Biochar

ORIGINAL RESEARCH

Open Access

Machine learning predicting and engineering the yield, N content, and specific surface area of biochar derived from pyrolysis of biomass

Lijian Leng[1], Lihong Yang[1], Xinni Lei[1], Weijin Zhang[1], Zejian Ai[1], Zequn Yang[1], Hao Zhan[1], Jianping Yang[1], Xingzhong Yuan[2], Haoyi Peng[1*] and Hailong Li[1*]

Abstract

Biochar produced from pyrolysis of biomass has been developed as a platform carbonaceous material that can be used in various applications. The specific surface area (SSA) and functionalities such as N-containing functional groups of biochar are the most significant properties determining the application performance of biochar as a carbon material in various areas, such as removal of pollutants, adsorption of CO_2 and H_2, catalysis, and energy storage. Producing biochar with preferable SSA and N functional groups is among the frontiers to engineer biochar materials. This study attempted to build machine learning models to predict and optimize specific surface area of biochar (SSA-char), N content of biochar (N-char), and yield of biochar (Yield-char) individually or simultaneously, by using elemental, proximate, and biochemical compositions of biomass and pyrolysis conditions as input variables. The predictions of Yield-char, N-char, and SSA-char were compared by using random forest (RF) and gradient boosting regression (GBR) models. GBR outperformed RF for most predictions. When input parameters included elemental and proximate compositions as well as pyrolysis conditions, the test R^2 values for the single-target and multi-target GBR models were 0.90–0.95 except for the two-target prediction of Yield-char and SSA-char which had a test R^2 of 0.84 and the three-target prediction model which had a test R^2 of 0.81. As indicated by the Pearson correlation coefficient between variables and the feature importance of these GBR models, the top influencing factors toward predicting three targets were specified as follows: pyrolysis temperature, residence time, and fixed carbon for Yield-char; N and ash for N-char; ash and pyrolysis temperature for SSA-char. The effects of these parameters on three targets were different, but the trade-offs of these three were balanced during multi-target ML prediction and optimization. The optimum solutions were then experimentally verified, which opens a new way for designing smart biochar with target properties and oriented application potential.

Highlights

- Yield, N content, and specific surface area of biochar predicted by machine learning
- Gradient boosting regression outperformed random forest, with test R^2 of 0.81–0.95

*Correspondence: penghaoyicsu@126.com; hailongli18@gmail.com

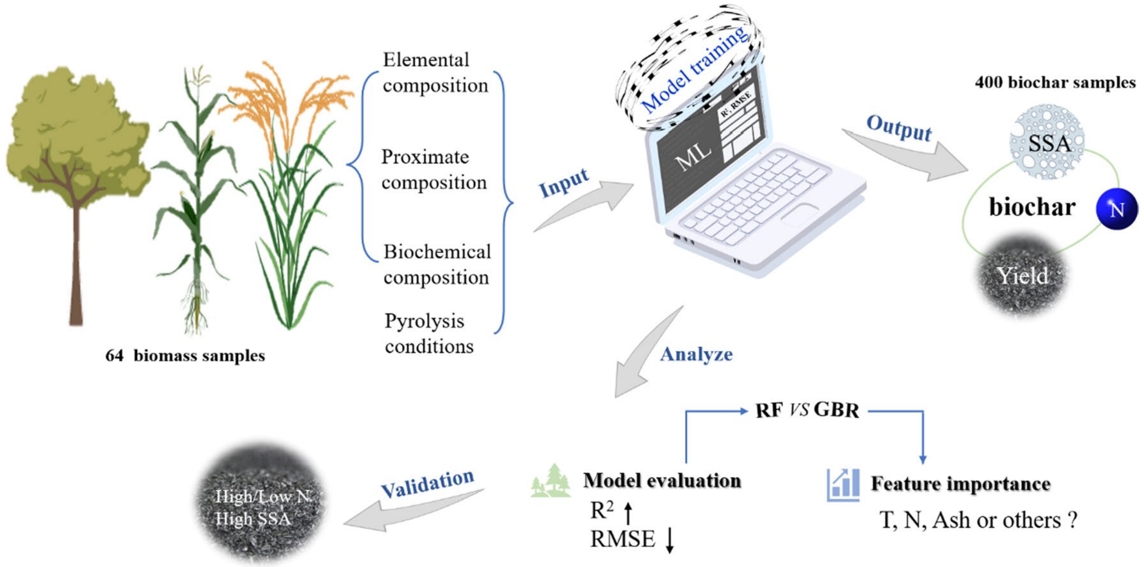
[1] School of Energy Science and Engineering, Central South University, Changsha 410083, Hunan, People's Republic of China
Full list of author information is available at the end of the article

Springer

- Temperature, nitrogen, and ash were top features for predicting the three targets
- The yield and properties of biochar were engineered and experimentally verified

Keywords: Specific surface area, Nitrogen, Biochar, Pyrolysis, Machine learning

Graphical Abstract



1 Introduction

Biochar, a platform carbonaceous material or pyrogenic organic matter produced from the pyrolysis of biomass, has attracted increasing interest among researchers, practitioners, and policymakers worldwide due to its wide application potentials in a variety of areas (Liu et al. 2015; Wang and Wang 2019; Chen et al. 2021). Biochar was first introduced as a potential carbon sequester since it can store the biochar carbon or black carbon in the soil for more than 100 years, helping mitigate climate change (Amalina et al. 2022; Parthasarathy et al. 2022). In the recent decade, it has been widely used in various domains such as adsorption of pollutants, energy storage, chemical reaction catalysis, and soil remediation and amendment (Leng et al. 2019; Huang et al. 2019; Qiu et al. 2021; Liu et al. 2022). The potential for the application of biochar in different areas mainly depends on its surface characteristics, including specific surface area (SSA) and functionalities, etc. (Cao et al. 2020; He et al. 2021; Leng et al. 2022; Yang et al. 2022).

SSA generally determines the number of sites, particularly active sites that work effectively in different applications, and the higher SSA means higher availability of the material's surface, and thus better application performance for most cases. For example, SSA ranked second (only after pressure), among the nine influencing factors studied in a machine learning (ML) model, to predict the hydrogen uptake capacity of porous carbon materials (Cao et al. 2016; Maulana Kusdhany and Lyth 2021). This is reasonable since SSA is determined based on the gas uptake capability using N_2 or sometimes CO_2 under a compressed atmosphere. Therefore, SSA is decisive to the CO_2 adsorption capacity of biochar, and it was more important than other biochar variables in ML models (Yuan et al. 2021). SSA was also listed as the top factor in ML models determining the adsorption of organic pollutants such as pharmaceuticals and personal care products on biochar and activated carbon, while the effects of other variables seemed negligible (Zhu et al. 2021, 2022). Moreover, many other ML predictions studied the adsorption coefficient of various organics on biochar, and they indicated that SSA was the most important descriptor (Sigmund et al. 2020; Pathy et al. 2020; Zhao et al. 2021, 2022). In another ML study, SSA was ranked the most significant property for carbonaceous membrane electrode materials of proton exchange membrane fuel cells among the 26 studied variables that might affect the current density and specific power of the fuel cells (Huo et al. 2021). Therefore, SSA is among the most important properties that determine the application performance of carbon materials such as biochar, although the importance level varies between applications and between studies.

N-containing functional groups on biochar such as pyridinic-N, pyrrolic-N, quaternary-N, and amine-N (proteinic-N) have been reported to be responsible for the favorable performance of biochar in catalysis of oxygen reduction, electric catalysis, and adsorption of various pollutants, etc. (Leng et al. 2021b, 2022). For example, the N content of the electrode materials ranked the second (following SSA) for determining the performance of proton exchange membrane fuel cells (Huo et al. 2021). The atomic ratio (O + N)/C of biochar and activated carbon, representing the quantity of O-/N-containing functional groups, also ranked after SSA to predict the adsorption of pharmaceuticals and personal care products (Zhu et al. 2022). In a recent ML study, the N content of biochar (N-char) even ranked the first among the 20 variables that affect immobilization of heavy metal by biochar in soil, while the effect of SSA was negligible (ranked 17) (Palansooriya et al. 2022). The presence of N-containing functional groups on the biochar surface provides active sites for heavy metals immobilization through strong covalent bonding, H bonding, chelation, and electrostatic attraction (Cao et al. 2021; Palansooriya et al. 2022), and the general sites of biochar (represented by SSA) should only have a low affinity to heavy metals. Among various N-containing functional groups, pyridinic-N was proved particularly efficient in catalyzing oxygen reduction reaction in an experimental study reported in Science (Guo et al. 2016), which received extensive attention among researchers.

Based on the above studies, it is hypothesized that the fabrication of highly oriented pyrolytic biochar with specific surface characteristics, such as high SSA or high content of N (functional groups), or both, can significantly enhance the performance of biochar applications. According to our previous reviews, which overviewed the influence of various factors on N content (Leng et al. 2020a) and SSA of biochar (Leng et al. 2021a), there are so many variables, such as compositions and properties of biomass (elemental composition, etc.), pyrolysis parameters (temperature, residence time, etc.), and additional treatments (Leng et al. 2020b, 2021a; Chen et al. 2022), that have considerable effects on these two properties as well as the yield of biochar (Yield-char). Therefore, engineering biochar with either high SSA or high N content with experimental screening method is difficult, not to mention mediating these two biochar characteristics with Yield-char together.

However, the prediction of the SSA-char or N-char by ML has not yet been revealed, and the multi-target prediction of the yield with either SSA or N-char, or both, has also not been reported. The achievement of any of the above-mentioned predictions can be crucial to engineer smart biochar carbon materials, and simultaneously mediating two of the three targets or regulating the three altogether can find its practical use in various applications. Therefore, in the present study, the single-target and multi-target predictions of the Yield-char, SSA-char, and N-char will be addressed first by varying models (algorithms) and model hyper-parameters. Then, the optimization will be applied to guide the design of pyrolysis parameters for selected biomass to produce desired biochar with experimental validations.

2 Methodologies

2.1 Datasets collected

Data of 400 biochar samples were collected from 50 SCI journal papers (references provided in the Additional file 1 and detail biochar samples listed in the Additional file 9), and the biochar samples were produced from pyrolysis of 64 biomasses, including forestry waste, agricultural waste, manure, food waste, algae, grass, sludge, and their mixtures, etc. The following 14 input variables were collected: (i) the elemental composition of biomass (the element contents of carbon, hydrogen, nitrogen, and oxygen, C-H-N-O); (ii) the proximate composition of biomass (ash, fixed carbon, and volatile matter, Ash-FC-VM); (iii) the structural or biochemical composition of biomass (cellulose, hemicellulose, lignin, and extractive, Cel-Hem-Lig-Ext), and (iv) the pyrolysis conditions (pyrolysis temperature, heating rate, and residence time, T-HR-RT). The characteristics of biochar (outputs) included 3 targets, namely Yield-char, N-char, and SSA-char. Because the prediction target varied for single-target and multi-target models, the collected data were divided into 7 datasets according to the 7 different prediction tasks: (i) 3 single-target prediction tasks (predicting Yield-char, N-char, and SSA-char individually, datasets #1-#3 from Additional files 2, 3, and 4), (ii) 3 two-target prediction tasks (predicting either of the two from Yield-char, N-char, and SSA-char, datasets #4-#6 from Additional files 5, 6, and 7), and (iii) 1 three-target prediction task (dataset #7 from Additional file 8). Since not each piece of data set contains full information of all the above input and output variables and the effects of different input variables on prediction performance are well worthy of study, every dataset from datasets #1-#7 was split into 6 sub-datasets (datasets #n_m, n = 1–7 and m = 1–6) depending on input variables (combination of different biomass variables with the same pyrolysis parameters): (i) C-H-N-O, T-RT-HR (m = 1); (ii) Ash-FC-VM, T-RT-HR (m = 2); (iii) Cel-Hem-Lig-Ext, T-RT-HR (m = 3); (iv)

C-H-N-O, Ash, T-RT-HR (m = 4); (v) C-H-N-O, Ash-FC-VM, T-RT-HR (m = 5); (vi) C-H-N-O, Ash-FC-VM, Cel-Hem-Lig-Ext, T-RT-HR (m = 6). A total of 42 datasets were split, and correspondingly, 42 ML prediction tasks were conducted.

