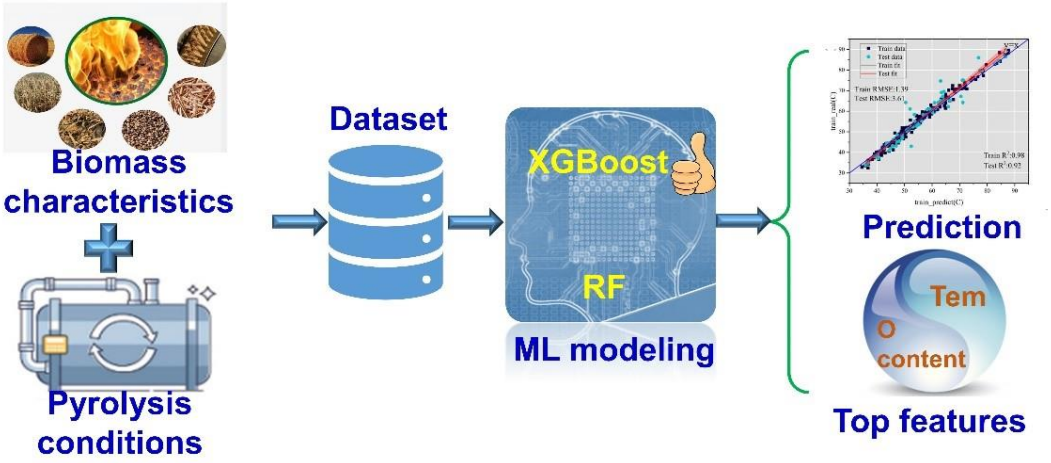
Using Machine Learning to Predict Biochar Yield and Carbon Content: Enhancing Efficiency and Sustainability in Biomass Conversion

Qingsheng Xu,[a,b] Long Du,[a] and Rui Deng [a,b,*]

** Corresponding author: dengrui@hfut.edu.cn (R. Deng)*

DOI: 10.15376/biores.19.3.6545-6558

GRAPHICAL ABSTRACT



Using Machine Learning to Predict Biochar Yield and Carbon Content: Enhancing Efficiency and Sustainability in Biomass Conversion

Qingsheng Xu,[a,b] Long Du,[a] and Rui Deng [a,b,*]

The production of biochar through pyrolysis of biomass is expected to reduce dependence on traditional energy sources and mitigate global warming. However, current predictive models for biochar yield and composition are computationally intensive, complex, and lack accuracy for extrapolative scenarios. This study utilized machine learning to develop predictive models for biochar yield and carbon content based on pyrolysis data from lignocellulosic biomass. Assessing the importance of input features revealed their significant role in predicting biochar properties. The findings indicate that eXtreme Gradient Boosting (XGBoost) algorithms can accurately forecast biochar yield and carbon content based on biomass characteristics and pyrolysis conditions. This research contributes new insights into understanding biomass pyrolysis and enhancing biochar production efficiency.

DOI: 10.15376/biores.19.3.6545-6558

Keywords: Biomass; Biochar; Pyrolysis; Machine learning; Feature importance

Contact information: a: School of Resources and Environmental Engineering, Hefei University of Technology, Hefei, Anhui 230009, China; b: Key Laboratory of Nanominerals and Pollution Control of Anhui Higher Education Institutes, Hefei University of Technology, Hefei, Anhui 230009, China;
** Corresponding author: dengrui@hfut.edu.cn (R. Deng)*

INTRODUCTION

To cope with the dual pressures of increasing energy demand and environmental pollution caused by fossil fuels worldwide, biomass has received much attention as one of the most abundant and promising renewable raw materials (Antar *et al*. 2021; Hassan *et al*. 2024). The global biomass reservoir stands at an impressive 540 million tons of standard coal, with 280 million tons suitable for energy production (Xu *et al*. 2021). Biomass lends itself to conversion into bioenergy through biological and thermochemical means (Wang and Wu 2023).

Pyrolysis is the dominant technology for converting solid biomass into bio-oils, gaseous products and biochar in anoxic or anoxic environments. Among them, biochar has been used in multidisciplinary fields in recent years due to its high surface area, rich carbon content, and surface functional groups (Li *et al*. 2023). The yield and characteristics of biochar are affected by the nature of the biomass and the specific situation of the pyrolysis process, and biochar with different characteristics has corresponding applications. Notably, the most favorable settings for biochar manufacture are not universal but are tailored to the specific biochar attributes and intended end-uses sought after (Tomczyk *et al*. 2020). Safarian (2023) conducted an extensive review of numerous environmentally sustainable techniques for biochar generation, examining them through the lens of process efficiency. Their findings highlighted a substantial variation in both the output quantities and the

qualitative features of the products derived from thermochemical biomass conversion methods. This observed disparity can largely be attributed to disparities in raw material compositions (feedstocks), operational parameters, and the specific methodologies employed during application (Safarian 2023). The studies of Leng and Huang (2018) further revealed that the variables in biochar manufacturing, encompassing the composition and traits of the original biomass as well as the precise pyrolysis circumstances, exert diverse impacts on the resultant biochar stability. Of these factors, the pyrolysis temperature emerges as the paramount determinant in regulating biochar stability. Elevated temperatures during processing render biochar notably consistent, accomplishing the biomass variety and supplementary pyrolysis conditions inconsequential. Conversely, in scenarios where milder temperatures are employed, alternative process parameters assume a more significant role in influencing biochar stability (Leng and Huang 2018). As a result, scholars have delved into biochar manufacturing processes that use large amounts of biomass and sought strategies to increase biochar production to meet industrial demand. While focusing on the critical role of biochar in carbon capture and sequestration, the focus is not just on maximizing yield; the carbon content of biochar itself is equally important, underscoring the dual focus of research efforts (Deng *et al*. 2024).

