nanomaterials

MDPI

Article

A Study on Machine Learning Methods' Application for Dye Adsorption Prediction onto Agricultural Waste Activated Carbon

Seyedehmaryam Moosavi [1,*], Otilia Manta [2,3], Yaser A. El-Badry [4], Enas E. Hussein [5], Zeinhom M. El-Bahy [6], Noor fariza Binti Mohd Fawzi [7], Jaunius Urbonavičius [1] and Seyed Mohammad Hossein Moosavi [8]

[1] Department of Chemistry and Bioengineering, Vilnius Gediminas Technical University, 10223 Vilnius, Lithuania; jaunius.urbonavicius@vilniustech.lt
[2] Romanian Academy, Center for Financial and Monetary Research "Victor Slavescu", 050711 Bucharest, Romania; otilia.manta@rgic.ro
[3] Research Department, Romanian-American University, 012101 Bucharest, Romania
[4] Chemistry Department, Faculty of Science, Taif University, Khurma, P.O. Box 11099, Taif 21944, Saudi Arabia; y.elbadry@tu.edu.sa
[5] National Water Research Centre, P.O. Box 74, Shubra EI-Kheima 13411, Egypt; enas_el-sayed@nwrc.gov.eg
[6] Chemistry Department, Faculty of Science, Al-Azhar University, Cairo 11884, Egypt; zeinelbahy@azhar.edu.eg
[7] Nanotechnology & Catalysis Research Centre (NANOCAT), Institute for Advanced Studies (IAS), University for Malaya (UM), Kuala Lumpur 50603, Malaysia; farizafawzi@um.edu.my
[8] Faculty of Engineering, Centre for Transportation Research (CTR), University of Malaya (UM), Kuala Lumpur 50603, Malaysia; mh.moosavi65@gmail.com
* Correspondence: m.moosavi1987@gmail.com

check for updates

Citation: Moosavi, S.; Manta, O.; El-Badry, Y.A.; Hussein, E.E.; El-Bahy, Z.M.; Mohd Fawzi, N.f.B.; Urbonavičius, J.; Moosavi, S.M.H. A Study on Machine Learning Methods' Application for Dye Adsorption Prediction onto Agricultural Waste Activated Carbon. *Nanomaterials* **2021**, *11*, 2734. https://doi.org/10.3390/nano11102734

Academic Editors: Gabriella Salviulo and George Z. Kyzas

Received: 30 August 2021
Accepted: 11 October 2021
Published: 15 October 2021

Abstract: The adsorption of dyes using 39 adsorbents (16 kinds of agro-wastes) were modeled using random forest (RF), decision tree (DT), and gradient boosting (GB) models based on 350 sets of adsorption experimental data. In addition, the correlation between variables and their importance was applied. After comprehensive feature selection analysis, five important variables were selected from nine variables. The RF with the highest accuracy (R^2 = 0.9) was selected as the best model for prediction of adsorption capacity of agro-waste using the five selected variables. The results suggested that agro-waste characteristics (pore volume, surface area, agro-waste pH, and particle size) accounted for 50.7% contribution for adsorption efficiency. The pore volume and surface area are the most important influencing variables among the agro-waste characteristics, while the role of particle size was inconspicuous. The accurate ability of the developed models' prediction could significantly reduce experimental screening efforts, such as predicting the dye removal efficiency of agro-waste activated carbon according to agro-waste characteristics. The relative importance of variables could provide a right direction for better treatments of dyes in the real wastewater.

Keywords: machine learning; wastewater treatment; dye adsorption; agricultural waste; activated carbon

1. Introduction

Approximately 10,000 dyes are commercially available, and annually, about 1.6 million tons of dyes are produced for industrial use [1], of which 10–15% of these dyes are disposed of as wastewater [2]. This pollution is caused by the use of dyes in the clothing, paper, dyeing, and plastics industries. Because the dyes are very stable and solvable in water, failed dye treatment and disposal of these wastes into receiving waters causes huge damages to the environment: affecting photosynthetic activity [3]; being toxic to aquatic life due to the presence of metals, chlorides, etc. [4]; and inherent toxicity, mutagenicity, and carcinogenicity [5]. In addition, overexposure to dyes has resulted in potentially life threatening complications such as skin harms, respiratory problems, and the probability of human carcinoma [6].

Most of these synthetic dyes are chemically and thermally stable, non-biodegradable, and quite toxic [7]. Due to toxicity of dyes, it is necessary to remove them from wastewater before them discharge to the natural environment. Adsorption is generally considered to be the most prominent approach due to its effectiveness [8], economy, and simplicity for quickly lowering the concentration of dissolved dyes in an effluent, as it does not require a pretreatment step before its application [9]. This method is very dependent upon the type of the adsorbent used (e.g., activated carbon (AC), biomass, polymer, nanomaterial, etc.) [10]. The cost of adsorbent production is not the only factor involved in developing an excellent adsorbent. The adsorption performance, regeneration ability, and adsorbent separation are other important features of an effective adsorbent [11].

Among the large number of adsorbents, activated carbons (ACs) exhibited advantages over other adsorbents for their high surface area, microporous character, chemical nature of their surface, and high adsorption capacity when used on wastewater with different dye molecules [12]. Most of the activated carbon materials used for adsorption research come from fruit peel [13], rubber tires [14], textile sludge [15], crab shell [16], and egg shell [17]. Recently, agricultural wastes have received considerable attention due to their abundant surface functional groups, porous structures, additional inorganic minerals, and high surface area. Many researchers study the dyes' adsorption on various agro-waste materials such as corn cobs [18], oil cake [19], rice husk [20], sugarcane bagasse [21], gram husk [22], sawdust [23], pine cone [24], tobacco residue [25], white sugar [26], Astragalus bisulcatus tree [27], tea residue [28], vinasse [29], groundnut shell [30], and so on. This class of activated carbons have a valuable potential for the wastewater dye removal. Agro-waste ability in the treatment of wastewater containing dyes such as MB dyes could reach the maximum value of 2251 mg/g [29]. Most of current research followed similar methods for AC for preparation under appropriate pyrolysis temperature, followed by measuring the removal value of any types of dyes at different environmental conditions like solution pH and initial dye concentration. The adsorption kinetic, isotherm, and maximum dye removal value were subsequently modelled and confirmed based on the obtained data. The adsorption capacity of adsorbate onto the adsorbent is determined at ambient condition (under certain experimental conditions). The procedure to obtain relative contributions of adsorption mechanisms (like initial concentration) using different agro-waste AC characteristic (like surface area, pore volume, pH, and particle size) were time consuming and complex. In the above, holistic adsorption mechanisms were not considered. In order to better deal with dyes in real water and wastewater, it is essential to understand the relative importance of each variable so as to gain the right solution for improving the adsorption capacity.

