Machine learning modelling and analysis of biohydrogen production from wastewater by dark fermentation process

Ahmad Hosseinzadeh[a], John L. Zhou[a*], Ali Altaee[a], Donghao Li [b]

[a]Centre for Green Technology, School of Civil and Environmental Engineering, University of Technology Sydney, NSW 2007, Australia

[b]Department of Chemistry, Yanbian University, Park Road 977, Yanji 133002, Jilin Province, China

Abstract

The fermentation process for wastewater treatment and H_2 production simultaneously is gaining attention. In this study, machine learning (ML)-assisted procedures were used to analyze and model H_2 production from wastewater by this process. Different ML-assisted procedures were assessed based on mean square error (MSE) and R2 to select the most robust models for modelling the fermentation process. The research showed that gradient boosting machine (GBM), support vector machine (SVM), random forest (RF) and AdaBoost were the most appropriate, which were optimized by grid search and deeply analyzed by permutation variable importance (PVI) to identify the relative importance of the variables. All four models demonstrated promising performances in predicting H_2 productions with determination coefficient values of 0.893, 0.885, 0.902 and 0.889. The MSE of these models were 0.015, 0.015, 0.016 and 0.015, respectively. Moreover, RF-PVI demonstrated better performance in variables' relative importance showing that acetate (A), butyrate (B), A/B, ethanol, Fe and Ni have a higher importance in decreasing order.

Keywords: Dark fermentation; Bio-hydrogen; Machine learning; Process modelling,

1. Introduction

The explosion of the world population and urbanization and industrialization have caused serious challenges in energy deficiency, freshwater shortage, and environmental pollution (Hosseinzadeh et al., 2021). Fossil fuels have long been the dominating source of energy generation, which has led to growing emission of various pollutants (e.g. NO_x, CO, PM) and greenhouse gases (e.g. CO_2) into the atmosphere resulting in deteriorating air quality and global warming (Hosseinzadeh et al., 2020b; Huang et al., 2019). Based on the reports (Alassi et al., 2019; Mai-Moulin et al., 2021), renewable energy currently provides less than 25% of the total global energy requirement, which will be increased to more than half in 2040. Currently, bioenergy represents the highest portion of renewable energy (Gómez-Marín and Bridgwater, 2021). In addition, wastewater and solid wastes are regarded as one of the main sources of health and environmental challenges (Alidadi et al., 2016; Zorpas, 2020). In order to tackle the current challenges, different technologies can be adopted, e.g. generating energy from renewable sources (Hosseinzadeh et al., 2021), sorting and recycling of solid wastes (Alidadi et al., 2016), and advanced oxidation processes for wastewater treatment (Bao et al., 2020a; Bao et al., 2020b; Kamranifar et al., 2021). Furthermore, developing technologies that can simultaneously address the mentioned problems is exciting and rewarding, supporting individual nations to meet the UN Sustainable Development Goals (Hosseinzadeh et al., 2021).

Dark fermentation is when the microorganisms syntrophically treat wastewater and produce biohydrogen simultaneously (Sekoai et al., 2021). Therefore, the dark fermentation process can address all three challenges mentioned, i.e. energy deficiency, freshwater shortage and environmental pollution. As a sustainable process, it has received extensive attention owing to several merits, e.g. considerable capability in the consumption of various substrates, no need for light, cheap and simple reactor configurations, and the ability to produce biohydrogen under ambient temperature and pressure (Baeyens et al., 2020; Pradhan et al., 2016; Sekoai et al., 2021). However, the performance of the process is affected by different operating conditions, e.g. pH, temperature, substrate, process type, hydraulic retention time (HRT) and the metabolites produced during the process (Wong et al., 2014). Solution pH can influence the performance of the process through different ways, e.g. in the selection of a suitable microbial community (Toquero and Bolado, 2014; Zhao et al., 2015), maintaining surface charge on the microbial membrane simplifying the nutrient absorption by the microorganism and providing an appropriate environment for the enzymes' activity catalyzing H_2 production (Liu et al., 2012; Wong et al., 2014). Temperature can affect the physiological activities of the microorganisms in H_2 production, and the higher the temperature, the lower the solubility of H_2, and consequently, the lower the consumption of the produced H_2 by H_2 consumer microorganisms in the process (Wong et al., 2014). The type of substrate plays a crucial role in the H_2 production by this process. Each mole of glucose and lactose can theoretically produce 12 moles and 23 moles H_2, respectively; however, the process is less efficient in practice (Wong et al., 2014). Most of the thermal enthalpies are consumed to produce volatile fatty acids (VFAs), the most important metabolites in this process. Correspondingly, the common maximum H2 production efficiency is 4 moles and 2 moles H2 per mole of glucose using the acetate and butyrate pathways. The acetate to butyrate ratio determines the type of the dominant H_2 production pathway. If the ratio is more than one, the pathway will be via acetate; otherwise, the pathway will be via butyrate. Moreover, providing all co-factors by the substrate required for H_2 producing bacteria is another aspect of substrate effectiveness (Wong et al., 2014). For example, the hydrogenase enzymes catalyzing H_2 are categorized into [Fe-Fe] and [Ni-Fe], based on the metals at their active sites. Therefore, Fe^{2+} and Ni^{2+} are two of the key ingredients of the enzymes used for H_2 production, which should be provided by the substrates (Karadag and Puhakka, 2010). In addition, the loss of the adapted inoculum and avoiding the trace elements deficiency over the process are the other effective factors affected by the process mode, the hydraulic retention time (HRT) and the inoculum proportion (Cao et al., 2019; Li et al., 2020).

Therefore, optimizing the dark fermentation process is key to its success, which the experimental and numerical procedures can accomplish. The numerical modelling of the process is highly complementary and usually faster and more economical than the experimental approach. In comparison, there has been a wide range of experimental studies conducted to optimize the fermentation process. Yet, there is a lack of studies regarding the application of the modelling procedures in the fermentation process. In addition, to the best of our knowledge, there is no study yet to consider all of those parameters together to study the fermentation process, which is very important to pre-design the process before the experimental study. More importantly, the relative importance of the effective factors should be determined to support the experimental design and optimization of operating conditions, which will reduce the number of experiments for achieving the intended outcome.

Machine learning (ML)-assisted approaches are vigorous techniques to learn and model the complicated correlations among the dependent and independent variables in various processes or phenomena. These approaches do not need to understand all complicated background mechanisms of the processes to master the potential correlations. Various types of such approaches can model different types of processes; however, the performances of these approaches can be different in various applications. So far, there is a major knowledge gap regarding the application of these approaches in H_2 production from wastewater by the fermentation process. More importantly, there is no study to systematically investigate the application of various ML-assisted approaches in the fermentation process to select the most vigorous ones for modelling and analysis purposes.

Therefore, this study aims first to apply different ML-assisted procedures, i.e. gradient boosting machine (GBM), support vector machine (SVM), random forest (RF), AdaBoost, multilayer perceptron (MLP), linear regression (LR) and Ridge in H_2 production from wastewater through the fermentation process. Key parameters including Fe, Ni, biomass proportion, acetate (A), butyrate (B), A/B, ethanol, pH, HRT and COD are considered inputs to select the more robust procedures, which are then used to carefully model and analyze the process. Finally, the performances of the chosen models will be compared using the outcomes, and the relative importance of the effective factors will be studied by permutation variable importance procedure.

2. Materials and Methods

2.1. Data collection and processing

To model and analyze H_2 production from wastewater by fermentation process, a detailed literature review was accomplished by considering a wide spectrum of factors, e.g. reporting the acetic and butyric acids proportions over the process, the presence of Fe and Ni as cofactors and enzymatic metals, comparable units presenting H_2 production, the application of same inoculum in the process and the other operating condition in wastewater treatment by dark fermentation process. A schematic setup for the production of H_2 from wastewater by dark fermentation process is presented in Fig. 1. Based on the literature search, 211 data points were selected and extracted from the published papers (Dessì et al., 2020; Karadag and Puhakka, 2010). The extraction of the experimental data was carried out by Plot Digitizer. In addition, experimental data were normalized to a range of 0-1, using Eq. 1, to avoid overfitting and reduce the computation complexity (You and Zhang, 2017):

$$\text{Normalized value } (X) = \frac{\mathrm{x_i} - \text{minimum value of data}}{\text{maximum value of data} - \text{minimum value of data}} \times (1-0) + 0.1 \qquad \text{(Eq. 1)}$$

where x_i is any data.

