Journal of Environmental Chemical Engineering 8 (2020) 103952

Contents lists available at ScienceDirect

Journal of Environmental Chemical Engineering

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

Random Forest as a promising application to predict basic-dye biosorption process using orange waste

Arthur Pontes de Miranda Ramos Soares[a], Frede de Oliveira Carvalho[a], Carlos Eduardo de Farias Silva[a,*], Andreza Heloiza da Silva Gonçalves[b], Ana Karla de Souza Abud[c]

[a] Technology Center, Federal University of Alagoas, Maceió, Brazil
[b] Institute of Chemistry and Biotechnology, University of Alagoas, Maceió, Brazil
[c] Food Technology Department, Federal University of Sergipe, São Cristovão, Brazil

ARTICLE INFO

Editor: G.L. Dotto

Keywords:
Machine learning
Algorithm
Artificial Neural Networks
Wastewater treatment
Methylene blue
Modelling
Python

ABSTRACT

In the present study, adsorption of methylene blue dye in residual agricultural biomass (orange bagasse) was modelled using o machine learning algorithm Random Forest (RF) and compared with the traditional Artificial Neural Networks (ANN) approach. The Machine Learning was performed using Python, a free and open source programming language. The models were built and validated with a combination of 202 independent experiments aimed at separately predicting the final concentration of methylene blue (C_f), adsorption capacity (Q) and adsorbate percentage removal ($R\%$), having as input variables: Temperature, pH, adsorbent dosage, contact time, salinity, initial methylene blue concentration and rotation. The validation process of the models was carried out using the Coefficient of Determination (R^2) and the Mean Squared Error (MSE). According to the obtained results, both RF and ANN models exhibited similar performances, as shown by their respective R^2 values of 0.9739 and 0.9734 for C_f; 0.9932 and 0.9919, for Q; 0.9318 and 0.9257 for $R\%$, as well as their respective MSE values of 0.0012 and 0.0016 for C_f; 0.0005 and 0.0007 for Q; 0.0015 and 0.0019 for $R\%$. However, RF stood out due to its capacity to better capture data variation. Finally, it was possible to point out that both methods resulted in models able to satisfactorily predict all three response variables, thereby allowing less experimental effort.

1. Introduction

The use of adsorption in the removal of toxic compounds in effluents is widely disseminated given the high efficiency and simplicity of treatment, such as dye removal using biomass (methylene blue by orange bagasse) [1–3]. In addition, the development of models capable of describing the adsorption of impurities from effluents is essential, as more information can be provided regarding the behaviour of the process, allowing a more in-depth study of its variables and the relationship between them [4]. Nevertheless, the design of phenomenological models can be quite complex, as a result of the non-linear and highly complex relationship between input and output variables of the adsorption process [5].

One alternative for modelling systems with a high degree of complexity is the use of machine learning techniques. These empirical modelling techniques are characterised by total knowledge independence regarding the phenomenological nature of the process considered, being capable of mapping non-linear relationships between a group of input and output variables. Several algorithms have been used in the development of empirical models for adsorption, such as Artificial Neural Networks (ANN) [2], the Adaptative Neuro Fuzzy Inference System (ANFIS) [6], Support Vector Machine (SVM) [7] and Random Forest (RF) [8]. Ghaedi and Vafaei [9] highlight the applicability of ANN for modelling the adsorption process, being considered a well-established technique in this area. Moreover, authors point out that new techniques must be sequentially tested and validated in similar studies within the adsorption field, thus, demonstrating a suitable and less bureaucratic applicability of other models compared to ANNs, for instance.

With this in mind, RF has emerged as a promising alternative, due to its ability of identifying relationships between variables in a relatively small set of data, which is common in adsorption processes, and the tolerance to a great number of input variables [10,11].

As previously pointed out, ANNs have been widely used in the

* Corresponding author.
E-mail address: eduardo.farias.ufal@gmail.com (C.E. de Farias Silva).

https://doi.org/10.1016/j.jece.2020.103952
Received 14 February 2020; Received in revised form 31 March 2020; Accepted 13 April 2020
Available online 18 April 2020
2213-3437/ © 2020 Elsevier Ltd. All rights reserved.

empirical modelling of adsorption processes, given their capability of describing non-linear relationships between variables, combined with the capacity of generalising experimental data [2,12]. This technique was recently applied in the adsorption study of metals such as chromium IV from commercial activated carbon and NiO nanoparticles [13,14], with zinc II from palm kernel shell based activated carbon [15], lead II using rice husk carbon and polyamine-polyurea polymer modified with pyromellitic dianhydride [2,5], cadmium with natural walnut carbon [16], mercury using AC@Fe_3O_4-NH_2-COOH [17], nickel II with perlite [18] and copper with magnetic nanocomposites [19].

These networks have also been applied in the adsorption of dyes such as methylene blue with zinc sulfide nanoparticles with activated carbon (ZnS-NP-AC), activated spent tea, peanut sticks based activated carbon, natural walnut carbon and *Abelmoschus esculentus* seed [4,16,20–22], brilliant green on ZnS-NP-AC [23], malachite green with nanoscale zerovalent zinc [24], disperse blue with aluminium-based water treatment residuals [25], crystal violet using reduced-graphene-oxide-supported bimetallic Fe/Ni nanoparticles [26], amido black from polyaniline/SiO_2 nanocomposite [27] and dye mixtures and mesoporous activated carbons of low-cost agricultural bio-wastes [28,29].

On the other hand, RF is an ensemble machine learning technique which consists in the combination of response variables from a combination of simple estimators, in this case, decision trees. This technique is characterised by its bootstrapping (sampling with replacement) and randomised variable selection method, which reduces the correlation between decision trees, also decreasing the variance of the model [30]. As it is a new application in the description and prediction of adsorption processes, not many studies have addressed the use of RF in the empirical modelling of metals and dyes in effluents. The studies found involve the removal of metals such as lead II from Tamarisk based activated carbon [31], cadmium, nickel, arsenic, copper and zinc using different biochars [32], as well as methylene blue dye from Tamarisk based activated carbon [31], brilliant green dye from ZnS-NP-AC [23], malachite green using jackfruit seeds [8], Congo red from tin sulfide nanoparticles loaded on activated carbon [33], chrysoidine G by copper sulfide nanoparticles loaded on activated carbon [34] and bromophenol blue using *Astragalus bisulcatus* based activated carbon [30].

However, besides the reduced number of works published when compared with ANNs, only some studies have applied RF for modelling dye adsorption from residual biomass, such as orange bagasse, which is discussed in this work. Moreover, the publications found tend to present a limited number of experiments (set of data) or analyse only a few variables. In the case of Ghaedi et al. [30] and Heydari et al. [31], significant datasets are used – 241 and 360 experiments, respectively. Nevertheless, only three variables are considered. On the other hand, Kooh et al. [8] consider five variables for developing the model, though a smaller dataset is used, with 144 experiments. Therefore, the need of more studies regarding this subject is clearly evident, especially for taking advantage of the capacity of RF in dealing with a large dataset and comparing the results obtained in an attempt of verifying the efficiency of the method with a larger or smaller amount of data and/or variables. In the present article, 202 independent experiments, are used, being carried out in triplicate with the analysis of seven variables.