Measuring these properties with traditional analytical methods proves complex, time-consuming, and resource-intensive. Additionally, understanding optimal conditions and mechanisms for biochar production and utilization remains elusive, with few comprehensive studies on biomass characteristics and process parameters' effects on biochar properties. Biomass comprises cellulose, hemicellulose, and lignin, reacting differently during pyrolysis based on temperature and time (Chen *et al*. 2022). The process of biomass pyrolysis encompasses intricate transformations, both chemical and physical, including dehydration and depolymerization, which significantly mold the biochar's characteristics, notably its elemental composition and pH (Vuppaladadiyam *et al*. 2023). Nonetheless, deciphering the complexity of biomass pyrolysis remains a formidable task, owing to the intricate, non-linear correlations that exist between the inherent qualities of the biomass and the resultant biochar characteristics. Current biomass pyrolysis models lack reliability and generalizability due to limited experimental data. Consequently, employing big data analytics, machine learning algorithms, and data mining techniques to scrutinize the pyrolysis behavior of materials, in conjunction with a thorough examination of the original feedstock attributes and pyrolysis parameters, is vital for conducting a comprehensive evaluation of how these variables collectively impact biochar output and elemental composition (Zhu *et al*. 2019).

Machine learning is capable of finding patterns and extracting insights from large data sets. By using historical data on feedstock properties, process conditions, and properties of the resulting biochar, machine learning models can learn and predict how changes in input parameters affect output metrics. This not only deepens our understanding of the basic mechanisms of biochar production, but also empowers stakeholders to optimize processes, thereby increasing efficiency and reducing costs. RF and XGBoost both fall within the realm of Ensemble Learning, which enhances the generalization and robustness of a solitary learner by integrating the predictions of multiple base learners for predicting complex nonlinear targets in engineering. Although RF is more frequently used to simulate pyrolysis, XGBoost has higher efficiency and could reduce over-fitting (Cruz *et al*. 2022; Yang *et al*. 2023).

The yield of biochar helps to assess the commercial practicality and scalability of the manufacturing process, and the carbon content is closely related to the carbon capture capacity of the material and its continuous enhancement of soil health over time. In this study, two advanced machine learning algorithms, RF and XGBoost, were used to verify their efficacy in predicting biochar yield and carbon content and to explore the different significance of factors such as raw material properties and pyrolysis environment during pyrolysis. With the XGBoost model providing accurate predictions and revealing new perspectives behind those predictions, this study provides a comprehensive understanding of biochar production through biomass pyrolysis, thereby providing engineers with strategic directions for optimizing biochar yield or its carbon enrichment.

MATERIALS AND METHODS

Data Collection and Preprocessing

During data collection and preprocessing, searches were conducted on Web of Sciences using the keywords "biochar" and "pyrolysis". This extensive exploration led to the assembly of 355 different datasets from a compilation of 23 academic publications that are the basis for building our data-intensive predictive model, summarized in "Supplementary Materials".

To ensure broad applicability, the analysis encompassed a range of biochar attributes, specifically: (1) Elemental composition comprising carbon (C), hydrogen (H), oxygen (O), nitrogen (N), sulfur (S), fixed carbon (FC), volatile matter (VM), and Ash content.; (2) Lignocellulosic composition, including cellulose (CL), hemicellulose (HC), and lignin (LG); (3) Characteristics of the biomass feedstock, such as type and sample mass (SM); (4) Pyrolysis operational parameters, namely reaction temperature (Temp), residence time (RT), and heating rate (HR); (5) Yield of the resulting biochar and its Higher heating value (HHV). However, certain variables were deliberately excluded from the machine learning model. These variables had minimal impact on the torrefaction process and including them would unduly complicate the model structure and elevate computational challenges without significant benefit.

In instances where data availability ranged from over 80% to below 100% for both inputs and outputs, the missing entries were supplanted with the average values pertinent to each respective variable. Consequently, the refined dataset employed for model development comprised 15 predictor variables—pertaining to the composition of biomass feedstock and pyrolysis conditions, including H, O, N, S, FC, VM, ash content, CL, HC, LG, SM, HHV, Temp, RT, and HR—and 2 response output variables associated with biochar characteristics (carbon content and yield). Prior to modelling, the dataset underwent preprocessing to identify and address any missing or duplicated records. To ensure data integrity, a dual approach combining Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a robust clustering technique, and the Isolation Forest (iForest) algorithm, rooted in decision tree models, was implemented to identify, and exclude outliers. Given the notable variability and unstable distribution of Heating Rate (HR), Residence Time (RT), and Sample Mass (SM), which could potentially undermine model performance or hinder convergence, as per Zhou *et al*. (2021), these variables were omitted. Thus, the final dataset tailored for model construction consisted of 12 predictor variables linked to biomass feedstock composition and torrefaction parameters (H, O, N,

S, FC, VM, Ash, CL, HC, LG, HHV, and Temp) and maintained the same 2 output variables linked to biochar attributes (Carbon content and biochar yield).

Considering the extensive variation and scales across different feature dimensions, the dataset underwent preprocessing with StandardScaler from the scikit-learn library in Python to standardize the variables, adhering to the methodology outlined by Pedregosa *et al*. (2011). Furthermore, to delve into the relationships between the predictors and the predicted outcomes, Pearson's Correlation Coefficient (PCC) was employed. With the aim to facilitate quicker algorithm convergence and enhance prediction accuracy by achieving a consistent scale among all variables, the empirical dataset was subjected to Min-Max normalization, transforming values to fall within the interval of 0 to 1, utilizing Eq. 1 for this transformation process.

$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X_i is the value of the input feature; X_i' is the normalized value of initial X_i; X_{min}, and X_{max} are the minimum and maximum values of X_i, respectively.

Construction and Evaluation of Machine Learning Models

In this study, two machine learning algorithms were employed, namely XGBoost and Random Forest (RF), to construct models predicting carbon content and biochar yield. These models utilized 17 different variables as input features. The implementation of these algorithms was facilitated through Python, leveraging libraries such as sklearn for general machine learning tasks, XGBoost specifically for the XGBoost algorithm, and keras for potential deep learning components. Prior to the modeling process, the dataset was divided randomly into two subsets: 80% of the data was allocated for training the models, while the remaining 20% served as a test dataset to assess the predictive accuracy of the models developed.

