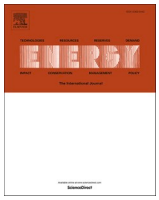
Energy 324 (2025) 136087

Contents lists available at ScienceDirect

Energy

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



Prediction of product yields and heating value of bio-oil from biomass fast pyrolysis: Explainable predictive modeling and evaluation

Longfei Li, Zhongyang Luo *, Liwen Du, Feiting Miao, Longyi Liu

State Key Laboratory of Clean Energy Utilization, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Lignocellulosic biomass
Bio-oil
Ensemble learning
Product yield
Fast pyrolysis
Hyperparameter optimization

ABSTRACT

In this study, optimized ensemble learning algorithms were employed to predict and analyze the product distribution and higher heating value (HHV) of bio-oil from biomass fast pyrolysis, based on feedstock characteristics, operating conditions, and reactor parameters. The results reveal that pyrolysis temperature, biomass carbon and hydrogen content, and feedstock volatile matter are the most influential factors for achieving high bio-oil yield, while deoxygenation pretreatment and moderate pyrolysis temperatures (approximately 500 °C) are critical for enhancing HHV. SHapley Additive exPlanations (SHAP) and Partial Dependence Plot (PDP) analyses further elucidated the complex interactions among these parameters, providing actionable insights for optimizing pyrolysis processes. Additionally, the developed ML models demonstrated robust predictive accuracy, with R^2 values exceeding 0.93 for bio-oil yield prediction, and a user-friendly graphical user interface (GUI) was developed to facilitate practical applications. Finally, when evaluated on the external dataset, the optimized LightGBM model demonstrates a moderate linear relationship between predicted and true values, achieving an accuracy of approximately 80 %, with a peak of 84 %. The residual distribution reflects strong generalization capabilities, validating the effectiveness of the optimization strategy. This work provides a comprehensive understanding of biomass pyrolysis behavior and valuable guidance for industrial process optimization.

Nomenclature

Abbreviations	
C	Carbon content
H	Hydrogen content
O	Oxygen content
N	Nitrogen content
M	Moisture content
Ash	Ash content
V	Volatile matter content
FC	Fixed carbon content
Cel	Cellulose content
Hem	Hemicellulose content
Lig	Lignin content
HT	Heating temperature
GFR	Gas flow rate
FR	Feed rate
H/D	Height-to-Diameter ratio
H/C	Hydrogen-to-Carbon ratio
O/C	Oxygen-to-Carbon ratio
HHV	Higher Heating Value
Eo	Energy conversion efficiency
PCA	Principal Component Analysis
5-CV	5-fold Cross-Validation
ML	Machine learning
RF	Random Forest
MLR	Multiple Linear Regression
AdaBoost	Adaptive Boosting
GBDT	Gradient Boosting Decision Tree
XGBoost	eXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
CatBoost	Categorical Boosting
NGBoost	Natural Gradient Boosting
SHAP	SHapley Additive exPlanations
PDP	Partial Dependence Plot
GUI	Graphical User Interface
R^2	The coefficients of determination
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
HPO	Hyperparameter Optimization

(*continued on next column*)

(*continued*)

* Corresponding author.
E-mail address: zyluo@zju.edu.cn (Z. Luo).

https://doi.org/10.1016/j.energy.2025.136087
Received 13 August 2024; Received in revised form 13 March 2025; Accepted 9 April 2025
Available online 10 April 2025

1. Introduction

The decline in oil reserves, rising demand in emerging markets, and the political and environmental complexities of fossil fuels highlight the urgent need for cost-effective, energy-efficient technological innovations. Optimizing the energy structure and ensuring national energy security are critical tasks. Biomass energy, a key renewable resource, is the only form of renewable energy that can be directly converted into carbon-based liquid fuels [1–4]. Given the global shift towards low-carbon energy and industrial progress, the efficient conversion of biomass into liquid fuels or chemicals is of great significance [5,6].

Biomass fast pyrolysis is a crucial thermochemical process for producing sustainable fuels and chemicals [7,8]. Fluidized bed reactors offer several advantages for the thermochemical conversion of biomass, including improved heat transfer, high mass transfer rates, and effective mixing, which enhance product yield and quality [9–12]. Biomass pyrolysis in fluidized bed reactors produces bio-oil, biochar, and syngas, each with distinct characteristics and applications. Predicting the yields and properties of these products is essential for optimizing process parameters, enhancing product quality, and maximizing system efficiency [13,14]. Traditional methods for predicting product yields and characteristics often rely on complex mathematical models based on empirical correlations and experimental data. These methods may struggle to fully capture the complexities of biomass pyrolysis processes and often require extensive experimental validation.

Recently, machine learning (ML) has emerged as a valuable tool for precisely predicting complex chemical processes. Traditional methods such as Multiple Linear Regression (MLR) and simple neural networks, for predicting product yields and properties in biomass pyrolysis often rely on empirical correlations and complex mathematical models, which may struggle to fully capture the complexities of the process and often require extensive experimental validation. In contrast, advanced ML models can analyze extensive experimental datasets to uncover intricate relationships between input variables and output responses [15]. Fig. 1 illustrates the application and workflow of ML in biomass thermal conversion. By training ML models on extensive datasets from biomass pyrolysis experiments in fluidized bed reactors, predictive tools can effectively predict the yields and properties of bio-oil, biochar, and syngas under various conditions. For instance, the Random Forest (RF) algorithm has been employed to predict the yield and carbon content of slow pyrolysis biochar, demonstrating its effectiveness in biomass pyrolysis predictions [16,17]. Tang et al. utilized RF and Multiple Linear Regression (MLR) algorithms to predict bio-oil yield and hydrogen content. RF outperformed MLR, demonstrating its superior effectiveness in predicting bio-oil properties [13]. Leng et al. estimated the calorific value of bio-oil and the distribution of three-phase products from fast pyrolysis of lignocellulosic biomass. The RF model outperformed other algorithms, demonstrating its superior predictive accuracy. [18]. Zhang et al. used the RF algorithm to examine how different pyrolysis conditions and biomass compositions affect bio-oil yield, viscosity, calorific value, Hydrogen-to-Carbon ratio (H/C), and Oxygen-to-Carbon ratio (O/C). Their findings highlighted the superior performance of models based on elemental analysis [19].

Unlike previous studies that focused on specific aspects of biomass pyrolysis (e.g., bio-oil properties or single-product yields), this study aims to develop universal ML models capable of predicting product yields (bio-oil, biochar, syngas) and bio-oil heating value during biomass fast pyrolysis in a fluidized bed reactor. The comprehensive dataset of 424 samples integrates diverse factors, including biomass composition, pyrolysis conditions, and reactor geometry, to ensure wide applicability and robustness.

Seven ensemble learning algorithms—RF, Adaptive Boosting (AdaBoost), Gradient Boosting Decision Tree (GBDT), Extreme Gradient

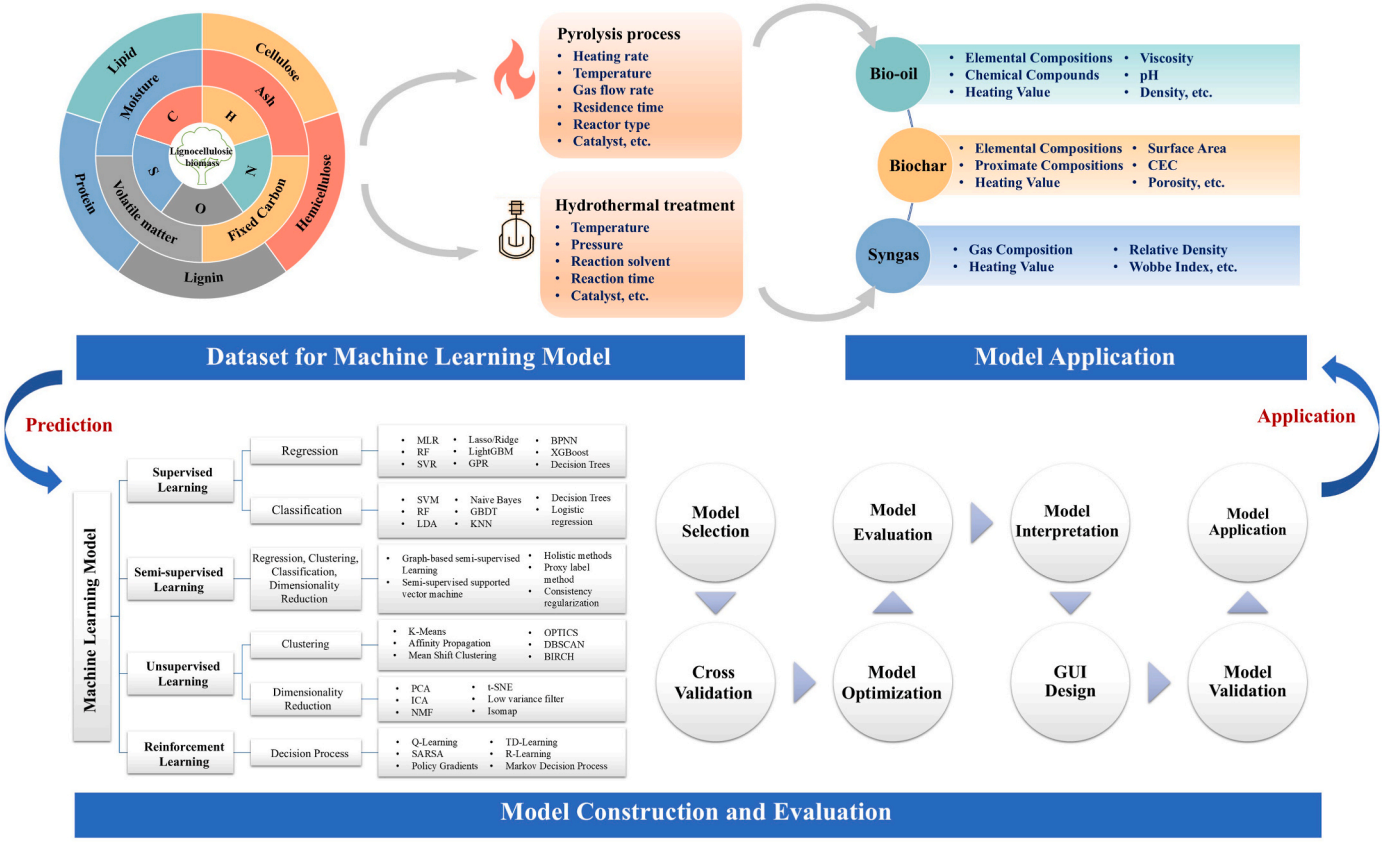


Fig. 1. ML-assisted thermochemical conversion of biomass.

Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), and Natural Gradient Boosting (NGBoost)—were constructed and optimized using OPTUNA. By employing the OPTUNA framework for hyperparameter tuning, this work achieves enhanced predictive accuracy and generalization compared to existing studies. SHapley Additive exPlanations (SHAP) was employed to assess feature importance, and Partial Dependence Plot (PDP) illustrated the impact of different variables on target values. These methods provide actionable insights into the process, bridging the gap between black-box predictions and process optimization. To facilitate real-world application, this study embeds the optimized ML models in a user-friendly GUI. This feature supports researchers and practitioners in simulating pyrolysis outcomes and making data-driven decisions, a capability not addressed in previous studies.

Through the integration of explainability methods and a robust dataset, this study provides new insights into how biomass composition and pyrolysis conditions influence product yield and properties. For instance, it highlights critical parameters such as volatile matter, pyrolysis temperature, and hydrogen content, offering a deeper understanding of their impact on bio-oil yields and quality.

2. Methodology

2.1. Dataset establishment and preprocessing

Fast pyrolysis is a thermochemical conversion process aimed at decomposing biomass feedstocks under anaerobic or low-oxygen conditions to produce liquid bio-oil, gaseous syngas, and solid biochar. This process typically occurs at elevated temperatures (approximately 400–700 °C), with rapid heating rates (10–200 °C/s) and short residence times (<2 s) to minimize further cracking and polymerization of the organic components in the biomass, thereby achieving a higher yield of bio-oil (75–80 %) [7,12,20]. Data on fast pyrolysis of lignocellulosic biomass in fluidized beds were sourced from published literature, with information directly extracted from tables or digitized from images using Web Plot Digitizer software. The dataset includes a diverse range of biomass feedstocks such as wood, agricultural residues, and energy crops, ensuring a broad representation of biomass types. However, it is important to note that the dataset may not cover all possible biomass feedstock types, as some niche or less-studied biomass sources might be missing.

