Vertical flow constructed wetlands using expanded clay and biochar for wastewater remediation: A comparative study and prediction of effluents using machine learning

Xuan Cuong Nguyen,[a,b+] *Quang Viet Ly,*[c+] *Wanxi Peng,*[d] *Van-Huy Nguyen,*[e,,f] *Dinh Duc Nguyen,*[g,h] *Quoc Ba Tran,*[a,b] *Thi Thanh Huyen Nguyen,*[a,b] *Christian Sonne,*[i] *Su Shiung Lam,*[j] *Huu Hao Ngo,*[k] *Peter Goethals,*[l] *Quyet Van Le,*[a,]*

[a]Laboratory of Energy and Environmental Science, Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

[b]Faculty of Environmental and Chemical Engineering, Duy Tan University, Da Nang 550000, Vietnam

[c]School of Environmental Science and Engineering, Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Sun Yat-sen University, Guangzhou, 510275, China

[d]Henan Province Engineering Research Center for Biomass Value-added Products, School of Forestry, Henan Agricultural University, Zhengzhou 450002, China

[e]Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam; nguyenvanhuy@tdtu.edu.vn

[f]Faculty of Applied Sciences, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[g]Faculty of Environmental and Food Engineering, Nguyen Tat Thanh University, 300A Nguyen Tat Thanh, District 4, Ho Chi Minh City, 755414, Vietnam

[h]Department of Environmental Energy Engineering, Kyonggi University, Suwon 16227, Republic of Korea

[i]Aarhus University, Department of Bioscience, Arctic Research Centre (ARC), Frederiksborgvej 399, PO Box 358, DK-4000 Roskilde, Denmark

[j]Higher Institution Centre of Excellence (HICoE), Institute of Tropical Aquaculture and Fisheries (AKUATROP), Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

[k]Centre for Technology in Water and Wastewater, School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, NWS, 2007, Australia

[l]Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Jozef Plateaustraat 22, B-9000 Ghent, Belgium

Corresponding authors: Email: levanquyet@dtu.edu.vn (Q. V. Le)

Abstract

This study evaluated and compared the performance of two vertical flow constructed wetlands (VF) using expanded clay (VF_1) and biochar (VF_2), of which both are low-cost, eco-friendly, and exhibit potentially high adsorption as compared to conventional filter layers. Both VFs achieved relatively high removal for organic matters (i.e. Biological oxygen demand during 5 days, BOD_5) and nitrogen, accounting for 9.5 – 10.5 $g \cdot BOD_5 \cdot m^{-2} \cdot d^{-1}$ and 3.5 – 3.6 $g \cdot NH_4\text{-}N \cdot m^{-2} \cdot d^{-1}$, respectively. The different filter materials did not exert any significant discrepancy to effluent quality in terms of suspended solids, organic matters and NO_3-N (P>0.05), but they did influence NH_4-N effluent as evidenced by the removal rate of that by VF_1 and VF_2 being of 82.4 ± 5.7 and 84.6 ± 6.4%, respectively (P<0.05). The results obtained from the designed systems were further subject to machine learning to clarify the effecting factors and predict the effluents. The optimal algorithms were random forest, generalized linear model, and support vector machine. The values of the coefficient of determination (R^2) and the root mean square error (RMSE) of whole fitting data achieved 74.0% and 5.0 $mg \cdot L^{-1}$, 80.0% and 0.3 $mg \cdot L^{-1}$, 90.1% and 2.9 $mg \cdot L^{-1}$, and 48.5% and 0.5 $mg \cdot L^{-1}$ for BOD_5_VF_1, NH_4-N_VF_1, BOD_5_VF_2, and NH_4-N_VF_2, respectively.

Keywords: Biochar; Constructed wetland; Expanded clay; Machine learning; Vertical flow.

Abbreviations:

ML	:	Machine learning
BOD_5	:	Biological oxygen demand during 5 days,
SVM	:	Support vector machine
RMSE	:	Root mean square error
R^2	:	Coefficient of determination
GLM	:	Generalized linear model
VF_1	:	Expanded clay vertical flow constructed wetland
VF_2	:	Biochar vertical flow constructed wetland
KNN	:	K nearest neighbor
RF	:	Random forest
CW	:	Constructed wetland
ExC	:	Expanded clay
TSS	:	Suspended solids
R^2	:	Coefficient of determination
CV	:	Cross-validation
HLR	:	Hydraulic loading rate
LM	:	Linear regression model

1. Introduction

Wastewater reclamation is widely recognized as one of the most promising approaches to achieve sustainable water management worldwide (Zhang et al., 2020). In this context, constructed wetland (CW) emerge as natural and low-cost technology extensively applied for wastewater treatment for decades. Vertical flow constructed wetland (VF), one of the most common CW types, is used frequently as the central unit in multi-stage CW system due to their merits of high treatment efficacy concerning organic matters (Biological oxygen demand during 5 days - BOD_5, Chemical oxygen demand - COD), nutrients (Nitrogen, Phosphorus) and pathogenic microorganisms (Abou-Elela and Hellal, 2012, Cooper, 2005, Vymazal, 2007). However, the requirement for vast land is a critical hurdle for widespread application especially in a densely populated area (Ilyas and Masih, 2017). Therefore, the improvement of the VF performance is of critical importance to tackle these issues in a bid to warrant the success of the overall CW system for sewage treatment. Of those factors, the filter substrate is a foremost factor that remarkably influences VF efficacy (Wu et al., 2014). Previous studies demonstrated that expanded clay (ExC) enhanced the removal efficiency (e.g. phosphate) accredited to the porous matrix that provided great adsorption sites for biofilm development (Calheiros et al., 2009b) and hydraulic conductivity (Mlih et al., 2020). Dordio and Carvalho (2013) revealed ExC-CW system obtained an overall high capacity (>80%) to treat typical pollutants (i.e. TSS, COD, and nitrogen) and hazardous matters such as polyphenols, pharmaceutical, and a pesticide, with a retention time of 3 and 9 days. Recently, much interests have been drawn for biochar, which outperform conventional filter layers as they could spur plant growth, improve soil quality, and adsorb pollutants from water (Kasak et al., 2018), leachate (Joseph et al., 2020), secondary wastewater effluent (Odedishemi Ajibade et al., 2021), and domestic wastewater (Jia et al., 2020). Besides CW packed biochar demonstrated the high reduction of nitrate (78%), phosphate (70%), and COD (65%), it also supported the removal of contaminants and odor from intensified leachate (COD 4000 – 14000 $mg \cdot L^{-1}$, ammonia 760 – 900 $mg \cdot L^{-1}$) (Joseph et al., 2020). The system of Fe-biochar – CW illustrated the superiority than normal CW in pollution removal (NH_4-N 86.33%, and COD 63.36%) and abundances of genes involved in nitrogen removal (Jia et al., 2020). Besides, flow regime in CW is another important factor affecting CW systems such as recirculation (Decezaro et al., 2019, Prost-Boucle and Molle, 2012, Torrijos et al., 2016), two-stage system (Kim et al., 2014, Nguyen et al., 2018, Saeed et al., 2019) and step – feeding operation (Patil and Chakraborty, 2017, Wang et al., 2020).

Meanwhile, from the available results of the pilot CW systems, the questions were raised what factors influence the effluents and whether to rely on that data to predict the performance and from that support to design new treatment systems. To model the treatment system, numerous mathematical modeling strides have been made, including the first-order model (Cooper et al., 1996, Kadlec, 2000), Monod kinetics with different flow patterns, or combined Monod first-order model (Rousseau et al., 2004). These first-order models depend on influent/effluent and do not rely on some factors, such as hydraulic loading rate (HLR) and environmental conditions. However, the efficacy of the CW system is also largely governed by biological processes and time with highly nonlinear characteristics (Guo et al., 2015). As a consequence, these linear kinetic models are unable to describe the observed mechanisms, and thus could not be used for system design (Kadlec, 2000). To address this drawback, several approaches developed to model the CW were introduced, e.g., multiple regression (Babatunde et al., 2011, Murray-Gulde et al., 2008, Nguyen et al., 2018), artificial neural networks and principal component analysis (Akratos et al., 2008). Of those, machine learning (ML) has paid considerable attention to wastewater treatment, particularly for CW systems. This tool offers superior benefits since it directly predict output values from the input of complex treatment systems with high accuracy. As an example, Hijosa-Valsero et al. (2011) used four statistical models, including two ML algorithms (clustering tree diagrams and regression trees), to predict the removal of organic matter and pharmaceuticals by CW. This study indicated that the removal efficiency of many parameters of water quality was not linear with input variables indicating the low values of R^2 of predictive model while some others achieved only 0.5 – 0.65 in R^2. Other comprehensive studies consisting of training and validation, which used and compared ML algorithms for evaluating the effluent concentration and the water quality have also been launched (Chen et al., 2020, Manu and Thalla, 2017, Wu et al., 2015). The random forest (RF) algorithm was also of interest in given aquatic systems. Zhou et al. (2019) developed the RF to predict the influent flow of two wastewater treatment plants that obtained R^2 of 0.58 (Humber plant) and 0.72 (confidential plant) for testing data. Time series forecasting of chlorophyll-a in these two water bodies using RF, achieving a range of 0.36 to 0.52 in R^2 also reported by Yajima and Derot (2017). Until now, there have not yet been studied using those ML algorithms for VF. Moreover, despite extensive studies to investigate ML algorithms that have been made for various water and wastewater, the importance of variables and feature selection, which increases the algorithms' accuracy, has been still unsolved effectively. In other words, the ML application in the field of wastewater has yet been comprehensively understood in terms of technique and tool.

