Journal of Environmental Management 301 (2022) 113868

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Developing a new approach for design support of subsurface constructed wetland using machine learning algorithms

Xuan Cuong Nguyen [a,b], Thi Thanh Huyen Nguyen [a,b], Quyet V. Le [c], Phuoc Cuong Le [d], Arun Lal Srivastav [e], Quoc Bao Pham [f], Phuong Minh Nguyen [g], D. Duong La [h], Eldon R. Rene [k], H. Hao Ngo [l], S. Woong Chang [j], D. Duc Nguyen [i,j,*]

[a] Laboratory of Energy and Environmental Science, Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam
[b] Faculty of Environmental Chemical Engineering, Duy Tan University, Da Nang, 550000, Viet Nam
[c] Department of Materials Science and Engineering, Institute of Green Manufacturing Technology, Korea University, Seoul 02841, Republic of Korea
[d] Department of Environmental Management, Faculty of Environment, The University of Danang-University of Science and Technology, Danang, 550000, Viet Nam
[e] Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India
[f] Institute of Applied Technology, Thu Dau Mot University, Binh Duong Province, Viet Nam
[g] Faculty of Environmental Sciences, University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan, Hanoi, Viet Nam
[h] Institute of Chemistry and Materials, Nghia Do, Cau Giay, Hanoi, Viet Nam
[i] Faculty of Environmental and Food Engineering, Nguyen Tat Thanh University, 300A Nguyen Tat Thanh, District 4, Ho Chi Minh City, 755414, Viet Nam
[j] Department of Environmental Energy Engineering, Kyonggi University, Suwon 442-760, Republic of Korea
[k] Department of Environmental Engineering and Water Technology, IHE Delft Institute for Water Education, 2601DA Delft, the Netherlands
[l] Center for Technology in Water and Wastewater, School of Civil and Environmental Engineering, University of Technology Sydney, Australia

ARTICLE INFO

Keywords:
Ammonium
Constructed wetland
Design
Machine learning
Organic matter

ABSTRACT

Knowing the effluent quality of treatment systems in advance to enable the design of treatment systems that comply with environmental standards is a realistic strategy. This study aims to develop machine learning - based predictive models for designing the subsurface constructed wetlands (SCW). Data from the SCW literature during the period of 2009–2020 included 618 sets and 10 features. Five algorithms namely, Random forest, Classification and Regression trees, Support vector machines, K-nearest neighbors, and Cubist were compared to determine an optimal algorithm. All nine input features including the influent concentrations, C:N ratio, hydraulic loading rate, height, aeration, flow type, feeding, and filter type were confirmed as relevant features for the predictive algorithms. The comparative result revealed that Cubist is the best algorithm with the lowest RMSE (7.77 and 21.77 mg.L^{-1} for NH_4–N and COD, respectively) corresponding to 84% of the variance in the effluents explained. The coefficient of determination of the Cubist algorithm obtained for NH_4–N and COD prediction from the test data were 0.92 and 0.93, respectively. Five case studies of the application of SCW design were also exercised and verified by the prediction model. Finally, a fully developed Cubist algorithm-based design tool for SCW was proposed.

1. Introduction

Constructed wetlands (CWs) are man-made ecological engineering systems (i.e., wetlands) designed and constructed to mimic and enhance the natural functions of wetlands for wastewater treatment or water remediation (Kadlec and Wallace, 2009; Vo et al., 2018). Two commonly effective types include vertical and horizontal flows and are called subsurface constructed wetlands (SCWs). Although research and the use of CWs to ameliorate water quality have been employed for a long time, this still remains a developing technology. The optimal design of CWs to fulfill the environmental requirements as well as the available resources requires considerable research attention.

In general, three methods have been used to design SCWs, namely the rule of thumb, kinetics, and statistical models. The design of SCWs based on the rule of thumb is the simplest approach that is derived from the observations, design manual, or the experiences of the CW

* Corresponding author. Faculty of Environmental and Food Engineering, Nguyen Tat Thanh University, 300A Nguyen Tat Thanh, District 4, Ho Chi Minh City, 755414, Viet Nam.
E-mail address: nguyensyduc@gmail.com (D.D. Nguyen).

https://doi.org/10.1016/j.jenvman.2021.113868
Received 21 June 2021; Received in revised form 7 September 2021; Accepted 26 September 2021
Available online 7 October 2021

engineers. However, the design rule of the thumb approach also originated from modeling or simulation programs (Moran, 2018). According to Rousseau et al. (2004), the rule of thumb used in calculating the area of horizontal flow (HF) seemed to be a conservative design method that resulted in low effluents but a high investment cost. Kinetic models are used to obtain insights into the SCW system regarding the reaction rate and hydraulic loading pattern. Thus, the first-order and Monod models with various hydraulic loading pattern assumptions (i.e., ideal plug flow, continuously stirred tank reactors background concentration, and tanks - in - series) were developed to simulate the results (Kadlec and Knight, 1996; Kadlec and Wallace, 2009; Rousseau et al., 2004; Wood, 1995). In addition, the compartmental model that simulates bacterial growth and metabolism in SCWs was presented by Wynn and Liehr (2001). These kinetic models contain a lot of assumptions, empirical relations, and are too sensitive to changes in individual parameters (Rousseau et al., 2004; Wynn and Liehr, 2001).

Another approach of SCW design is based on the statistical model called the regression method, which models the relationship between one or more input variables (influent concentrations, operational parameters, system characteristics, etc.) and output variables (i.e., effluents or removal rate). This may ignore the underlying mechanisms and hydraulic patterns of SCWs, paying more attention to data and the performance of the system (Rousseau et al., 2004). Although regression-based design does not directly reflect the mechanisms of the SCW system, the high performance of regression equations (i.e., high coefficient of determination–R^2) is helpful for design purposes. A wide range of R^2 values have been reported using the regression method for modeling SCWs. For example, high R^2 values of 0.99 (COD) and 0.96 (NH_4–N) were achieved with inlet loading rate as the input variable and the removal rate/efficiency as the output (Albuquerque et al., 2009; Chang et al., 2015) while other studies revealed R^2 values of 0.73 and 0.61 for the COD and TP, respectively (Haberl et al., 1995). An investigation using multiple regression for a hybrid SCW by Nguyen et al. (2018) to predict the removal rate and effluent concentration (NH_4–N and BOD_5) reported R^2 values of 0.90 and 0.64–0.67, respectively. Regarding the low correlation index, Li et al. (2018) concluded that the diverse nature of the data of SCWs may be a reason for the low regression coefficient values between the design parameters and pollutant removal efficiencies, which poses new challenges and research topics for SCW design.

The rapid increase in the number of studies on SCWs has added more data to the field. This opens up opportunities for new approaches, namely data-driven design–decisions based on quantitative data. This is vitally useful because in engineering practice, designers and engineers need to roughly estimate the system performance based on the input data available (Romeo et al., 2020). In addition, appropriate and diverse sources of data from different types of CW in terms of inlet pollutant concentrations, operations, size, configuration, feeding strategy, and aeration can provide valuable information on SCWs via comparative analysis and correlation. An effective data-driven model to cope with abundant data available on SCWs is machine learning (ML). ML uses mathematical algorithms that can automatically learn from data to build the model (Chollet and Allaire, 2018; Haupt et al., 2009). A powerful capacity of ML, which is useful in the environmental field, is prediction. Nitrogen removal in vertical flow (VF) was predicted early with an artificial neural network (ANN) model including three layers with a structure of 9 (i.e., inputs)–6 (i.e., neurons in the hidden layer)–1 (i.e., output) and achieved low accuracy with an R^2 of 0.69 (Akratos et al., 2009). ANN and SVM algorithms used previously for forecasting the total nitrogen (TN) effluent in a food wastewater plant showed that SVM had a higher prediction performance than ANN, but the R^2 values for validation was rather low, ranging from 0.46 to 0.47 (Guo et al., 2015). These low predictions could possibly be due to the small dimensions of the data, inappropriate input variables, and the algorithm used or the lack of hyper-parameter optimization.

From the above analysis, it can be concluded that an effective tool based on ML to support the design of SCWs is necessary and is a promising work, especially if it inherits the previous results of SCW. Applying the predictive power of ML to predict and assist SCW design on the basis of caring about discharge standards is a new attempt. Thus, this study aims to develop an approach that supports SCW design based on a ML method. It includes collecting datasets of SCWs from the literature (618 sets and 10 attributes), choosing the appropriate input variables, comparing the five ML algorithms from two groups of supervised learning algorithms (non-linear vs ensemble methods), turning, and predicting the effluents of SCW. Finally, five case studies of the application of SCW design were also applied and verified by the prediction model.

