


ORIGINAL RESEARCH

Open Access



A novel intelligent system based on machine learning for hydrochar multi-target prediction from the hydrothermal carbonization of biomass

Weijin Zhang¹, Junhui Zhou¹, Qian Liu¹, Zhengyong Xu³, Haoyi Peng¹, Lijian Leng^{1,2*} and Hailong Li^{1*} 

Abstract

Hydrothermal carbonization (HTC) is a thermochemical conversion technology to produce hydrochar from wet biomass without drying, but it is time-consuming and expensive to experimentally determine the optimal HTC operational conditions of specific biomass to produce desired hydrochar. Therefore, a machine learning (ML) approach was used to predict and optimize hydrochar properties. Specifically, biochemical components (proteins, lipids, and carbohydrates) of biomass were predicted and analyzed first via elementary composition. Then, accurate single-biomass (no mixture) based ML multi-target models (average $R^2 = 0.93$ and RMSE = 2.36) were built to predict and optimize the hydrochar properties (yield, elemental composition, elemental atomic ratio, and higher heating value). Biomass composition (elemental and biochemical), proximate analyses, and HTC conditions were inputs herein. Interpretation of the model results showed that ash, temperature, and the N and C content of biomass were the most critical factors affecting the hydrochar properties, and that the relative importance of biochemical composition (25%) for the hydrochar was higher than that of operating conditions (19%). Finally, an intelligent system was constructed based on a multi-target model, verified by applying it to predict the atomic ratios (N/C, O/C, and H/C). It could also be extended to optimize hydrochar production from the HTC of single-biomass samples with experimental validation and to predict hydrochar from the co-HTC of mixed biomass samples reported in the literature. This study advances the field by integrating predictive modeling, intelligent systems, and mechanistic insights, offering a holistic approach to the precise control and optimization of hydrochar production through HTC.

Highlights

- Biochemical components of biomass were first predicted by elemental composition.
- The multi-target ML model accurately predicted hydrochar with $R^2 = 0.93$.
- The ash, T , N, and C were the most critical factors affecting hydrochar properties.
- An online intelligent system based on optimal models was posted and verified.

Handling editor: Hailong Wang

*Correspondence:

Lijian Leng

lljchs@126.com

Hailong Li

hailongli18@gmail.com

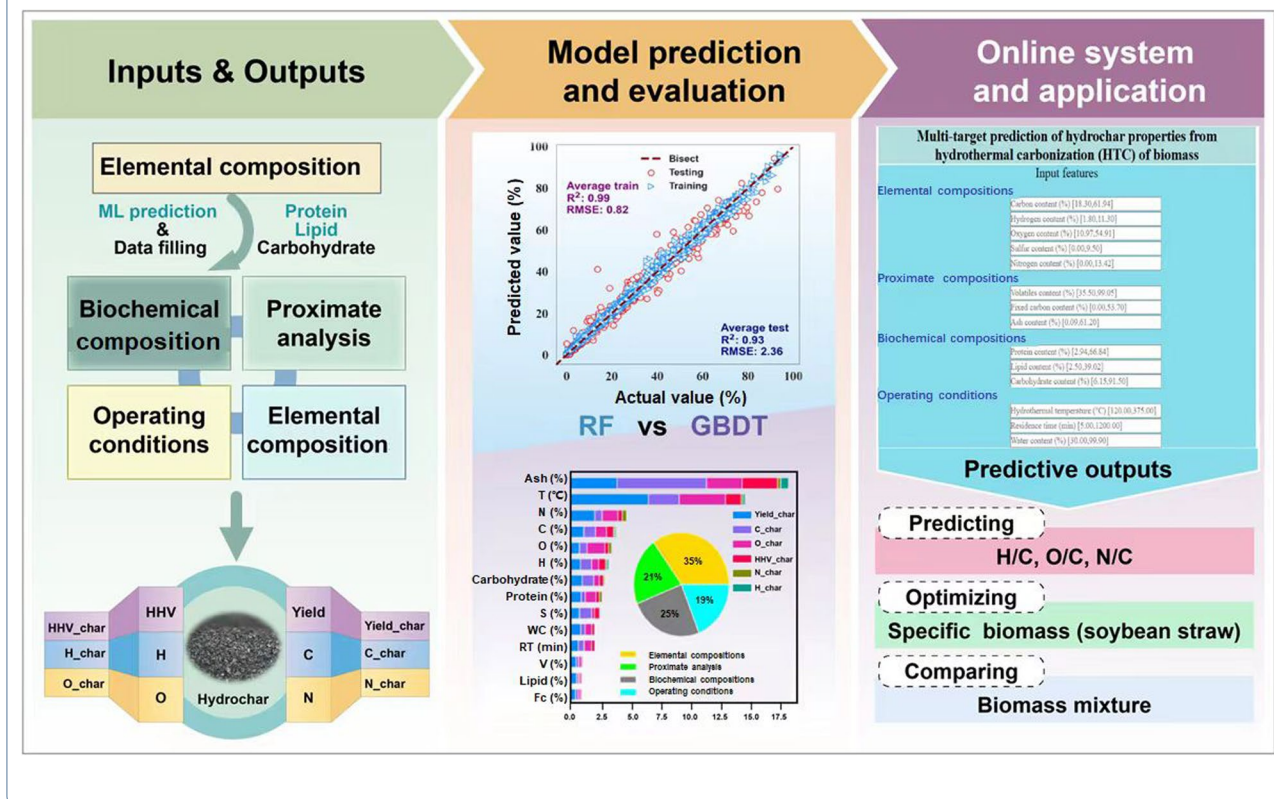
Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Keywords Biomass, Hydrothermal carbonization, Hydrochar, Machine learning, Intelligent prediction system

Graphical Abstract



1 Introduction

Rapid industrialization and economic growth have led to an energy consumption rate that far exceeds its production rate. Most current energy demands are met by fossil fuels, which has resulted in global climate change, energy crisis, and environmental pollution (Zhang et al. 2023a). Therefore, it is necessary to develop renewable energy. Biomass is considered the only carbon-based renewable energy source (Seo et al. 2022) that is abundant and has a high energy potential. Biomass (especially biowaste) is the most commonly used renewable energy source because it is carbon neutral, environmentally friendly, and economically viable.

Hydrothermal carbonization (HTC), which uses subcritical water as the reaction solvent, is a promising biomass treatment and recycling method for converting wet biomass into hydrochar and chemicals without the need for pre-drying. The properties of hydrochar are affected by various factors, such as the composition of biomass feedstock and reaction parameters (Zhang et al. 2023b). Different biomass combinations have been studied to

explore the effects of biomass types and reaction parameters (temperature, reaction time, and water content) on the hydrochar yield and fuel characteristics. The results have shown that temperature is the most important factor affecting hydrochar yield and energy density (Toptas Tag et al. 2018). Compared with temperature, the residence time and water content have similar but weaker effects on the yield of solid products (Xu et al. 2018). The introduction of a catalyst can directionally regulate the deoxygenation and denitrification reactions to improve the yield and quality of hydrochar (Marzbali et al. 2021; Xu et al. 2022), thereby generating functional carbon-based materials. Traditional approaches mainly measure the physical and chemical properties of hydrochar products and HTC conditions, but these experimental approaches are expensive and time-consuming.

Machine learning (ML) methods can address the above-mentioned problems by predicting and regulating the mining of experimental data to assist in the production and optimization of biofuels, chemicals, and materials in the HTC system (Li et al. 2023). Recently, several

researchers have used ML to investigate the HTC process of biomass, and their current limitations are summarized in Table 1. For example, the yield and energy recovery of hydrochar can be predicted (average $R^2=0.90$) by the random forest (RF) and artificial neural network (ANN) models (Li et al. 2020b; Gao et al. 2022), but these studies merely explored the impact of the biomass components on hydrochar. The fuel and structural properties of hydrochar have also been predicted for some specific types of biomass (e.g., sewage sludge, food waste, manure Mu et al. 2022; Zhu et al. 2023), or kitchen waste (Zhang et al. 2023c). A support vector regression (SVR) model was used to construct relationships between biomass properties/HTC conditions and hydrochar composition (Shafizadeh et al. 2023). Most previous research only established one or several properties of hydrochar prediction models separately but did not carry out model verification, interpretation, and application, which were limited to the collected dataset. Thus, it is necessary to develop reliable ML models for predicting and optimizing multiple properties of hydrochar with clear mechanisms. Moreover, the biochemical composition was used to predict bio-oil properties during the hydrothermal liquefaction of biomass, which showed that lipids were the

major component affecting the bio-oil yield and energy recovery (Sheng et al. 2018; Li et al. 2021a). The impacts of biochemical composition on synergism/antagonism of the bio-oil yield were more than that of reaction conditions (Yu et al. 2023), which may be suitable for HTC processes. The nitrogenous components were mainly derived from proteins (Zhang et al. 2021; Leng et al. 2022a). Generally, three primary biochemical components of biomass (i.e., proteins, lipids, and carbohydrates) were not given simultaneously with the elementary composition of HTC, even though they are principal factors in the formation of hydrochar (Li et al. 2019). However, few studies have considered the effects of the biochemical components on the hydrochar properties.

In this study, biochemical components (i.e., proteins, lipids, and carbohydrates) of biomass were predicted first by elemental composition, which contributed to biomass characterization. The impacts of biochemical components on the hydrochar properties were explored by ML models, guiding the HTC process design. Then, ML multi-target models were constructed to predict and optimize the hydrochar properties (yield, elementary composition, elemental atomic ratios, and HHV). Single-biomass composition (elementary and

Table 1 Comparison of research designs with applied ML models between previous studies and this work in the HTC process

Dataset size	Input	Output	Models	Remarks	References
649 and 475, from references published in 2010–2017	13 input features: (i) C, H, O, (ii) Fc, Ash, V, (iii) RT, T, WC, (iv) cellulose, hemicellulose, lignin, (v) polarity	yield, ER	RF	Single-target prediction No feature importance analysis Not well-interpreted No GUI was available	Li et al. (2020b)
248–536, from references published in 2010–2022	9 input features: (i) C, H, N, O, (ii) Fc, Ash, V, (iii) RT, T, WC	yield, HHV, ER, CR, N/C, H/C, O/C, C_char, H_char, N_char, S_char, O_char, Ash _{hy} , V _{hy} , Fc _{hy}	SVR, RF DNN ANN PSO-NN DTR	Small datasets Limited biomass types Unreasonable data and data noise Not well-interpreted Single-target prediction mostly No GUI was available mostly	Li et al. (2021b); Gao et al. (2022); Mu et al. (2022); Shafizadeh et al. (2023)
138–300, from references published in 2013–2022 and experiments	4 input features: (i) protein, lipid, carbohydrate, (ii) reaction intensity	HHV, ER, fuel ratio; OMC, pH, CEC; SSA, PV, MPS	RF	One specific biomass type Limited input variables Single-target prediction No model interpretation No GUI was available	Zhang et al. (2023c)
648, from references published in 2010–2022	14 input features: (i) C, H, N, O, S, (ii) Fc, Ash, V, (iii) RT, T, WC (iv) protein, lipid, carbohydrate	9 outputs: yield, N_char, O_char, C_char, H_char, HHV_char, N/C, O/C, H/C	RF, GBDT	Large dataset 6 common types of biomass Multi-target prediction Biochemical compositions as inputs Good interpretation An online GUI system	This work

RF: random forest; SVR: support vector regression; DNN: deep neural network; ANN: artificial neural network; PSO-NN: particle swarm optimization for artificial neural network; Fc: fixed carbon; V: volatiles; RT: reaction time; T: temperature; WC: water content; HHV: higher calorific value; ASH_{hy}: The ash content of hydrochar; OMC: organic matter content; CEC: cation exchange capacity; SSA: specific surface area; PV: pore volume; MPS: mean pore size; GUI: graphical user interface

biochemical), proximate analyses, and HTC conditions were used as inputs. The in-depth feature analysis through ML provides insights into the mechanisms of hydrochar formation. An intelligent system was also formed and shared based on a multi-target model, which was verified by predicting elemental atomic ratios (N/C, O/C, and H/C). It was not only used to guide the optimal production of hydrochar derived from HTC from single-biomass samples. Finally, hydrochars from the co-HTC of mixed biomass samples reported in previous references were compared.