According to our best knowledge there is a lack of comparison between biochar and ExC and the flow regimes in the evaluation of CW performance particularly in VF systems for sewage treatment. In this work, the performance of two VF tanks for wastewater treatment and the feasibility of predicting VF's effluents using six ML algorithms were elucidated. In addition, a sequence of techniques was further taken into account including descriptive statistics and visualization, feature selection, algorithm evaluation and tuning. They help eliminate any unnecessary variables, improve the accuracy of the model, and reduce the computation time, and thus the overall expenses.

2. Materials and methods

2.1. Pilot-scale treatment system

2.1.1. Description of the treatment system

Two VF tanks run in parallel with the same wastewater influent presented in Fig. 1. The wastewater was pumped directly from the internal sewer system into two VFs via perforated pipes. The tanks were steel-made rectangular prisms with 0.5 × 0.5 × 1.0 m (length × width × height). Each VF comprised four material layers with a total working height of 0.8 m. While filter layers in the VF_1 followed the order from the top to the bottom of the sandy soil, sand, ExC and gravel with the corresponding height of 10, 20, 40 and 10 cm (Nguyen et al., 2020a), respectively, that in the VF_2 was placed in the same order (top-bottom) of sandy soil, sand, biochar, and gravel with the height of 10, 20, 40, and 10 cm (Nguyen et al., 2020b), respectively. Sandy soil is top, where creating the substrate for plants to grow and sand layer helps to stabilize the soil substrate. Main layers of biochar and ExC were placed in the middle of the tanks as the main treatment areas. Also, the bottom gravel of VF plays a role in the drainage layer. The sandy soil collected from river mudflats was a mixture of the majority of sand (~87%) and a little part of soil (humus, ~13%). The gravel size varied from 2 to 3 cm and sand had a smaller diameter of 2 mm. The biochar was produced from wattle bark by heating the material to 500 °C at the heating rate of 10 °C/min in a furnace for 2 hours. The mean diameter of the biochar fell within 1–3 cm. The ExC purchased from a local factory (Dang Gia Trang Co., Ltd, Vietnam) was fabricated in the furnace at 1,200 °C and had an average diameter of 0.2–1.0 cm with a density of 600 kg/m^3. Biochar and ExC are effective materials supporting the wastewater treatment processes in CW that partly deployed in the prior investigations. A biochar from waste embedded subsurface CW used to treat leachate that achieved 78 and 65% in the removal of nitrate and COD, respectively (Joseph et al., 2020). In addition, the reduction of 89.1 and 90.2% for COD and nitrogen, respectively, reported by Odedishemi Ajibade et al. (2021) in the system of non-aerated biochar amended VF treating secondary wastewater. Furthermore, a combination of intermittent aeration, biochar, and Fe-modified biochar investigated for enhancing treatment performance and identifying the potential risk of substrate clogging (Zhou et al., 2020) and improving microbial nitrogen removal capability in horizontal subsurface (Jia et al., 2020). The substrate of expanded clay integrated in horizontal subsurface flow CW for treating tannery wastewater (Calheiros et al., 2009a) and agriculture effluent (Dordio and Carvalho, 2013). These systems enhanced the removal of pollutants from wastewater (e.g. phosphate, ammonium) through adsorption and biological pathways (Mlih et al., 2020). The local tree elephant ear (*Colocasia esculenta*) was planted in both VF_1 and VF_2. Seedling of plants was collected from a home garden, subsequently cut into the length of 25 cm, and planted with 10 cm of space (16 seedlings in each tank). More detailed information on the VF tanks could be found elsewhere (Nguyen et al., 2020a, Nguyen et al., 2020b).

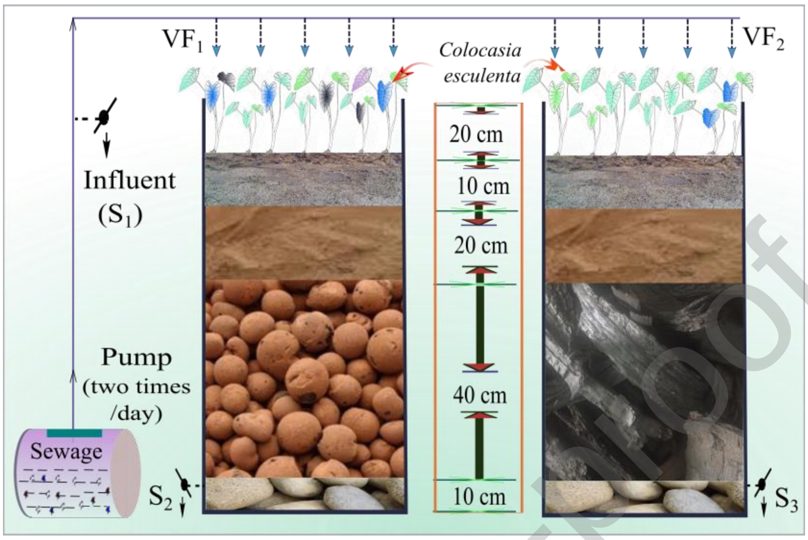

Fig. 1. Diagram of two vertical flow constructed wetland tanks

2.1.2. Sample analysis and operation

Wastewater samples were taken at different VF sites, i.e., the influent (S_1) and effluents of the VF_1 (S_2) and VF_2 (S_3) (Fig. 1). For each tank, a total of 4, 12, 12, and 10 sample sets were collected during stages of I, II, III, and IV, respectively. BOD_5 (5210B), COD (5220D), NH_4-N (4500-NH_3 F), NO_3-N (4500 NO_3-B), and TSS (2540D) were analyzed using standard methods (APHA/WEF/AWWA, 2012). Besides, pH was measured using a multi-parameter water quality meter (HQ40D; Hach, USA). The pollutants used for assessing the VF capacity are presentative from domestic wastewater. These are nitrogen (NH_3-N, NO_3-N), organic matter (BOD_5, COD), and suspended solids (TSS) that regulated in the national effluent standards.

The wastewater was directed from the internal dormitory sewer to the VF_1 and VF_2 systems. The treatment tanks operated in four different flow periods that consisted of a start-up period with a mixing ratio of wastewater and tap water (1:1) to reach an HLR of 0.02 $m{\cdot}d^{-1}$

and the next three stages, which were operated at increasing HLRs, i.e., 0.04, 0.06 and 0.12 $m \cdot d^{-1}$ for stage II, III and IV, respectively. The HLR was calculated by the volumetric flow rate divided by area. It is an important parameter commonly used for the CW design; higher HLR meaning higher hydraulic retention time.

2.3. Machine learning algorithms and metrics

Three groups determining learning algorithms including classification, regression, and ensemble methods. Classification algorithms are applied for categorical outcome variables while the regression counterparts are used for real value outcome variables. The ensemble group is a kind of combined algorithms in one model. However, some algorithms, such as the k - nearest neighbors (KNN) and support vector machine (SVM) can be used for both categorical and real value outcome variables. Within the scope of this article, their applications were focused instead of going into details of mathematical algorithms. An illustration describing the algorithms and applications is given in Table 1.

Table 1. Summary of the used algorithms and their applications

Group method	Algorithm	Characteristic	Application	Reference
Linear	Linear regression model (LM)	LM is represented as a line in the form of $y_i = \beta_0 + \beta_1 x_i + e_i$ y_i and x_i are numeric and normal distribution.	Regression	(Kuhn and Johnson, 2013, Spath, 1992)
	Generalized linear model (GLM)	GLM is a flexible generalization of ordinary LM, with the probability distributions. It composes LM, ANOVA, Poisson regression, log-linear models etc.	Classification and Regression	(Dobson, 2002, McCullagh, 1989)
Non-linear	Support vector machines (SVM)	SVM plots each data item as a point in n-dimensional space (n is number of features) and searching the hyper-plane which best segregates the two classes.	Classification and Regression	(Tong and Koller, 2002)
	K-nearest neighbors	The KNN predicts a new sample using the K nearest neighbor	Classification and Regression	(Beyer et al., 1997, Guo et al.,

	(KNN)	samples from the training set		2003)
Ensemble methods	CUBIST	The tree grows and the endpoint leaf contains a linear regression model for prediction. By means of this, a series of trees are produced to establish the Cubist model.	Regression	(Quinlan, 1992),
	Random forest (RF)	RF is a combination of tree predictors using randomly the bootstrapped sample	Classification and Regression	(Breiman, 2001, Liaw and Wiener, 2001)

The coefficient of determination (R^2) and the root mean square error (RMSE) metrics reflect two sides of the algorithm's accuracy. The former is referred to how well the model fitted the data or the proportion of the variance explained by the regression model, i.e., the perfection extent increase from 0 to 100% (Ghatak, 2017) while the latter gives the idea of how wrong the model reflected the data, with the absolute perfection sets at 0. When an outcome variable is a number, RMSE is used for the model's predictive capabilities (Kuhn and Johnson, 2013). The RMSE and R^2 were determined as equations of (1) and (2), respectively (Ait-Amir et al., 2015).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - x_i)^2}{n}} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2}$$

Where y_i is the true value of the response, x_i is the predicted response by the model, $\bar{y}_i$ is average observed value, and n is the number of samples.