In all datasets, the S content in biomass was not considered as input in ML prediction because S content is either very low compared to the contents of other elements (basically negligible) or is not provided in the collected data. Proximate analysis was unified based on a dry basis, while the elemental composition was unified based on ash-free and dry basis, and O was calculated by difference (Eqs. 1–3). The Yield-char, and SSA-char collected were on dry basis, while N-char was on ash-free basis.

$$\mathrm{O} = 100 - C - H - N - S \text{ (if available)} \tag{1}$$

$$Ash + FC + VM = 100 \tag{2}$$

$$Cel + Hem + Lig + Ext = 100 \tag{3}$$

To confirm input and output parameters obey normal distribution and are on a proper scale, the collected datasets were normalized before training the ML model based on Eq. (4):

$$x_i^* = \frac{x_i - \mu}{S} \tag{4}$$

where x_i indicates the value of input value i; x_i^* is the normalized value of origin x_i; s and μ represent the standard deviation and the mean value of input variable x_i, respectively.

The linear correlation between two variables was measured through Pearson correlation coefficient (PCC) calculated according to Eq. (5):

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

where ρ_{xy} is the value of PCC for two variables; $\bar{x}$ and $\bar{y}$ are the mean of input variable x and output variable y. The value of ρ_{xy} is between 1 and −1, where a positive or negative number means positive or negative correlation, and 0 means no linear correlation. The PCC was applied to detect collinearity relationship between any two input variables and linear correlation between input variables and target variables in this study (Were et al. 2015).

2.2 ML models and hyper-parameters

Two classical predictive models based on ML algorithms, including random forest (RF) and gradient boosting regression (GBR), were applied to explore the relationship between the characteristics of biomass and biochar and predict the yield and properties of biochar. RF, a ML algorithm building and combining decision trees, can deal with the nonlinear regression problems between variables (Li et al. 2020a). GBR algorithm is one of the ensemble learning algorithms, which is trained by adopting boosting strategy (Tan et al. 2021). The typical conception of boosting is to integrate a bunch of weak learning algorithms to form a strong model with good predictive performance through an iterative method, in which every learning algorithm learns from the mistakes of the previous one (Cai et al. 2020; Rzychoń et al. 2020). Compared to other ML models, these two ML models showed many advantages, including strong model interpretability, strong generalization ability, fast convergence during training, great adaptability for multiple feature data, etc. (Li et al. 2021b). Moreover, some of the published literatures have indicated that these two algorithms have extraordinary potential in predicting biochar (Zhu et al. 2019a; Li et al. 2020b, 2021a).

The 5-fold cross-validation method was applied to optimize hyper-parameters of two kinds of ML models, which would improve model's prediction ability during training. A detailed description about 5-fold cross-validation can be found in the link (https://scikit-learn.org/stable/modules/cross_validation.html#multimetric-cross-validation). Two important hyper-parameters in RF, namely the numbers of trees ranging from 2 to 150 and the max depth of each tree varying from 2 to 128, were adjusted to train the model. The two hyper-parameter ranges of GBR were the same as those of RF. Besides, the learning rate of the GBR model is noteworthy, which shrinks the contribution of each tree by learning rate. The most appropriate learning rate of 0.1 was determined through testing and adjustment. Too small a learning rate will slow down the convergence rate, while too high a learning rate causes the gradient to vibrate around the minimum, or even fail to converge.

2.3 Model training and evaluation

ML algorithms are to solve the minimization of loss function for the training dataset, which evaluates the quality of the model through repeated cross-validation. Every dataset was randomly split into a training dataset and a test dataset. 80% of the total data points were identified as the training dataset to train the model using the 5-fold cross-validation method, which avoids the overfitting of the trained model in the hyper-parameter optimization process. And the test dataset (20% of data points) is suitable for measuring the accuracy of the model.

The correlation coefficient (R^2) is used as a statistical measure to evaluate the degree of fitness. In general, R^2 shows the relationship between the independent variable and dependent variable. The root mean squared error (RMSE) indicates the deviation between the experimental value and predictive value. The evaluation index of the regression model performance is based on R^2 and RMSE, as shown in Eqs. (6) and (7):

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

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

where Y_i^{exp} and Y_i^{pred} represent the experimental value and the predicted value, respectively. Y_{ave}^{exp} means the average of the experimental values.

2.4 Feature importance analysis

ML models could deal with the complex relationship between the dependent variable and independent variable, and feature importance analysis can explain the contributions of input parameters in the prediction task, for any "black-box" (Islam et al. 2019). Feature importance analysis for models obtained with better predictive performance (GBR models for this study) was conducted. Based on the model prediction, it is feasible to verify the importance of the feature parameters and obtain the score of each feature. The principle is to change the value of a feature variable into a random number and calculate its influence on the accuracy of the model. The importance of all input parameters is measured in terms of the average drop precision value calculated multiple times. The higher the value, the higher importance the variable shows (Genuer et al. 2010). The value of the feature closer to zero indicates the feature parameter under consideration has a weak effect on the model prediction. The big positive value means that the feature is so important that a small change of the value would have a significant influence on the model prediction result. The feature importance only presents the influence of a certain feature on model prediction in a numerical value way, while partial dependent plots (PDP) could reflect how the feature affects the prediction, and therefore, the PDP of important features were provided.

2.5 Forward optimization (biochar engineering) and experimental verification

The forward optimization to engineer target biochar (e.g., biochar of high yield, high N content and high SSA) from three given feedstocks (namely rice husk, sawdust, and corn stalk) with known characteristics (Additional file 1: Table S1) was carried out based on the optimum model (model based on dataset #7_5). Specifically, only the pyrolysis conditions need to be optimized during forward optimization, therefore, iteration step size of 50 °C for T from 200 to 800 °C, of 30 min for RT from 30 to 120 min, and step size of 2 °C min^{-1} for HR from 2 to 10 °C min^{-1} according to the pyrolysis conditions of the dataset and engineering implementation considerations. The result of these iterations will result in vast results of Yield-char, N-char, and SSA-char corresponding to different pyrolysis conditions and feedstocks, from which the optimal solutions will be screened by comparing contours of pyrolysis conditions (T, RT, and HR) against the three targets.

Then, the optimal conditions to achieve the targets were verified by pyrolysis experiments. In brief, the pyrolysis experiments were conducted in a tube furnace, and the T, RT, and HR were selected and set according to the screened optimal solutions, with 99.999% N_2 purge gas of flow rate 100 ml min^{-1}. The yield of biochar was calculated based on the solid in a pyrolysis vessel before and after pyrolysis. N-char was measured by elemental analyzer (Elementar, Vario, Germany), in which the temperature of the combustion tube of which was 1150 °C with O_2 gas flow rate 15 ml min^{-1} and He purge gas flow rate 200 ml min^{-1}. The Brunauer–Emmett–Teller (BET) SSA-char was determined by the N_2 adsorption method with 120 °C of degassing temperature and 6 h of degassing time using a Quantasorb SI instrument (Quantachrone, USA).

3 Results and discussions

3.1 Statistical analysis of datasets

Table 1 shows all the data collected, which is the dataset before splitting into datasets #1-#7). The data size for each variable ranged from 148 (for biochemical composition) to 400 (for pyrolysis conditions). The mean values and the standard deviations (SD) were used to reflect the dispersion degree of each variable, while the minimum and maximum values displayed the scope of each variable, with quartile ranges being provided to show additional measures of location. Both lignocellulosic biomasses, such as wheat straw, polar wood, and husk which own relatively high contents of cellulose, hemicellulose, and lignin, and other biomasses, such as algae and animal manure which have a high content

Table 1 The parameter distribution and parameter scale in the whole dataset

Item		Count	Mean	SD	Min	25%	50%	75%	Max
Elemental composition[a]	C (%)	371	51.52	9.58	33.77	47.17	48.91	53.04	87.62
	H (%)	371	6.69	1.47	3.42	6.00	6.35	7.11	13.67
	N (%)	371	1.76	1.82	0.05	0.51	1.17	1.99	9.78
	O (%)	371	39.75	10.31	0.87	37.84	42.62	45.54	52.86
Proximate composition[b]	VM (%)	299	74.30	8.55	29.86	70.78	75.82	79.80	94.16
	Ash (%)	375	8.94	9.41	0.30	2.33	5.31	12.00	45.54
	FC (%)	284	17.06	5.77	3.37	13.20	17.15	20.90	37.77
Biochemical composition	Cel (%)	148	35.24	11.29	11.87	28.00	32.05	44.40	66.25
	Hem (%)	148	28.84	11.07	10.21	17.22	29.70	40.39	55.42
	Lig (%)	148	20.30	11.39	2.11	13.20	22.40	27.18	54.23
	Ext (%)	148	15.62	13.15	2.19	7.35	12.12	24.99	72.12
Pyrolysis conditions	T (°C)	400	489	150	150	400	500	600	800
	RT (min)	400	88	82	6	30	60	120	480
	HR (°C min^{-1})	400	25	47	1	10	10	20	300
Properties of biochar	Yield-char (%)	296	37.99	14.57	9.60	26.86	35.53	48.81	93.50
	N-char (%)	260	2.24	2.11	0.09	0.70	1.57	3.21	9.20
	SSA-char (m^2 g^{-1})	284	77.20	108.24	0.02	5.55	25.89	94.00	525.86

[a] On dry ash-free basis

[b] On dry basis

of extractives such as protein, were included. Pyrolysis temperatures ranged from 150 to 800 °C with most between 400 and 600 °C, and residence times varied from 6 to 480 min, with most in the range of 30–120 min. Meanwhile, 75% of the heating rate values were below 20 °C min^{-1} (mostly 10–20 °C min^{-1}), although a heating rate of up to 300 °C min^{-1} was collected. Biochar yields varied from 9.6% to 93.5%, with most being in the range of 26–48%. Meanwhile, the N content in biochar fluctuated between 0.09% and 9.20% (mostly in the range of 0.7–3.2%), and the much high N content existed in biochar from *Chlorella Vulgaris* and penicillin mycelial dreg. The SSA of biochar ranged more sparsely from 0.02 to 525.86 m^2 g^{-1}, although most values concentrated within 5–94 m^2 g^{-1} (Table 1). The high SSA values were mainly found for biochar from the softwood and hardwood (Hassan et al. 2020). The detailed information for each of datasets #1-#7 was shown in Additional file 1: Table S2. And the range of data distribution (datasets #1-#3 from Additional files 2, 3, and 4) was performed by box plots (Additional file 1: Fig. S1). As can be seen from the box plots, only a few outliers existed in these datasets, which were within the normal range of biomass properties and pyrolysis conditions. Therefore, it is not necessary to delete these outliers.