2 Materials and methods

2.1 Dataset preparation and preprocessing

The related keywords (e.g., hydrothermal carbonization, biomass, algae, lignocellulose*, and hydrochar) were retrieved from literature databases such as Web of Science and Google Scholar for searching amounts of SCI papers. The criterion for selecting suitable literature was that the data in the chosen articles must contain inputs (biomass composition and HTC conditions) and outputs (yield and elementary composition of the hydrochar). The hydrochar dataset was organized from the experiments of 87 selected articles among reviewed publications since 2010. All data points were collected from tables or read from figures with PlotDigitizer (a data extraction software). The detailed datasets #1–6 are provided in the Additional file 2. The collected data were divided into two datasets: hydrochar from single biomass (dataset #1) for model development, and hydrochar from mixed biomass (dataset #6) for model validation. In detail, the elementary composition (i.e., C, H, N, O, S, and ash), proximate composition (fixed carbon (Fc), volatiles (V), and ash), and operational conditions (residence time (*RT*), temperature (*T*), and water content of biomass (*WC*)) were treated as the input variables. The output variables of the HTC experiments including the hydrochar yield (Yield_{char}) and elementary composition of hydrochar. Finally, 51 types of forestry leftovers, 43 types of domestic waste, 29 types of crop straw, and 17 types of organic waste were collected in dataset #1. To further explore the effect of biomass composition and HTC conditions on hydrochar quantity and quality, 199 kinds of biomass from 59 publications with listed elemental and biochemical composition were compiled in dataset #2. This consisted of 108 types of microalgae, 41 types of macroalgae, 20 types of sludge, 12 types of manure, 16 types of food waste, and 2 types of lignocellulose. The units of input and output variables in all datasets were standardized using international units to improve the consistency of units and better compare each feature in the datasets. The biochemical components were predicted from a model built based on dataset #2, and they were input into dataset #3

in combination with the HTC conditions to predict the hydrochar properties. Then, the predicted biochemical components were added to dataset #1 to form dataset #4. In addition, the atomic ratios were calculated according to the elemental compositions of hydrochar and then compiled to form dataset #5. The data of hydrochar related to the co-HTC of mixed biomass from references were integrated into dataset #6. The detailed information of all datasets is listed in Table 2.

The composition of biomass in datasets containing biochemical, proximate, and elemental composition was normalized using the following three equations (dry basis), respectively:

$$\text{Protein} + \text{Lipid} + \text{Carbohydrate} + \text{Ash} = 100 \quad (1)$$

$$V + Fc + \text{Ash} = 100 \quad (2)$$

$$C + H + N + O + S + \text{Ash} = 100 \quad (3)$$

where *V*, *Fc*, *C*, *H*, *N*, *O*, and *S* are the contents (wt%) of volatiles, fixed carbon, carbon, hydrogen, nitrogen, oxygen, and sulfur in biomass, respectively.

The elementary compositions of hydrochar in datasets were treated on a dry basis and unified with the following formula:

$$O_{char} = 100 - C_{char} - H_{char} - N_{char} - \text{Ash}_{char} \quad (4)$$

where *O_{char}*, *C_{char}*, *H_{char}*, and *N_{char}* are the elemental contents (wt%) of oxygen, carbon, hydrogen, and nitrogen of hydrochar, respectively. The elementary composition of biomass and hydrochar in datasets were directly collected from the literature or calculated on a dry basis using Eqs. 3 and 4, respectively.

Through data preprocessing, the original data can be made more standardized, accurate, and usable, and the quality of the dataset can be improved, thereby improving the accuracy of data mining and prediction results (Kirchner et al. 2016).

Pearson correlation analysis was conducted to understand correlations between inputs and outputs (Leng et al. 2023), and the Pearson correlation coefficient (PCC, *r* value) was calculated using Eq. 5:

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

where *r_{xy}* represents the PCC value of the two variables, and \bar{x} and \bar{y} are the mean values of the input feature *x* and another variable *y*, respectively. The value of *r_{xy}* is between −1 and 1, where 0 indicates no linear correlation between the two variables. A higher PCC indicates a

Table 2 Detailed description of various compiled datasets used in this study

Dataset	Data point	Input feature	Outputs
#1	648	Elementary composition and proximate analysis of single biomass + HTC operating conditions	Yield_char N_char O_char C_char H_char HHV_char
#2	199	Elementary composition of single biomass	Protein Lipid Carbohydrate
#3	648	Predicted biochemical components + HTC operating conditions	Yield_char N_char O_char C_char H_char HHV_char
#4	648	Elementary composition, biochemical components, and proximate analysis of single biomass + HTC operating conditions	Yield_char N_char O_char C_char H_char HHV_char
#5	648		N/C O/C H/C
#6	27	Elementary composition and proximate analysis of mixed biomass + HTC operating conditions	Yield_char N_char O_char C_char H_char HHV_char

higher correlation between two variables, and a negative value indicates a negative correlation and vice versa.

2.2 Machine learning models establishment

The scope of experimental values of various features varied widely. To obtain a consistent range of orders of magnitude between different variables without changing the distribution relationships of each feature, all input features were normalized using the *Z*-score method before training using Eq. 6:

$$Z_i^* = \frac{Z_i - \mu}{\sigma} \quad (6)$$

where Z_i^* and Z_i are the standardized and actual experimental values of each input variable, respectively; σ and μ express the standard deviation and mean values of each Z_i , respectively. All input features were treated with similar scales of importance to avoid generating significant

errors from range inconsistency or outlier points (Zhang et al. 2024).

The data points in the dataset were split into two parts (the training dataset and the testing dataset) during the ML model development process. Herein, 80% of data points in the whole dataset were randomly chosen and included in the training dataset, while the remaining 20% were used as the testing dataset to estimate ML predictive models.

Two classical tree-based ensemble learning ML algorithms, random forest (RF) and gradient boosting decision tree (GBDT), were used to train the predictive models (Deng and Lin 2022; Xiong et al. 2023), which could provide explainable prediction process. During RF learning, bagging or bootstrap aggregation was applied to develop hundreds of decision trees with the input dataset, where each tree was a regressor (Fang et al. 2022). The final model prediction result for RF was the average output value of all individual decision trees. The GBDT

algorithm was developed using the boosting method (Natekin and Knoll 2013), in which a leaf (an initial predicted value for obtaining the initial errors) was obtained from the mean of the observed output values. The gradient boost strategy involves building the first regressor tree to predict errors according to changes between the actual and original values. Subsequently, each tree minimizes the errors from the prior tree to obtain a highly accurate and efficient GBDT model (Yuan et al. 2021). The optimal hyperparameters of these two algorithms were determined through fivefold cross-validation combined with a simulated annealing algorithm (SSA) based on maximal mean regression coefficient (R^2) during model training for fast optimization (Leng et al. 2022b). The training dataset was randomly split into five parts, of which four parts were used to determine the optimal hyperparameters of the ML model during training, and the remaining part was used to evaluate the training performance. For SSA, the hyperparameters of the RF or GBDT models were used as inputs, and the average mean R^2 of the fivefold cross-validation was the output (Additional file 1: Fig. S1). The best hyperparameters of these two were finally confirmed by the values corresponding to maximized mean R^2 (Additional file 1: Fig. S1). In this present study, the tuned hyperparameters of the RF were the number of trees (n_estimators) and the maximum depth of trees (max_depth), while those for the GBDT were n_estimators, max_depth, learning rate, and loss function. The model and optimization algorithm employed the *scikit-learn* library (scikit-learn.org/stable) and *Hyperopt* library (Putatunda and Rama 2018) in the Python programming platform, respectively.

2.3 Model evaluation

To comprehensively predict and evaluate the output characteristics, multi-target models were applied to predict the biochemical components of biomass, the biomass yield, and the properties of hydrochar in this work. A total loss function (L_M) of the multi-target predictive model was specified as follows (Liu et al. 2017):

$$L_M = \sum_{t=1}^M \alpha_t L_t(y_n^t, \hat{y}_n^t) \quad (7)$$

where M is the number of targets; α_t represents the weight factor for each target t , which was used to underline the degree of relative importance for this task, where all targets in this work were given equal weight; L_M indicates the loss function for each t ; y_n^t is the actual value for each t ; \hat{y}_n^t is the predicted value for each t . All outputs were treated equally in the multi-target predictions (i.e., with the same weight).

The determination coefficient (R^2), a statistical index, was employed to evaluate the model, which generally represents the fitting degree of the relationship between input and output variables.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i^* - \hat{y}_i)^2}{\sum_{i=1}^N (y_i^* - \bar{y})^2} \quad (8)$$

where N indicates the number of test samples; y_i^* and \hat{y}_i indicate the i th experimental value and corresponding predicted value, respectively; \bar{y} indicates the mean of N experimental values.

In addition, the root mean square error (RMSE) was calculated to evaluate the model's prediction performance by representing the degree of dispersion between the actual and predicted values using Eq. 9.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^* - \hat{y}_i)^2} \quad (9)$$

Since biochemical components were both outputs and inputs for ML models, residue analysis was conducted on biochemical composition predictive models. The Shapiro–Wilk (S–W) test (Shapiro and Wilk 1965; Yap and Sim 2011; Rodrigues et al. 2023) (confidence interval = 95%, median residues value = 0) was analyzed based on the obtained residues to calculate significant values (p_{S-W}) of predicted and actual values. If $p_{S-W} > 0.05$, the residues conform to the normal distribution (Shapiro and Wilk 1965), which shows high prediction feasibility and reliability of ML models (Leng et al. 2023).

2.4 Model interpretation

After the ML models were built and their performance was assessed, the impact of input features on targets was analyzed to understand relationships between crucial input variables and output variables. Embedded feature analysis [e.g., feature importance and partial dependence plot (PDP)] and Shapley additive explanation (SHAP) (Ribeiro et al. 2016) for input features on targets were used to understand the ML predictions to understand the underlying ML model. The formula of the SHAP value is as follows:

$$\phi_i = \sum \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \quad (10)$$

where M and S are the dimensions of the set of all features and a subset of M , respectively; ($S \in M$); $f_x(S)$ and $f_x(S \cup \{i\})$ are the mean sample values calculated by the models trained with and without output variables, respectively (Zhu et al. 2023). The PDP of ML models can also display visual relationships (e.g., linear, monotonic, or more complex correlations) between one or two

feature input and output variables (Palansooriya et al. 2022).

2.5 Model applications and verification

2.5.1 Graphical user interface

An intelligent system based on an online graphical user interface (GUI) website was developed to provide a convenient platform for other researchers to use. Herein, the predictive ML model was uploaded to the GUI website by using *hypertext mark-up language* (HTML), *cascading style sheets* (CSS), Python program (version 3.8), and *Flask* framework (version 1.1) (Leng et al. 2022a).

2.5.2 Prediction of elemental atomic ratios (N/C, O/C, and H/C)

The elemental atomic ratios (N/C, O/C, and H/C) provide indicators for measuring the fuel properties of hydrochar. After the system was developed, we aimed to evaluate whether the elemental composition of the hydrochar were precise or not. Hence, the elemental atomic ratios were calculated by the values predicted by the multi-target ML model, and the results were compared with the actual results.