A pre-screening process accounted for variability in feedstock sources, experimental conditions, and repeatability, resulting in 294 datasets for predicting product yields and 130 datasets for predicting bio-oil heating value. Proximate analysis of biomass includes moisture content (M, wt.%), ash content (Ash, wt.%), volatile matter (V, wt.%), and fixed carbon (FC, wt.%). The primary components of biomass are cellulose (Cel, wt.%), hemicellulose (Hem, wt.%), and lignin (Lig, wt.%). Pyrolysis yields can be seen as the sum of the individual components' pyrolysis, making their composition crucial. Elemental composition—carbon (C, wt.%), hydrogen (H, wt.%), oxygen (O, wt.%), and nitrogen (N, wt.%)—significantly impacts product quality. Biomass particle size (PS, mm) affects yield by influencing heat and mass transfer during pyrolysis. Pyrolysis conditions include temperature (HT, °C), gas flow rate (GFR, m^3/h), and feed rate (FR, kg/h). The reactor's Height-to-Diameter ratio (H/D, m/m) is also considered an input parameter, as it affects heat and mass transfer, impacting product quantity and quality. While fast pyrolysis is generally carried out at elevated temperatures, this study examined a broader temperature range (277–700 °C) to explore the variations in biomass pyrolysis behavior under extreme conditions. This approach aids in comprehending the influence of temperature on the yield and composition of pyrolysis products, thereby providing more comprehensive data to support industrial applications.

The output variables are biochar yield, bio-oil yield, bio-gas yield, and the higher heating value (HHV, MJ/kg) of bio-oil. The energy conversion efficiency (Eo) related to product yield and heating value is discussed and can be calculated using Eq. (1).

$$Eo = \frac{HHV_{biooil}}{HHV_{biomass}} \times \frac{Yield\ of\ biooil}{Mass\ of\ biomass} \tag{1.}$$

In this study, all data are presented based on a dry basis. In instances where HHV values were absent in the initial dataset, they were estimated through the application of an empirical formula as outlined by Eq. (2) [21].

$$HHV_{dry}(MJ/kg) = 0.3491C + 1.1738H + 0.1005S - 0.1034O - 0.0151N - 0.0211A \tag{2.}$$

The C, H, S, O, N and A in Eq. (2) are contents of carbon, hydrogen, sulphur, oxygen, nitrogen, and ash in bio-oil on dry basis.

Statistical analysis was utilized in conjunction with important metrics such as measures of central tendency, quartiles, and evaluations of normality to examine the structure and distribution of the data. To mitigate potential performance challenges associated with ML estimators when features exhibit substantial deviations from a normal distribution, all data points were standardized using the StandardScaler function available in the Scikit-learn library. The standard scores (z) of the samples were computed using Eq. (3).

$$z = \frac{x - \mu}{s} \tag{3.}$$

where z represents the normalized value of each input feature, x denotes the original value of the input feature, and μ and s are the respective mean and standard deviation of each input feature.

Correlation coefficients were used to explore relationships between features and the target variable, as well as among various features. The analysis examined three common correlation coefficients: Pearson, Spearman, and Kendall. Given the normality assumption and outlier sensitivity of Pearson's coefficient, Spearman's correlation was chosen to evaluate correlations in the dataset [22]. The Spearman correlation coefficient is expressed in Eq. (4).

$$r_s = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)} \tag{4.}$$

where N is the sample size and d_i is the difference between the grades of the two variables being observed.

In this research, Principal Component Analysis (PCA) was used as an unsupervised learning method to extract and compare the fundamental principles of two distinct subsets. The basic implementation of PCA is as follows:

Given a dataset X with m samples and n features, each row represents a sample and each column represents a feature.

1) Data centering: Data centering is achieved by subtracting the mean of each feature. Let $\bar{x}_j$ denote the mean of the j-th feature. The centered dataset $\widehat{X}$ can be calculated using Eq. (5):

$$\widehat{X}_{ij} = X_{ij} - \bar{x}_j \tag{5.}$$

Here, $\widehat{X}_{ij}$ denotes elements from the centered dataset, and X_{ij} denotes elements from the original dataset.

2) Compute the covariance matrix: The element C_{ij} of the covariance matrix can be calculated using Eq. (6):

$$C_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}\left(\widehat{X}_{ki}\cdot\widehat{X}_{kj}\right) \quad (6.)$$

Here, C_{ij} represents an element of the covariance matrix, while $\widehat{X}_{ki}$ and $\widehat{X}_{kj}$ denote elements from the centered dataset.

3) Eigendecomposition: Perform eigendecomposition on the covariance matrix to obtain eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and their corresponding eigenvectors $v_1, v_2, \ldots, v_n$.
4) Selecting principal components involves sorting eigenvalues in descending order and choosing the eigenvectors corresponding to the top k eigenvalues, where k is the number of principal components selected (typically chosen to achieve a cumulative contribution rate of 80 % or 90 %).
5) Projection onto Principal Components: Project the centered dataset $\widehat{X}$ onto the matrix composed of the selected top k eigenvectors to obtain the reduced-dimensional dataset Y:

$$Y = \widehat{X}\cdot V_k \quad (7.)$$

Here, V_k is the matrix containing the top k eigenvectors.

2.2. ML models and evaluation

2.2.1. Model construction

The modeling and evaluation were carried out utilizing the Scikit-learn library within the Python programming environment (Python 3.10.5). Before constructing the ML models, the dataset was randomly partitioned into train and test sets at a ratio of 4:1, and coupled with 5-fold Cross-Validation (5-CV) for training the ML models, while the testing set was employed for the final assessment of the models. To gauge the predictive efficacy of the ML models and identify the most suitable one for further interpretability analysis, the coefficients of determination (R^2), Root Mean Square Error (*RMSE*), and Mean Absolute Error (*MAE*) were employed to compare the predictive precision and generalization capacity of each model. In essence, higher R^2 values and lower *RMSE* and *MAE* values are indicative of enhanced predictive accuracy. The aforementioned evaluation metrics are calculated using Eq. (8), Eq. (9), and Eq. (10).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i}^{N}\left(Y_i^{exp} - Y_{ave}^{exp}\right)^2} \quad (8.)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{N}^{i=1}\left(Y_i^{exp} - Y_i^{pred}\right)^2} \quad (9.)$$

$$MAE = \frac{1}{N}\sum_{N}^{i=1}\left|Y_i^{exp} - Y_i^{pred}\right| \quad (10.)$$

where Y_i^{exp} and Y_i^{pred} are experimental and predicted values, and Y_{ave}^{exp} is the average of experimental values.

2.2.2. Hyperparameter optimization

Hyperparameter optimization is a critical step in the training process of machine learning models. It involves tuning the non-data-driven parameters of a model to enhance its generalization ability and predictive accuracy. In this study, we employed OPTUNA, an open-source hyperparameter optimization (HPO) framework, which utilizes Bayesian optimization to automatically search for the optimal combination of hyperparameters. OPTUNA was selected for its efficiency in navigating large search spaces and its flexibility in handling complex model configurations.

For each model, we defined a specific hyperparameter search space, including key parameters such as learning rate, number of trees, and tree depth. For the optimization strategy, we employed Bayesian optimization. This technique builds a probabilistic model of the hyperparameter space, which guides the search process by balancing exploration (searching areas with high uncertainty) and exploitation (focusing on areas known to produce good results). This allows OPTUNA to find the optimal solution in fewer iterations compared to traditional grid search or random search methods.

The optimization objective was to minimize the *RMSE* on the training set, with the goal of improving model accuracy and generalization. To evaluate the performance of each hyperparameter combination, we used 5-CV, ensuring that the models would generalize well to unseen data and preventing overfitting.

Further details on the hyperparameter optimization process, including Python code snippets used for defining the search space and implementing the optimization procedure, are provided in the Supplementary Materials.

2.2.3. Visual interpretation

Assessing interpretability in ML models is crucial for optimizing production processes via reverse engineering. This involves analyzing feature importance using SHAP values. SHAP, a retrospective explanation technique, calculates the incremental effect of features on model outputs, providing explanations for opaque models at both global and local levels [23]. Additionally, the PDP method is used to quantify the relationship between feature variables and the target variable, aiding in process enhancement. Specifically, one-way PDP reveals the influence patterns of individual features on the target value, while two-way PDP shows the interaction effects between pairs of features.

3. Results and discussion

3.1. Exploratory analysis of data

The dataset first underwent imputation to address missing values, considering three methods: mean, median, and mode imputation. Mean imputation proved most advantageous for model development (Fig. S1). A violin plot of the dataset is shown in Fig. 2. Additional statistical summaries can be found in Tables S2 and S3 of the Supplementary Materials. Notably, the concentrations of C, H, O, and M in the biomass feedstock range from 39.8 wt% to 59.17 wt%, 3.77 wt% to 8.10 wt%, 29.6 wt% to 53.64 wt%, and 1.14 wt% to 18.6 wt%, respectively. Key process parameters such as pyrolysis temperature (277–700 °C), feed rate (0.07–10 kg/h), gas flow rate (0.06–14 m^3/h), and equipment parameter H/D (2–28.57) cover common experimental conditions in contemporary pyrolysis studies. These diverse datasets offer opportunities for developing robust ML models with strong generalization capabilities.

Spearman correlation analysis visually represents the level of connection between variables, showing both the intensity and direction of the correlation. Fig. 3 depicts a strong correlation among biomass elemental compositions. Oxygen composition, often reported as the residual variance to sum up C, N, H, and other components to 100 %, shows this trend. In proximate analysis, a significant negative correlation exists between FC and V. Additionally, a notable correlation is identified between the HHV of bio-oil and the H/C and O/C ratios. The relationship between inputs and yields shows a moderately strong linear connection between pyrolysis temperature and the yields of biochar and bio-gas. As temperature increases, char yield decreases while gas yield increases, consistent with previous studies [20,24]. It is essential to highlight that without corresponding tests for significance levels, this initial examination of linear relationships is not conclusive but serves as a valuable reference for subsequent analyses on model feature importance and PDP analysis.

Analyzing operational parameters and elemental composition is

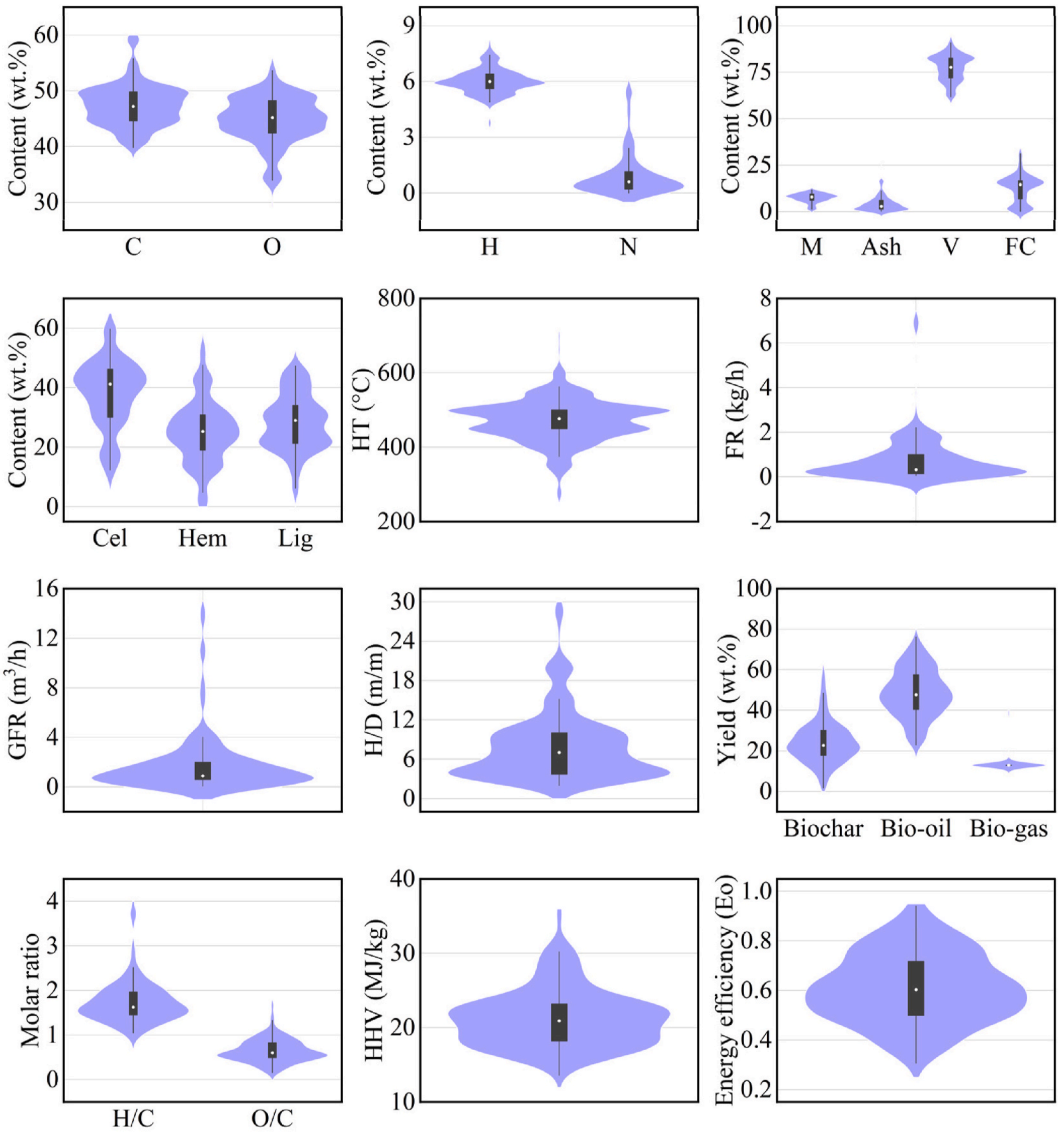


Fig. 2. Violin-plot distribution of important parameters.