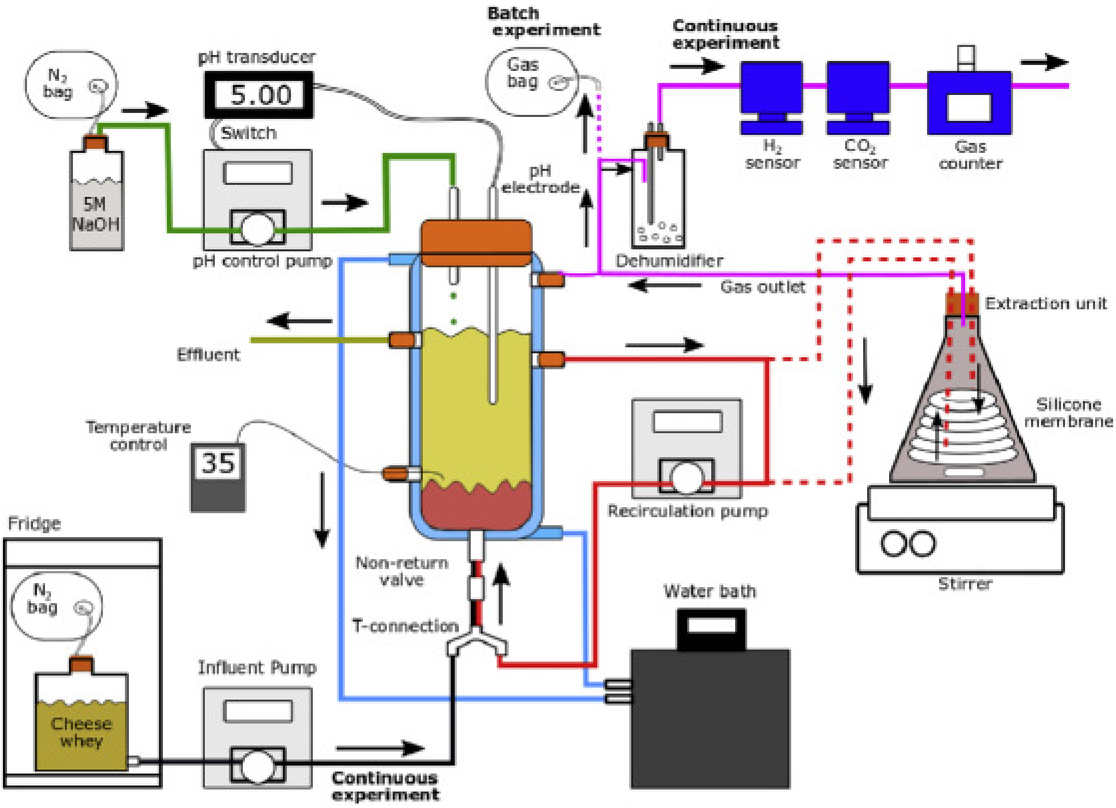

Fig. 1. Schematic setup for H_2 production from wastewater by dark fermentation (Dessì et al., 2020).

2.2. Pearson correlation coefficient

In order to compute the linear correlation or relation validity between two parameters affecting the H_2 production in the fermentation process, the Pearson correlation coefficient (*r*) was used. Pearson correlation coefficient was calculated by Eq. 2 (Hasheminasab et al., 2020).

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad \text{(Eq. 2)}$$

where Y and X are the peer parameters, $\bar{Y}$ and $\bar{X}$ are the means of the peer parameters studied for their linear correlations, and n is the sample size.

2.3. Selection of ML-assisted procedures and modelling generality

Regarding the Occam's Razor's principle stating that "a model should be as simple as possible, and as complex as needed" (Baeten et al., 2018), along with the different performances of the various ML-assisted procedures in different applications (Hosseinzadeh et al., 2020a; Hosseinzadeh et al., 2020c; Zaghloul et al., 2021), the selection of the most appropriate procedures will be crucial. Therefore, seven ML-assisted procedures, including gradient boosting machine (GBM), support vector machine (SVM), random forest (RF), AdaBoost, multilayer perceptron (MLP), linear regression (LR) and Ridge from *Scikit-learn* library were pre-screened by considering the default hyperparameters which may be obtained from to find the more proper approaches. The mean square error (MSE) and determination coefficient (R^2) were used to evaluate the outcomes of pre-screened approaches. To pre-screen and conduct deeply modelling, all datasets were randomly partitioned into training datasets (80%) and test datasets (20%). To avoid wasting the data and overfitting, cross-validation with 5-folds was used to check the validation of the developed models. The test dataset was applied to monitor the generalization performance of the developed model (Serfidan et al., 2020). To tune the hyperparameters, a grid search was defined for each of the selected procedures. Finally, the tuned hyperparameters were used in developing and testing the models. MSE (Eq. 3) and R^2 (Eq. 4) were used to assess and choose each procedure's most proper developed models. It is worth highlighting that the average of the statistical indices in all folds was considered to evaluate the performances of the validation phase over the modelling process.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_{prd,i} - y_{Act,i}\right)^2 \quad \text{(Eq. 3)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{prd,i} - y_{Act,i})}{\sum_{i=1}^{N}(y_{prd,i} - y_m)} \quad \text{(Eq. 4)}$$

where $y_{prd,i}$ nnd $y_{Act,i}$ are the predicted and real proportions of H_2 production, consecutively; y_m is the mean of real H_2 production, and N is the total number of data points.

2.4. Support vector machine (SVM)

SVM was proposed by Cortes and Vapnik as a supervised and well known machine learning approach designed according to the minimization of the structural risk and the theory of statistical learning (Cortes and Vapnik, 1995). This approach has been efficaciously applied for different applications, e.g. regression problems, text detection, troubleshooting and image retrieval. There are three different layers in the SVM structure network, i.e. input, hidden and output layers (Zendehboudi et al., 2018). The independent and dependent variables are located in the input and output layers, respectively. In the hidden layer, kernels are defined as a collection of the mathematical functions getting the inputs and converting them into the required forms. SVM algorithms make benefit from various types of kernels. Finding a hyperplane through nonlinear mapping to properly train the model/classify the data is the key gist of this procedure. The nonlinear input area is transferred into a high dimensional feature space. According to the reports, SVM has demonstrated better performance than the conventional statistical models in all regression analysis, pattern recognition and classification fields. When SVM is used for regression and function approximation, it is called support vector regression (SVR). General kernel functions, e.g. linear, radial basis function (rbf), and polynomial (poly) are commonly applied in various SVMs (Zendehboudi et al., 2018).

The independent variables were regarded as the inputs to develop an SVR model for H2 production from wastewater by the dark fermentation process. The generality of the modelling was based on the condition in section 2.2. However, to tune the hyperparameters and selection of the best kernel, a grid search was defined to tune and optimize all the hyperparameters, i.e. C (1, 100 and 50), epsilon (0.01, 0.1, 0.15, 0.3, 0.8, 1 and 2), and degree (2, 3, 4 and 5), to find the best condition of the hyperparameters for each of the kernels. Then, the tuned hyperparameters along with the related kernel were used to develop the models. In the end, the most appropriate one was selected.

2.5. Gradient boosting machine (GBM)

GBM is an ensemble and powerful supervised machine learning approach proposed first by Friedman and can model and analyze data for regression and classification problems (Cai et al., 2020; Friedman, 2001). In GBMs, which are from the decision tree category, there are three elements, i.e. weak and strong learner, loss function and additive model. The weak or base learner is introduced as the initial decision trees, having at least rarely better prediction strength than the random guess; the strong one is a learner whose performance in prediction is considerable and created with a combination of several weak learners. GBMs use training decision trees in a gradual, additive and serial method to model and analyze the processes by boosting the weak learners into the strong ones. In order to reduce the total error or loss function, new weak or base learners are added and trained to decrease the error of the model. Meanwhile, the present weak learners in the model will not be altered (Grillone et al., 2020; Nguyen et al., 2021). To develop a GBM for this process, a grid search was employed to find the best condition of the hyper-parameters in a grid. Although finding the hyperparameters' proper proportions in a grid sometimes needs unacceptable time, it can assure to find the optimal conditions of the hyper-parameters (Zhou et al., 2021). Some of the main hyper-parameters considered in this procedure were the number of gradient boosted trees (*n_estimator*), a minimum number of samples per leaf (*min_samples_leaf*) and required to split an internal node (*min_samples_split*), maximum depth of trees of GBM (*max_depth*) and the number of features for best split (*max_features*). These parameters were tuned in the ranges (100-1000), (2, 3, 4, 5, 6 and 7), (2, 3, 4, 5, 6 and 7), (1, 2, 3, 4 and 5) and (2, 3, 4, 5, 6 and 7) consecutively.