In recent years, machine learning (ML), playing an important role in computer science, artificial intelligence, chemistry, and biomedicine, has attracted people's attention [31]. Unlike empirical models, such as the Freundlich and Langmuir models used to detect the adsorption equilibrium, which can barely predict conclusions and make the relationship between operating conditions and adsorption capacity unavailable [26]; today, the machine learning (ML) method is preferred through modeling and learning the behavior of adsorption on agro-waste [32]. It may be preferred to resolve the problem through modeling and learning the adsorption behavior of dyes onto agro-wastes. High-quality machine learning models used to predict the adsorption efficiency have the ability to reduce the complexity, numbers, and time of experiments and to find a non-linear mathematical relationship between dependent and independent input variables. In recent studies, developing ML models could remarkably decrease the material resources and amount of manpower for future experiments and research. The emerging machine learning models, especially the decision tree (DT), random forest (RF), and gradient boosting models, have been successfully proved to have the merit to model and predict complex and non-linear mathematical relationships between dependent and independent factors. Therefore, the three machine learning models were selected and compared to (i) develop models for predicting the efficiency of dye adsorption onto the agro-wastes AC, based on the agro-waste characteristics, initial concentration, type of dyes, and conditions of adsorption; (ii) evaluate and select the relative importance of each variables and select the most important variables; and (iii) predict dyes adsorption efficiency and find the intrinsic information behind the models based on the agro-waste characteristics, which is valuable for reducing unnecessary repetitive experiments.

The models developed in this study are used to predict dye adsorption efficiency in wastewater based on measurable agro-wastes AC characteristics such as surface area, pH, pore volume, and particle size. This study with the aid of machine learning, which would be valuable for future applications with the increasing accumulation of big data in the scientific literature, while detecting the relative importance of each factor in improving adsorption efficiency; it provides a comprehensive understanding of dye removal using agro-wastes and proposed guidelines for the treatment of wastewater and contaminated water containing dyes.

2. Materials and Methods

2.1. Data Collection

The dye adsorption experimental data on the agro-waste AC were obtained from previous studies. The adsorption value data were collected directly from the tables or extracted from the Supplementary Material data and graphs with Getdata 2.21 in the published papers. Ultimately, 350 sets of adsorption experimental data of different dyes on the agro-waste AC were selected [15,16,22,25,26,30,33–39]. Some of data were missing and needed to be deleted, because dealing with missing data could cause errors in our whole dataset. The detailed data were shown in the Table S1 in Supplementary Materials data file. All the 39 adsorbents were produced all from agro-wastes (16 kinds of agro-wastes) at the temperature range of 400–800 °C. The characteristics of the agro-wastes were varied due to different feedstock and pyrolysis conditions, where the statistical distributions of the factors related to the agro-waste characteristics were acquired via boxplots.

In this study, the adsorption process was modelled and the dyes' adsorption capacity onto the agro-waste AC was predicted. Ten variables were considered as influencing factors and divided into four parts: (i) agro-waste characteristics, including surface area (SA, m^2/g), agro-waste pH in water (pHH_2O), pore volume (PV, cm^3/g), and particle size (PS, mm); (ii) adsorption conditions, including adsorption temperature (T, °C) and pH of the solution (pHsol); (iii) dye initial concentration (C_0, mg/g); and (iv) type of dyes. These explainable variables and adsorption efficiency values are summarized in Table S1.

2.2. Data Pre-Processing (Pre-Processing) and Model Estimators

The data processing algorithm was developed using the SK-learn module in Python 3.6 (Python Software Foundation, Beaverton, OR, USA). To eliminate the measurement unit, the predictors were normalized by mean and standard deviation before fitting in regression. In this paper, three methods were mainly used to compare model performance, gradient boosting (GB), decision tree (DT), and random forest (RF) models. To evaluate the accuracy of the model, the correlation coefficient (R), the mean squared error (MSE), and the root mean squared error were used.

$$R = \sqrt{1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i-1}^{N} (\hat{y}_i - \overline{y})^2}} \quad (1)$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

The Pearson correlation coefficient (PCC) measured the linear dependences between any two selected variables or between each feature and the target variable, which were calculated with Equation (4):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\sum_{i=1}^{n}(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4}$$

where $\overline{x}$ or $\overline{y}$ = mean of factors x or y, respectively. Then, each variable's data were normalized into the range of 0 and 1 with Equation (5):

$$y = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \tag{5}$$

where y = normalized value of initial x_i,and x_{max} and x_{min} = maximum and minimum value of x_i, respectively.

2.3. Models Built with Decision Tree, Gradient Boosting, and Random Forest Methods

Decision tree classifiers (DTCs) are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition, to name only a few. Perhaps the most important feature of DTCs is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution that is often easier to interpret [40]. Decision trees classify inputs into branch-like segments by taking paths from the root node (implies the prediction or that gives the best split of the target class values) through internal nodes to leaf nodes [41]. Each internal node contains splits and holds two or more child nodes, and further, splitting is applied iteratively to the subgroups until leaf nodes are obtained. The decision tree process is completed in one of these aspects: (a) the class label of the leaf node is the same as the target class value, (b) every prediction is used to split a partition, and (c) there are no more records for a particular value of a prediction or variable.

The training process is used to build and evaluate the decision tree model by minimizing the difference between the measured and predicted outputs. These procedures are applied to estimate the accuracy of the decision tree by comparing predicted outputs with actual data. Finally, a decision tree can be employed for classification and prediction purposes using a new dataset [41].

The random forest (RF) model, which works on the basis of bootstrap aggregation and multiple decision trees, is known as a supervised ensemble ML technique. It is included in the scope of the ensemble learning field. RF is in fact a bagging algorithm-based additive model. It should be noted that unlike bagging, RF builds each tree with the use of a random sample predictor before each node segmentation. This strategy results in a significant reduction in bias. The RF algorithm is a technique of classification, which makes use of a CART decision tree as the base classifier. In this model, every decision tree is produced in parallel, which could be either a regression tree or a classification tree. Each node within a decision tree is divided with the help of the optimal features capable of producing the optimum solution amongst all the available features. RF has been widely used in the literature in order to extract valuable, but there is hidden information in large volumes of data. This algorithm produces the training sets first with the application of the bootstrap method; afterward, it constructs a decision tree for each training set.