3. Results and discussion

3.1. Comparison of two treatment systems

3.1.1 Removal performance

Table 2 presents the efficacy of two VFs for removing different pollutants. Regardless of VFs, the removal rate consistently followed the order of COD > TSS > BOD_5 > NH_4-N. Interestingly, the two VFs showed a discernable increase of NO_3-N ($p<0.05$). It should be noted that removal of nitrogen by VF treatment could occur via concomitant routes; either nitrification or denitrification process (Zhou et al. 2018, Li et al. 2019). Hence, such an increase of NO_3-N after VF treatment could be ascribed to the effective removal of NH_4-N by nitrification and COD, which acted as the carbon sources for microbial respiration. As a result, the leftover COD is insufficient for heterotrophic bacteria to complete denitrification processes. Our results were well supported by another work (Li et al. 2019). Table 2 indicates that except for NH_4-N (P =2.7.10^{-3}), there are no significant difference of effluents between two VF tanks (P >0.05), implying that the filter materials (biochar vs. ExC) did not pose any significant discrimination to effluent quality in terms of TSS and organic matters. For NH_4-N, the VF tank using biochar performed a higher removal capacity than that filled with ExC as being of 84.6 ± 6.4% and 82.4 ± 5.7%, respectively. The high adsorption ability of biochar and microbial cultivation in the porous media might enhance the nitrogen removal in VF_2. The effects of media types in VF influenced on nitrogen removal could also be observed in many previous studies. Comparing two media types of zeolite and bauxite, Stefanakis and Tsihrintzis (2012) concluded VF-embedded zeolite achieved more effective in nitrogen and organic matter removal with rates of more than 90%. In addition, VF – packed activated alumina demonstrated better NH_4-N removal than VF – embedded shale ceramsite (Tan et al., 2020). In this study, mass removal of NH_4-N in two VF tanks were 3.5±2.5 $g \cdot m^{-2} \cdot d^{-1}$ in VF_1

and 3.6 ± 2.5 $g^{.}m^{-2.}d^{-1}$ in VF_2, significantly higher than the range of mass removal between 1.4 and 1.8 $g.m^{-2}.d^{-1}$ previously reported (Abdelhakeem et al., 2016, Paing et al., 2015). However, the results were comparable to previous work i.e., NH_4-N mass removal of 3.0 – 4.0 $g^{.}m^{-2.}d^{-1}$, which implemented pilot-scale VF units filled with fine sand and medium gravel for domestic wastewater (Bohorquez et al., 2017).

Fig. 2 demonstrates that the effluent values of BOD_5 exhibit considerable fluctuation with more outliers than that of COD, suggesting that BOD_5 is a sensitive parameter. Average BOD_5 effluents were 22.6 ± 9.9 and 18.0 ± 9.04 $mg^{.}L^{-1}$ for VF_1 and VF_2, respectively. These averaged values met the discharge limit of Vietnam's technical standards (QCVN 14:2008 & 08:2015/BTNMT) for water transportation and other low-quality water uses. The removal of BOD_5 by VF_2 was slightly higher than that of VF_1 with a removal rate of 75.4 ± 12.0 and 69.4 ± 13.0%, respectively (P=0.04). This suggests that biochar with high porous, large specific surface area and functional groups contributed the improvement of organic matter removal (Joseph et al., 2020, Odedishemi Ajibade et al., 2021, Tran et al., 2020). These results of BOD_5 removal are agreement with Kizito et al. (2017), which achieved 75% (phase I), and 86% (phase II) with VF packed biochar. Previous studies using alternative materials in VF achieved variable results for removing organic matters. VF – embedded activated alumina resulted in relatively high reduction for COD with 74.7 - 93.1% (Tan et al., 2020) while VF with pyrite, and limestone accounted for 53.3 - 56%, and 49.7 – 53.2%, respectively (Ge et al., 2019).

The removal efficiency of organic matters and nitrogen is widely adopted as the critical proxy to evaluate the capacity of CW's system. From the above interpretations, both VF_1 and VF_2 achieved a relatively high efficiency for removing organic matters and nitrogen in wastewater. The high adsorption capacity of material (i.e., biochar) packed in CW exposed the promising potential for absorbing nitrogen (i.e., ammonia) in wastewater. Other factors such as gravel, soil, sand, types of plants, microorganisms, and operational conditions in VFs could also contribute to the degradation of polluted matters. Their roles in VF should not be ruled out, and thus is also an interesting topic in future research.

Table 2. Treatment performance and statistical analysis of two VF tanks

Parameter	Influent	Effluent ($mg\cdot L^{-1}$)		Mass removal ($g\cdot m^{-2}\cdot d^{-1}$)		Hypothesis test	
		VF_1	VF_2	VF_1	VF_2	t value	*P*
TSS ($mg\cdot L^{-1}$)	128.1±19.9	35.9±20.8	42.3±26.4	17.2±7.4	15.3±42.0	1.2	0.24
BOD_5 ($mg\cdot L^{-1}$)	74.1±11.6	22.6±9.9	18.0±9.0	9.5±4.1	10.5±4.7	-2.1	0.04
COD ($mg\cdot L^{-1}$)	146.7±32.2	57.2±18.1	50.1±19.1	17.3±9.5	18.5±9.2	-1.7	0.10
NH_4-N ($mg\cdot L^{-1}$)	20.2±5.1	3.4±0.6	2.9±0.7	3.5±2.5	3.6±2.5	-3.1	$2.7.10^{-3}$
NO_3-N($mg\cdot L^{-1}$)	1.4±0.3	8.6±2.2	8.6±2.2	-	-	0.10	0.92

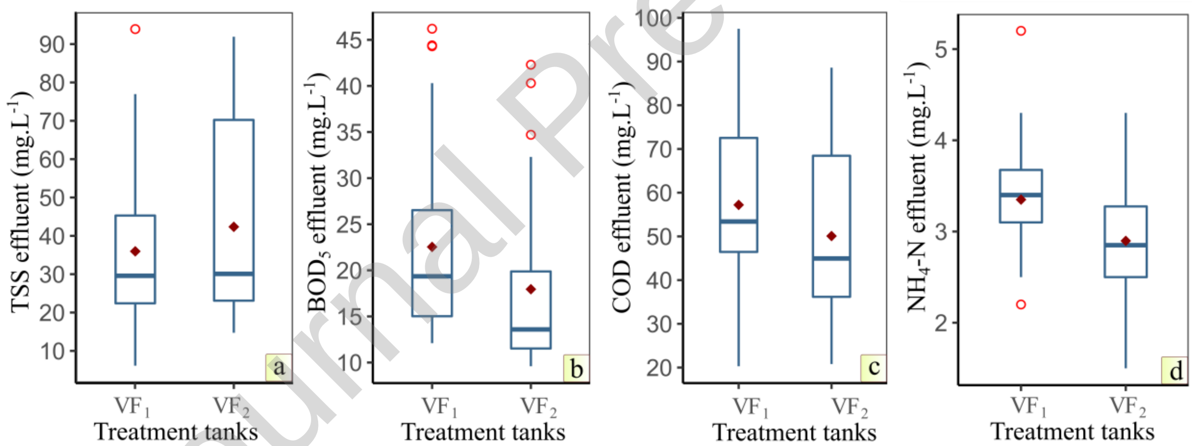

Fig. 2. Comparison of treatment tank's effluents: TSS (a), BOD_5 (b), COD (c) and NH_4-N (d). Dark red rectangle and red circle symbols are the mean and outlier values, respectively.

3.1.2. Effects of hydraulic loading rate

The effects of HLR on the efficacy of VF tanks are shown in Fig. 3. Except for NH_4-N, the higher HLR was, the less treatment efficiency i.e., COD, BOD_5 and TSS, was observed for the two VFs. This indicates the water retention time in VFs played a vital role in the removal processes of organic matters and suspended solids. More precisely, the shorter retention time of water kept in treatment tanks (i.e. high HLR) reduced the removal capacity of VF, especially in stage III and IV (Fig. 3-b&c). This finding is in good agreement with Ghosh and Gopal (2010), which also stated that an increase in hydraulic retention time from 1.0 to 2.0 days led to a rise of nearly 3-folds in the efficiency of BOD_5 and COD removal. High HLR in stage IV led to the rising of water velocity, which reduce the settlement of suspended solids, resulting in the washout of TSS with the VF effluents (Fig. 3-a). Interestingly, HLR did not pose any significant change for the NH_4-N removal rate. The stable effluents of NH_4-N over operational time might be due to the mature state of VFs, where bacterial communities in material layers and plant roots responsible for ammonia removal, was established. Some studies even highlighted the enhancement of nitrogen removal in CW with increased HRT (Ghosh and Gopal, 2010, Vymazal, 2011, Zhang et al., 2012).