2.6. Random Forest (RF)

RF is a supervised machine learning approach that models both classification and regression phenomena (Li et al., 2018), which Breiman first proposed to work according to the regression trees (Ma and Cheng, 2016). RF produces a wide range of decision trees as a function of regression so that the ultimate proportion of the response variable is the mean of all decision trees (Li et al., 2018). As a single regression tree is insufficient to develop a proper model in most items, the RF algorithm was suggested to resolve the problem (Ma and Cheng, 2016). In developing the RF model, the generality of the modelling was conducted based on section 2.2

and in a grid search. The hyperparameters, i.e. number of gradients boosted trees (*n_estimator*), a minimum number of samples per leaf (*min_samples_leaf*) and required to split an internal node (*min_samples_split)* and the number of features for best split (*max_features*) were tuned in the ranges (100-1000), (1, 2, 3, 4, 5, 6, 7 and 8), (0.5, 1 2, 3, 4, 5 and 6) and (2, 3, 4, 5, 6, 7 and 8) consecutively.

2.7. AdaBoost

The AdaBoost procedure can be applied for classification and regression problems (Min and Luo, 2016). This procedure is classified as an ensemble machine learning based on finding a promising predictor from a number of weak predictors (Min and Luo, 2016). The generality of the AdaBoost model development for this process was according to the mentioned condition in section 2.2. However, to tune the hyperparameters and selection of the best loss function, a grid search was defined to tune and optimize all the hyperparameters, i.e. several gradients boosted trees (*n_estimator*) and learning rate in the ranges (20-500) and (0.1, 0.5, 1, 2, 3, 4 and 5) respectively. In addition, like all three other models (SVR, GBM and RF), the learning curve were prepared to show the goodness of fit of the models.

2.7. Variable importance evaluation

Permutation variable importance (PVI) proposed by Breiman (2001) is a procedure to inspect any fitted model in the tabular data. This procedure considers the developed model's errors in predicting the output with a random permutation of the considered input. So that the more the errors, the more the importance of the feature (Mohammadifar et al., 2021). Regarding the errors, MSE was used to measure the relative importance of the features. There are various merits for PVI procedure, e.g. fast and easy to calculate, a general method, considering both individual and interactive effects of each variable (Altmann et al., 2010; Antoniadis et al., 2021; Wei et al., 2015). To identify the relative importance of the input variables in H_2 production from wastewater through dark fermentation process, PVI procedure was used for all the developed GBM, SVR, RF and AdaBoost models.

2.8. Comparison of model performance

Four statistical indices, determination coefficient (R^2), MSE and MAE (Eq. 5) were used to compare the performances and strengths of the developed SVR, GBM, RF and AdaBoost models to predict the H_2 production from wastewater by the fermentation process. It is worth mentioning that the test datasets were used to calculate the mentioned statistical parameters.

$$MAE = 1 - \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \qquad \text{(Eq. 5)}$$

where y_i , x_i and n are predicted value, actual value and total number of data points, respectively.

3. Results and discussion

3.1. Selection of ML-assisted procedures

The performances of various ML-assisted procedures in modelling H_2 production from wastewater using dark fermentation were assessed, with their results presented in Table 1. Based on the statistical indices' values indicating the models' prediction strengths (Table 1), GBM, SVR, RF and AdaBoost were selected as the most efficient modelling procedures. Furthermore, various studies demonstrate promising performances of GBM, SVR, RF and AdaBoost in different applications (Almuhtaram et al., 2021; Thompson and Dickenson, 2021; Xia et al., 2020; Xing et al., 2019). Therefore, these four procedures were used in this study to model the H2 production by fermentation process deeply.

Table 1. Performances of different ML-assisted procedures in modelling H_2 production during dark fermentation process

	GBM	RF	AdaBoost	SVR	MLP	LR	Ridge
Total-Train R^2	0.985	0.976	0.910	0.853	0.737	0.766	0.750
Total-Test R^2	0.802	0.805	0.805	0.734	0.685	0.693	0.670
Train MSE	0.002	0.004	0.014	0.023	0.038	0.037	0.040
Test MSE	0.023	0.023	0.023	0.032	0.038	0.037	0.039

3.2. SVR

3.2.1. Kernel selection and tunning the hyperparameters

To select the most appropriate kernel, the different conditions of the hyperparameters were tuned by grid search with each mentioned kernels. All values of the tuned hyperparameters and their MSE and R2 values in different modelling phases are listed in Table 2. As can be observed, rbf was shown as the best kernel with C, degree and epsilon of 11, 2 and 0.01 consecutively. The MSE and R^2 of the training and validation phases were 0.021 and 0.864, and 0.024 and 0.845 correspondingly. 20% of the unseen data points regarded as the test dataset were used to test the performance of the developed model in H_2 production by fermentation process as well. As observed in Table 2 and Fig. 2, the prediction strength of this model was 88.5%, with an MSE of 0.016. Moreover, the prediction strength of the SVR model in the test phase showed the considerable performances of this model in this field. In addition, Chen et al. (2015) used SVM to model the production of iturin A through the fermentation process. Using asparagine concentration, glutamic acid and proline as inputs, they introduced SVM as a proper model with a relatively low root MSE of 466.13, which agrees with the present study.