2. Method

2.1. Subsurface constructed wetland

Here, two popular types of CW in terms of configuration, flow, filter materials, plants, and other characteristics, including VF and HF, are presented in Fig. 1. Normally, VF distributes water across the surface, while HF flows from the inlet to the outlet, and the water level in the treatment tank remains below the surface of the filter. In addition, a SCW has several modifications such as a baffled surface (Wang et al., 2012; Zhai et al., 2011), different zones with plants and no plants (Benny and Chakraborty, 2020), and separate input areas (Ge et al., 2019). In VF, intermittent flow is frequently used to get more oxygen from the air to the bed, but some combinations of tidal flow, unsaturated down-flow, and recirculation have been investigated (Foladori et al., 2013; Kadlec and Wallace, 2009; Sklarz et al., 2009). The height of the filter in VF is normally higher than that in HF, but there have been exceptions (Abou-Elela et al., 2013). In addition, the ratio of length: height in HF is larger than that in VF, for instance, the rate of 4.3 reported in the study of HF by Benny and Chakraborty (2020). The application of VF and HF types was diverse, including the treatment of domestic, industrial, agricultural, and storm wastewaters.

2.2. Dataset

2.2.1. Characteristics of the dataset

For the purpose of conforming to the goal of a SCW, design-related variables from the literature were compiled. The design parameters selected for CWs vary widely, including the hydraulic retention time, hydraulic loading rate (HLR), areal requirements, and maximum BOD_5 loading rate (Kadlec and Knight, 1996; Rousseau et al., 2004; USEPA, 2000; Wood, 1995). The categorical variables of a SCW (e.g., flow type, feeding regime, etc.), which had not been previously used for the design because of the prediction tool, were gathered and employed in this work to improve the prediction efficiency.

The cleaning dataset included 618 rows and 10 columns (attributes or features) which were derived from 31 studies, published from 2009 to 2020. The type and characteristics of the data are summarized in Tables 1S and 2S, respectively (see Supplementary Data). There are two common types of data: numerical and categorical variables and nominal variables, but the ML algorithm does not use the data directly. The 10 inputs used in this study were NH_4–N influent (Inf_NH_4), COD influent (Inf_COD), TN influent (Inf_TN), carbon/TN ratio (Inf_C–N), HLR, filter height (Height), types of SCW (Type), aeration or not (Aeration), inlet feeding (Feeding), and the types of filter (Filter), while the COD of the effluent (Eff_COD) and NH_4–N (Eff_NH_4) were the target variables.

The input variables were selected based on their potential to influence and control the SCW system as well as on the accuracy of the predictive model. Here, we chose input variables with three aspects: influent, operation, and design, based on the understanding of the SCW system and the availability of published data about SCW. The critical factors of SCW have been previously considered, such as plant, wastewater type, HLR, hydraulic retention time, water depth, feeding (Ghosh

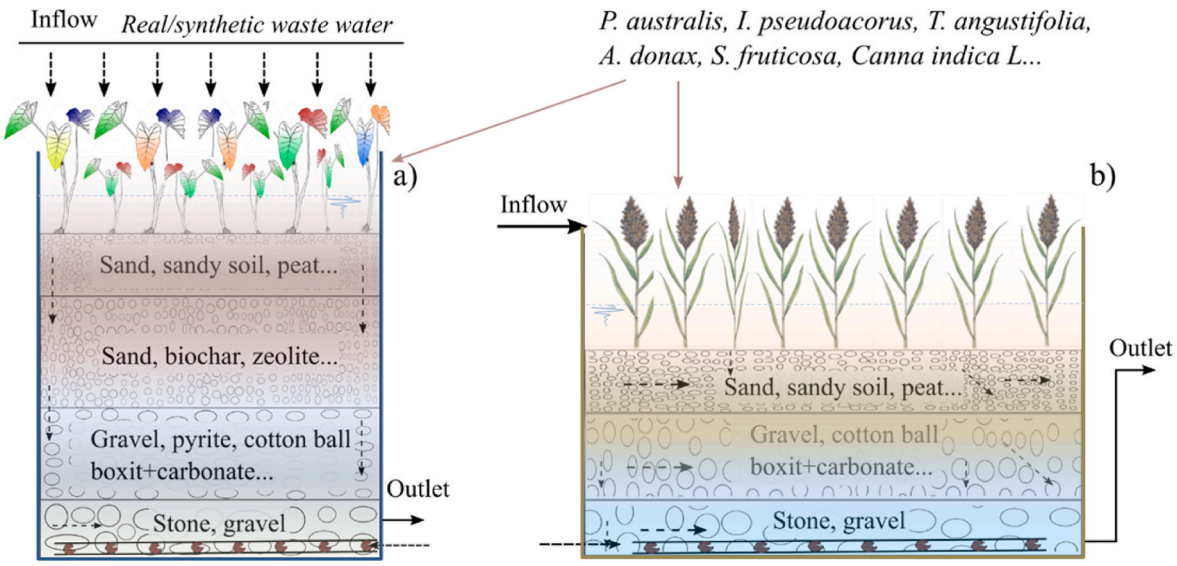

Fig. 1. Schematic of the subsurface constructed wetlands: Vertical flow constructed wetland (a) and horizontal flow constructed wetland (b).

and Gopal, 2010; Stefanakis and Tsihrintzis, 2012; Wu et al., 2015), C:N ratio (Wang et al., 2020a), and oxygen or air supply (Fan et al., 2013; Ilyas and Masih, 2017; Jia et al., 2018).

2.2.2. Data collection and tools

For this study, we collected experimental data on SCWs from previously published literature that had sufficient information. The balance among categories, for example, VF versus HF, was also considered when selecting the source of data. These studies were published between 2009 and 2020. Given the input variables, data were taken directly from the study's figures or extracted from graphs. R language (version 4.0, https://cran.r-project.org) was used to process the data and run the ML algorithms. In addition, the packages applied to visualize and process the data as well as to run the ML algorithms include "Random Forest," "Caret," "ggplot 2," "psych," "Cubist," and "dplyr,".

2.3. Design of constructed wetlands

2.3.1. Design methods for constructed wetlands

2.3.1.1. Rule of thumb design. The rule of thumb used in SCW design is derived from experimental knowledge and is a relatively conservative design method that tends to result in the extension of tank volume (Rousseau et al., 2004). Some variables of this approach used for SCW design are HLR, hydraulic retention time (HRT), the areal requirement, and the removal rate/efficiency of the pollutants. For instance, the bed area and HLR of a SCW recommended for practical application were reported as 3–5 $m^2.PE^{-1}$ and 0.08–0.3 m d^{-1}, respectively (Kadlec and Knight, 1996). In addition, the HLR value of the SCW is normally less than 5 m d^{-1} (Wu et al., 2015).

2.3.1.2. Model-based design. This approach develops models based on the reaction rate and hydraulic patterns. The reaction rate of zero order, first order, Monod model, and flow patterns of an ideal plug flow and a plug flow with dispersion are integrated into the design models. Zero-order-contaminant reduction is a constant rate of removal that is not real in SCWs, while several models with Monod-type, variable-order, and plug flow with dispersion contain more assumptions and empirical relations. Wetland designers use a first-order areal constant (k, m/d) and/or background concentration (C*) to calculate the required area of the system (Luo et al., 2020; Rousseau et al., 2004). To provide an overview of the model for SCW design, we refer to three models as follows (Eqs. (1)–(3)).

First-order model with ideal plug flow (Kadlec and Knight, 1996):

$$\frac{C_e}{C_i} = e^{-k/q} \tag{1}$$

First-order model with ideal plug flow and C* (i.e., the k – C* model) (Kadlec and Knight, 1996):

$$\frac{C_e - C^*}{C_i - C^*} = e^{-k/q} \tag{2}$$

First-order model with tanks - in - series (Kadlec and Wallace, 2009)

$$\frac{C_e - C^*}{C_i - C^*} = \frac{1}{\left(1 + \frac{k}{p \cdot q}\right) \cdot P} \tag{3}$$

where, C_o is the outlet concentration, (mg.L^{-1}); C_i is the inlet concentration (mg.L^{-1}); C^* is the background concentration (mg.L^{-1}); k is the modified first-order areal constant (m/d); P is the apparent number of tanks in series; q is the hydraulic loading rate (m.d^{-1}).