2.5.3 Forward optimization of hydrochar from specified biomass and experimental verification

To further verify the generality of the model and system, the input HTC conditions were adjusted by taking soybean straw from our laboratory as the raw material (C=41.05%, H=5.52%, N=2.90%, O=41.39%, S=0.28%, ash=8.87%, FC=74.38%, and V=16.75%). Among them, the temperature was set to 5 °C, from 140 °C to 310 °C; the residence time was 5 min, from 5 to 240 min; the water content step was 5%, from 75% to 99%. The optimal hydrochar objective function was set to maximize yield, C_char, and H_char and to minimize N_char and O_char, belonging to multi-objective optimization (Liu et al. 2022). Experimental validation was conducted using the optimal iteration outputs to guide the experiments to generate the desired hydrochar. Soybean straw was collected from a farm and was smashed into <50 μm particles during HTC experiment validation. Then, it was dried to constant weight at 105 °C and stored in a glass container. The detailed operating procedures (e.g., reactor setup, reaction program, procurement of products, and analysis) are provided in our previous studies (Zhang et al. 2021; Leng et al. 2022c). Briefly, the reactor was heated to a set temperature at a rate of 4–5 °C min⁻¹, and HTC was reacted at a given residence time. The solid product was filtered from the mixture after the reactor was cooled to room

temperature, and the solid product was dried at 105 °C for 24 h before characterization.

2.5.4 Prediction of hydrochar from co-HTC

The system was established using hydrochar derived from single biomass HTC without considering a biomass mixture. To explore the feasibility of using this system for mixed biomass, the validation dataset of hydrochar characteristics from different biomass mixtures in the literature was collected. The values predicted by the system were compared with the actual experimental values from previous studies using the same input variables.

3 Results and discussion

3.1 Dataset integration, filling, and analysis

The box plots show the minimum, average, median, maximum, and quartile values of the input and output features. Figure 2a–c show the distribution of each feature in dataset #1. Elements C and O were the two major components of the elementary composition of biomass, while V was the major component in proximate composition (Fig. 2a). The sludge generally contained low amounts of C and V and high contents of N, S, and ash, which are shown as outliers in Fig. 1a. For the HTC operating conditions, the most common values of *RT* ranged from 5 to 240 min, *T* ranged from 140 °C to 310 °C, and *WC* ranged from 75.00% to 99.90%, respectively (Fig. 2b). For the output variables, the median yield, C, and O were 55.26%, 50.60%, and 23.46% (Fig. 2c). Moreover, the Van Krevelen diagrams of H/C and O/C showed that three reactions (i.e., decarboxylation, dehydration, and demethylation) could occur in the HTC (Fig. 2d), of which dehydration and decarboxylation were dominant. The overall distributions of all six output variables were approximately normal (Additional file 1: Fig. S2), which expressed that data randomly selected from the literature satisfied the laws of probability. The normal distributions of outputs also examined the reliability of hydrochar samples, which validated the assumption that normality is of fundamental concern for machine learning modeling and analysis (Yap and Sim 2011). A total of 27 hydrochar data from the HTC of mixed biomass were compiled, and the range of input variables is shown in Fig. 2e and f, where the distribution ranges of input variables of mixed biomass were within the data range of single biomass (Fig. 2). The range of mixed biomass was more concentrated without obvious outliers. In addition to elemental and proximate composition, the biochemical components were used as another indicator to identify the type of biomass. However, biochemical components (i.e., proteins, lipids, and carbohydrates) are often ignored or unknowable during the HTC process. To fill the missing data in dataset #1,

the new dataset #2 was compiled using the elemental and biochemical composition to develop ML models (details provided in Sect. 3.2.1). Figure 2g represents the numerical intervals of elemental and biochemical composition. The data points covered nearly the whole three-phase graph space, which indicated that the collected dataset #2 covered most biomass types (Fig. 2h). Then, the biochemical components from dataset #1 were inserted using the above prediction model to compile a new dataset #4.

3.2 Model predictions and feature analysis

3.2.1 Predictions and impact analysis of biochemical components of biomass

Based on dataset #2, RF and GBDT single-target prediction models were used to predict the protein, lipid, and carbohydrate contents using biomass elementary composition as the inputs. The hyperparameters and training results of the three prediction models are shown in Additional file 1: Table S1, Fig. S3, and Table 3. A comparison of the training results showed that both RF and GBDT had better predictive performance for proteins and lipids, and the prediction effect ($R^2 > 0.90$) of RF was slightly better than that of GBDT ($R^2 = 0.88$). The R^2 of RF and GBDT models for carbohydrates was only about 0.80, possibly because values of carbohydrates may have come from various sources, including experimentally measured values and calculated values (Eq. 1). The measurement methods vary from one carbohydrate to another, including anion-exchange chromatography (Hoekman et al. 2011), nuclear magnetic resonance spectroscopy (Yuan et al. 2011), and the phenol-sulfuric acid method (DuBois et al. 1956). The use of different methods led to a sizeable systematic error in the carbohydrate data, ultimately resulting in poor predictive accuracy and efficiency of the model for carbohydrates. It could be seen that most residues of the data points were concentrated between -5 and 5 , and the residual distributions of protein and lipid prediction models were closer to zero than that of the carbohydrate prediction model (Fig. 4a–c). According to the S–W test results, the residues of the training and testing datasets obeyed the normal distribution for all RF models for predicting protein, lipid, and carbohydrate ($p_{S-W} > 0.05$, Fig. 4). However, the residues of the testing data in the carbohydrate prediction model were relatively large and diffused (Fig. 4c), indicating poor generalization ability. To more accurately determine the carbohydrate content and reduce prediction errors, the protein and lipid contents predicted by the RF model and those experimentally determined from ash were used to calculate the carbohydrate content by difference (Eq. 1). The calculated carbohydrate contents fitted well with the actual values in dataset #2, with $R^2 = 0.92$ and $RMSE = 5.77$ (Fig. 3f), showing the

model's good predictive performance for proteins and lipids (Fig. 3a and b). Moreover, the residual distribution of calculated carbohydrate contents was similar to that of the training dataset in the prediction model (Fig. 4c and d, $p_{S-W} > 0.05$), validating that the calculated carbohydrate contents could reduce errors compared with direct prediction. Various ML models were compared for their ability to predict carbohydrate contents (i.e., cellulose, hemicelluloses, and lignin) of syngas from biomass fluidized gasifier as the input. However, the predictive performance was mediocre, and the highest R^2 was 0.77 among all models (Kim et al. 2023). Carbohydrates from previous studies were obtained mainly by difference (Duan et al. 2013; Liu et al. 2021). Hence, it is reasonable to employ a similar method to obtain carbohydrates in this study.

To accurately establish the relationship between elemental components and biochemical composition, the RF model was selected for subsequent model evaluation and prediction. Figure 3 shows the predictions of the RF model for proteins and lipids based on dataset #2 and the ranking of their respective feature importance. The protein content was concentrated between 20% and 60%, and the lipid content was mainly below 30%. The ML model showed more accurate predictions for data ranges, while the confidence interval widened and the validated prediction accuracy decreased for intervals with relatively little data (Fig. 3a–c). The feature analysis exhibited that the N content of biomass significantly affected the protein content (Fig. 3c), while other elements did not (relative importance < 0.05) because proteins are nitrogen-containing compounds (Leng et al. 2020). In addition, the C content of biomass had the greatest influence on the lipid content, followed by the H content (Fig. 3d) because lipids are mainly composed of C and H atoms.

The elemental composition, proximate analysis, and HTC conditions in dataset #1 were applied as inputs for the above optimal RF models to predict proteins and lipids, thereby also determining carbohydrates by difference.

3.2.2 Multi-target predictions and impact analysis of hydrochar properties by biochemical composition and HTC conditions

After obtaining 648 data points of biochemical composition, the corresponding HTC conditions (T , RT , and WC) and hydrochar properties were compiled into dataset #3 to predict the properties of hydrochar. The ternary phase diagram displays the various data distribution relationships between biochemical composition and hydrochar properties (Additional file 1: Fig. S4) caused by interactions between biochemical components (Duan et al. 2018). The collected biomass types in

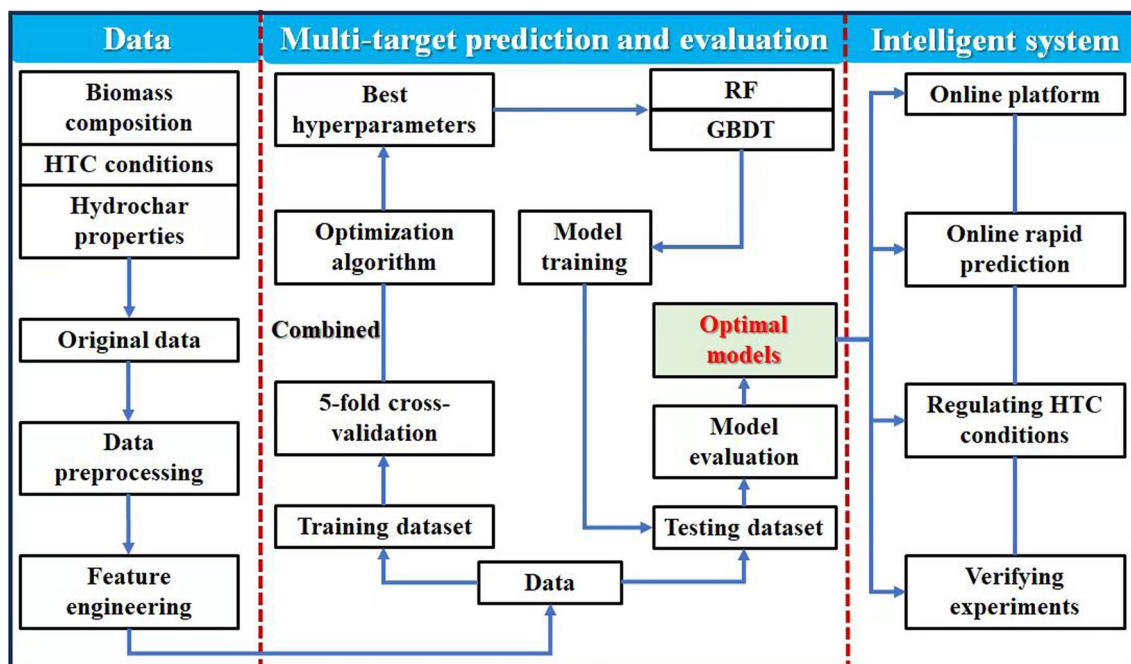


Fig. 1 Method flowchart of this study

dataset #1 were limited, so the ranges of predicted biochemical components (Additional file 1: Fig. S4) were narrower than those in dataset #2 (Fig. 2h). Table 3 shows the performance of the ML model in terms of using biochemical components and HTC conditions as inputs. The optimal hyperparameters are determined and listed in Additional file 1: Table S1, in which all hydrochar properties were predicted with relative accuracy (training $R^2 > 0.97$ and test $R^2 > 0.82$ of each output). The biochemical components showed good linear correlation with hydrochar properties ($|r| > 0.40$) except Yield_char and T was correlated with Yield_char among HTC conditions (Fig. 5), resulting in high predictive performance of hydrochar properties (average test $R^2 \sim 0.90$). The overall feature importance of hydrochar and the individual PDP of each biochemical component to various hydrochar properties are shown in Fig. 5a, Additional file 1: Fig. S5, S6. In particular, ash was the dominant feature of hydrochar, followed by T , while RT and WC had only slight impacts on hydrochar (Additional file 1: Fig. S6a). One reason for the major impact of ash on Yield_char was that the ash contents were used to calculate Yield_char in the literature. A higher carbohydrate content promoted O_char, C_char, H_char, and HHV_char. When the carbohydrate content rose from 10% to $\sim 40\%$, Yield_char gradually increased (Additional file 1: Fig. S5), indicating that carbohydrate hydrolysis may have occurred in this interval to form hydrochar. Upon further increasing the carbohydrate

content, the average C_char and O_char also increased from 20–25% to 45–52%, respectively, resulting in the polymerization of precursors (e.g., α -carbonyl aldehyde) from carbohydrates (Shi et al. 2019). Moreover, proteins are N-containing components and could participate in the Maillard reaction with carbohydrates, which affected the Yield_char and N_char (Fig. 5). Lipids exhibited a steady trend to Yield_char, N_char, and O_char, indicating that they had little contribution (Additional file 1: Fig. S4). However, lipids preferentially generated C_char, H_char, and HHV_char, with nearly linear relationships (Fig. 5 and Additional file 1: Fig. S5). Although the ML model here performed well, the biochemical components and HTC conditions (six inputs) were insufficient to describe the HTC process (six outputs), especially Yield_char. Hence, it will be necessary to consider more input descriptors to train more accurate ML models.