Regarding the application of machine learning techniques and data analysis, Python is one of the most well-established programming languages used, as a result of its great archive of scientific libraries. In addition, Python stands out due to its accessibility, as it is available free of charge, being an open source programming language and with simple syntax, similar to spoken language. From the Scikit-Learn programming library, Python offers high-level functions for application in several algorithms, among them ANN and RF [35,36].

Accordingly, this work was aimed at modelling the adsorption process of methylene blue in residual agricultural biomass (orange bagasse) using machine learning algorithm RF and comparing to ANN, already traditionally used to model adsorption process. The modelling procedure was carried out using Python as a programming language and a high number of input variables (7, including rotation and salinity; which are not often studied, such as solution pH and temperature; but of great importance in the adsorption process) and a large set of independent data (more than 200 experiments).

2. Materials and methods

2.1. Adsorbent preparation and variables analysed

The oranges used in this study were collected from Santana do Mandaú, in the Brazilian State of Alagoas, Brazil. The fruits were initially peeled and squashed, in order to obtain the juice, and the residues were cut and disinfected in sodium hypochlorite solution (100 $mg.L^{-1}$) for 15 min. After being rinsed, the material was placed in an air-circulation furnace at 50 °C under constant weight and the dry material was crushed with a Wylle (30 mesh) cutting mill and packed in hermetically plastic flasks at room temperature.

The cationic dye used in the adsorption trials was methylene blue (Sigma®), suitable as a model for cationic dyes removal from aqueous solutions [37]. A stock solution of 1000 $mg.L^{-1}$ of methylene blue was initially prepared and subsequently diluted to prepare 50, 100, 250, 350, 500, 750 and 850 $mg.L^{-1}$ solutions, used in the adsorption studies. Adsorption capacity and removal rate were calculated by Eqs. 1 and 2, with Q being the adsorption capacity ($mg.g^{-1}$), R the removal rate (%), Ci the initial concentration, C_f the concentration at (final) equilibrium ($mg.L^{-1}$), V the volume of the solution (L) and S the mass of adsorbent (g).

$$Q = \frac{(C_i - C_f)V}{S} \quad (1)$$

$$R\ (\%) = \frac{(C_i - C_f).\ 100}{C_i} \quad (2)$$

After the adsorption process, centrifugation (Hettich Universal 320/320R) was performed to remove the solid phase and the liquid phase analysed by UV–vis spectrophotometer (BEL SP2000UV) at 653 nm using a calibration curve of methylene blue as standard.

2.2. Batch adsorption experiments

The influence of initial solution pH (natural pH just dissolving the dye in water) on the adsorption capacity was subsequently evaluated to verify if, during the adsorption process over time, it changed significantly affecting the process with respect to the pH, with the biomass being in contact with solutions of $50-1{,}000\ mg.L^{-1}$ in their initial solution pH range for 60 min (0, 5, 10, 15, 20, 30, 40, 50 and 60 min, sample times) using 1% of biomass ($w.v^{-1}$), in a rotating incubator (TECNAL TE-424) at 30 °C and 100 rpm, to only evaluate the effect of initial dye concentration. The pH was measured using a glass-electrode digital pH meter (PHTEK PHS-3B), previously set with buffer solutions at a pH of 4 and 7.

The evaluation of the solution pH was made by changing the solution pH between 1.5–12.5 using NaOH 2 N and H_2SO_4 2.5 N with the experiments conducted at 100 rpm, 30 °C, 1% of biomass load and 60 min of contact time based on Silva et al. [1]. Then, considering the adequate pH range, the influence of biomass dosage was evaluated at a range of $10-30\ g.L^{-1}$ (1–3 % $w.v^{-1}$). The trials were carried out using a range of dye concentration between 50–1,000 $mg.L^{-1}$ at 100 rpm and 30 °C for 60 min.

The study of the rotation influence was performed for the rates of 60, 80 and 100 rpm, with biomass load of 1% ($w.v^{-1}$) at 30 °C, using a range of dye concentration between 50–1,000 $mg.L^{-1}$ in contact with the dye solution for 60 min. The effect of temperature was evaluated in the same conditions as the rotation study, except for the use of 100 rpm only, and the temperature changed between 30, 45 and 60 °C. In addition, salinity experiments were performed, which considered 30 °C

and 100 rpm, changing sodium chloride concentration (NaCl) between 1 − 10 g.L^{-1} (0.1–1 % w.v^{-1}). All experimental results and adsorbent characterisation are shown in detail in Silva et al. [3], totalising 202 independent experiments (different experimental conditions which were performed in triplicate, totalising 606experiments).

2.3. Artificial neural network modelling

ANNs are machine learning algorithms that attempt to replicate the human brain, which, in the form of supervised learning, learn cause and effect relationships based on past experiences. The architecture of these networks is based on a network of simple processing structure, the artificial neurons, divided into layers interconnected with each other. Each neuron is responsible for processing signals originating from the connections received, through a non-linear function, the activation function. Therefore, the ANN is able to map the relationships between input and output variables of complex phenomena [38,39].

In the present work, three models were built for separately predicting the final concentration of methylene blue (C_f in mg.L^{-1}), adsorption capacity (Q in mg.g^{-1}) and the adsorbate removal percentage ($R\%$). These three parameters are sequentially referred with legislation properties of discharge in environment, adsorbent efficiency to dye removal in solid-phase and process efficiency, thus, however they be correlated, mean different aspects. In order to determine these parameters, the following variables were used as input variables in each model: Temperature (°C), pH, adsorbent dosage (% w.w^{-1}), contact time (min), salinity (g.L^{-1} of NaCl), initial methylene blue concentration (Ci mg.L^{-1}) and rotation (rpm). Fig. 1 shows a schematic representation of the models developed.

In order to determine the best backpropagation neural network architeture to be used, different combinations were tested between the activation functions and the number of neurons in the hidden layer. Values between 8 and 22 were tested for the number of neurons in the output layer, according to the empirical formula described in Eq. 3 [12].

$$N_H = 2N_i + 1 \tag{3}$$

where N_H is the number of neurons in the hidden layer and N_i the number of neurons in the input layer, in terms of the number of input variables. In addition, four activation functions were tested: the identity, hyperbolic tangent, logistic and ReLu functions – showed in more in the supplementary material [40].

The logistic function was used as the activation function in the output layer and Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) solver as a training algorithm, with the latter being part of the quasi-Newton family methods. The adsorption process was modelled using Python 3.6.8 as a programming language, with the Scikit-Learn programming library, version 0.21.3, using the *MLPRegressor* function [35].

The normalisation of input and output data is necessary in order to avoid the generation of weights with very different magnitudes, thus, increasing the efficiency of the training process [30,41]. Therefore, the data used for training, validation and for testing the networks were normalised in the range between 0 and 1, according to Eq. 4.

$$y = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4}$$

Where y is the normalized value of x_i, with x_{min} and x_{max} being the maximum and minimum value of a certain variable, respectively. For the models training and testing steps, a set of 202 experimental data was randomly divided, with 80 % being used in the training step (161) and 20 % for testing [41].