2.3.1.3. Statistical model-based design. A common statistical model deployed for SCW design is linear regression, which focuses on the input-output of a SCW compared to other internal processes. This lumps the system of a SCW into some parameters with an oversimplification, for example, simple regression has only one input parameter, and multiple regression has more than two parameters. A linear regression model includes an output (i.e., the dependent variable) and several independent variables (i.e., the predictors).

2.3.2. Machine learning-based design

ML has proved to be highly effective in predicting the effluent quality of wastewater treatment systems. In engineering practice, it is more common for engineers to roughly estimate some basic performance parameters on the basis of input data available during the process or later in the design phase (Romeo et al., 2020). For a complex system such as SCWs that have heterogeneous input parameters, prediction of effluent quality can be time-consuming and it might result in large errors. Due to some limits of the approach of model-based design, the rule of thumb for SCW design is practical and feasible to create the draft project quickly. The power of ML is now harnessed to accurately forecast the output and from there, to fine-tune the design of SCW. The difference to the previous approaches is that this method combines the rule of

thumb method and the powerful performance of ML to predict the effluent quality. It needs no hydraulic assumption or empirical relations (as kinetic methods), but paying a lot of attention to effluents of SCW system – to fulfill discharge standards. Additionally, it minimizes SCW area waste, similar to the rule of thumb method, by design adjustments based on effluent forecasting in advance. Therefore, it can be said that the ML approach will support the efficient design of SCWs. A combination of the predictive model using ML and SCW design was applied in this work and is presented in Figs. 2S and Fig. 5, respectively.

2.4. Procedure and machine learning algorithms

2.4.1. Feature engineering

2.4.1.1. Data transformation and resampling. Owing to the large range of values of the features, *Scale* and *Box-cox* techniques were applied to mutate the raw data into the proper format for the ML. The large scale of different data was converted into a common scale-by-scale technique. For some ML algorithms such as regression or Euclidean distances (e.g., KNN or K-Means), the *Scale* method is useful. The *Box-cox* technique was used to transform the skewed distribution data into normalized data – a common assumption of many statistical methods. Furthermore, *One-hot coding* was applied to encode two-factor variables (Type, Feeding Filter, and Aeration). After this step, the data was split into 496 sets as training data (80% of the total), and 122 sets as testing data (20% of the total). The training dataset was employed to "train" and finalize the optimal model, while the testing dataset was used to check the accuracy of the final model.

2.4.1.2. Feature selection. The collected data were normally qualitative based on the understanding that the system may have a large dimension with many variables. Furthermore, we had no prior knowledge of which input factors were significant or unimportant to the predictive models. Therefore, the assessment of the importance and selection of variables is necessary to achieve high efficiency in the application of predictive modeling. Boruta is a wrapper built by the RF classification that runs first in the R program (Kursa and Rudnicki, 2010) via a package of "Random Forest" (Liaw and Wiener, 2001). It ranks variables and confirms or rejects them automatically on a given dataset.

2.4.2. Machine learning algorithms

2.4.2.1. Cubist. The Cubist algorithm is a prediction-oriented regression model described in Quinlan (1992) with additional corrections based on nearest neighbors in the training set, as described in Quinlan (1993). The Cubist models are combined using a linear combination, as follows (Eq. (4)) (Kuhn and Johnson, 2013):

$$\widehat{y}_{par} = a \cdot x \cdot \widehat{y}_k + (1-a) \cdot x \cdot \widehat{y}_p \tag{4}$$

where, $\widehat{y}_k$ is the prediction from the current algorithm; $\widehat{y}_p$ is from the parent model above it in the tree; a is the smoothing coefficient.

Cubist is a model tree that simultaneously constitutes a set of rules (i.e. committee) and a certain linear regression model. In other words, Cubist uses an "if and then" statement, in which "if" is a condition and "then" is a linear equation for calculating the outcome. Two parameters of Cubist can be default or tunable, including committees - the number of boosting operations and neighbors - the number of instances used to correct the rule-based prediction (Kuhn and Johnson, 2013; Nguyen et al., 2021).

2.4.2.2. CART – Classification and Regression trees. This algorithm was first described by Breiman et al (1984) and has been developed with modern variations such as RF. CART can be applied to classification or regression predictive problems. The basic concept behind this algorithm is a straightforward but a reliable forecasting model (Krzywinski and Altman, 2017). In this method, data are partitioned along the predictor axes into subsets with values of the dependent variable, where a decision tree is produced to make predictions from the new observations. The CART model is represented as a binary tree. Each root node represents a continuous predictor variable (x) and a split point on that variable. From the root node, the leaf nodes of the tree are generated with an output variable (y) that is used to create the prediction. The process continues to reach a final prediction, in other words, until a suitable tree is constructed.

2.4.2.3. Random forest. RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). RF is a type of ensemble algorithm called bootstrap aggregation and is one of the most popular ML methods. An RF performs the prediction based on input variables (x) that occur in the following steps (Cutler et al., 2011): (1) Take a bootstrap sample and randomly select a subset of the supervised learning dataset; (2) Fit a tree using binary recursive partitioning; (3) Start with all observations in a single node and repeat the following steps recursively for each unsplit node until the stopping criterion is met. In this step, two turning parameters, mtry (the number of variables randomly sampled as candidates at each split) and ntree (the number of trees) were used. Lastly, make a prediction of a new point x. For regression, the mean of a number of regression trees is the final result.

2.4.2.4. Support vector machines. SVMs plot each data item as a point in n-dimensional space (n is the number of features) and estimates the hyper-plane that best segregates the two classes (Kuhn and Johnson, 2013). SVM learning combines instance-based nearest-neighbor learning, lazy learning – classification using nearest neighbors, and the linear regression model (Lantz, 2019). SVMs detect an optimal hyper-plane that assists in classifying new points of a dataset. This algorithm can be used for both classification and numerical prediction. As a practical technique for using SVM in practice, a kernel trick and function are utilized; a kernel trick is used to fix non-linear input spaces.

2.4.2.5. K-nearest neighbors. KNN is a type of clustering algorithm, a supervised learning technique used to classify new data points based on their position to nearby data points (Walker, 2018). The KNN predicts a new sample using the K-nearest neighbor samples from the training set (Guo et al., 2003). KNN has no model that simply stores the entire dataset, and therefore, there is no learning required for the algorithm. It uses complex data structures such as k – d trees and observes and connects new patterns during the prediction step. For each instance in the test data, the function will identify the KNN using the Euclidean distance (distance between two points in a plane), where *k* is a user-specified number (Lantz, 2019). KNN has a low error rate with large data, as it optimally finds the nearest neighbor of a point and the low number of features (less dimensions) (Forsyth, 2019).

2.4.3. Metrics and residuals

Two metrics were used to evaluate the accuracy of the algorithms, namely, the root mean squared error (RMSE) and R^2. These metrics were determined as follows (Eqs. (5) and (6)):

$$RMSE = \sqrt{\frac{1}{m}\sum\nolimits_{i\in \text{data}}^{m} \left(y_i - f_{\widehat{w}}(x_i)\right)^2} \tag{5}$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{6}$$

where, m is the number of samples; $f_{\widehat{w}}\ (x_i)$ is the estimated model that predicts the response; y_i is the actual value of the response; *RSS* is the residual sum of the squares; *TSS* is the total sum of the squares; y and $\widehat{y}$

are the actual value and predicted value, respectively; $\overline{y}$ is the mean value.

Residuals and standardized residuals were calculated as follows (Eq. (7)) and (Eq. (8)) (Dobson and Barnett, 2008):

$$r = y - \widehat{y} \tag{7}$$

$$r_s = \frac{y - \widehat{y}}{\widehat{\sigma}} \tag{8}$$

where, r is the raw residual; y and $\widehat{y}$ are the actual value and predicted values, respectively; $\widehat{\sigma}$ is the predicted standard deviation; r_s is the standardized residual.