3.2.3 Multi-target predictions and impact analysis of hydrochar properties by biomass components and HTC conditions

After filling dataset #1 with the predicted biochemical components, the GBDT model was applied to predict the yield, elemental composition, and HHV of hydrochar according to 14 input features. These features included the elemental composition, proximate analysis, biochemical components of biomass, and HTC conditions (Additional file 1: Fig. S6).

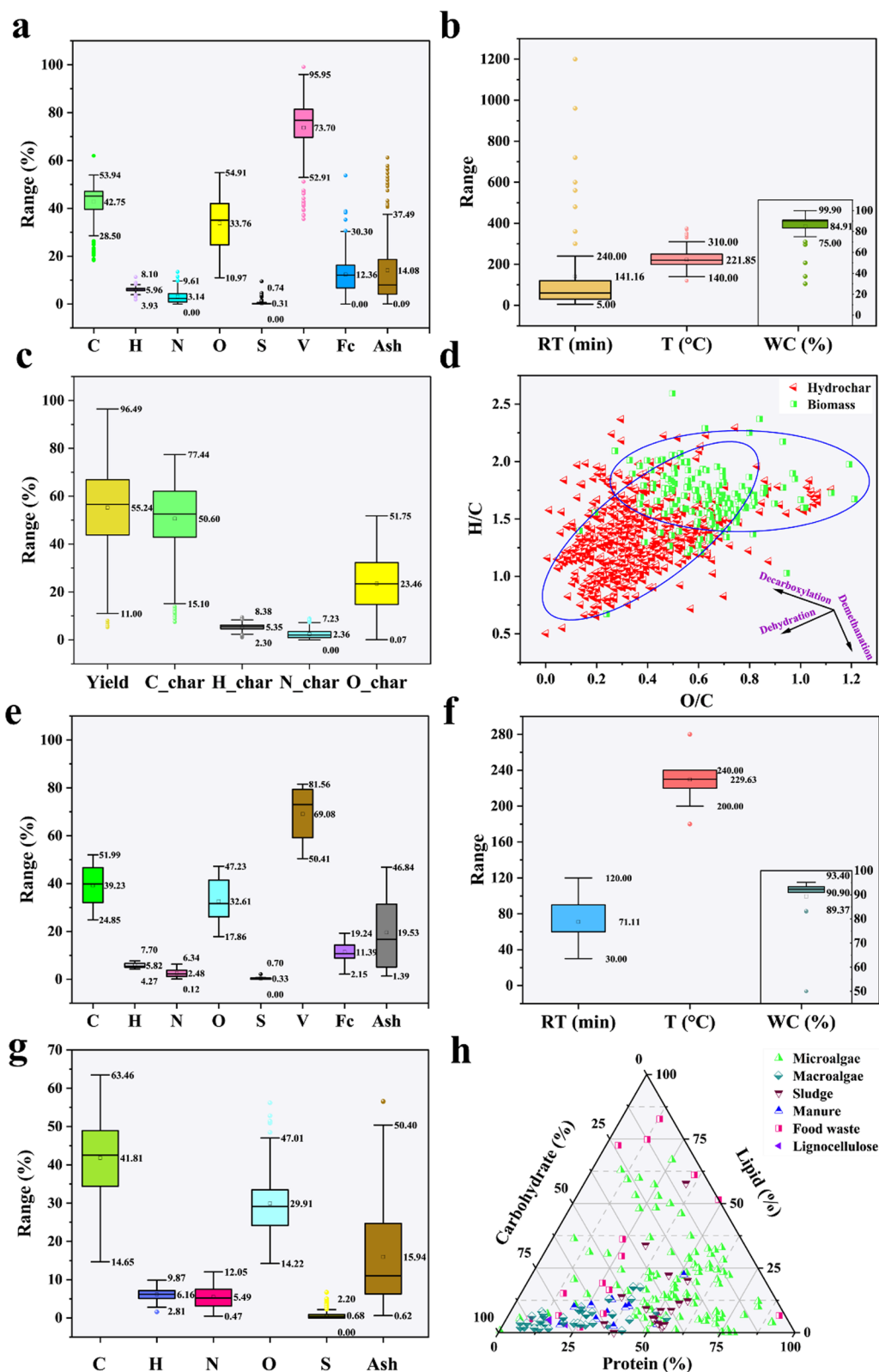


Fig. 2 Statistical data visualization of elementary and proximate composition (a), operating conditions of the HTC (b), hydrochar characteristics (c) and Van Krevelen diagram (d) in dataset #1, input variables (e, f) in dataset #6, and elementary composition (g) and biochemical components (h) of biomass in dataset #2. The three components (proteins, lipids, and carbohydrates) were unified to ensure that they added up to 100% (data unified with a dry-ash-free basis)

Table 3 Result of ML prediction model performance under different HTC datasets

Dataset	Input	Output	Model	Train R ²	Avg ^a	Train RMSE	Avg	Test R ²	Avg	Test RMSE	Avg	
#1	Elementary composition + proximate analysis + HTC operating conditions	Yield_char	GBDT	0.95	0.96	4.93	1.89	0.82	0.88	7.18	2.82	
		N_char		0.97		0.32		0.90		0.50		
		O_char		0.96		2.28		0.89		3.84		
		C_char		0.97		2.28		0.91		3.00		
		H_char		0.97		0.26		0.89		0.39		
		HHV_char		0.96		1.28		0.89		1.99		
		Protein	GBDT	0.97	_b	3.42	-	0.88	-	6.02	-	
			RF	0.97	-	3.37	-	0.90	-	5.49	-	
		Lipid	GBDT	0.97	-	2.44	-	0.88	-	6.85	-	
			RF	0.93	-	3.91	-	0.91	-	6.40	-	
#2	Elementary composition	Carbohydrate	GBDT	0.93	-	5.21	-	0.80	-	8.25	-	
			RF	0.92	-	5.59	-	0.82	-	7.88	-	
		Yield_char	GBDT	0.95	0.97	3.73	1.47	0.82	0.87	7.23	2.89	
		N_char		0.99		0.23		0.88		0.66		
		O_char		0.97		1.98		0.86		4.11		
		C_char		0.99		1.63		0.90		3.10		
#3	Biochemical components + HTC operating conditions	H_char		0.97		0.23		0.88		0.45		
		HHV_char		0.98		1.01		0.90		1.80		
		Yield_char	GBDT	0.99	0.99	2.08	0.82	0.88	0.93	5.74	2.36	
		N_char		0.99		0.12		0.96		0.42		
		O_char		0.99		1.07		0.91		3.37		
		C_char		0.99		1.00		0.96		2.87		
		H_char		0.99		0.11		0.92		0.37		
		HHV_char		0.99		0.55		0.94		1.40		
		N/C	Calculated	0.99	0.99	0.003	0.022	0.98 ^c	0.93^c	0.005 ^c	0.044^c	
		O/C		0.99		0.025		0.97 ^c		0.026 ^c		
#4	Elementary composition + biochemical components + proximate analysis + HTC operating conditions	H/C		0.98		0.038		0.83 ^c		0.101 ^c		
		N/C	GBDT	0.99	0.98	0.003	0.027	0.95	0.86	0.009	0.066	
		O/C		0.98		0.039		0.78		0.123		
		H/C		0.97		0.039		0.84		0.066		

^a Avg; Average; ^bNot applicable; ^cValidation

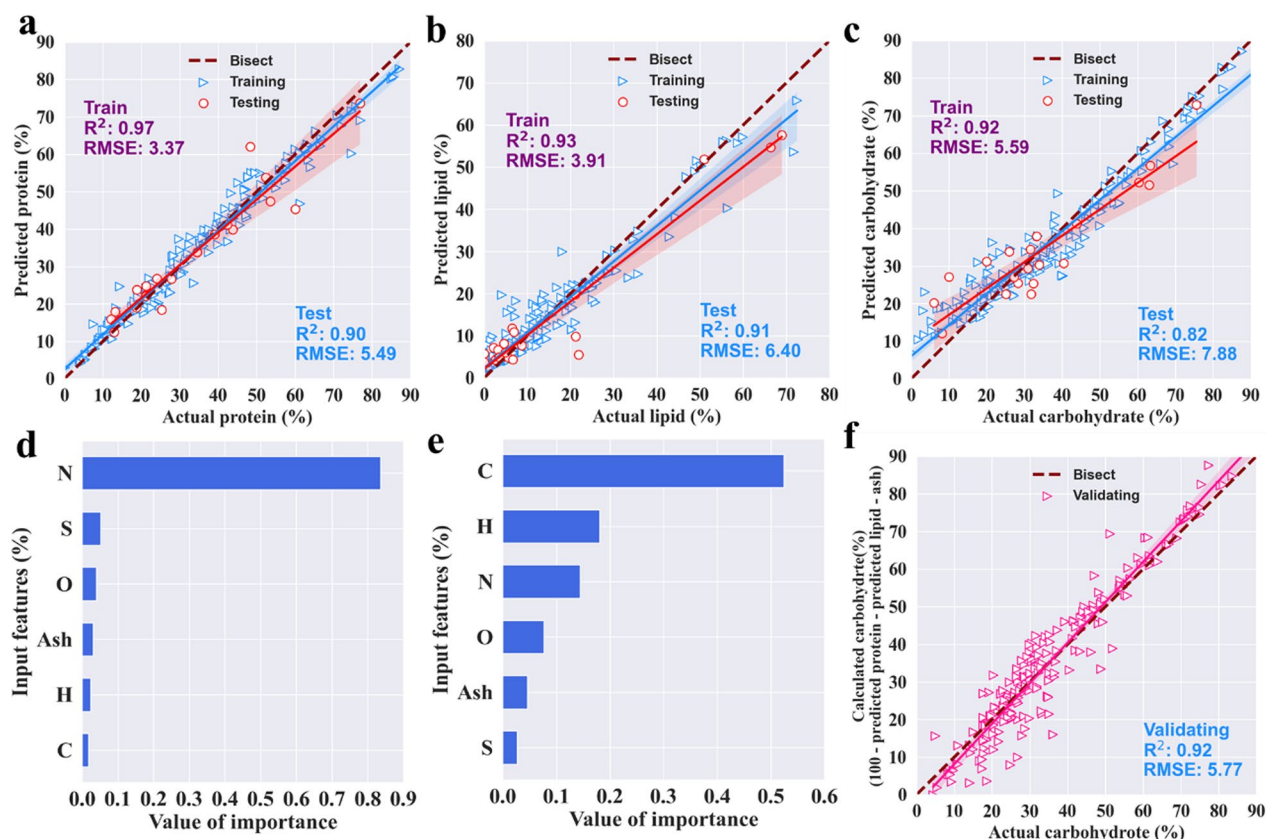


Fig. 3 Model performance and feature importance analysis for predicting (a, d) protein, (b, e) lipid, and carbohydrate (c) training and (f) validating based on RF models (dataset #2)

The PCC matrix of input and output based on dataset #4 (Fig. 5) shows that C_{char} , H_{char} , and HHV_{char} of hydrochar were positively correlated with the C, H, and V contents of biomass ($r > 0.50$), and strongly negatively correlated with ash content ($r < -0.60$). The N and protein contents of biomass were strongly positively correlated with N_{char} and N/C ($r > 0.75$), which indicates that the N content of the raw material was the main factor affecting the N content of hydrochar. The N/C ratio of hydrochar was negatively correlated with the O content and carbohydrate content ($r < -0.70$). Temperature showed a weak negative correlation with hydrochar, excluding C_{char} and HHV_{char} . The output variables such as $Yield_{char}$, H/C , and O/C had no significant linear correlation with the input characteristics. The target features also showed correlations, e.g., the atomic ratios of hydrochar were positively or negatively correlated with the corresponding elemental composition. Therefore, the elemental atomic ratios (H/C , O/C , and N/C) were classified as a separate category to be predicted to avoid similar information with the elemental composition of hydrochar. HHV_{char} was positively correlated with C_{char} and H_{char} ($r > 0.70$), and $Yield_{char}$

was negatively correlated with C_{char} ($r = -0.39$) and HHV_{char} ($r = -0.42$). In summary, the input variables were mostly correlated with the outputs, with PCC values mostly between -0.5 and 0.5 (except C_{char} , H_{char} , and HHV_{char}), showing that these descriptors were adequately representative. Thus, the feasibility of using multi-target ML prediction was demonstrated, showing it could be used to draw generalized conclusions about typical correlations.