crucial for understanding phenomena such as elemental migration during biomass pyrolysis. This understanding helps optimize biomass composition and operational conditions to improve product yield and quality. Fig. 4 illustrates how element composition and pyrolysis temperature affect bio-oil quantity and quality. As shown in Fig. 4(A) and (B), the carbon content in bio-oil, averaging 51.7 wt% with a maximum of 72.7 wt%, is significantly higher than in the original biomass, which averages 47.4 wt% with a minimum of 39.8 wt%. Hydrogen content in bio-oil, averaging 7.1 wt% and peaking at 11.2 wt%, also surpasses that of the parent biomass, which averages 6.2 wt% and reaches a maximum of 8.1 wt%. Conversely, bio-oil's minimum oxygen content of 15.3 wt% is much lower than the raw biomass, which ranges from 29.6 wt% to 53.6 wt%. The significant reduction in oxygen content, due to deoxygenation and decarboxylation, raises the bio-oil's calorific value to an average of 23.7 MJ/kg, with a maximum of 35.3 MJ/kg. Using suitable catalysts can further enhance the heating value of bio-oil, making it a more viable option for alternative transportation fuels [25–27].

In the context of biomass and bio-oil, a decrease in biomass oxygen content leads to an increase in the HHV of bio-oil, as shown in Fig. 4(C) and (D). After fast pyrolysis, bio-oil exhibits a significantly higher H/C ratio compared to raw biomass, peaking at 3.77. However, there is a direct relationship between the H/C and O/C ratios of bio-oil, indicating that an increase in H/C is accompanied by a rise in O/C. Thus, using catalysts with strong deoxygenation capabilities is essential to reduce bio-oil oxygen content and enhance its heating value. Additionally, higher ash content and increased heteroatoms (such as nitrogen and oxygen) in biomass can decrease bio-oil yield and affect its heating value, as illustrated in Fig. 4(E) and (F). This challenge arises from the production of additional gaseous byproducts like CO_2 and CO with elevated levels of biomass heteroatoms [28–30].

The ash content in feedstock also promotes the formation of biochar and bio-gas by hindering the interaction between organic matter and the reaction medium, thereby reducing bio-oil yield [31]. Fig. 4(G) demonstrates the effect of pyrolysis temperature on the yield and HHV of bio-oil. Typically, the highest bio-oil yield is achieved within the temperature range of 500–550 °C, accompanied by a relatively high HHV. Additionally, a negative correlation between bio-oil yield and HHV value is observed. Maintaining an HHV value between 20 and 30 MJ/kg and a yield above 60 wt%, the energy conversion efficiency reaches its peak value, as depicted in Fig. 4(H). This observation provides valuable insights for optimizing process parameters to achieve the highest energy conversion efficiency.

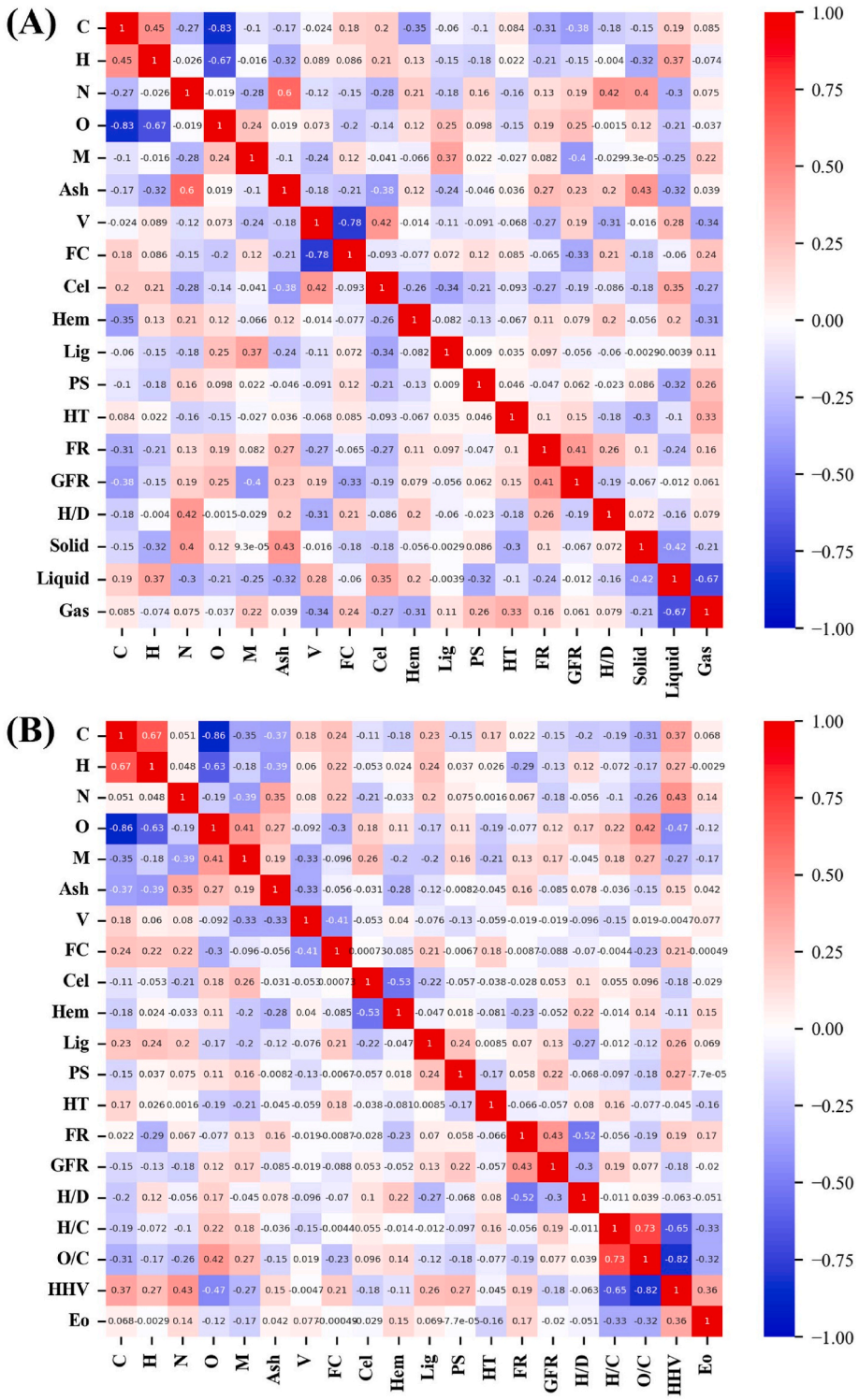


Fig. 3. Spearman correlation coefficient matrix between the input and output variables: (A) with product yield; (B) with bio-oil properties.

3.2. Principal component analysis

PCA is used to visually represent data distribution across multiple dimensions and to gain insights into the original dataset. It reduces dimensionality by calculating the primary 'eigenvalues', which provide detailed information on raw material composition, process parameters, and product composition, as illustrated in Fig. 5. The cumulative variance results in Fig. 5(A) and (C) show that the cumulative variance of the first 12 principal components exceeds 95 %, capturing all relevant information about biomass pyrolysis products and bio-oil properties. This demonstrates the effectiveness of PCA in dimensionality reduction and extracting crucial insights from the dataset.

The initial and subsequent principal components of the product yield dataset explain 23 % and 17 % of the total variance, respectively, with the first four principal components accounting for 60 % of the variance. Principal Component 1 predominantly includes characteristics such as the elemental composition of raw materials (C, H, O) and chemical composition (Cel, Lig), while Principal Component 2 mainly comprises

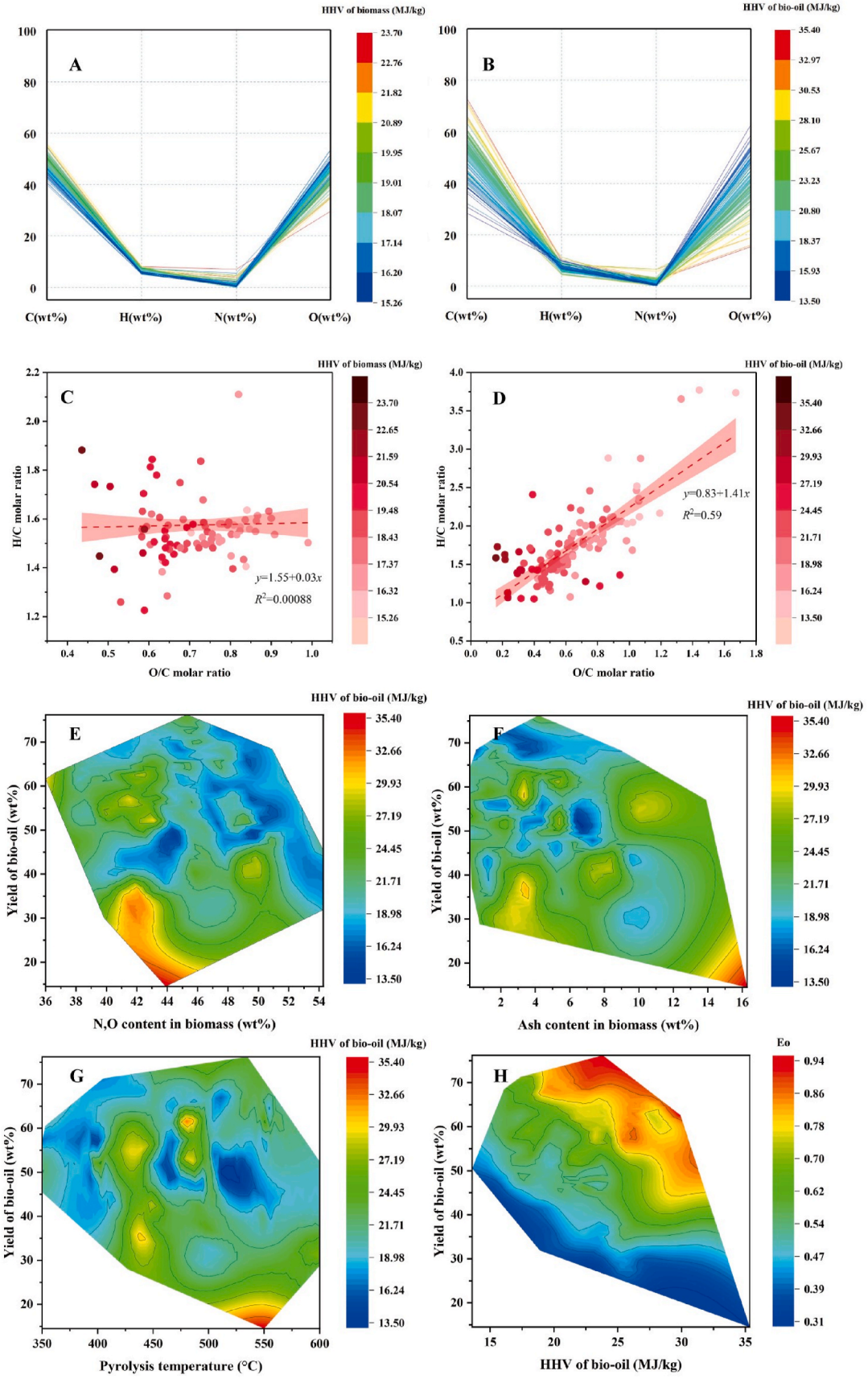


Fig. 4. The influence of element composition and pyrolysis temperature on the yield and HHV of bio-oil.