3. Results and discussion

3.1. Statistical characteristics and test

3.1.1. Characteristics and correlation of variables

A summary of the numerical inputs and target outputs is shown in Table 1. It can be observed that the mean and range of influent concentrations for COD, NH_4–N, and TN were relatively high, with mean values of 159.33, 51.86, and 61.97 mg.L^{-1}, respectively. The C: N ratio was in the range of 0.52–77.74, which was large, and HLR varied from 0.01 to 0.78 m/d with a mean of 0.13 m d^{-1}. This average HLR is common in lab-scale studies or in real applications of SCWs. The reported values of HLR of VF ranged from 0.2 to 0.8 m d^{-1} in China and 0.2–0.3 m d^{-1} in western countries, while the average HF in China was 0.5 m d^{-1} (Li et al., 2018). Ammonium and COD removal averaged 70.00% and 69.67%, and the concentrations in the effluent were 14.74 mg.L^{-1} and 61.50 mg.L^{-1}, respectively. High COD influent concentrations (e.g., 400 mg.L^{-1}) may disrupt the growth of plants in SCWs; some lab-scale investigations, on the other hand, were able to deal with this difficulty and conduct the trials with SCWs at high COD concentrations.

The correlation between numerical inputs of SCW and effluent concentrations, as reflected by the Pearson correlation coefficients (r), regression plots, and probability values (P), are presented in Fig. 2S. Two relationships were identified here: mutual input correlations and inputs–target variable correlations. The first one is to detect the highly correlated variables that may have to be removed and the latter is to provide the initial idea regarding the importance of the input variables. In addition, for the purpose of designing SCWs, the negative or positive correlation between input variables and targets (i.e., Eff_NH_4–N and Eff_COD) is a pre-requisite in order to adjust the inputs and obtain the desired outputs.

As shown in Fig. 2S, the NH_4–N effluents correlated positively with the NH_4–N ($r = 0.74$, $P < 0.0005$) and COD influents ($r = 0.58$, $P < 0.005$). This shows that the influent concentrations of ammonium and COD have a large impact in controlling the effluent concentrations. The inverse relationships between NH_4–N effluent and C: N ratio, HLR, and height reflect the effect of the operational and filter factors on the SCW effluent quality. An increased adjustment of the filter height may enhance the SCW capacity, resulting in a decrease in the NH_4–N effluent. For the COD effluent, the correlation values are relatively low. The highest correction was found between the COD influent and effluent with $r = 0.49$, while the height of the filter had a negative correlation with the COD effluent ($r = -0.36$). Unexpectedly, the correlation between the HLR and effluents had an r value of −0.09. Except for the relatively high interrelationships between the influent and effluent concentrations, the remaining factors had low correlation values for the target variables. This may be due to the erratic nature of pollutant concentrations and factors (e.g., algal growth, dead zone) in the system. In general, the mutual input correlations of the SCW were low ($r < 0.56$), meaning that this data is less likely to cause multicollinearity.

Table 1
Descriptive statistics of the input and output variables.

Input	Unit	Mean ± SD	Min – Max	Meanings
Inf_NH_4	mg. L^{-1}	51.86 ± 54.37	0.09–380.72	NH_4–N influent concentration
Inf_TN	mg. L^{-1}	61.97 ± 58.28	0.62–359.80	Total nitrogen influent concentration
Inf_COD	mg. L^{-1}	259.33 ± 177.74	7.12–770.41	COD influent concentration
HLR	m.d^{-1}	0.13 ± 0.09	0.01–0.78	Hydraulic loading rate
C: N	–	5.79 ± 5.16	0.52–77.74	The ratio of COD and total nitrogen
Height	M	0.76 ± 0.21	0.14–1.20	Height of the filter
Eff_NH_4	mg. L^{-1}	14.74 ± 19.14	0.00–155.00	NH_4–N effluent concentration
Eff_COD	mg. L^{-1}	61.50 ± 71.21	2.72–696.00	COD effluent concentration

3.1.2. Statistical test

To determine the influence of the four input variables of SCW on the removal rate, a statistical test (i.e., Student's t-test) was performed. Here, we tested the difference between two sub-categories of each categorical feature to determine whether they were statistically significant or random. The significant differences were calculated between the means of the removal rate of two sub-categories: Type (VF versus HF), Aeration (Aeration versus no Aeration), Filters (Advance versus Normal), and Feeding (intermittent versus continuous), as presented in Table 2.

Based on the t and P values, it can be concluded that the types of filters and aeration did not differ from the COD and NH_4–N removal rates. On the contrary, the types of SCW (i.e., VF versus HF) and feeding forms (i.e., intermittent versus continuous) remarkably influenced the removal rates of COD and ammonium. In addition, the different types of filters and forms of aeration (i.e. natural or artificial aeration using pumps) significantly influenced the removal rates of ammonium and COD, respectively. For SCWs packed with the advanced and normal materials, COD removal rates were 23.36 and 24.85 g COD.m^{-2}.d^{-1}, respectively, while the SCWs with aeration achieved a slightly higher ammonium removal rate of 4.70 g NH_4–N.m^{-2}.d^{-1} when compared to that without aeration of 4.03 g NH_4–N.m^{-2}.d^{-1}. The ammonium removal rate of the SCW with an advanced filter (6.57 g m^{-2}.d^{-1}) was more than two times higher compared to the one that used a normal filter (3.07 g m^{-2}.d^{-1}).

As showin in Table 2, VF achieved noticeably high removal rates for both COD and ammonium of 35.22 and 6.50 g m^{-2}.d^{-1}, respectively; while those for HF were only 10.83 and 1.31 g m^{-2}.d^{-1}, respectively. The relatively common combination of VF, a higher filter, and intermittent flow may enhance the removal efficiency of VF compared to that of HF. For the feeding types, the COD and NH_4–N removal rates of SCW with the continuous inlet flow were approximately four times lower than that of the one with intermittent flow. Specifically, the continuous and intermittent types of SCW obtained removal rates of 11.88 and 51.70 g m^{-2}.d^{-1}, respectively, for COD, and 2.40 and 8.12 g m^{-2}.d^{-1}, respectively, for NH_4–N.

However, the two consequences in Table 2 are unusually different from those in previous reports of SCW. The SCWs supporting artificial aeration resulted in a lower ammonium removal rate (18.07 g m^{-2}.d^{-1}) in comparison to that with no aeration (26.48 g m^{-2}.d^{-1}), while the SCW with advanced materials showed a reduced COD removal efficiency than the normal CW. This explains why the number of SCWs without aeration and with a normal filter may fall into the VF type, which exhibited a higher capacity (347 sets of VF versus 278 sets of HF). In addition, the imbalanced datasets of the variables that were presented by 424 sets of normal versus 201 sets of advanced and 468 sets of No (without aeration) versus 157 sets of Yes (aeration), may affect the final comparison. The previous comparative investigations concluded that a combination of the advanced materials or artificial aeration, among others, to the SCW notably increased the removal efficiency (Fan et al., 2013; Feng et al., 2020; Ge et al., 2019; Hou et al., 2018; Huang et al.,

Table 2
Significant differences between the means of tanks, types of aeration, filters, and feeding for the COD and NH_4–N removal rate.

Variables	Sub-variables	N	COD removal rate ($g.m^{-2}.d^{-1}$)			NH_4–N removal rate ($g.m^{-2}.d^{-1}$)		
			Mean	*t* value	*P*	Mean	*t* value	*P*
Type	VF	347	35.22	13.78	2.2×10^{-16}	6.50	13.78	2.2×10^{-16}
	HF	278	10.83			1.31		
Filter	Advance	201	23.36	−0.58	0.565	6.57	5.43	1.5×10^{-7}
	Normal	424	24.85			3.07		
Feeding	Continuous	429	11.88	−18.49	2.2×10^{-16}	2.40	−9.48	2.2×10^{-16}
	Intermittent	195	51.70			8.12		
Aeration	Yes	157	18.07	−3.58	4×10^{-4}	4.70	1.86	0.06
	No	468	26.48			4.03		

2018; Uggetti et al., 2016).