Selected optimal hyperparameters are shown in Additional file 1: Table S2, and the comprehensive prediction performance is presented in Table 3. In Fig. 6, the horizontal coordinate represents the original experimental data, and the vertical coordinate represents the data predicted by the model. The closer the scatter point to the bisect line (function $y=x$), the more accurate the model; otherwise, the model error will be larger. After adding the elemental composition and proximate analysis into dataset #3, the training accuracy of hydrochar properties was improved, with both training R^2 values reaching 0.99 and the average testing R^2 increasing from 0.87 to 0.93 (Table 3 and Fig. 6). This value was superior to

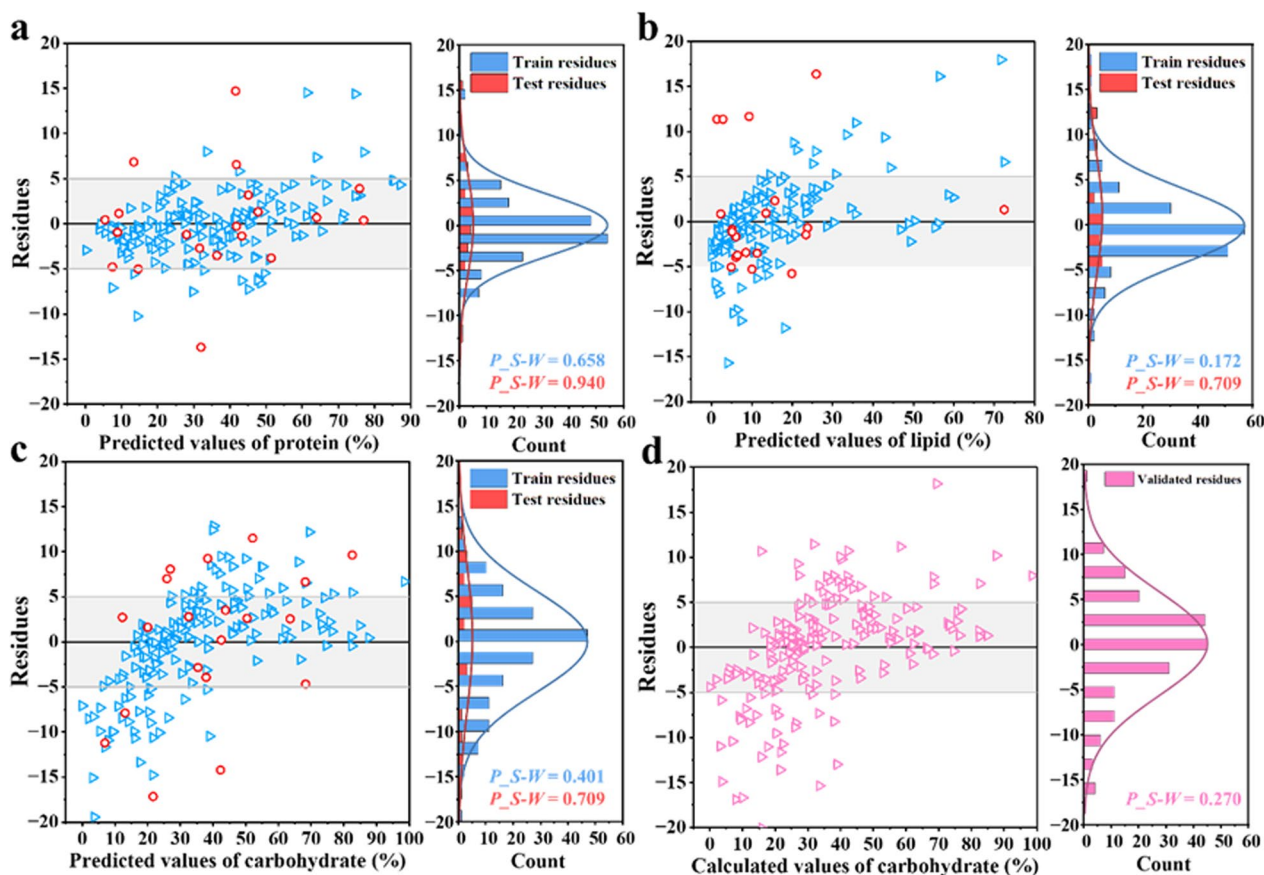


Fig. 4 The residue analysis for predicted values and actual values of (a) protein, (b) lipid, (c) carbohydrate, and (d) calculated carbohydrate based on RF models (dataset #2)

that (0.86–0.91) of a previous study that did not consider biochemical components (Li et al. 2020a). The results showed that adding biochemical components improved the prediction performance of hydrochar properties to a certain extent, providing a means to characterize biomass raw materials. Therefore, considering more key features of biomass components (containing elemental composition, proximate analysis, and biochemical components) could comprehensively represent biomass and further improve the prediction accuracy of ML models for hydrochar properties, due to correlations between inputs and outputs variables (Fig. 5).

As illustrated in Additional file 1: Fig. S7, the input features had various effects on the different targets of hydrochar. SHAP assigned an importance value to each feature variable for each predicted data point, where a positive SHAP value indicates that the input features are positive to the predicted target. A red or blue point indicates a high or low original value of the input features, respectively. It can be assumed that the input feature is positively related to the output target when the data points of the input feature change from blue to red upon

increasing the SHAP value. Moreover, the mean absolute SHAP ($|\text{SHAP}|$) values of each input feature were aggregated for the six characteristics of hydrochar. The importance ranking of input features during the multi-task prediction process was conducted according to the sum of $|\text{SHAP}|$ values for the six outputs. The one-way PDP provided more specific functional information of the input variables to each output (Additional file 1: Figs. S8–12).

In the *Yield_char* prediction, the HTC temperature, N content, and O content were identified as the three most significant variables, among which the biomass N content and temperature negatively affected the *Yield_char* (Additional file 1: Fig. S8). The low O content (<20%) increased *Yield_char*. When the O content was between 20% and 26%, the relationship between O and *Yield_char* was linear negative, and further increasing the O content did not significantly change *Yield_char* (Additional file 1: Fig. S8). The N and protein contents were the two most critical contributors to *N_char*, showing that the N atoms in hydrochar were mainly derived from N-containing components of biomass (Fig. 7b). The positive

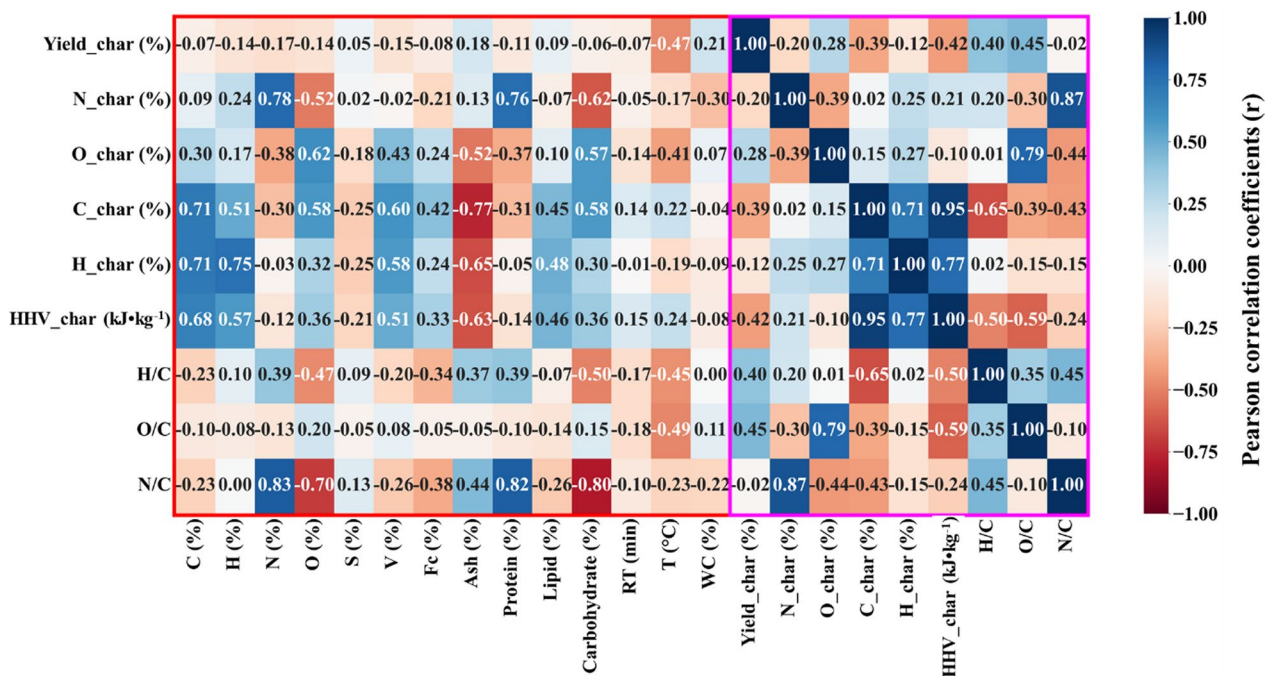


Fig. 5 Pearson correlation coefficients (r) matrix of inputs features and targets (dataset #4)

correlation between N and N_char was stronger than that between protein and N_char because protein was mainly predicted and affected by the N content (Fig. 3d and Additional file 1: Fig. S9). In O_char prediction, the O content, T, and RT were the most critical input features, in which O_char was negatively correlated with T and RT and positively correlated with the O content (Additional file 1: Fig. S10). The cumulative O content indicated a reduction in the organic matter content and O_char aggravation. A higher T and RT accelerated deoxidation and dehydrogenation reactions (Additional file 1: Fig. S10) during the HTC process. The biomass ash was slightly linearly negatively correlated with C_char and H_char (Additional file 1: Figs. S11–12) due to an increase in inorganic matter (Eqs. 1–3). Therefore, the ash content was the key characteristic affecting C_char (Fig. 7d), which might have been conducive to decarboxylation and aromatization reactions during HTC (Buss et al. 2019). For HHV_char prediction, the H content, ash content, and temperature had the greatest influence, where the H content (6–8%) and temperature both positively impacted HHV_char, while ash and HHV_char were negatively correlated.

Furthermore, ash, T, and N and C contents had the greatest overall impact on the six prediction objectives (Fig. 7b). In general, insoluble substances in ash decreased the yield and quality of hydrochar. However, the release of alkaline earth and alkaline earth metals (e.g., Ca) from the ash to the hydrochar surface increased

the yield by catalyzing the hydrolysis of biomass to hydrochar (He et al. 2022a, 2022b). Thus, the impact of ash on Yield_char might have been neutralized, showing a steady trend (Additional file 1: Fig. S7). Therefore, temperature and C content could be identified as significant features because temperature affected the degree of thermal degradation of biomass, and C was the main energy source of hydrochar. However, the N content of biomass was another key factor that affected the properties of hydrochar, especially during the prediction of Yield_char.

