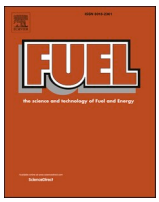
Fuel 366 (2024) 131346

Contents lists available at ScienceDirect

Fuel

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



Full Length Article

Enhancing biomass Pyrolysis: Predictive insights from process simulation integrated with interpretable Machine learning models

Douglas Chinenye Divine [a], Stell Hubert [a], Emmanuel I. Epelle [b], Alaba U. Ojo [c], Adekunle A. Adeleke [d], Chukwuma C. Ogbaga [i,j], Olugbenga Akande [e], Patrick U. Okoye [f], Adewale Giwa [g], Jude A. Okolie [h,*]

[a] Department of Process Engineering and Energy Technology, Hochschule Bremerhaven
[b] School of Engineering, Institute for Infrastructure and Environment, The University of Edinburgh, Edinburgh EH9 3JL, UK
[c] Department of Chemical Engineering, University of South Carolina, Columbia, SC, USA
[d] Nile University of Nigeria, Abuja. Address: Plot 681, Cadastral Zone C-OO, Research & Institution Area Nigeria, Airport Rd, Jabi 900001, Abuja, Nigeria
[e] Department of Advanced Convergence, Handong Global University, 558 Handong-ro, Heunghae-eup, Buk-gu, Pohang, Gyeongsangbuk-do 37554, Republic of Korea
[f] Instituto de Energías Renovables (IER-UNAM), Privada Xochicalco s/n Col. Centro, Temixco, Morelos 62580, Mexico
[g] Chemical and Water Desalination Engineering Program, Department of Mechanical, and Nuclear Engineering, P.O. Box 27272, University City, University of Sharjah, Sharjah, United Arab Emirates
[h] Gallogly College of Engineering, University of Oklahoma, USA
[i] Departments of Biotechnology, Microbiology, and Biochemistry, Philomath University, Kuje, Abuja, Nigeria
[j] Department of Biological Sciences, Coal City University, Enugu, Nigeria

A B S T R A C T

Waste biomass pyrolysis is a promising thermochemical conversion process for the production of biofuels and sustainable materials. However, it is challenging to accurately predict the properties and yield of products formed during pyrolysis. Machine learning (ML) is a useful tool for predicting the performance of a process. In the present study, ML algorithms integrated with process simulation were explored to accurately model waste biomass pyrolysis based on properties such as H/C, O/C, oil yield, gas yield, and char yield. Six different ML models including Random Forest (RF), Gradient Boosting Regressor (GBR), eXtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Artificial Neural Network (ANN), and Stochastic Gradient Descent (SGD) were used to model pyrolysis process. It was found that the out-of-the-box (without optimization) models for RF, XGBoost, ANN, and GBR performed the best and did not benefit from hyperparameter optimization. The GBR was identified as the most effective among various ML models. It accurately predicted yields of gas, biochar, bio-oil yields, and their H/C and O/C compositions. GBR effectively demonstrated the complex relationships between these variables. The box plot showing the root mean squared logarithmic error (RMSE) revealed that the GBR model had the best overall performance with a value less than 0.03. Also, the partial dependence plot and SHAP feature importance were evaluated to better understand each feature's effect on the output. Lastly, a shareable graphical user interface (GUI) was created to enable researchers explore and predict pyrolysis yield.

1. Introduction

The use of non-renewable fossil fuels has led to several challenges, such as energy insecurity and environmental pollution [1]. To address these challenges, significant attention has been devoted to exploring alternative energy sources that are both abundant and renewable. One such source of energy is biomass, which is obtained from organic matter such as plants and animals [2]. Biomass is a promising energy source due to its plentiful supply and renewable nature. It plays a role in a balanced carbon cycle, wherein the carbon dioxide (CO_2) emitted during its combustion is offset by the CO_2 absorbed by plants in photosynthesis, potentially resulting in a neutral effect on atmospheric CO_2 levels [3].

Biomass can be transformed into sustainable fuels and chemicals via thermochemical and biological conversion processes [4]. While both biological and thermochemical conversion processes for biomass are promising, each has its challenges. Biological processes often require longer residence times and the cultivation of microorganisms [4]. On the other hand, thermochemical conversion is efficient and cost-effective for converting biomass into biofuels. These biofuels can then be synthesized into various chemicals or used directly as fuel [5]. One major thermochemical process used for the conversion of biomass to green fuel and value-added process is pyrolysis. It involves thermally decomposing biomass in an oxygen-free environment, resulting in various products like bio-oil, biochar, and gases, depending on specific reaction conditions. This process stands out for its efficiency in biomass conversion.

* Corresponding author at: Engineering Pathways, Gallogly College of Engineering, University of Oklahoma, Norman United States.
E-mail address: Jude.okolie@ou.edu (J.A. Okolie).

https://doi.org/10.1016/j.fuel.2024.131346
Received 26 November 2023; Received in revised form 17 January 2024; Accepted 26 February 2024
Available online 1 March 2024

Nomenclature

Abbreviations and symbols

AdaBoost	Adaptive boosting
ANN	Artificial neural networks
Ash	Ash content
C	Carbon
FC	Fixed carbon
GAN	Generative Adversarial Network
GBR	Gradient boosting for regression
H	Hydrogen
H/C	Ratios of hydrogen to carbon
HT	Heating temperature
IQR	Interquartile range
MAE	Mean absolute error
ML	Machine learning
N	Nitrogen
O	Oxygen
O/C	Ratios of oxygen to carbon
PDP	Partial dependence plot
PIML	Physics-Informed Machine Learning
PS	Particle Size
PT	Pyrolysis temperature
R^2	Regression coefficient
RF	Random forest
RMSE	Root-mean-square error
RMSLE	Root mean squared logarithmic error
S	Sulphur
SCC	Spearman correlation coefficient
SGD	Stochastic gradient descent
SHAP	SHapley Additive exPlanations
SVM	Support vector machines
VAE	Variational Autoencoders
Vm	Volatile matter
XGBoost	eXtreme gradient boosting

Bio-oil, as a complex product of biomass pyrolysis, has attracted significant attention due to its potential as a versatile fuel source. Its potential uses include being a boiler fuel, transportation fuel, and a precursor for chemical products. This versatility positions bio-oil as a promising alternative to traditional fossil fuels [6]. The chemical composition of bio-oil is diverse, containing over 350 different compounds, including acids, aldehydes, ketones, esters, alkanes, and various aromatic compounds [7]. These compounds are primarily composed of carbon, hydrogen, and oxygen elements, with the hydrogen and oxygen contents serving as important indicators of bio-oil quality. The hydrogen content of bio-oil is crucial as it influences the fuel's calorific value and chemical structure [8]. However, bio-oil typically has a lower calorific value compared to other transportation fuels. This is due to its higher water content and a larger proportion of oxygenated chemicals, necessitating further treatment for its upgrading and optimization as a fuel source [9].

It is important to note that the properties of biomass precursors could influence the composition and yield of bio-oil as well as other products during pyrolysis. The relationship between biomass properties and product yield and characteristics has been reported by several studies [10,11]. However, the relationship between biomass composition and pyrolysis product yield and characteristics is still unclear. Numerous studies have been conducted using thermogravimetry and scanning calorimetry to develop kinetic models that investigate the relationship between the characteristics of bio-oil (specifically yield and compositions), biomass compositions and pyrolysis conditions [12,13]. One such study, conducted by Lv et al.[14] established a kinetic model for biomass pyrolysis by simulating the fluidized bed design of biomass pyrolysis. At very high temperatures (800 ℃), their model's calculated data was consistent with experimental data from pine sawdust, lignin, and cellulose pyrolysis [14]. However, it is important to note that a kinetic study conducted on a particular type of biomass and its corresponding pyrolysis conditions is not universally applicable, and the resulting model cannot be directly applied to other biomass systems. Thus, it is imperative to establish a universal relationship between bio-oil characteristics, biomass compositions and pyrolysis conditions.

Traditional models like thermodynamics, kinetics, computational fluid dynamics (CFD), and process modelling have been used to establish and represent complex relationships during thermochemical conversion processes. However, they often face challenges in accurately capturing input–output relationships during pyrolysis and gasification [15]. These models rely on assumptions and simplifications, limiting their practical implementation. CFD models need substantial computational resources and specific parameters for reliable simulations. Kinetic models depend heavily on estimating complex reaction rates, and their reaction mechanisms may be incomplete or poorly understood [15]. Thermodynamic models usually assume equilibrium states, which are rarely achieved in practice. Process modelling often requires costly software. Thus, there is a need for a more reliable, accurate, and efficient modelling approach for thermochemical processes.

Recent advancements in artificial intelligence have opened new avenues for research, leading to promising outcomes in modelling input–output relationships [16]. Machine learning (ML) techniques such as random forest (RF), support vector machines (SVM), and gradient boosting (GB) have been developed to predict various parameters related to thermochemical conversion processes [17,18]. Previous studies have demonstrated the effectiveness of ML in thermochemical conversion processes for optimization and evaluating the relationship between product yield and feedstock properties. For instance, Zhu et al. established a correlation between biomass structural composition, pyrolysis conditions, and biochar yield and successfully applied the RF method to predict biochar yield [19]. Xing et al. also utilized the RF model to predict the chemical compositions of different biomasses and compared their results with existing relevant data and experimental data [20]. While numerous studies have explored the connection between bio-oil yield and biomass properties (Table 1), studies evaluating the simultaneous influence of operating conditions and biomass composition are scarcely reported. Therein lies the motivation for this study.

The present study combines data from literature, a publicly available database and process simulation to develop a ML model. The model was then used to comprehensively evaluate the impact of biomass properties (proximate, ultimate, and compositional analysis), pyrolysis operating conditions (heating rate (HR), biomass particle size (PS), and pyrolysis temperature (PT) on product yield and bio-oil properties. This study's innovative integration of process simulation with interpretable machine learning models in waste biomass pyrolysis could significantly advance biomass conversion efficiency, potentially revolutionizing renewable energy production. It opens avenues for precise, optimized pyrolysis processes, laying the groundwork for future research in sustainable energy and practical applications in green fuel production.

2. Methodology

2.1. Data collection and preprocessing

This section meticulously describes the steps employed in processing the input data for the ML models. These include sources of data and the collection strategies, methods employed in cleaning the data and

Table 1
An overview of recent research and reviews on ML applications in biomass pyrolysis.