In order to refine the predictive power of our machine learning models, a 5-fold cross-validation methodology was adopted to calibrate hyperparameters during the training phase. The XGBoost model is grounded in the Gradient Boosting Decision Tree (GBDT) algorithm, which is renowned for its capability to adeptly and proficiently manage both regression and classification tasks. The hyperparameter tuning process for XGBoost entailed meticulous consideration of key elements such as the quantity of decision trees, the utmost tree depth, and the learning rate (Chen and Guestrin 2016). Concurrently, Random Forest (RF), as another decision tree-driven machine learning architecture, delves into the intricate, nonlinear correlations between inputs and outputs. Hyperparameter optimization for RF was carried out by carefully modulating the count and the maximal depth of the decision trees, thereby further enhancing model performance (Breiman 2001).

Post hyperparameter optimization, the machine learning models underwent retraining with the designated training dataset to derive the most efficacious versions capable of precisely forecasting carbon content and biochar yield. To rigorously assess the predictive precision and the broader applicability of these finely-tuned models, a trio of metrics was employed: the coefficient of determination (R^2) for goodness of fit, the root mean square error (RMSE) to quantify prediction errors, and the mean squared error (MSE) to offer a complementary perspective on the models' accuracy, all utilizing the separate test dataset. The R^2, RMSE, and MSE were calculated with Eqs. 2, 3, and 4, respectively.

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

$$MSE = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{n} \quad (4)$$

where n is the number of total data points for the training set or testing set; y_i is the experimental value of the output variable; $\hat{y}_i$ is the corresponding predicted value; $\overline{y_i}$ is the average value of output variable.

Method for Feature Importance Analysis

In an effort to decipher the relative significance and interplay among diverse input variables affecting the output in inherently opaque machine learning models, a feature importance analysis was initiated to illuminate the pivotal aspects governing the intricate biological mechanisms at play. This investigation employed the Shepley Additive Explanations (SHAP), a game-theoretic explanation framework for machine learning, to systematically rank and assess these features. The SHAP analysis methodically attributed an individual, quantifiable measure of predicted importance to every input feature, thereby encapsulating its localized impact on the prediction of the output variable (Lundberg and Lee 2017).

Features were deemed more influential in forecasting the carbon content and biochar yield within the biochar pyrolysis system as their positive importance values escalated, signifying a stronger predictive contribution. Conversely, a feature associated with a negative importance value suggested that randomizing this variable could potentially enhance prediction outcomes. Additionally, features with an importance score nearing zero had a negligible impact on the overall model performance, indicating their limited relevance to the prediction task.

RESULTS AND DISCUSSION

Statistical Analysis of Multiple Variables

An ensemble of 355 datasets was assembled from 23 publications documenting the proximate composition of biomass, pyrolysis parameters, and the elemental composition, yield, and Higher Heating Value (HHV) of derived biochars. Figure 1 illustrates the span of this amassed data. Analysis of these datasets showed the typical carbon and oxygen concentrations in biomass to be 54.81 wt% and 34.88 wt%, respectively, implying a slight dominance of carbon. Characteristically, biomass is rich in volatile matter (VM) yet scant in ash, averaging an HHV around 20.41 MJ/kg. Regarding the chemical composition, lignin content across samples varied widely, primarily spanning from 1.4% to 98.64%. Cellulose and hemicellulose fractions, both composed of carbohydrate structures, ranged from 0.07% to 57.39% and 0.86% to 56.29%, respectively. Notably, lignin, distinguished by its aromatic structure, exhibited the greatest thermal stability among these components. Pyrolysis temperatures reported varied extensively from 180 to 900 °C, with most studies concentrating within the 200 to 300 °C range. Residence times (RT) were typically brief, extending from 1.0 min to 2 h, though certain experiments prolonged RTs to enhance biochar uniformity. The aggregate mean yield of biochar production was recorded at 56.2 wt%.

According to the Pearson correlation coefficient matrix depicted in Fig. 2, the strength and direction of linear associations between any pair of variables were assessed. Here, the p-value served as an indicator of the statistical confidence in these linear relationships. Notably, an inverse pattern emerged between the two output measures; the biochar yield exhibited a strong negative correlation with fixed carbon (FC), evidenced by a correlation coefficient of -0.75, whereas carbon (C) content demonstrated a distinct positive correlation with FC, with a coefficient of 0.59. This inverse implies that conditions favoring one output tend to be detrimental to the other. The pyrolysis temperature followed a similar trend, facilitating the degradation of organic substances into volatiles and promoting the formation of a carbon-rich residue, leading to reduced yields but enhanced C content. Apart from thermal conditions, biochar yield revealed a positive association with cellulose content and a contrasting negative correlation with hemicellulose and lignin contents, with the magnitude of these correlations decreasing in that order. This observation aligns with previous findings, suggesting that at lower temperatures, hemicellulose and cellulose are more readily broken down compared to lignin, resulting in lower conversion rates to biochar (Zhang *et al*. 2024). This is due to the preferential degradation of lignin, hemicellulose and cellulose at the temperature level, where hemicellulose begins to decompose at lower temperatures around 170 to 300 °C, followed by the initial stages of cellulose degradation at similar temperatures, whereas the lignin component undergoes thermal decomposition over a wider temperature range between 170 and 560 °C, with the more advanced stages of cellulose degradation occurring at higher temperatures between 300 and 360 °C, with more advanced stages of cellulose degradation occurring at higher temperatures between 300 and 36 0°C (Zhang *et al*. 2021). Furthermore, the Pearson correlation matrix also enabled the exploration of relationships among input variables.