Each training subset was used to train various classifiers of the same type. After that, simple majority voting was applied to combining the individual classifiers. RF was implemented through three steps as follows: (1) performing the sampling process in a random way through dividing datasets into a number of subsamples; (2) training the decision trees with various subsamples, where each tree grew to the maximum degree on the basis of a bootstrap replicate of the training data, and each leaf node resulted the mean of all label values in the node; and (3) obtaining the final prediction through averaging the performance of all trees. To obtain the most desired model, the trial-and-error approach was used to optimize a number of tuning hyper-parameters, e.g., the number of trees (Ntree), the maximum number of features at each node (Nfeature), and the other input parameters. For the N case of each input variable, the relative importance was measured by means of the mean decrease impurity method. In this method, one can calculate how much each feature reduces the weighted impurity in a tree, and then they can average the reduction in the impurity from each feature and rank it for the developed forest.

Gradient boosting (GB) work based on the boosting principle, that is, to combine models with low variance error and high bias in a way to decrease bias and, at the same time, keep a low level of variance. Boosting learns multiple classifiers through altering the samples weight in the course of each training process. Then, it linearly integrates these classifiers with the aim of improving the classification performance. More specifically, boosting trees does not employ deep trees and various training datasets, rather, they prefer to make use of shallow trees trained in the same dataset; in this system, each tree is specialized in a definite characteristic of the relation between the input and output. In other words, succeeding shallow trees are trained in series, where the nth tree is trained in order to decrease the estimation errors of the former ($n-1$)th trees.

GB is essentially aimed at developing an additive model to minimize the loss function. First, at the initialization step, GB starts with a constant value minimizing the loss function. After that, in each iterative training process, the negative gradient of the loss function is predicted as the residual value in the current model, and then a novel regression tree is trained in a way to be fitting the current residual. Then, the final step involves the addition of the current regression tree to the former model and updating the residual. The algorithm's operation is continued iteratively until the stopping criterion (reaching the maximum number of iterations) is met. GB has been successful in improving the former poor performance of data through persistently employing the regression tree for the purpose of fitting the residuals. In the following, the GB algorithm is described briefly.

2.4. Statistical Evaluations

To evaluate the accuracy of the model, the correlation coefficient (R) and the root mean squared error (RMSE) were used.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(Y_i^{exp} - \overline{Y}_{ave}^{exp}\right)^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2} \tag{7}$$

where Y_i^{exp} and Y_i^{pred} = experimental value and predicted values, respectively, and $\overline{Y}_{ave}^{exp}$ = average of the experimental value.

3. Results and Discussion

Comparison of the Models

Table 1 presents the nine independent variables' weight based on output of the DT, RF, and GB techniques. According to this table, some mutual important variables were detected (for example, adsorption capacity, agro-waste pH, and surface area with highest weights in all three methods). The importance rank of the nine independent variables based on the DT, RF, and GB techniques is shown (in different colors) and presented in Figure 1. In this study, we adopted different feature selection methods to select only the important variables and develop our model based on selected variables. The main reason behind reducing the number of variables (based on their level of importance and correlations) is to reduce the complexity and improve the applicability of the final model. As shown in Figure 1, the weight values of each variable were summed up and compared. The accumulated weight values of the nine variables were sorted out from highest to lowest values in Table 1. As Table 1 presents, the initial concentration, surface area, and pore volume were detected as the most important variables, which is confirmed by previous studies. Therefore, the variables' importance is used based on three different tree-based supervised machine learning techniques and the five most important variables according to the sum of three models are selected. Table 2 presented the variable importance of the five selected variables based on feature selection criteria.

Table 1. Importance score (weight) of variables based on three machine learning (ML) methods.

DT		RF		GB	
Attribute	**Weight**	**Attribute**	**Weight**	**Attribute**	**Weight**
C_0	0.281	Pore volume	0.291	C_0	0.279
Surface area	0.122	C_0	0.267	Pore volume	0.086
Agro-waste pH	0.042	Surface area	0.131	Agro-waste pH	0.082
Pyrolysis temperature	0.03	Agro-waste pH	0.067	Surface area	0.055
Particle size	0.029	Adsorption pH	0.031	Dye type	0.032
Adsorption temperature	0.027	Particle size	0.029	Adsorption temperature	0.026
Adsorption pH	0.025	Pyrolysis temperature	0.022	Particle size	0.025
Pore volume	0.007	Adsorption temperature	0.013	Pyrolysis temperature	0
Dye type	0.004	Dye type	0.002	Adsorption pH	0

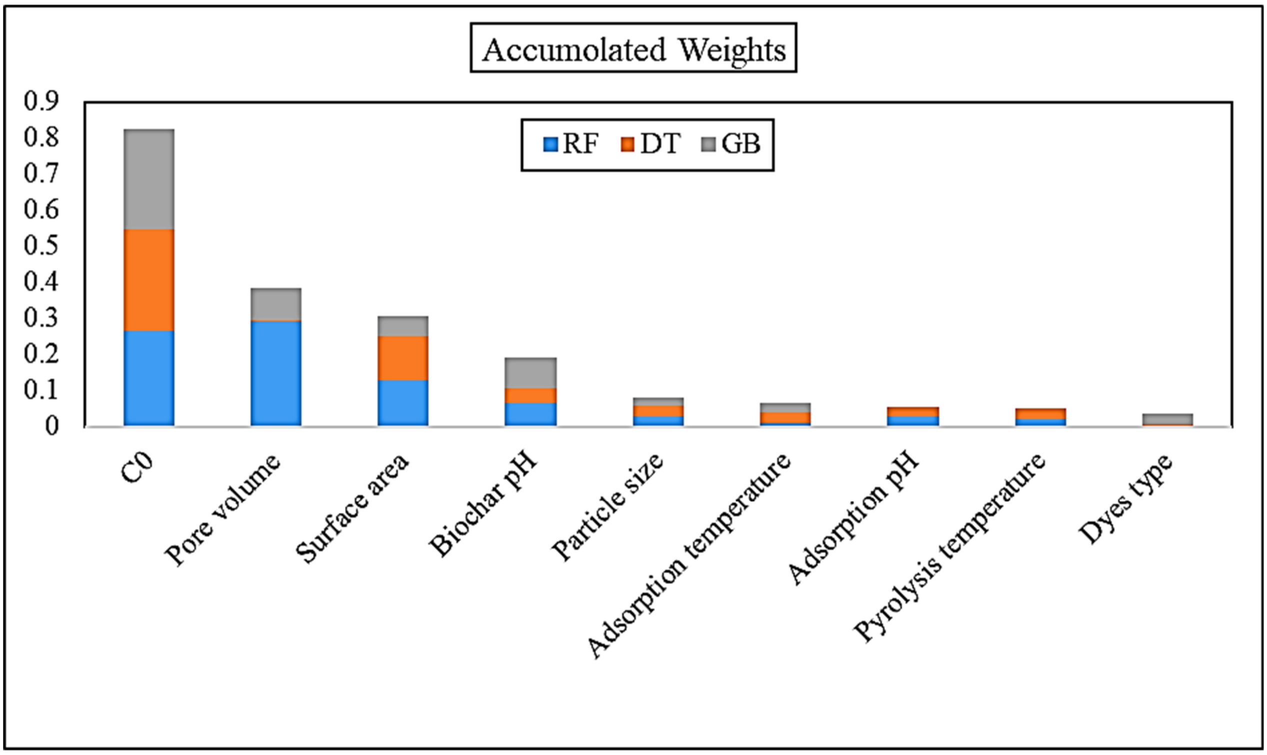

Figure 1. Accumulated weights of variables based on three ML methods.