Research objectives	Key findings	References
The research aims to predict biochar yield during pyrolysis using input parameters such as biomass properties and pyrolysis conditions	The ANN model, combined with the Rao-2 algorithm, was the most effective in predicting biochar yield. The integrated model achieved high accuracy ($R^2 \sim 0.93$) and low error (RMSE ~ 1.74 %). Analysis revealed that the most influential factors for this performance were pyrolysis temperature (56 %), residence time (23 %), and heating rate (8 %).	Ullah et al. [21]
The study aims to evaluate the performance of Machine learning models for the prediction of product yield and composition during microwave pyrolysis	In a comparison of three machine learning models (support vector regressor, random forest regressor, and gradient boost regressor) using 14 descriptors, the gradient boost regressor model outperformed the others. It achieved better prediction accuracy, as indicated by its higher R^2 value (over 0.822), lower RMSE (less than 12.38), and lower RMSE (below 0.765).	Yang et al. [22]
The study presents a comprehensive review of machine learning applications in advancing thermochemical conversion processes.	Machine learning can advance thermochemical conversion process development, especially in the area of predictive performance and optimization.	Li et al. [23]
The study aims to characterize wastewater sludge pyrolysis using machine learning models.	The random forest regression model was the most accurate among the tested models, with a correlation coefficient over 0.813 and a relative mean squared error under 12.51. SHAP analysis using this model identified the five most significant factors affecting bio-oil yield: ash content, fixed carbon content, operating temperature, and volatile matter content.	Shahbeik et al. [24]
The study proposed a machine learning model for predicting product yield during co-pyrolysis of biomass and coal.	In predicting the co-pyrolysis of coal and biomass, both the Extra Trees (ET) and Random Forest (RF) models perform well. However, the ET model generally surpasses the RF in terms of accuracy, better generalization capabilities, and lesser tendency to overfit.	Wei et al. [25]
The study highlights ML applications in pyrolysis process optimization and control, predicting product yield, real-time monitoring, life-cycle assessment (LCA), and techno-economic analysis (TEA).	ML models can help reduce uncertainties in the economic and environmental assessments of pyrolysis. Furthermore, the use of ML can advance the commercialization of various biomass pyrolysis technologies.	Akinpelu et al. [26]
The present study combines data from literature, a publicly available database and process simulation to develop a ML model. The model was then used to comprehensively evaluate the impact of biomass properties and pyrolysis operating conditions on product yield and bio-oil properties.	The GBR was identified as the most effective among various machine learning models. It accurately predicts yields of gas, biochar, and bio-oil, and their H/C and O/C compositions, using partial dependence plots (PDP) for analysis. The SHAP importance score analysis of the GBR model result revealed that the ultimate analysis data of the biomass feedstock and pyrolysis conditions had a significant impact on the bio-oil, biochar, and gas yield	This work

statistical analysis used in understanding the data.

2.1.1. Data collection

The data utilized in this study were obtained from a combination of existing literature [10,20,27–29], Phyllis database [30] and process simulation as illustrated in Fig. 1.

Various experimental datasets were gathered from the literature on pyrolysis yield including bio-oil, gases and biochar yield, bio-oil H/C ratio, and O/C ratio. Additional data related to compositional analysis of the biomass as well as the proximate and ultimate analysis were also compiled. Aspen plus process simulation was used to determine some product yields that were not obtainable from the literature. Aspen plus process simulator was employed to model gas yield, bio-oil yield, and biochar yield under the same pyrolysis conditions and raw material compositions found in the literature. The simulation results were also expressed as percentages to mitigate potential simulation errors and ensure data consistency. It should be mentioned that the combined experimental and process simulation dataset appears diverse and representative, capturing a range of biomass compositions and pyrolysis conditions. The variability in key components like cellulose (Cel), hemicellulose (Hem), lignin (Lig), volatile matter (Vm%), ash (Ash%), and fixed carbon (FC%) is well-represented, as indicated by the standard deviations. Moreover, the dataset's broad range in values suggests it captures the variability in biomass compositions and pyrolysis conditions effectively. The dataset's broad range of values for each variable is crucial as it allows the model to learn from a diverse array of data, enhancing its generalization ability. Normalization was performed to help address any biases and potential variability in the data.

The diagrammatic representation of the process simulation in Aspen Plus is presented in Fig. 2. The pyrolysis process was simulated with the introduction of 100 kg/h of biomass sent to the dryer. The dryer helps to reduce the moisture content of the waste tires to < 3 wt%. Following this, the dried biomass feedstock was introduced into a fluidized bed reactor, serving as the site for pyrolysis. Augmenting this stage, a nitrogen flow rate of 49, 1 kg/h, compressed to a pressure of 20 bar, is infused into the reactor to create an inert atmosphere required for pyrolysis. A RYield reactor operating at 1 bar was used to model the dynamics of the pyrolysis reactor guided by the study of Liu et al. [31]. The resultant bio-oil composition primarily encompasses esters, alcohols, acids, furans, phenolics, aromatics, and ketones.

It should be mentioned that the Aspen Plus simulation was steady state and isothermal. Each biomass feedstock was defined as a non-conventional stream based on their individual proximate and ultimate analysis information. In order to define a non-conventional stream, RYield reactor block was used to decompose the biomass materials into conventional components including C, H_2, O_2, N_2, Cl_2, H_2O, ash and S [32]. While a calculator block was incorporated into the RYield block, and its function is executed using a programmed FORTRAN subroutine statement. Details of the non-conventional reactor modelling in Aspen Plus can be found in our previous study [32]. The pyrolysis reactor is modeled using an RYield reactor operating at 450 °C and 1 bar in the presence of nitrogen which is used to create the inert environment.

The resulting char is a mixture of solid carbon and ash. This char is separated using a solid separator and then burned to generate heat, as combustion is an exothermic reaction. This process not only produces heat but also significantly contributes to the total energy output of the plant. Meanwhile, the oil and gas fractions are carefully cooled until

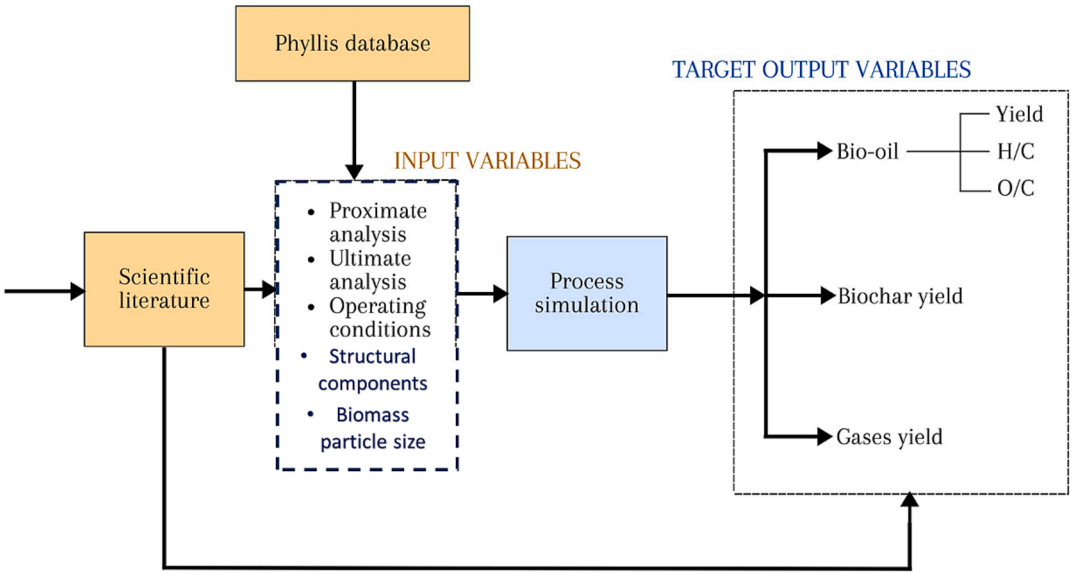


Fig. 1. Schematics of the data collection methodology.

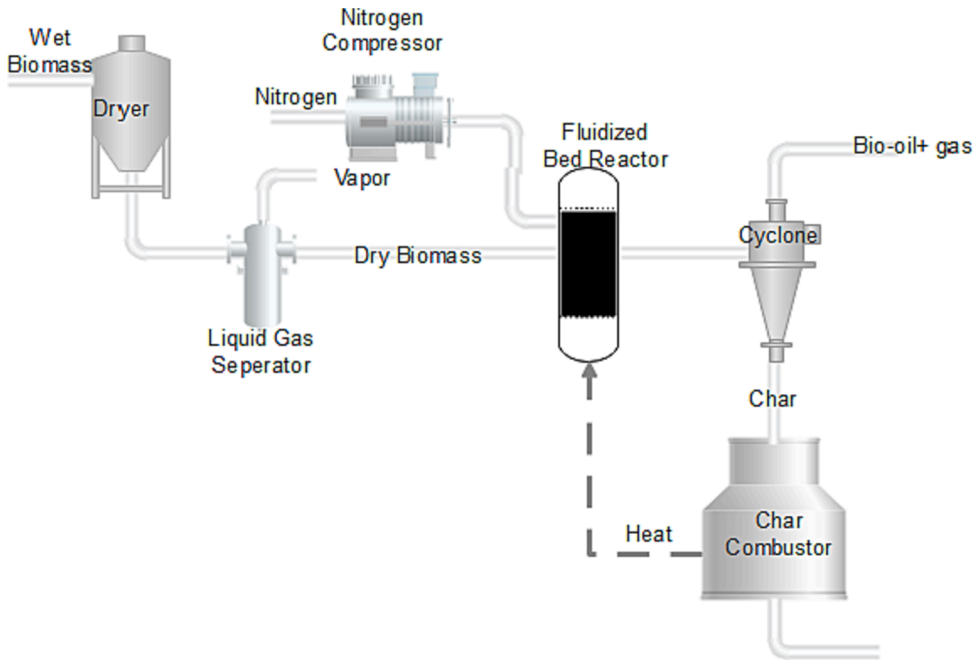


Fig. 2. Schematics of the Aspen plus process simulation for biomass pyrolysis.

they reach a temperature of 25 °C.

2.1.2. Data classification

The factors influencing biomass pyrolysis and product yield, as well as composition in this study, are categorized into five groups [3327]:

- The structural or chemical components of biomass, comprising the relative proportions of lignin, cellulose, and hemicellulose.
- The elemental composition (ultimate analysis) of biomass, including carbon, hydrogen, oxygen, and nitrogen (C–H-O-N).
- The proximate analysis of the biomass materials, including the amounts of volatiles, ash, and fixed carbon contents (VM-Ash-FC).
- Biomass particle size (PS).
- Pyrolysis conditions.

To improve the data quality and reduce experimental inaccuracies associated with the data, the results of the proximate and ultimate analyses of the raw materials were presented in percentages.

2.1.3. Data cleaning

Data cleaning is an imperative data preparation step before building and training any model. This guarantees maximum performance and confidence in the developed models. In this study, all data were cleaned, and any missing values were replaced with the average of the entire dataset, as many ML algorithms require data to be in a uniform and specific format. This is necessary because most algorithms cannot process missing or invalid entries, which must be addressed beforehand.

Some ML models prefer normally distributed features; thus, each feature was standardized by subtracting the mean and scaling samples to unit variance. Standardization is helpful because many ML algorithms

assume that the data is normally distributed, and features are on the same scale. It also helps to reduce the impact of outliers since these will be scaled down along with the rest of the data [34]. The performance of each model was evaluated using the StandardScaler, MinMaxScaler, and RobustScaler provided by sklearn. Preprocessing [35].

The models' performance was assessed using various scalers provided by sklearn. preprocessing library. Among them, the MinMaxScaler demonstrated superior performance, prompting its selection. MinMaxScaler is a ML preprocessing technique that transforms a dataset's features into a specific range, typically between 0 and 1. This is achieved by subtracting the minimum feature value from each data point and dividing the result by the feature's range (i.e., the difference between the minimum and maximum values). This process normalizes the data and is particularly beneficial for algorithms sensitive to the scale of input features, such as k-nearest neighbors and artificial neural networks.