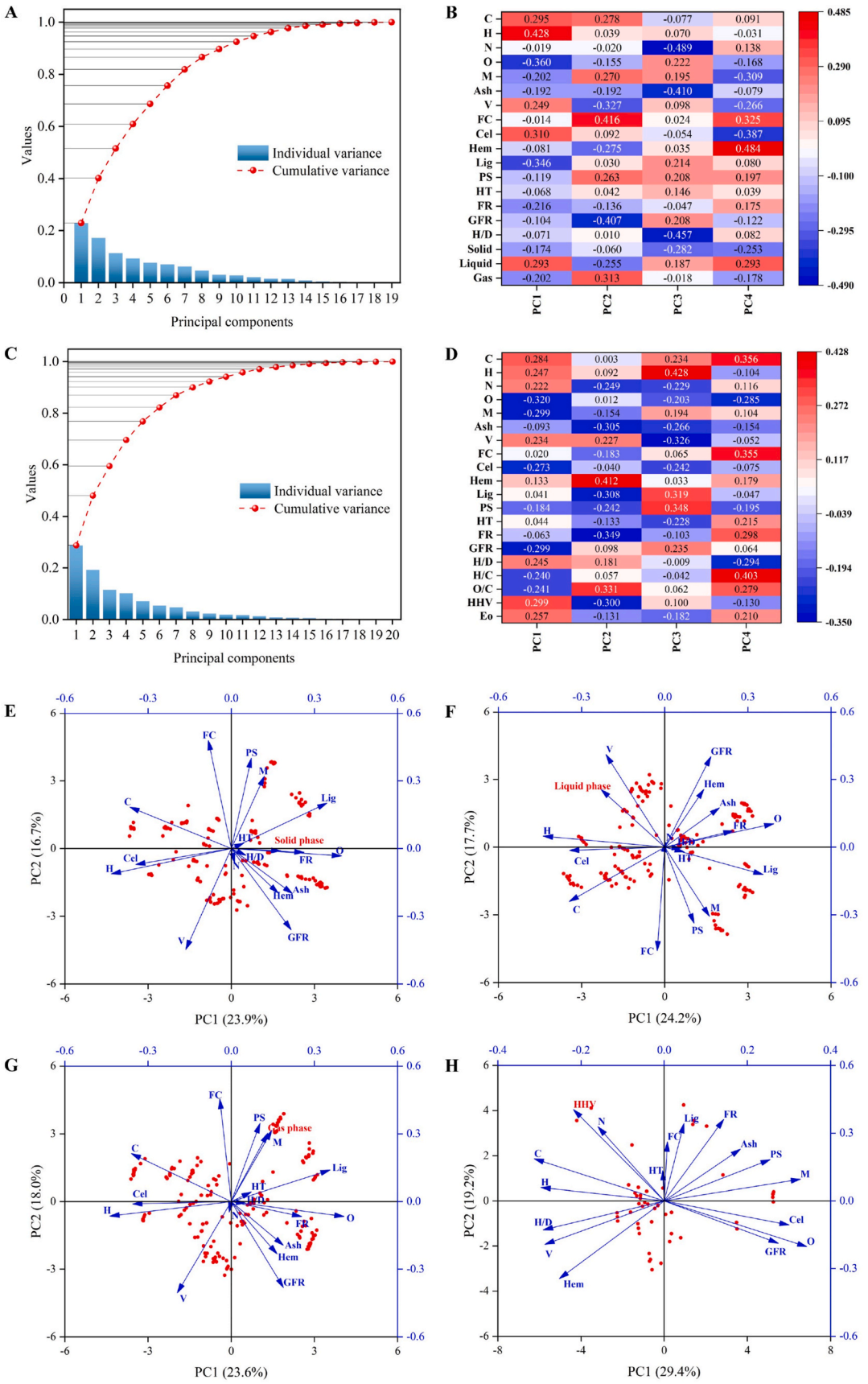


Fig. 5. Principal component analysis of the whole primary dataset. Variance of each component on (A) product yields and (C) bio-oil properties; Correlation of features with the top-four components on (B) product yields and (D) the HHV of bio-oil. Effect of input variables on (E) biochar yield, (F) bio-oil yield, (G) syngas yield, and (H) the HHV of bio-oil in the space of the first two components.

features related to raw material properties (V, FC, M) and operational parameters (GFR).

In the bio-oil properties dataset, the first and second principal components explain 29 % and 19 % of the total variance, respectively, with the first four principal components capturing 60 % of the variance. Principal Component 1 primarily includes attributes related to raw material properties (C, O, M, Cel) and operational conditions (GFR), whereas Principal Component 2 encompasses features associated with raw material properties (Ash, Hem, Lig) and operational conditions (FR). Both datasets underscore the significance of raw material properties.

The projections of the experimental dataset onto the first two principal components are shown in Fig. 5. During thermal decomposition, the contents of C, H, V, FC, and Cel in the raw material positively affect bio-oil yield, while oxygen content negatively correlates with liquid product yield but positively correlates with biochar and gas production, as CO and CO_2 are the main gaseous products, and biochar contains oxygen-functional groups.

Fig. 5(H) shows the correlation of various factors with HHV, indicating that increasing the content of C, H, and V in raw materials improves HHV. Additionally, there is a positive correlation between the reactor parameter (H/D) and HHV, though its correlation with yield is not significant. However, Cahanap et al. found that gas yield increases with the H/D ratio, as increasing reactor height leads to more volatilization and gasification of biomass particles [32]. Notably, these correlations may vary based on the composition of biomass and operational parameters.

3.3. Evaluation and optimization of ML models

Nine ensemble learning models and two combined models were optimized using OPTUNA for hyperparameter tuning, and their performances and intrinsic characteristics were compared. Table S4 summarizes the optimized hyperparameters obtained from the 5-CV on the training dataset for each model, with the visualization of the optimization process available in Fig. S1 of the Supporting Materials. Table 1 presents the performance metrics of each model with both default and optimized hyperparameters. After optimization, the R^2 of each model improved by 5 %–20 %.

LightGBM exhibited the best performance in predicting biochar yield, bio-oil yield, and the HHV of bio-oil, with R^2 values of 0.9079, 0.9305, and 0.8542, *RMSE* values of 3.5161, 4.0549, and 1.9153, and *MAE* values of 2.3697, 2.9929, and 1.5377, respectively. The combined model of XGBoost and LightGBM performed best in predicting biogas yield with R^2 value of 0.9166, *RMSE* value of 5.8364, and *MAE* value of 4.2924. Overall, the LightGBM model demonstrated the best predictive performance, followed by XGBoost, with the combined model of LightGBM and XGBoost also showing good performance (Fig. 6).

The superior performance of the LightGBM model can be attributed to its advanced training mechanism and noise resistance, resulting in reliable output predictions [33,34]. LightGBM employs techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which significantly enhance the model's efficiency and accuracy when handling large-scale datasets [35]. These techniques allow LightGBM to reduce the computational burden and memory usage, making it particularly well-suited for large and high-dimensional datasets, like those used in biomass pyrolysis.

In contrast, XGBoost incorporates regularization terms to control

Tabel 1
Comparative evaluation of nine ML models based on the test set.

Objectives	ML Models	Under default hyperparameters			Fine-tuned by OPTUNA		
		RMSE	*MAE*	R^2	*RMSE*	*MAE*	R^2
Solid yield	RF	3.8315	2.7934	0.8361	4.1685	2.9096	0.8706
	AdaBoost	4.9375	4.0110	0.7278	5.4895	4.4824	0.7756
	GBDT	3.6940	2.7499	0.8476	4.0034	2.7266	0.8806
	XGBoost	3.6542	2.7422	0.8509	3.6514	2.4034	0.9007
	LightGBM	4.0309	2.9501	0.8186	3.5161	2.3697	0.9079
	CatBoost	3.9914	2.7223	0.8221	3.8745	2.8468	0.8882
	NGBoost	3.6529	2.7283	0.8510	3.7410	2.5696	0.8958
	XGBoost-LightGBM	4.1659	2.9738	0.8062	3.7899	3.0848	0.8930
	CatBoost-NGBoost	3.7932	2.7812	0.8393	3.7649	2.9174	0.8944
Liquid yield	RF	6.0353	4.4552	0.8243	4.9297	3.7241	0.8973
	AdaBoost	7.1985	6.1678	0.7501	7.1879	5.8695	0.7817
	GBDT	6.6101	4.8855	0.7893	4.9875	3.7060	0.8949
	XGBoost	7.5745	5.0881	0.7233	5.2647	4.0703	0.8829
	LightGBM	6.0146	4.6601	0.8255	4.0549	2.9929	0.9305
	CatBoost	5.3682	3.7266	0.8610	5.0633	3.9790	0.8917
	NGBoost	6.5574	4.9785	0.7926	4.9974	3.7023	0.8945
	XGBoost-LightGBM	8.0516	5.6665	0.6873	6.0697	4.6128	0.8444
	CatBoost-NGBoost	6.1630	4.7633	0.8168	6.0078	4.8375	0.8475
Gas yield	RF	6.2250	3.9951	0.8044	6.1982	4.1428	0.8803
	AdaBoost	8.5712	6.9619	0.5283	7.9755	6.2145	0.6708
	GBDT	6.8201	4.6347	0.7380	6.7294	4.6773	0.8232
	XGBoost	7.2092	4.5172	0.6955	6.7809	4.2017	0.8174
	LightGBM	6.5693	4.6604	0.7709	6.1567	4.2634	0.8846
	CatBoost	6.9975	4.5557	0.7189	6.7587	4.3558	0.8199
	NGBoost	7.5960	4.7416	0.6510	6.5753	4.7391	0.8403
	XGBoost-LightGBM	5.9098	4.2559	0.8336	5.8364	4.2924	0.9166
	CatBoost-NGBoost	6.4553	4.6497	0.7757	6.2804	4.4886	0.8718
HHV of Bio-oil	RF	2.7449	2.1644	0.6866	2.0465	1.7960	0.8025
	AdaBoost	2.7187	2.0473	0.6983	2.0660	1.8585	0.7929
	GBDT	2.7685	2.1011	0.6760	2.0012	1.6260	0.8242
	XGBoost	3.3108	2.3840	0.5958	2.2514	1.7219	0.7501
	LightGBM	2.9028	2.2396	0.6376	1.9153	1.5377	0.8542
	CatBoost	2.0609	1.6519	0.7954	2.4884	1.8820	0.7959
	NGBoost	2.2360	1.8319	0.7060	2.6384	2.0241	0.7333
	XGBoost-LightGBM	2.8191	2.1945	0.6530	2.5408	1.9275	0.7745
	CatBoost-NGBoost	2.4106	2.0005	0.6097	2.6276	2.0046	0.7379