3.2 Pearson correlation coefficient

Pearson correlation coefficient (PCC) matrix, which was applied to describe the relationship between variables, is shown in Fig. 1. The magnitude of the coefficient indicates the degree to which one parameter impacts the other. Input parameters, including N content ($p<0.01$, p-value), Ash ($p<0.01$), FC ($p<0.01$), and HR ($p<0.01$), had high correlations with three target parameters, while RT showed weak correlation or irrelevancy to three targets. Other inputs showed varied correlations with the outputs. According to the PCC matrix, the Yield-char showed a negative correlation with T ($p<0.01$, $r=-0.523$). The previous research also proved that the high temperature would contribute to the rapid decomposition of lignocellulosic ingredients, leading to a decrease in biochar content (Hassan et al. 2020). The HR showed a tendency similar to T, indicating a negative correlation between HR and Yield-char ($p<0.01$, $r=-0.220$) or N-char ($p<0.01$, $r=-0.164$) but with lower correlation coefficients than those between T and Yield-char. At a high HR, biomass is more likely to depolymerize into volatile fractions, leading to lower Yield-char and N-char (Tripathi et al. 2016). Besides, it was found that the Yield-char was positively correlated with C ($p<0.05$, $r=0.123$) and N ($p<0.01$, $r=0.332$), Ash ($p<0.01$, $r=0.399$) as well as Lig ($p<0.01$, $r=0.444$). Higher Lig leads to higher Yield-char due to the high thermal stability of the aromatic monomers which are resistant to volatilization (Yang et al. 2020; Hassan et al. 2020). And the Ash of biomass is generally treated as part of biochar (not for ash-free Yield-char, this study collected

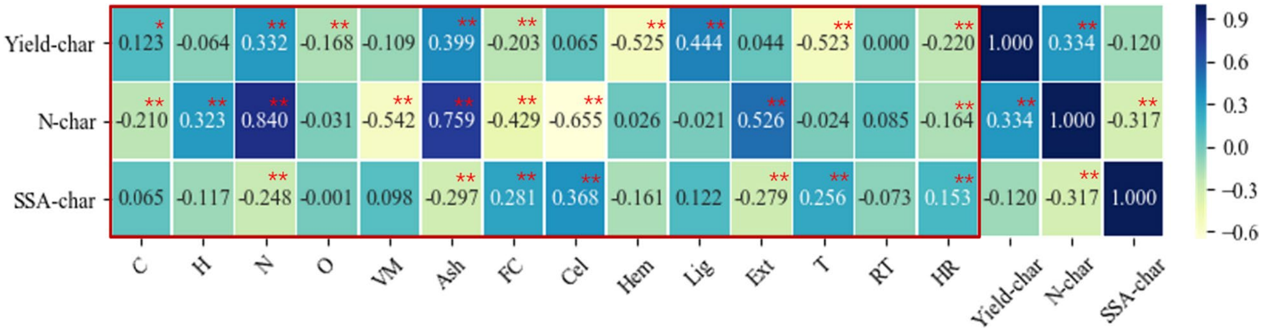


Fig. 1 Heatmap about characteristics of biomass and biochar through Pearson coefficient analysis based on the whole data collection of 400 pieces of data (* means p-value < 0.05 and ** means p-value < 0.01)

data of Yield-char with ash included), and thus it can increase Yield-char (Fahmi et al. 2008; Li et al. 2017). The Yield-char held negative correlation with O ($p < 0.01$, $r = - 0.168$), FC ($p < 0.01$, $r = - 0.203$), and Hem ($p < 0.01$, $r = - 0.525$). Hem in biomass contributes to oxygen-containing functional groups, and Hem is quite unstable and easy to decompose during pyrolysis, and thereby is negatively related to Yield-char (Hassan et al. 2020).

In terms of N-char, the N and Ash of biomass showed significant positive correlations compared to other variables, with correlation coefficients being 0.840 ($p < 0.01$) and 0.759 ($p < 0.01$), respectively. This phenomenon indicates the initial N content in biomass feedstock plays a decisive role in the N-char (Xu et al. 2021). Besides, the N-char showed a positive correlation with the Ext ($p < 0.01$) since the content of protein (N source of N-char) was included in the Ext. And the N-char had a negative correlation with C ($p < 0.01$, $r = - 0.210$), VM ($p < 0.01$, $r = - 0.542$), FC ($p < 0.01$, $r = - 0.429$), Cel ($p < 0.01$, $r = - 0.655$), and HR ($p < 0.01$, $r = - 0.164$).

As for SSA-char, T was positively correlated ($p < 0.01$, $r = 0.256$). The increase in T would lead to more porous structures, which contributes to a greater SSA (Yang and Hanping Chen 2007). This may be due to the increased micropore volume from the removal of the volatilized residual material that blocked micropores upon heating at higher T (Lee et al. 2010). In addition to T, the SSA-char had a positive correlation with FC ($p < 0.01$, $r = 0.281$), Cel ($p < 0.01$, $r = 0.368$), and HR ($p < 0.01$, $r = 0.153$), and FC and Cel should be the structures that are favorable to biochar surface and pores. It had a negative correlation with N ($p < 0.01$, $r = - 0.248$), Ash ($p < 0.01$, $r = - 0.297$), as well as Ext ($p < 0.01$, $r = - 0.279$), and Ash may block pores of biochar while N and Ext either are easy to decompose or would form N-containing functional groups that may block pores of biochar as well.

However, not all correlations can be well explained by the current knowledge. To further explore the quantitative effects of biomass characteristics on the biochar yield, N-char and SSA-char, it is important to establish suitable prediction models by different machine learning algorithms.

3.3 ML predictions

3.3.1 Hyper-parameter tuning

Figures 2a and b show the average R^2 based on RF and GBR models, respectively, for hyper-parameter tuning regarding the single-target prediction of Yield-char after 5-fold cross-validations. When the max_depth of RF was less than 4, the average R^2 showed no improvement with the increasing n_estimators (Fig. 2a). However, when the max_depth of RF was over 4, the average R^2 had a remarkable increase with the n_estimators rising from 4 to 32. The n_estimators and max_depth of 32 and 16, respectively, afforded the maximum average R^2 value (Fig. 2a). For GBR, the average R^2 of the model increased considerably when the n_estimators increased from 2 to 16 irrespective of the max_depth (Fig. 2b). The further increase of n_estimators and decrease of max_depth increased the R^2, with the optimal n_estimators and max_depth being 150 and 2, respectively. The tendencies for the effect of hyper-parameter on average R^2 of multi-target prediction were similar to those of single-target prediction, and the optimum hyper-parameters are noted in Fig. 2c, d. The optimum hyper-parameters for the other 80 ML models were obtained in the same way, and the values are provided in Additional file 1: Tables S3, S4, and S5.

3.3.2 Model evaluation

Based on the optimum hyper-parameters obtained from the hyper-parameter tuning of the 84 ML models, the whole dataset was retrained, and Additional file 1: Tables S3, S4, and S5 show each model's performance, including

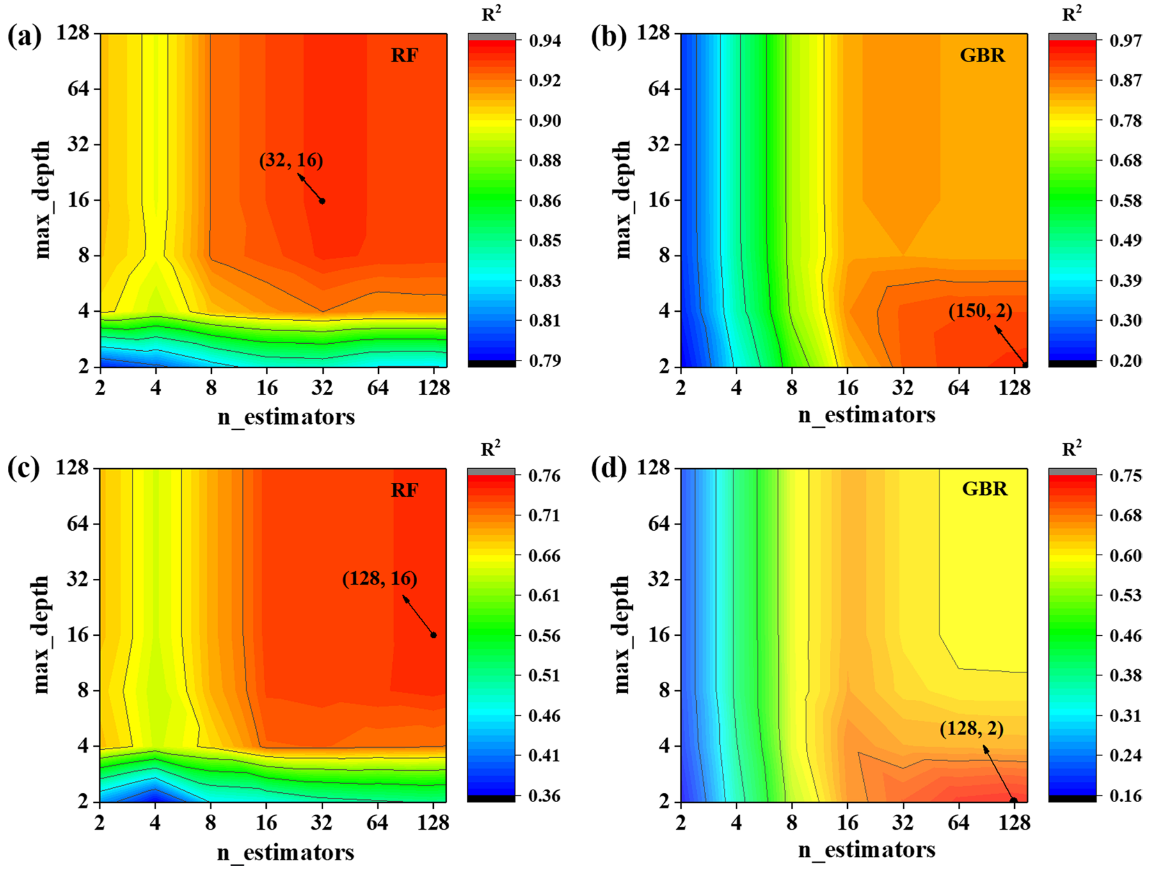


Fig. 2 Hyper-parameter tuning for the prediction of Yield-char from dataset #1_6 (**a**, **b**) and that of three targets from dataset #7_6 (**c**, **d**)

input and output details as well as R^2 and RMSE, which can be used as the evaluation criteria of model accuracy. The R^2 value closer to 1 or the smaller RMSE value for the same single-target or multi-target means the higher accuracy of the prediction model. For comparison, the test R^2 and RMSE of the 84 models (dots in plots) are presented in Fig. 3, with the average test R^2 and the average test RMSE being presented as bars.

In the single-target predictions, the test R^2 values of RF and GBR models were in the range of 0.83–0.97 and averaged at around 0.90 (test RMSE in the range of 2.32–6.68) for Yield-char from dataset #1 (Fig. 3a and Additional file 1: Table S3), indicating better performance than, if not comparable with, previously reported ML predictions for Yield-char (Table 2). For the single-target prediction of N-char and SSA-char, the test R^2 values were generally around 0.85, and the test RMSE for N-char was 0.56–0.97 and that for SSA-char was 26–81 (Fig. 3a and Additional file 1: Table S3). There were several cases obtained with test R^2 values lower than 0.80 for single-target prediction of N-char and SSA-char, with four cases being even lower than 0.70 mainly because of small data sizes or deficient variables (Fig. 3a and Additional file 1: Table S3). On the other hand, the predictive performance of multi-target tasks (Fig. 3b) was slightly better than the single-target ones (Fig. 3a); all multi-target models achieved R^2 higher than 0.70 (Fig. 3b). The co-prediction of Yield-char and N-char by GBR achieved the highest average test R^2 (~ 0.92), while the co-prediction of N-char and SSA-char received the lowest average test R^2 of ~ 0.81 by RF, and the co-prediction of three target obtained average test R^2 values of ~ 0.81 and ~ 0.84 by RF and GBR, respectively (Fig. 3b and Additional file 1: Tables S4 and S5).