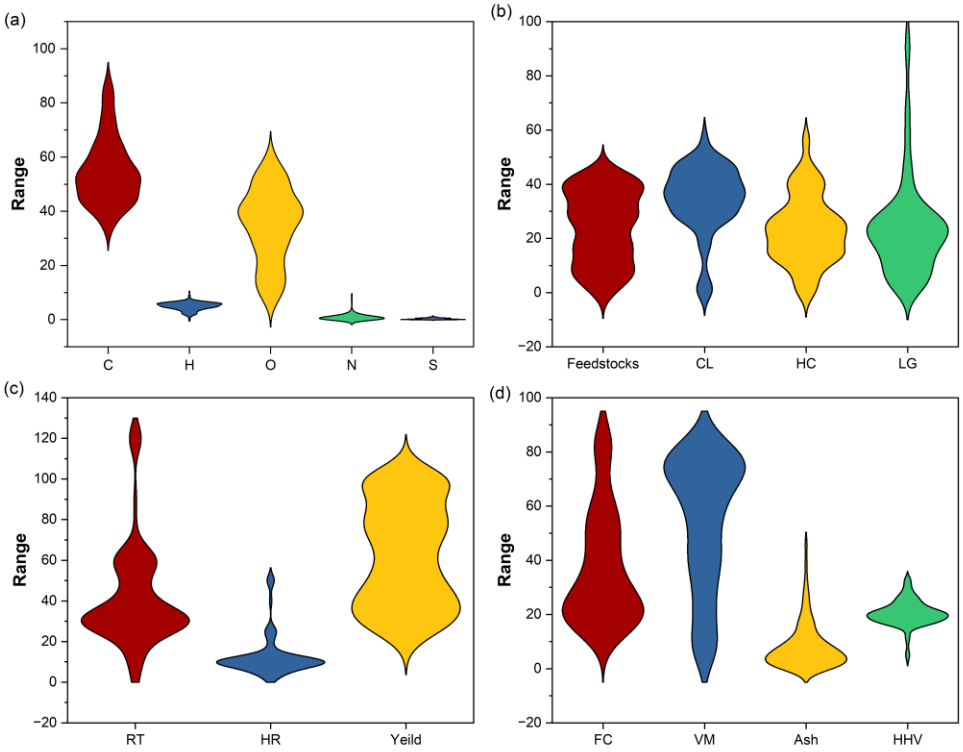


Fig. 1. Violin plot of (a) elemental composition, (b) proximate composition and HHV of biochar, (c) proximate composition of biomass, and (d) torrefaction conditions and biochar yield

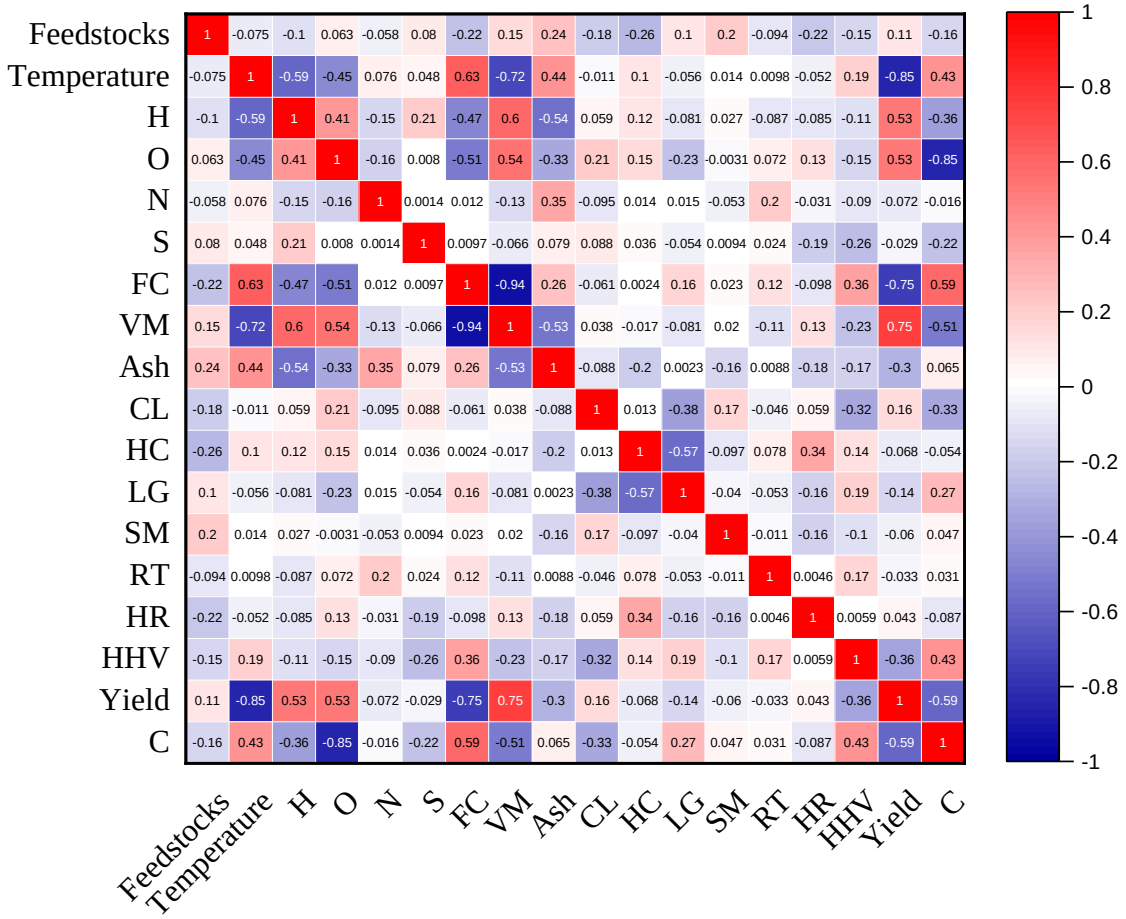


Fig. 2. Pearson correlation matrix between any two variables

As an illustration, carbon content was positively correlated with the Higher Heating Value (HHV), with a correlation coefficient of 0.43, highlighting a direct link between these properties. Collectively, these correlations provide insights into the intricate dependencies within the biochar production process, influenced by both feedstock characteristics and process parameters.