The testing or validation step aimed at assessing the generalizability of the developed models, based on the comparison between experimental values (y_o) and predicted values by the model (y_p). This technique is also called Holdout validation, as it is performed from a set of data not included in the model training process, and it is also used to

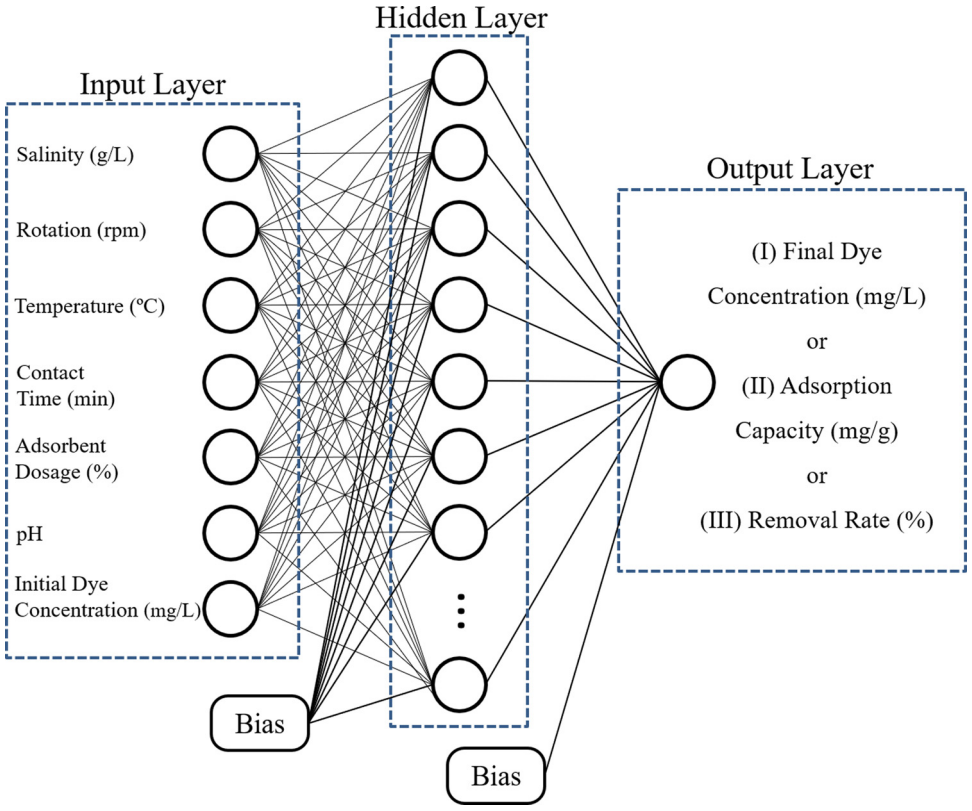

Fig. 1. Schematic representation of the three models developed using ANN for predicting: (I) Final concentration (C_f), (II) Adsorption capacity (Q) and (III) Removal percentage ($R\%$), as a function of the input variables.

prevent model overfitting [40]. In addition, the models training and validation procedures were repeated extensively, in order to reduce the effect of randomness. To quantify the models forecasting capacity, Mean Square Error (MSE) and Coefficient of Determination (R^2), represented by Eqs. 3 and 4, were used.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{p,i} - y_{o,i})^2}{\sum_{i=1}^{N}(y_{p,i} - y_m)^2} \quad (5)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(|y_{p,i} - y_{o,i}|)^2 \quad (6)$$

2.4. Random forest modelling

RF is an ensemble machine learning algorithm, which consists in the association of various decision trees, aimed at improving the predictive power and reduce the variance of the model [31]. The result is reached through the randomised generation of n decision trees with bootstrapping (resampling with replacement), as well as by the aggregation of the results of each of the trees through the average of the results. In addition, the RF algorithm randomly selects a subset of variables from the process that is used in the development of each tree, reducing the correlation between them and decreasing the generalisation error [42].

Analogous to the procedure adopted for ANNs, three models were built for separately predicting C_f, $Q$and $R\%$, using the same input variables. Fig. 2 shows a schematic representation of the models developed.

Similar to ANNs, RF also has some hyperparameters that can influence the quality of the model. Aimed at finding a better architecture for each of the models, different values were tested for the number of decision trees (n) and for the maximum number of variables in each tree (N_{var}). Regarding the number of trees, values between 100 and 2000 trees were tested, while the maximum number of variables tested ranged between 1 and 7 [32].

For the implementation of the RF method, Python 3.6.8 was used as a programming language, as well as the Scikit-Learn programming library, 0.21.3. The function used was the *RandomForestRegressor*, with the remaining hyperparameters in their standard configuration [35].

Table 1
Optimal configurations for ANN models and their respective R^2 and MSE.

Variables Studied	Activation Function	Neurons	Training		Testing	
			R^2	MSE	R^2	MSE
C_f	Tanh	11	0.9801	0.0009	0.9734	0.0016
Q	Tanh	11	0.9929	0.0005	0.9919	0.0007
$R\%$	Tanh	12	0.9419	0.0016	0.9257	0.0019

The normalisation applied to the data for developing the RF models was identical to that used in the treatment of data when developing the ANN models, considering Eq. 4 and maintaining the same proportion of data used in the training and testing steps – 80 % and 20 %, respectively. Likewise, the same metrics and validation procedure were used for assessing the generalisation capability of the models (R^2 and MSE), according to Eqs. 5 and 6.

3. Results and discussion

In this work, machine learning algorithms were utilised for modelling the adsorption process of methylene blue in orange bagasse. A dataset of 202 experiments (different experimental conditions performed in triplicate, totalising 606 experiments), were applied for developing the models to predict the C_f, Q and $R\%$, using ANN and RF. Finally, the RF models were compared with ANN models, which is the traditional approach for machine learning modelling of adsorption process.

3.1. Analysis of the ANN models

The influence of topology is a central point in the development of ANN models [28]. In this work, backpropagation ANNs with three layers were used for developing the models. According to Liu et al. [12], the architecture of the networks with three layers is suitable for describing the adsorption process, being commonly used in similar works published in the literature [14,20,24–26].