Figure 3 shows that the predictive performance of GBR was generally better than RF when the average test R^2 was considered except for the single-target prediction of N-char. When the test R^2 for each sub-dataset was considered, predictive performance for datasets #n_2 was generally better than datasets #n_1, meaning the use of proximate analysis of biomass as input is better than the elemental composition of biomass. The major difference between the sub-datasets in each of datasets #1-#7 is the input variables of biomass, namely the descriptors for biomass. The use of biochemical composition as inputs

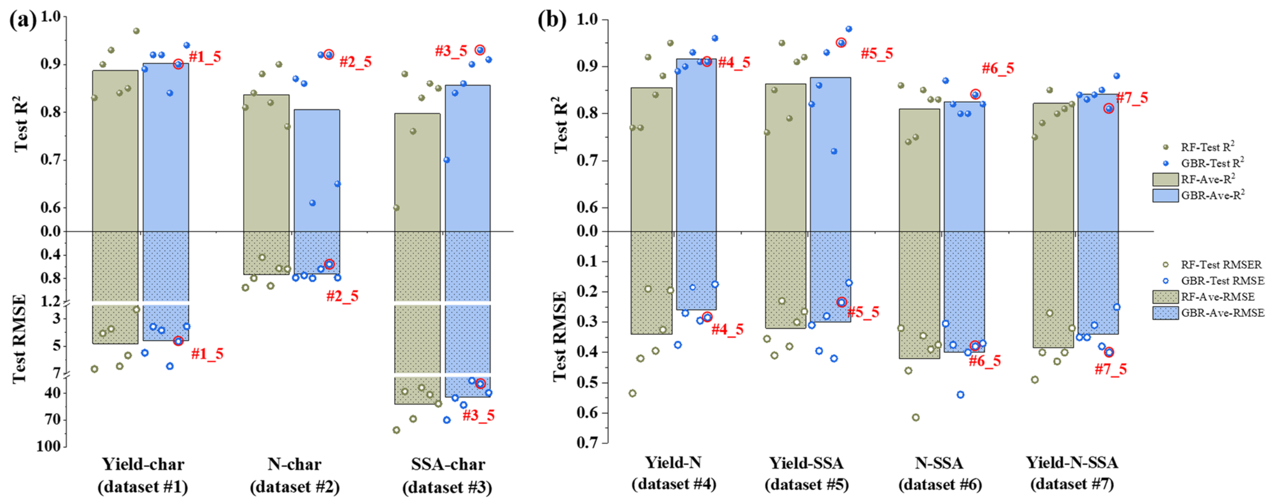


Fig. 3 Test R^2 and RMSE of the 42 ML predictions and the average test R^2 of datasets #1-#7 for **a** single-target predictions and **b** multi-target predictions based on RF and GBR (Details refer to Additional file 1: Tables S3, S4, and S5)

Table 2 Comparison of predictive performance for biochar yield in this study and in previous studies

Input parameters	Dataset points	Model	Test R^2	Test RMSE	Refs.
C, H, N, O, T, RT, and HR	282	GBR (RF)	0.89 (0.83)[a]	5.49 (6.68)[b]	This study
Ash, FC, VM, T, RT, and HR	206	GBR (RF)	0.92 (0.90)	3.57 (4.07)	This study
Cel, Hem, Lig, T, RT, and HR	133	GBR (RF)	0.92 (0.93)	3.84 (3.73)	This study
C, H, N, O, Ash, T, RT, and HR	282	GBR (RF)	0.84 (0.84)	6.48 (6.48)	This study
C, H, N, O, Ash, VM, FC, T, RT, and HR	196	GBR (RF)	0.90 (0.85)	4.66 (5.69)	This study
C, H, N, O, Ash, VM, FC, Cel, Hem, Lig, T, RT, and HR	118	GBR (RF)	0.94 (0.97)	3.56 (2.32)	This study
C, H, N, O, T, RT, and HR	91	XGB	0.73	–[c]	(Pathy et al. 2020)
C, H, N, O, Ash, FC, VM, and T	91	XGB	0.77	–	(Pathy et al. 2020)
C, H, N, O, Ash, FC, VM, T, RT, and HR	91	XGB	0.75	–	(Pathy et al. 2020)
C, H, N, O, Ash, PS, T, RT, and HR	245	RF	0.80	3.97	(Zhu et al. 2019a)
Cel, Hem, Lig, PS, T, RT, and HR	245	RF	0.85	3.50	(Zhu et al. 2019a)
C, H, N, O, Ash, Cel, Hem, Lig, PS, T, RT, and HR	245	RF	0.85	3.40	(Zhu et al. 2019a)

[a] Test R^2 based on Random Forest

[b] Test RMSE based on Random Forest

[c] Not available

PS Particle size of biomass

(datasets #n_3, including datasets #1_3, #2_3, #3_3, #4_3, #5_3, #6_3, and #7_3) afforded better predictive performance than the former two (datasets #n_1 and #n_2) for predicting Yield-char and SSA-char, but not for N-char. Nevertheless, the combination of proximate and elemental compositions of biomass (datasets #n_5) can obtain satisfactory test R^2 and RMSE values for all 7 tasks. The further incorporation of biochemical composition to the input variables (datasets #n_6) seemed also to improve the predictive performance except for the single-target task for the prediction of N-char. However, datasets #n_6 contain only limited data points, 118 for predicting Yield-char and 63–68 for the other predictions, and the models from datasets #n_6 may not be as reliable as the others.

Figures 4a–c and 5 depict the predicted values versus the actual experimental data derived from published literature in the form of scatter diagram based on the GBR models obtained from datasets #n_5 (i.e., #1_5, #2_5, #3_5, #4_5, #5_5, #6_5, and #7_5, red-line circled in Fig. 3). The input & output variables' scales and the predictive performance of the 7 models (based on #n_5) for the single-target and multi-target predictions are shown in Table 3 and Table 4, respectively, and those for models

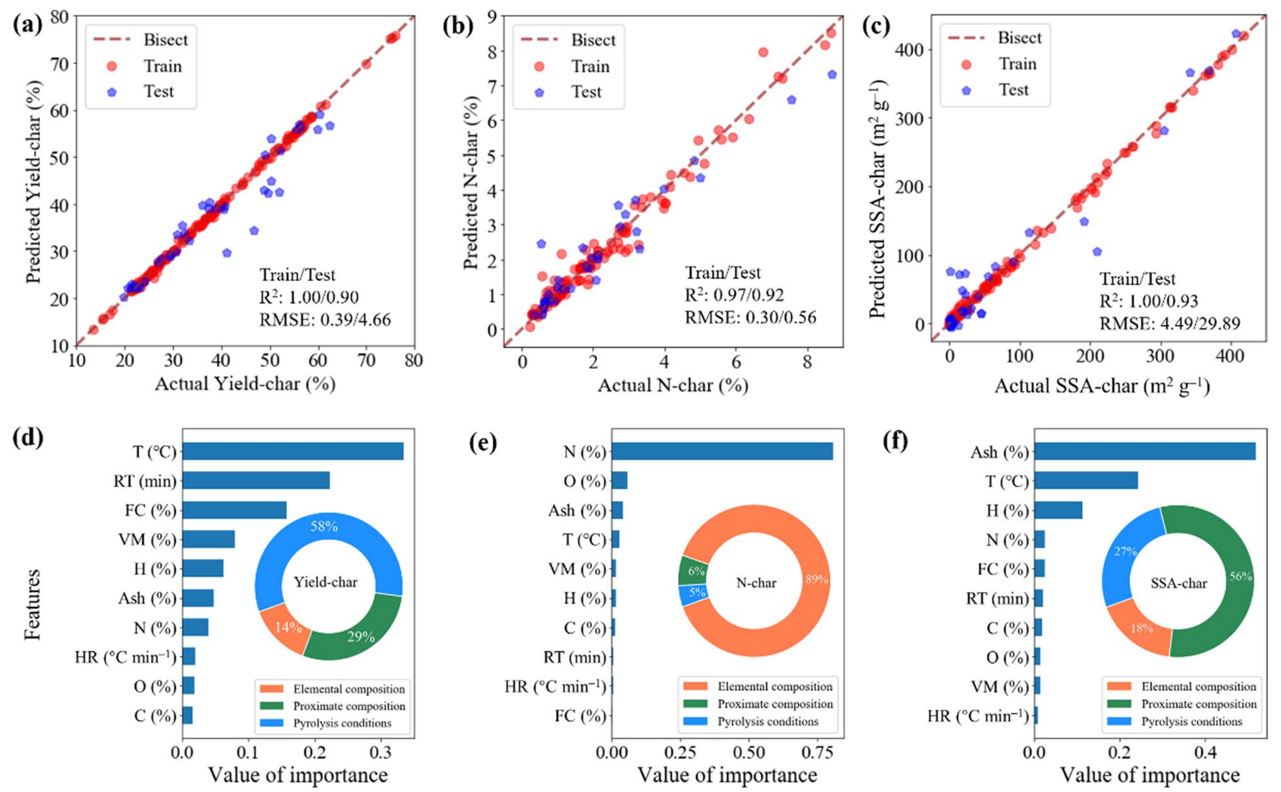


Fig. 4 Single-target prediction figures of **a** Yield-char, **b** N-char, and **c** SSA-char, and the relative contribution of each feature for **d** Yield-char, **e** N-char, and **f** SSA-char based on GBR model ((**a** and **d**), (**b** and **e**), (**c** and **f**) from dataset #1_5, #2_5, and #3_5, respectively)

based on #n_6 are shown in Additional file 1: Table S6 (not detailed in the main body). The test data points around the 45-degree line (y = x) in Figs. 4 and 5 indicate that models based on datasets #n_5 are credible in the prediction of biochar yield and properties. The distribution of data points for the training dataset was relatively more concentrated compared to the test dataset. Besides, the R^2 of the training dataset was closer to 1 and the RMSE was smaller, as compared to that of the test dataset. It indicates the predictive performance based on the GBR model presented a slight weakness in the test dataset. Figure 4a–c present that the performance of GBR based on datasets #n_5 is appropriate for single-target prediction of Yield-char, N-char, and SSA-char without normalization, and test R^2 of 0.90, 0.92, and 0.93 and test RMSE of 4.66, 0.56, and 29.89 were obtained for these three targets, respectively. For multi-target, the normalization was applied to map all output parameters to the same scale due to the order of magnitude differences of Yield-char, N-char, and SSA-char. Figure 5 provides the predictive performance for the two-target or three-target tasks under normalization, the test R^2 of 0.84–0.95 and RMSE of 0.24–0.38 were obtained for two-target tasks, while the test R^2 of 0.81 and RMSE of 0.41 were obtained for the three-target task.

3.4 Feature importance analysis of GBR models

The feature importance was applied to estimate the relative contribution of each feature parameter for GBR models predicting Yield-char (based on dataset #1_5), N-char (dataset #2_5), and SSA-char (dataset #3_5) individually (Fig. 4d–f). The elemental & proximate compositions and the pyrolysis conditions were applied to analyze the feature importance of the three targets. The feature variables were ranked according to their respective importance, and the percentage of importance for different compositions was also shown. Among all the pyrolysis parameters, T was the most important, and it accounted for 33.40% and 24.26% feature importance level during the prediction of Yield-char and SSA-char, respectively (Figs. 4d and f). And other important variables included RT, N-biomass, and Ash (Figs. 4d–f).