Model Development and Evaluation

RF and XGBoost are cutting-edge analytical tools in the realm of data mining, frequently employed for unraveling intricate non-linear associations among variables (Souaissi *et al*. 2023). This is due to the fact that they are powerful and flexible enough to handle a wide range of data types and sizes while providing good predictive performance. They are commonly used in situations where the goal is to achieve high accuracy without sacrificing much interpretability (Banik and Biswas 2024). Consequently, within the scope of this investigation, these sophisticated models were harnessed to forecast both the carbon content and the yield of biochar, capitalizing on their prowess in navigating complex interdependencies.

Model hyperparameters are key factors that affect the accuracy and generalisation of model predictions. Therefore, setting reasonable hyperparameters for the model is an important step in constructing the model (Bischl *et al*. 2023). To ensure the development

of high-performing machine learning models, hyperparameter optimization was carried out leveraging the training dataset. This process entailed employing a grid search methodology in conjunction with k-fold cross-validation, with the primary goal of minimizing Mean Squared Error (MSE) values. Lower MSE scores were indicative of enhanced predictive capabilities for the models, thus guiding us towards the most advantageous hyperparameter configurations (Zhao *et al*. 2024). Moreover, employing the average of multiple folds in n-fold cross-validation mitigates the variability in model assessment that can arise from arbitrary data partitioning, thereby diminishing the likelihood of inflated performance estimates due to chance and reducing the risk of overfitting subsequent to hyperparameter tuning. For the XGBoost model, Max depth, Min child weight, and subsample bytree are the key hyperparameters that determine its prediction accuracy; whereas the performance of the Random Forest model is highly dependent on the hyperparameter settings of n_estimators, Max depth, Min samples leaf, and Min samples split. Therefore, in this section, a grid search tuning approach was used to search the entire hyperparameter space within the model hyperparameter space given above and determine the model hyperparameter combination that possesses the best predictive performance through the MSE results of the 5-fold cross-validation in order to improve the accuracy and generalisation of the XGBoost model for predicting the biochar yield and C content. The results of well-tuned hyperparameters are listed in Table 1.

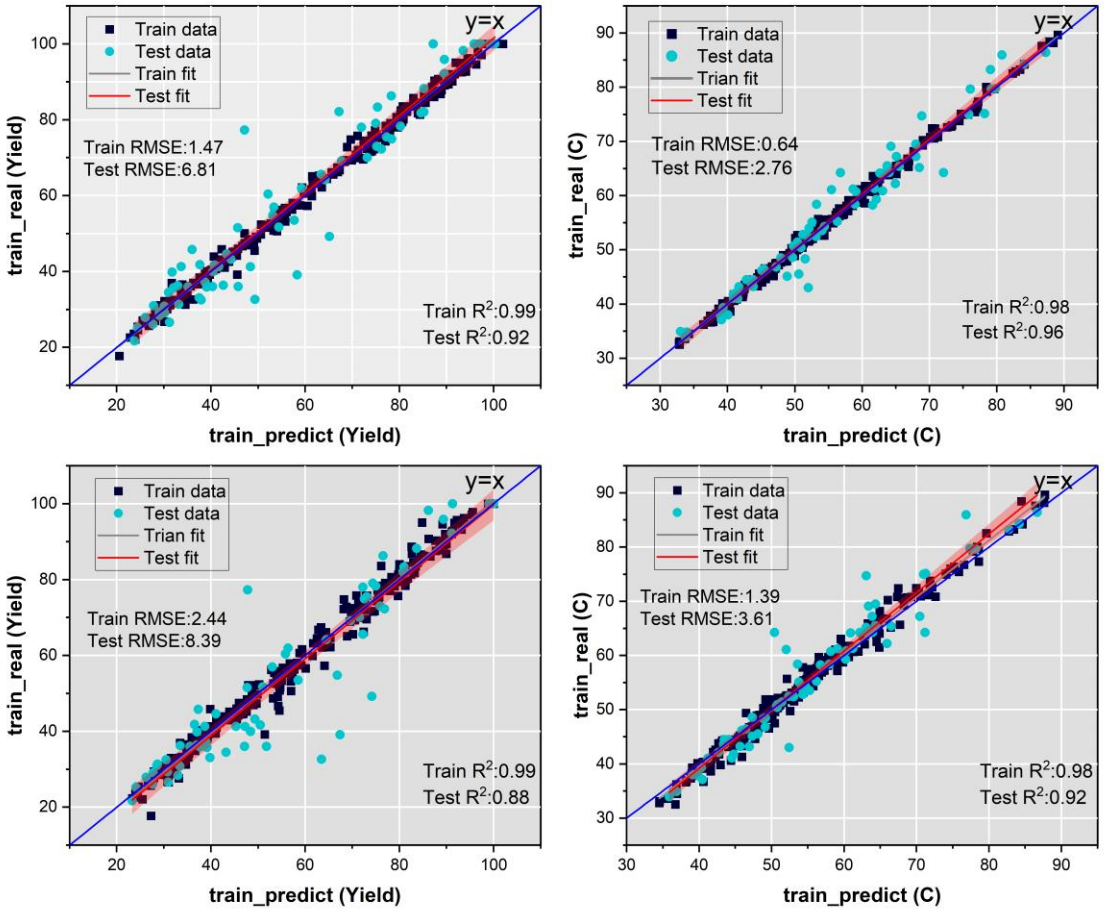


Fig. 3. Comparison of predicted biochar yield/C content and actual values using collected data: (a-b) XGBoost modeling, (c-d) RF modeling

Table 1. Well-tuned Hyper-parameters of ML Models for Predicting C-content and Biochar Yield

Model	Hyperparameters	Range	Optimized Value
RF	*n*_estimators	10 to 80	50
	Max depth	1 to 9	9
	Min samples leaf	1 to 9	1
	Min samples split	2 to 9	3
XGBoost	Max depth	1 to 10	5
	Min child weight	1 to 8	1
	Sub sample	0.1 to 0.9	0.7
	Col sample by tree	0.1 to 0.9	0.8
	Learning rate	0.1 to 0.9	0.25
	Gamma	0 to 0.5	0
	Reg_alpha	0 to 0.5	0.001
	Reg_lambda	0 to 0.1	0.1

In pursuit of refining the models to their optimal state, a variety of machine learning models were recalibrated utilizing the training dataset, now equipped with meticulously adjusted hyperparameters. Subsequently, their efficacy was gauged against the test dataset. The outcomes illuminated the proficiency of both XGBoost and RF in anticipating biochar yield and carbon content, as attested by their high mean training R^2 scores (0.98 and 0.99, respectively) and similarly robust mean testing R^2 scores (0.91 for XGBoost and 0.94 for RF). Given that XGBoost marginally outperformed RF in terms of simulation accuracy, it was elected as the prime candidate for the succeeding feature importance examination.