Additionally, to improve the model's performance, outliers were removed from the dataset. The Spearman correlation coefficient (SCC) was used to evaluate the linear relationship between pairs of variables. Utilizing the SCC in the development of a ML model is advantageous for various reasons: it uncovers non-linear relationships between variables, determines which features are most strongly associated with the outcome, and identifies outliers, namely variables that show a weak correlation with the target variable. The method for calculating the SCC is presented in Equation (1) [36]:

$$SCC = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (1)$$

Where SCC = Spearman's rank correlation coefficient,
d_i = difference between the two ranks of each observation,
n = number of observations.

2.2. Model development

An overview of different ML model types employed in this study is described herein. These include the tree-based models as well as the ANN models. Hyperparameter optimization strategies as well as the metrics used in evaluating the performance of the ML models are also presented in this section.

2.2.1. Model types

In this study, a range of model types were compared, recognizing that the most suitable model often depends on the specific nature of the problem and the data involved. Among these, tree-based models are a crucial family of models. These models operate by deriving simple decision rules from the features in the dataset to predict a target variable. In practical applications, many models employ a collection of trees to mitigate the variance that typically arises from using a single decision tree. Random Forest (RF) is one such method.

Bootstrap aggregating, commonly referred to as bagging, operates by averaging the outcomes of multiple independent learners. This approach effectively mitigates the noise issues typically encountered in individual trees by averaging the results from numerous decision trees. In contrast, boosting algorithms focus on training and adjusting the weights of several weak learners to construct a robust ensemble model. While boosting is centered on refining and combining weak learners, bagging emphasizes averaging the outputs of various independent models. This study evaluated three algorithms in this category: Gradient Boosting for Regression (GBR), XGBoost, and AdaBoost. Although modern deep neural networks often surpass SVMs and some other model types in performance when large datasets are available, the algorithms discussed here may still be among the most accurate for the size of the dataset used in this study [24]. Consequently, a simpler Multilayer Perceptron neural network was also considered as an alternative to a deep neural network. Notably, Artificial Neural Networks (ANN) have been the most popular machine learning method for modeling biomass and waste gasification.

The ML models used in this study including RF, GBR, XGBoost, AdaBoost, ANN, and SGD were selected for several reasons. These models have shown promising performance for the prediction of biofuel yield during thermochemical processes. For instance, Khan et al. showed that ANN coupled with metaheuristic algorithms could effectively predict biochar yield during pyrolysis [21]. Okolie et al. reported the promising predicting performance of RF and GBR in predicting the structural composition of biomass for subsequent thermochemical conversion processes [37]. Moreover, the RF model offers robustness and accuracy for complex datasets while the GBR model is often used to handle unbalanced datasets. XGBoost provides high performance and speed with scalability, AdaBoost is effective for boosting weak classifiers and improving accuracy. On the other hand, the ANN model is versatile for modelling non-linear relationships while SGD is efficient in large-scale and sparse data learning [38]. Each of these models presents varying advantages and limitations, which is why they are used in the study.

2.2.2. Model optimisation and comparison

While some algorithms, like RF and AdaBoost, are well known for their robust out-of-the-box performance, others, like ANN, require careful hyperparameter tuning to maximize their performance [39]. In the past, ML models were often fine-tuned manually through trial and error. However, more sophisticated approaches have emerged for identifying the best hyperparameter combinations. [40]. This study utilized a search algorithm to fine-tune each model type. A parameter grid was formulated for each model type, comprising the optimized hyperparameters and the range of options or values each parameter could assume. Subsequently, the algorithm explored the parameter space to identify the most efficient hyperparameter combination.

2.2.3. Model evaluation metrics

Regression coefficient (R^2) and root-mean-square error (RMSE) are often used as regression analysis evaluation indexes. To evaluate the prediction performance of the ML models, three evaluation indicators, including mean absolute error (MAE), root-mean-square error (RMSE), and R^2 are used. These indicators are meticulously described in equations (2), (3), and (4).

$$MAE = \frac{\sum_{i=1}^{N}\left|y_i^{exp} - y_i^{pred}\right|}{N} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(y_i^{exp} - Y_i^{exp}\right)^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{pred}\right)^2}{N}} \quad (4)$$

Where y_i^{exp} and y_i^{pred} are the actual and predicted values, respectively. Y_i^{exp} is the mean of the actual values [41][42].

Mean Absolute Error (MAE) measures how far the predictions from a model are from the actual values. MAE is computed by determining the absolute discrepancies between the predicted and true values for every individual point in the dataset and then calculating the mean of these differences. The MAE is expressed in the same units as the original data, which makes it easy to interpret.

MAE is scale-dependent, indicating that its value increases with the magnitude of the target variable in comparison to the predictors. To mitigate this, normalizing the data prior to model training is advisable. On the contrary, the RMSE gauges the spread of errors around zero, with a higher sensitivity to outlier values. The R^2 assesses how well the model fits the observed data. MAE is favored in regression analysis due to its straightforward interpretability and its robustness to outliers, as it treats all deviations equally regardless of their extent.

The MAE, RMSE, and R^2 were calculated for test sets and cross-

validated models. The data was divided into a training and test set, with an 84 % to 16 % split for model development. Cross-validation is another method for determining a model's generalization capability. This study used k-fold cross-validation, which divides the dataset into k parts. The model was then trained on k-1 folds and tested on the final fold. This process was then repeated k times to test each fold once. Cross-validation provides an independent measure of model performance while using all available data for model development. In this case, 5-fold cross-validation was used.

2.3. Interpretability analysis

Interpretability and explainability of ML models involve uncovering patterns in data and understanding the relationships derived from these models [43]. Interpretability refers to the capacity of a human to grasp why a ML model makes a certain prediction [44]. Explainability, a closely related concept, pertains to the capability of offering a clear and comprehensible explanation for a model's decisions [45]. A model that is highly interpretable is straightforward to comprehend, and its results can be consistently anticipated by a human.

Many ML models are often considered black boxes due to a common lack of understanding of their internal mechanics. To address this issue, one approach is to use models that are inherently more interpretable [46]. A more contemporary strategy involves model-agnostic methods, which are developed to interpret any black box model. An example of this type of method is permutation feature importance. This is a simple technique where the values of a feature in the dataset are randomly shuffled [47].

Another approach is the global method such as the Gini index, also known as Gini impurity. Global methods refer to interpretation techniques that provide an overall understanding of the model's behavior, rather than explanations for individual predictions. Global methods can be model-specific (like the Gini index for tree-based models) or model-agnostic (like the aforementioned permutation feature importance) [34]. One drawback of employing the Gini index for evaluating the significance of variables is its inclination towards favoring inputs with a greater number of categories. This bias is less significant when dealing with predominantly continuous features with minimal correlations. Additionally, it's worth noting that this approach does not apply to other model types [34].

Global methods provide insights into the model's overall behavior but fall short of clarifying specific predictions. To address this, local surrogate models such as LIME and SHAP are employed. LIME approximates an interpretable model locally. These local approaches focus on elucidating singular predictions without the necessity of accurately reflecting the global model. Despite LIME's potential, it encounters several challenges, particularly with structured data. Adjusting its hyperparameters can be complex, with the outcomes often varying based on the chosen width of the smoothing kernel. Additionally, LIME's explanations can be inconsistent, leading to considerable variation upon repetition of the same explanation process [48,49].

Permutation feature importance is another technique that can be utilized to interpret a ML model [50]. This method shuffles the values of each feature one by one, observing the effect on the model's accuracy. Each feature undergoes this process, and the one causing the largest reduction in accuracy upon shuffling is deemed the most crucial. Permutation feature importance serves both as a global and a local interpretative method for ML models. In its global application, the importance of a feature is determined by taking the average of its impact on accuracy across the entire dataset. This global perspective presumes uniform feature importance across all data points, providing a general view of which features are most influential.

When used locally, each sample's feature importance is computed separately by only permuting the feature on the sample and measuring the effect on the prediction. This approach gives a more detailed view of feature importance and can show how feature importance varies across samples. Local feature importance can be more informative in some cases where feature importance varies between samples.

SHAP is also a useful method that can be used for both global interpretation and individual prediction explanation [51]. It has a solid theoretical foundation in game theory and estimates the importance of features by allocating optimal credits based on Shapley values. SHAP force plots visually represent how various features influence an individual prediction. SHAP has one advantage over the Gini and permutation feature importance assessment methods for global interpretation in that it provides information about the importance of features and their relationship with the output. Furthermore, SHAP's predictions are fairly distributed across feature values. These factors are critical in ensuring trust in the method.

Permutation and SHAP feature importance assessment were used in this work to provide a global interpretation of the developed models. A partial dependence plot (PDP) will be used to understand the local interpretation of the model. A PDP shows the relationship between a feature and the predicted outcome of a model while holding other features constant. It provides insight into how a specific feature affects the model's predictions for a specific range of values and can help identify non-linear relationships or interactions between features. Because it focuses on the relationship between a single feature and the model's predictions, it can be considered a local method instead of a global method, which would show the relationship between all features and the predictions [52]. The analysis results were graphically represented for interpretation. Global methods were illustrated using boxplots for permutation feature importance and bar plots for SHAP by displaying the importance scores for all model outputs or targets in the different figures. This maximizes the amount of information contained in each figure.

3. Results and discussion

3.1. Dataset description and statistical analysis

A quantified the statistical characteristics of the cleaned data which include both input (Cel, Hem, Lig, Vm%, Ash%, FC%, C-%, H-%, O-%, N-%, Size, HR, PT, Temp) and output (H/C, O/C, Oil_Yield%, Gas_Yield%, Char_Yield%) variables were presented in Tables 2 and 3. The compositions of carbon (C), hydrogen (H), oxygen (O), and nitrogen (N) in biomass typically range from 0.99 to 79.17 %, 0.90–11.34 %, 0.004–74.04 %, and 0.08–18.29 %, in that order. Moreover, the concentrations of C, H, and O are relatively high, with average percentages being 47.31 %, 6.03 %, and 38.84 %, respectively, while the N content is relatively lower in comparison. The structural components of biomass, including cellulose, hemicellulose, and lignin, vary between 8.00 % and 69.00 %, 4.00 % to 79.5 %, and 2.70 % to 79 %, respectively. Additionally, the FC, VM, and ash contents of the biomass dataset range from 0.55 % to 82.47 %, 1.15 % to 97.25 %, and 0.11 % to 86.84 %, respectively, with average contents of FC, VM, and ash at approximately 18.78 %, 75.35 %, and 5.87 %, respectively. The variation in the ranges of composition analysis could be due to the wide variety of biomass sources as well as the heterogeneity of individual biomass. It should be mentioned that the pyrolysis conditions including the PS, HT and RT cover a broad spectrum, ranging from 0.08 to 42 mm for PS, 1.9 to 500 °C/min for HT, and 100 to 800 °C for PT.

The linear correlation between any two variables, was assessed using the SCC and the result is depicted using a heatmap (refer to Fig. 3). An SCC value close to 0 indicates a weak correlation between variables, whereas an SCC value near $\pm$ 1 indicates a very strong correlation. Notably, among the input parameters, strong correlations (SCC $\geq$ 0.3) were noted for several pairs, such as FC versus VM with an SCC of -0.68, C versus ash with an SCC of -0.34, VM versus ash with an SCC of -0.53, and C versus O content with an SCC of -0.6. These findings are consistent with other studies reported in the literature that have identified correlations between the proximate and ultimate compositions of

Table 2
Statistical summary of input features used in the ML models.