Pyrolysis conditions were the most important parameters for predicting Yield-char (accounting for 58% of the importance level), and T played the most important role in the prediction of Yield-char (Fig. 4d). The Yield-char decreased with increasing T, especially at 200–600 °C

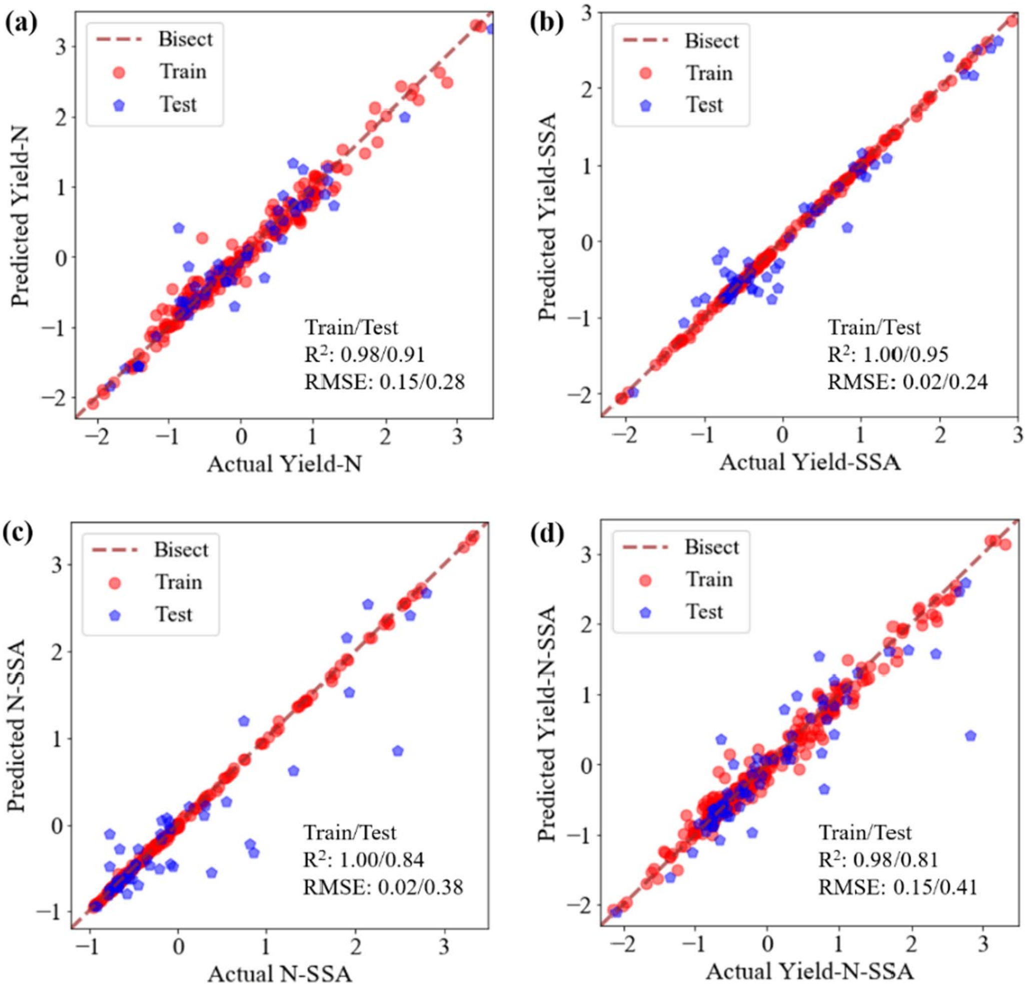


Fig. 5 Multi-target prediction figures based on GBR model for **a** Yield-char and N-char, **b** Yield-char and SSA-char, **c** N-char and SSA-char, and **d** Yield-char, N-char, and SSA-char from dataset #4_5, #5_5, #6_5, #7_5, respectively

(the left plot of Fig. 6a), which is consistent with PCC. It is because a high T would lead to the heavier decomposition of lignocellulosic components, which contributes to a lower Yield-char. T was also ranked as the first feature to predict biochar yield in previous ML studies (Zhu et al. 2019a; Li et al. 2020a). T was followed by RT; the Yield-char showed a slight increase when the RT was more than 100 min (the middle plot of Fig. 6a), although RT had no correlation with Yield-char in PCC (Fig. 1). In addition, RT showed very low importance and is negligible in many previous studies about predicting biochar yield (Zhu et al. 2019a; Li et al. 2020a; Pathy et al. 2020). The third important feature was FC, the Yield-char would decrease with incremental FC in the range of 21- 23% (the right plot of Fig. 6a). However, HR and C were the second and third features to biochar yield prediction in the aforementioned two ML studies (Zhu et al. 2019a; Li et al. 2020a). One of the reasons may be due to the differences in input variables, for example, FC was not included in the study carried out by Zhu et al. (Zhu et al. 2019a). In addition, the data difference can be another important reason since the data between this study and the previous ones are totally different.

Elemental composition dominated the prediction of N-char (accounting for 89% of the importance level), among which the N content of biomass was the dominant one (Fig. 4e). N-char increased with the incremental biomass N content (the left plot of Fig. 6b); this is because N-char only comes from biomass N (Hassan et al. 2020; Xu et al. 2021). The N-char slightly decreased as the O content of biomass rises (the middle plot of Fig. 6b), and O content also showed a low negative correlation with N-char in PCC analysis (Fig. 1). The N content of biochar decreased during pyrolysis of the biomass with high O content owning to decarboxylation and dehydration process (Gascó et al. 2018). N-char slightly decreased

Table 3 The parameter scale of the elemental composition, the proximate composition and the pyrolysis conditions for single-target cases in GBR models

Item	Dataset #1_5	Dataset #2_5	Dataset #3_5
Elemental composition[a]			
C (%)	40.02–61.57	33.77–61.57	33.77–61.57
H (%)	4.00–10.76	4.00–13.67	4.00–13.67
N (%)	0.23–9.78	0.10–8.21	0.23–8.21
O (%)	27.36–52.86	27.36–52.86	27.36–52.86
Proximate composition[b]			
VM (%)	49.36–94.16	49.36–94.16	49.36–94.16
Ash (%)	0.40–42.02	0.40–42.02	0.44–42.02
FC (%)	3.37–32.03	3.37–32.03	3.37–32.03
Pyrolysis conditions			
T (°C)	150–800	200–800	150–800
RT (min)	6–180	6–180	10–180
HR (°C min^{-1})	1–300	1–300	1–300
Properties of biochar (outputs)			
Yield-char (%)	13.50–81.66	–	–
N-char (%)	–[c]	0.26–9.20	–
SSA-char (m^2 g^{-1})	–	–	0.02–418.73
Data size			
Number of data points	196	161	183
Hyper-parameter			
N_estimators, max_depth	150, 4	150, 2	150, 4
Predictive performance			
Train R^2 (Test R^2)	1.00 (0.90)	0.97 (0.92)	1.00 (0.93)
Train RMSE (Test RMSE)	0.39 (4.66)	0.30 (0.56)	4.49 (29.89)

[a] On dry ash-free basis

[b] On dry basis

[c] Not available

(Ash < 2%) and then increased (Ash > 2%) with the incremental Ash (the right plot of Fig. 6b). Interestingly, T only ranked fourth in the model, and it seems negligible to predict N-char.

On the other hand, the proximate composition was more important than elemental composition and pyrolysis conditions (accounting for 56% of the importance level) (Fig. 4f) to SSA-char. Ash and T showed high importance to the prediction of SSA-char (Fig. 4f). Ash was more significant, and it showed a prominent negative effect on SSA-char when the ash content was smaller than 2% (the left plot of Fig. 6c). The major reason is that the pore formation is closely related to the release of volatile matter (interestingly, VM and HR were the least important features, Fig. 4f), while high Ash would result in blocking of a large number of pores, and ash itself is not porous structure and makes less contribution to SSA-char. The SSA data collected are dry basis and thus the dry basis SSA of char can be much higher than the collected SSA for high-ash biochar samples. A recent study reported that the addition of ash to biomass during pyrolysis decreased SSA-char considerably, which was mainly a result from the dilution effect of ash (Grafmüller et al. 2022). The washing of biochar to remove ash component (from ~ 16% to ~ 6.5%) can significantly increase SSA-char, for example from 81–201 to 311–493 m^2 g^{-1} (Thomas Klasson et al. 2014). However, Ash higher than 2% resulted in only limited influence on SSA-char. This means the dominant role of Ash is mainly because a certain couple of biomass samples with ash yield < 2% resulting in the production of biochar with very high SSA. For example, wood hybrid poplar and douglas fir wood with ash yield of 1.0% and 0.3% produced biochar with SSA of 181–223 m^2 g^{-1} and 478–526 m^2 g^{-1}, respectively (Suliman et al. 2016). In another study, poplar wood (*Populus alba*) with ash yield of 1.13% was pyrolyzed and produced SSA-char of 200–400 m^2 g^{-1} depending on pyrolysis conditions (Chen et al. 2016). Nevertheless, low ash yield does not necessarily mean high SSA-char; for example, biochar from water oak has SSA-char of 2.9–52.9 m^2 g^{-1} (Li et al. 2018). That is because other parameters such as T and H would also affect SSA-char considerably (Fig. 4f). SSA-char experienced a significant increase when T increased from 400 °C to 700 °C (the middle plot of Fig. 6c). During the conversion of amorphous carbon to crystalline carbons, more volatiles are removed, forming sparse regions, which bring about cracks in biochar, thus promoting formation of more pores with incremental T (Fu et al. 2009; Keiluweit et al. 2010). Hydrogen content in biomass had a slightly negative effect on SSA-char, corresponding to PCC (Fig. 1). However, the mechanism of the effect of the H element on SSA-char is unclear and needs to be further explored. The value of importance for the multi-target prediction model based on dataset #7_5 followed the trend of N > Ash > T > RT > FC, and other variables were less important (Additional file 1: Fig. S2a).