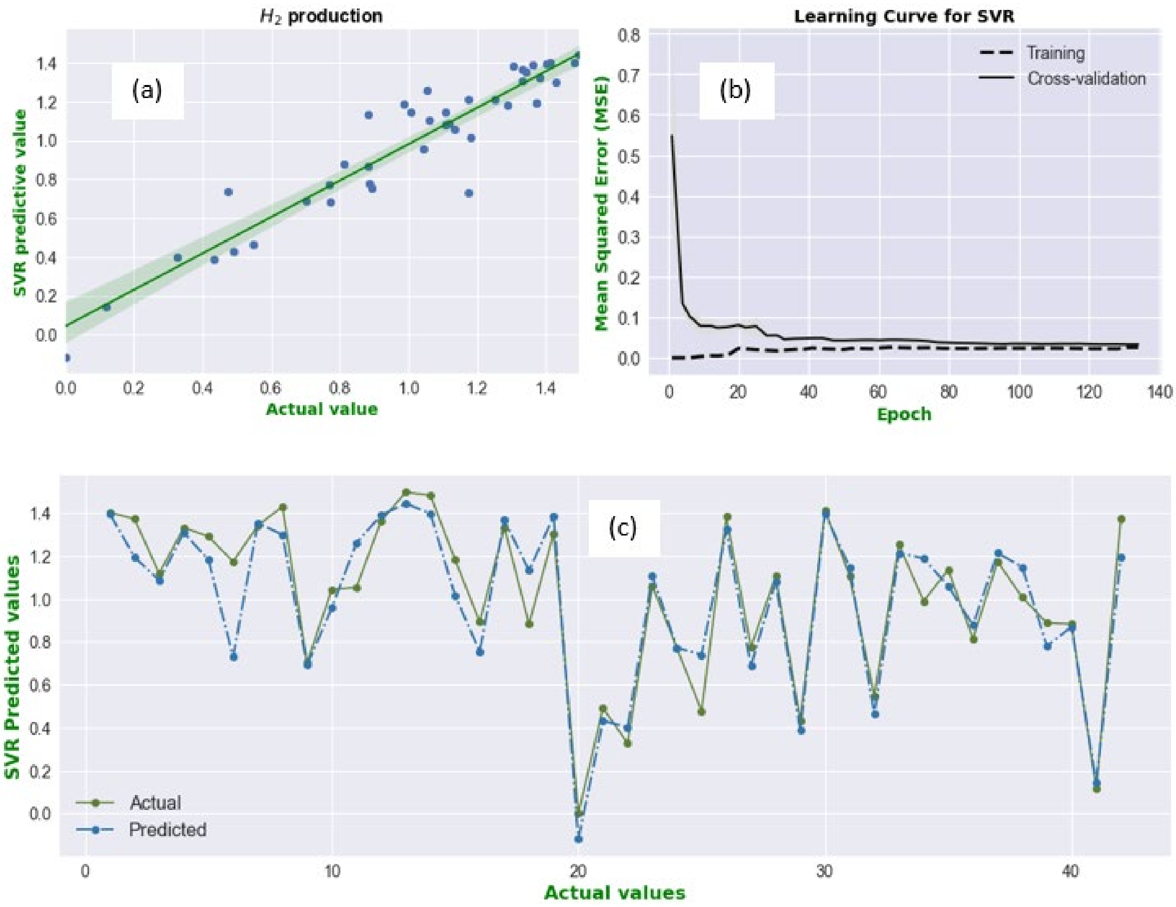

Fig. 2. The presentation of the SVR model. a) correlation coefficient of the model in test phase, b) learning curve of the developed model, and c) prediction strength of the model in test phase.

Table 2. SVR model outcomes using various kernels with tuned hyperparameters

	Tuned hyper-parameters by grid search			Determination coefficient (R^2)				MSE			
	C	degree	epsilon	Train	validation	Total-Train	Test	Train	validation	Total-Train	Test
Linear	91	2	0.15	0.747	0.713	0.746	0.745	0.038	0.043	0.039	0.035
rbf	**11**	**2**	**0.01**	**0.864**	**0.845**	**0.863**	**0.885**	**0.021**	**0.024**	**0.021**	**0.016**
poly	21	2	0.01	0.856	0.814	0.855	0.874	0.021	0.029	0.022	0.017

3.2.2. SVR learning curve

Underfitting and overfitting are two main problems, which can be observed in models developed by machine learning procedures. In underfitting, the model cannot learn the process, while overfitting is more complicated, according to which the generalizability of the model will not be acceptable; that is, the developed model only memorizes the train dataset and cannot predict the unseen dataset (Bejani and Ghatee, 2021). Since demonstrating the fact that there is no underfitting and overfitting in the developed models is regarded as a very important part of the modelling process, the learning curve, which is deemed as an effective tool to show underfitting/overfitting/good fitting condition of the models was provided for the developed model. The learning curve is an efficient tool showing the performance of the model in training and validation phases over different epochs (Braga et al., 2019). The learning curve of the SVR model in train and validation phases are depicted in Fig. 2. Based on which MSEs of the validation decreased approximately to epoch 70, followed a stable and consistent condition with a small gap with train minimum MSEs pointing out that there is no overfitting and underfitting.

3.3. GBM

To develop the GBM, the considered hyperparameters, i.e. number of gradient boosted trees, maximum depth of trees of GBM, number of features for best split, a minimum number of samples per leaf, minimum number of samples per split were tuned in a grid search, and the obtained best condition of these parameters were 100, 5, 6, 3 and 6 respectively. With respect to this condition, the training and cross-validation were conducted, and the R^2 values (0.996 and 0.813) and MSE values (0.0005 and 0.027) were obtained for these two phases correspondingly. The obtained R^2 for the total train (train and validation) along with the test phases were 0.995 and 0.893, and MSEs of 0.0008 and 0.015, respectively, showing that the model has considerable prediction strength (89.3%) in H_2 prediction from wastewater by the fermentation process. Fig. 3 depicts the test dataset's fitting condition in the model's test phase, demonstrating promising prediction strength of the model. In addition, Zhuang et al. (2021) applied GBM to model a membrane bioreactor performance in COD, NH_4-N and TN removal. Their GBM model could show a considerable performance with R^2 of 0.847, 0.792 and 0.851 correspondingly. Therefore, these findings show the considerable potential of GBM in different applications.

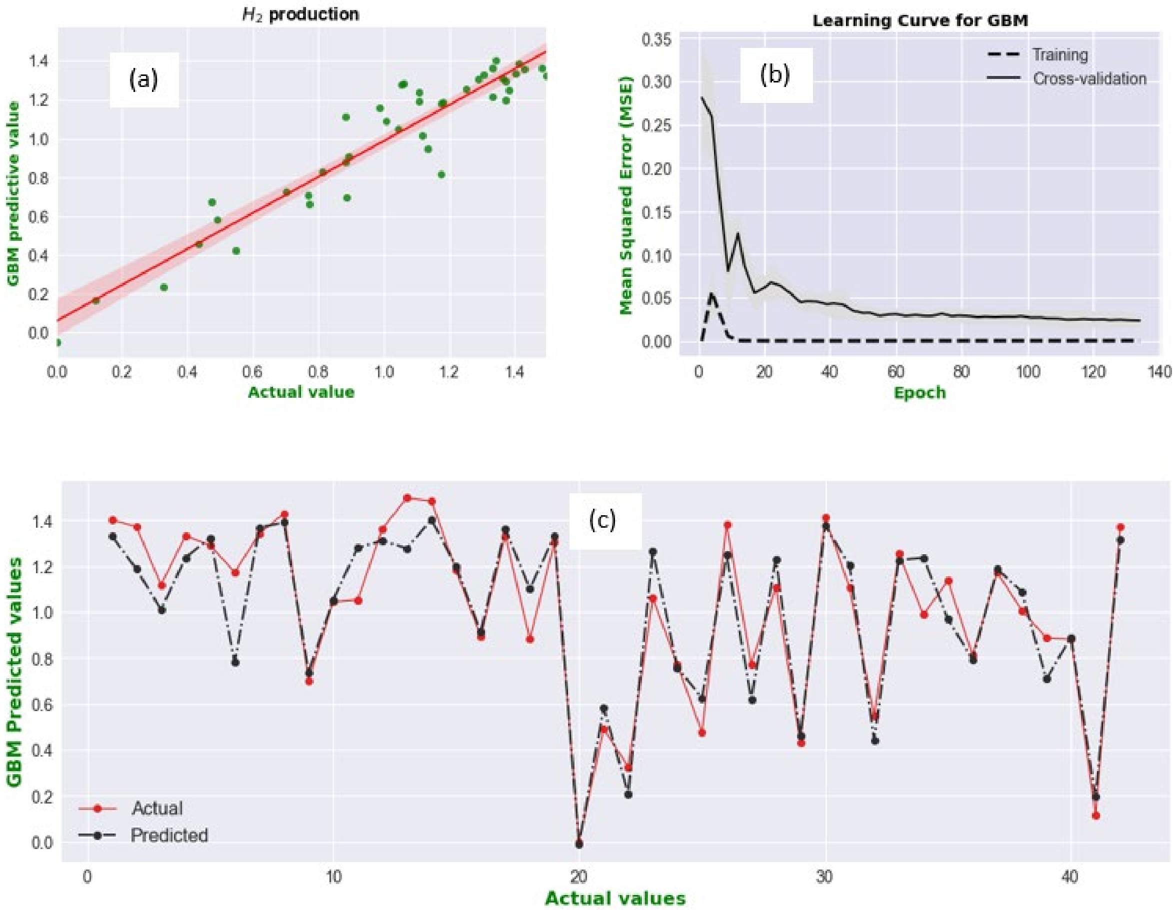

Fig. 3. The presentation of the GBM model; a) correlation coefficient of the model in test phase; b) learning curve of the developed model, and c) prediction strength of the model in the test phase.

In addition, in order to check the good fitting condition of the model and showing there is no overfitting in the model, as observed in Fig. 3, the MSEs of the training and validation phases experienced a decreasing trend with the same pattern with a small gap between themselves pointing out that there is no overfitting. Approximately from epoch 60 there is a constant and stable condition in the MSEs of these phases.