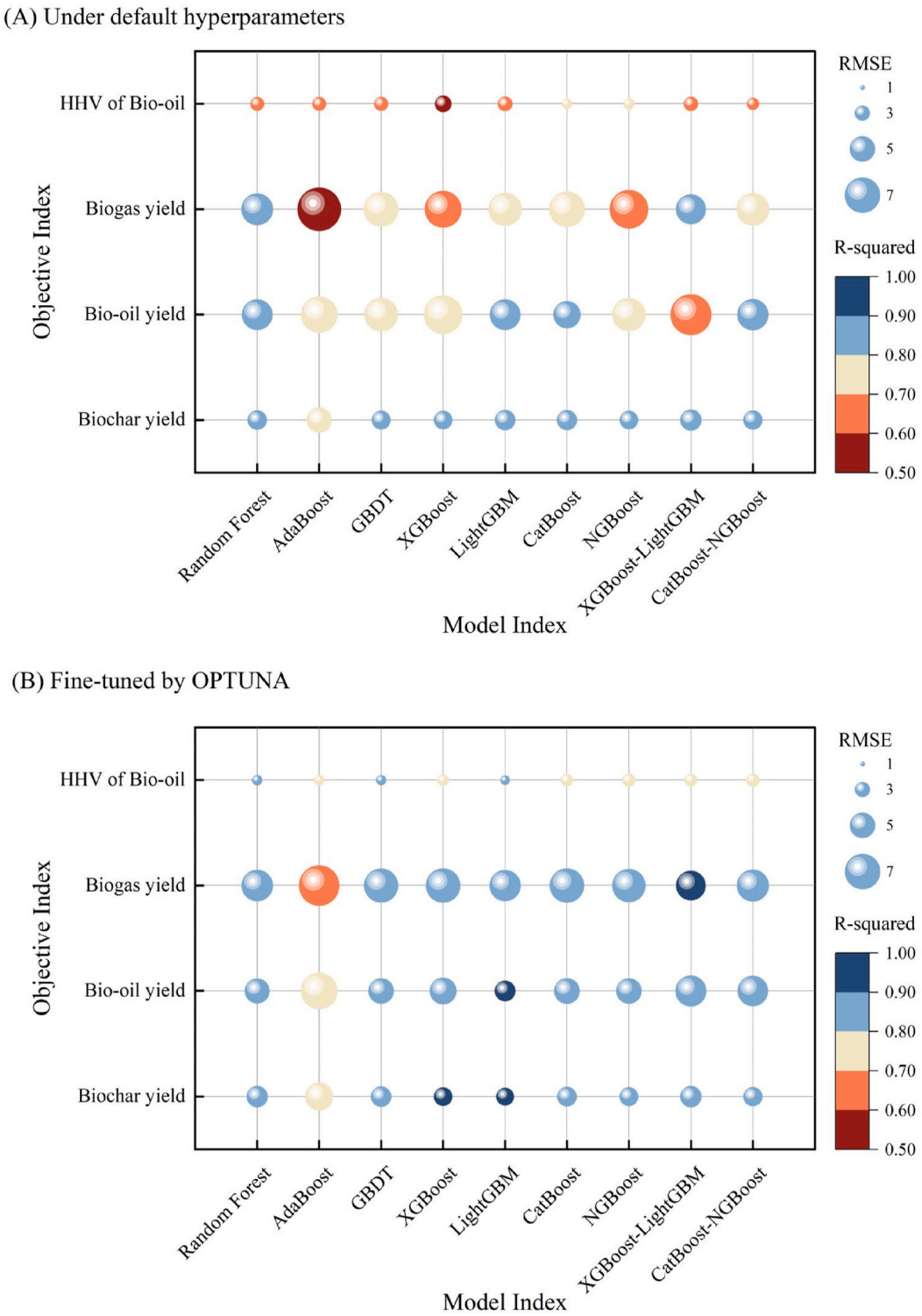


Fig. 6. Comparison of various ML models based on R^2 score and *RMSE* developed the prediction of product yields and HHV of bio-oil.

model complexity and reduce overfitting, which helps improve generalization [36]. While XGBoost is effective for handling nonlinear features, it may require more computational resources and longer training times compared to LightGBM, especially when dealing with large datasets. RF improves prediction stability by constructing multiple decision trees and aggregating results through averaging or majority voting. However, it can struggle with very large datasets and high-dimensional feature spaces, as it does not incorporate mechanisms like GOSS or EFB.

On large datasets, LightGBM demonstrates faster training speed and lower memory consumption due to its histogram-based algorithm optimization. Both XGBoost and LightGBM are capable of learning nonlinear features effectively. However, LightGBM excels in handling high-dimensional sparse data, which is particularly crucial for the biomass pyrolysis dataset used in this study, where many features exhibit sparse distributions [37]. This gives LightGBM a distinct advantage in terms of both model performance and computational efficiency for bio-oil yield prediction.

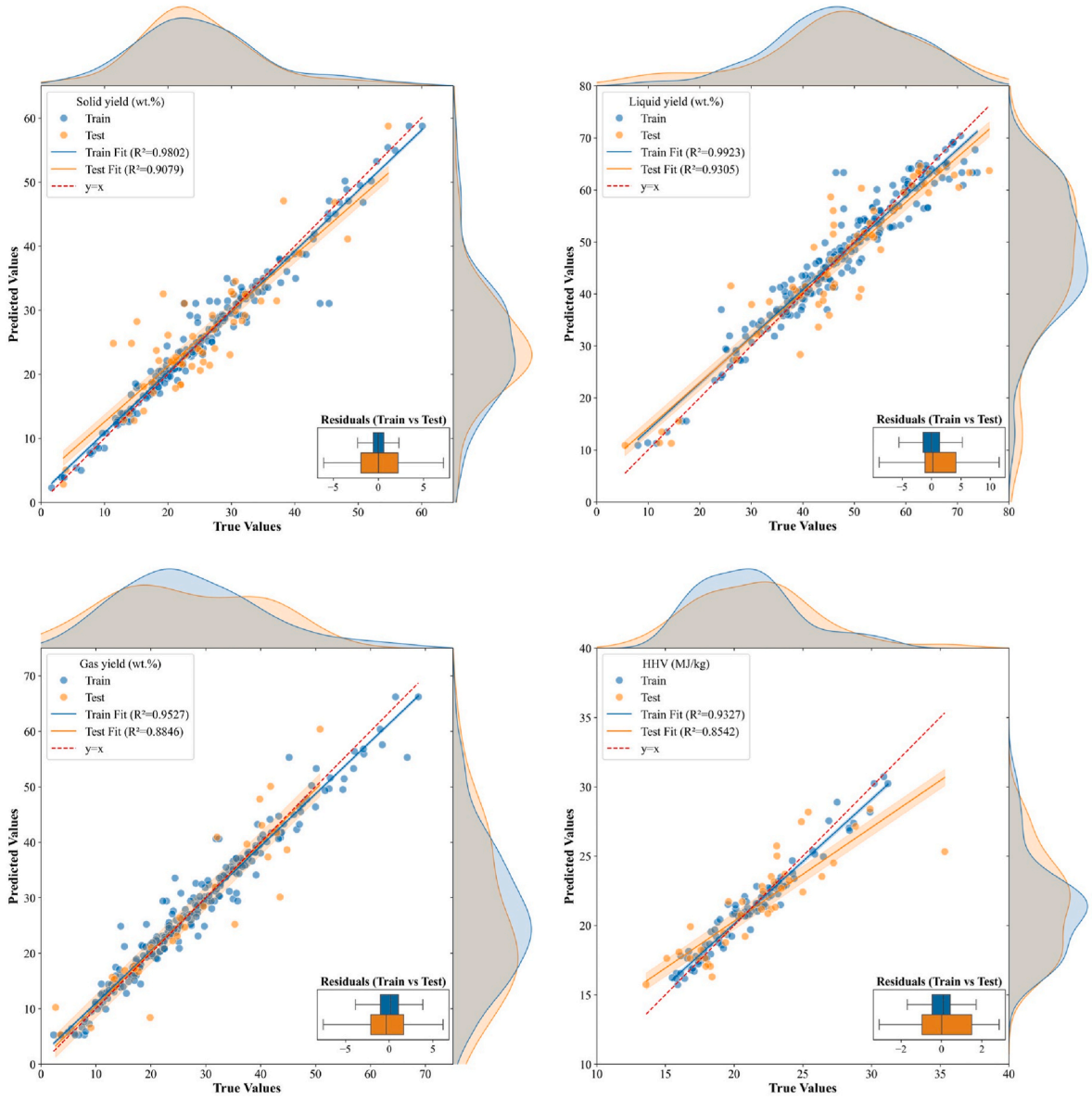


Fig. 7. Plot of predicted product yields and HHV of bio-oil for the optimal LightGBM model (The orange and blue lines represent linear fitting curves, the light filled area is the 95 % confidence interval).

The combined model improves prediction accuracy and robustness by averaging the predictions of individual models [38]. Fig. 7 shows the joint scatter plot of actual and predicted values for product yields and the HHV of bio-oil determined by the LightGBM model. The scatter plot illustrates the relationship between predicted values and true values for two datasets: the training set and the test set. Linear fit lines for each dataset highlight the overall trend between predictions and actual outcomes. The diagonal line $y=x$ serves as a benchmark for perfect predictions, where points lying on this line represent a complete match between predicted and true values. The results reveal that the fit line for the training set is closer to the diagonal, indicating a superior goodness-of-fit compared to the test set. Although performance on the test set shows a slight decline, it still achieves a high predictive accuracy ($R^2 = 0.9305$), demonstrating the model's capacity for generalization. Additionally, the 95 % confidence intervals of the fit lines are shown, reflecting the uncertainty in the model's predictions. The narrower confidence interval for the training set suggests greater stability, while the broader interval for the test set points to reduced prediction stability on unseen data.

The residual box plot, positioned in the bottom right corner of the main figure, provides insights into the residual distributions for both datasets. Residuals, defined as the difference between true and predicted values, measure the deviation of the model's predictions from actual values. For the training set, the residual box plot is concentrated and exhibits a narrow range, indicating small prediction errors. Conversely, the residuals for the test set are more dispersed and include potential outliers, reflecting larger prediction errors on unseen data. This increased variability in the test set residuals may suggest that the model failed to capture certain features of the data during training. Future improvements could include applying regularization techniques or enhancing the dataset through augmentation to further improve the model's generalization ability.

In summary, the minimal difference in R^2 between the training and test sets indicates the good fit and predictive performance of the LightGBM model. Subsequent analyses using SHAP and PDP will be based on the LightGBM model.

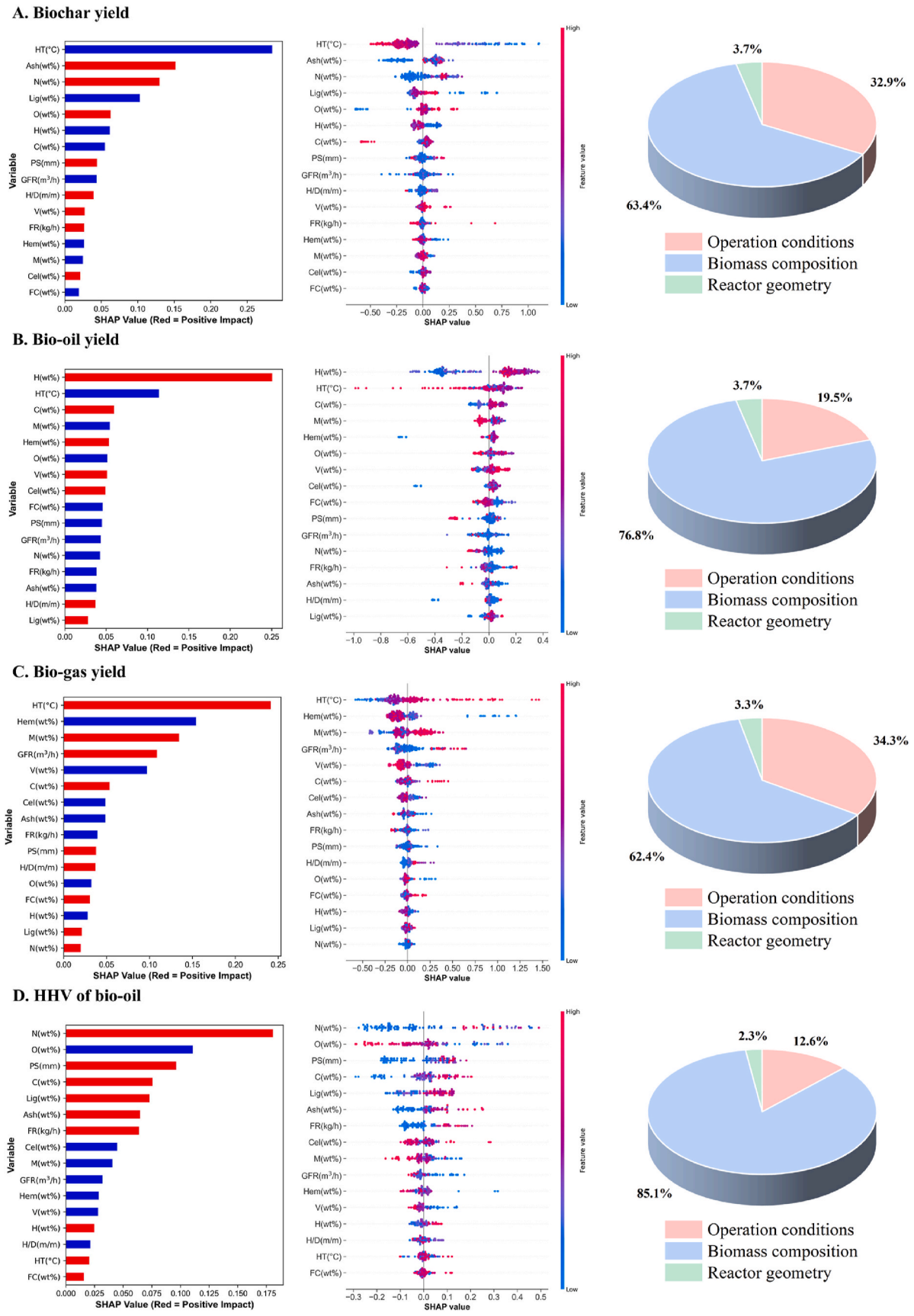


Fig. 8. SHAP value obtained from the LightGBM model representing the effect of input descriptors on each output target.