Feature importance analysis of GBR models based on datasets #1_6, #2_6, and #3_6 was also conducted. After adding biochemical compositions as inputs, all the top features changed, from T to H for Yield-char, N to C for N-char, and from Ash to FC for SSA-char (comparing Figs. 4d-f and S2b-d). However, that H, C, and FC became the top features is more likely to be a result from several particular samples according to the partial dependence plots shown in Additional file 1: Fig. S3. And the small data size of datasets #n_6 enables data from particular samples to affect the model trends considerably. Lig became the third feature, after H and T, in the prediction of Yield-char. The high content of lignin, which is the most stable component of biomass,

Table 4 The parameter scale of the elemental composition, the proximate composition and the pyrolysis conditions for multi-target cases in GBR models

Item	Dataset #4_5	Dataset #5_5	Dataset #6_5	Dataset #7_5
Elemental composition[a]				
C (%)	40.02–61.57	40.02–61.57	33.77–61.57	40.02–61.57
H (%)	4.00–10.76	4.00–10.76	4.00–13.67	4.00–10.76
N (%)	0.23–8.21	0.23–8.21	0.23–8.21	0.23–8.21
O (%)	27.36–52.86	27.36–52.86	27.36–52.86	27.36–52.86
Proximate composition[b]				
VM (%)	49.36–94.16	49.36–94.16	49.36–94.16	49.36–94.16
Ash (%)	0.40–42.02	0.44–42.02	0.44–42.02	0.44–42.02
FC (%)	3.37–32.03	3.37–32.03	3.37–32.03	3.37–32.03
Pyrolysis conditions				
T (°C)	200–800	200–800	200–800	200–800
RT (min)	6–180	10–180	10–180	10–180
HR (°C min^{-1})	1–300	1–300	1–300	1–300
Properties of biochar (outputs)				
Yield-char (%)	13.50–81.66	15.38–81.66	–	15.38–81.66
N-char (%)	0.24–8.48	–	0.24–9.20	0.24–8.48
SSA-char (m^2 g^{-1})	–[c]	0.49–418.73	0.56–418.73	0.73–418.73
Data size				
Number of data points	136	131	134	120
Hyper-parameter				
N_estimators, max_depth	150, 2	150, 4	150, 4	150, 2
Predictive performance				
Train R^2 (Test R^2)	0.98 (0.91)	1.00 (0.95)	1.00 (0.84)	0.98 (0.81)
Train RMSE (Test RMSE)	0.15 (0.28)	0.02 (0.24)	0.02 (0.38)	0.15 (0.41)

[a] On dry ash-free basis

[b] On dry basis

[c] Not available

improved Yield-char, which has been reported in previous studies (Yang et al. 2007; Tripathi et al. 2016). Theoretically, Cel and Hem can help capture N by char or oil products by Maillard reaction during pyrolysis, and their effects can be different (Leng et al. 2020b, a). When Hem was more than 40%, the N-char would increase with the incremental Hem (Additional file 1: Fig. S3b), although it had no prominent correlation with N-char in PCC (Fig. 1). On the contrary, Cel had a slightly negative effect on the N-char (Additional file 1: Fig. S3b), and it was consistent with PPC (Fig. 1). Note that the test R^2 for dataset #2_6 is only 0.65 (Fig. 3a and Additional file 1: Table S3), the effects of inputs on N-char may not be credible. Overall, the feature importance analysis of GBR models based on datasets #n_6 is not as reliable as that of datasets #n_5 because of data size limitations (only 7, 9, and 11 kinds of biomasses, and a total of 67, 68, and 118 data points because of varied pyrolysis conditions, are available for use to predict N-char, SSA-char, and Yield-char, respectively, Additional file 1: Table S3). A similar phenomenon was observed for GBR models from datasets #1_3, #2_3, and #3_3 (with biochemical composition of biomass and pyrolysis conditions as inputs) although the data points for these datasets were slightly more than #1_6, #2_6, and #3_6, respectively (Additional file 1: Fig. S4 and Table S3).

However, the feature importance results for GBR models from other datasets (#n_m, n = 1, 2, 3, m = 1, 2, 4, Additional file 1: Fig. S4) echoed those from datasets #1_5, #2_5, and #3_5 (Figs. 4d-f): T, RT, and FC were the top three features to predict Yield-char unless the feature was not included as input (i.e., FC is not in dataset #1_1 and FC is not in #1_4); N and ash were among the top three features to predict N-char (Ash is not in dataset #2_1 and N is not in #2_2); Ash and T were the top features to predict SSA-char (Ash is not in dataset #3_1, and T ranks third, after N and RT in dataset #3_1). The feature importance analysis results from the two- and three-target models are similar to those from single-target models and are not reported here. In summary, the

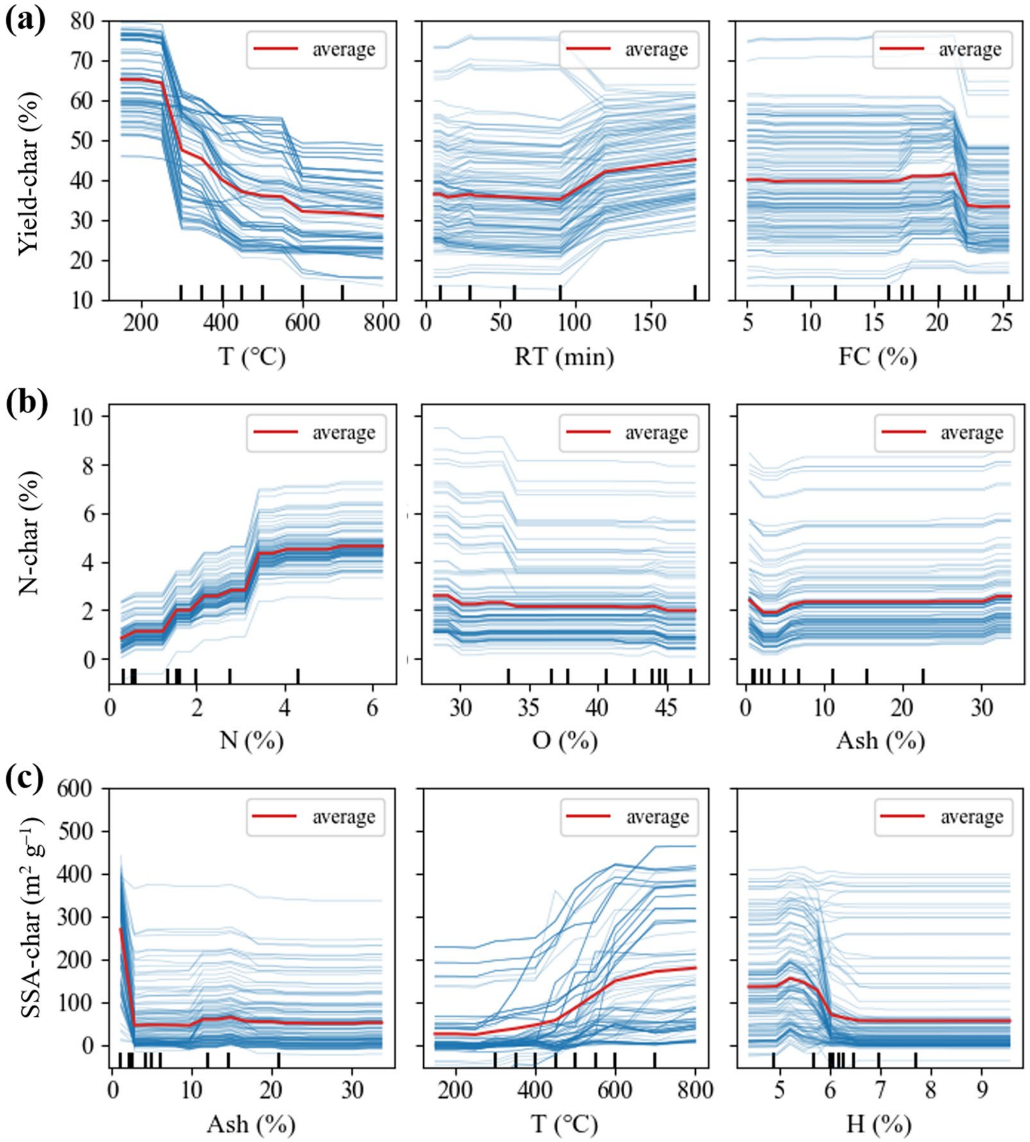


Fig. 6 Partial dependence plots of the top-three inputs to the prediction of **a** Yield-char **b** N-char, and **c** SSA-char based on improved GBR model developed from dataset #1_5, #2_5, #3_5, respectively

feature importance results analyzed according to datasets #n_m (m = 1, 2, 4, 5) are credible and may be used to help engineer Yield-char, N-char, and/or SSA-char individually or simultaneously.

3.5 Biochar engineering and experimental validations

Based on the optimal GBR model from dataset #7_5, T, RT, and HR were optimized to obtain the optimal pyrolysis conditions for producing the desired biochar, namely biochar with high N content or low N content and high SSA. Figures 7a, b, and c show the contours of the iteration to find optimal pyrolysis conditions (i.e., T, RT, and HR). When T, HR and RT were 600 °C, 6 °C min^{-1} and 90 min, biochar from rice husk (Run 1) showed high N content and high SSA (Yield-char of 32.69%, N-char of 1.12, and SSA-char of 98.56, Figs. 7 a-c). And When T, HR and RT were 800 °C, 8 °C min^{-1} and 30 min, biochar from rice husk (Run 2) showed low N content and high SSA (Yield-char of 32.17%, N-char of 0.75% and SSA-char of 185.07 m^2 g^{-1}, Figs. 7a-c). The optimal schemes of sawdust and corn stalk are shown in Additional file 1: Fig. S5.

A total of four optimal solutions for the desired biochar were conducted according to N-char and SSA-char (Additional file 1: Table S1). Differences between the

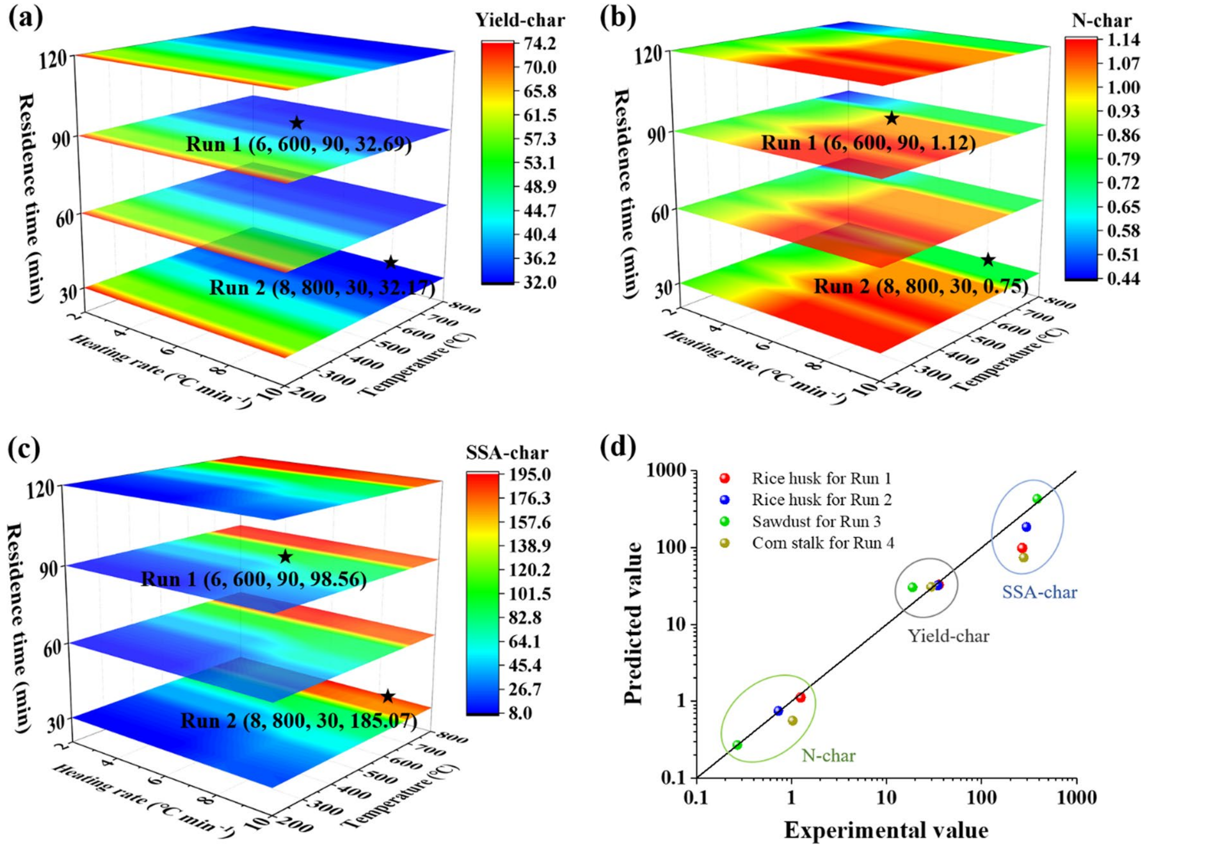


Fig. 7 The optimal schemes of three target parameters derived from **a**–**c** rice husk and **d** the experimental verification of four optimal solutions

experimental verification and GBR predicted values of the three targets were analyzed. The experimental validations were acceptable but not completely satisfactory. The experimental and predicted values for Yield-char or N-char were similar, meaning good validation performance, while those for SSA were considerably different for some cases (Fig. 7d). The large variation may be because of the large RMSE of the prediction model (should be similar to RMSE of model based on dataset #3_5, which is 29.89, Fig. 4c). Besides, the good validation for Yield-char and N-char may lead to a bad performance of SSA-char due to weight coefficients of the three targets (equal weight for the three in this study), and vice versa, and the trade-offs (weight coefficients) optimization needs further investigation.