3.4. RF

The hyperparameters were tuned in a grid search to construct an RF model. Following optimization, the appropriate conditions were determined to be 7, 1000, 1 and 2 for the number of features for best split, several gradients boosted trees, a minimum number of samples per leaf, and a minimum number of samples per split, respectively. Regarding the conditions attained, the R^2 and MSE for the training phase (0.973, 0.004) and validation phase (0.823, 0.025) were obtained in the same order. The attained R^2 for the total train (train and validation) coupled with the test phases were 0.975 and 0.902, with MSE of 0.004 and 0.016, correspondingly demonstrating that the model has considerable prediction strength (86.3%) in H_2 prediction from wastewater by the fermentation process. Fig. 4 depicts the fitting condition of the test dataset in the test phase of the model. All of the provided information presents an acceptable model for this process.

In addition, as observed in Fig. 4, the MSE of the various epochs during training and validation phases approximately experienced a decreasing trend and showed that there is no overfitting on the developed RF model. It can be seen that these MSEs follow the same pattern with a minimum gap between themselves for training and validation phases from almost epoch 60, pointing out that the prediction strengths and errors of the condition in these two phases experience stability and consistency.

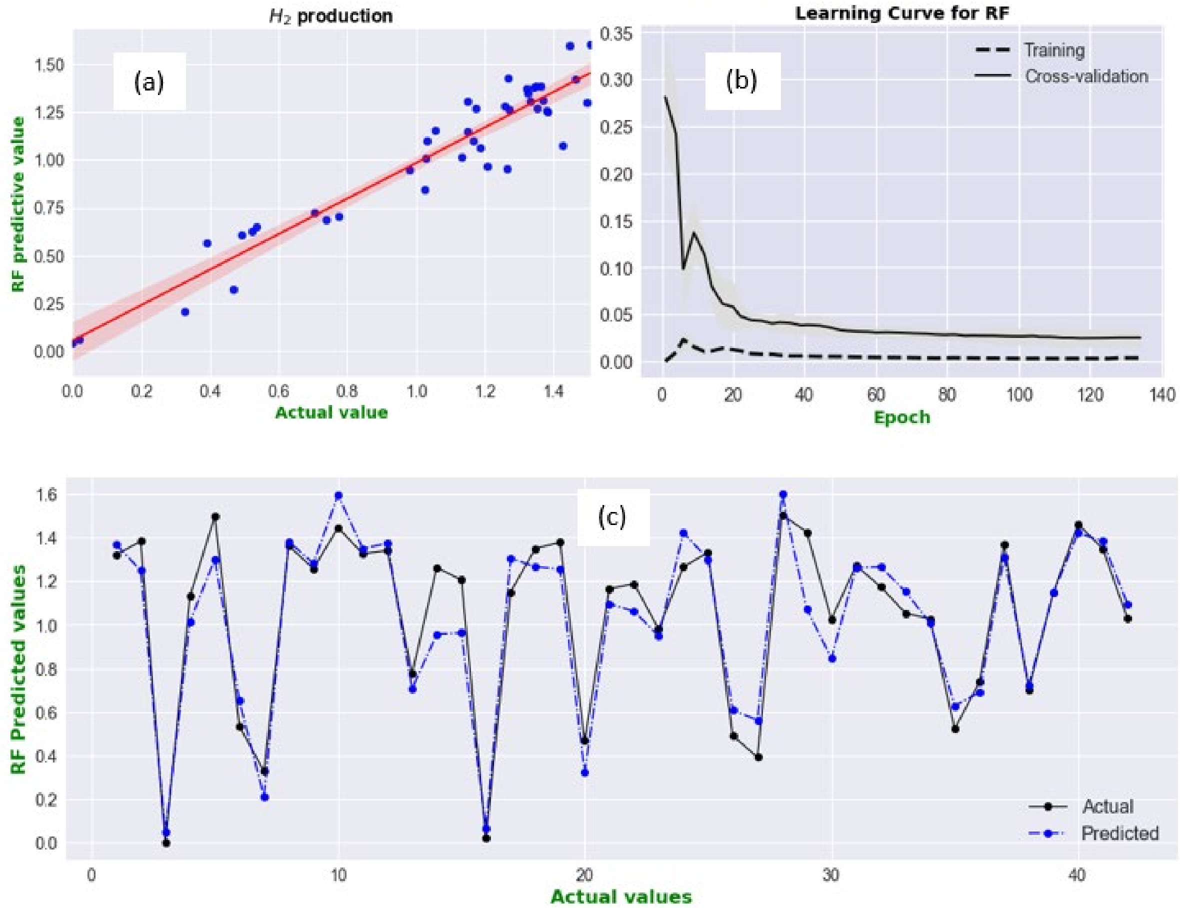

Fig. 4. The presentation of the RF model; a) correlation coefficient of the model in test phase; b) learning curve of the developed model, and c) prediction strength of the model in test phase.

3.5. AdaBoost

3.5.1. Loss function selection and tunning the hyperparameters

To select the most appropriate loss function, the mentioned different conditions of the hyperparameters were tuned by grid search with each mentioned loss function. All values of the tuned hyperparameters and their MSE and R2 values in different modelling phases have been listed in Table 3. As can be seen, linear was the best loss function with *n_estimator* and learning rate of 200 and 0.1, respectively. The MSE and R^2 of the training and validation phases were (0.014 and 0.027) and (0.901 and 0.801) correspondingly.

Table 3. AdaBoost model outcomes using various loss functions with tuned hyperparameters

	Grid search		(R^2)	(R^2)	(R^2)	(R^2)	MSE	MSE	MSE	MSE
	n-estimator	learning rate	Train	validation	Total-Train	Test	Train	validation	Total-Train	Test
Linear	**200**	**0.1**	**0.901**	**0.801**	**0.889**	**0.888**	**0.014**	**0.027**	**0.015**	**0.023**
Square	80	1.0	0.911	0.816	0.909	0.844	0.012	0.025	0.013	0.029
Exponential	260	0.1	0.914	0.813	0.906	0.847	0.013	0.027	0.014	0.024

Like the previous models, 20% of the unseen data points regarded as the test dataset were used to test the performance of the developed model in H_2 production by the fermentation process. As observed in Table 3 and Fig. 5, the considerable prediction strength of this model in the test phase was obtained 87.4% with an MSE of 0.023. In addition, Thompson and Dickenson (2021) applied AdaBoost to detect *de facto* reuse in water. In a way that TOC, turbidity, temperature, ORP, conductivity, pH, UVA_{254} and tryptophan-like fluorescence were used as inputs to model the quality of a surface water resource before intake for drinking purpose to produce proper alerts for the operators to perform required actions to intake water with better quality. The model developed could successfully work with an accuracy of more than 99%, demonstrating the high potential of AdaBoost in other different applications.