Table 2. Variable importance of the five selected variables based on the feature selection criteria.

Attribute	RF	DT	GB
C_0	0.267	0.281	0.279
Pore volume	0.291	0.007	0.086
Surface area	0.131	0.122	0.055
Biochar pH	0.067	0.042	0.082
Particle size	0.029	0.029	0.025

The modeling using RF, GB, and DT was performed using nine variables (Table 3) and five selected variables (Table 4). According to Table 3, the R^2 has a maximum value of 0.92 using the RF models and minimum R^2 using the GB model. In addition, the RMSE and error have the lowest values using the RF models. Therefore, the overall RMSE value and R^2 value developed by the RF algorithm were acceptable and more accurate compared to the other two models. The prediction ability of RF model for the pollutant adsorption onto the agro-waste was proved in another study [42].

In this study, the ML methods including RF, DT, and GB were used and discussed to determine the deep relationship between adsorption capacity and five selected influencing variables. Table 4 presents the modeling performance with the five variables. According to Table 4, RF had highest accuracy (R^2 = 0.90) and lowest RMSE (0.0148) and absolute error (0.092) among the other two models. As it was expected, with decreasing the number of variables using feature selection, the accuracy decreased but only by 0.02. The modeling performance using the five variables was more valuable as we could reduce the complexity, and 2% difference is not much difference. Therefore, RF was competent and chosen as the most accurate model.

Table 3. Modeling performance with nine variables.

		Train			Test	
Index	**R^2**	**RMSE**	**Absolute Error**	**R^2**	**RMSE**	**Absolute Error**
RF	0.92	0.116	0.076	0.84	0.127	0.098
G.B	0.84	0.162	0.076	0.76	0.081	0.100
D.T	0.83	0.178	0.098	0.71	0.130	0.167

Table 4. Modeling performance with five variables.

		Train			Test	
Index	**R^2**	**RMSE**	**Absolute Error**	**R^2**	**RMSE**	**Absolute Error**
RF	0.90	0.148	0.092	0.81	0.150	0.098
G.B.	0.83	0.164	0.100	0.72	0.169	0.124
D.T	0.82	0.178	0.105	0.72	0.188	0.114

According to the Pearson correlation matrix (Table 5), the adsorption capacity was found to be the adsorption efficiency positively correlated with initial concentration, surface area, pore volume, and particle size and negatively correlated with the agro-waste pH. Some internal connections between the agro-waste characteristics were also detected with the Pearson correlation coefficient (PCC): (1) The particle size of agro-waste showed positive correlation with surface area and pore volume. This is because finer particles have a larger external surface exposed to heat, resulting in an extensive pore widening; thus, the surface area and micropore volume decreased as the latter transformed into mesopores [43]. (2) The agro-waste pH had an inverse correlation with pore volume and surface area. According to the previous studies, the higher carbonization degree meant removal of volatile matter agro-waste (has direct influence on carbon and nitrogen dynamics in soil) while higher ash percentage could reduce the surface area by filling micropores [44]. Moreover, ash content plays a main role in pH [45]. Therefore, the surface area of agro-waste showed a negative relation with the agro-waste pH (Table 5). In order to determine the deep relationship between the adsorption efficiency and these influencing factors, ML methods were used in the study and will be discussed in the next sections.

Table 5. Pearson correlation matrix between any two features of agro waste, and between any variable and the target variable (adsorption efficiency).

Attributes	Adsorption Efficiency	Biochar pH	C_0	Particle Size	Pore Volume	Surface Area
Adsorption efficiency	1	−0.558	0.739	0.084	0.666	0.690
Biochar pH	−0.558	1	−0.424	0.102	−0.5401	−0.252
C_0	0.739	−0.424	1	0.095	0.546	0.664
Particle size	0.084	0.102	0.095	1	0.154	0.277
Pore volume	0.666	−0.540	0.546	0.154	1	0.669
Surface area	0.690	−0.252	0.664	0.277	0.669	1

The Random Forest model was performed using the five selected variables and the importance of the variables is presented in Table 6. According to this table, the adsorption efficiency is highly affected by initial concentration (0.347), surface area (0.287), and pore volume (0.131), which is in line with previous findings [46,47]. The initial dye concentration has the highest effect on the adsorption capacity [47,48]. The increase of dye concentration gradients between the aqueous solution and adsorbent surface facilitated the adsorption of dye onto the agro-waste. Ins addition, the contribution of the main agro-waste characteristics was furtherly assessed in detail. It was found that pore volume was the main factor of agro-waste over the other properties of agro-waste. According to Table 6, the surface area is the second important factor of agro-wastee. A higher surface area could afford more active sites and an enhanced carbon/substrate interaction to improve the adsorption efficiency [49–51]. The impact of surface area on the adsorption capacity continuously increased within the range below ~600 m^2/g, but a further increase in the range of 600–1700 m^2/g showed a restrained trend, and over 1700 m^2/g there was a sharp decrease in adsorption efficiency. A very large surface area (over 1700 m^2/g) might have a negative impact on the other physicochemical characteristics of agro-waste like limited micropore accessibility, slow mass transfer, and diminished surface functional groups, which reciprocally influence the overall adsorption performance [39,52,53]. The evaluation of results in this study can contribute to present the adsorption of relevant water pollutants as dyes on adsorbents with high surface areas (like agro-waste).

Table 6. Importance of variables based on RF model.