The importance of elemental composition, biochemical components, proximate analysis, and operating conditions to the six properties of hydrochar was 35%, 25%, 21%, and 19%, respectively (Fig. 7b). These results indicate that elemental composition and biochemical components were fundamental characteristics that affected the properties of the resulting hydrochar (the total relative importance was about 60%). The relative importance of biomass components (elemental composition, biochemical components, and proximate analysis, >80%) was much higher than that of HTC conditions (<20%), showing hydrochar properties were more dependent on the composition of feedstocks. Because biochemical composition was predicted, some of the information about the biochemical components was similar to that of the elementary composition. Therefore, it was reasonable that the elemental composition was more important than the biochemical components. Furthermore, the feature importance ranking and analysis were almost the same as

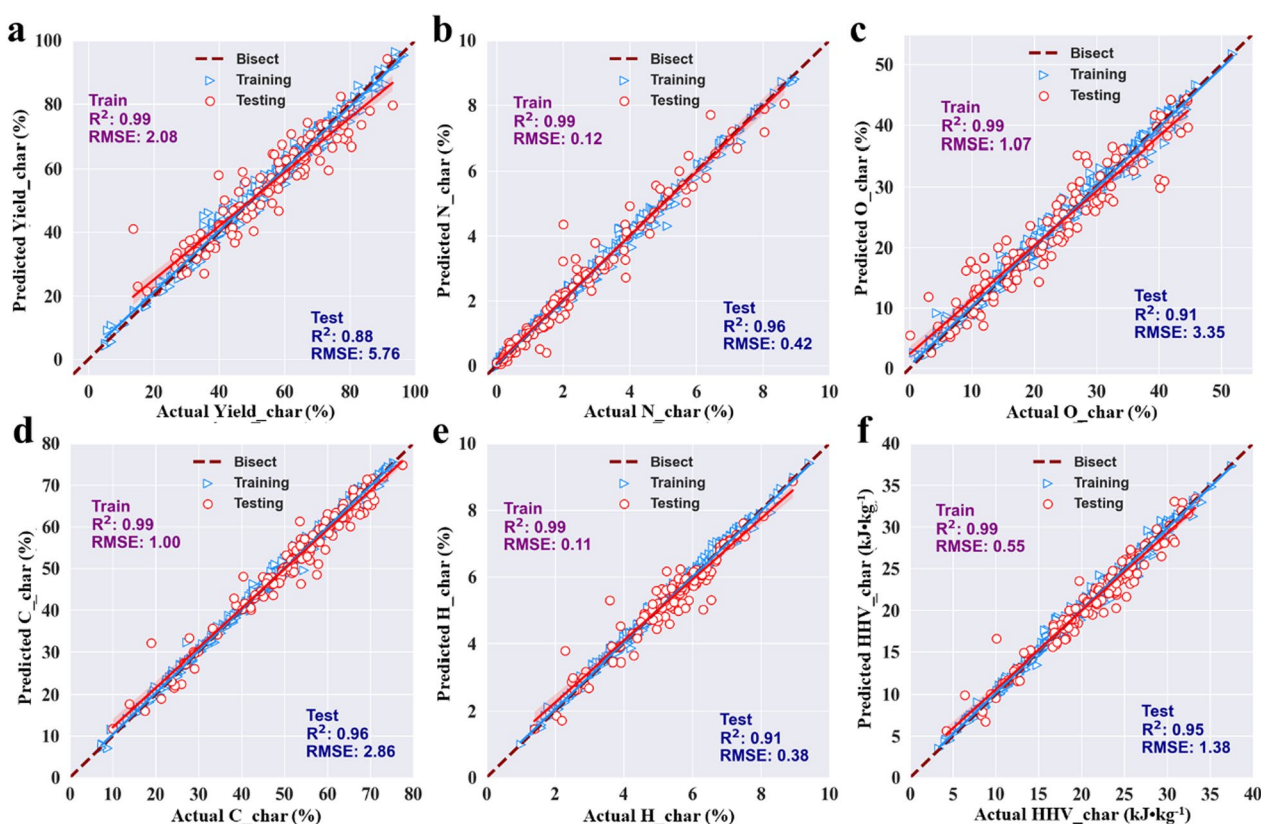


Fig. 6 Multi-target predicted data versus the actual data based on the optimized GBDT for predicting (a) yield, (b) N content, (c) O content, (d) C content, (e) H content, and (f) HHV of hydrochar (dataset #4)

the models based on dataset #3 in which the ash and T were still the top two features (Additional file 1: Fig. S6). This was different from the previous results (Additional file 1: Fig. S6a), which showed that among biochemical components, carbohydrates were the main contributors to hydrochar formation (Additional file 1: Fig. S6b), consistent with previous research results (Li et al. 2019). N_{char} increased with the carbohydrate and protein contents (Fig. 7 and Additional file 1: Fig. S9) due to the Maillard reaction between proteins and carbohydrates during the HTC of biomass.

3.3 Practical validation for the online intelligent system

An online intelligent system (website link: <http://43.139.222.20:8000/>) for predicting and optimizing hydrochar's properties based on the above multi-target ML models was proposed and made accessible to other researchers. The homepage of the website is shown in Additional file 1: Fig. S13. More critically, the reliability and generality of the proposed system should be verified by practical data other than the training and testing dataset.

3.3.1 Verification of the atomic ratios (N/C, O/C, and H/C) with the predicted elementary composition of hydrochar from the model

A multi-objective prediction model of elemental atomic ratios was established using dataset #5, which used biomass components and the HTC conditions as the inputs. The optimal hyperparameter and predictive performance results are shown in Table 3 and Additional file 1: Table S2, respectively. The overall predicted results were accurate (the average training R^2 was 0.98, and the average test R^2 was 0.86). Still, there was a significant difference between the test results of the three atomic ratios. Only the test R^2 of N/C reached 0.94, while those of O/C and N/C were lower than 0.85 (Table 3), which might be because dozens of the H/C and O/C values of hydrochar were dispersed outside the concentrated area in the Van Krevelen diagram (Fig. 2).

Compared with the simultaneous prediction of all three atomic ratios, the predictive performance of the multi-target models for predicting six properties of hydrochar (dataset #4) was slightly improved by using the training dataset (R^2 from 0.98 to 0.99; RMSE from 0.027 to 0.022).

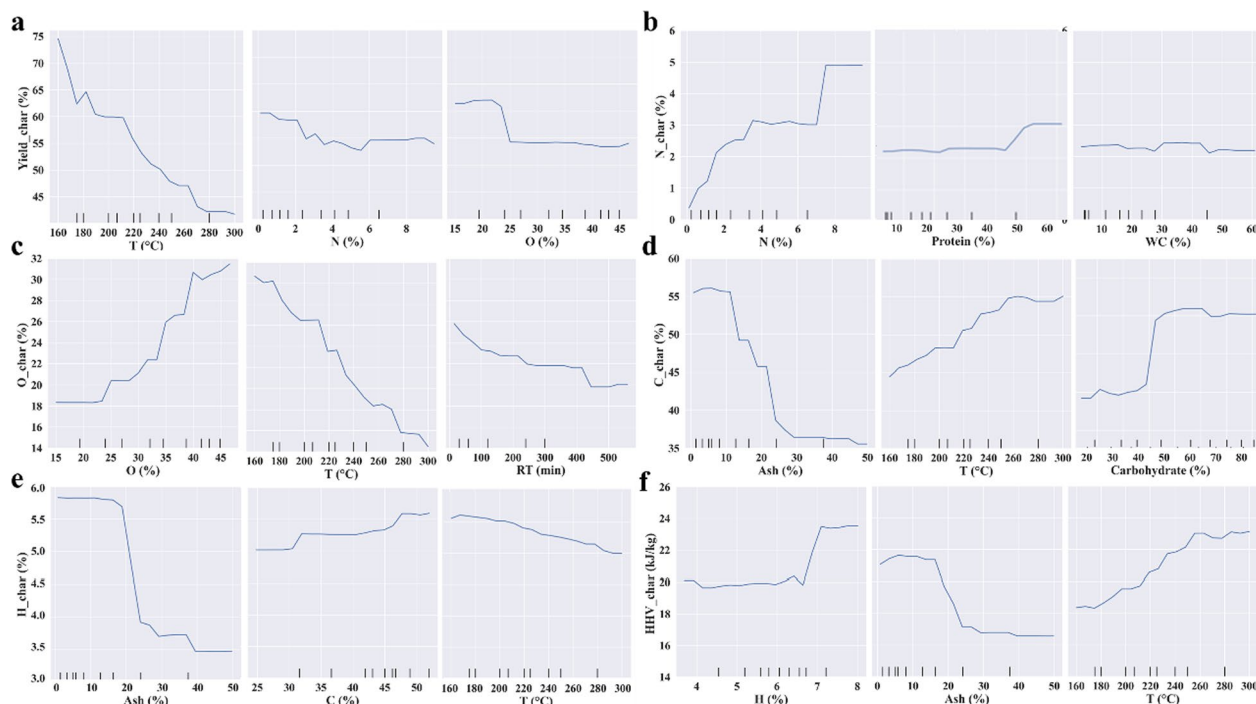


Fig. 7 Average impact of the top three important features by one-way PDP for (a) Yield_char, (b) N_char, (c) O_char, (d) C_char, (e) H_char, and (f) HHV_char with GBDT model developed based on dataset #4

On the test dataset, the predictive accuracy of N/C and O/C was improved, with the R^2 of O/C increased significantly from 0.78 to 0.97 (Table 3). The predictive accuracy of H/C slightly decreased (Table 3), possibly because the predictive performance of H_char (test $R^2=0.89$) was worse than that of N_char and O_char (test $R^2>0.90$) in the multi-target prediction model for hydrochar (Fig. 6). The direct prediction of elemental composition from the multi-target prediction model could reduce transfer errors (Qureshi et al. 2017) between H & O and C during calculations. Overall, the accuracy of predicting the elemental composition and then calculating the atomic ratios using the multi-objective prediction model for hydrochar properties was much better than that of predicting the atomic ratios directly. Therefore, the multi-target model based on dataset #4 accurately predicted the hydrochar elemental composition and provided good generalization performance for elemental composition. The system can be used to accurately calculate and predict the atomic ratios of hydrochar samples.

3.3.2 Forward optimization for hydrochar with specified biomass and experimental verification

From many prediction results from HTC of soybean straw with the system (12,420 iteration results, Additional file 1: Fig. S14), eight optimal solutions (maximized $optimal_{char}$) with different operating conditions were

screened for experimental verification (Table 4). Herein, $optimal_{char}$ was determined as follows:

$$Optimal_{char} = Maximum(Yield_{char} + C_{char} + H_{char} - O_{char} - N_{char}) \quad (11)$$

The WC was set at 75% or 90% in the forward experiments because the feature impact of WC was less significant for hydrochar than T and RT (Additional file 1: Fig. S6). In all cases, the absolute error of Yield_char between experiment and predicted values was $<5\%$. The elementary composition of hydrochar (N_char, C_char, and H_char) was experimentally verified, with an absolute error of $<1\%$. The absolute values of relative errors for all cases were lower than 15% (except one of 21.89%, negligible). In fact, most cases were within 10%, and one-third of cases were $<5\%$ (Table 4), indicating satisfactory validating performance. However, the predicted values of O_char were higher than some experimental values (Runs 2, 3, and 6), in which the O_char calculated via difference (Eq. 4) might also be a possible reason for the deviation. Overall, the relative errors between the predicted and experimental values of hydrochar properties were small, and the generality of the model was verified by optimization experiments. This confirms that the intelligent system based on multi-target modeling prediction can be used for the aforementioned single biomass.