Table 1 shows the best topologies obtained for three predictive

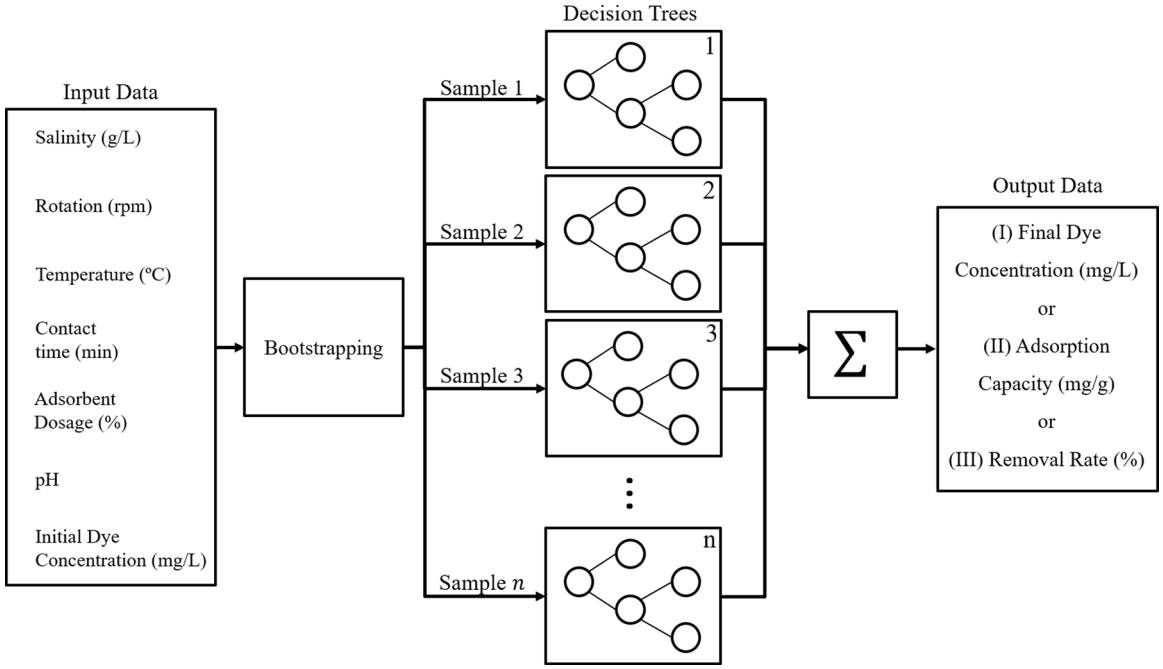

Fig. 2. Schematic diagram of the three models developed using RF for predicting: (I) Final concentration (C_f), (II) Adsorption capacity (Q) and (III) Removal percentage ($R\%$), as a function of input variables.

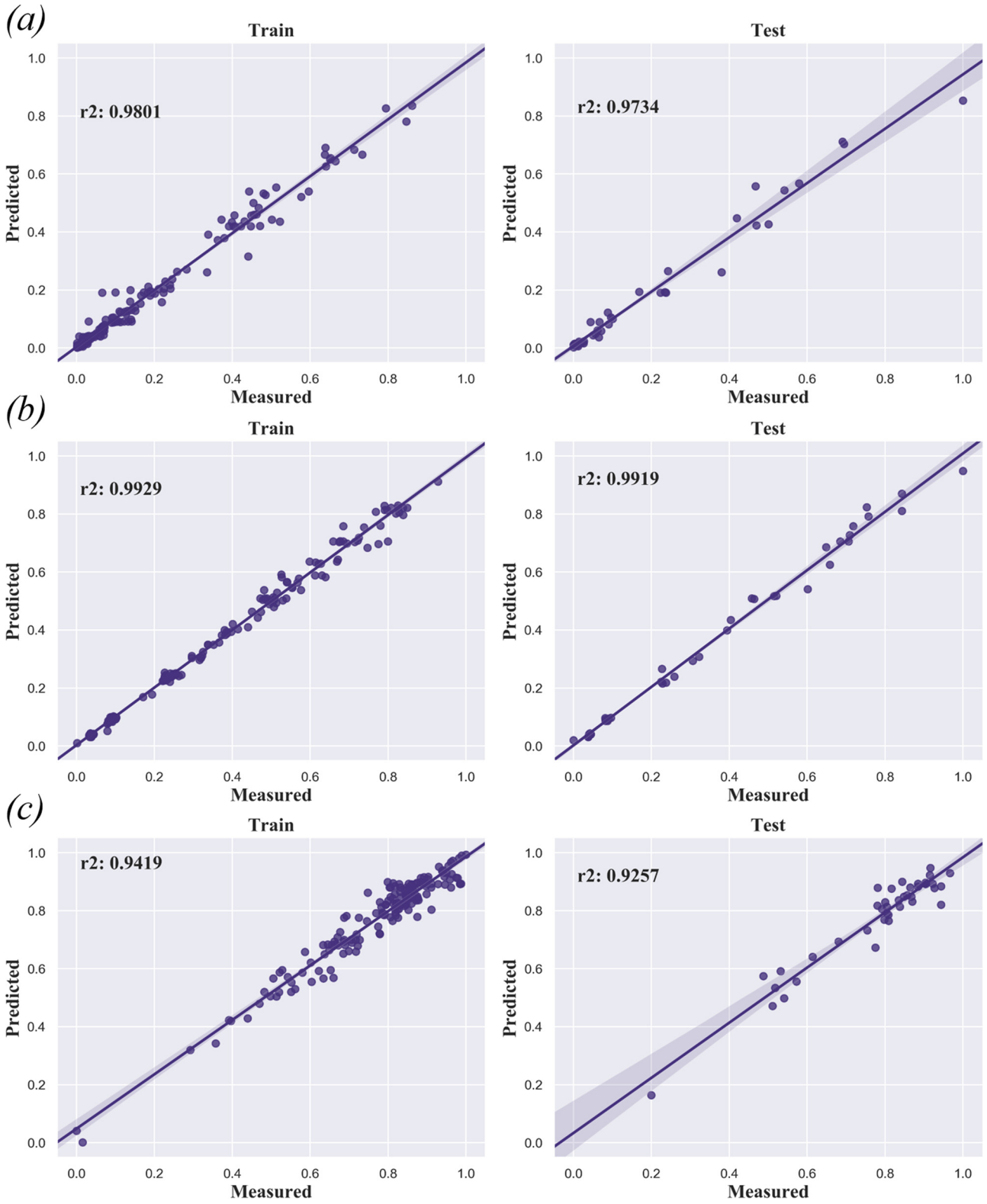

Fig. 3. Comparison between the experimental and predicted values for *(a)* final concentration of methylene blue (C_f); *(b)* adsorption capacity (Q); *(c)* removal percentage ($R\%$), using ANN. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

models, in terms of the activation function and number of neurons in the hidden layers, as well as model predictive capacity in terms of the R^2 and MSE.

Among the four activation functions tested, the hyperbolic tangent function (Tanh) was the most suitable for describing the adsorption process, taking into account the estimation of three variables of interest. Regarding the number of neurons, very similar values were also obtained for three models, which resulted in almost identical topologies. It is important to point out that the number of neurons was tested in the range between 8 and 22. Therefore, the results show that an increase in the number of neurons does not necessarily intensify the predictive capacity of the model. Besides, considering the results of R^2

Table 2
Details of different works found in literature regarding the modelling of methylene blue adsorption process using artificial neural networks.

	Adsorbent	Input Variables	Predicted Variable	Dataset Size	Network Topology	Activation Function
Asfram et al. [4]	ZnS-NPs-AC	pH, m (g), C_o ($mg.L^{-1}$), t_c (min)	*R%*	30	4:8:1	Logistic
Babaei et al. [20]	Activated Spent Tea (AST)	C_i ($mg.L^{-1}$), pH, D ($g.L^{-1}$), T (K), t_c (min)	*R%*	81	5:10:1	Linear
Dil et al. [28]	ZnO-NR-AC	C_{i1}, C_{i2}, C_{i3}, C_{i4} ($mg.L^{-1}$), m (g), t_c (min)	*R%*	54	6:6:1	Tanh
Karimi and Ghaedi [21]	Activated Carbon	pH, R_t (rpm), D ($g.L^{-1}$), C_i ($g.L^{-1}$), t_c (min)	*R%*	108	5:5:1	Tanh
Zeinali et al. [44]	Graphite oxide	pH, m (mg), and t_c (min)	C_f	144	3:10:1	Tanh

Table 3
Optimal configurations for RF models and their respective values of R^2 and MSE.

