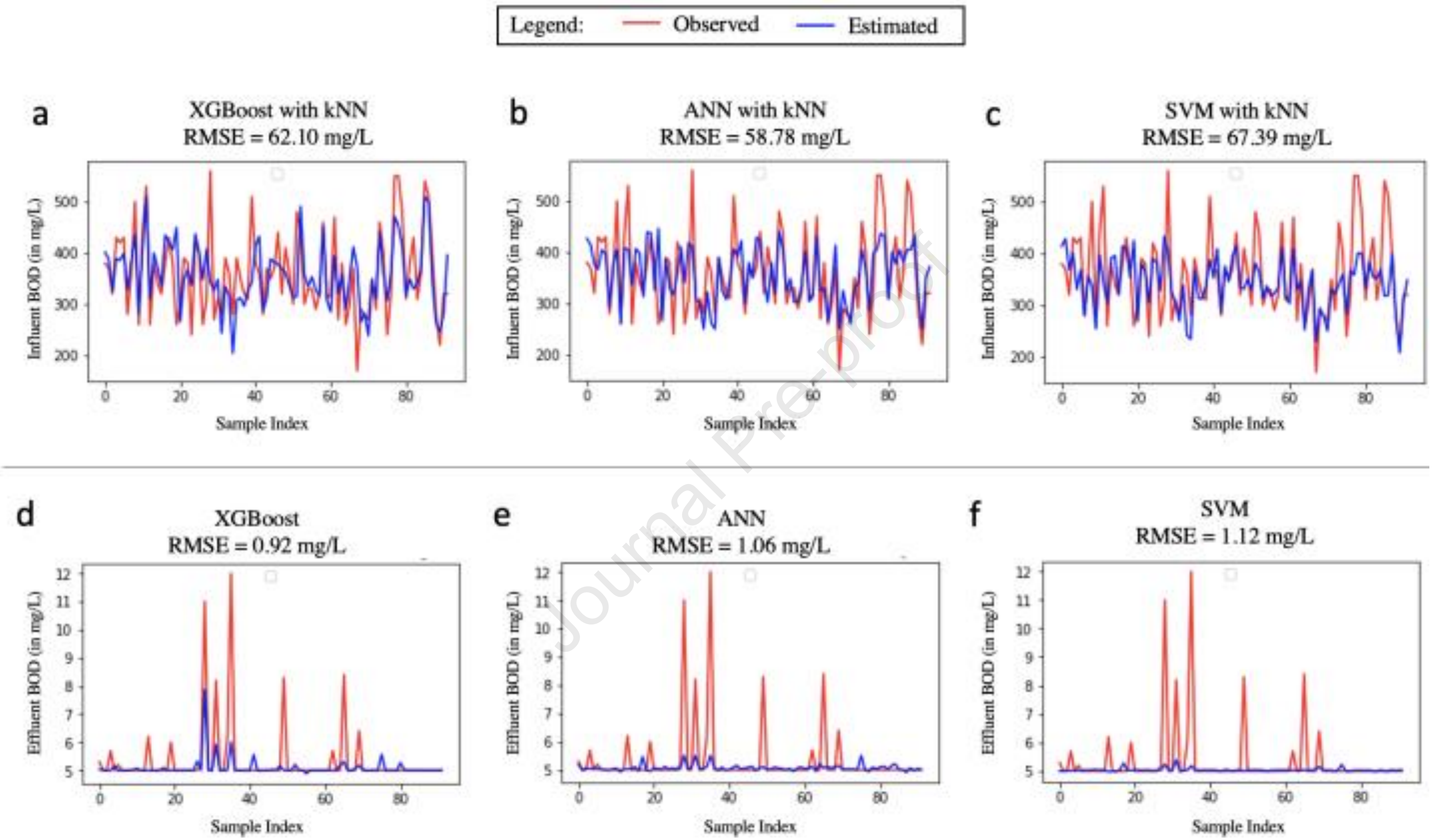

Figure 5 Visual comparison and RMSE (mg/L) of (a) BOD_{inf} estimated by XGBoost; (b) BOD_{inf} estimated by ANN with kNN; (c) BOD_{eff} estimated by SVM with kNN; (d) BOD_{eff} estimated by XGBoost; (e) BOD_{eff} estimated by ANN with kNN; and (e) BOD_{eff} estimated by SVM with kNN, modelled using the Hong Kong dataset.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this review paper.