3.4. SHAP analysis

In this investigation, SHAP values were used to evaluate the significance of input variables on the output responses of the LightGBM model. Fig. 8 displays the SHAP values for input features specific to the outputs. These features are arranged in descending order based on their average absolute SHAP values, with higher rankings indicating greater importance. Data points are represented as points in the bee colony diagram, with red dots highlighting descriptors with high values and significant predictive effects. In the accompanying bar graph, red bars indicate a positive influence of an input feature on the target variable, while blue bars denote negative influences. The length of each bar corresponds to the SHAP value on the horizontal axis, emphasizing the relative importance of each input feature to the target variable.

The characteristics of biomass exert a more significant influence on product yields than operating conditions and apparatus geometric parameters, contributing over 60 % collectively. Specifically, ash content in biomass contributes over 14 % to biochar yield, hydrogen content contributes over 25 % to bio-oil yield, and hemicellulose content contributes over 13 % to gas yield. Additionally, operating conditions such as heating temperature can contribute up to 26 % to biochar yield. Higher temperatures were found to increase the production of syngas while decreasing the yield of biochar. By raising the operating temperature, secondary cracking reactions are facilitated, leading to enhanced gaseous product formation and reduced char production [39]. The contribution of biomass composition to the HHV of bio-oil exceeds 80 %, with the content of C, O, and N elements in biomass being the primary influencing factors. The effect of operating temperature on HHV is minimal. This detailed comprehension highlights the intricate relationships within the pyrolysis process and underscores the potential for customized process optimization guided by predictive insights.

3.5. PDP analysis

To better understand the dependencies between input and target variables, PDP will be constructed based on the importance of each input variable, as shown in Fig. 9. One-way PDP visualize the relationship between selected input features and product yields, keeping all other features at their average values. Bivariate PDP illustrate interactions between two features. These plots, incorporating the model's predicted values, allow for assessing whether the model has accurately captured trends observed in prior experiments.

The yield of biochar significantly decreases with increasing temperature, demonstrating a strong negative correlation (correlation coefficient of −0.30) as shown in Fig. 3. This trend aligns with the high feature importance score for pyrolysis temperature. Additionally, the ash content of the initial biomass material shows high feature importance, as depicted in Fig. 8, and the associated PDP in Fig. 9 reveals a substantial increase in biochar yield with higher ash content. While previous research has indicated that higher lignin content promotes char formation [40]. This dataset shows no strong positive correlation between lignin content and biochar yield. Instead, a negative correlation between lignin content and biochar yield is observed, supported by the Spearman correlation coefficient in Fig. 3 and the feature importance scores in Fig. 8. Conversely, the oxygen content in raw biomass, with a correlation coefficient of 0.12, enhances char yield, as indicated by the PDP. These findings suggest that biomass materials with elevated ash and oxygen contents are more likely to achieve higher char yields at lower pyrolysis temperatures.

As pyrolysis temperature increases, bio-oil production rises, peaking around 500 °C, consistent with prior studies [41–43]. However, further temperature increases lead to a significant decline in bio-oil yield due to secondary cracking reactions above 500 °C, which generate gases, water, or water-soluble compounds [44]. Additionally, bio-oil yield initially increases with higher carbon content in biomass, as elevated carbon content generally indicates higher lignin levels, which produce more liquid products like phenols and oligomers during pyrolysis [45]. Increased hydrogen content further enhances liquid yield by promoting hydrocarbon formation. Conversely, higher moisture content impedes heat and mass transfer, adversely affecting liquid production. The two-way PDP also shows that bio-oil yield decreases with higher gas flow rate (GFR). This decline is attributed to the impact of GFR on fluidization, vapor residence time, and heat transfer, as well as secondary reactions like thermal cracking and re-polymerization [46,47]. While a higher GFR improves bubble movement and mixing, excessive flow rates can lead to larger bubbles, reducing solid mixing and heat transfer efficiency, thereby decreasing bio-oil production [33]. Equipment parameters, such as the height-to-diameter ratio (H/D), have a minimal effect on bio-oil yield.

As shown in Fig. 9(C), the yield of gaseous products increases approximately linearly with temperature, while the yields of bio-oil and biochar decrease correspondingly. This trend is attributed to the decomposition of volatile compounds and secondary cracking of bio-oil [48]. However, the rate of increase in gas yield slightly diminishes at higher temperatures, likely due to the hydrolysis of biomass by water molecules [35] and the increased difficulty in decomposing pyrolysis residues [45]. Gas yield initially decreases and then increases with higher moisture content. Additionally, when carbon content exceeds 50 %, it promotes gas formation, while hemicellulose content above 10 wt % significantly reduces gas yield in favor of bio-oil production. Increasing temperature and gas flow rate facilitates the formation of smaller molecular gases, thereby enhancing gas yield.

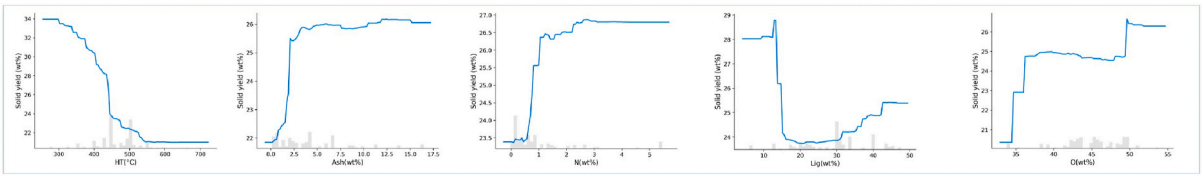
The heating value is a key indicator of bio-oil quality. While pyrolysis conditions affect product yields, the HHV of bio-oil is more influenced by feedstock properties. As shown in Fig. 9(D), the HHV increases approximately linearly with carbon content (43–50 wt%) and lignin content (25–30 wt%), while it decreases with increasing oxygen content. Higher carbon content in feedstock reduces the O/C ratio in bio-oil, enhancing its HHV. Research shows a strong positive correlation between hydrogen content in biomass and hydrogen content in bio-oil, while increased oxygen content in feedstock lowers the HHV [13,49]. This relationship is also described by Eq. (2). Additionally, Pütün et al. [50] found that higher hydrogen content in feedstock promotes aromaticity and cycloalkane formation, which mitigates the impact of oxygen on bio-oil and improves its HHV. Thus, deoxygenation pretreatment of biomass is beneficial for enhancing bio-oil quality [51,52].

Feedstock characteristics significantly influence the HHV of bio-oil, but pyrolysis temperature also plays a crucial role. The optimal HHV is achieved at approximately 450 °C, which is near the peak temperature for bio-oil production. Bok et al. found that the HHV and viscosity of biocrude increase between 400 °C and 450 °C, but decline with higher temperatures [53]. Similarly, Salehi et al. observed that for wood pyrolysis, oxygen content in bio-oil decreases and then increases as temperature rises from 425 °C to 550 °C [54]. Elevated oxygen content in bio-oil, which results in higher levels of acidic oxygen-containing compounds, reduces its HHV [55]. Thus, a temperature of around 450 °C is optimal for producing high-calorific bio-oil in fluidized bed pyrolysis. This differential sensitivity of bio-oil yield and HHV to temperature accounts for variations in energy conversion efficiency.

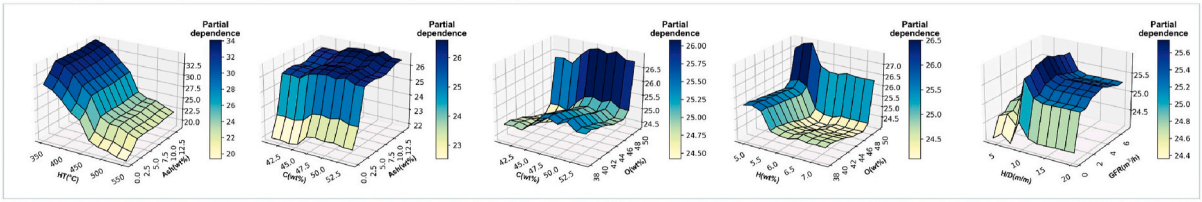
3.6. Further validation and GUI design

To evaluate the model's generalization ability, a new dataset from previous research was used for further validation (Supplementary Materials). Fig. 10 presents the actual experimental data and the errors between experimental and predicted values. The validation of the optimized LightGBM model on external data demonstrates a moderate linear relationship between predicted and true values, with an accuracy of approximately 80 %, peaking at 84 %. The reasonable distribution of residuals indicates that the model possesses a certain level of generalization capability. These results confirm the effectiveness of the optimization strategy. Future work will focus on expanding the dataset to

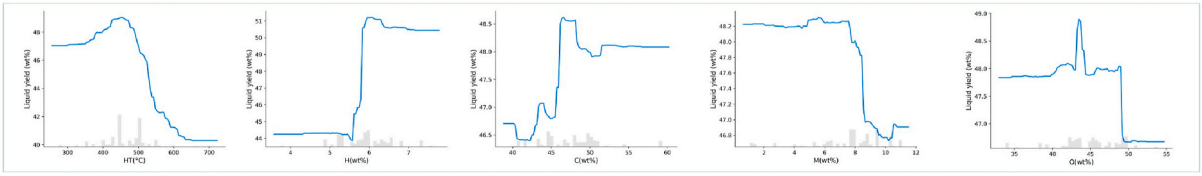
A1. One-way partial dependence plots for biochar yield (wt%) prediction



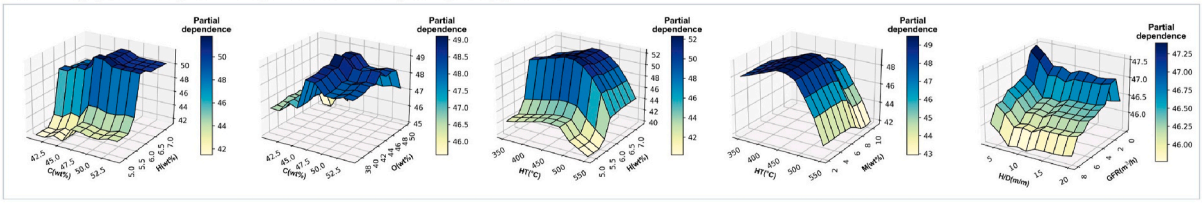
A2. Two-way partial dependence plots for biochar yield (wt%) prediction



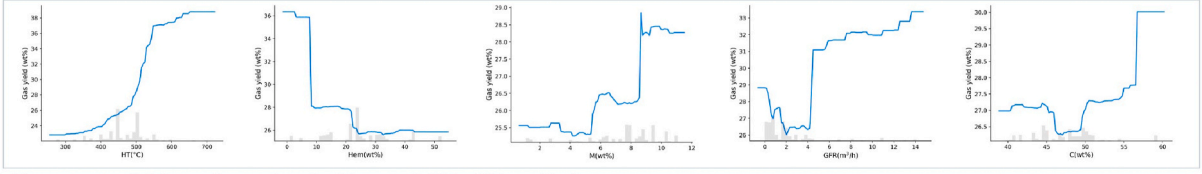
B1. One-way partial dependence plots for bio-oil yield (wt%) prediction



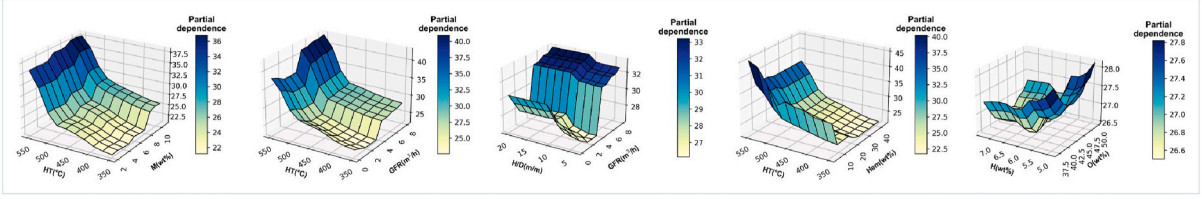
B2. Two-way partial dependence plots for bio-oil yield (wt%) prediction



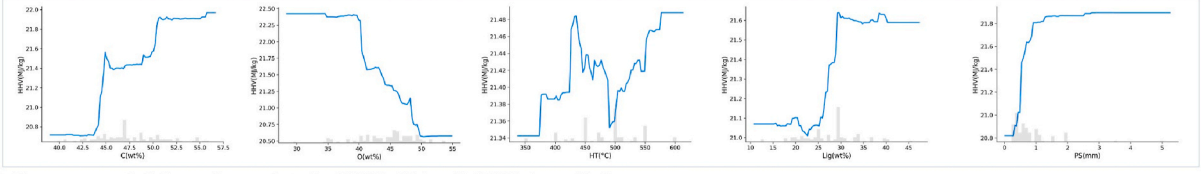
C1. One-way partial dependence plots for bio-gas yield (wt%) prediction



C2. Two-way partial dependence plots for bio-gas yield (wt%) prediction



D1. One-way partial dependence plots for HHV of bio-oil (MJ/kg) prediction



D2. Two-way partial dependence plots for HHV of bio-oil (MJ/kg) prediction

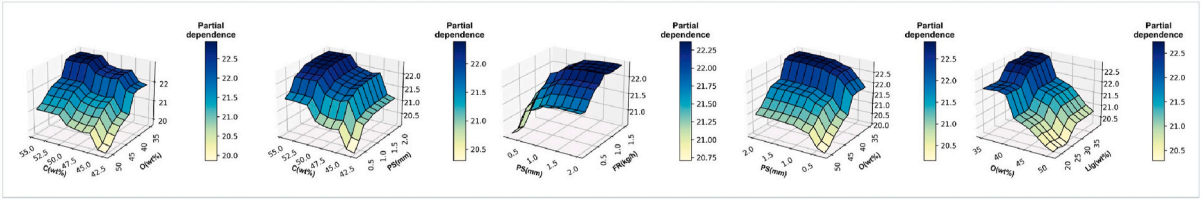


Fig. 9. One-way and two-way PDP for product yields and HHV of bio-oil.