Within the confines of this research, XGBoost excelled in forecasting both biochar yield and carbon content, exhibiting the utmost values for both training and testing R^2, alongside the minimal values for training and testing RMSE at 1.06 and 4.79, respectively (depicted in Figs. 3(a) and (b)). This distinguished XGBoost as a highly resilient machine learning model for biochar yield predictions, boasting exceptional capability to generalize beyond the training data. Given its standout performance, there's a compelling case for a deeper exploration into the underlying dynamics. Consequently, the SHAP method was employed to elucidate the model's workings, thereby capitalizing on XGBoost's robust predictive power.

Interpretability Results

To delve deeper into the interpretability of machine learning models, particularly the best-performing XGBoost model in the study, a comprehensive feature importance and SHapley Additive exPlanations (SHapley Additive exPlanations) analysis was carried out. These assessments cover the full range of input attributes, including the elemental and approximate composition of the biomass feedstock, as well as the operational parameters of the pyrolysis process. In this case, feature importance becomes a key diagnostic tool, making it possible to measure the role that each predictor plays in forming the model's predictions of the target variables (in this case, biochar yield and carbon content). By elucidating the correlation between each feature and the outcome, one can gain a clearer understanding of the underlying mechanisms that drive the model's predictive accuracy (Saarela and Jauhiainen 2021). As presented in Fig. 4(a), given its critical role in regulating the carbonization and volatilization processes that occur in biochar formation, temperature

was the most dominant factor affecting the elemental composition of biochar (Su and Jiang 2024). Furthermore, temperature emerged as the paramount influence on both FC and VM, while the ash content of the resultant biochar directly mirrored that of the initial biomass, which is consistent with the patterns observed in Pearson correlation coefficients. Indeed, temperature acts as the chief catalyst in the pyrolytic reactions, exerting a substantial sway over biochar yield. An increase in the pyrolysis temperature triggered profound alterations in the biomass's physical and chemical characteristics, triggering the liberation of volatile components and consequently leading to a diminished yield of biochar (Tomczyk *et al*. 2020). Concurrently, the rise in temperature facilitated the progressive carbonization of the residual solid biomass, thereby exerting a profound effect on the compositional makeup and structural attributes of the biochar produced.

Nonetheless, the outcomes from PCC and feature important analyses diverge in their implications. A case in point is that PCC suggests the H and O contents in the biomass as the principal determinants of biochar yield, whereas feature importance highlights temperature as the vital element. This disparity stems from the differing lenses through which PCC and feature importance view variable relationships. PCC essentially quantifies linear associations between pairs of variables, and thus, it may fall short in fully encapsulating interactions that exhibit non-linear patterns or higher degrees of intricacy. Conversely, feature importance adopts a broader perspective, assessing the cumulative impact of each feature on the predictive efficacy of the machine learning model without being constrained by assumptions of linearity. It comprehensively evaluates contributions, even in the presence of complex, non-linear relationships, offering a more nuanced understanding of the relative significance of each feature (Rengasamy *et al*. 2022).

Illustrated in Figs. 4(b) and 4(c), SHAP analysis elucidates the individual impact of each input feature on biochar yield results. By assigning a quantitative SHAP value to each data entry, it measures the extent to which each input variable contributes to the predicted output. Positive SHAP scores represent features that enhance the predicted value, while color coding (red for high and blue for low original eigenvalues) helps to understand the magnitude and direction of the impact visually. Both temperature and oxygen (O) content adversely affect biochar yield and carbon content, and temperature further dominates the approximate analysis of biochar properties. Although increasing pyrolysis temperature enhances the carbon stability in biochar by increasing the decomposition of biomass, it is negatively correlated with biochar yield. In contrast, a higher calorific value (HHV) in biomass indicates a higher carbon-oxygen ratio, which is more conducive to the pyrolysis process and thus increases biochar production (Seow *et al*. 2022). This reflects that biomass with higher energy density, which is characterized by higher carbon content and less oxygen, can be more efficiently converted into biochar during pyrolysis, resulting in higher yields. More specifically, in the biochar manufacturing process, temperature regulation and oxygen management are two core elements that directly affect the yield of biochar and its carbon content. An increase in temperature accelerates the pyrolysis process of biomass, leading to the precipitation of volatiles and the conversion of biomass into biochar with a high density carbon structure (Vuppaladadiyam *et al*. 2022). However, high temperatures enhance the carbon stability of the biochar, but they reduce the overall biochar yield due to the conversion of large amounts of organic matter into gases and other non-solid products. The involvement of oxygen then catalyzes the oxidation reaction, consuming the combustible fraction of the biomass and reducing the carbon source for biochar formation. Therefore, limiting the supply of oxygen during the pyrolysis stage prevents complete oxidation of the biomass, thus retaining more carbon for biochar generation (Zou *et al*.

2024). In addition, the calorific value (HHV) of the feedstock biomass is also an important factor. A high HHV often implies a high carbon-oxygen ratio, *i.e.*, the biomass contains a high energy density inside, which is reflected in a high carbon and low oxygen content; such biomass is more favorable to be converted into biochar under pyrolysis conditions because its high carbon and low oxygen characteristics are conducive to the construction of a stable carbon skeleton (Liu *et al*. 2023).

In summary, the combination of feature importance and SHAP analyses proves instrumental in pinpointing the salient factors shaping biochar attributes during pyrolysis. Specifically, temperature alongside biomass properties, notably O and ash content, stand out as paramount influencers of biochar characteristics. Elevating the temperature accelerates both the release of volatile components and the carbonization process in biomass, thereby driving transformative shifts in the resultant biochar's properties.