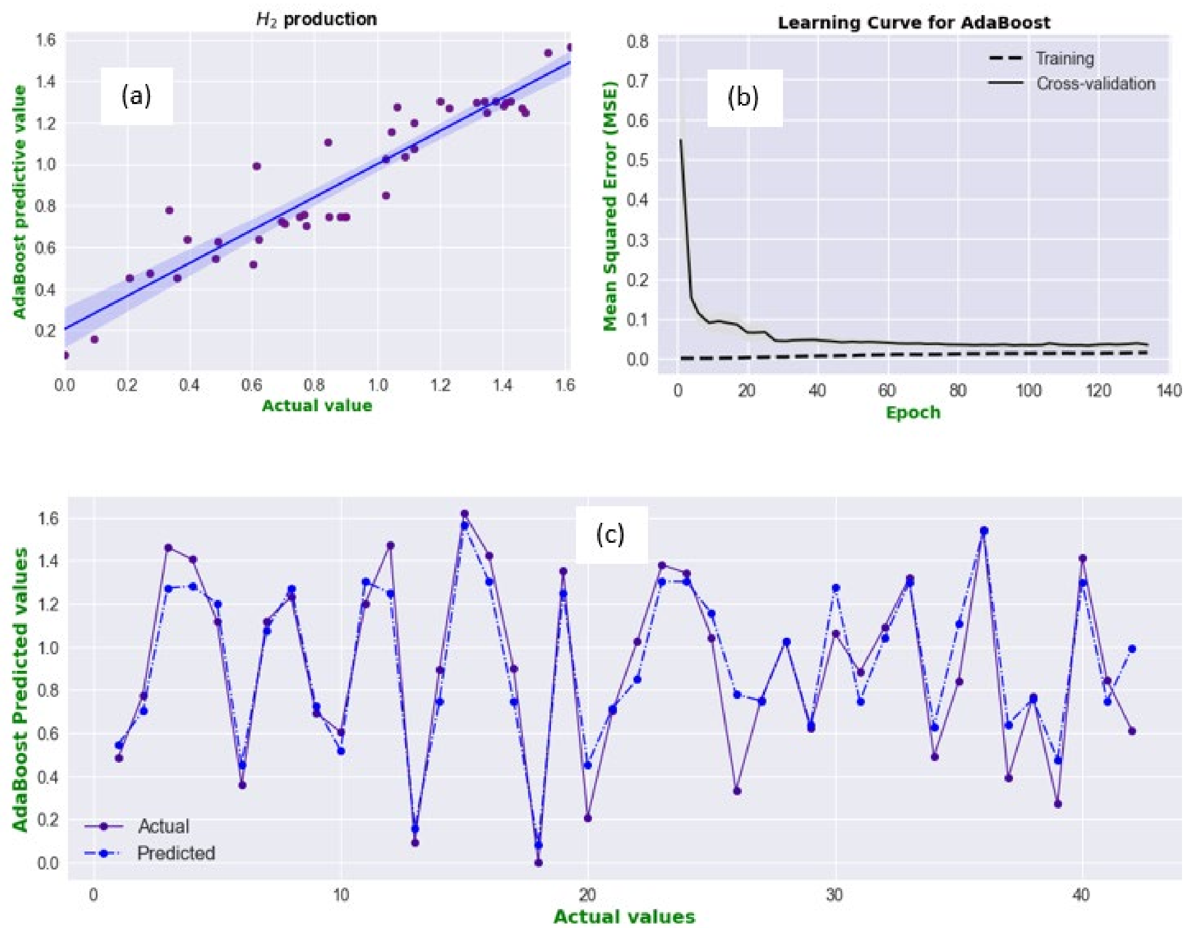

Fig. 5. The presentation of the AdaBoost model; a) correlation coefficient of the model in test phase; b) learning curve of the developed model, and c) prediction strength of the model in test phase.

Furthermore, the learning curve of the developed AdaBoost model (Fig. 5) points out the training and validation learning condition of the developed model in different epochs, representing no overfitting in the developed model.

3.6. Relative importance of the variables

The obtained relative importance of the variables by PVI procedure in the developed GBM, SVR, RF and AdaBoost is demonstrated in Fig 6. As observed, different importance values w were obtained for the inputs by PVI of each of these four models. Regarding the results, ethanol shows more importance in H_2 production from wastewater by the dark fermentation process, which can be completely justified. In strict anaerobic processes, solventogenesis and acidogenesis are two main pathways producing solvent, e.g. ethanol, and acid, e.g. acetate and butyrate, respectively. Since ethanol as a solvent can undesirably affect some of the H_2 producing bacteria, and solventogenesis is not a friendly pathway for H_2 production, the considerable importance of this variable can completely be reasonable (Wong et al., 2014). In addition, the demonstrated higher importance of A/B ratio and acetate and butyrate by the SVR-PVI and RF-PVI can be justified because A/B ascertains whether the fermentation pathway is acetate or butyrate one. The proportion of the produced H2 from one mole glucose in the acetate pathway is two folds higher than that of the butyrate one (Liu et al., 2006; Wong et al., 2014). Regarding the importance of Fe and Ni as cofactors of the H_2 production pathways in the dark fermentation process, since [Fe-Fe] and [Ni-Fe] are two groups presented in the H_2 catalyzing enzymes, basically, it seems that the higher importance of the Fe can be more justifiable than Ni (Karadag and Puhakka, 2010). However, it is obvious that the considerable importance of biomass, COD and pH cannot be ignored in the dark fermentation process because, without biomass and COD, the biological activity leading to H_2 production will not be possible. At alkaline pH, hydrogen-producing bacteria will not properly activate and produce H_2 (Durán et al., 2020). The less importance of these three variables in Fig. 5, especially in RF-PVI, can be attributed to the fact that the optimum range of pH, COD and biomass in the dark fermentation process have been cleared. Most researchers consider the optimum condition, so there is a limit range of values for these variables resulting in these outcomes. Therefore, among all these four analyses, RF-PVI and SVR-PVI procedures pointed out more accurate conditions in comparison to the others. Overall, RF-PVI can be more better option than SVR-PVI as well.

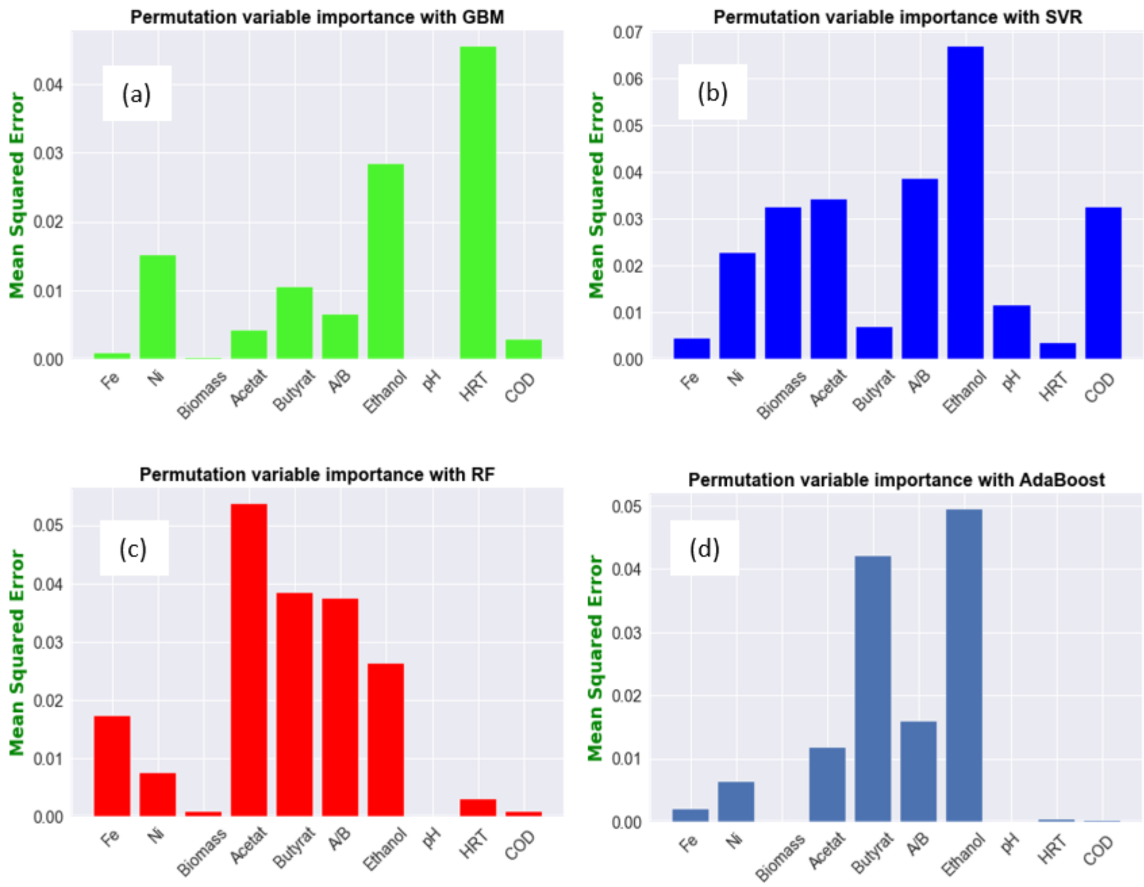

Fig. 6. Permutation variable importance through a) GBM; b) SVR; c) RF; d) AdaBoost models

3.7. Comparison of the models

In order to assess the performances of the developed models, i.e. GBM, SVR, RF and AdaBoost, three various statistical indexes showing the strengths of the models in H_2 production from wastewater by fermentation process were employed. Based on the results in Table 4, approximately all four developed models pointed out the same strengths; however, there were a few differences between these models. From the errors, both MSE and MAE, SVR had the lowest one followed by GBM, RF and AdaBoost with increasing order. The residual errors of these models in the test phase are presented in Fig 7. Furthermore, regarding the R^2 of these models, RF showed rarely better performance than the others did. Generally with considering both R^2 and errors, SVR and GBM and RF demonstrated promising performance compared to the AdaBoost one.