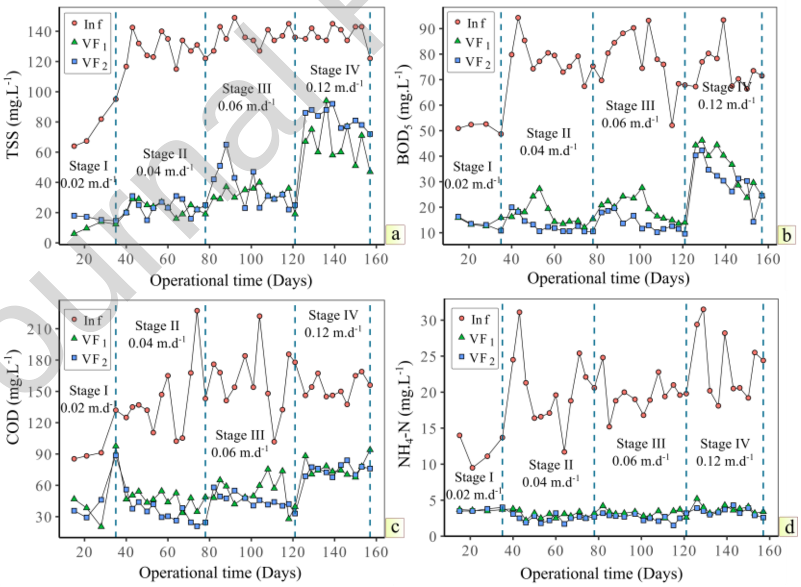

Fig. 3. Influent and effluent concentrations of wastewater in VF tanks over operational time

3.2. Predictive machine learning models

3.2.1. Input, correlation and feature selection

Although highly correlated independent variables, which cause problems (i.e. multicollinearity) is fixed by modern ML algorithms, correlation measurement is necessary because it presents the concept of linear association between variables and how attributes relate to each other. Too low correlation coefficients of explanatory can be removed in choosing variables for fitting the model (Pires et al., 2008). Similarly, highly correlated attributes of input variables also need to eliminate to make the model more accurate. This step, a so-called feature (variable) selection, is the process of choosing the variables to propose the accurately predicted variable or eliminating features, which may reduce the accuracy of algorithms. This process used *Recursive feature elimination* technique that fits an algorithm, and then removes the weakest feature or highly correlated attributions until the specified number of features is reached. This process gives the optimal features, which will be subsequently used as input variables for ML model.

Histogram plot presented the distribution of the data is given in Fig. 4. It shows that most influent features, excluding NH_4-N and NO_3-N, are not a normal distribution. The comb and skewed distribution of data may not be suitable for some statistical tests such as t-test, z-test, and ANOVA test. From distributions of data, Fig. 4 also indicates that the pre-process (i.e. *scale*, *center*…) of data may be essential to enhance the accuracy of the result.

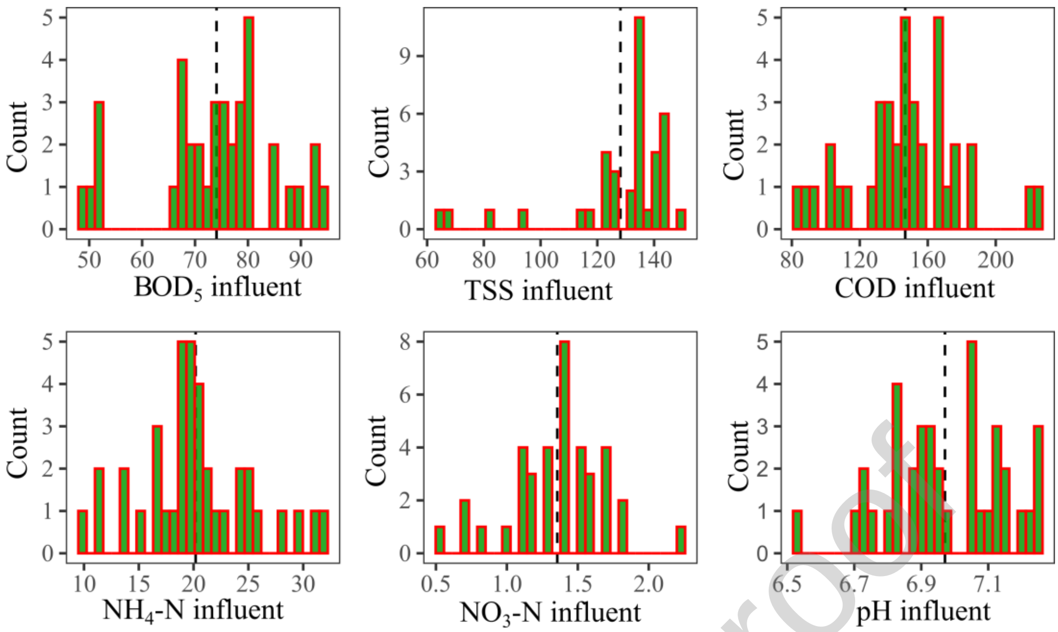

Fig. 4. Frequency distribution histogram for influents of VF systems

Fig. 5 shows that most variables represent weak to medium correlation. The highest correlations found to be 0.81 - 0.82 between BOD_5 effluent and HLR while these values for NH_4-N and HLR are only 0.42, and 0.34 at VF_1 and VF_2, respectively. The outcome variables (BOD_5 and NH_4-N effluents) correlate weakly to predictor variables, excluding HLR.

The correlation between the input-output of BOD_5 in this study is lower than that observed in the earlier study conducted by Babatunde et al. (2011), which presented an r-value of 0.79. However, there is no clear trend for NH_4-N between the two studies. As an example, while r values of NH_4-N between influent and effluent achieved 0.35 and 0.05 for VF_1 and VF_2 respectively, that in previous report was 0.18 (Babatunde et al., 2011). The low correlations between influent – effluent BOD_5 and NH_4-N mean that using simple linear regression methods (i.e., first-order kinetic model) may not reflect effectively the performance of VF tanks. Besides, a low correlation between influent and effluent indicates that the VF operated stably and efficiently, and less being shocked by the influent load of nutrients and organic matters.

Fig. 5 show that there are positive significant correlations between targeted outcome variables (BOD_5 and NH_4-N) and HLRs. This is comparable to previous findings by (Lian-sheng et al., 2006) showing a strong correlation (r> 0.93) between NH_4-N and BOD_5 removal rates and its loading rates .

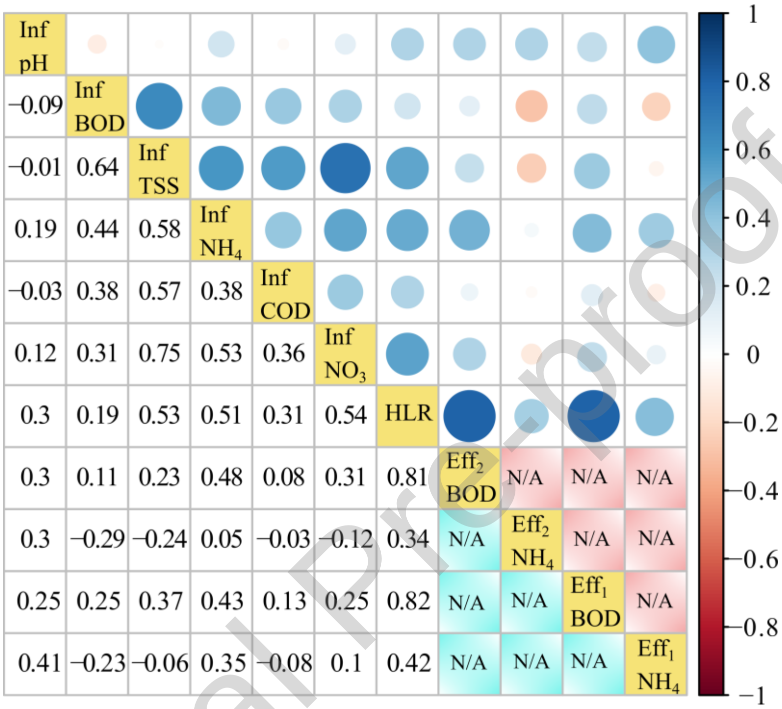

Fig. 5. Correlation matrix plot of the target outcome variables and influents. The correlations for each pair of attributes in terms of value and color level (the blue dots are positive, and red dots are negative correlation). The higher deviation from zero indicates the higher correlation magnitude, which could range from absolute positive, i.e., 1 to absolute negative, i.e, -1. The acronyms of Inf and Eff denote the influent and the effluent parameters, respectively.