Attribute	RF
C_0	0.347
Pore volume	0.287
Surface area	0.131
Agro-waste pH	0.056
Particle size	0.033

Additionally, the random forest (RF) model was optimized by simultaneously adjusting N_{tree} ranging from 20 to 140 and the maximal depth from 2 to 7 with a step size of 1 (Table 7). Model assessment was repeated 12 times using a different number of trees from 20 to 140, and the results are shown in Table 7. The minimum value of error rate was acquired when the N_{tree} and maximal depth were set as 140 and 7, respectively, which was easy to understand because the RF performance generally ameliorates with maximal depth increment due to higher number of available features to consider [54]. However, the best optimal model was acquired with the N_{tree} of 20 and maximal depth of 7 (no. 9 in Table 7). As shown in Table 7, the lowest error rate belonged to optimal model no. 12

(error rate = 0.347), which is very close to the error rate of model no. 9 (error rate = 0.351). Optimal model no. 9 was selected as the best model, because of it had the lowest N_{trees} (20) in comparison with N_{trees} of 140 (No. 12), and the error rate difference was only 0.004. The N_{tree} = 20 and maximal depth = 7 was used to assess the lowest error rate, as shown in Figure 2.

Table 7. The optimization process of RF models with error rate.

No.	Number of Trees	Maximal Depth	Error Rate
1	20.0	2.0	0.672
2	60.0	2.0	0.671
3	100.0	2.0	0.673
4	140.0	2.0	0.679
5	20.0	4.0	0.405
6	60.0	4.0	0.392
7	100.0	4.0	0.398
8	140.0	4.0	0.402
9	**20.0**	**7.0**	**0.351**
10	60.0	7.0	0.348
11	100.0	7.0	0.350
12	140.0	7.0	0.347

The predicted results in the test groups plotted versus the corresponding experimental data with the RF models are presented in Figure 3. The blue line represented the regression line. As it can be seen in Figure 3, the predicted outputs were satisfactory and the RF model could present valuable overall predicted ability (R^2 = 0.90).

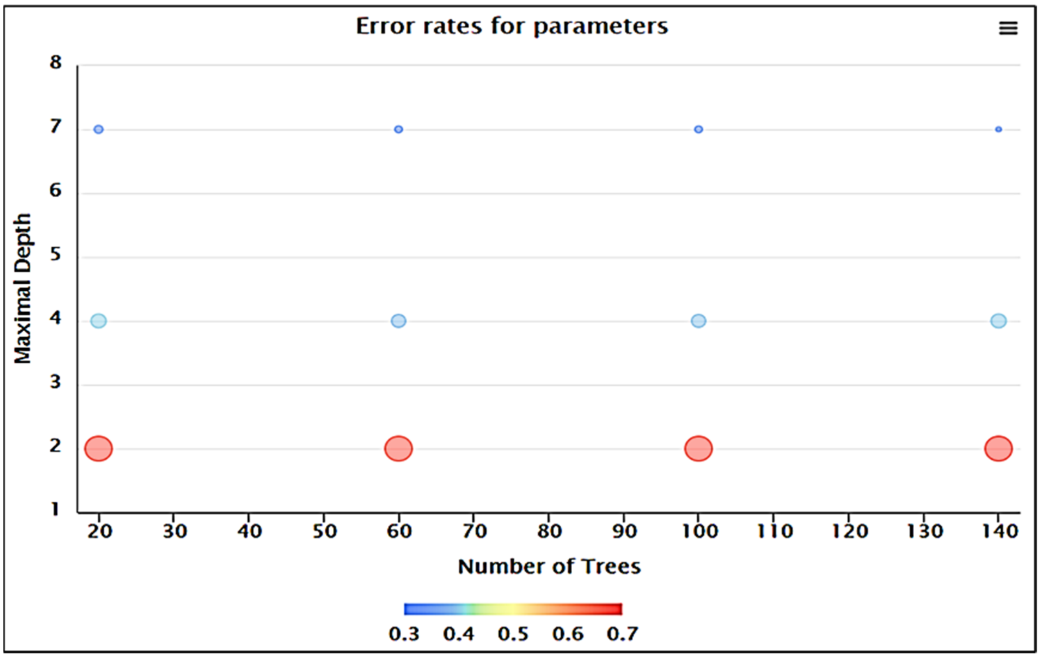

Figure 2. 3D optimization process of RF models with error rates.

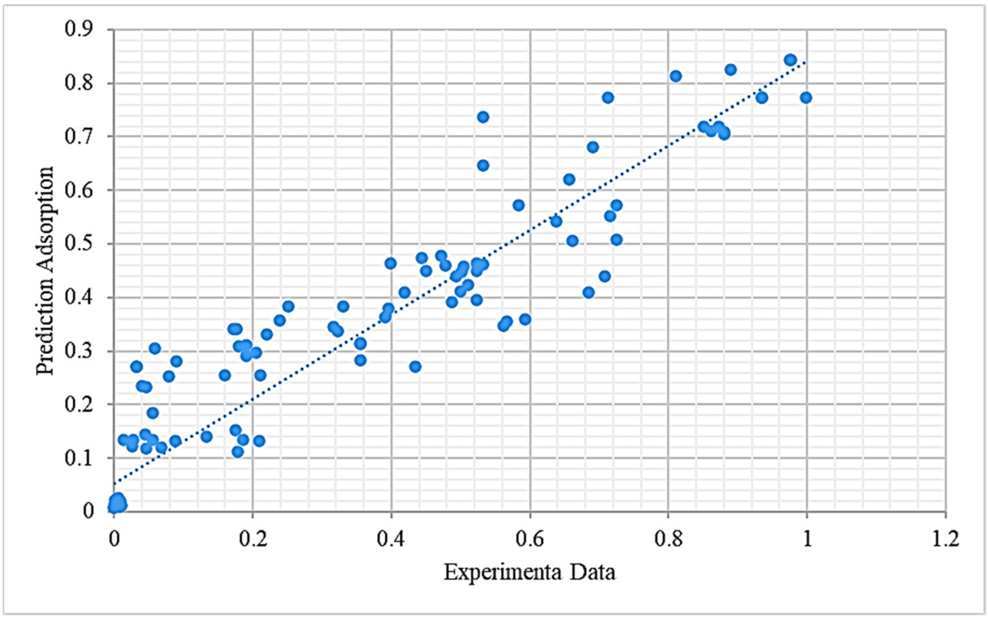

Figure 3. Predicted adsorption efficiency versus experimental data using random forest (RF) model.

For us, the accurate ability of the models' prediction was valuable, but the underlying models' information was more important based on such accurate predictions. For example, pore volume, surface area, and pH of agro-waste were illustrated as the most efficient variables for dyes' adsorption, so this class of agro-wastes are quite appropriate to be used for wastewater treatment with dyes. For the target pollutants (dyes), the selected model narrowed down the target of searching to get the best agro-waste AC adsorbents and decreased the unnecessary attempts with adsorption experiments. Additionally, this could decrease the potential harm for environmental systems and researchers in doing routine experiments. In addition, the RF, DT, and GB only took 3, 8, and 32 s for one job in the study, respectively, which could significantly accelerate the research and applications of dye adsorption onto the agro-waste. Table 8 showed the optimized value of the selected variables' values.