3.6 Implications

The N-char and SSA-char are important properties for biochar as a carbonaceous material in a wide range of applications. The prediction model and the use of the model for experimentally engineering biochar properties can facilitate the screening of biomass feedstock and pyrolysis parameters for the production of target biochar materials. This study provides a new way, instead of the trial-and-error method which is time-, cost-, and labor-consuming and ineffective, to material engineering and biomass (biowaste) valorization. Other properties of biochar, such as cation exchange capacity, which is the dominant property of biochar contributing to heavy metal adsorption (Zhu et al. 2019b), and total pore volume and pore distribution, which are the most important features to predict CO_2 capture capacity (Zhang et al. 2019; Zhu et al. 2020), may also be predicted and optimized in a similar manner.

Adding more input variables does not necessarily improve predictive performance. Some inputs may be redundant and can lower the predictive performance and weaken the model generalization ability and robustness. Therefore, feature importance and PCC or other techniques may be used to screen only meaningful variables to improve predictive performance and reduce the calculation cost (more variables mean higher cost). In addition, data availability has a great influence on the predictive performance and the ranking of feature importance. Input variables with small data size or limited coverage (e.g., the biochemical composition of biomass in this study) may bias the model prediction and the feature importance level of variables, and result in a model with low generalization ability, and such variables should be considered seriously.

ML-basis multi-target prediction and optimization can deal with trade-off problems effectively. The functional relationship between multiple targets and the weight coefficient for each target can help find the optimum recipe by ML algorithms to achieve the prediction of multiple targets and find the complementary points (optimal solutions) to obtain favorable trade-offs. For example, the trade-offs between high Yield-char (negatively correlated with T, Fig. 6a) and high SSA-char (positively correlated with T, Fig. 6c) can be balanced by weight coefficient mediation. Using equal weight coefficients for the three targets in this study did not yield with completely satisfactory results, and more studies should be carried out in this direction.

This study also has other limitations:

i) The quantity and quality of the data collected are not completely satisfactory, and model bias may exist due to the fact that the limited available particular cases may have a significant influence on model prediction, e.g., the biochemical composition of biomass in this study.

ii) The dataset differences between different models make it hard to compare the predictive performance of models strictly, and datasets with full and uniform information can help increase the model comparability in addition to the model accuracy.

iii) The input variables should be expanded (e.g., molecular ratio H/C and O/C and particle size of biomass feed, pyrolysis method (Meyer et al. 2011; Cha et al. 2016), and SSA analysis method (Maziarka et al. 2021)), and the target variables (output) may be replaced (e.g., daf-basis SSA and particular N-containing functional groups such as pyridinic-N, pyrrolic-N, and quaternary-N/graphic-N Ebikade et al. 2020; Xu et al. 2021)) to obtain more closely correlated variables or more useful properties; on the other hand, other indicators or properties (e.g., elemental composition and molecular ratio as well as proximate composition and other properties such as pH) of biochar can be used to predict SSA-char with higher accuracy, although in this way, the model cannot be used directly to instruct experimental studies.

iv) High SSA of biochar is generally obtained by physical and/or chemical activation either by adding activating agent during or after pyrolysis (Gao et al. 2020), but it was not considered in this study; future study should be exerted to predict and optimize SSA-char by including activation parameters.

4 Conclusions

The prediction of Yield-char, N-char, and SSA-char was conducted by using ML methods such as RF and GBR models. GBR outperformed RF for most predictions. When input parameters include elemental and proximate composition as well as pyrolysis conditions, the test R^2 values for the single-target and multi-target GBR models are 0.90–0.95 except for the two-target prediction of Yield-char and SSA-char, which has a test R^2 of 0.84 and the three-target prediction model, which has a test R^2 of 0.81. Taking into account of R^2 and RMSE values, the prediction accuracies of these models are acceptable. PCC between variables and the feature importance of these models indicates that T, RT, and FC are the top three features to predict Yield-char; N and ash are the most important features to predict N-char; and Ash and T are the two significant features to predict SSA-char. Yield-char, SSA-char, and N-char were optimized and engineered simultaneously by a multi-target ML prediction model with experimental verification. The experimental and ML predicted values for Yield-char, N-char and SSA-char were close, meaning the success of using ML to aid the prediction and optimization of biochar properties. This study provided new insight into biochar designing and valorization of biomass or biowastes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s42773-022-00183-w.

Additional file 1. Table S1. The optimal solution and corresponding predicted and experimental outputs from ML models. **Table S2**. Detailed parameter distribution and parameter scale in the seven datasets. **Table S3**. Single-task prediction for biochar yield, N content and specific surface area of biochar based on RF and GBR. **Table S4**. Multi-task prediction for biochar yield and N content of biochar, N content and specific surface area, biochar yield and specific surface area based on RF and GBR. **Table S5**. Multi-task prediction for biochar yield, N content in biochar and specific surface area of biochar based on RF and GBR. **Table S6**. Detailed parameter distribution and parameter scale in the seven sub-datasets. **Fig. S1**. The data distribution of biomass properties and biochar properties from 400 samples. **Fig. S2**. The relative contribution of each feature for (a) Yield-char, N-char and SSA-char with all inputs except biochemical composition (b) Yield-char with all inputs (c) N-char with all inputs, (d) SSA-char with all inputs based on GBR model. **Fig. S3**. Partial dependence plots of the top-four inputs to the prediction of (a) Yield-char (from dataset #1_6), (b) N-char (from dataset #2_6) and (c) SSA-char (from dataset #3_6) based on improved GBR model developed. **Fig. S4**. The relative contribution of each feature for (a) Yield-char (b) N-char, (c) SSA-char based on GBR model. **Fig. S5**. The optimal schemes of three target parameters derived from (a-c) sawdust and (d-f) corn stalk.

Additional file 2. dataset #1 for predicting biochar yield.

Additional file 3. dataset #2 for predicting N content of biochar.

Additional file 4. dataset #3 for predicting specific surface area of biochar.

Additional file 5. dataset #4 for predicting biochar yield and N content of biochar.

Additional file 6. dataset #5 for predicting biochar yield and specific surface area of biochar.

Additional file 7. dataset #6 for predicting N content and specific surface area of biochar.

Additional file 8. dataset #7 for predicting biochar yield, N content and specific aurface area of biochar.

Additional file 9. dataset for review.

Acknowledgements

We appreciate Miss. Yaoting Deng and Mr. Tong Li for assisting program implementation.

Author contributions

The first draft was written by Lijian Leng and Lihong Yang. Supervision, Conceptualization, and Project administration were addressed by Lijian Leng. Data collection and Methodology were performed by Zejian Ai and Weijin Zhang. Software, Formal analysis and Visualization were performed by Lihong Yang. Resources, Validation were performed by Xinni Lei. Writing -Review & Editing was performed by Hao Zhan, Zequn Yang, Jianping Yang, Hailong Li, and Xingzhong Yuan. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (2021YFE0104900), the National Natural Science Foundation of China (51906247), Hunan Provincial Natural Science Foundation of China (2022JJ20064), and the Science and Technology Innovation Program of Hunan Province (2021RC4005).

Data availability

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author details

[1]School of Energy Science and Engineering, Central South University, Changsha 410083, Hunan, People's Republic of China. [2]College of Environmental Science and Engineering, Hunan University, Changsha 410083, Hunan, People's Republic of China.

Received: 28 June 2022 Accepted: 26 September 2022
Published online: 29 November 2022

References

Amalina F, Razak ASA, Krishnan S et al (2022) A comprehensive assessment of the method for producing biochar, its characterization, stability, and potential applications in regenerative economic sustainability—a review. Clean Mater 3:100045. https://doi.org/10.1016/j.clema.2022.100045

Cai J, Xu K, Zhu Y et al (2020) Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Appl Energy 262:114566. https://doi.org/10.1016/j.apenergy.2020.114566

Cao H, Xin Y, Yuan Q (2016) Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. Bioresour Technol 202:158–164. https://doi.org/10.1016/j.biortech.2015.12.024

Cao H, Wu X, Syed-Hassan SSA et al (2020) Characteristics and mechanisms of phosphorous adsorption by rape straw-derived biochar functionalized with calcium from eggshell. Bioresour Technol 318:124063. https://doi.org/10.1016/j.biortech.2020.124063

Cao H, Milan YJ, Mood SH et al (2021) A novel elemental composition based prediction model for biochar aromaticity derived from machine learning. Artif Intell Agric 5:133–141. https://doi.org/10.1016/j.aiia.2021.06.002

Cha JS, Park SH, Jung SC et al (2016) Production and utilization of biochar: a review. J Ind Eng Chem 40:1–15. https://doi.org/10.1016/j.jiec.2016.06.002

Chen D, Li Y, Cen K et al (2016) Pyrolysis polygeneration of poplar wood: Effect of heating rate and pyrolysis temperature. Bioresour Technol 218:780–788. https://doi.org/10.1016/j.biortech.2016.07.049

Chen J, Wang P, Ding L et al (2021) The comparison study of multiple biochar stability assessment methods. J Anal Appl Pyrolysis 156:105070. https://doi.org/10.1016/j.jaap.2021.105070

Chen J, Ding L, Wang P et al (2022) The estimation of the higher heating value of biochar by data-driven modeling. J Renew Mater 10:1555–1574. https://doi.org/10.32604/jrm.2022.018625

Ebikade EO, Wang Y, Samulewicz N et al (2020) Active learning-driven quantitative synthesis–structure–property relations for improving performance and revealing active sites of nitrogen-doped carbon for the hydrogen evolution reaction. React Chem Eng 5:2134–2147. https://doi.org/10.1039/D0RE00243G

Fahmi R, Bridgwater AV, Donnison I et al (2008) The effect of lignin and inorganic species in biomass on pyrolysis oil yields, quality and stability. Fuel 87:1230–1240. https://doi.org/10.1016/j.fuel.2007.07.026

Fu P, Hu S, Sun L et al (2009) Structural evolution of maize stalk/char particles during pyrolysis. Bioresour Technol 100:4877–4883. https://doi.org/10.1016/j.biortech.2009.05.009

Gao Y, Yue Q, Gao B, Li A (2020) Insight into activated carbon from different kinds of chemical activating agents: a review. Sci Total Environ 746:141094. https://doi.org/10.1016/j.scitotenv.2020.141094