Statistics criterion	Cel	Hem	Lig	Vm%	Ash%	FC%	C-%	H-%	O-%	N-%	Size	HR	PT	Temp
count	262	262	262	262	262	262	262	262	262	262	262	262	262	262
mean	37.15	27.81	23.74	75.3	5.87	18.78	47.3	6.03	38.84	1.95	1.47	37.87	506	44.38
std	7.82	8.346	8.74	14.9	6.89	13.45	8.69	1.51	8.861	2.32	2.92	71.57	93.4	50.7
min	8	4	2.7	1.15	0.11	0.55	0.99	0.9	0.004	0	0.08	1.9	100	10
25 %	37.15	27.33	22.68	73.8	1.75	13.75	43.5	5.38	35.66	0.42	0.5	7	452	30
50 %	37.15	27.81	23.74	78.8	4.98	16.39	47.9	5.92	40.58	1.08	0.95	20	500	44.38
75 %	37.15	27.81	23.74	82.3	7.66	18.64	50.6	6.59	44.32	2.67	1.4	30	550	44.38
max	69	79.5	79	97.3	86.8	82.47	79.2	11.3	74.04	18.3	42	500	800	550

Table 3
Statistical summary of the target features used in the ML model.

Statistics criterion	H/C	O/C	Oil yield (wt%)	Gas yield (wt%)	Char yield (wt%)
count	262	262	262	262	262
mean	3.997	0.3686	27.69	32.23	40.09
std	2.319	0.4954	6.435	6.870	10.05
min	0.2	0.06	0	2.721	2.160
25 %	2.268	0.1	25.33	28.78	35.19
50 %	3.998	0.19	28.45	33.19	38.77
75 %	4.527	0.3686	30.76	36.02	42.29
max	17.5	3.57	60.46	47.13	86.84

biomass feedstock [2119]. Furthermore, the SCC values between inputs and outputs highlight the influence of an input variable on the prediction of outputs. For example, Fig. 3 demonstrates that the input variable PT has a significant impact on the gas yield, with an SCC of 0.71.

The positive SCC value indicates that an increase in PT leads to higher gas yield. Conversely, Fig. 3 shows a strong negative correlation between PT and biochar yield (SCC = -0.75), suggesting that higher PT reduces biochar yield. Additionally, a higher C content in biomass correlates with increased biochar yield (SCC = 0.55). Fig. 3 also highlights significant correlations between various output-input pairs, such as an SCC of 0.78 for char yield versus O content and 0.71 for gas yield versus PT. These correlations reveal that PT and O notably affect outputs like oil, char, and gas yields. The ratios of hydrogen to carbon (H/C) and oxygen to carbon (O/C) are influenced by factors like lignin content (Lig) for H/C, and temperature, cellulose content (Cel), hemicellulose content (Hem), and N content for O/C. Hence, uncertainties in these key factors can greatly impact the predictive accuracy of the data-driven model.

Such uncertainties in the ML model, arising from data quality, model complexity, and algorithmic randomness, can lead to reduced accuracy.

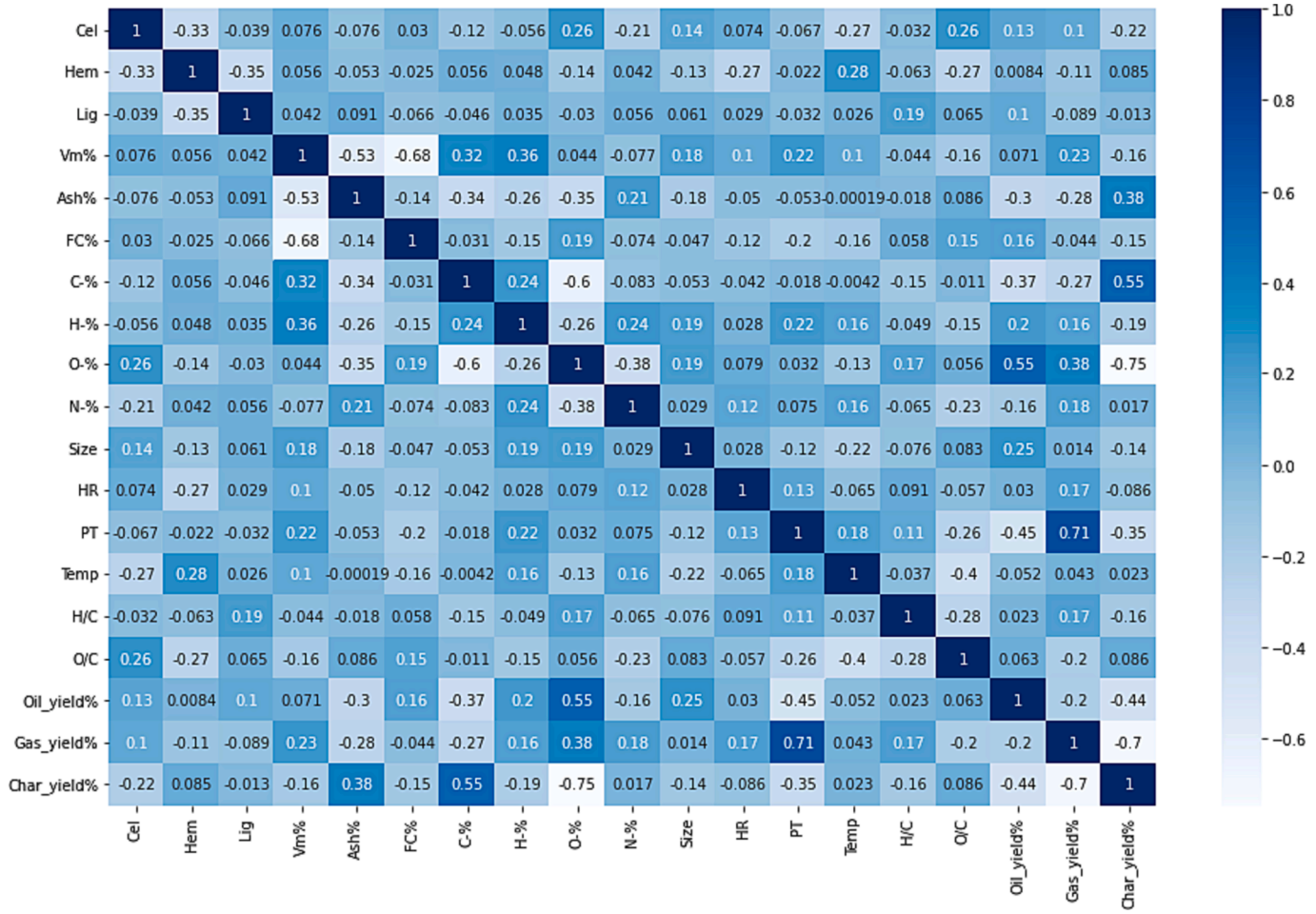


Fig. 3. Spearman correlation coefficient between any two variables of interest.

To mitigate data uncertainty, a robust data preprocessing strategy was employed in this study. Additionally, the heatmap in Figure S1 of the supplementary materials, which uses significance levels or p-values, indicates the strength of these correlations, with higher p-values signifying stronger correlations.

3.2. Selection of optimal preprocessing steps

This section outlines results from data scaling and outlier removal as well as the performance of hyperparameter optimization and various ML models.

3.2.1. Scaling of data

Data preparation and preprocessing are essential steps in the creation of ML models, leading to the development of two principal methods for transforming data: normalization and standardization. When evaluating these methods in terms of their impact on model performance, it was found that models performed better when utilizing normalized data compared to standardized data. Consequently, normalized data was chosen for the construction of the ML model.

3.2.2. Outliers removal

Following the data preprocessing step, which involved removing columns with a significant amount of missing data, 262 data points were retained. An outlier threshold was established using Sklearn, and based on this criterion, 17 data points from the cleaned dataset were excluded, as shown in Fig. 4. The red dashed line likely represents the threshold for outlier removal. Observations to the right of this line are considered normal, while those to the left are identified as potential outliers. The majority of the data clusters near zero, indicating a skew towards higher normality values. After the removal of outliers, the performance of each model was evaluated, with the results presented in Tables 4 and 5. It is important to note that the elimination of outliers is undertaken to enhance the accuracy of ML predictions.

3.2.3. Model optimization and comparison of model performance

The performances of both optimized and out-of-the-box models, along with the best preprocessing steps including normalization and outlier removal, were compared, and presented in Tables 4 and 5. It should be mentioned that evaluation metrics in the table are calculated for all output variables collectively. This means the reported values are aggregate measures of the model's performance across all predicted variables. The table indicate that out-of-the-box models for RF, XGBoost, ANN, and GBR performed optimally without needing hyperparameter optimization. In contrast, the SGD and AdaBoost models showed improved performance after hyperparameter optimization compared to their out-of-the-box counterparts.

For example, the AdaBoost model's R^2 increased from 0.81 to 0.90 because of hyperparameter optimization; however, the R^2 of the GBR model marginally declined from 0.97 to 0.93. Additionally, the R^2 of the SGD model increased from −0.01 to 0.50. In general, optimization was demonstrated to have a negligible impact on ensemble and tree-based models. However, hyperparameter optimization can efficiently get the optimum model structure when the computing demands during model training are not a concern. Hyperparameter optimization is a process of fine-tuning the settings used in ML algorithms to obtain the best performance. When computing demands during model training are not a concern, this optimization can efficiently find the optimum model structure for several reasons. It is easier to explore a wide range of hyperparameter values without computing demand bottlenecks. This includes performing exhaustive searches like grid search, which systematically works through multiple combinations of parameter values, or randomized search, which samples a large number of parameter combinations. Furthermore, higher computational capacity allows for experimenting with more complex models, which might have numerous hyperparameters. Complex models often have the potential to achieve higher performance, provided they are well-tuned. Moreover, with the availability of computational resources, there is a flexibility to iteratively refine hyperparameters. Algorithms like Bayesian optimization can benefit from this, as they iteratively update the understanding of the parameter space to find the best values. It should be mentioned that the contour plots showing the hyperparameters selected for optimization, their corresponding parameter ranges, and the resulting optimized hyperparameter combinations have been included in figures S3 – S8 of the supplementary materials.

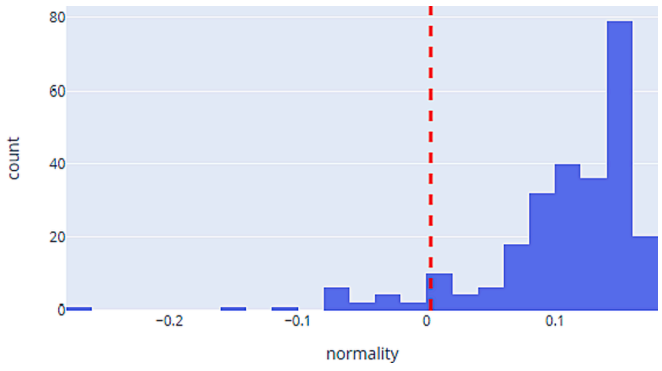


Fig. 4. Diagrammatic representation of the outlier removal threshold.