Table 4. Performance comparison of GBM, SVR, RF and AdaBooost models developed for H_2 production from wastewater by fermentation

Models	Statistical indices		
	R^2	MSE	MAE
GBM	0.893	0.015	0.097
SVR	0.885	0.015	0.092
RF	0.902	0.016	0.098
AdaBoost	0.889	0.015	0.117

Fig. 7. The residual errors of the developed GBM, SVR, RF and AdaBoost models in prediction of H_2 production from wastewater by fermentation process.

3.8. Pearson correlation coefficient

This analysis shows the linear relationships between the variables. However, it is worth highlighting that if the correlation of both variables considered is input, the *R* will be near 0; however, the opposite statement is incorrect (Nguyen et al., 2021). As shown in Fig. 8 pointing out colour map correlation matrix and pair-wise scatter correlation plots of the variables, Fe and Ni have a strong negative correlation showing that the more the biomass, the more the consumption of the Fe and Ni. It is clear that the correlation between the Fe-biomass with 0.27 is rarely higher than the Ni-biomass correlation, which can be attributed to the fact that the enzymes catalyzing the biohydrogen production are [Fe-Fe] and [Ni-Fe] groups requiring more Fe than the Ni (Karadag and Puhakka, 2010). Furthermore, the strong negative correlation between the pH and COD can be observed because the more the COD, the more the production of the VFAs reduces the pH.

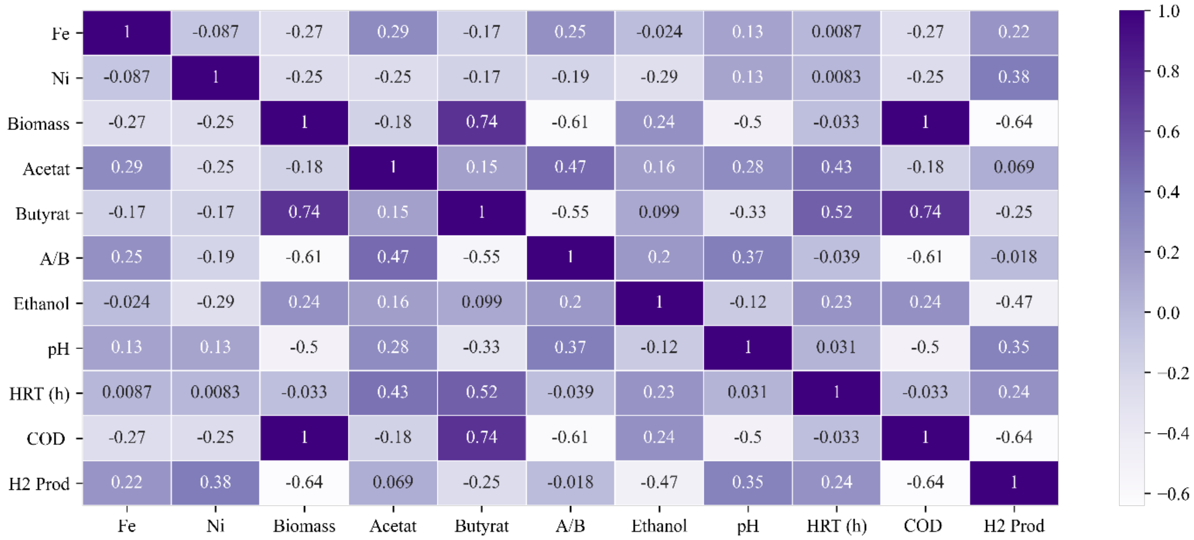

Fig. 8. Correlation coefficients of the independent variables affecting H_2 production from wastewater in the fermentation process.

4. Conclusions

H_2 production from wastewater by dark fermentation process is regarded as an interesting process. Seven different types of machine learning approaches were pre-screened to model this process to find the most appropriate ones for this application. Based on the results, the SVR, GBM, RF and AdaBoost were selected and deeply model this process. The results showed that all four developed models showed approximately the same performance to the dark fermentation process of H2 production from wastewater. Regarding permutation relative variable importance, the RF-PVI demonstrated better outcomes, based on which acetate, butyrate, A/B ratio, ethanol, Fe and Ni were identified as the most important ones with a decreasing order.

Acknowledgements

We thank the University of Technology Sydney for an International Research Scholarship and UTS President’s Scholarship.

References

Alassi, A., Bañales, S., Ellabban, O., Adam, G. and MacIver, C. 2019. HVDC Transmission: Technology Review, Market Trends and Future Outlook. Renewable and Sustainable Energy Reviews 112, 530-554.

Alidadi, H., Hosseinzadeh, A., Najafpoor, A.A., Esmaili, H., Zanganeh, J., Dolatabadi Takabi, M. and Piranloo, F.G. 2016. Waste recycling by vermicomposting: Maturity and quality assessment via dehydrogenase enzyme activity, lignin, water soluble carbon, nitrogen, phosphorous and other indicators. Journal of Environmental Management 182, 134-140.

Almuhtaram, H., Zamyadi, A. and Hofmann, R. 2021. Machine learning for anomaly detection in cyanobacterial fluorescence signals. Water Research 197, 117073.

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T. 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26(10), 1340-1347.

Antoniadis, A., Lambert-Lacroix, S. and Poggi, J.-M. 2021. Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety 206, 107312.

Baeten, J.E., van Loosdrecht, M.C.M. and Volcke, E.I.P. 2018. Modelling aerobic granular sludge reactors through apparent half-saturation coefficients. Water Research 146, 134-145.

Baeyens, J., Zhang, H., Nie, J., Appels, L., Dewil, R., Ansart, R. and Deng, Y. 2020. Reviewing the potential of bio-hydrogen production by fermentation. Renewable and Sustainable Energy Reviews 131, 110023.

Bao, T., Damtie, M.M., Hosseinzadeh, A., Frost, R.L., Yu, Z.M., Jin, J. and Wu, K. 2020a. Catalytic degradation of P-chlorophenol by muscovite-supported nano zero valent iron composite: Synthesis, characterization, and mechanism studies. Applied Clay Science 195, 105735.

Bao, T., Damtie, M.M., Wei, W., Phong Vo, H.N., Nguyen, K.H., Hosseinzadeh, A., Cho, K., Yu, Z.M., Jin, J., Wei, X.L., Wu, K., Frost, R.L. and Ni, B.-J. 2020b. Simultaneous adsorption and degradation of bisphenol A on magnetic illite clay composite: Eco-friendly preparation, characterizations, and catalytic mechanism. Journal of Cleaner Production, 125068.

Bejani, M.M. and Ghatee, M. 2021. A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review.

Braga, D., Madureira, A.M., Coelho, L. and Ajith, R. 2019. Automatic detection of Parkinson's disease based on acoustic analysis of speech. Engineering Applications of Artificial Intelligence 77, 148-158.

Breiman, L. 2001. Random forests. Machine learning 45(1), 5-32.

Cai, J., Xu, K., Zhu, Y., Hu, F. and Li, L. 2020. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied Energy 262, 114566.

Cao, S., Sun, F., Lu, D. and Zhou, Y. 2019. Characterization of the refractory dissolved organic matters (rDOM) in sludge alkaline fermentation liquid driven denitrification: Effect of HRT on their fate and transformation. Water Research 159, 135-144.

Chen, F., Li, H., Xu, Z., Hou, S. and Yang, D. 2015. User-friendly optimization approach of fed-batch fermentation conditions for the production of iturin A using artificial neural networks and support vector machine. Electronic Journal of Biotechnology 18(4), 273-280.

Cortes, C. and Vapnik, V. 1995. Support-vector networks. Machine Learning 20(3), 273-297.

Dessì, P., Asunis, F., Ravishankar, H., Cocco, F.G., De Gioannis, G., Muntoni, A. and Lens, P.N. 2020. Fermentative hydrogen production from cheese whey with in-line, concentration gradient-driven butyric acid extraction. International Journal of Hydrogen Energy 45(46), 24453-24466.