Table 8. The predicted importance of variables on the adsorption efficiency.

Attribute	Optimum Values (Normalized)	Optimum Values (Actual Value)	Maximum Adsorption	Maximum Adsorption (Actual Value)
C_0 (mg/L)	0.357	355	0.81	1813
Pore volume (cm^3/g)	0.889	1.027		
Surface area (m^2/g)	0.863	2106		
Agro-waste pH	0.272	2.53		
Particle size (mm)	0.112	0.013		

4. Conclusions

This study modelled the dye adsorption capacity onto agro-waste activated carbon using a machine learning approach based on nine variables: agro-waste characteristics, including surface area, agro-waste pH in water, pore volume, and particle size; (ii) adsorption conditions, including the adsorption temperature and pH of the solution; (iii) dye initial concentration; and (iv) type of dyes. RF, GB, and DT models were performed with a high accuracy of 0.92, 0.84, and 0.83, respectively. Five variables (initial concentration, pore volume, surface area, agro-waste pH, and particle size) were selected as the most effective variables on adsorption capacity. The three models were performed again on the five selected variables. The RF model with the highest accuracy (R^2 = 0.9) was selected as the best model for the prediction of adsorption capacity on agro-waste AC. The results showed that agro-waste characteristics (pore volume, surface area, agro-waste pH, and particle size) accounted for a 50.7% contribution for adsorption capacity. The pore volume and surface area are the most effective variables among the agro-waste characteristics, while the role of particle size was inconspicuous. The accurate prediction for adsorption efficiency of dyes and the intrinsic information behind the models based on the agro-waste characteristics was valuable to reducing unnecessary repetitive experiments and the rational design and optimal selection of agro-waste material for dye removal from industrial wastewater with minimal experimental screening efforts.

5. Future Perspectives

The future research suggestions of ML models in the field of dye adsorption for carrying out extensive studies are as follows:

- The dye adsorption capability using other ML models such as the group method of data handling (GMDH).
- The application of machine learning models on dye adsorption using raw bio-waste.
- Specific study on the capability of ML models on the adsorption of pollutants like BOD and COD.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/nano11102734/s1. Table S1. A Study on Machine Learning Methods' Application for Dye Adsorption Prediction onto Agricultural Waste Activated Carbon.

Author Contributions: Conceptualization, S.M. and S.M.H.M.; Methodology: S.M., S.M.H.M. and N.f.B.M.F.; software: S.M.H.M.; validation, S.M.H.M.; formal analysis, E.E.H.; investigation, Z.M.E.-B., and J.U.; data curation: Z.M.E.-B.; writing—original draft preparation: S.M.; writing—review and editing: S.M., O.M., Z.M.E.-B.; J.U., Y.A.E.-B. and E.E.H.; supervision, J.U.; project administration, O.M. and Y.A.E.-B.; funding acquisition, Y.A.E.-B. and O.M. All authors have read and agreed to the published version of the manuscript.

Funding: Project [TURSP-2020/106] Taif University and the Research Council of Lithuania (LMTLT) S-MIP-19-61. Also, the authors acknowledge the financial support of Taif University Researchers Supporting Project number (TURSP-2020/106), Taif University, Taif, Saudi Arabia.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Kausar, A.; Iqbal, M.; Javed, A.; Aftab, K.; Nazli, Z.; Nawaz, H.; Nouren, S. Dyes adsorption using clay and modified clay: A review. *J. Mol. Liq.* **2018**, *256*, 395–407. [CrossRef]
2. Hunger, K. *Industrial Dyes: Chemistry, Properties, Applications*; Wiley-Vch: Frankfurt, Germany, 2007; ISBN 3-527-30426-6.
3. Regti, A.; Laamari, M.R.; Stiriba, S.-E.; El Haddad, M. Use of response factorial design for process optimization of basic dye adsorption onto activated carbon derived from Persea species. *Microchem. J.* **2017**, *130*, 129–136. [CrossRef]
4. Abdolali, A.; Guo, W.S.; Ngo, H.H.; Chen, S.S.; Nguyen, N.C.; Tung, K.L. Typical lignocellulosic wastes and by-products for biosorption process in water and wastewater treatment: A critical review. *Bioresour. Technol.* **2014**, *160*, 57–66. [CrossRef]
5. Asfaram, A.; Ghaedi, M.; Agarwal, S.; Tyagi, I.; Gupta, V.K. Removal of basic dye Auramine-O by ZnS: Cu nanoparticles loaded on activated carbon: Optimization of parameters using response surface methodology with central composite design. *RSC Adv.* **2015**, *5*, 18438–18450. [CrossRef]
6. Lellis, B.; Fávaro-Polonio, C.Z.; Pamphile, J.A.; Polonio, J.C. Effects of textile dyes on health and the environment and bioremediation potential of living organisms. *Biotechnol. Res. Innov.* **2019**, *3*, 275–290. [CrossRef]
7. León, O.; Muñoz-Bonilla, A.; Soto, D.; Pérez, D.; Rangel, M.; Colina, M.; Fernández-García, M. Removal of anionic and cationic dyes with bioadsorbent oxidized chitosans. *Carbohydr. Polym.* **2018**, *194*, 375–383. [CrossRef]
8. Qi, X.; Wei, W.; Li, J.; Zuo, G.; Pan, X.; Su, T.; Zhang, J.; Dong, W. Salecan-based pH-sensitive hydrogels for insulin delivery. *Mol. Pharm.* **2017**, *14*, 431–440. [CrossRef]
9. Moosavi, S.; Lai, C.W.; Gan, S.; Zamiri, G.; Akbarzadeh Pivehzhani, O.; Johan, M.R. Application of efficient magnetic particles and activated carbon for dye removal from wastewater. *ACS Omega* **2020**, *5*, 20684–20697. [CrossRef] [PubMed]
10. Moosavi, S.; Lai, C.W.; Akbarzadeh, O.; Johan, M.R. Recycled Activated Carbon-Based Materials for the Removal of Organic Pollutants from Wastewater. *Waste Recycl. Technol. Nanomater. Manuf.* **2021**, 513–539.