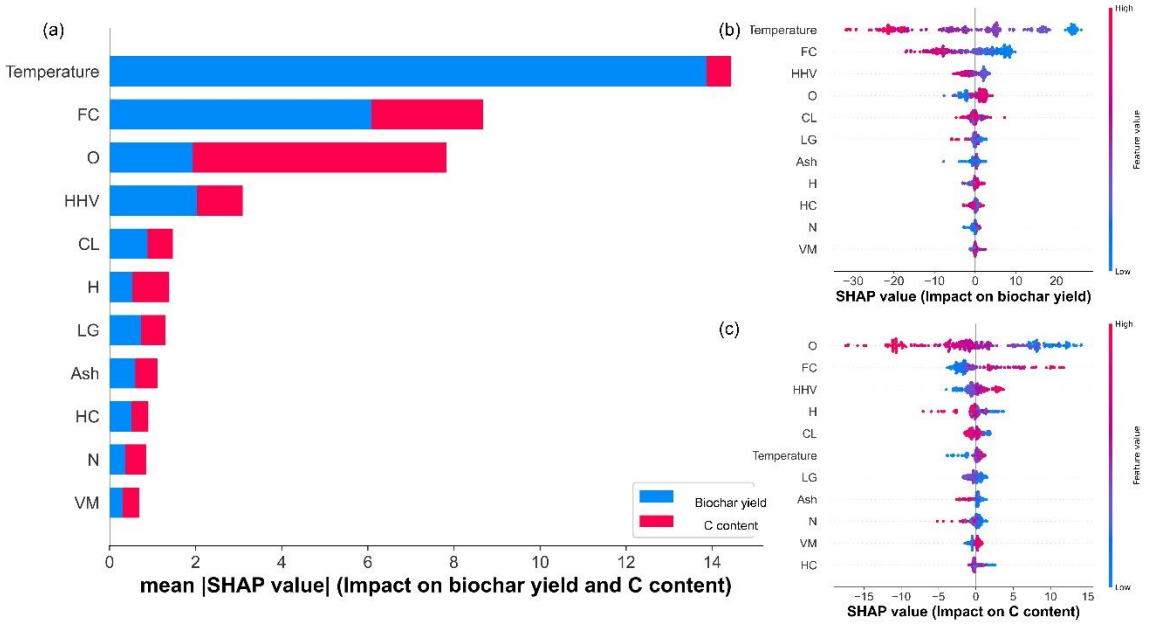


Fig. 4. Effects of input variables on biochar yield and C content based on XGBoost model using (a) feature importance and (b-c) SHAP analyses.

CONCLUSIONS

The production and characteristics of biochar are governed by the choice of biomass and the specifics of the pyrolysis process.

1. This study evaluated two machine learning algorithms to anticipate biochar yields and carbon content *via* XGBoost models and demonstrated the utmost predictive precision, boasting R^2 scores ranging from 0.92 to 0.98.
2. Through meticulous examinations of feature importance and SHAP analyses, it was elucidated that temperature plays a pivotal role in dictating biochar yield, while O-content is a crucial determinant of C-content within the biochar.
3. Collectively, the XGBoost model provided accurate predictions of biochar yield and C content and revealed the effects of feedstock and pyrolysis temperature on

pyrolysis. This study has provided a comprehensive understanding of biochar production through biomass pyrolysis, thus providing engineers with a strategic direction to optimise biochar yield or its carbon enrichment.

ACKNOWLEDGEMENTS

This research was funded by the Natural Science Foundation of Anhui Province (2308085QD122) and Fundamental Research Funds for the Central Universities (JZ2023HGTA0198).

REFERENCES CITED

Antar, M., Lyu, D., Nazari, M., Shah, A., Zhou, X., and Smith, D. L. (2021). "Biomass for a sustainable bioeconomy: An overview of world biomass production and utilization," *Renewable and Sustainable Energy Reviews* 139, article 110691. DOI:10.1016/j.rser.2020.110691

Banik, R., and Biswas, A. (2024). "Enhanced renewable power and load forecasting using RF-XGBoost stacked ensemble," *Electrical Engineering*. DOI: 10.1007/s00202-024-02273-3

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A. L., *et al*. (2023). "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery* 13(2), article e1484. DOI: 10.1002/widm.1484

Breiman, L. (2001). "Random forests," *Machine Learning* 45(1), 5-32.

Chen, D., Cen, K., Zhuang, X., Gan, Z., Zhou, J., Zhang, Y., and Zhang, H. (2022). "Insight into biomass pyrolysis mechanism based on cellulose, hemicellulose, and lignin: Evolution of volatiles and kinetics, elucidation of reaction pathways, and characterization of gas, biochar and bio-oil," *Combustion and Flame* 242, article 112142. DOI:10.1016/j.combustflame.2022.112142

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting," in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cruz, I. A., Chuenchart, W., Long, F., Surendra, K., Andrade, L. R. S., Bilal, M., Liu, H., Figueiredo, R. T., Khanal, S. K., and Ferreira, L. F. R. (2022). "Application of machine learning in anaerobic digestion: Perspectives and challenges," *Bioresource Technology* 345, article 126433. DOI: 10.1016/j.biortech.2021.126433

Deng, X., Teng, F., Chen, M., Du, Z., Wang, B., Li, R., and Wang, P. (2024). "Exploring negative emission potential of biochar to achieve carbon neutrality goal in China," *Nature Communications* 15(1), article 1085. DOI: 10.1038/s41467-024-45314-y

Hassan, Q., Viktor, P., J. Al-Musawi, T., Mahmood Ali, B., Algburi, S., Alzoubi, H. M., Khudhair Al-Jiboory, A., Zuhair Sameen, A., Salman, H. M., and Jaszczur, M. (2024). "The renewable energy role in the global energy transformations," *Renewable Energy Focus* 48, article 100545. DOI:10.1016/j.ref.2024.100545

Leng, L., and Huang, H. (2018). "An overview of the effect of pyrolysis process parameters on biochar stability," *Bioresource Technology* 270, 627-642.