Variables Studied	*n*	N_{var}	Training		Testing	
			R^2	MSE	R^2	MSE
C_f	200	6	0.9782	0.0011	0.9739	0.0012
Q	100	6	0.9934	0.0005	0.9932	0.0005
R%	100	6	0.9532	0.0013	0.9318	0.0015

and MSE, it can be observed that the models chosen had no indication of overfitting, a phenomenon noticed when the results of training are considerably superior to those of the testing step.

In general, it can be argued that the models obtained for the three variables studies exhibited a good predictive capacity, despite the relatively significant difference between the values obtained for removal and the remaining variables. Accordingly, it can be said that the backpropagation ANN with three layers was an efficient technique for building the models regarding the adsorption of methylene blue as an (adsorbent), within the experimental range used. Fig. 3 shows a comparison between the target and predicted values for the three models, in the training and testing steps.

Other works in the literature discussed the modelling of the adsorption process of methylene blue using ANN, with different adsorbents and input variables. Table 2 presents some of these works, showing the adsorbent used, as well as the input variables, the size of the dataset applied for building the model and the best topology found for the network.

By observing the results obtained in the literature, it can be noted that the network topology used by the other authors is similar to that used in the present work. In addition, it can be observed that the hyperbolic tangent function (Tanh) is commonly used for modelling the removal of methylene blue by adsorption. Nevertheless, it is important to point out that the choice of activation function and of the number of neurons in the hidden layer depends on the set of experimental data used, as well as on the variables analysed, which varies in each case.

Therefore, the results found for ANNs proved to be adequate and suitable for a comparison with the test carried out with the RF method, as subsequently shown.

3.2. Analysis of the RF models

The results regarding the three predictive RF models are presented below, in a similar way to the results presented for the ANN models. Table 3 shows the best configuration obtained during model development in terms of the best set of tuning parameters (*n* and N_{var}), regarding each variable as well as their respective values of R^2 and MSE.

According to Breiman [43], the increase in the *n* does not lead to overfitting in RF models, different to the behaviour shown in terms of the number of neurons in the hidden layers, in the case of ANNs. Nonetheless, the predictive quality of the model reaches its limit as the number of trees increases. This behaviour could be observed in the three models built, in which the best obtained values for *n* were 200, 100 and 100 trees, for C_f, *Q* and *R*, respectively. Although values in the range between 100 and 2000 decision trees were tested, there was no significant difference between the lower and higher values. For the modelling of the adsorption process discussed, it can be argued that the best values obtained for the number of decision trees were capable of describing the process at a faster rate in terms of algorithm execution, besides being in accordance with most works found in the literature [23,30–34].

N_{var} is a parameter responsible for limiting the size of the subset of data available in the construction of the trees, thus, varying between 1 and N_i, where N_i is the number of input variables in the model. According to Ahmad et al. [11], the highest values of this parameter are expected to lead to an increase in the predictive quality of the RF method, as each tree would have a greater number of variables available. In the present work, the tests carried out showed that considerably low values of N_{var} hampered the predictability of the model, while values close to the maximum number of variables, 7, presented better results. Zhu et al. [32] obtained similar values in their study regarding the adsorption of metals in biochars. Of the 14 variables used in the modelling, the authors indicated that a maximum of 13 variables resulted in a more efficient model.

According to the results obtained regarding the R^2 and MSE, it can be argued that the RF method presented a satisfactory predictive capacity for the three models built. Therefore, the method can be considered a viable alternative for modelling the adsorption process of methylene blue by orange bagasse, within the experimental range considered. Fig. 4 shows a comparison between the target and predicted values for the three models, for the training and testing steps.

As it is a method not yet widely used for modelling adsorption processes, especially when compared to ANNs, not many works were found in the literature regarding the RF method. Among the recent works found, it is important to highlight the study carried out by Heydari et al. [31], who used RF for modelling the adsorption of methylene blue and lead (II) using activated carbon. The input variables used were initial concentration – C_i ($mg.L^{-1}$), amount of adsorbent (g) and contact time – t_c (min), while the output was the removal – *R(%)*. From the set of 360 experimental data, the authors stated that the best configuration for the RF model in terms of tuning hyperparameters consisted of 100 decision trees and a maximum of 2 variables in each tree. Taking into account that the authors used only three input variables for building the model, it can be said that the results obtained by Heydari et al. [31] are in accordance with those obtained in the present work. Finally, the authors verified the efficiency of RF for modelling blue methylene and lead (II), based on the agreement between experimental and predicted data.

3.3. Comparison between RF and ANN

When analysing the results obtained, as pointed out in Tables 1 and 3, it is possible to point out that both machine learning methods were suitable for predicting the three variables studied, the C_f, *Q* and *R%*. Regarding C_f, the models built with RF (0.9739) and ANN (0.9734) were not significantly different in terms of their predictive capacity, with both models considered efficient. In terms of, the RF model (0.9932) was only significantly different from the ANN model (0.9919). Among the three variables studied, *R%* exhibited the greatest difference between the models studied. In this case, the RF model (0.9318) showed to be slightly better than the ANN model (0.9257), although