3.2. Feature selection

As mentioned above, the feature selection step was used to assess and select relevant features for the ML predictive model. The results from the Boruta and REF algorithms are plotted in Fig. 2. All input variables for NH_4–N (Fig. 2a) and COD (Fig. 2b) were confirmed as relevant features for the predictive model. The importance of input variables for predicting ammonium effluent was highest for NH_4–N influent, followed by the HLR, COD influent, height, C:N ratio, aeration, type, feeding, and filter (Fig. 2a). Similarly, five input features were found to be of the highest importance for the COD predictive model, including the influents of ammonium and COD, HLR, height, and C:N ratio (Fig. 2b). The influent concentration was previously considered as a vital input variable for forecasting the effluent concentration, especially for predictive regressions (Babatunde et al., 2011; Chan et al., 2008; Nguyen et al., 2018) and it also ranked the highest in this work. The second important input variable was the HLR, which revealed a high correlation and significant influence on the organic matter and ammonium effluents of sewage wastewater (Ghosh and Gopal, 2010; Nguyen et al., 2018; Yadav et al., 2018). The height of the tank's filter - a design factor of SCW rarely mentioned in previous studies, was rated as important in this study. For a certain surface area of SCW, the increase in material height enhances the pollutant removal by extending the filter's capacity and microbial growth substrate as well as the hydraulic retention time. In addition, the ratio of COD and TN contributed significantly to the predictive model for both COD and NH_4–N. The C:N ratio expresses the source of carbon in relation to nitrogen existing in the SCW, which is essential for optimizing the oxygen supply and the COD:N ratio for increasing the nitrogen removal capacity (Wang et al., 2020a). The available oxygen in the SCW promoted artificially by aeration with the aim of enhancing the ammonium removal was ranked as a relevant input variable; fifth for NH_4–N and ninth for COD. The contribution of aeration in enhancing nitrogen and organic matter has been indicated in many previous studies (Ilyas and Masih, 2017; Jia et al., 2018).

3.3. Prediction of effluents

3.3.1. Comparison of algorithms

A comparison of the results among the five algorithms on the basis of the metrics is presented in Fig. 3. The mean RMSE and R^2 in Fig. 3 indicate that the Cubist algorithm performed the best on the training data for predicting both COD and NH_4–N effluents. Approximately 84% of the variance in the effluent variables (i.e., COD and ammonium) were explained by the selected input variables using the Cubist, in comparison with 42%–82% using other models. The RF also achieved a good accuracy for both target outputs, especially for NH_4–N effluent with an RMSE of 7.99 $mg.L^{-1}$ as compared to 7.77 $mg.L^{-1}$ for Cubist (Fig. 3a). The KNN, CART, and SVM had less prediction efficiency with high RMSE and low R^2 values.

For predicting TN removal in the HF, a simple ANN algorithm (three layers) achieved an R^2 of 0.69 (Akratos et al., 2009). Moreover, the COD effluent from wastewater treatment plants predicted by ANN (three layers: 6: 50: 1) achieved R^2 of 0.87 (Naceureddine and Aziez, 2018) and 0.71 using the RF algorithm (Xin et al., 2020) for the training data. The results of using RF, Cubist, and SVM for predicting different variables from previous research is relatively varied. A prediction of methane production in anaerobic digestion revealed that KNN performed with the best accuracy, i.e. with an RMSE of 26.6, followed by the RF and SVM algorithms as shown by Wang et al. (2020b). Cubist and RF have demonstrated high capacity for estimating the shear strength from the characteristics of the rock-fill material, accounting for 0.96–0.97 in R^2 (Zhou et al., 2019). In addition, RF for the forecasting of metal adsorption onto biochars was found to have a high R^2 of 0.97, as reported by Zhu et al. (2019).

The comparison of forecasting capacity among studies using ML models is qualitative because of the variation in output, specific characteristics, and structure of the data. For instance, KNN worked well with sufficient training data and a sufficiently low dimension (fewer

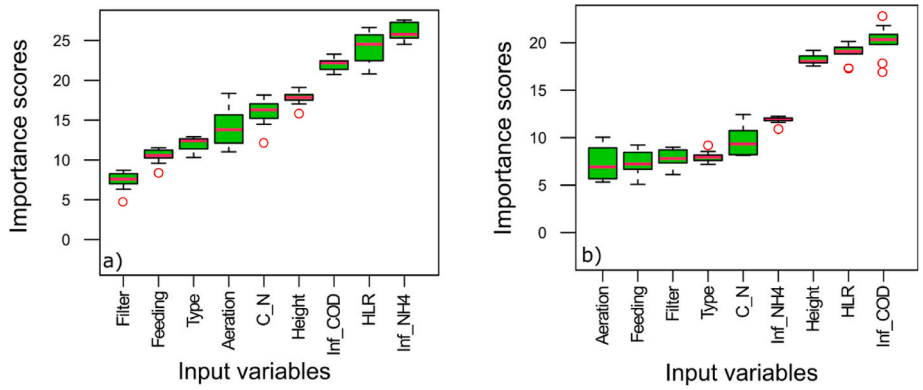

Fig. 2. Results of variables importance for NH_4–N (a) and COD (b). The importance of input features decreased from Inf_NH_4 to filter (Fig. 2a) and Inf_COD to aeration (Fig. 2b).

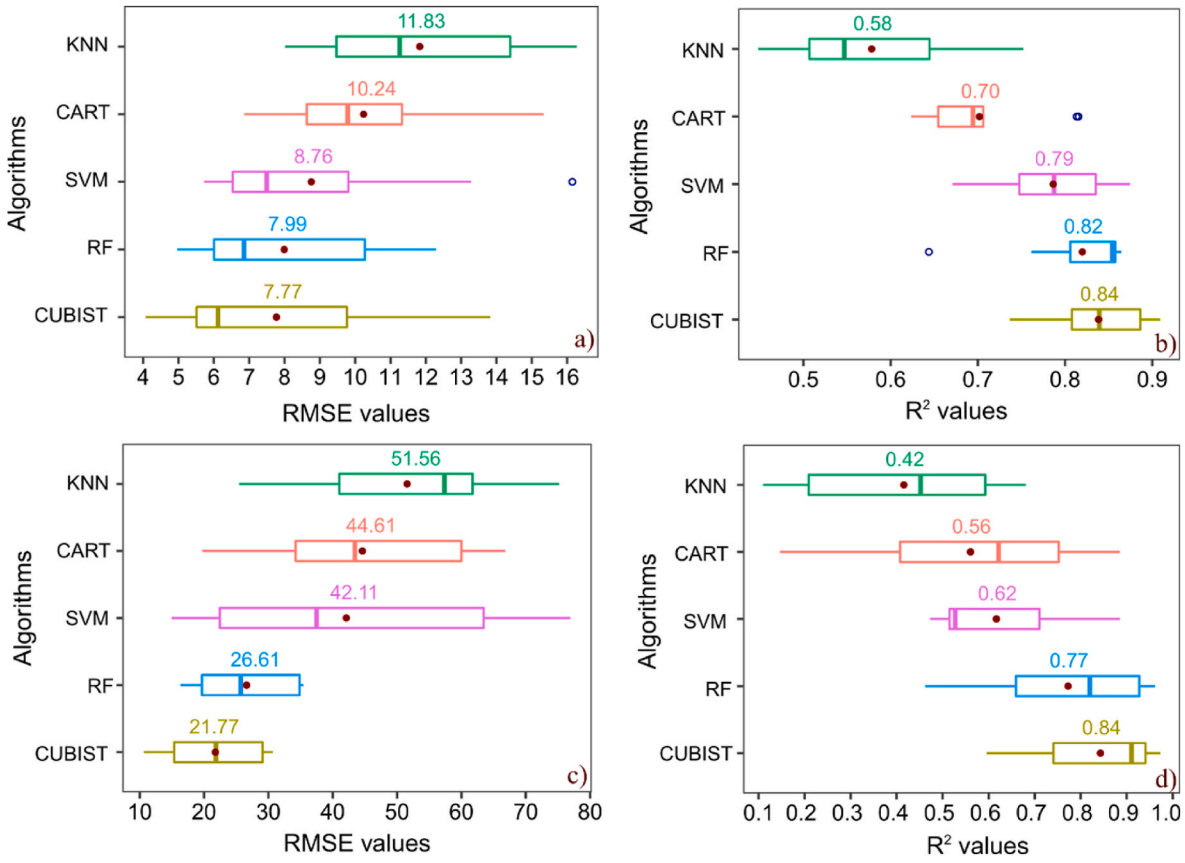

Fig. 3. Comparisons of algorithms with resampling of 10-fold Cross-validation (10 runs): RMSE of NH_4–N (a), and COD (c); R^2 of NH_4–N (b) and COD (d). Red points on the boxplots represent the mean of RMSE of algorithms. The lower the RMSE, the more efficient the model performed and vice versa for R^2. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

input variables) (Forsyth, 2019) while RF has proven to have a superior prediction ability with small sample sizes (Qi, 2012). Therefore, it may be more difficult to predict accurately the output in a complex system.