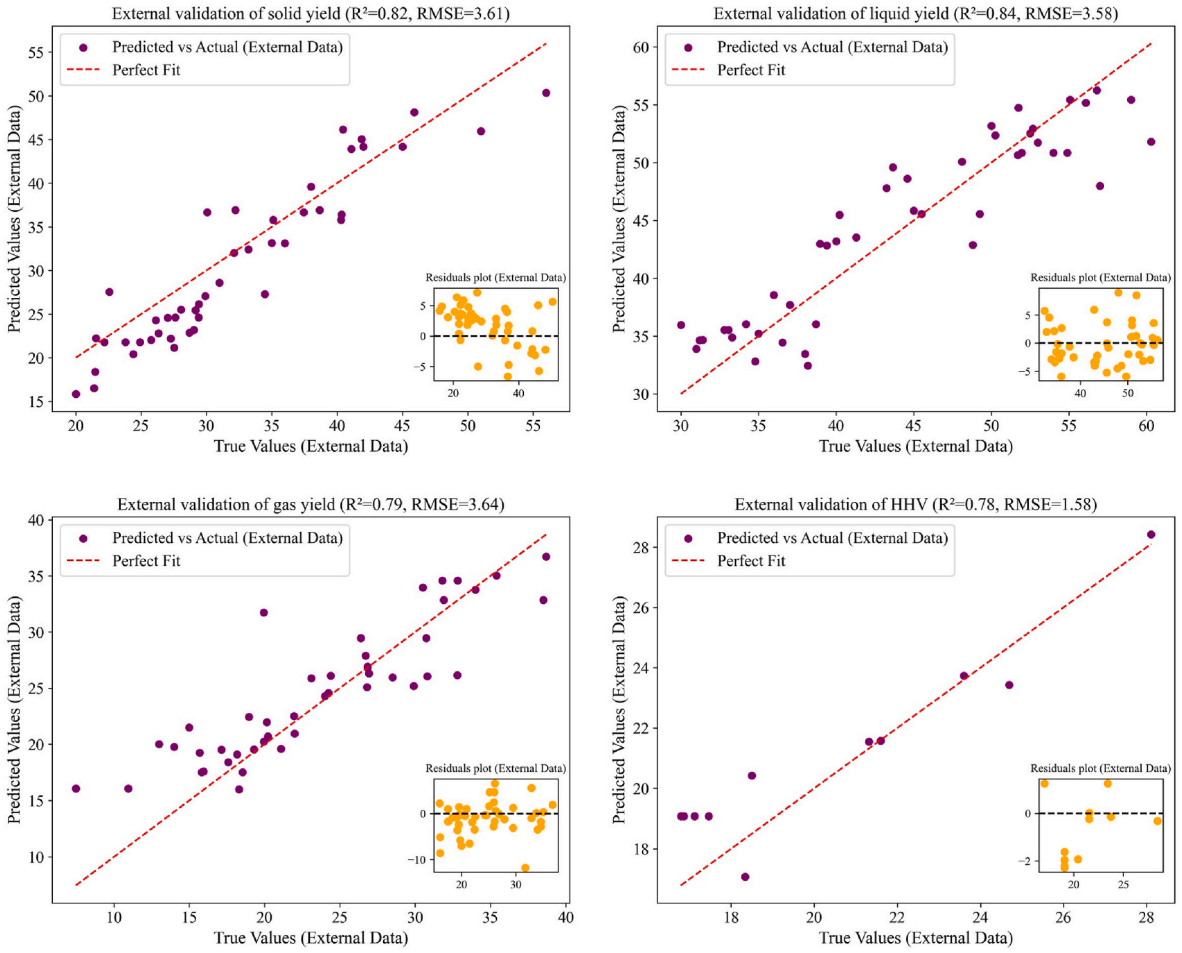


Fig. 10. Further validation results of the optimized LightGBM model with new experimental data.

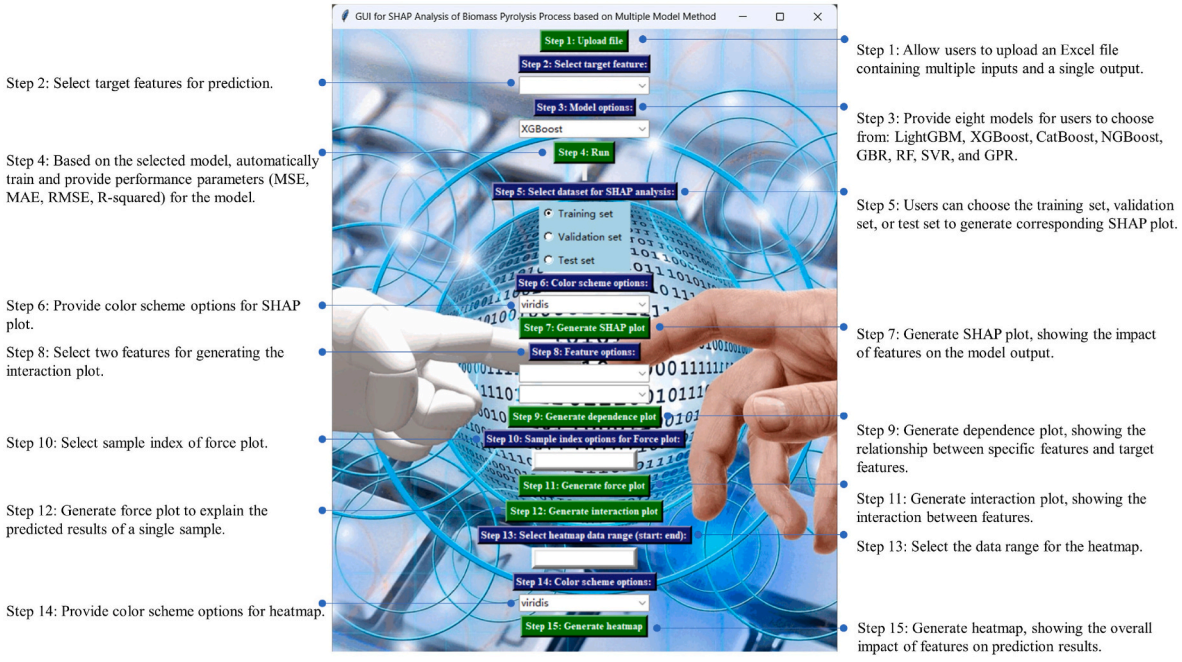


Fig. 11. GUI for SHAP analysis of biomass pyrolysis process based on multiple model method.

include a wider range of biomass sources, thereby enhancing the model's applicability and robustness under diverse pyrolysis conditions. This will involve collecting additional experimental data from different biomass types and incorporating them into the model training process.

Finally, a GUI for SHAP analysis was developed (Fig. 11). This GUI supports a range of ML models, including RF, GBDT, XGBoost, LightGBM, CatBoost, and NGBoost, with options to incorporate additional models. It provides functions for model selection and evaluation, enhancing the assessment and choice of predictive models. Additionally, the GUI integrates SHAP for advanced model interpretability. The developed GUI shows promising effectiveness, offering new insights for biomass pyrolysis research and other interdisciplinary applications.

4. Conclusion

This study developed and optimized ensemble ML models to predict the yields and heating value of bio-oil from biomass fast pyrolysis. Key findings include.

(1) Parameter Influence: Pyrolysis temperature, carbon and hydrogen content in biomass, and volatile matter are critical parameters influencing bio-oil yield and HHV. Moderate temperatures (~500 °C) and deoxygenation pretreatment significantly enhance bio-oil quality.
(2) Model Performance: The optimized LightGBM model achieved high predictive accuracy, with R^2 exceeding 0.93 for bio-oil yield. External validation demonstrated the model's strong generalization ability.
(3) Process Insights: SHAP and PDP analyses uncovered the influence and interactions of parameters, offering actionable insights for optimizing pyrolysis processes.
(4) Practical Application: The developed GUI bridges the gap between academic research and practical implementation, enabling efficient evaluation and optimization of pyrolysis conditions.

These findings advance the understanding of biomass pyrolysis and provide valuable guidance for industrial applications. To ensure the reliability of our results, we implemented outlier detection and removal, standardization and other preprocessing steps to identify and eliminate potential systematic errors. Additionally, we employed 5-CV and hyperparameter optimization to further reduce the impact of systematic errors on model training and prediction. Future work will focus on expanding the dataset to include a broader spectrum of biomass types and further refining the model to enhance its applicability and robustness under diverse pyrolysis conditions. This will involve collecting additional experimental data from different biomass types and incorporating them into the model training process to improve its generalization capability, as well as exploring the integration of online learning techniques to improve adaptability to dynamic environments.

CRediT authorship contribution statement

Longfei Li: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Zhongyang Luo:** Validation, Supervision, Methodology, Funding acquisition, Conceptualization. **Liwen Du:** Writing – review & editing, Investigation. **Feiting Miao:** Writing – review & editing, Investigation. **Longyi Liu:** Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China, No.52236011.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.energy.2025.136087.

Data availability

Data will be made available on request.