Table 4 Forward optimization and experimental verification of hydrochar properties

Run	T °C	RT min	WC %	Yield_char %	RE ^c %	N_char %	RE %	O_char %	RE %	C_char %	RE %	H_char %	RE %
1	200	30	90	59.48 ^a (58.18) ^b	-2.19	2.01 (2.45)	21.89	35.39 (35.00)	-1.10	48.97 (49.09)	0.25	6.26 (6.10)	-2.56
2	210	30	90	58.06 (58.69)	1.08	2.87 (2.57)	-10.45	27.04 (30.65)	13.35	45.03 (41.78)	-7.22	5.45 (5.40)	0.92
3	210	120	90	51.08 (55.56)	8.77	2.80 (2.63)	14.40	22.87 (25.27)	10.49	46.21 (44.00)	-4.78	5.12 (5.10)	-0.39
4	220	240	90	50.10 (52.92)	5.63	2.84 (2.92)	3.04	25.19 (28.68)	13.85	46.59 (43.78)	-6.03	5.17 (5.20)	0.58
5	220	60	90	49.70 (51.23)	3.08	2.88 (2.64)	-7.64	20.60 (19.41)	-5.78	44.89 (46.11)	2.72	4.71 (4.80)	1.91
6	230	60	90	45.21 (44.94)	-0.60	3.09 (2.64)	-14.56	25.57 (29.00)	13.41	45.10 (46.00)	2.44	5.81 (5.90)	1.55

^a Values from experiments

^b Predicted values from the model

^c RE: relative error

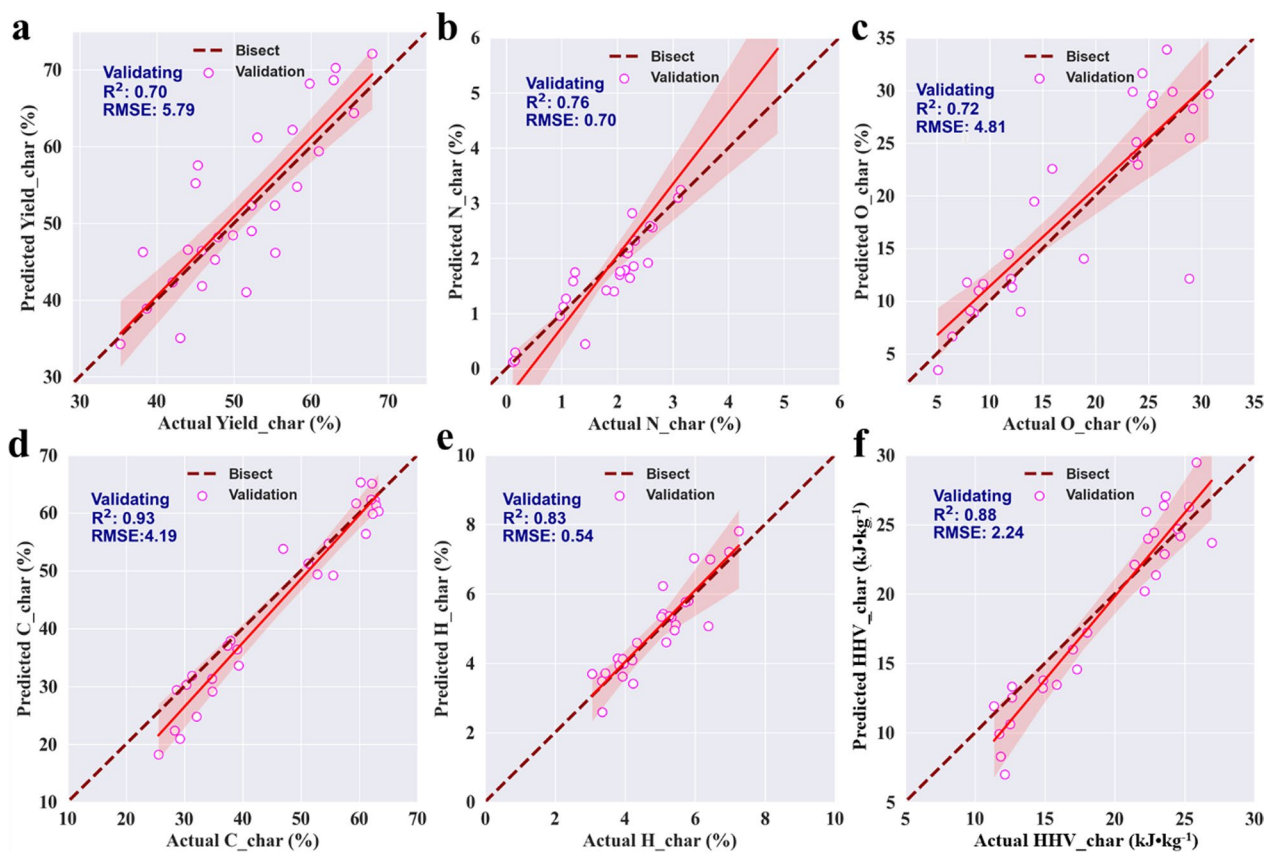


Fig. 8 Validation performance comparison of predicted values by the multi-target model (dataset #4) and experimental values of hydrochar obtained from a biomass mixture

3.3.3 Verification of hydrochar from co-HTC

During actual production, it is necessary to choose the appropriate ratio according to the biomass raw materials to produce high-yield and high-quality hydrochar under optimal HTC conditions. To reduce the number of experiments, the system can be further applied to predict and adjust the HTC process of biomass mixtures.

Figure 8 shows the predictions of six hydrochar properties using mixed biomass-related input variables as inputs into the multi-target model. The validated R^2 values of the GBDT multi-target model for the Yield_char, N_char, O_char, C_char, H_char, and HHV_char from mixed biomass were 0.70, 0.76, 0.72, 0.93, 0.83, and 0.88, respectively, with an average R^2 of 0.79 (Fig. 8). The RMSE values were 5.79, 0.70, 4.81, 4.19, 0.54, and 2.24 for these six targets, respectively (Fig. 8). The validation result of Yield_char was the worst among all prediction targets, which resulted in the predictive performance of Yield_char (test R^2 of 0.88) being lower than other hydrochar properties, with all test $R^2 \geq 0.91$ (Fig. 6). Moreover, the synergistic effects of different biomass could either promote or inhibit Yield_char during co-HTC process, which may have caused the Yield_char predicted by the multi-target model for single biomass to be lower or higher than the actual values (Fig. 8a). The validating performance of O_char was also unsatisfactory, similar to the forward experiments of single biomass (Table 4), due to the different acquisition methods of O_char values. The C_char, H_char, and HHV_char derived from biomass mixtures were predicted with validating $R^2 > 0.83$ because T and ash were the same important features for all of them (Additional file 1: Fig. S6), which are generally provided accurately. In summary, the online intelligent system could be employed to preliminarily predict hydrochar properties in real industrial applications only considering the single-biomass composition without considering the synergistic effect from co-HTC.

3.4 Future perspective and outlook

Although the system developed in this study was satisfactory, there is still room for improvement. More detailed input features could be considered to describe the HTC process, such as catalysts, solvents, ash compounds, pressure, and water recirculation. More descriptors for hydrochar, such as structural composition (functional groups), fuel properties (density), and material properties (e.g., surface area, surface morphology, pore volume, and cation exchange capacity), could be measured to more comprehensively assess the value of hydrochar for certain applications. The above descriptors are not considered in the proposed model, which is one limitation of this intelligent system based on the model.

The generality of this system for predicting hydrochar properties from biomass mixtures was inferior to that for single biomass, possibly because of the synergistic effects of biomass mixtures. Hence, more data related to hydrochar from biomass mixtures could be added to obtain datasets to improve the generality of our system. Additionally, more intelligent functions of the developed system could be exploited and upgraded, for example, the automated online reading of input datasets, the prediction of batches of data, reverse optimization, and error calculation.

4 Conclusions

In this study, biochemical components (i.e., proteins, lipids, and carbohydrates) of biomass were predicted first by their elementary composition. A novel intelligent system based on a multi-target ML prediction model (average test R^2 of 0.93; RMSE of 2.36) was established to predict and optimize the yield, elementary composition (C_char, H_char, O_char, and N_char), and HHV of hydrochar by using single biomass components and HTC operating conditions as inputs. Interpretation of the model's results showed that the ash content, temperature, and N and C contents of biomass were the most important factors affecting the six properties of hydrochar. Finally, an intelligent system was constructed based on a multi-target model and verified by applying it to predict the atomic ratios (N/C, O/C, and H/C), optimize hydrochar yield from the HTC of single biomass samples, and predict hydrochar properties from the co-HTC of mixed biomass samples. The proposed intelligent system provided accurate predictions and has practical generality, showing its great potential for industrial applications.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s42773-024-00303-8>.

Additional file 1: Table S1. Determined hyperparameters of different ML models. **Figure S1.** A basic illustration of fivefold cross-validation methods for developing the ML model (a) and (b) the flow chart of the SAA for hyperparameters tuning. **Figure S2.** The numerical distribution of output variables: (a) Yield_char, (b) N_char, (c) O_char, (d) C_char, (e) H_char, and (f) HHV_char in the dataset #2. **Figure S3.** Hyper-parameters tuning process of (a-c) RF and (d-h) GBR models for protein prediction based on dataset#2. **Figure S4.** Biochemical composition distribution of dataset #3 for outputs (a) yield, (b) N content, (c) O content, (d) C content, (e) H content, and (f) HHV of hydrochar (all data containing protein, lipid, and carbohydrate was unified with dry-ash-free basis to ensure them sum to 100). **Figure S5.** One-way PDP/sensitivity analysis of biochemical composition for (a) Yield_char, (b) N_char, (cc) O_char, (d) C_char, (e) H_char, and (f) HHV_char with GBDT model developed based on dataset #3. **Figure S6.** Overall feature importance for hydrochar based on the explanation of models with (a) dataset #3 and (b) dataset #4. **Figure S7.** Impacts of input features on each individual characteristic: (a) yield, (b) N content, (c) O content, (d) C content, (e) H content, and (f) HHV of hydrochar based on the explanation of GBDT model using SHAP values with dataset #4. **Figure S8.** One-way partial dependence plots of 14 inputs: (a-e) elemental

composition, (f–h) proximate composition, and (i–k) biochemical compositions of biomass, and (l–n) HTC operating conditions to Yield_char prediction. **Figure S9.** One-way partial dependence plots of 14 inputs: (a–e) elemental composition, (f–h) proximate composition, and (i–k) biochemical composition of biomass, and (l–n) HTC operating conditions to N_char prediction. **Figure S10.** One-way partial dependence plots of 14 inputs: (a–e) elemental compositions, (f–h) proximate compositions, and (i–k) biochemical compositions of biomass, and (l–n) HTC operating conditions to O_char prediction. **Figure S11.** One-way partial dependence plots of 14 inputs: (a–e) elemental composition, (f–h) proximate composition, and (i–k) biochemical compositions of biomass, and (l–n) HTC operating conditions to C_char prediction. **Figure S12.** One-way partial dependence plots of 14 inputs: (a–e) elemental composition, (f–h) proximate composition, and (i–k) biochemical composition of biomass, and (l–n) HTC operating conditions to H_char prediction. **Figure S13.** Online system based on multi-target models for predicting and optimizing hydrochar (website link: <http://43.139.222.20:8000/>). **Figure S14.** Optimal operating conditions (the black stars in the plots) for hydrochar production with soybean straw of maximized Yield_char (a), C_char (d), H_char (e), and HHV_char (f), minimized N_char (b), and O_char (c) based on multi-target prediction system.

Additional file 2. Dataset #1–6 for predicting biochemical composition of biomass and yield, N content, O content, C content, H content, and HHV of the hydrochar.

Acknowledgements

The authors appreciate Miss. Qingyue Chen and Lihong Yang for assisting program implementation.

Author contributions

The first draft was written by WZ. Supervision, Conceptualization, and Project administration were addressed by LL and HL. Data collection and Methodology were performed by JZ and QL. Software, Formal analysis and Visualization were performed by WZ. Resources, Validation were performed by ZX. Writing-Review & Editing was performed by HP, LL, and HL. 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 Open Project of Xiangjiang Laboratory (22xj03003), and the Science and Technology Innovation Program of Hunan Province (2021RC4005, 2021GK1210).