The selected Cubist algorithm was refined to check whether the accuracy improved by removing the outliers of the data, resampling, transformation, and model tuning. The outliers (values at an abnormal distance from the other values in the data) were removed but the RMSE value did not change (7.77 for COD and 21.77 for NH_4–N). Resampling of 10-fold CV yielded high NH_4–N and COD prediction, while Repeat-CV, Bootstrap, and Leave-one-out CV did not change the accuracy of the algorithm. Data transformation improves the Cubist algorithm, in which standardizing (i.e., center and scale) reduced the RMSE from 7.77 to 7.30 mg mg.L^{-1} for NH_4–N prediction and from 21.77 to 19.50 mg.L^{-1} for COD. The two hyper-parameters of the Cubist algorithm can be tuned, including committee and neighbor. Neighbor is the number of instances used to correct the rule-based prediction, and the committee is the number of boosting operations. Two hyper-parameters were tuned with an extensive range of committees and neighbors, but the model's metric did not change. The optimal Cubist was established at 20 of committee and 5 of neighbor for NH_4–N prediction, and 10 of committee and 5 of neighbor for COD prediction, respectively.

3.3.2. Prediction capacity

To minimize bias in constructing a predictive model, we evaluated the algorithm's capacity using testing data comprising of 20% of the total data, and this data was previously unknown to the model. Thus, 122 sets of data were used to check the Cubist's generalization. Generalization for forecasting the NH_4–N effluent accounted for 0.92 in R^2 and 6.00 mg.L^{-1} in RMSE. The predictive capacity of the Cubist for NH_4–N was enhanced notably, presenting an increase of 8.7% in R^2 and a decrease of 17.8% in RMSE, as compared with the testing data. Similarly, the improvement in the predictive capacity corresponded to COD in the testing data that increased from 0.84 to 0.93 for R^2 and reduced from 19.50 to 12.08 for RMSE. This result indicates that the developed Cubist algorithm will work efficiently with any new SCW data. This is an outstanding feature of this algorithm, compared to other predictive studies using ML that significantly reduced the generalization when fitted with the training data. For example, Zhu et al. (2019) reported a reduction in R^2 of 34% and 4% for ANN and RF, respectively, when fitting with the test data or 15.5% for the ANN model stated by Xin et al. (2020).

The generalization of the developed algorithm is visually shown in terms of the raw residual and standardized Pearson residual, plotted in Fig. 4. The values of standardized residual provides further information to identify the performance of the predicted model, besides the R^2. In general, there is a relatively strong correlation between the Cubist's predicted and actual values. However, the results also indicate some points with standardized residuals more than ±3, meaning the unsatisfying performance of predictive models. The min-max of the residual for predicting the NH_4–N and COD effluents were −18.35 – 15.42 and −65.81 – 52.10, respectively. The predicted values for the NH_4–N effluent tended to be higher than the actual data and expressed more negative residuals (53%), while the residuals for the COD prediction were balanced.

From the results derived from developing the model with the training and testing data, a final Cubist algorithm was established with the entire data set. This Cubist model for predicting the NH_4–N effluent includes 20 committees, five neighbors, *10-fold CV, Center,* and *Scale* (transformation), and the COD effluent consists of 10 committees, 5 neighbors, *10-fold CV*, *Center*, and *Scale*. The categorical input variables were also

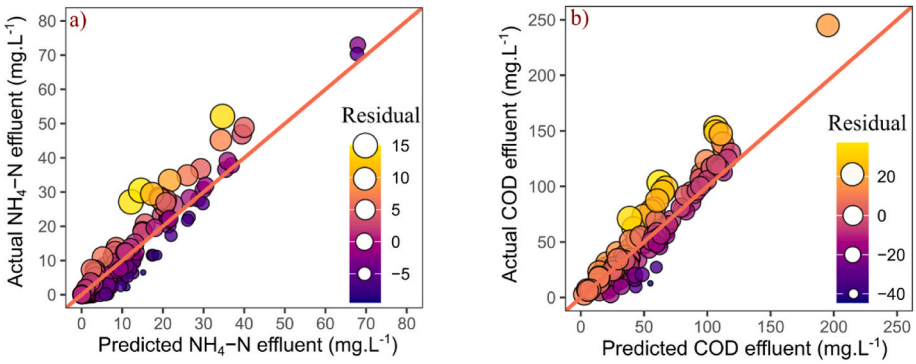

Fig. 4. The diagnostic plots of predicted and real values for prediction of NH_4–N (a) and COD effluents (b). The size of circles and color tones denote the standardized residuals. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

converted into numerical variables by *One-hot coding*. Thus, 13 predictors were used for this Cubist algorithm to predict the effluent NH_4–N and COD concentrations.

3.3.3. Case studies and proposed design tool

To elucidate the application of the ML algorithm in the SCW design support, we used the case studies in Table 3S. The value and information of the input variables, such as influent concentration, flow rate (Q, $m^3.d^{-1}$), and surface area (A, m^2) were derived from the range of characteristics of the SCW data and were exerted for developing the ML model. To assess the potential to meet the general discharge limits for the NH_4–N and COD effluents, we used the standards of 5 and 50 $mg.L^{-1}$, respectively, in-line with those regulated in several countries' water discharge guidelines (Nguyen et al., 2017, 2020).

An example of four runs (i.e., run 1, 2 …) were employed with the adjustment of the input variables to obtain the desired outputs, as shown in Table 3S. The Cubist algorithm developed above was used to predict the COD and NH_4–N effluents. Some rules of thumb was applied for choosing the HLR and the random selection of SCW characteristics was applied to establish the Cubist's input variables. For run 1, only in case 2 did the effluents fulfill the discharge limits, showing lower NH_4–N and COD values than 5 $mg.L^{-1}$ and 50 $mg.L^{-1}$, respectively. In run 2, by adjusting the HLR, type, aeration, height, and feeding, the effluents improved with case 5 in terms of both and cases 3 and 4 for NH_4–N regarding meeting the discharge standards. Run 3 was performed with alterations of the inputs and it achieved significant improvements when the effluents of all cases were lower than the discharge limits. Finally, run 4 was implemented to determine the chance of reducing the cost of SCW, with a decrease in the surface area, and the filter replacement in case 3 and the final output was deemed acceptable.

From the above results of developing the predictive model and demonstrating the application with five design case studies of SCW, we suggest an ML-based design support tool, which is illustrated in Fig. 5. The design inputs of SCW consist of influent concentrations and the flow rate of the wastewater, while environmental requirements and resources cover effluent standards, potential finance, land availability, and local conditions, such as local materials, plants, and esthetic requirements. These are vital inputs for a projected plan of a SCW which uses a simple and fast way, namely the rule of thumb, to draw the design parameters. From the projected plan with the selected variables, the ML algorithm was used to forecast the effluents of SCW in terms of COD and NH_4–N. This prediction is compared to the discharge limits to decide whether to accept or reject the projected plan. In case of dissatisfaction with the discharge standards, a process of adjustment of the projected plan will be implemented. Otherwise, a rough plan is used to optimize the final design of SCW. The relationship between the effluent concentrations or removal rate and the input features elucidated above (Section 3.1) will assist in the adjustment of the projected plan. Although there was a low

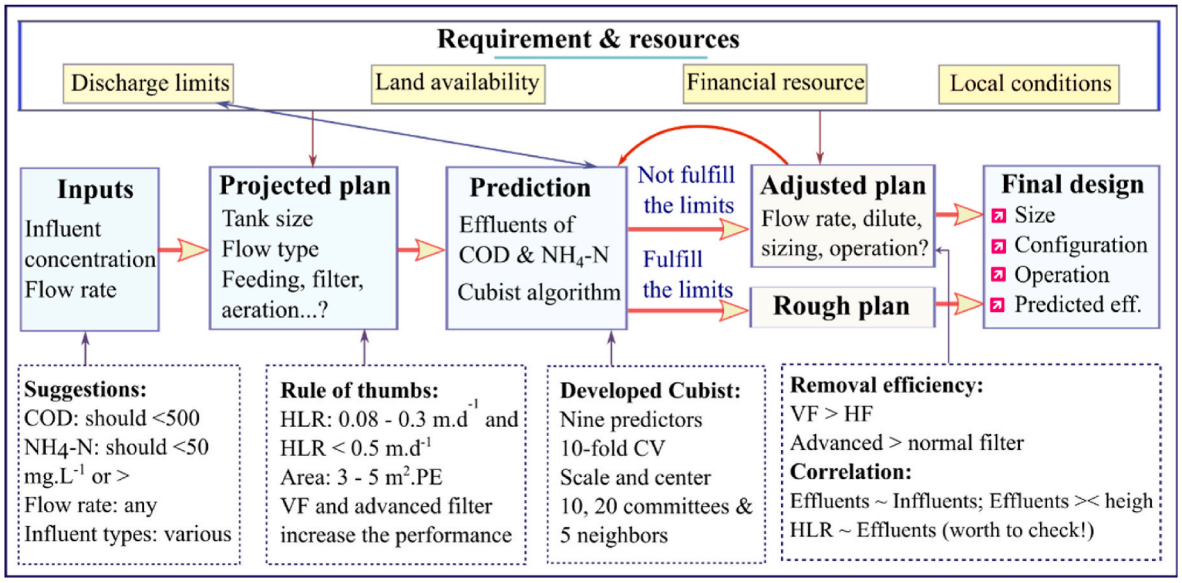

Fig. 5. Proposed tool for ML-based subsurface constructed wetland design. A design cycle that goes from left to right and dashed boxes provide tools and additional information for design.