Figs. 6a-d show the magnitude of the importance of input variables. According to Fig. 6, HLR displays the highest correlation towards BOD_5 of $VF_{1\&2}$, and NH_4-N of VF_2. NH_4-N influent is the most important parameter of VF_1. In sharp contrast, pH and NO_3-N influents demonstrate the least importance to the predictive models. The results of feature selection plotted in Fig. 6e-h represent the value of RMSE corresponding to the number of features and thus recommends the optimal number of features (i.e. the purple circles on the plots in Fig. 6e-h). Generally, VF_1 needs more predictors to achieve the lowest RMSE, whereas VF_2 requires one or two predictors. To get the best prediction for BOD_5 and NH_4-N in VF_1, six influent predictors, i.e., HLR, NH_4-N, COD, TSS, BOD_5, NO_3-N and five influent predictors, i.e., HLR, NH_4-N, COD, pH, BOD_5 were selected, respectively. For VF_2, two predictors including HLR, and NH_4-N influents were nominated for predicting BOD_5 effluents. Despite only HLR was selected automatically by REF technique as a predictor for predicting NH_4-N effluents in VF_2, through the fitting of ML algorithm, we decided to use two predictors, which are HLR and TSS for further assessment.

Based on the ML's results, it can be seen that the flexibility and sensitivity in the VF systems are significant. For example, the contribution of HLR to NH_4-N effluents or the magnitude of variables contributes to effluents in the two tanks is different. In addition, the relationship among parameters including influent, design and the operation conditions was elucidated by REF method. This confirms that the use of ML has supported the analysis and clarification of VF's capacity and operation.

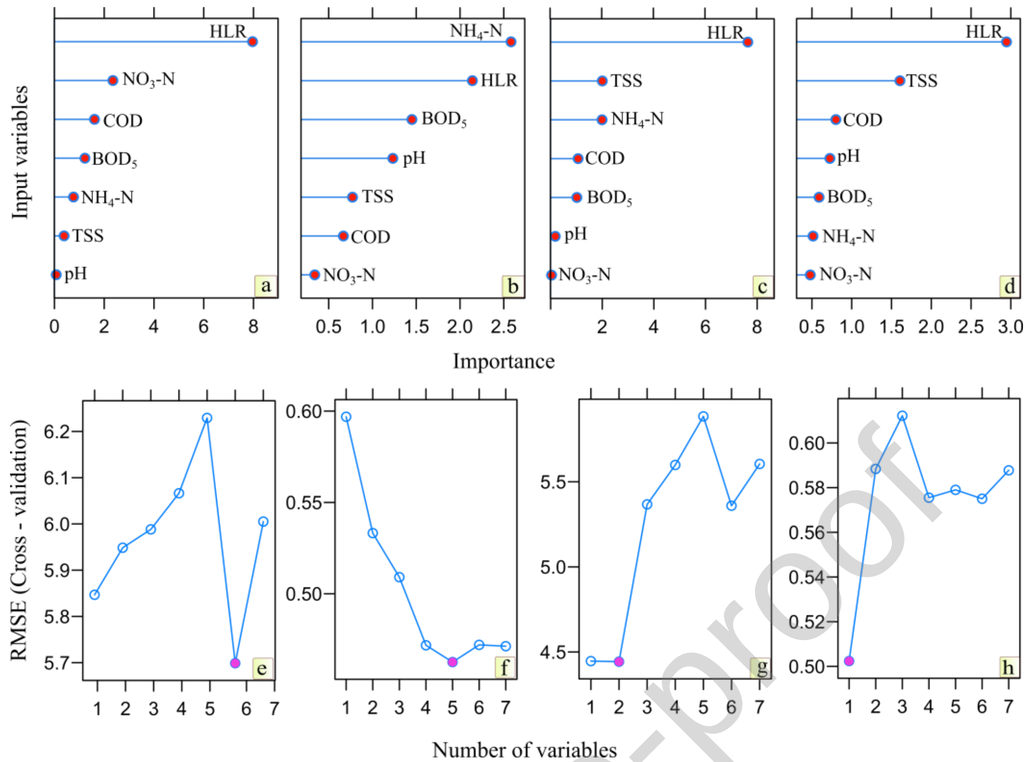

Fig. 6. Results of variable importance for BOD_5_VF_1 (a), NH_4-N_VF_1 (b), BOD_5-VF_2 (c), NH_4-N_VF_2 (d) and feature selection for BOD_5_VF_1 (e), NH_4_N_VF_1 (f); BOD_5_VF_2 (g), NH_4-N_VF_2 (h)

3.2.2. Comparison of the algorithms

Because input and output variables are numeric, RMSE was used to evaluate the accuracy and the fitting of the algorithms. The model with a smaller value of RMSE would be rated as the better one. Besides, R^2, an indicator expressing the observed variation explained by the inputs, is also introduced to clarify the results clearly. The input data were comprised of 80% training and 20% of testing. The technique of repeated cross-validation (Repeat CV) with the number of 10 of CV repeating triplicate was used for training the models. The estimated performance of a model through predictive error (i.e., RMSE) is the facile way to know how well it performed upon an unseen dataset.

For each prediction, six algorithms were run using the training data and resampling method of Repeat CV. The comparative results of six algorithms are presented in Table 3 and Fig. 7. For BOD_5, KNN performs the least performance with the highest RMSE values (8.5 and 7.1 $mg{\cdot}L^{-1}$) (Table 3). GLM and RF algorithms attain the lowest RMSE values for $BOD_5\_VF_1$ and $BOD_5\_VF_2$, respectively. The RMSE values for NH_4-N manifest the low magnitude between the algorithms. RF with RMSE of 0.48 $mg{\cdot}L^{-1}$ and SVM with RMSE of 0.46 $mg{\cdot}L^{-1}$ are the best algorithms for NH_4-N_VF_1 and NH_4-N_VF_2, respectively. The robust capacity of SVM was also confirmed by Manu and Thalla (2017), which predicted the nitrogen removal in wastewater treatment plant, concluding that SVM was better than a neuro-fuzzy inference system. For predicting solid waste generation, the previous works stated that SVM was considered a good predictive model (Abbasi et al., 2013, Abbasi and El Hanandeh, 2016). Moreover, Kumar et al. (2018) compared RF and SVM, and found that both models had a quite similar metric in terms of R^2 and RMSE. Table 3 also indicates that LM and GLM do not have a significant difference in RMSE, and KNN may be the least efficient model. Low effective algorithm of LM implies that linear relationship or linear kinetics between influent and effluent failed in describing mechanisms, efficiency, and designing the system.

Table 3. The average RMSE of the algorithms.

ML algorithms	**$BOD_5\_VF_1$**	**$BOD_5\_VF_2$**	**NH_4-N_VF_1**	**NH_4-N_VF_2**
RF	5.9	4.7	0.48	0.50
SVM	5.7	5.1	0.49	0.46
KNN	8.5	7.1	0.52	0.58
GLM	4.9	5.8	0.48	0.52
LM	5.1	5.8	0.50	0.50
CUBIST	6.1	5.1	0.49	0.51

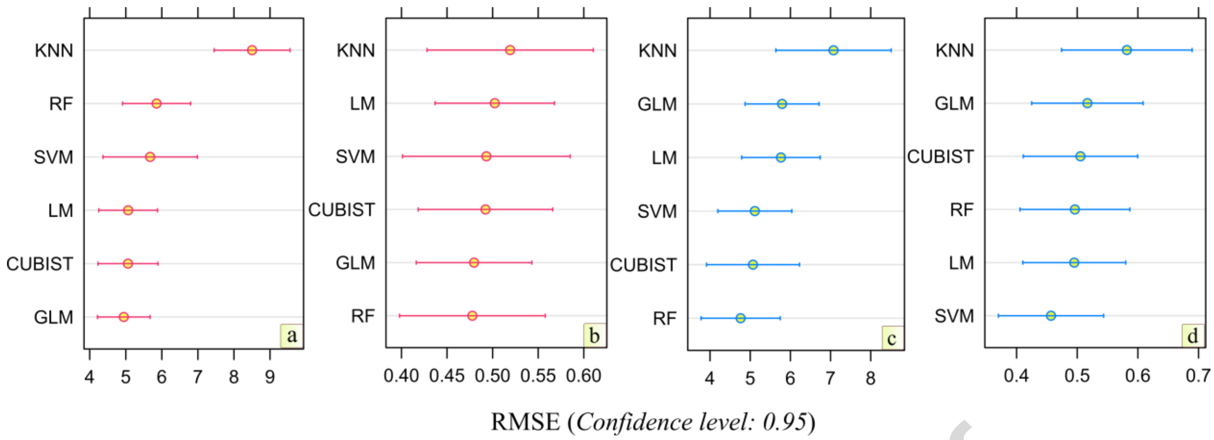

Fig. 7. Comparative results of six ML algorithms regarding RMSE values: BOD_5_VF_1 (a), NH_4-N_VF_1 (b), BOD_5_VF_2 (c) and NH_4-N_VF_2 (d).