11. Moosavi, S.; Li, R.Y.M.; Lai, C.W.; Yusof, Y.; Gan, S.; Akbarzadeh, O.; Chowhury, Z.Z.; Yue, X.G.; Johan, M.R. Methylene blue dye photocatalytic degradation over synthesised fe3 o4 /ac/tio2 nano-catalyst: Degradation and reusability studies. *Nanomaterials* **2020**, *10*, 2360. [CrossRef]
12. Moosavi, S.; Gan, S.; Zakaria, S. Functionalized cellulose beads with activated carbon Fe_3O_4/CoFe $_2O_4$ for cationic dye removal. *Cellul. Chem. Technol.* **2019**, *53*, 815–825. [CrossRef]
13. Zhang, B.; Ma, Z.; Yang, F.; Liu, Y.; Guo, M. Adsorption properties of ion recognition rice straw lignin on $PdCl_4{}^{2-}$: Equilibrium, kinetics and mechanism. *Colloids Surf. A Physicochem. Eng. Asp.* **2017**, *514*, 260–268. [CrossRef]
14. Al-Saadi, A.A.; Saleh, T.A.; Gupta, V.K. Spectroscopic and computational evaluation of cadmium adsorption using activated carbon produced from rubber tires. *J. Mol. Liq.* **2013**, *188*, 136–142. [CrossRef]
15. Wong, S.; Yac'cob, N.A.N.; Ngadi, N.; Hassan, O.; Inuwa, I.M. From pollutant to solution of wastewater pollution: Synthesis of activated carbon from textile sludge for dye adsorption. *Chin. J. Chem. Eng.* **2018**, *26*, 870–878. [CrossRef]
16. Gao, Y.; Xu, S.; Yue, Q.; Wu, Y.; Gao, B. Chemical preparation of crab shell-based activated carbon with superior adsorption performance for dye removal from wastewater. *J. Taiwan Inst. Chem. Eng.* **2016**, *61*, 327–335. [CrossRef]
17. Aravind, M.; Amalanathan, M. Structural, morphological, and optical properties of country egg shell derived activated carbon for dye removal. *Mater. Today Proc.* **2021**, *43*, 1491–1495. [CrossRef]
18. Milenković, D.D.; Bojić, A.L.; Veljković, V.B. Ultrasound-assisted adsorption of 4-dodecylbenzene sulfonate from aqueous solutions by corn cob activated carbon. *Ultrason. Sonochem.* **2013**, *20*, 955–962. [CrossRef] [PubMed]
19. Ajmal, M.; Rao, R.A.K.; Khan, M.A. Adsorption of copper from aqueous solution on Brassica cumpestris (mustard oil cake). *J. Hazard. Mater.* **2005**, *122*, 177–183. [CrossRef]
20. Chen, S.; Qin, C.; Wang, T.; Chen, F.; Li, X.; Hou, H.; Zhou, M. Study on the adsorption of dyestuffs with different properties by sludge-rice husk biochar: Adsorption capacity, isotherm, kinetic, thermodynamics and mechanism. *J. Mol. Liq.* **2019**, *285*, 62–74. [CrossRef]
21. Fideles, R.A.; Ferreira, G.M.D.; Teodoro, F.S.; Adarme, O.F.H.; da Silva, L.H.M.; Gil, L.F.; Gurgel, L.V.A. Trimellitated sugarcane bagasse: A versatile adsorbent for removal of cationic dyes from aqueous solution. Part I: Batch adsorption in a monocomponent system. *J. Colloid Interface Sci.* **2018**, *515*, 172–188. [CrossRef]
22. Gupta, K.; Gupta, D.; Khatri, O.P. Graphene-like porous carbon nanostructure from Bengal gram bean husk and its application for fast and efficient adsorption of organic dyes. *Appl. Surf. Sci.* **2019**, *476*, 647–657. [CrossRef]
23. Suganya, S.; Saravanan, A.; Ravikumar, C. Computation of adsorption parameters for the removal of dye from wastewater by microwave assisted sawdust: Theoretical and experimental analysis. *Environ. Toxicol. Pharmacol.* **2017**, *50*, 45–57.
24. Pholosi, A.; Naidoo, E.B.; Ofomaja, A.E. Batch and continuous flow studies of Cr (VI) adsorption from synthetic and real wastewater by magnetic pine cone composite. *Chem. Eng. Res. Des.* **2020**, *153*, 806–818. [CrossRef]
25. Archin, S.; Sharifi, S.H.; Asadpour, G. Optimization and modeling of simultaneous ultrasound-assisted adsorption of binary dyes using activated carbon from tobacco residues: Response surface methodology. *J. Clean. Prod.* **2019**, *239*, 118136. [CrossRef]
26. Xiao, W.; Garba, Z.N.; Sun, S.; Lawan, I.; Wang, L.; Lin, M.; Yuan, Z. Preparation and evaluation of an effective activated carbon from white sugar for the adsorption of rhodamine B dye. *J. Clean. Prod.* **2020**, *253*, 119989. [CrossRef]
27. Ghaedi, M.; Ghaedi, A.M.; Negintaji, E.; Ansari, A.; Vafaei, A.; Rajabi, M. Random forest model for removal of bromophenol blue using activated carbon obtained from Astragalus bisulcatus tree. *J. Ind. Eng. Chem.* **2014**, *20*, 1793–1803. [CrossRef]
28. Jain, S.N.; Tamboli, S.R.; Sutar, D.S.; Jadhav, S.R.; Marathe, J.V.; Shaikh, A.A.; Prajapati, A.A. Batch and continuous studies for adsorption of anionic dye onto waste tea residue: Kinetic, equilibrium, breakthrough and reusability studies. *J. Clean. Prod.* **2020**, *252*, 119778. [CrossRef]
29. Li, L.; Wu, M.; Song, C.; Liu, L.; Gong, W.; Ding, Y.; Yao, J. Efficient removal of cationic dyes via activated carbon with ultrahigh specific surface derived from vinasse wastes. *Bioresour. Technol.* **2021**, *322*, 124540. [CrossRef]
30. Alshabib, M.; Oluwadamilare, M.A.; Tanimu, A.; Abdulazeez, I.; Alhooshani, K.; Ganiyu, S.A. Experimental and DFT investigation of ceria-nanocomposite decorated AC derived from groundnut shell for efficient removal of methylene-blue from wastewater effluent. *Appl. Surf. Sci.* **2021**, *536*, 147749. [CrossRef]
31. Mullainathan, S.; Spiess, J. Machine learning: An applied econometric approach. *J. Econ. Perspect.* **2017**, *31*, 87–106. [CrossRef]
32. Zhu, X.; Wang, X.; Ok, Y.S. The application of machine learning methods for prediction of metal sorption onto biochars. *J. Hazard. Mater.* **2019**, *378*, 120727. [CrossRef] [PubMed]
33. Lu, P.; Lin, H.; Yu, W.; Chern, J. Chemical regeneration of activated carbon used for dye adsorption. *J. Taiwan Inst. Chem. Eng.* **2011**, *42*, 305–311. [CrossRef]
34. Shokry, H.; Elkady, M.; Hamad, H. Nano activated carbon from industrial mine coal as adsorbents for removal of dye from simulated textile wastewater: Operational parameters and mechanism study. *J. Mater. Res. Technol.* **2019**, *8*, 4477–4488. [CrossRef]
35. Mei, S.; Gu, J.; Ma, T.; Li, X.; Hu, Y.; Li, W.; Zhang, J.; Han, Y. N-doped activated carbon from used dyeing wastewater adsorbent as a metal-free catalyst for acetylene hydrochlorination. *Chem. Eng. J.* **2019**, *371*, 118–129. [CrossRef]
36. Ravenni, G.; Cafaggi, G.; Sárossy, Z.; Nielsen, K.T.R.; Ahrenfeldt, J.; Henriksen, U.B. Waste chars from wood gasification and wastewater sludge pyrolysis compared to commercial activated carbon for the removal of cationic and anionic dyes from aqueous solution. *Bioresour. Technol. Rep.* **2020**, *10*, 100421. [CrossRef]
37. Rani, K.M.; Palanisamy, P.N.; Gayathri, S.; Tamilselvi, S. Adsorptive removal of basic violet dye from aqueous solution by activated carbon prepared from tea dust material. *Int. J. Innov. Res. Sci. Eng. Technol.* **2015**, *4*, 6845–6853.