correlation between the effluents and HLR in this analysis, many individual studies of SCW proved that the decrease in HLR (i.e., reduce inlet flow rate per area of treatment or rise in HRT) resulted in improved water quality (Ghosh and Gopal, 2010; Sgroi et al., 2018). Furthermore, the great contribution of HLR, only after influent concentration, for the predictive ML model for both ammonium and COD effluents is confirmed above. As a result, it is necessary to alter the HLR based on the surface area in order to establish whether or not the effluent concentrations would vary.

In summary, SCW experiments for wastewater treatment are costly and time-consuming. Experiments using SCW are a waste of money and natural resources unless the research aims to discover something new. For treating general wastewater by SCW, in which organic matter and ammonium concentrations are the common parameters, the ML predictive model based on available data would be practical for supporting the design. Discharge standards of certain nations or areas are different. Thus the application of effluent quality prediction-based SCW designs will reduce the environmental violations on discharge and contribute to sustainable environmental management. It is considered that this work is fast and easy with little cost compared to launching a new experiment of SCW.

Although the proposed technique is valuable for SCW design, it has its own limitations. The generalization of the developed Cubist algorithm for NH_4–N and COD predictions achieved relatively high (R^2 = 0.92–0.93); however, the results of standardized residual presented several outliers – meaning the low performance of the predictive model in some cases. This could be due to the inherent complexity of the SCW system, or because the data size and its scale was not great enough. In addition, this study only focuses on two main effluents of the SCW. Therefore, more water quality parameters such as phosphate, pathogens, and heavy metals should be included in future research.

4. Conclusions

Data from the literature of SCWs (i.e., 618 sets and 10 features) was used to train and develop different data driven predictive models. From the comparative results of the five ML algorithms, the Cubist was found to be the optimal algorithm. The Cubist for fitting the training data achieved the lowest RMSE (7.77 mg mg.L^{-1} for NH_4–N and 21.77 mg mg.L^{-1} for COD) which corresponded to 84% of the variance in the effluents explained. The high Cubist's generalization for predicting NH_4–N and COD using the testing data was found to be 0.92 and 0.93 in R^2, respectively. Finally, a developed Cubist algorithm and an ML-based design tool for SCW was proposed that is fast and easy with little cost compared to launching a new experiment of SCW.

Author contribution statement

Xuan Cuong Nguyen Conceptualization, Writing – original draft. Thi Thanh Huyen Nguyen Conceptualization, Writing – original draft. Quyet V. Le Validation, Writing – review & editing Le Phuoc Cuong Methodology, Validation Arun Lal Srivastav Validation, Writing – review & editing Quoc Bao Pham Normal analysis, Investigation, Data curation Phuong Minh Nguyen Formal analysis, Writing – review & editing Duc Duong La Methodology, Investigation Eldon R. Rene Formal analysis, Writing – review & editing H. Hao Ngo Conceptualization, Writing – review & editing S. Woong Chang Funding acquisition, Project administration D. Duc Nguyen Supervision, Project administration, Writing – review & editing.

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 paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2021.113868.

References

Abou-Elela, S.I., Golinielli, G., Abou-Taleb, E.M., Hellal, M.S., 2013. Municipal wastewater treatment in horizontal and vertical flows constructed wetlands. Ecol. Eng. 61, 460–468.

Akratos, C.S., Papaspyros, J.N.E., Tsihrintzis, V.A., 2009. Total nitrogen and ammonia removal prediction in horizontal subsurface flow constructed wetlands: use of artificial neural networks and development of a design equation. Bioresour. Technol. 100, 586–596.

Albuquerque, A., Oliveira, J., Semitela, S., Amaral, L., 2009. Influence of bed media characteristics on ammonia and nitrate removal in shallow horizontal subsurface flow constructed wetlands. Bioresour. Technol. 100, 6269–6277.

Babatunde, A.O., Zhao, Y.Q., Doyle, R.J., Rackard, S.M., Kumar, J.L., Hu, Y.S., 2011. Performance evaluation and prediction for a pilot two-stage on-site constructed wetland system employing dewatered alum sludge as main substrate. Bioresour. Technol. 102, 5645–5652.

Benny, C., Chakraborty, S., 2020. Continuous removals of phenol, organics, thiocyanate and nitrogen in horizontal subsurface flow constructed wetland. Journal of Water Process Engineering 33, 101099.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., J, F., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman and Hall, New York.

Chan, S.Y., Tsang, Y.F., Cui, L.H., Chua, H., 2008. Domestic wastewater treatment using batch-fed constructed wetland and predictive model development for NH3-N removal. Process Biochem. 43, 297–305.

Chang, J.J., Liang, K., Wu, S.Q., Zhang, S.H., Liang, W., 2015. Comparative evaluations of organic matters and nitrogen removal capacities of integrated vertical-flow constructed wetlands: domestic and nitrified wastewater treatment. J Environ Sci Health A Tox Hazard Subst Environ Eng 50, 757–766.

Chollet, F., Allaire, J.J., 2018. Deep Learning with R. Manning Publications Co., New York, United States.

Cutler, A., Cutler, D., Stevens, J., 2011. Random forests. In: Zhang, C., Ma, Y. (Eds.), Ensemble Machine Learning: Methods and Applications. Springer US, Boston, MA, pp. 157–176.

Dobson, A., Barnett, A., 2008. An Introduction to Generalized Linear Models, third ed.

Fan, J., Zhang, B., Zhang, J., Ngo, H.H., Guo, W., Liu, F., Guo, Y., Wu, H., 2013. Intermittent aeration strategy to enhance organics and nitrogen removal in subsurface flow constructed wetlands. Bioresour. Technol. 141, 117–122.

Feng, L., Wang, R., Jia, L., Wu, H., 2020. Can biochar application improve nitrogen removal in constructed wetlands for treating anaerobically-digested swine wastewater? Chem. Eng. J. 379, 122273.

Foladori, P., Ruaben, J., Ortigara, A.R.C., 2013. Recirculation or artificial aeration in vertical flow constructed wetlands: a comparative study for treating high load wastewater. Bioresour. Technol. 149, 398–405.

Forsyth, D., 2019. Applied Machine Learning. Springer International Publishing, Cham, Switzerland.

Ge, Z., Wei, D., Zhang, J., Hu, J., Liu, Z., Li, R., 2019. Natural pyrite to enhance simultaneous long-term nitrogen and phosphorus removal in constructed wetland: three years of pilot study. Water Res. 148, 153–161.

Ghosh, D., Gopal, B., 2010. Effect of hydraulic retention time on the treatment of secondary effluent in a subsurface flow constructed wetland. Ecol. Eng. 36, 1044–1051.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. KNN Model-Based Approach in Classification.

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y.M., Park, J.-p., Kim, J.H., Cho, K.H., 2015. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. J. Environ. Sci. 32, 90–101.

Haberl, R., Perfler, R., Mayer, H., 1995. Constructed wetlands in europe. Water Sci. Technol. 32, 305–315.

Haupt, S., Pasini, A., Marzban, C., 2009. Artificial Intelligence Methods in the Environmental Sciences.

Hou, J., Wang, X., Wang, J., Xia, L., Zhang, Y., Li, D., Ma, X., 2018. Pathway governing nitrogen removal in artificially aerated constructed wetlands: impact of aeration mode and influent chemical oxygen demand to nitrogen ratios. Bioresour. Technol. 257, 137–146.

Huang, J., Yan, C.-N., Cao, C., Peng, C., Liu, J.-L., Guan, W.-Z., 2018. Performance evaluation of Iris pseudacorus constructed wetland for advanced wastewater treatment under long-term exposure to nanosilver. Ecol. Eng. 116, 188–195.