3.2.3. Improvement of algorithms

The accuracy of the chosen ML algorithms can be improved by several techniques, including transformation, resampling, and tuning. This work used four popular methods of data transformation, also known as "center", "scale", "box-cox", and "range" and four resampling techniques, including Repeat CV, K- fold CV, Leave-one-out CV, and Bootstrap. In addition, the next step of tuning will be performed based on the hyperparameters of a certain algorithm.

Table 4 shows the output of the improvement step for choosing algorithms. As can be seen, data transformed do not make the algorithms more accurate (no change in RMSE). Only the performance of the RF algorithm for BOD_5_VF_2 is enhanced by the resampling of K-fold CV and tuning with mtry of 2.0 and ntree of 50.0. From Repeat CV to K-fold CV, RMSE of RF (BOD_5_VF_1) decreases from 4.8 to 4.5 $mg{\cdot}L^{-1}$, and further reduces to 4.0 $mg{\cdot}L^{-1}$ by tuning. Through the improvement step for ML algorithm, it can be concluded that Repeat CV is the most effective resampling method and tuning is applicable for RF. In addition, GLM cannot be tuned because it does not have a tuning parameter.

Table 4. Result of the improvement of algorithms

Algorithm	Transformation	Resampling	Tuning	Final parameters of the model
GLM (BOD_5_VF_1)	No change	No change	No tuning	Repeat CV; RMSE = 5.0 mg·L^{-1}, R^2 = 71.2%
RF (NH_4-N_VF_1)	No change	No change	Good with tuning	Repeat CV mtry = 1.0 and ntree = 50.0 RMSE 0.42 mg·L^{-1}; R^2= 50.8%
RF (BOD_5_VF_2)	No change	Good with K-fold CV	Good with tuning	K-fold CV; mtry = 2.0 and ntree = 50.0 RMSE = 4.0 mg·L^{-1}, R^2 = 76.9%
SVM (NH_4-N_VF_2)	No change	No change	No change	Repeat CV; sigma = 5.6 and C = 1; RMSE = 0.50 mg·L^{-1}; R^2= 70.4%

Fig. 8 presents the change of RMSE when the hyperparameters, including mtry and ntree of the tuned RF. With ntree of 50.0 along mtry of 1.0 and 2.0, the lowest of RMSE was achieved for BOD_5_VF_2 and NH_4-N_VF_1, respectively. RMSE, and R^2 of RF algorithm for NH_4-N_VF_1 (mtry = 1.0, ntree = 50.0, and Repeat CV) account for 0.42 mg·L^{-1}, and 50.8%, while those for BOD_5_VF_2 (mtry = 3.0, ntree = 50.0, and K-fold CV) are 4.0 mg·L^{-1}, and 76.9%, respectively. These hyperparameters of RF were used to build the final model for predicting the performance of VF tanks.

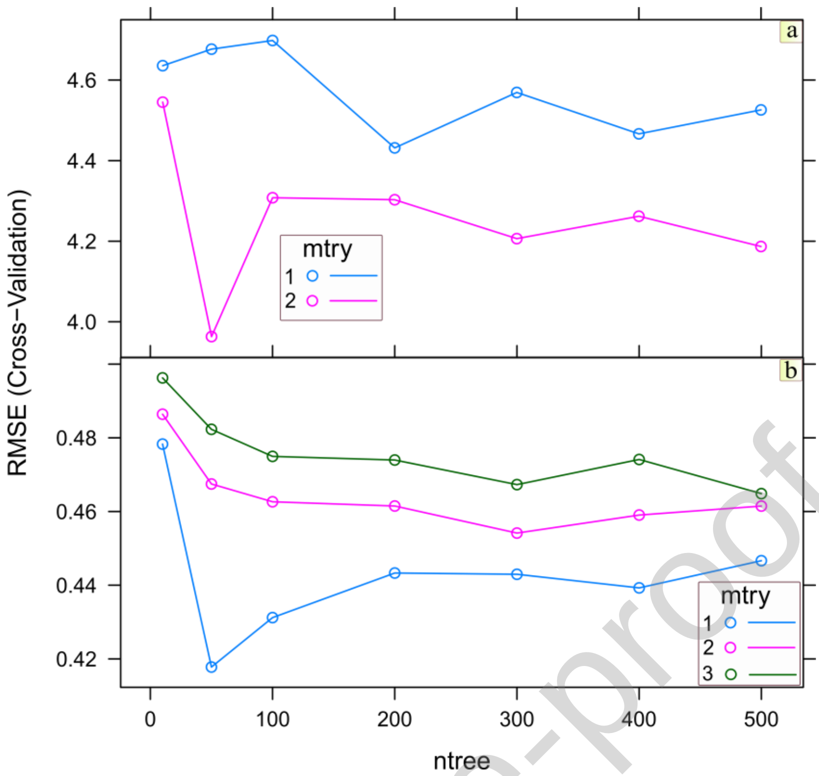

Fig. 8. The tuning results of RF algorithm: a) BOD_5_VF_2 and b) NH_4-N_ VF_1

3.2.4. Prediction of effluents

To evaluate the accuracy and feasibility of the prediction of VFs' effluent, the final selected models were fitted with the testing sub-data and whole data. Table 5 indicates the RMSE values increase from training data to testing data for all algorithms, except for RF of BOD_5_VF_2. The lower RMSE for testing data, which accounts for 20% of total data and is considered as the blind input, indicates that RF is a robust model even with unseen data. Moreover, the high correlation between the actual and predicted values in Fig. 8-c and R^2 with 90.1% (Table 5) reinforce that RF performs effectively in predicting the effluents of VF treatment tanks. The results are comparable with the previous studies used RF. For example, Zhou et al. (2019) presented high R^2 of 54.5 – 97.1% for training and testing data, and Ahmed et al. (2019) showed R^2 of 67.1% with four input variables.

The prediction error of the GLM algorithm significantly increases from 4.9 mg·L^{-1} for training data to 7.9 mg·L^{-1} for the testing data, while R^2 of SVM algorithm achieves the lowest value of 48.5%. The low R^2 value may be ascribed to the small number of input variables and data noise (Guo et al., 2015). In addition, the high variability of effluents in the certain treatment tanks could be the reason of low R^2 of predictive model. The scale of standardized residuals presented in Fig. 9a-c illustrates that the predicted and actual data of NH_4-N_$VF_{1\&2}$ and BOD_5_VF_2 are normally distributed (95.0% of standardized residuals fall into -2.0 to +2.0), meaning that predictive model is robust. However, many points (60.0%) outside the ±2.0 limits were found for BOD_5_VF_1. In terms of R^2, SVM's performance previously reported was found varying. For example, Manu and Thalla (2017) stated an R^2 value of 82.5%, while R^2 of 34.6% was noted elsewhere Ahmed et al. (2019). In addition, the accuracy of the algorithms were examined by comparing the RMSE with the range of effluent values (i.e. output variable). For instance, the RMSE values of BOD_5 for VF_1 and VF_2 were 5.0, and 2.9 mg·L^{-1}, respectively, as compared to range of BOD_5 effluents of 12.1 – 46.2 mg·L^{-1} (VF_1), and 9.6 - 42.3 mg·L^{-1} (VF_2), indicating that ML algorithms used have relatively high accuracy. This robust prediction exposes a feasible way to use ML algorithm to support the design of VF systems. Moreover, from the raw inputs, predictive ML model can draw the results of output, offering the manager to make the decisions regarding the CW construction investment as well as adjusting to fulfill the discharge limits.

Table 5. The results of algorithms' metric for the testing and whole dataset.

Algorithm	**Training data (RMSE)**	**Testing data (RMSE)**	**Whole data**		**The range of effluents (mg·L^{-1})**
			RMSE (mg·L^{-1})	**R^2 (%)**	
GLM (BOD_5_VF_1)	4.9	7.9	5.0	74.0	12.1 – 46.2
RF (NH_4-N_VF_1)	0.4	0.5	0.3	80.0	2.2. – 5.2
RF (BOD_5_VF_2)	4.0	3.6	2.9	90.1	9.6 - 42.3
SVM (NH_4-N_VF_2)	0.5	0.7	0.5	48.5	1.5 - 4.3