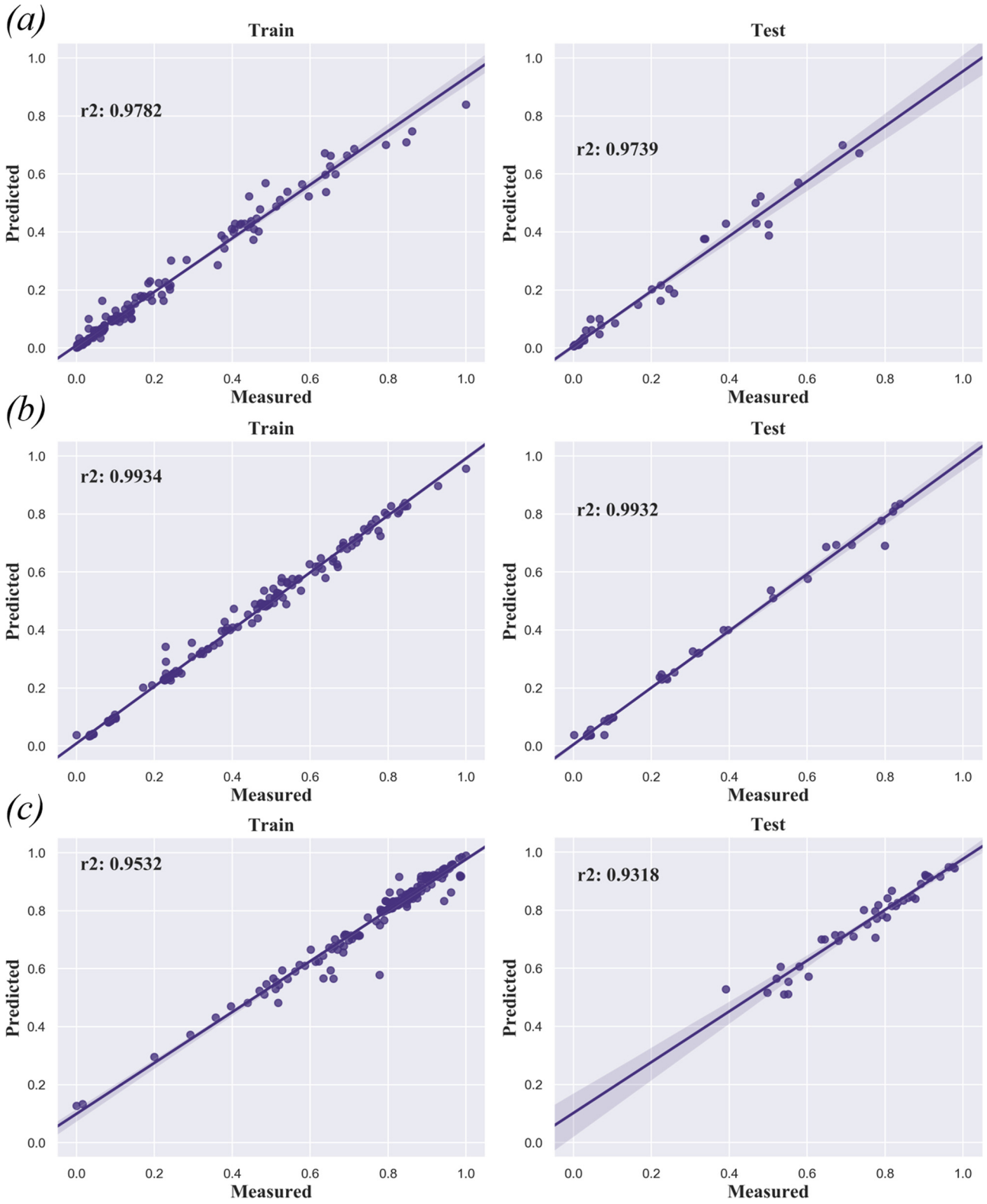

Fig. 4. Comparison between the experimental and predicted values for *(a)* final concentration of methylene blue (C_f); *(b)* adsorption capacity (Q); *(c)* removal percentage ($R\%$), using RF. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

with a minor difference.

Despite both algorithms presenting similar results regarding R^2 and MSE, the RF method stands out when compared with ANN. Fig. 5 shows a comparison between experimental and predicted data regarding the three studied variables. As it can be seen in Fig. 5A–B, for final concentration (C_f), the RF-based model tends to better capture the variation in concentration values, in contrast to the RNA-based model that presents smoother curve for predicted values, especially in lower concentration. The same behavior can be observed in Fig. 5C–D for Q and 5E-F for $R\%$. The experimental data and predicted values for each variable and model can be found in the supplementary material.

Another point to be highlighted was the ease in implementing

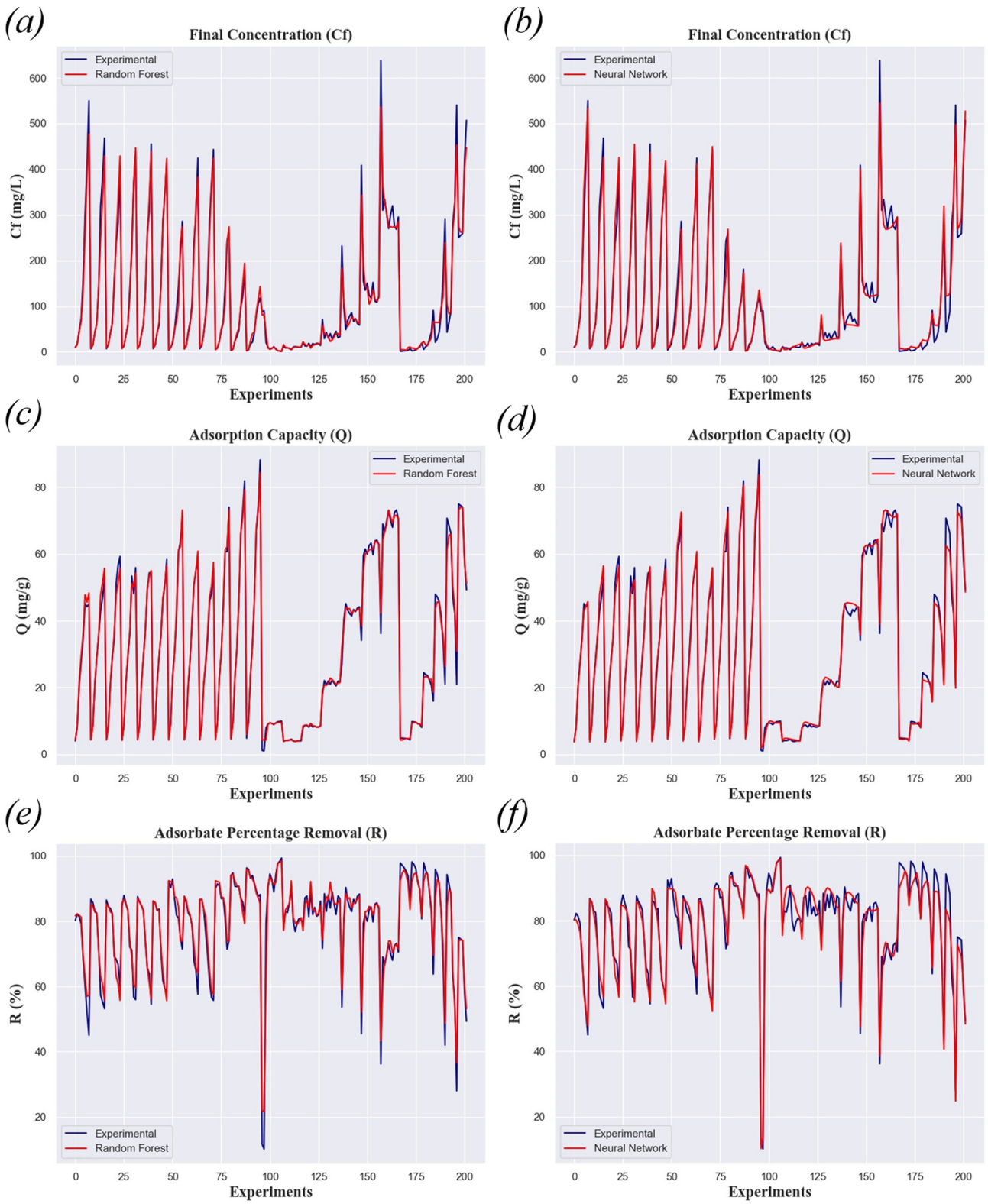

Fig. 5. Comparison between experimental and predicted data for Final concentration of methylene blue (C_f) and (a) RF (b) ANN; adsorption capacity (Q) and (c) RF (d) ANN; removal percentage ($R\%$) and (e) RF (f) ANN. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

machine learning algorithms in Python. Like other commercial software, Python has many machine learning libraries with the advantage of being free and open source. In addition, Python is known for its user-friendly syntax, which combined with its high-level functions, results in cleaner code and greater readability. The models development code can be found in the supplementary material.