Durán, F., Robles, Á., Giménez, J.B., Ferrer, J., Ribes, J. and Serralta, J. 2020. Modeling the anaerobic treatment of sulfate-rich urban wastewater: Application to AnMBR technology. Water Research 184, 116133.

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gómez-Marín, N. and Bridgwater, A.V. 2021. Mapping bioenergy stakeholders: A systematic and scientometric review of capabilities and expertise in bioenergy research in the United Kingdom. Renewable and Sustainable Energy Reviews 137, 110496.

Grillone, B., Danov, S., Sumper, A., Cipriano, J. and Mor, G. 2020. A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings. Renewable and Sustainable Energy Reviews 131, 110027.

Hasheminasab, M., Kermani, M.J., Nourazar, S.S. and Khodsiani, M.H. 2020. A novel experimental based statistical study for water management in proton exchange membrane fuel cells. Applied Energy 264, 114713.

Hosseinzadeh, A., Baziar, M., Alidadi, H., Zhou, J.L., Altaee, A., Najafpoor, A.A. and Jafarpour, S. 2020a. Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions. Bioresource Technology 303, 122926.

Hosseinzadeh, A., Zhou, J.L., Altaee, A., Baziar, M. and Li, D. 2020b. Effective modelling of hydrogen and energy recovery in microbial electrolysis cell by artificial neural network and adaptive network-based fuzzy inference system. Bioresource Technology 316, 123967.

Hosseinzadeh, A., Zhou, J.L., Altaee, A., Baziar, M. and Li, X. 2020c. Modeling water flux in osmotic membrane bioreactor by adaptive network-based fuzzy inference system and artificial neural network. Bioresource technology 310, 123391.

Hosseinzadeh, A., Zhou, J.L., Navidpour, A.H. and Altaee, A. 2021. Progress in osmotic membrane bioreactors research: Contaminant removal, microbial community and bioenergy production in wastewater. Bioresource Technology, 124998.

Huang, Y., Surawski, N.C., Organ, B., Zhou, J.L., Tang, O.H. and Chan, E.F. 2019. Fuel consumption and emissions performance under real driving: Comparison between hybrid and conventional vehicles. Science of the Total Environment 659, 275-282.

Kamranifar, M., Al-Musawi, T.J., Amarzadeh, M., Hosseinzadeh, A., Nasseh, N., Qutob, M. and Arghavan, F.S. 2021. Quick adsorption followed by lengthy photodegradation using FeNi3@SiO2@ZnO: A promising method for complete removal of penicillin G from wastewater. Journal of Water Process Engineering 40, 101940.

Karadag, D. and Puhakka, J.A. 2010. Enhancement of anaerobic hydrogen production by iron and nickel. International Journal of Hydrogen Energy 35(16), 8554-8560.

Li, Y., Zhang, Z., Xia, C., Jing, Y., Zhang, Q., Li, S., Zhu, S. and Jin, P. 2020. Photo-fermentation biohydrogen production and electrons distribution from dark fermentation effluents under batch, semi-continuous and continuous modes. Bioresource Technology 311, 123549.

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J.C.W., van den Bossche, P., Van Mierlo, J. and Omar, N. 2018. Random forest regression for online capacity estimation of lithium-ion batteries. Applied Energy 232, 197-210.

Liu, D., Liu, D., Zeng, R.J. and Angelidaki, I. 2006. Hydrogen and methane production from household solid waste in the two-stage fermentation process. Water Research 40(11), 2230-2236.

Liu, H., Wang, J., Liu, X., Fu, B., Chen, J. and Yu, H.-Q. 2012. Acidogenic fermentation of proteinaceous sewage sludge: Effect of pH. Water Research 46(3), 799-807.

Ma, J. and Cheng, J.C.P. 2016. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. Applied Energy 183, 193-201.

Mai-Moulin, T., Hoefnagels, R., Grundmann, P. and Junginger, M. 2021. Effective sustainability criteria for bioenergy: Towards the implementation of the european renewable directive II. Renewable and Sustainable Energy Reviews 138, 110645.

Min, H. and Luo, X. 2016. Calibration of soft sensor by using Just-in-time modeling and AdaBoost learning method. Chinese Journal of Chemical Engineering 24(8), 1038-1046.

Mohammadifar, A., Gholami, H., Comino, J.R. and Collins, A.L. 2021. Assessment of the interpretability of data mining for the spatial modelling of water erosion using game theory. Catena 200, 105178.

Nguyen, H., Vu, T., Vo, T.P. and Thai, H.-T. 2021. Efficient machine learning models for prediction of concrete strengths. Construction and Building Materials 266, 120950.

Pradhan, N., Dipasquale, L., d'Ippolito, G., Fontana, A., Panico, A., Pirozzi, F., Lens, P.N.L. and Esposito, G. 2016. Model development and experimental validation of capnophilic lactic fermentation and hydrogen synthesis by Thermotoga neapolitana. Water Research 99, 225-234.

Sekoai, P.T., Ghimire, A., Ezeokoli, O.T., Rao, S., Ngan, W.Y., Habimana, O., Yao, Y., Yang, P., Yiu Fung, A.H., Yoro, K.O., Daramola, M.O. and Hung, C.-H. 2021. Valorization of volatile fatty acids from the dark fermentation waste Streams-A promising pathway for a biorefinery concept. Renewable and Sustainable Energy Reviews 143, 110971.

Serfidan, A.C., Uzman, F. and Türkay, M. 2020. Optimal estimation of physical properties of the products of an atmospheric distillation column using support vector regression. Computers & Chemical Engineering 134, 106711.

Thompson, K.A. and Dickenson, E.R.V. 2021. Using machine learning classification to detect simulated increases of de facto reuse and urban stormwater surges in surface water. Water Research, 117556.

Toquero, C. and Bolado, S. 2014. Effect of four pretreatments on enzymatic hydrolysis and ethanol fermentation of wheat straw. Influence of inhibitors and washing. Bioresource Technology 157, 68-76.

Wei, P., Lu, Z. and Song, J. 2015. A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system. Environmental Modelling & Software 70, 178-190.

Wong, Y.M., Wu, T.Y. and Juan, J.C. 2014. A review of sustainable hydrogen production using seed sludge via dark fermentation. Renewable and Sustainable Energy Reviews 34, 471-482.

Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., Jia, X., Yang, C., Liu, C., Ma, S., Lin, J., Wang, X., Hou, X., Zhang, K., Gao, X., Duan, P. and Qian, C. 2020. River algal blooms are well predicted by antecedent environmental conditions. Water Research 185, 116221.

Xing, J., Luo, K., Wang, H. and Fan, J. 2019. Estimating biomass major chemical constituents from ultimate analysis using a random forest model. Bioresource Technology 288, 121541.

You, H. and Zhang, X. 2017. Sustainable livelihoods and rural sustainability in China: Ecologically secure, economically efficient or socially equitable? Resources, Conservation and Recycling 120, 1-13.

Zaghloul, M.S., Iorhemen, O.T., Hamza, R.A., Tay, J.H. and Achari, G. 2021. Development of an ensemble of machine learning algorithms to model aerobic granular sludge reactors. Water Research 189, 116657.

Zendehboudi, A., Baseer, M.A. and Saidur, R. 2018. Application of support vector machine models for forecasting solar and wind energy resources: A review. Journal of Cleaner Production 199, 272-285.

Zhao, J., Wang, D., Li, X., Yang, Q., Chen, H., Zhong, Y. and Zeng, G. 2015. Free nitrous acid serving as a pretreatment method for alkaline fermentation to enhance short-chain fatty acid production from waste activated sludge. Water Research 78, 111-120.

Zhou, L., Fujita, H., Ding, H. and Ma, R. 2021. Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting. Applied Soft Computing 110, 107672.

Zhuang, L., Tang, B., Bin, L., Li, P., Huang, S. and Fu, F. 2021. Performance prediction of an internal-circulation membrane bioreactor based on models comparison and data features analysis. Biochemical Engineering Journal 166, 107850.

Zorpas, A.A. 2020. Strategy development in the framework of waste management. Science of The Total Environment 716, 137088.