The RMSLE performance for each model was visualized using a boxplot, as shown in Fig. 5. RMSLE was chosen as the metric for hyperparameter optimization because it disproportionately penalizes larger differences between predicted and actual values, especially when the predicted values are much higher than the actual ones. This attribute is particularly beneficial in datasets with a few extreme outliers, as it helps to prevent the model from being excessively influenced by these outliers. Additionally, RMSLE offers improved interpretability since it is in the same unit as the target variable, thereby simplifying the understanding of error magnitude.

The results presented in Tables 4 and 5 indicate that the GBR model exhibited the most superior predictive performance. Notably, the GBR model with its default parameters outperformed the version with optimized parameters, leading to a preference for the default GBR model. With mean R^2 scores exceeding 0.80 for the training data, the RF, XGBoost, and AdaBoost models were deemed to perform satisfactorily. Despite the ANN models showing promise in previous studies [40], they fell short in this study. While the data set utilized in this study is still more significant than the ones used in prior works, ANNs require a lot of data; a potential explanation for this is the comparatively small quantity of data available to train the model [53,54,55]. The comparison of the GBR model's predictions for each output (H/C, O/C, gas yield, char yield, and biochar yield) fitted from the training and testing datasets are shown in Fig. 6.

It should be mentioned that the 12 % deviation line in the figure indicates the boundary within which the predicted values are considered to be within 12 % of the actual measured values. This line serves as a visual guide to assess the accuracy of the model's predictions. The data points that fall within the area between the 12 % deviation lines are within an acceptable range of error, while points outside of this area deviate from the actual values by more than 12 %. This helps to quickly identify the model's predictive performance and the proportion of predictions that significantly deviate from the true values. Moreover, 12 % was set as a benchmark based on a meticulous literature search.

More cluster points at the 45-degree line suggested that the GBR model performed optimally. The GBR model outperformed the other developed ML models in the training dataset, with MAE, RMSE, and R^2 having values of 0.011, 0.016, and 0.97, respectively. The training performance of the XGBoost model was the second-best, and the SGD and SVM models had the worst performance. In the testing dataset, among the developed ML models, the GBR and XGBoost models gave

Table 4
Performance of all out-of-box models (i.e., after removing outliers without optimization).

Models	R^2_train	R^2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
SVM	0.57	0.5269	0.0678	0.0694	0.0521	0.0552
RF	0.93	0.7588	0.0267	0.0499	0.016	0.0326
XGB	0.96	0.7972	0.0189	0.0446	0.0129	0.03
ANN	0.96	0.6011	0.0193	0.0598	0.0123	0.042
GBR	0.97	0.809	0.0166	0.043	0.0113	0.0286
ADA	0.81	0.6904	0.0461	0.056	0.0375	0.043
SGD	−0.01	−0.083	0.103	0.1005	0.0724	0.0691

Table 5
Performance of all models after removing outliers and optimization.

Models	R^2_train	R^2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
RF	0.86	0.7265	0.0388	0.0535	0.0234	0.037
XGB	0.92	0.7481	0.0292	0.0504	0.0186	0.0338
ANN	0.58	0.5939	0.0653	0.0681	0.0456	0.046
GBR	0.93	0.7908	0.022	0.0465	0.0045	0.0279
ADA	0.90	0.7381	0.0301	0.0516	0.0224	0.0373
SGD	0.50	0.4722	0.0714	0.0779	0.0433	0.0456

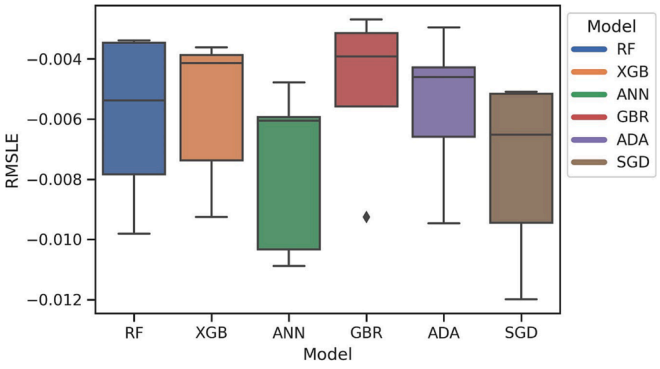


Fig. 5. RMSLE comparison for each model after hyperparameter optimization.

better testing performances than other models, with the MAE, RMSE, and R^2 having values of 0.0281, 0.0430, 0.80, and 0.0286, 0.0428, and 0.81, respectively.

To further evaluate the robustness of the GBR model, Fig. 7 shows the sample relative error distribution of each output. The relative error is the absolute difference between the simulated and experimental values divided by the experimental value. From the overall relative error distribution of the GBR model for each output, the relative error of the oil yield, gas yield, and biochar yield was the smallest, followed by H/C. The relative error for O/C was the highest. The sample relative error distribution range of the GBR model in the training and testing datasets for the oil yield, gas yield, and biochar yield was less than 15 %, and most of the samples were within 2 %. While the relative error distribution ranges of the other outputs (bio-oil H/C and O/C) varied significantly, most of the samples were less than 100 %. The GBR model excelled due to its adeptness at managing diverse data types and numerous features, capturing non-linear relationships and interactions among variables effectively. Its incremental improvement approach—focusing on areas of poor performance—contributes to enhanced

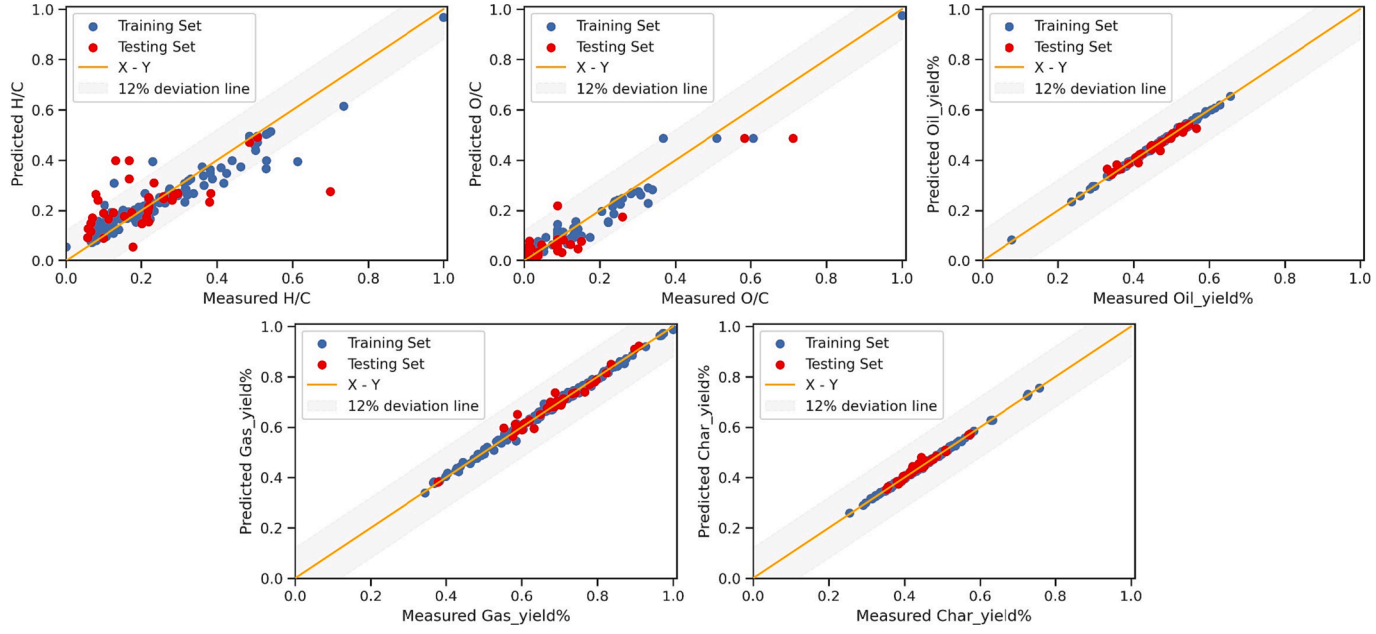


Fig. 6. Comparison of the predictions of the GBR model for each output fitted from the training and testing set.

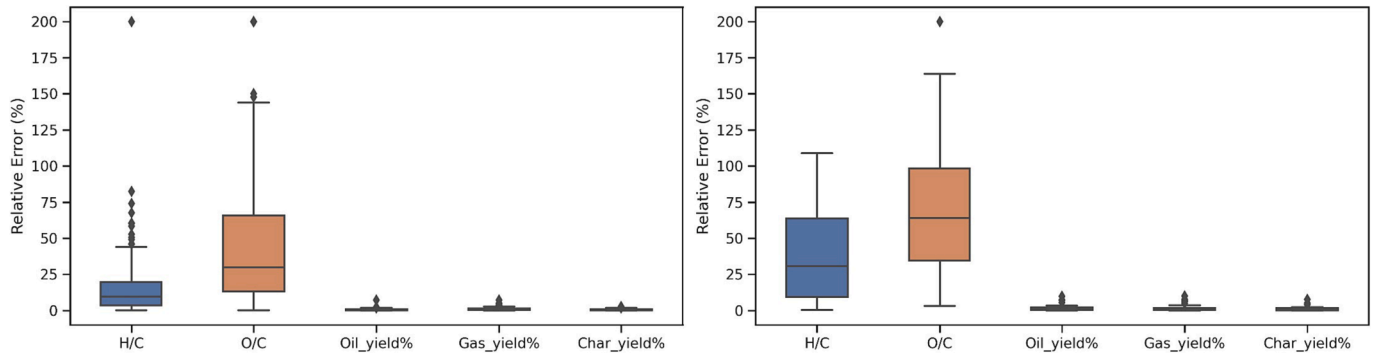


Fig. 7. Relative error distribution of the GBR model for each output in the training and testing set.

overall accuracy. Consequently, this method yields a more robust training process, reducing the risk of overfitting and minimizing the impact of noisy data.

3.2.4. *Feature importance analysis*

Feature selection stands as a vital step in the preprocessing phase of developing ML models. It plays a key role in pinpointing critical Fig. 8 evaluates the relative importance of the input features for the model that delivers optimal performance. This evaluation is performed in the context of predicting ratios such as H/C and O/C in bio-oil, as well as gas yield, oil yield, and biochar yield, employing the feature importance methodology of the GBR model.

The impurity-based feature importance used by the GBR can assign high significance to features that may not actually predict the target variable, particularly if those features contribute to overfitting. Moreover, there appears to be no direct correlation between the feature importance as determined by the GBR model and the linear correlations of features to targets for H/C and O/C ratios, as described by the Spearman analysis. The GBR model indicates that the composition of the feedstocks exerts the strongest influence. Specifically, the H/C and O/C characteristics of the bio-oil are more heavily dependent on the results from ultimate analysis and the chemical composition of the biomass materials rather than the pyrolysis operating conditions. It should be mentioned that the outcomes of SCCs and the GBR model are inconsistent with the pyrolysis conditions. This is because impurity-based feature importance suffers from being computed from the training dataset. Therefore, it is necessary to remove correlated or collinear data for the following reasons:

- Overfitting, which happens when a model is too complex and has too many degrees of freedom, can be brought on by collinear data. On new, untested data, overfitting might result in poor generalization performance.
- Correlated data can cause multicollinearity in linear regression models. This happens when two or more independent variables exhibit strong correlations, which might make it challenging to ascertain the proper relationship between the independent and dependent variables.
- Correlated data may complicate the model's interpretation. For instance, it could be challenging to determine which variable is responsible for the relationship with the target variable if the two variables are highly correlated.