References

[1] Rezania S, Oryani B, Cho J, Talaiekhozani A, Sabbagh F, Hashemi B, Rupani PF, Mohammadi AA. Different pretreatment technologies of lignocellulosic biomass for bioethanol production: an overview. Energy 2020;199:117457. https://doi.org/10.1016/j.energy.2020.117457.
[2] Kabir G, Hameed BH. Recent progress on catalytic pyrolysis of lignocellulosic biomass to high-grade bio-oil and bio-chemicals. Renew Sustain Energy Rev 2017;70:945–67. https://doi.org/10.1016/j.rser.2016.12.001.
[3] Luo Z, Zhu W, Miao F, Zhou J. Catalytic hydrodeoxygenation of pyrolysis bio-oil to jet fuel: a review. Front Energy 2024;18:550–82. https://doi.org/10.1007/s11708-024-0943-7.
[4] Luo Z, Qian Q, Sun H, Wei Q, Zhou J, Wang K. Lignin-first biorefinery for converting lignocellulosic biomass into fuels and chemicals. Energies 2023;16. https://doi.org/10.3390/en16010125.
[5] Vispute TP, Zhang H, Sanna A, Xiao R, Huber GW. Renewable chemical commodity feedstocks from integrated catalytic processing of pyrolysis oils. Science 2010;330:1222–7. https://doi.org/10.1126/science.1194218.
[6] Li G, Wang R, Pang J, Wang A, Li N, Zhang T. Production of renewable hydrocarbon biofuels with lignocellulose and its derivatives over heterogeneous catalysts. Chem Rev 2024;124:2889–954. https://doi.org/10.1021/acs.chemrev.2c00756.
[7] Wang S, Dai G, Yang H, Luo Z. Lignocellulosic biomass pyrolysis mechanism: a state-of-the-art review. Prog Energy Combust Sci 2017;62:33–86. https://doi.org/10.1016/j.pecs.2017.05.004.
[8] Bridgwater A. Fast pyrolysis processes for biomass. Renew Sustain Energy Rev 2000;4:1–73. https://doi.org/10.1016/S1364-0321(99)00007-6.
[9] Patlolla SR, Katsu K, Sharafian A, Wei K, Herrera OE, Mérida W. A review of methane pyrolysis technologies for hydrogen production. Renew Sustain Energy Rev 2023;181:113323. https://doi.org/10.1016/j.rser.2023.113323.
[10] Abbas-Abadi MS, Ureel Y, Eschenbacher A, Vermeire FH, Varghese RJ, Oenema J, Stefanidis GD, Van Geem KM. Challenges and opportunities of light olefin production via thermal and catalytic pyrolysis of end-of-life polyolefins: towards full recyclability. Prog Energy Combust Sci 2023;96:101046. https://doi.org/10.1016/j.pecs.2022.101046.
[11] Miao F, Luo Z, Zhou Q, Du L, Zhu W, Wang K, Zhou J. Study on the reaction mechanism of C8+aliphatic hydrocarbons obtained directly from biomass by hydropyrolysis vapor upgrading. Chem Eng J 2023;464. https://doi.org/10.1016/j.cej.2023.142639.
[12] Cai W, Luo Z, Zhou J, Wang Q. A review on the selection of raw materials and reactors for biomass fast pyrolysis in China. Fuel Process Technol 2021;221. https://doi.org/10.1016/j.fuproc.2021.106919.
[13] Tang Q, Chen Y, Yang H, Liu M, Xiao H, Wu Z, Chen H, Naqvi SR. Prediction of bio-oil yield and hydrogen contents based on machine learning method: effect of biomass compositions and pyrolysis conditions. Energy Fuels 2020;34:11050–60. https://doi.org/10.1021/acs.energyfuels.0c01893.
[14] Kostetskyy P, Broadbelt LJ. Progress in modeling of biomass fast pyrolysis: a review. Energy Fuels 2020;34:15195–216. https://doi.org/10.1021/acs.energyfuels.0c02295.
[15] Li L, Luo Z, Miao F, Du L, Wang K. Prediction of product yields from lignocellulosic biomass pyrolysis based on gaussian process regression. J Anal Appl Pyrolysis 2024;177:106295. https://doi.org/10.1016/j.jaap.2023.106295.
[16] Zhu X, Li Y, Wang X. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol 2019;288:121527. https://doi.org/10.1016/j.biortech.2019.121527.
[17] Xing J, Luo K, Wang H, Jin T, Fan J. Novel sensitivity study for biomass directional devolatilization by random forest models. Energy Fuels 2020;34:8414–23. https://doi.org/10.1021/acs.energyfuels.0c00822.
[18] Leng E, He B, Chen J, Liao G, Ma Y, Zhang F, Liu S, E J. Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning. Energy 2021;236:121401. https://doi.org/10.1016/j.energy.2021.121401.
[19] Zhang T, Cao D, Feng X, Zhu J, Lu X, Mu L, Qian H. Machine learning prediction of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions. Fuel 2022;312:122812. https://doi.org/10.1016/j.fuel.2021.122812.

[20] K N Y, T PD, P S, S Y K R K, Varjani S, AdishKumar S, Kumar G, J RB. Lignocellulosic biomass-based pyrolysis: a comprehensive review. Chemosphere 2022;286:131824. https://doi.org/10.1016/j.chemosphere.2021.131824.
[21] Channiwala SA, Parikh PP. A unified correlation for estimating HHV of solid, liquid and gaseous fuels. Fuel 2002;81:1051–63. https://doi.org/10.1016/S0016-2361(01)00131-4.
[22] Mu Y, Liu X, Wang L. A Pearson's correlation coefficient based decision tree and its parallel implementation. Inf Sci 2018;435:40–58. https://doi.org/10.1016/j.ins.2017.12.059.
[23] Sun Y, Li Y, Wang R, Ma R. Modelling potential land suitability of large-scale wind energy development using explainable machine learning techniques: applications for China, USA and EU. Energy Convers Manag 2024;302:118131. https://doi.org/10.1016/j.enconman.2024.118131.
[24] Angın D. Effect of pyrolysis temperature and heating rate on biochar obtained from pyrolysis of safflower seed press cake. Bioresour Technol 2013;128:593–7. https://doi.org/10.1016/j.biortech.2012.10.150.
[25] Rijo B, Soares Dias AP, Ramos M, Ameixa M. Valorization of forest waste biomass by catalyzed pyrolysis. Energy 2022;243:122766. https://doi.org/10.1016/j.energy.2021.122766.
[26] Onarheim K, Hannula I, Solantausta Y. Hydrogen enhanced biofuels for transport via fast pyrolysis of biomass: a conceptual assessment. Energy 2020;199:117337. https://doi.org/10.1016/j.energy.2020.117337.
[27] Du L, Luo Z, Wang K, Miao F, Zhou Q, Zhu W, Li L. Study on ex-situ catalytic pyrolysis of poplar to produce aromatics over a dual-catalyst system of hydroxyapatite and HZSM-5. J Energy Inst 2024;114. https://doi.org/10.1016/j.joei.2024.101597.
[28] Hu Y, Gong M, Feng S, C, Charles, Xu, Bassi A. A review of recent developments of pre-treatment technologies and hydrothermal liquefaction of microalgae for bio-crude oil production. Renew Sustain Energy Rev 2019;101:476–92. https://doi.org/10.1016/j.rser.2018.11.037.
[29] Leng L, Li T, Zhan H, Rizwan M, Zhang W, Peng H, Yang Z, Li H. Machine learning-aided prediction of nitrogen heterocycles in bio-oil from the pyrolysis of biomass. Energy 2023;278:127967. https://doi.org/10.1016/j.energy.2023.127967.
[30] Liu S, Zhao A, He Z, Li Y, Bi D, Gao X. Effects of temperature and urea concentration on nitrogen-rich pyrolysis: pyrolysis behavior and product distribution in bio-oil. Energy 2022;239:122443. https://doi.org/10.1016/j.energy.2021.122443.
[31] Huang Y, Sekyere DT, Zhang J, Tian Y. Fast pyrolysis behaviors of biomass with high contents of ash and nitrogen using TG-FTIR and Py-GC/MS. J Anal Appl Pyrolysis 2023;170:105922. https://doi.org/10.1016/j.jaap.2023.105922.
[32] Cahanap DR, Mohammadpour J, Jalalifar S, Mehrjoo H, Norouzi-Apourvari S, Salehi F. Prediction of three-phase product yield of biomass pyrolysis using artificial intelligence-based models. J Anal Appl Pyrolysis 2023;172:106015. https://doi.org/10.1016/j.jaap.2023.106015.
[33] Mahdaviara M, Sharifi M, Bakhshian S, Shokri N. Prediction of spontaneous imbibition in porous media using deep and ensemble learning techniques. Fuel 2022;329:125349. https://doi.org/10.1016/j.fuel.2022.125349.
[34] Zhu X, Shen X, Chen K, Zhang Z. Research on the prediction and influencing factors of heavy duty truck fuel consumption based on LightGBM. Energy 2024;296:131221. https://doi.org/10.1016/j.energy.2024.131221.
[35] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: a highly efficient gradient boosting decision tree. In: Neural inf. Process. Syst.; 2017. https://api.semanticscholar.org/CorpusID:3815895.
[36] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD int. Conf. Knowl. Discov. Data min. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[37] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67. https://doi.org/10.1007/s10462-020-09896-5.
[38] Sagi O, Rokach L. Ensemble learning: a survey. WIREs Data Min. Knowl Discov 2018;8:e1249. https://doi.org/10.1002/widm.1249.
[39] Wan Mahari WA, Azwar E, Foong SY, Ahmed A, Peng W, Tabatabaei M, Aghbashlo M, Park Y-K, Sonne C, Lam SS. Valorization of municipal wastes using co-pyrolysis for green energy production, energy security, and environmental sustainability: a review. Chem Eng J 2021;421:129749. https://doi.org/10.1016/j.cej.2021.129749.
[40] Wei L, Xu S, Zhang L, Zhang H, Liu C, Zhu H, Liu S. Characteristics of fast pyrolysis of biomass in a free fall reactor. Fuel Process Technol 2006;87:863–71. https://doi.org/10.1016/j.fuproc.2006.06.002.
[41] Duman G, Okutucu C, Ucar S, Stahl R, Yanik J. The slow and fast pyrolysis of cherry seed. Bioresour Technol 2011;102:1869–78. https://doi.org/10.1016/j.biortech.2010.07.051.
[42] Kan T, Strezov V, Evans TJ. Lignocellulosic biomass pyrolysis: a review of product properties and effects of pyrolysis parameters. Renew Sustain Energy Rev 2016;57:1126–40. https://doi.org/10.1016/j.rser.2015.12.185.
[43] Yang SI, Wu MS, Wu CY. Application of biomass fast pyrolysis part I: pyrolysis characteristics and products. Energy 2014;66:162–71. https://doi.org/10.1016/j.energy.2013.12.063.
[44] Zheng J, Yi W, Wang N. Bio-oil production from cotton stalk. Energy Convers Manag 2008;49:1724–30. https://doi.org/10.1016/j.enconman.2007.11.005.
[45] Stefanidis SD, Kalogiannis KG, Iliopoulou EF, Michailof CM, Pilavachi PA, Lappas AA. A study of lignocellulosic biomass pyrolysis via the pyrolysis of cellulose, hemicellulose and lignin. J Anal Appl Pyrolysis 2014;105:143–50. https://doi.org/10.1016/j.jaap.2013.10.013.
[46] Ateş F, Pütün E, Pütün AE. Fast pyrolysis of sesame stalk: yields and structural analysis of bio-oil. J Anal Appl Pyrolysis 2004;71:779–90. https://doi.org/10.1016/j.jaap.2003.11.001.
[47] Heidari A, Stahl R, Younesi H, Rashidi A, Troeger N, Ghoreyshi AA. Effect of process conditions on product yield and composition of fast pyrolysis of Eucalyptus grandis in fluidized bed reactor. J Ind Eng Chem 2014;20:2594–602. https://doi.org/10.1016/j.jiec.2013.10.046.
[48] Park HJ, Park Y-K, Kim JS. Influence of reaction conditions and the char separation system on the production of bio-oil from radiata pine sawdust by fast pyrolysis. Fuel Process Technol 2008;89:797–802. https://doi.org/10.1016/j.fuproc.2008.01.003.
[49] Bilgen S, Keleş S, Kaygusuz K. Calculation of higher and lower heating values and chemical exergy values of liquid products obtained from pyrolysis of hazelnut cupulae. Energy 2012;41:380–5. https://doi.org/10.1016/j.energy.2012.03.001.
[50] Pütün AE, Özcan A, Gerçel HF, Pütün E. Production of biocrudes from biomass in a fixed-bed tubular reactor: product yields and compositions. Fuel 2001;80:1371–8. https://doi.org/10.1016/S0016-2361(01)00021-7.
[51] Qian Q, Luo Z, Sun H, Wei Q, Shi J, Li S. Comparing physicochemical characteristics and depolymerization behaviors of lignins derived from different pretreatment processes. Fuel Process Technol 2023;250. https://doi.org/10.1016/j.fuproc.2023.107921.
[52] Chen D, Gao A, Cen K, Zhang J, Cao X, Ma Z. Investigation of biomass torrefaction based on three major components: Hemicellulose, cellulose, and lignin. Energy Convers Manag 2018;169:228–37. https://doi.org/10.1016/j.enconman.2018.05.063.
[53] Bok JP, Choi HS, Choi JW, Choi YS. Fast pyrolysis of Miscanthus sinensis in fluidized bed reactors: characteristics of product yields and biocrude oil quality. Energy 2013;60:44–52. https://doi.org/10.1016/j.energy.2013.08.024.
[54] Salehi E, Abedi J, Harding T. Bio-oil from sawdust: effect of operating parameters on the yield and quality of pyrolysis products. Energy Fuels 2011;25:4145–54. https://doi.org/10.1021/ef200688y.
[55] Li C, Nishu, Yellezuome D, Li Y, Liu R, Cai J. Enhancing bio-aromatics yield in bio-oil from catalytic fast pyrolysis of bamboo residues over bi-metallic catalyst and reaction mechanism based on quantum computing. Fuel 2023;336:127158. https://doi.org/10.1016/j.fuel.2022.127158.