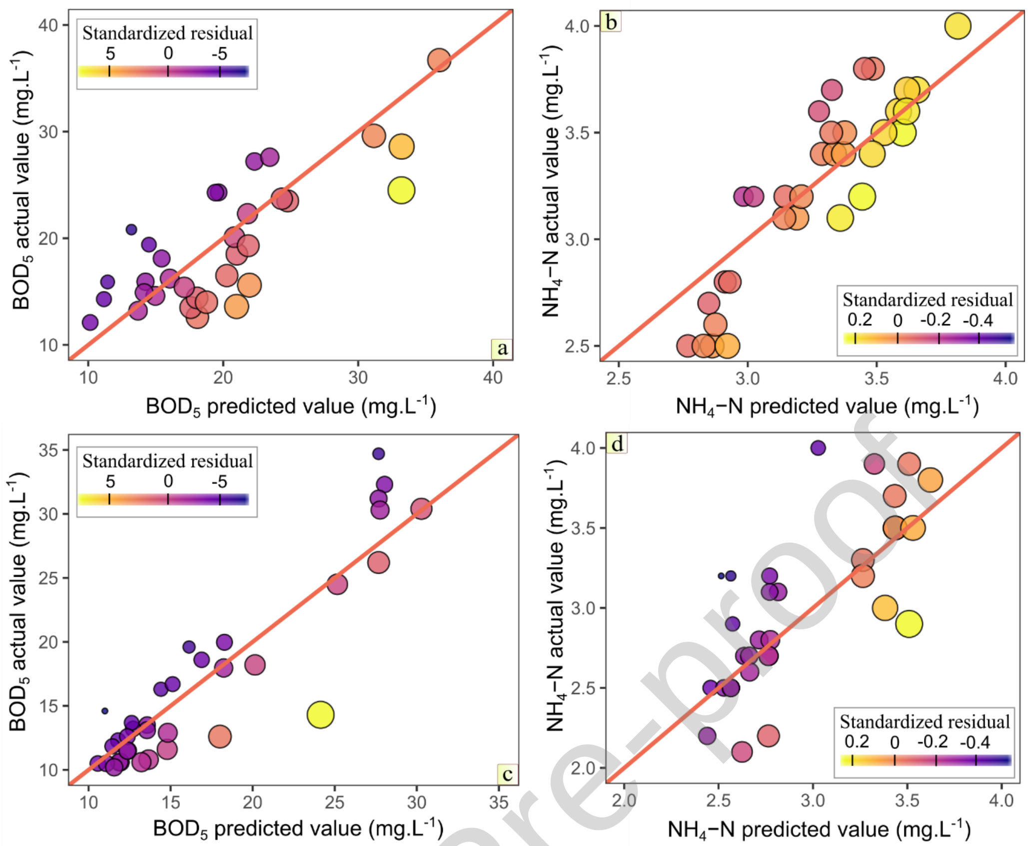

Fig. 9. The results of fitting and residual values of BOD_5 at VF_1 (a) and VF_2 (c); NH_4-N at VF_1 (b), and VF_2 (d)

4. Conclusions

Two VF tanks packed with biochar and ExC tested for wastewater treatment for 21 weeks. The comparison results indicate that the different materials did not affect significantly the effluent concentrations of TSS, organic matters and NO_3-N, except for NH_4-N. More precisely, the removal rate of NH_4-N by VF_2 was much higher than that by VF_1 being of 84.6 ± 6.4 and 82.4 ± 5.7%, respectively. The high adsorption capacity of material (i.e. biochar) packed in CW could be possibly ascribed to nitrogen elimination in wastewater. The best algorithms selected for predicting effluents of VF tanks were RF, GLM, and SVM. The values of R^2 of whole fitting data achieve 74.0, 80.0, 90.1, and 48.5% for BOD_5_VF_1, NH_4-N_VF_1, BOD_5_VF_2, and NH_4-N_VF_2, respectively. The study demonstrated that ML could be a promising tool to predict the efficiency of the CW systems for wastewater treatment. Further works should be carried out to expand this application for a full-scale system.

Acknowledgment

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.99-2019.25.

References

Abbasi, M., M.A. Abduli, B. Omidvar, A. Baghvand, (2013). Forecasting Municipal Solid waste Generation by Hybrid Support Vector Machine and Partial Least Square Model. Int. J. Environ. Res. 7, 27-38.

Abbasi, M., A. El Hanandeh, (2016). Forecasting municipal solid waste generation using artificial intelligence modelling approaches. Waste Management 56, 13-22.

Abdelhakeem, S.G., S.A. Aboulroos, M.M. Kamel, (2016). Performance of a vertical subsurface flow constructed wetland under different operational conditions. Journal of Advanced Research 7, 803-814.

Abou-Elela, S.I., M.S. Hellal, (2012). Municipal wastewater treatment using vertical flow constructed wetlands planted with Canna, Phragmites and Cyprus. Ecological Engineering 47, 209-213.

Ahmed, U., R. Mumtaz, H. Anwar, A. Shah, R. Irfan, J. García-Nieto, (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. Water 11, 2210.

Ait-Amir, B., P. Pougnet, A. El Hami, (2015). 6 - Meta-Model Development, in: A. El Hami, P. Pougnet (Eds.) Embedded Mechatronic Systems 2, Elsevier, pp. 151-179.

Akratos, C.S., J.N.E. Papaspyros, V.A. Tsihrintzis, (2008). An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands. Chem. Eng. J. 143, 96–110.

APHA/WEF/AWWA, Standard Methods for the Examination of Water & Wastewater, Centennial Edition. 22 ed, in, American Public Health Association, the American Water Works Association, and the Water Environment Federation, Washington DC, USA, 2012.

Babatunde, A.O., Y.Q. Zhao, R.J. Doyle, S.M. Rackard, J.L. Kumar, Y.S. Hu, (2011). On the fit of statistical and the k-C* models to projecting treatment performance in a constructed wetland system. Journal of environmental science and health. Part A, Toxic/hazardous substances & environmental engineering 46, 490-499.

Beyer, K., J. Goldstein, R. Ramakrishnan, U. Shaft, (1997). When Is "Nearest Neighbor" Meaningful? ICDT 1999. LNCS 1540.

Bohorquez, E., D. Paredes, C.A. Arias, (2017). Vertical flow-constructed wetlands for domestic wastewater treatment under tropical conditions: effect of different design and operational parameters. Environmental technology 38, 199-208.

Breiman, L., (2001). Random Forests. Machine Learning 45, 5-32.

Calheiros, C.S., A.F. Duque, A. Moura, I.S. Henriques, A. Correia, A.O. Rangel, Castro, P.M., (2009a). Changes in the bacterial community structure in two-stage constructed wetlands with different plants for industrial wastewater treatment. Bioresoure Technology 100, 3228-3235.

Calheiros, C.S.C., A.F. Duque, A. Moura, I.S. Henriques, A. Correia, A.O.S.S. Rangel, P.M.L. Castro, (2009b). Substrate effect on bacterial communities from constructed wetlands planted with Typha latifolia treating industrial wastewater. Ecological Engineering 35, 744-753.

Chen, K., H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, H. Ren, (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research 171, 115454.

Cooper, P., (2005). The performance of vertical flow constructed wetland systems with special reference to the significance of oxygen transfer and hydraulic loading rates. Water science and technology : a journal of the International Association on Water Pollution Research 51, 81-90.

Cooper, P.F., G.D. Job, M.B. Green, R.B.E. Shutes, (1996). Reed Beds and Constructed Wetlands for Wastewater Treatment. WRc Publications, Medmenham, UK.

Decezaro, S.T., D.B. Wolff, C. Pelissari, R.J.M.G. Ramírez, T.A. Formentini, J. Goerck, L.F. Rodrigues, P.H. Sezerino, (2019). Influence of hydraulic loading rate and recirculation on oxygen transfer in a vertical flow constructed wetland. Science of The Total Environment 668, 988-995.

Dobson, A.J., (2002). An introduction to generalized linear models / Annette J. Dobson. Chapman & Hall/CRC, Boca Raton.

Dordio, A., A.J. Carvalho, (2013). Constructed wetlands with light expanded clay aggregates for agricultural wastewater treatment. The Science of the total environment 463-464, 454-461.

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

Ghatak, A., (2017). Machine Learning with R. Springer Nature, Singapore.

Ghosh, D., B. Gopal, (2010). Effect of hydraulic retention time on the treatment of secondary effluent in a subsurface flow constructed wetland. Ecological Engineering 36, 1044-1051.

Guo, G., H. Wang, D. Bell, Y. Bi, K. Greer, (Year). KNN Model-Based Approach in Classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Springer Berlin Heidelberg, 986-996.

Guo, H., K. Jeong, J. Lim, J. Jo, Y.M. Kim, J.P. Park, J.H. Kim, K.H. Cho, (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of environmental sciences (China) 32, 90-101.

Hijosa-Valsero, M., R. Sidrach-Cardona, J. Martín-Villacorta, M.C. Valsero-Blanco, J.M. Bayona, E. Bécares, (2011). Statistical modelling of organic matter and emerging pollutants removal in constructed wetlands. Bioresour. Technol. 102, 4981–4988.

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

Jia, W., X. Sun, Y. Gao, Y. Yang, L. Yang, (2020). Fe-modified biochar enhances microbial nitrogen removal capability of constructed wetland. Science of The Total Environment 740, 139534.

Joseph, S.M.R., P. Wijekoon, B. Dilsharan, N.D. Punchihewa, B.C.L. Athapattu, M. Vithanage, (2020). Anammox, biochar column and subsurface constructed wetland as an integrated system for treating municipal solid waste derived landfill leachate from an open dumpsite. Environmental Research 189, 109880.

Kadlec, R.H., (2000). The inadequacy of first-order treatment wetland models. Ecological Engineering 15, 105-119.

Kasak, K., J. Truu, I. Ostonen, J. Sarjas, K. Oopkaup, P. Paiste, M. Kõiv-Vainik, Ü. Mander, M. Truu, (2018). Biochar enhances plant growth and nutrient removal in horizontal subsurface flow constructed wetlands. Sci. Total Environ. 639, 67-74.