DOI:10.1016/j.biortech.2018.09.030

Li, Y., Gupta, R., Zhang, Q., and You, S. (2023). "Review of biochar production via crop residue pyrolysis: Development and perspectives," *Bioresource Technology* 369, article 128423. DOI:10.1016/j.biortech.2022.128423

Liu, J., Chen, X., Chen, W., Xia, M., Chen, Y., Chen, H., Zeng, K., and Yang, H. (2023). "Biomass pyrolysis mechanism for carbon-based high-value products," *Proceedings of the Combustion Institute* 39(3), 3157-3181. DOI: 10.1016/j.proci.2022.09.063

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," *International Conference on Neural Information Processing Systems* 30, 4768-4777.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al*. (2011). *Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res*. 12, 2825-2830.

Rengasamy, D., Mase, J. M., Kumar, A., Rothwell, B., Torres, M. T., Alexander, M. R., Winkler, D. A., and Figueredo, G. P. (2022). "Feature importance in machine learning models: A fuzzy information fusion approach," *Neurocomputing* 511, 163-174. DOI: 10.1016/j.neucom.2022.09.053

Saarela, M., and Jauhiainen, S. (2021). "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences* 3(2), 272.

Safarian, S. (2023). "Performance analysis of sustainable technologies for biochar production," *A Comprehensive Review. Energy Reports* 9, 4574-4593.

Seow, Y. X., Tan, Y. H., Mubarak, N. M., Kansedo, J., Khalid, M., Ibrahim, M. L., and Ghasemi, M. (2022). "A review on biochar production from different biomass wastes by recent carbonization technologies and its sustainable applications," *Journal of Environmental Chemical Engineering* 10(1), article 107017. DOI:10.1016/j.jece.2021.107017

Souaissi, Z., Ouarda, T. B. M. J., and St-Hilaire, A. (2023). "Non-parametric, semi-parametric, and machine learning models for river temperature frequency analysis at ungauged basins," *Ecological Informatics* 75, article 102107.

Su, G., and Jiang, P. (2024). "Machine learning models for predicting biochar properties from lignocellulosic biomass torrefaction," *Bioresource Technology* 399, article 130519. DOI:10.1016/j.biortech.2024.130519

Tomczyk, A., Sokołowska, Z., and Boguta, P. (2020). "Biochar physicochemical properties: Pyrolysis temperature and feedstock kind effects," *Reviews in Environmental Science and Bio/Technology* 19(1), 191-215.

Vuppaladadiyam, A. K., Varsha Vuppaladadiyam, S. S., Sikarwar, V. S., Ahmad, E., Pant, K. K., Pandey, A., Bhattacharya, S., Sarmah, A., and Leu, S. Y. (2023). "A critical review on biomass pyrolysis: Reaction mechanisms, process modeling and potential challenges," *Journal of the Energy Institute* 108, article 101236. DOI: 10.1016/j.joei.2023.101236

Vuppaladadiyam, A. K., Vuppaladadiyam, S. S. V., Awasthi, A., Sahoo, A., Rehman, S., Pant, K. K., Murugavelh, S., Huang, Q., Anthony, E., Fennel, P., *et al*. (2022). "Biomass pyrolysis: A review on recent advancements and green hydrogen production," *Bioresource Technology* 364, article 128087. DOI: 10.1016/j.biortech.2022.128087

Wang, Y., and Wu, J. J. (2023). "Thermochemical conversion of biomass: Potential future prospects," *Renewable and Sustainable Energy Reviews* 187, article 113754.

Xu, L., Saatchi, S. S., Yang, Y., Yu, Y., Pongratz, J., Bloom, A. A., Bowman, K., Worden, J., Liu, J., Yin, Y., *et al*. (2021). "Changes in global terrestrial live biomass over the 21st century," *Science Advances* 7(27), article eabe9829. DOI: 10.1126/sciadv.abe9829

Yang, X., Yuan, C., He, S., Jiang, D., Cao, B., and Wang, S. (2023). "Machine learning prediction of specific capacitance in biomass derived carbon materials: Effects of activation and biochar characteristics," *Fuel* 331, article 125718. DOI: 10.1016/j.fuel.2022.125718

Zhang, C., Chao, L., Zhang, Z., Zhang, L., Li, Q., Fan, H., Zhang, S., Liu, Q., Qiao, Y., Tian, Y., *et al*. (2021). "Pyrolysis of cellulose: Evolution of functionalities and structure of bio-char versus temperature," *Renewable and Sustainable Energy Reviews* 135, article 110416. DOI: 10.1016/j.rser.2020.110416

Zhang, S., Mei, Y., and Lin, G. (2024). "Pyrolysis interaction of cellulose, hemicellulose and lignin studied by TG-DSC-MS," *Journal of the Energy Institute* 112, article 101479.

Zhao, Y., Zhang, W., and Liu, X. (2024). "Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting," *Applied Soft Computing* 154, article 111362.

Zhou, Z. C., Wu, Z., and Jin, T. (2021). "Deep reinforcement learning framework for resilience enhancement of distribution systems under extreme weather events," *International Journal of Electrical Power and Energy Systems* 128, article 106676. DOI:10.1016/j.ijepes.2020.106676

Zhu, X., Li, Y., and Wang, X. (2019). "Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions," *Bioresource Technology* 288, article 121527. DOI:10.1016/j.biortech.2019.121527

Zou, X., Debiagi, P., Amjed, M. A., Zhai, M., and Faravelli, T. (2024). "Impact of high-temperature biomass pyrolysis on biochar formation and composition," *Journal of Analytical and Applied Pyrolysis* 179, article 106463. DOI: 10.1016/j.jaap.2024.106463

Article submitted: June 14, 2024; Peer review completed: July 11, 2024; Revised version received: July 15, 2024; Accepted: July 19, 2024; Published: July 26, 2024.
DOI: 10.15376/biores.19.3.6545-6558