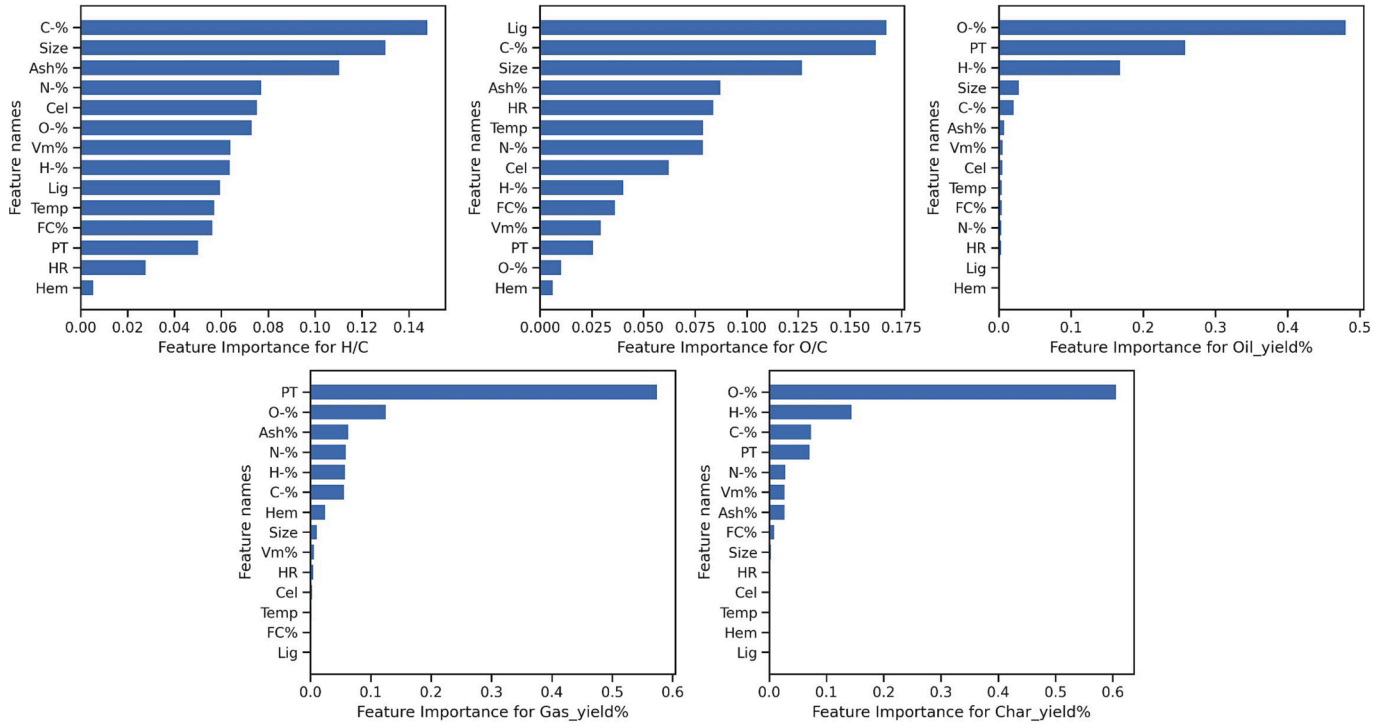


Fig. 8. Feature importance plots showing the influence of each feature on the output (target) variables.

- Correlated data can increase the computational cost and slow the model's training process.

A feature selection technique that uses hierarchical clustering to group similar features together was used alongside the SCC to remove the correlated data. By thresholding the linkage matrix formed, flat clusters are created. Clusters with the highest number of features are selected. Fig. 9a shows a hierarchical clustering diagram, a tree-like visual representation of a hierarchical clustering model, and a heat-map of the SCC (Fig. 9b).

The hierarchical clustering diagram shows the structure of the clusters and the relationships between them. The height of the branches represents the distance or similarity between the clusters, and the dendrogram can be cut at a certain height to form a specified number of clusters. The dendrogram displays the correlation between the input features. Each colour group has a minimum of one feature chosen. O, PT, H, N, Lig, Size, and Ash were selected as the features. The GBR model was then trained using these features, and its performance was evaluated using the testing dataset. The model's R^2 increased by 5 %, from 0.80 to 0.84.

The results indicate that a suitable combination of the PT, PS, proximate analysis (Ash), chemical components (Lignin), and element composition of biomass (O, H, and N, since there is a strong correlation between C and O) are necessary features for predicting the H/C, O/C, gas yield, oil yield, and biochar yield of a given biomass.

3.3. Interpretability analysis

The global interpretability of the GBR models was studied as it was found to be the best-performing model. Permutation feature importance was computed for the ML model using all the input features in Fig. 10. Additionally, the best features obtained from feature selection, using SCC and hierarchical clustering were selected for permutation feature importance in Fig. 11. The permutation feature importance scores were further visualized using the optimal features to understand the relationship between the features and each target variable. Although it received a high score, the C feature was excluded from the feature selection process due to its significant correlation with O, which was done to avoid issues with multicollinearity. Furthermore, C was later incorporated into the feature set, and the model was retrained to reinforce this decision. However, the resulting R^2 value dropped from 0.84 to 0.83 upon the inclusion of C.

The permutation feature important plot suggests that O exerted the greatest influence on the bio-oil H/C, O/C, gas yield, oil yield, and biochar yield. Following closely was the pyrolysis temperature, PT. The result aligns with experimental findings reported in the literature for biomass pyrolysis [18,19,21]. High oxygen content in biomass can lead to increased production of water and CO_2 during pyrolysis, which affects the yield and composition of bio-oil and gas. For example, studies show that the presence of oxygen influences the reactive species during thermochemical conversion, which in turn affects the H/C and O/C ratios in the resulting bio-oil, a crucial determinant of its quality and energy content [56,57]. Similarly, PT has been found to be a critical factor; higher temperatures generally increase gas yield and decrease biochar yield, as thermal decomposition intensifies, leading to more significant fragmentation of the biomass components [58]. This combination of high O content and optimal PT can therefore dictate the efficiency and output of the biomass conversion process.

An assessment of SHAP feature importance was conducted to gain deeper insights into the impact of each feature on the output. The relevance of each feature in predicting the respective target values for the biomass is illustrated in Fig. 12. From the diagram, PT and O had more of an influence on the model's ability to estimate gas yield than did the N content of the biomass. O had a greater influence on the model for char yield, while PT, H, and N had less impact. The model is not significantly affected by Ash, PS, or Lig. Furthermore, PT, O, and H played a significant role in predicting oil yield. Ash, Lig, and O had a more significant impact on the model's ability to predict H/C. Compared to the permutation feature importance, the SHAP feature importance result seemed more interpretable and offered a better insight into the impact of each feature on the model.

To gain a deeper understanding of how each feature influences a specific target variable, a partial dependence analysis was conducted. The optimal model's feature importance determined the relevance of the target variable, and partial correlation analysis was performed on the relevant feature variables related to bio-oil H/C and O/C ratio, gas yield, oil yield, and biochar production. The Partial Dependence Plot (PDP) analysis unveils how each feature impacts the target variables. In this approach, while keeping the remaining features at their mean values, one feature is adjusted to predict the target variables, which include H/

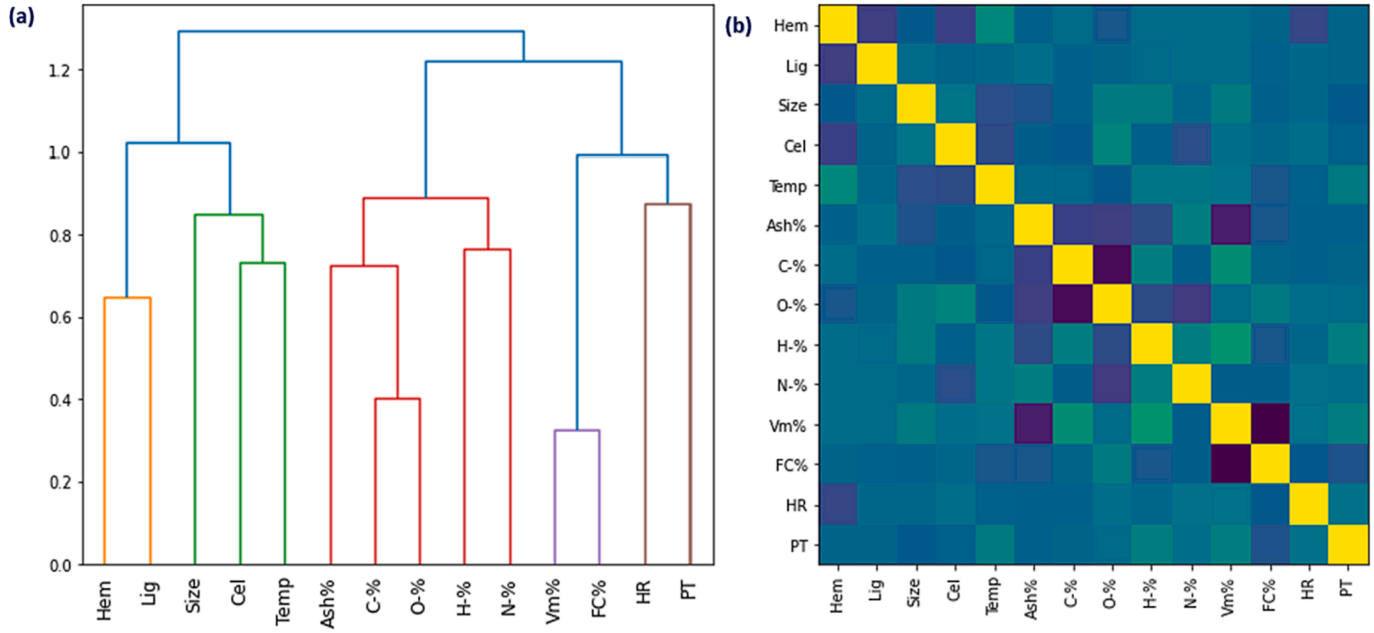


Fig. 9. Hierarchical clustering showing the correlation between input features. (a) hierarchical clustering diagram (b) heat map.