Ilyas, H., Masih, I., 2017. Intensification of constructed wetlands for land area reduction: a review. Environ. Sci. Pollut. Res. Int. 24, 12081–12091.

Jia, L., Wang, R., Feng, L., Zhou, X., Lv, J., Wu, H., 2018. Intensified nitrogen removal in intermittently-aerated vertical flow constructed wetlands with agricultural biomass: effect of influent C/N ratios. Chem. Eng. J. 345, 22–30.

Kadlec, R.H., Knight, R.L., 1996. Treatment Wetlands. Lewis Publishers, Boca Raton, FL.

Kadlec, R.H., Wallace, S.D., 2009. Treatment Wetlands, second ed. CRC Press, Florida, USA.

Krzywinski, M., Altman, N., 2017. Classification and regression trees. Nat. Methods 14, 757–758.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York, NY, pp. 27–59.
Kursa, M., Rudnicki, W., 2010. Feature selection with Boruta package. J. Stat. Software 36, 1–13.
Lantz, B., 2019. Machine Learning with R: Expert Techniques for Predictive Modeling. Packt Publishing.
Li, X., Ding, A., Zheng, L., Anderson, B.C., Kong, L., Wu, A., Xing, L., 2018. Relationship between design parameters and removal efficiency for constructed wetlands in China. Ecol. Eng. 123, 135–140.
Liaw, A., Wiener, M., 2001. Classification and Regression by RandomForest, vol. 23. Forest.
Luo, P., Liu, F., Zhang, S., Li, H., Chen, X., Huang, X., Xiao, R., Wu, J., 2020. Nitrogen removal performance and needed area estimation of surface-flow constructed wetlands using a probabilistic approach. J. Environ. Manag. 255, 109881.
Moran, S., 2018. Appendix 4 - selection and sizing of unit operations. In: Moran, S. (Ed.), An Applied Guide to Water and Effluent Treatment Plant Design. Butterworth-Heinemann, pp. 381–398.
Naceureddine, B., Aziez, Z., 2018. Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant. Manag. Environ. Qual. Int. J. 30.
Nguyen, X.C., Chang, S.W., Nguyen, T.L., Ngo, H.H., Kumar, G., Banu, J.R., Vu, M.C., Le, H.S., Nguyen, D.D., 2018. A hybrid constructed wetland for organic-material and nutrient removal from sewage: process performance and multi-kinetic models. J. Environ. Manag. 222, 378–384.
Nguyen, X.C., Ly, Q.V., Li, J., Bae, H., Bui, X.-T., Nguyen, T.T.H., Tran, Q.B., Vo, T.-D.-H., Nghiem, L.D., 2021. Nitrogen Removal in Subsurface Constructed Wetland: Assessment of the Influence and Prediction by Data Mining and Machine Learning. Environmental Technology & Innovation, p. 101712.
Nguyen, X.C., Nguyen, D.D., Thi Loan, N., Chang, S.W., 2017. Potential of integrated vertical and horizontal flow constructed wetland with native plants for sewage treatment under different hydraulic loading rates. Water Sci. Technol. 76, 434–442.
Nguyen, X.C., Tran, T.C.P., Hoang, V.H., Nguyen, T.P., Chang, S.W., Nguyen, D.D., Guo, W., Kumar, A., La, D.D., Bach, Q.-V., 2020. Combined biochar vertical flow and free-water surface constructed wetland system for dormitory sewage treatment and reuse. Sci. Total Environ. 713, 136404.
Qi, Y., 2012. Random forest for bioinformatics. In: Zhang, C., Ma, Y. (Eds.), Ensemble Machine Learning: Methods and Applications. Springer US, Boston, MA, pp. 307–323.
Quinlan, R., 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference On Artificial Intelligence, Australian, pp. 343–348.
Quinlan, R., 1993. Combining instance-based and model-based learning. Proceedings of the Tenth International Conference on Machine Learning (United States).
Romeo, L., Loncarski, J., Paolanti, M., Bocchini, G., Mancini, A., Frontoni, E., 2020. Machine learning-based design support system for the prediction of heterogeneous machine parameters in industry 4.0. Expert Syst. Appl. 140, 112869.
Rousseau, D.P.L., Vanrolleghem, P.A., De Pauw, N., 2004. Model-based design of horizontal subsurface flow constructed treatment wetlands: a review. Water Res. 38, 1484–1493.
Sgroi, M., Pelissari, C., Roccaro, P., Sezerino, P.H., García, J., Vagliasindi, F.G.A., Ávila, C., 2018. Removal of organic carbon, nitrogen, emerging contaminants and fluorescing organic matter in different constructed wetland configurations. Chem. Eng. J. 332, 619–627.
Sklarz, M.Y., Gross, A., Yakirevich, A., Soares, M.I.M., 2009. A recirculating vertical flow constructed wetland for the treatment of domestic wastewater. Desalination 246, 617–624.
Stefanakis, A.I., Tsihrintzis, V.A., 2012. Effects of loading, resting period, temperature, porous media, vegetation and aeration on performance of pilot-scale vertical flow constructed wetlands. Chem. Eng. J. 181–182, 416–430.
Uggetti, E., Hughes-Riley, T., Morris, R.H., Newton, M.I., Trabi, C.L., Hawes, P., Puigagut, J., García, J., 2016. Intermittent aeration to improve wastewater treatment efficiency in pilot-scale constructed wetland. Sci. Total Environ. 559, 212–217.
USEPA, 2000. Constructed Wetlands Treatment of Municipal Wastewaters. United States Environmental Protection Agency, USA.
Vo, T.-D.-H., Bui, X.-T., Nguyen, D.-D., Nguyen, V.-T., Ngo, H.-H., Guo, W., Nguyen, P.-D., Nguyen, C.-N., Lin, C., 2018. Wastewater treatment and biomass growth of eight plants for shallow bed wetland roofs. Bioresour. Technol. 247, 992–998.
Walker, J.S., 2018. Machine learning for beginners: your ultimate guide to machine learning for absolute beginners, machine learning guide, scikit-learn, deep learning, TensorFlow, data analytics, Python, data science. CreateSpace Independent Publishing Platform.
Wang, J., Hou, J., Xia, L., Jia, Z., He, X., Li, D., Zhou, Y., 2020a. The combined effect of dissolved oxygen and COD/N on nitrogen removal and the corresponding mechanisms in intermittent aeration constructed wetlands. Biochem. Eng. J. 153, 107400.
Wang, L., Long, F., Liao, W., Liu, H., 2020b. Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. Bioresour. Technol. 298, 122495.
Wang, W., Gao, J., Guo, X., Li, W., Tian, X., Zhang, R., 2012. Long-term effects and performance of two-stage baffled surface flow constructed wetland treating polluted river. Ecol. Eng. 49, 93–103.
Wood, A., 1995. Constructed wetlands in water pollution control: fundamentals to their understanding. Water Sci. Technol. 32, 21–29.
Wu, S.-q., Zhang, J., Ngo, H.H., Guo, W., Hu, Z., Liang, S., Fan, J., Liu, H., 2015. A review on the sustainability of constructed wetlands for wastewater treatment: design and operation. Bioresour. Technol. 175, 594–601.
Wynn, T.M., Liehr, S.K., 2001. Development of a constructed subsurface-flow wetland simulation model. Ecol. Eng. 16, 519–536.
Xin, C., Shi, X., Wang, D., Yang, C., Li, Q., Liu, H., 2020. Multi-grained cascade forest for effluent quality prediction of papermaking wastewater treatment processes. Water Sci. Technol. 81, 1090–1098.
Yadav, A., Chazarenc, F., Mutnuri, S., 2018. Development of the “French system” vertical flow constructed wetland to treat raw domestic wastewater in India. Ecol. Eng. 113, 88–93.
Zhai, J., Xiao, H.W., Kujawa-Roeleveld, K., He, Q., Kerstens, S.M., 2011. Experimental study of a novel hybrid constructed wetland for water reuse and its application in Southern China. Water Sci. Technol. 64, 2177–2184.
Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., Jahed Armaghani, D., 2019. Random forests and cubist algorithms for predicting shear strengths of rockfill materials. Appl. Sci. 9, 1621.
Zhu, X., Wang, X., Ok, Y.S., 2019. The application of machine learning methods for prediction of metal sorption onto biochars. J. Hazard Mater. 378, 120727.