Gascó G, Paz-Ferreiro J, Álvarez ML et al (2018) Biochars and hydrochars prepared by pyrolysis and hydrothermal carbonisation of pig manure. Waste Manag 79:395–403. https://doi.org/10.1016/j.wasman.2018.08.015

Genuer R, Poggi J-M, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett 31:2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Grafmüller J, Böhm A, Zhuang Y et al (2022) Wood ash as an additive in biomass pyrolysis: effects on biochar yield, properties, and agricultural performance. ACS Sustain Chem Eng 10:2720–2729. https://doi.org/10.1021/acssuschemeng.1c07694

Guo N, Li M, Wang Y et al (2016) Soybean root-derived hierarchical porous carbon as electrode material for high-performance supercapacitors in ionic liquids. ACS Appl Mater Interfaces 8:33626–33634. https://doi.org/10.1021/acsami.6b11162

Hassan M, Liu Y, Naidu R et al (2020) Influences of feedstock sources and pyrolysis temperature on the properties of biochar and functionality as adsorbents: a meta-analysis. Sci Total Environ 744:140714. https://doi.org/10.1016/j.scitotenv.2020.140714

He X, Zheng N, Hu R et al (2021) Hydrothermal and pyrolytic conversion of biomasses into catalysts for advanced oxidation treatments. Adv Funct Mater 31:2006505. https://doi.org/10.1002/adfm.202006505

Huang J, Zhao B, Liu T et al (2019) Wood-derived materials for advanced electrochemical energy storage devices. Adv Funct Mater 29:1902255. https://doi.org/10.1002/adfm.201902255

Huo W, Li W, Zhang Z et al (2021) Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection. Energy Convers Manag 243:114367. https://doi.org/10.1016/j.enconman.2021.114367

Islam SR, Eberle W, Bundy S, Ghafoor SK (2019) Infusing domain knowledge in AI-based "black box" models for better explainability with application in bankruptcy prediction Sheikh. https://doi.org/10.48550/arXiv.1905.11474

Keiluweit M, Nico PS, Johnson MG, Kleber M (2010) Dynamic molecular structure of plant biomass-derived black carbon (biochar). Environ Sci Technol 44:1247–1253. https://doi.org/10.1021/es9031419

Lee JW, Kidder M, Evans BR et al (2010) Characterization of biochars produced from cornstovers for soil amendment. Environ Sci Technol 44:7970–7974. https://doi.org/10.1021/es101337x

Leng L, Huang H, Li H et al (2019) Biochar stability assessment methods: A review. Sci Total Environ 647:210–222. https://doi.org/10.1016/j.scitotenv.2018.07.402

Leng L, Xu S, Liu R et al (2020a) Nitrogen containing functional groups of biochar: An overview. Bioresour Technol 298:122286. https://doi.org/10.1016/j.biortech.2019.122286

Leng L, Yang L, Chen J et al (2020b) A review on pyrolysis of protein-rich biomass: Nitrogen transformation. Bioresour Technol 315:123801. https://doi.org/10.1016/j.biortech.2020.123801

Leng L, Xiong Q, Yang L et al (2021a) An overview on engineering the surface area and porosity of biochar. Sci Total Environ 763:144204. https://doi.org/10.1016/j.scitotenv.2020.144204

Leng L, Yang L, Leng S et al (2021b) A review on nitrogen transformation in hydrochar during hydrothermal carbonization of biomass containing nitrogen. Sci Total Environ 756:143679. https://doi.org/10.1016/j.scitotenv.2020.143679

Leng L, Liu R, Xu S et al (2022) An overview of sulfur-functional groups in biochar from pyrolysis of biomass. J Environ Chem Eng 10:107185. https://doi.org/10.1016/j.jece.2022.107185

Li W, Dang Q, Brown RC et al (2017) The impacts of biomass properties on pyrolysis yields, economic and environmental performance of the pyrolysis-bioenergy-biochar platform to carbon negative energy. Bioresour Technol 241:959–968. https://doi.org/10.1016/j.biortech.2017.06.049

Li S, Barreto V, Li R et al (2018) Nitrogen retention of biochar derived from different feedstocks at variable pyrolysis temperatures. J Anal Appl Pyrolysis 133:136–146. https://doi.org/10.1016/j.jaap.2018.04.010

Li J, Pan L, Suvarna M et al (2020a) Fuel properties of hydrochar and pyrochar: Prediction and exploration with machine learning. Appl Energy 269:115166. https://doi.org/10.1016/j.apenergy.2020.115166

Li J, Pan L, Suvarna M, Wang X (2021a) Machine learning aided supercritical water gasification for H2-rich syngas production with process optimization and catalyst screening. Chem Eng J. https://doi.org/10.1016/j.cej.2021.131285

Li J, Zhang W, Liu T et al (2021b) Machine learning aided bio-oil production with high energy recovery and low nitrogen content from hydrothermal liquefaction of biomass with experiment verification. Chem Eng J 425:130649. https://doi.org/10.1016/j.cej.2021.130649

Li J, Pan L, Suvarna M, et al (2020b) Machine learning prediction of syngas composition of hydrothermal gasification from wet organic wastes. Energy-proceedings

Liu W-J, Jiang H, Yu H-Q (2015) Development of biochar-based functional materials: toward a sustainable platform carbon material. Chem Rev 115:12251–12285. https://doi.org/10.1021/acs.chemrev.5b00195

Liu B, Tang C, Zhao Y et al (2022) Toxicological effect assessment of aged biochar on Escherichia coli. J Hazard Mater 436:129242. https://doi.org/10.1016/j.jhazmat.2022.129242

Maulana Kusdhany MI, Lyth SM (2021) New insights into hydrogen uptake on porous carbon materials via explainable machine learning. Carbon N Y 179:190–201. https://doi.org/10.1016/j.carbon.2021.04.036

Maziarka P, Wurzer C, Arauzo PJ et al (2021) Do you BET on routine? The reliability of N2 physisorption for the quantitative assessment of biochar's surface area. Chem Eng J. https://doi.org/10.1016/j.cej.2021.129234

Meyer S, Glaser B, Quicker P (2011) Technical, economical, and climate-related aspects of biochar production technologies: a literature review. Environ Sci Technol 45:9473–9483. https://doi.org/10.1021/es201792c

Palansooriya KN, Li J, Dissanayake PD et al (2022) Prediction of soil heavy metal immobilization by biochar using machine learning. Environ Sci Technol 56:4187–4198. https://doi.org/10.1021/acs.est.1c08302

Parthasarathy P, Al-Ansari T, Mackey HR et al (2022) A review on prominent animal and municipal wastes as potential feedstocks for solar pyrolysis for biochar production. Fuel 316:123378. https://doi.org/10.1016/j.fuel.2022.123378

Pathy A, Meher S, P B, (2020) Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Res 50:102006. https://doi.org/10.1016/j.algal.2020.102006

Qiu B, Tao X, Wang H et al (2021) Biochar as a low-cost adsorbent for aqueous heavy metal removal: A review. J Anal Appl Pyrolysis 155:105081. https://doi.org/10.1016/j.jaap.2021.105081

Rzychoń M, Żogała A, Róg L (2020) An Interpretable Extreme Gradient Boosting Model to Predict Ash Fusion Temperatures. Minerals 10:487. https://doi.org/10.3390/min10060487

Sigmund G, Gharasoo M, Hüffer T, Hofmann T (2020) Deep Learning Neural Network Approach for Predicting the Sorption of Ionizable and Polar Organic Pollutants to a Wide Range of Carbonaceous Materials. Environ Sci Technol 54:4583–4591. https://doi.org/10.1021/acs.est.9b06287

Suliman W, Harsh JB, Abu-Lail NI et al (2016) Influence of feedstock source and pyrolysis temperature on biochar bulk and surface properties. Biomass Bioenerg 84:37–48. https://doi.org/10.1016/j.biombioe.2015.11.010

Tan D, Suvarna M, Shee Tan Y et al (2021) A three-step machine learning framework for energy profiling, activity state prediction and production estimation in smart process manufacturing. Appl Energy 291:116808. https://doi.org/10.1016/j.apenergy.2021.116808

Thomas Klasson K, Uchimiya M, Lima IM (2014) Uncovering surface area and micropores in almond shell biochars by rainwater wash. Chemosphere 111:129–134. https://doi.org/10.1016/j.chemosphere.2014.03.065

Tripathi M, Sahu JN, Ganesan P (2016) Effect of process parameters on production of biochar from biomass waste through pyrolysis: A review. Renew Sustain Energy Rev 55:467–481. https://doi.org/10.1016/j.rser.2015.10.122

Wang J, Wang S (2019) Preparation, modification and environmental application of biochar: A review. J Clean Prod 227:1002–1022. https://doi.org/10.1016/j.jclepro.2019.04.282

Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic 52:394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

Xu S, Chen J, Peng H et al (2021) Effect of biomass type and pyrolysis temperature on nitrogen in biochar, and the comparison with hydrochar. Fuel 291:120128. https://doi.org/10.1016/j.fuel.2021.120128

Yang H, Hanping Chen RY (2007) The influence of temperature on biomass pyrolysis in fixed bed. Acta Energlae Solaris Sin 28:1152–1157. https://doi.org/10.3321/j.issn:0254-0096.2007.10.019

Yang H, Yan R, Chen H et al (2007) Characteristics of hemicellulose, cellulose and lignin pyrolysis. Fuel 86:1781–1788. https://doi.org/10.1016/j.fuel.2006.12.013

Yang X, Kang K, Qiu L et al (2020) Effects of carbonization conditions on the yield and fixed carbon content of biochar from pruned apple tree branches. Renew Energy 146:1691–1699. https://doi.org/10.1016/j.renene.2019.07.148

Yang L, Wang G, Liu T et al (2022) Experimental and ReaxFF molecular dynamic study on pyrolysis of phenylalanine. Fuel 324:124690. https://doi.org/10.1016/j.fuel.2022.124690

Yuan X, Suvarna M, Low S et al (2021) Applied machine learning for prediction of CO2 adsorption on biomass waste-derived porous carbons. Environ Sci Technol 55:11925–11936. https://doi.org/10.1021/acs.est.1c01849

Zhang Z, Schott JA, Liu M et al (2019) Prediction of carbon dioxide adsorption via deep learning. Angew Chemie Int Ed 58:259–263. https://doi.org/10.1002/anie.201812363

Zhao Y, Li Y, Fan D et al (2021) Application of kernel extreme learning machine and Kriging model in prediction of heavy metals removal by biochar. Bioresour Technol 329:124876. https://doi.org/10.1016/j.biortech.2021.124876

Zhao Y, Fan D, Li Y, Yang F (2022) Application of machine learning in predicting the adsorption capacity of organic compounds onto biochar and resin. Environ Res 208:112694. https://doi.org/10.1016/j.envres.2022.112694

Zhu X, Li Y, Wang X (2019a) Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol 288:121527. https://doi.org/10.1016/j.biortech.2019.121527

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

Zhu X, Tsang DCW, Wang L et al (2020) Machine learning exploration of the critical factors for CO2 adsorption capacity on porous carbon materials at different pressures. J Clean Prod 273:122915. https://doi.org/10.1016/j.jclepro.2020.122915

Zhu X, Wan Z, Tsang DCW et al (2021) Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. Chem Eng J. https://doi.org/10.1016/j.cej.2020.126782

Zhu X, He M, Sun Y et al (2022) Insights into the adsorption of pharmaceuticals and personal care products (PPCPs) on biochar and activated carbon with the aid of machine learning. J Hazard Mater 423:127060. https://doi.org/10.1016/j.jhazmat.2021.127060