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

¹School of Energy Science and Engineering, Central South University, Changsha 410083, China. ²Xiangjiang Laboratory, Changsha 410205, China. ³Hunan Modern Environmental Technology Co., LTD., Changsha 410000, China.

Received: 18 September 2023 Revised: 4 January 2024 Accepted: 11 January 2024

Published online: 01 March 2024

References

Buss W, Jansson S, Wurzer C, Mašek O (2019) Synergies between BECCS and biochar—maximizing carbon sequestration potential by recycling wood

ash. *ACS Sustain Chem Eng* 7:4204–4209. <https://doi.org/10.1021/acssuschemeng.8b05871>

- Deng Q, Lin B (2022) Automated machine learning structure-composition-property relationships of perovskite materials for energy conversion and storage. *Energy Mater* 1:100006. <https://doi.org/10.20517/energymater.2021.10>
- Duan P, Chang Z, Xu Y et al (2013) Hydrothermal processing of duckweed: effect of reaction conditions on product distribution and composition. *Bioresour Technol* 135:710–719. <https://doi.org/10.1016/j.biortech.2012.08.106>
- Duan PG, Yang SK, Xu YP et al (2018) Integration of hydrothermal liquefaction and supercritical water gasification for improvement of energy recovery from algal biomass. *Energy* 155:734–745. <https://doi.org/10.1016/j.energy.2018.05.044>
- DuBois M, Gilles KA, Hamilton JK et al (1956) Colorimetric method for determination of sugars and related substances. *Anal Chem* 28:350–356. <https://doi.org/10.1021/ac60111a017>
- Fang Y, Ma L, Yao Z et al (2022) Process optimization of biomass gasification with a Monte Carlo approach and random forest algorithm. *Energy Convers Manag* 264:115734. <https://doi.org/10.1016/j.enconman.2022.115734>
- Gao F, Shen Y, Brett Sallach J et al (2022) Predicting crop root concentration factors of organic contaminants with machine learning models. *J Hazard Mater* 424:127437. <https://doi.org/10.1016/j.jhazmat.2021.127437>
- He M, Cao Y, Xu Z et al (2022a) Process water recirculation for catalytic hydrothermal carbonization of anaerobic digestate: water-energy-nutrient nexus. *Bioresour Technol* 361:127694. <https://doi.org/10.1016/j.biortech.2022.127694>
- He M, Zhu X, Dutta S et al (2022b) Catalytic co-hydrothermal carbonization of food waste digestate and yard waste for energy application and nutrient recovery. *Bioresour Technol* 344:126395. <https://doi.org/10.1016/j.biortech.2021.126395>
- Hoekman SK, Broch A, Robbins C (2011) Hydrothermal carbonization (HTC) of lignocellulosic biomass. *Energy Fuels* 25:1802–1810. <https://doi.org/10.1021/ef101745n>
- Kim JY, Shin UH, Kim K (2023) Predicting biomass composition and operating conditions in fluidized bed biomass gasifiers: an automated machine learning approach combined with cooperative game theory. *Energy* 280:128138. <https://doi.org/10.1016/j.energy.2023.128138>
- Kirchner K, Zec J, Delibašić B (2016) Facilitating data preprocessing by a generic framework: a proposal for clustering. *Artif Intell Rev* 45:271–297. <https://doi.org/10.1007/s10462-015-9446-6>
- Leng L, Zhang W, Peng H et al (2020) Nitrogen in bio-oil produced from hydrothermal liquefaction of biomass: a review. *Chem Eng J* 401:126030. <https://doi.org/10.1016/j.cej.2020.126030>
- Leng L, Zhang W, Chen Q et al (2022a) Machine learning prediction of nitrogen heterocycles in bio-oil produced from hydrothermal liquefaction of biomass. *Bioresour Technol* 362:127791. <https://doi.org/10.1016/j.biortech.2022.127791>
- Leng L, Zhang W, Liu T et al (2022b) Machine learning predicting wastewater properties of the aqueous phase derived from hydrothermal treatment of biomass. *Bioresour Technol* 358:127348. <https://doi.org/10.1016/j.biortech.2022.127348>
- Leng S, Jiao H, Liu T et al (2022c) Co-liquefaction of Chlorella and soybean straw for production of bio-crude: effects of reusing aqueous phase as the reaction medium. *Sci Total Environ* 820:153348. <https://doi.org/10.1016/j.scitotenv.2022.153348>
- Leng L, Li T, Zhan H et al (2023) Machine learning-aided prediction of nitrogen heterocycles in bio-oil from the pyrolysis of biomass. *Energy* 278:127967. <https://doi.org/10.1016/j.energy.2023.127967>
- Li Y, Liu H, Xiao K et al (2019) Correlations between the physicochemical properties of hydrochar and specific components of waste lettuce: influence of moisture, carbohydrates, proteins and lipids. *Bioresour Technol* 272:482–488. <https://doi.org/10.1016/j.biortech.2018.10.066>
- 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 L, Flora JRV, Berge ND (2020b) Predictions of energy recovery from hydrochar generated from the hydrothermal carbonization of organic wastes.

- Renew Energy 145:1883–1889. <https://doi.org/10.1016/j.renene.2019.07.103>
- Li J, Zhang W, Liu T et al (2021a) 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, Zhu X, Li Y et al (2021b) Multi-task prediction and optimization of hydrochar properties from high-moisture municipal solid waste: application of machine learning on waste-to-resource. *J Clean Prod* 278:123928. <https://doi.org/10.1016/j.jclepro.2020.123928>
- Li H, Chen J, Zhang W et al (2023) Machine-learning-aided thermochemical treatment of biomass: a review. *Biofuel Res J*. 10:1786–1809. <https://doi.org/10.18331/BRJ2023.10.1.4>
- Liu A, Su Y, Nie W, Kankanhalli M (2017) Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Trans Pattern Anal Mach Intell* 39:102–114. <https://doi.org/10.1109/TPAMI.2016.2537337>
- Liu H, Basar IA, Nzihou A, Eskicioglu C (2021) Hydrochar derived from municipal sludge through hydrothermal processing: a critical review on its formation, characterization, and valorization. *Water Res* 199:117186. <https://doi.org/10.1016/j.watres.2021.117186>
- Liu Z, Cui Y, Wang J et al (2022) Multi-objective optimization of multi-energy complementary integrated energy systems considering load prediction and renewable energy production uncertainties. *Energy* 254:124399. <https://doi.org/10.1016/j.energy.2022.124399>
- Marzbali MH, Kundu S, Halder P et al (2021) Wet organic waste treatment via hydrothermal processing: a critical review. *Chemosphere* 279:130557. <https://doi.org/10.1016/j.chemosphere.2021.130557>
- Mu L, Wang Z, Wu D et al (2022) Prediction and evaluation of fuel properties of hydrochar from waste solid biomass: machine learning algorithm based on proposed PSO–NN model. *Fuel* 318:123644. <https://doi.org/10.1016/j.fuel.2022.123644>
- Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. *Front Neuro-robot*. <https://doi.org/10.3389/fnbot.2013.00021>
- 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>
- Putatunda S, Rama K (2018) A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. In: *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*. ACM press, New York, USA, pp 6–10
- Qureshi AS, Khan A, Zameer A, Usman A (2017) Wind power prediction using deep neural network based meta regression and transfer learning. *Appl Soft Comput* 58:742–755. <https://doi.org/10.1016/j.asoc.2017.05.031>
- Ribeiro MT, Singh S, Guestrin C (2016) “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM press, New York, USA, pp 1135–1144
- Rodrigues R, Souza D, Toebe M, Chuquel A (2023) Sample size and Shapiro–Wilk test: AN analysis for soybean grain yield. *Eur J Agron* 142:126666. <https://doi.org/10.1016/j.eja.2022.126666>
- Seo MW, Lee SH, Nam H et al (2022) Recent advances of thermochemical conversion processes for biorefinery. *Bioresour Technol* 343:126109. <https://doi.org/10.1016/j.biortech.2021.126109>
- Shafizadeh A, Shahbeik H, Rafiee S et al (2023) Machine learning-based characterization of hydrochar from biomass: implications for sustainable energy and material production. *Fuel* 347:128467. <https://doi.org/10.1016/j.fuel.2023.128467>
- Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52:591–611. <https://doi.org/10.2307/2333709>
- Sheng L, Wang X, Yang X (2018) Prediction model of biocrude yield and nitrogen heterocyclic compounds analysis by hydrothermal liquefaction of microalgae with model compounds. *Bioresour Technol* 247:14–20. <https://doi.org/10.1016/j.biortech.2017.08.011>
- Shi N, Liu Q, He X et al (2019) Molecular structure and formation mechanism of hydrochar from hydrothermal carbonization of carbohydrates. *Energy Fuels* 33:9904–9915. <https://doi.org/10.1021/acs.energyfuels.9b02174>
- Toptas Tag A, Duman G, Yanik J (2018) Influences of feedstock type and process variables on hydrochar properties. *Bioresour Technol* 250:337–344. <https://doi.org/10.1016/j.biortech.2017.11.058>
- Xiong T, Cui J, Hou Z et al (2023) Prediction of arsenic adsorption onto metal organic frameworks and adsorption mechanisms interpretation by machine learning. *J Environ Manage* 347:119065. <https://doi.org/10.1016/j.jenvman.2023.119065>
- Xu D, Lin G, Liu L et al (2018) Comprehensive evaluation on product characteristics of fast hydrothermal liquefaction of sewage sludge at different temperatures. *Energy* 159:686–695. <https://doi.org/10.1016/j.energy.2018.06.191>
- Xu Z, Ma X, Zhou J et al (2022) The influence of key reactions during hydrothermal carbonization of sewage sludge on aqueous phase properties: a review. *J Anal Appl Pyrolysis* 167:105678. <https://doi.org/10.1016/j.jaap.2022.105678>
- Yap BW, Sim CH (2011) Comparisons of various types of normality tests. *J Stat Comput Simul* 81:2141–2155. <https://doi.org/10.1080/00949655.2010.520163>
- Yu J, Zhong X, Huang Z et al (2023) Mining the synergistic effect in hydrothermal co-liquefaction of real feedstocks through machine learning approaches. *Fuel* 334:126715. <https://doi.org/10.1016/j.fuel.2022.126715>
- Yuan T-Q, Sun S-N, Xu F, Sun R-C (2011) Characterization of lignin structures and lignin-carbohydrate complex (LCC) linkages by quantitative ¹³C and 2D HSQC NMR spectroscopy. *J Agric Food Chem* 59:10604–10614. <https://doi.org/10.1021/jf2031549>
- 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 W, Li J, Liu T et al (2021) Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. *Bioresour Technol* 342:126011. <https://doi.org/10.1016/j.biortech.2021.126011>
- Zhang B, Biswal BK, Zhang J, Balasubramanian R (2023a) Hydrothermal treatment of biomass feedstocks for sustainable production of chemicals, fuels, and materials: progress and perspectives. *Chem Rev*. <https://doi.org/10.1021/acs.chemrev.2c00673>
- Zhang W, Chen Q, Chen J et al (2023b) Machine learning for hydrothermal treatment of biomass: a review. *Bioresour Technol* 370:128547. <https://doi.org/10.1016/j.biortech.2022.128547>
- Zhang X, Liu H, Yang G et al (2023c) Comprehensive insights into the application strategy of kitchen waste derived hydrochar: Random forest-based modelling. *Chem Eng J* 469:143840. <https://doi.org/10.1016/j.cej.2023.143840>
- Zhang S, Luo X, Mai B (2024) Multi-task machine learning models for simultaneous prediction of tissue-to-blood partition coefficients of chemicals in mammals. *Environ Res* 241:117603. <https://doi.org/10.1016/j.envres.2023.117603>
- Zhu X, Liu B, Sun L et al (2023) Machine learning-assisted exploration for carbon neutrality potential of municipal sludge recycling via hydrothermal carbonization. *Bioresour Technol* 369:128454. <https://doi.org/10.1016/j.biortech.2022.128454>