38. Mudyawabikwa, B.; Mungondori, H.H.; Tichagwa, L.; Katwire, D.M. Methylene blue removal using a low-cost activated carbon adsorbent from tobacco stems: Kinetic and equilibrium studies. *Water Sci. Technol.* **2017**, *75*, 2390–2402. [CrossRef]
39. Wang, H.; Li, Z.; Yahyaoui, S.; Hanafy, H.; Seliem, M.K.; Bonilla-Petriciolet, A.; Dotto, G.L.; Sellaoui, L.; Li, Q. Effective adsorption of dyes on an activated carbon prepared from carboxymethyl cellulose: Experiments, characterization and advanced modelling. *Chem. Eng. J.* **2021**, *417*, 128116. [CrossRef]
40. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man. Cybern.* **1991**, *21*, 660–674. [CrossRef]
41. Yu, Z.; Haghighat, F.; Fung, B.C.M.; Yoshino, H. A decision tree method for building energy demand modeling. *Energy Build.* **2010**, *42*, 1637–1646. [CrossRef]
42. Jiang, W.; Xing, X.; Li, S.; Zhang, X.; Wang, W. Synthesis, characterization and machine learning based performance prediction of straw activated carbon. *J. Clean. Prod.* **2019**, *212*, 1210–1223. [CrossRef]
43. Mui, E.L.K.; Cheung, W.H.; McKay, G. Tyre char preparation from waste tyre rubber for dye removal from effluents. *J. Hazard. Mater.* **2010**, *175*, 151–158. [CrossRef]
44. Gao, L.-Y.; Deng, J.-H.; Huang, G.-F.; Li, K.; Cai, K.-Z.; Liu, Y.; Huang, F. Relative distribution of Cd^{2+} adsorption mechanisms on biochars derived from rice straw and sewage sludge. *Bioresour. Technol.* **2019**, *272*, 114–122. [CrossRef]
45. Reyhanitabar, A.; Frahadi, E.; Ramezanzadeh, H.; Oustan, S. Effect of Pyrolysis Temperature and Feedstock Sources on Physicochemical Characteristics of Biochar. *J. Agric. Sci. Technol.* **2020**, *22*, 547–561.
46. Ahmad, H.; Nurunnabi, M.; Rahman, M.M.; Kumar, K.; Tauer, K.; Minami, H.; Gafur, M.A. Magnetically doped multi stimuli-responsive hydrogel microspheres with IPN structure and application in dye removal. *Colloids Surf. A Physicochem. Eng. Asp.* **2014**, *459*, 39–47. [CrossRef]
47. Zhu, X.; Wan, Z.; Tsang, D.C.W.; He, M.; Hou, D.; Su, Z.; Shang, J. Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. *Chem. Eng. J.* **2021**, *406*, 126782. [CrossRef]
48. Zhou, N.; Chen, H.; Xi, J.; Yao, D.; Zhou, Z.; Tian, Y.; Lu, X. Biochars with excellent Pb (II) adsorption property produced from fresh and dehydrated banana peels via hydrothermal carbonization. *Bioresour. Technol.* **2017**, *232*, 204–210. [CrossRef]
49. Wang, H.; Chu, Y.; Fang, C.; Huang, F.; Song, Y.; Xue, X. Sorption of tetracycline on biochar derived from rice straw under different temperatures. *PLoS ONE* **2017**, *12*, e0182776. [CrossRef] [PubMed]
50. He, J.; Dai, J.; Zhang, T.; Sun, J.; Xie, A.; Tian, S.; Yan, Y.; Huo, P. Preparation of highly porous carbon from sustainable α-cellulose for superior removal performance of tetracycline and sulfamethazine from water. *RSC Adv.* **2016**, *6*, 28023–28033. [CrossRef]
51. Selmi, T.; Sanchez-Sanchez, A.; Gadonneix, P.; Jagiello, J.; Seffen, M.; Sammouda, H.; Celzard, A.; Fierro, V. Tetracycline removal with activated carbons produced by hydrothermal carbonisation of Agave americana fibres and mimosa tannin. *Ind. Crops Prod.* **2018**, *115*, 146–157. [CrossRef]
52. Gonzalez-Serrano, E.; Cordero, T.; Rodriguez-Mirasol, J.; Cotoruelo, L.; Rodriguez, J.J. Removal of water pollutants with activated carbons prepared from H_3PO_4 activation of lignin from kraft black liquors. *Water Res.* **2004**, *38*, 3043–3050. [CrossRef] [PubMed]
53. Arampatzidou, A.C.; Deliyanni, E.A. Comparison of activation media and pyrolysis temperature for activated carbons development by pyrolysis of potato peels for effective adsorption of endocrine disruptor bisphenol-A. *J. Colloid Interface Sci.* **2016**, *466*, 101–112. [CrossRef] [PubMed]
54. Ahmad, M.W.; Mourshed, M.; Rezgui, Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build.* **2017**, *147*, 77–89. [CrossRef]