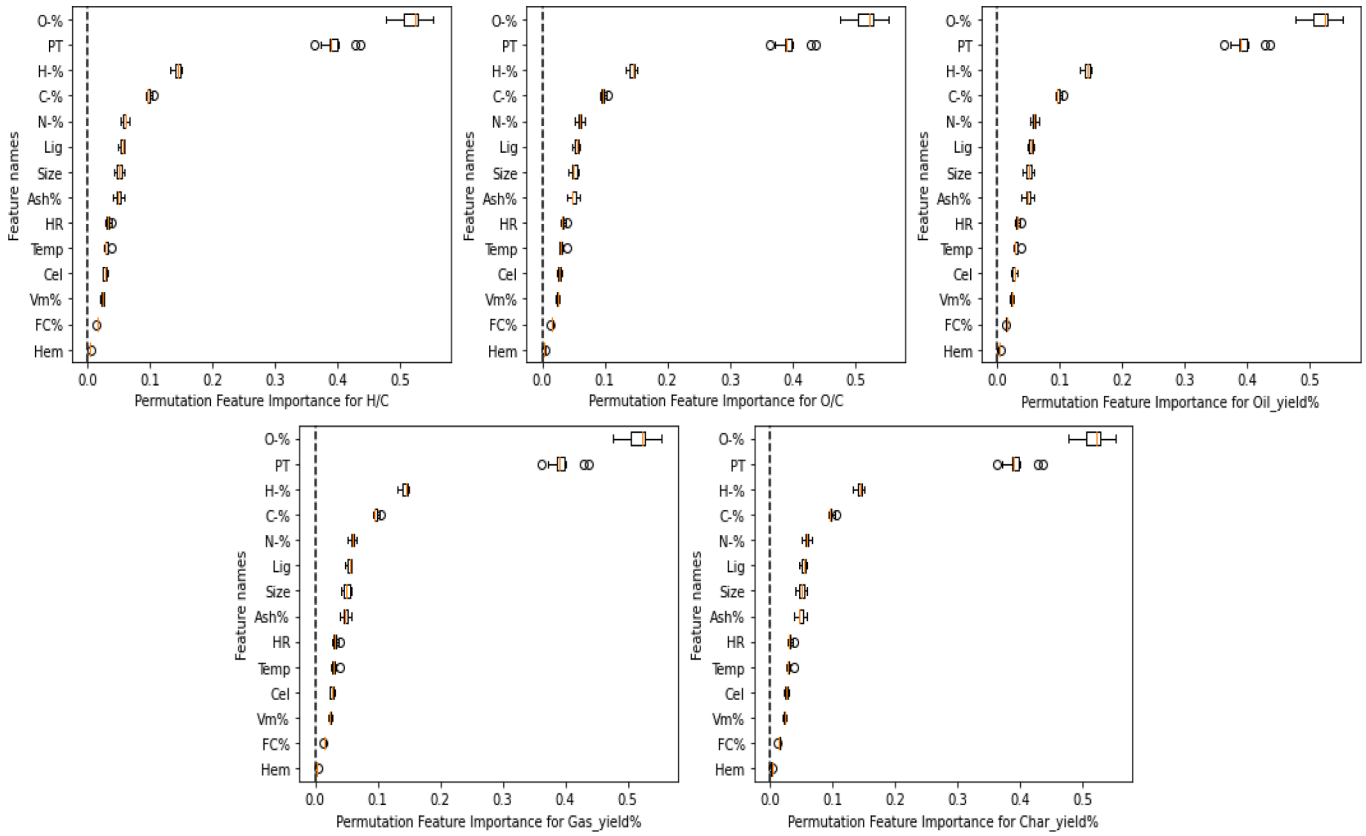


Fig. 10. Permutation feature importance without feature selection.

C, O/C, gas yield, oil yield, and biochar yield.

Figs. 13 - 17 illustrate the partial dependence of different feature variables on each respective target. From Fig. 13, the bio-oil H/C ratio remains unaffected by lignin values in biomass below 20; yet a modest uptick is noted when values range between 20 and 48, followed by a marked decline thereafter. Moreover, the H/C ratio in the bio-oil exhibits an increase as the biomass ash content rises within the 6–8 % bracket, only to taper off slightly beyond this range. The observed trends in the PDP regarding biomass composition and pyrolysis outcomes can be explained by the unique properties and reactions of lignin during the pyrolysis process. Lignin is a complex organic polymer found in plant cell walls that contributes to the structure and rigidity of plants. Its composition and the way it breaks down during pyrolysis can significantly influence the quality and yield of bio-oil [2]. For instance, it's known that lignin can produce higher amounts of bio-oil under certain conditions [59].

The content of lignin in the biomass and the pyrolysis temperature are both crucial. An increase in lignin content has been shown to affect the yield and composition of bio-oil. For example, a study demonstrated that increasing the lignin content in poplar from 17 % to 22 % at a pyrolysis temperature of 500 °C decreased the relative bio-oil yield and the yield of lignin-derived phenolic species [60].

Additionally, the thermal value and the C content of bio-oil are also impacted by the lignin content. Optimal yields of bio-oil with high C content and thermal value were obtained at a specific temperature, indicating that fewer aromatics in the liquid phase result in higher O/C and H/C ratios in the final products [61]. These findings suggest that the structure of lignin, the conditions under which pyrolysis occurs, and the specific methods used to process and upgrade the bio-oil all play a significant role in determining the final H/C and O/C ratios, as well as the overall yield of the pyrolysis products. The relationship between lignin content, pyrolysis conditions, and bio-oil characteristics is a nuanced interplay that has a profound impact on the efficiency and effectiveness of biomass conversion technologies.

Fig. 13 illustrates that the O content of biomass exerts the most pronounced influence on the H/C ratio in bio-oil. As the O content rises, there is a corresponding increase in the H/C ratio. This relationship can be attributed to the negative correlation between O and C contents; a higher O content typically means a lower carbon presence, leading to an elevated H/C ratio. H and N contents, on the other hand, did not show a significant effect on the H/C ratio. Additionally, the figure indicates that the H/C yield improves as the PS of the biomass exceeds 3 mm. Regarding the influence of PT, the H/C ratio remains unchanged at temperatures below 500 °C. However, an uptick in the H/C ratio is seen between 500 and 700 °C, after which it starts to decline. This suggests that only beyond a certain temperature threshold does the PT begin to affect the H/C ratio in bio-oil, reflecting the complex thermal dynamics of biomass pyrolysis.

The increase in the H/C ratio with a rise in biomass oxygen content is a consequence of the thermal decomposition of biomass constituents, where oxygen-rich components break down into water and carbon dioxide, thus enriching the H/C ratio. Larger biomass particles result in more incomplete carbonization, preserving more hydrogen. Temperatures between 500 °C and 700 °C facilitate the breakdown of biomass into simpler hydrocarbons, thus increasing the H/C ratio, while higher temperatures lead to more stable carbon structures and a lower H/C ratio [61].

The ratios of O/C in the bio-oil were not affected by the lower values of Lig below 15; however, an increase was observed for values in the range of 20–40 (Fig. 14). In addition, the O/C ratios increase significantly with ash content in the range of 8–14 %. No effect was observed for values less than 8 %. O has no significant impact on O/C. This suggests that while certain components of biomass, like lignin and ash at specific concentrations, contribute to the O/C ratio in the resulting bio-oil, the direct contribution of O content in the biomass does not alter this ratio significantly, pointing to complex interactions between the various

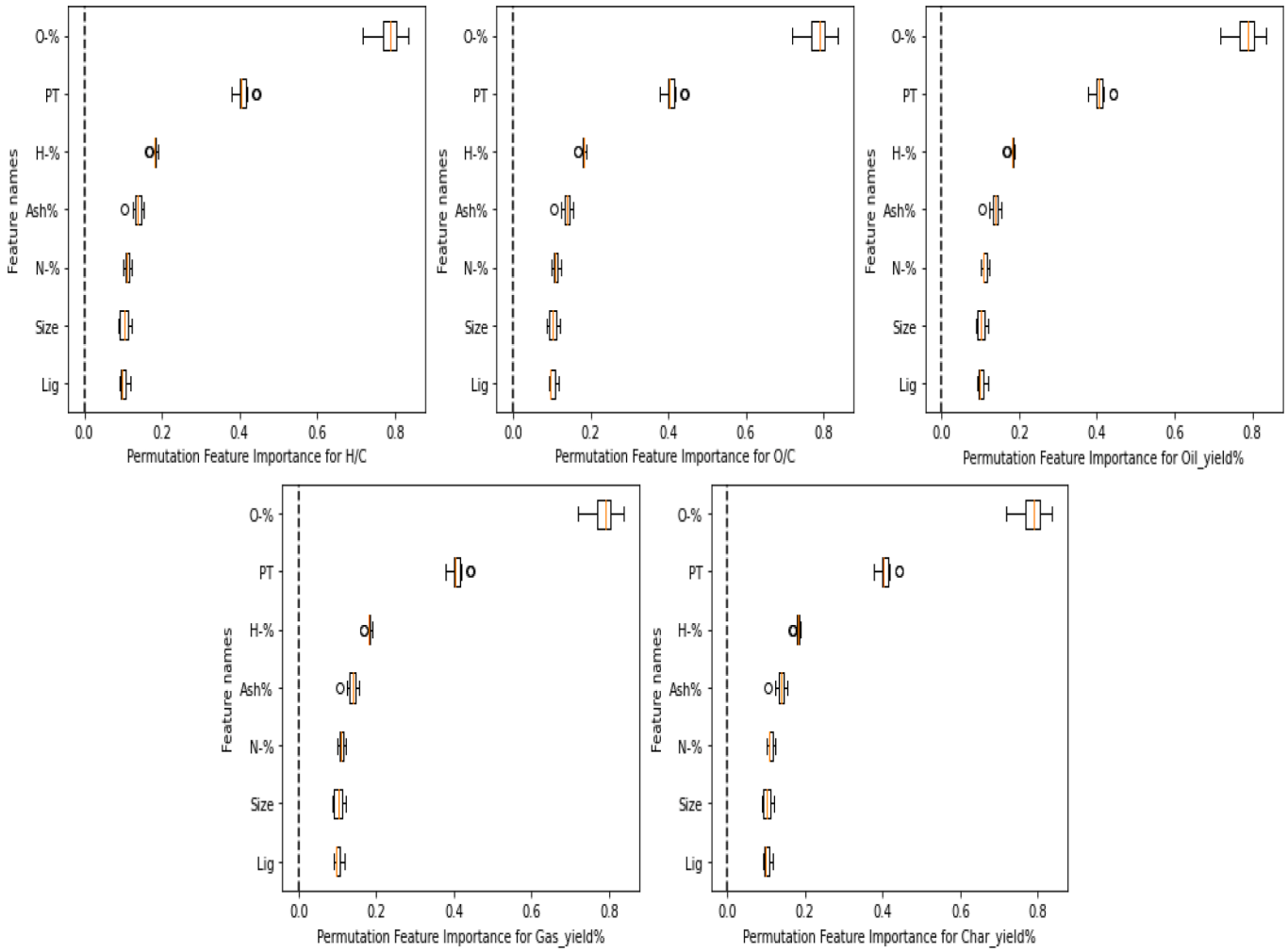


Fig. 11. Permutation feature importance after feature selection.

biomass constituents during the pyrolysis process.

The oil yield was not affected by Lig, ash, N, or the particle size of the biomass (Fig. 15). On the contrary, both the O and H content of biomass had a significant impact on the gas yield. As O and H increase, the oil yield also increases. The PT also significantly affected the oil yield. As PT increased above 500, a significant decrease in the oil yield was observed. An increase in oil yield was observed for PT values in the range of 400——500 °C. The trend where the oil yield increases with the O and H content in biomass but decreases above a PT of 500 °C can be explained by the chemical reactions occurring during pyrolysis. Oxygen and hydrogen present in the biomass contribute to the formation of volatile compounds during pyrolysis, which can enhance oil yield [59,61]. At optimal temperatures (400–500 °C), these volatiles are effectively converted into liquid bio-oil. However, at temperatures above 500 °C, further thermal decomposition leads to the cracking of these volatiles into non-condensable gases and char, thus reducing the oil yield [56].