4. Conclusions

The present work compared two machine learning algorithms to

model the adsorption of methylene blue in orange bagasse – RF and the traditional approach, ANN. Both methods were submitted to the same set of 202 independent experiments in triplicate, normalised within the range of [0,1] and at a proportion of 80 %–20 % for training and testing steps, respectively.

Based on the acquired results, it is possible to say that the RF-based models showed slightly better performance than the ANN-based models, presenting greater capacity to capture small variations in the studied variables. Nevertheless, both algorithms presented high performance models that could be used to model the studied process. It is also important to point out that the predictive capacity of machine learning algorithms depends on the process that is being modelled, as well as on the dataset available.

Author credit statement

A.P.M.R.S. and F.O.C. (Modelling implementation, writing, discussion and original draft), A.H.S.G. and A.K.S.A. (Experimental Results, article review and editing) and C.E.F.S (writing, discussion, original draft, experimental results, article review and editing).

Declaration of Competing Interest

Authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jece.2020.103952.

References

[1] C.E.F. Silva, A.H.S. Gonçalves, A.K.S. Abud, Treatment of textile industry effluents using orange waste: a proposal to reduce color and chemical oxygen demand, Water Sci. Technol. 74 (4) (2016) 994–1004.

[2] A. Kiraz, O. Canpolat, E.F. Erkan, C. Özer, Artificial neural networks modeling for the prediction of Pb (II) adsorption, Int. J. Environ. Sci. Technol. 16 (2018) 5079–5086, https://doi.org/10.1007/s13762-018-1798-4.

[3] C.E.F. Silva, B.M.V. Gama, A.H.S. Gonçalves, J.A. Medeiros, A.K.S. Abud, Basic-dye adsorption in albedo residue: effect of pH, contact time, temperature, dye concentration, biomass dosage, rotation and ionic strength, J. King Saud Univ. - Eng. Sci. (2019) In Press.

[4] A. Asfaram, M. Ghaedi, M.H.A. Azqhandi, A. Goudarzi, M. Dastkhoon, Statistical experimental design, least squares-support vector machine (LS-SVM) and artificial neural network (ANN) methods for modeling the facilitated adsorption of methylene blue dye, RSC Adv. 6 (46) (2016) 40502–40516, https://doi.org/10.1039/c6ra01874b.

[5] T. Khan, M.R.U. Mustafa, M.H. Isa, T.S.B.A. Manan, J.-W. Lim, Y.-C. Ho, N.Z. Yusof, Artificial Neural Network (ANN) for modelling adsorption of lead (Pb (II)) from aqueous solution, Water Air Soil Pollut. 228 (2017) 426, https://doi.org/10.1007/s11270-017-3613-0.

[6] K. Aghajani, H. Tayebi, Adaptive Neuro-Fuzzy Inference system analysis on adsorption studies of Reactive Red 198 from aqueous solution by SBA-15 / CTAB composite, Spectrochim. Acta A. Mol. Biomol. Spectrosc. 171 (2017) 439–448, https://doi.org/10.1016/j.saa.2016.08.025.

[7] M. Li, D. Wei, T. Liu, Y. Liu, L. Yan, Q. Wei, B. Du, W. Xu, EDTA functionalized magnetic biochar for Pb (II) removal: adsorption performance, mechanism and SVM model prediction, Sep. Purif. Technol. 227 (2019) 115696, , https://doi.org/10.1016/j.seppur.2019.115696.

[8] M.R.R. Kooh, M.K. Dahri, L.B.L. Lim, Jackfruit seed as low-cost adsorbent for removal of malachite green: artificial neural network and random forest approaches, Environ. Earth Sci. 77 (2018) 432, https://doi.org/10.1007/s12665-018-7618-9.

[9] A.M. Ghaedi, A. Vafaei, Applications of artificial neural networks for adsorption removal of dyes from aqueous solution: a review, Adv. Colloid Interface Sci. 245 (2017) 20–39, https://doi.org/10.1016/j.cis.2017.04.015.

[10] Y. Qi, Random forest for bioinformatics, Ensemble Machine Learning, Springer, 2012, pp. 307–323.

[11] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89, https://doi.org/10.1016/j.enbuild.2017.04.038.

[12] Z. Liu, F. Liang, Y. Liu, Artificial neural network modeling of biosorption process using agricultural wastes in a rotating packed bed, Appl. Therm. Eng. 140 (2018) 95–101, https://doi.org/10.1016/j.applthermaleng.2018.05.029.

[13] K. Anupam, S. Dutta, C. Bhattacharjee, S. Datta, Artificial neural network modelling for removal of chromium (VI) from wastewater using physisorption onto powdered activated carbon, Desalin. Water Treat. 57 (8) (2016) 3632–3641, https://doi.org/10.1080/19443994.2014.987172.

[14] S.K. Ashan, M.A. Behnajady, N. Ziaeifar, R. Khalilnezhad, Artificial neural network modelling of Cr(VI) surface adsorption with NiO nanoparticles using the results obtained from optimization of response surface methodology, Neural Comput. Appl. 29 (10) (2017) 969–979, https://doi.org/10.1007/s00521-017-3172-8.

[15] R.R. Karri, J.N. Sahu, Modeling and optimization by particle swarm embedded neural network for adsorption of zinc (II) by palm kernel shell based activated carbon from aqueous environment, J. Environ. Manage. 206 (2018) 178–191, https://doi.org/10.1016/j.jenvman.2017.10.026.

[16] H. Mazaheri, M. Ghaedi, M.H.A. Azqhandi, A. Asfaram, Application of machine/statistical learning, artificial intelligence and statistical experimental design for the modeling and optimization of methylene blue and Cd (ii) removal from a binary aqueous solution by natural walnut carbon, J. Chem. Soc. Faraday Trans. 19 (18) (2017) 11299–11317, https://doi.org/10.1039/C6CP08437K.

[17] M. Pazouki, M. Zabihi, J. Shayegan, M.H. Fatehi, Mercury Ion Adsorption on AC@Fe3O4-NH2-COOH from Saline Solutions: Experimental Studies and Artificial Neural Network Modeling 35 (2018), pp. 671–683, https://doi.org/10.1007/s11814-017-0293-9 3.

[18] S.M. Turp, Prediction of adsorption ef fi ciencies of Ni (II) in aqueous solutions with perlite via artificial neural networks, Arch. Environ. Prot. 43 (4) (2017) 26–32, https://doi.org/10.1515/aep-2017-0034.

[19] M. Fan, J. Hu, R. Cao, K. Xiong, X. Wei, Modeling and prediction of copper removal from aqueous solutions by nZVI/rGO magnetic nanocomposites using ANN-GA and ANN-PSO, Sci. Rep. 7 (2017) 18040, https://doi.org/10.1038/s41598-017-18223-y.

[20] A.A. Babaei, A. Khataee, E. Ahmadpour, M. Sheydaei, B. Kakavandi, Z. Alaee, Optimization of cationic dye adsorption on activated spent tea: equilibrium, kinetics, thermodynamic and artificial neural network modelling, Korean J. Chem. Eng. 33 (4) (2016) 1352–1361, https://doi.org/10.1007/s11814-014-0334-6.