Kim, B., M. Gautier, S. Prost-Boucle, P. Molle, P. Michel, R. Gourdon, (2014). Performance evaluation of partially saturated vertical-flow constructed wetland with trickling filter and chemical precipitation for domestic and winery wastewaters treatment. Ecological Engineering 71, 41-47.

Kizito, S., T. Lyu, S. Wu, Z. Ajmal, L. Hongzhen, R. Dong, (2017). Treatment of anaerobic digested effluent in biochar-packed vertical flow constructed wetland columns: Role of media and tidal operation. Sci. Total Environ. 592, 197-205.

Kuhn, M., K. Johnson, (2013). Applied Predictive Modeling. Springer, New York, NY.

Kumar, A., S.R. Samadder, N. Kumar, C. Singh, (2018). Estimation of the generation rate of different types of plastic wastes and possible revenue recovery from informal recycling. Waste Management 79, 781-790.

Lian-sheng, H., L. Hong-liang, X. Bei-dou, Z. Ying-bo, (2006). Effects of effluent recirculation in vertical-flow constructed wetland on treatment efficiency of livestock wastewater. Water Science & Technology 54, 137-146.

Liaw, A., M. Wiener, (2001). Classification and Regression by RandomForest. Forest 23.

Manu, D.S., A.K. Thalla, (2017). Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. Applied Water Science 7, 3783-3791.

McCullagh, P.a.N., J.A., (1989). Generalized Linear Models. Chapman and Hall, London.

Mlih, R., F. Bydalek, E. Klumpp, N. Yaghi, R. Bol, J. Wenk, (2020). Light-expanded clay aggregate (LECA) as a substrate in constructed wetlands – A review. Ecological Engineering 148, 105783.

Murray-Gulde, C.L., W.C. Bridges, J.H. Rodgers, (2008). Evaluating performance of a constructed wetland treatment system designed to decrease bioavailable copper in a waste stream. Environ. Geosci. 15, 21-38.

Nguyen, X.C., S.W. Chang, T.L. Nguyen, H.H. Ngo, G. Kumar, J.R. Banu, M.C. Vu, H.S. Le, D.D. Nguyen, (2018). A hybrid constructed wetland for organic-material and nutrient removal from sewage: Process performance and multi-kinetic models. Journal of Environmental Management 222, 378-384.

Nguyen, X.C., D.D. Nguyen, Q.B. Tran, T.T.H. Nguyen, T.K.A. Tran, T.C.P. Tran, T.H.G. Nguyen, T.N.T. Tran, D.D. La, S.W. Chang, R. Balasubramani, W.J. Chung, Y.S. Yoon, V.K. Nguyen, (2020a). Two-step system consisting of novel vertical flow and free water surface constructed wetland for effective sewage treatment and reuse. Bioresource Technology 306, 123095.

Nguyen, X.C., T.C.P. Tran, V.H. Hoang, T.P. Nguyen, S.W. Chang, D.D. Nguyen, W. Guo, A. Kumar, D.D. La, Q.-V. Bach, (2020b). Combined biochar vertical flow and free-water surface constructed wetland system for dormitory sewage treatment and reuse. Science of The Total Environment 713, 136404.

Odedishemi Ajibade, F., H.-C. Wang, A. Guadie, T. Fausat Ajibade, Y.-K. Fang, H. Muhammad Adeel Sharif, W.-Z. Liu, A.-J. Wang, (2021). Total nitrogen removal in biochar amended non-aerated vertical flow constructed wetlands for secondary wastewater effluent with low C/N ratio: Microbial community structure and dissolved organic carbon release conditions. Bioresource Technology 322, 124430.

Paing, J., A. Guilbert, V. Gagnon, F. Chazarenc, (2015). Effect of climate, wastewater composition, loading rates, system age and design on performances of French vertical flow constructed wetlands: A survey based on 169 full scale systems. Ecological Engineering 80, 46-52.

Patil, S., S. Chakraborty, (2017). Effects of step-feeding and intermittent aeration on organics and nitrogen removal in a horizontal subsurface flow constructed wetland. Journal of Environmental Science and Health, Part A 52, 403-412.

Pires, J.C.M., F.G. Martins, S.I.V. Sousa, M.C.M. Alvim-Ferraz, M.C. Pereira, (2008). Selection and validation of parameters in multiple linear and principal component regressions. Environmental Modelling & Software 23, 50-55.

Prost-Boucle, S., P. Molle, (2012). Recirculation on a single stage of vertical flow constructed wetland: Treatment limits and operation modes. Ecological Engineering 43, 81-84.

Quinlan, R., (Year). Learning with Continuous Classes. *Proceedings of the 5th Australian Joint Conference On Artificial Intelligence*, Australian, 343–348.

Rousseau, D.P.L., P.A. Vanrolleghem, N. DePauw, (2004). Model based design of horizontal subsurface flow constructed treatment wetlands: a review. Water Research 38, 1484–1493.

Saeed, T., N. Majed, T. Khan, H. Mallika, (2019). Two-stage constructed wetland systems for polluted surface water treatment. Journal of Environmental Management 249, 109379.

Spath, H., (1992). Mathematical algorithms for linear regression. Academic Press Professional, Inc.

Stefanakis, A., V. Tsihrintzis, (2012). Use of zeolite and bauxite as filter media treating the effluent of Vertical Flow Constructed Wetlands. Microporous and Mesoporous Materials 155, 106-116.

Tan, X., Y.-L. Yang, X. Li, Z.-W. Zhou, C.-J. Liu, Y.-W. Liu, W.-C. Yin, X.-Y. Fan, (2020). Intensified nitrogen removal by heterotrophic nitrification aerobic denitrification bacteria in two pilot-scale tidal flow constructed wetlands: Influence of influent C/N ratios and tidal strategies. Bioresource Technology 302, 122803.

Tong, S., D. Koller, (2002). Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2, 45–66.

Torrijos, V., O.G. Gonzalo, A. Trueba-Santiso, I. Ruiz, M. Soto, (2016). Effect of by-pass and effluent recirculation on nitrogen removal in hybrid constructed wetlands for domestic and industrial wastewater treatment. Water Research 103, 92-100.

Tran, T.V., L.X. Nong, H.-T.T. Nguyen, V.H. Nguyen, D.T.C. Nguyen, T.T. Nguyen, P.Q. Trang, D.H. Nguyen, T.D. Nguyen, (2020). Response surface methodology modeling for methylene blue removal by chemically modified porous carbon: Adsorption mechanism and role of surface functional groups. Separation Science and Technology 1-11.

Vymazal, J., (2011). Constructed Wetlands for Wastewater Treatment: Five Decades of Experience. Environmental Science & Technology 45, 61-69.

Vymazal, Y., (2007). Removal of nutrients in various types of constructed wetlands. Science of the Total Environment 380, 48–65.

Wang, Y., L. Shen, J. Wu, F. Zhong, S. Cheng, (2020). Step-feeding ratios affect nitrogen removal and related microbial communities in multi-stage vertical flow constructed wetlands. Science of The Total Environment 721, 137689.

Wu, S.-q., J. Zhang, H.H. Ngo, W. Guo, Z. Hu, S. Liang, J. Fan, H. Liu, (2015). A review on the sustainability of constructed wetlands for wastewater treatment: Design and operation. Bioresource Technology 175, 594-601.

Wu, S., P. Kuschk, H. Brix, J. Vymazal, R. Dong, (2014). Development of constructed wetlands in performance intensifications for wastewater treatment: A nitrogen and organic matter targeted review. Water Res 57, 40-55.

Yajima, H., J. Derot, (2017). Application of the Random Forest model for chlorophyll- a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. Journal of Hydroinformatics 20, jh2017010.

Zhang, C.-B., W.-L. Liu, J. Wang, Y. Ge, B.-H. Gu, J. Chang, (2012). Effects of plant diversity and hydraulic retention time on pollutant removals in vertical flow constructed wetland mesocosms. Ecol. Eng. 49, 244-248.

Zhang, S., Q.V. Ly, L.D. Nghiem, J. Wang, J. Li, Y. Hu, (2020). Optimization and organic fouling behavior of zwitterion-modified thin-film composite polyamide membrane for water reclamation: A comprehensive study. J Memb Sci 596, 117748.

Zhou, P., Z. Li, S. Snowling, B. Baetz, D. Na, G. Boyd, (2019). A random forest model for inflow prediction at wastewater treatment plants. Stochastic Environmental Research and Risk Assessment 33.

Zhou, X., Z. Chen, Z. Li, H. Wu, (2020). Impacts of aeration and biochar addition on extracellular polymeric substances and microbial communities in constructed wetlands for low C/N wastewater treatment: Implications for clogging. Chemical Engineering Journal 396, 125349.

Highlights

- Vertical flow packed biochar exposed the promising potential for absorbing nitrogen
- Different filters do not significantly influence effluents, except for NH_4-N
- Random forest, Generalized linear model, and Support vector machines selected.
- Low difference between RMSE and the range of experimental effluents (BOD_5 and NH_4-N)