[21] H. Karimi, M. Ghaedi, Application of artificial neural network and genetic algorithm to modeling and optimization of removal of methylene blue using activated carbon, J. Ind. Eng. Chem. 20 (4) (2014) 2471–2476, https://doi.org/10.1016/j.jiec.2013.10.028.

[22] A.K. Nayak, A. Pal, Green and efficient biosorptive removal of methylene blue by Abelmoschus esculentus seed: Process optimization and multi-variate modelling, J. Environ. Manage. 200 (2017) 145–159, https://doi.org/10.1016/j.jenvman.2017.05.045.

[23] M.H. Ahmadi-Azqhandi, M. Ghaedi, F. Yousefi, M. Jamshidi, Application of random forest, radial basis function neural networks and central composite design for modeling and / or optimization of the ultrasonic assisted adsorption of brilliant green on ZnS-NP-AC, J. Colloid Interface Sci. 505 (2017) 278–292, https://doi.org/10.1016/j.jcis.2017.05.098.

[24] W. Ruan, X. Shi, J. Hu, Y. Hou, M. Fan, R. Cao, X. Wei, Modeling of malachite green removal from aqueous solutions by nanoscale zerovalent zinc using artificial neural network, Appl. Sci. 8 (1) (2018) 3, https://doi.org/10.3390/app8010003.

[25] M.R. Gadekar, M.M. Ahammed, Modelling dye removal by adsorption onto water treatment residuals using combined response surface methodology-artificial neural network approach, J. Environ. Manage. 231 (2019) 241–248, https://doi.org/10.1016/j.jenvman.2018.10.017.

[26] W. Ruan, J. Hu, Qi J, Y. Hou, R. Cao, X. We, Removal of crystal violet by using reduced-graphene-oxide-supported bimetallic Fe/Ni nanoparticles (rGO/Fe/Ni): application of artificial intelligence modeling for the optimization process, Materials 11 (5) (2018) 865, https://doi.org/10.3390/ma11050865.

[27] M. Tanzifi, M.T. Yaroki, A.D. Kiadehi, S.H. Hosseini, M. Olazar, A.K. Bharti, S. Agawal, V.K. Gupta, A. Kazemi, Adsorption of Amido Black 10B from aqueous solution using polyaniline/SiO2 nanocomposite: experimental investigation and artificial neural network modelling, J. Colloid Interface Sci. 510 (2018) 246–261, https://doi.org/10.1016/j.jcis.2017.09.055.

[28] E.A. Dil, M. Ghaedi, A. Asfaram, F. Mehrabi, A.A. Bazrafshan, A.M. Ghaedi, Trace determination of safranin O dye using ultrasound assisted dispersive solid-phase micro extraction: artificial neural network-genetic algorithm and response surface methodology, Ultrason. Sonochem. 33 (2016) 129–140, https://doi.org/10.1016/j.ultsonch.2016.04.031.

[29] N.M. Mahmoodi, M. Taghizadeh, A. Taghizadeh, Mesoporous activated carbons of low-cost agricultural bio-wastes with high adsorption capacity: preparation and artificial neural network modeling of dye removal from single and multicomponent (binary and ternary) systems, J. Mol. Liq. 269 (2018) 217–228, https://doi.org/10.1016/j.molliq.2018.07.108.

[30] M. Ghaedi, A.M. Ghaedi, E. Negintaji, A. Ansari, A. Vafaei, A.M. Rajabi, Random forest model for removal of bromophenol blue using activated carbon obtained from Astragalus bisulcatus tree, J. Ind. Eng. Chem. 20 (4) (2014) 1793–1803, https://doi.org/10.1016/j.jiec.2013.08.033.

[31] F. Heydari, M. Ghaedi, A. Ansari, A.M. Ghaedi, Random forest model for removal of methylene blue and lead (II) ion using activated carbon obtained from Tamarisk, Desalin. Water Treat. 57 (41) (2015) 19273–19291, https://doi.org/10.1080/19443994.2015.1095124.

[32] X. Zhu, X. Wang, Y.S. Ok, The application of machine learning methods for prediction of metal sorption onto biochars, J. Hazard. Mater. 378 (2019) 120727, , https://doi.org/10.1016/j.jhazmat.2019.06.004.

[33] N. Dehghanian, M. Ghaedi, A. Ansari, A. Ghaedi, A. Vafaei, M. Asif, S. Agarwal, I. Tyagi, V.K. Gupta, A random forest approach for predicting the removal of Congo red from aqueous solutions by adsorption onto tin sulfide nanoparticles loaded on activated carbon, Desalin. Water Treat. 57 (20) (2016) 9272–9285, https://doi.org/10.1080/19443994.2015.1027964.

[34] A.R. Bagheri, M. Ghaedi, S. Hajati, A.M. Ghaedi, A. Goudarzi, A. Asfaram, Random forest model for the ultrasonic-assisted removal of chrysoidine G by copper sulfide nanoparticles loaded on activated carbon; response surface methodology approach, RSC Adv. 5 (73) (2015) 59335–59343, https://doi.org/10.1039/C5RA08399K.
[35] F. Pedregosa, G. Varoquoux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duches, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[36] A. Bogdanchikov, M. Zhaparov, R. Suliyev, Python to learn programming, J. Phys. Conf. Ser. 423 (1) (2013), https://doi.org/10.1088/1742-6596/423/1/012027.
[37] T. Liu, Y. Li, Q. Du, J. Sun, Y. Jiao, G. Yanga, Z. Wanga, Y. Xia, W. Zhang, K. Wang, H. Zhu, D. Wu, Adsorption of methylene blue from aqueous solution by graphene, Colloids Surf. B: Biointerface 90 (2012) 197–203.
[38] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Pearson Education, Inc, 2019.
[39] D.I. Mendoza-Castillo, H.E. Reynel-Ávila, F.J. Sánchez-Ruiz, R. Trejo-Valencia, J.E. Jaime-Leal, A. Bonilla-Petriciolet, Insights and pitfalls of artificial neural network modeling of competitive multi-metallic adsorption data, J. Mol. Liq. 251 (2018) 15–27, https://doi.org/10.1016/j.molliq.2017.12.030.
[40] A. Géron, Hands-on Machine Learning With Scikit-learn, Keras & TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, 2nd ed., O'Reilly Media, Inc., 2019.
[41] R.R. Karri, M. Tanzi, M. Tavakkoli, J.N. Sahu, Optimization and modeling of methyl orange adsorption onto polyaniline nano-adsorbent through response surface methodology and differential evolution embedded neural network, J. Environ. Manage. 223 (2018) 517–529, https://doi.org/10.1016/j.jenvman.2018.06.027.
[42] V. Rodriguez-Galiano, M. Sanchez-castillo, M. Chica-olmo, M. Chica-rivas, Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines, Ore Geol. Rev. 71 (2015) 804–818, https://doi.org/10.1016/j.oregeorev.2015.01.001.
[43] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[44] N. Zeinali, M. Ghaedi, G. Shafie, Competitive adsorption of methylene blue and brilliant green onto graphite oxide nano particle following: Derivative spectrophotometric and principal component-artificial neural network model methods for their simultaneous determination, J. Ind. Eng. Chem. 20 (2014) 3550–3558, https://doi.org/10.1016/j.jiec.2013.12.048.