This behavior aligns with the findings from literature wherein certain biomass constituents like lignin and ash do not directly influence oil yield, while the presence of oxygen and hydrogen is critical in determining the quantity and quality of bio-oil. Moreover, the thermal degradation of biomass components at varying temperatures has a significant effect on the distribution of pyrolysis products [62].

The gas yield was not affected by Lig, ash, or the PS of the biomass (Fig. 16). On the contrary, both O, PT, N, and H content significantly impacted the gas yield. As O, N, PT, and H increase, the gas yield also increases. Oxygen and nitrogen present in the biomass contribute to the formation of gas products through various pathways, including oxidative reactions and nitrogenous gas formation. The increase in PT facilitates thermal cracking, which breaks down biomass into simpler gas molecules, while hydrogen contributes to the formation of combustible gases. The combined increase in these elements and conditions enhances the decomposition and gasification of biomass, resulting in higher gas yields.

The pyrolysis temperature, PT, significantly affected the oil yield as shown in Fig. 16. As PT increased above 400 °C, a significant increase in the gas yield was observed. The PS of the biomass and the presence of Lig did not influence the char yield (Fig. 17). However, O, PT, N, and H content notably affected the gas yield. An increase in O, N, PT, and H content corresponded to a decrease in char yield. Notably, the pyrolysis temperature, PT, substantially impacted the oil yield, with a marked increase in gas yield observed when PT exceeded 400 °C. No observable effect was seen for PT values below 300 °C. For O levels below 30 %, there was no discernible impact on char yield, but for O contents exceeding 30 %, a significant reduction in char yield was noted. Furthermore, an increase in ash content resulted in a higher char yield.

To further enhance the implementation of the ML model among a wide variety of individuals an app was created in huggingface using Gradio. Details of the app are presented in figure S2 of the supplementary materials and the link to the app can be found in the Github repository (see data availability section).

4. Conclusions and prospects

This research employed a ML method to comprehensively predict the

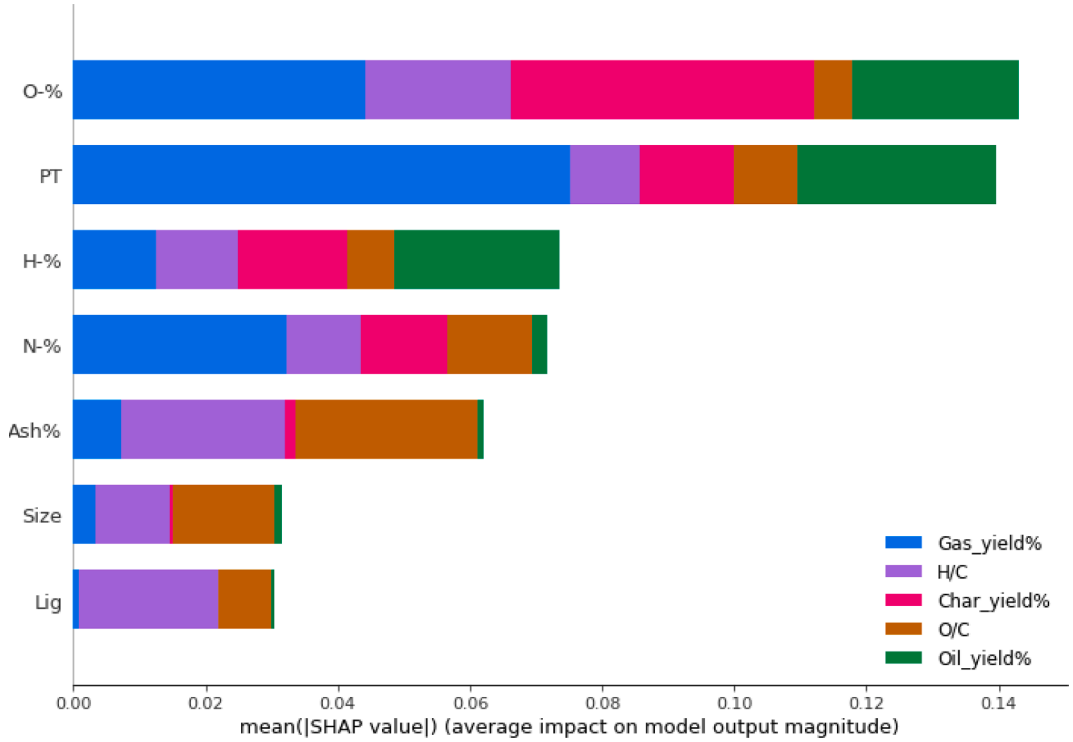


Fig. 12. A diagrammatic representation of the SHAP Importance Scores.

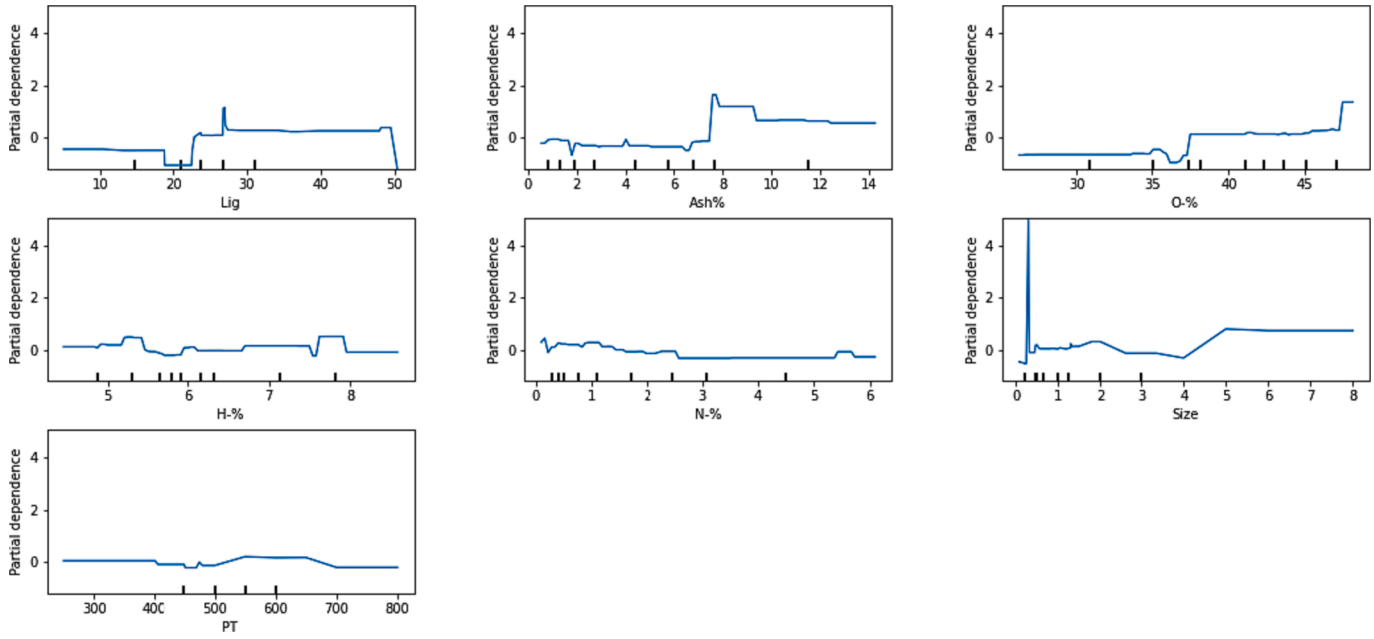


Fig. 13. Partial dependence plot for bio-oil H/C ratio.

yield of bio-oil, biochar, and gases produced by biomass pyrolysis and the H/C and O/C contents of bio-oil. The novel contributions lie in the integration of process modelling with ML to develop a comprehensive pyrolysis dataset for modelling complex input-out pyrolysis relationships. The GBR was identified as the most effective among various machine learning models. It accurately predicted yields of gas, biochar, and bio-oil, and their H/C and O/C compositions. GBR effectively demonstrated the complex relationships between these variables visually. The SHAP importance score analysis of the GBR model result revealed that the ultimate analysis data of the biomass feedstock and pyrolysis conditions had a significant impact on the bio-oil, biochar, and gas yield, while the chemical composition data of the biomass feedstock and the proximate analysis data had a greater influence on the O/C and H/C of bio-oil.

Based on the findings of this research, it is recommended that ML methods, such as the GBR, if trained on a huge dataset, be utilized in predicting the yield of bio-oil, biochar, and gas produced from biomass pyrolysis, along with their H/C and O/C compositions and other important characteristics like the calorific value and viscosity. The use of PDP and SHAP feature analysis can provide valuable insights into the

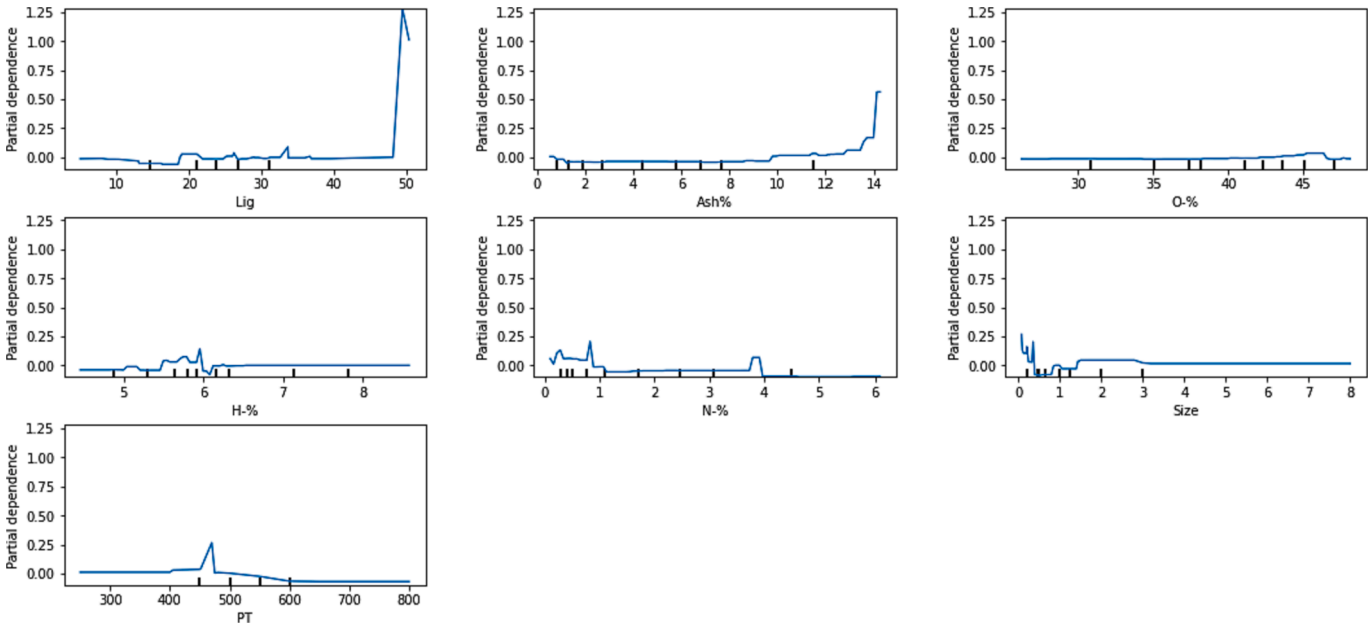


Fig. 14. Partial dependence plots for O/C.

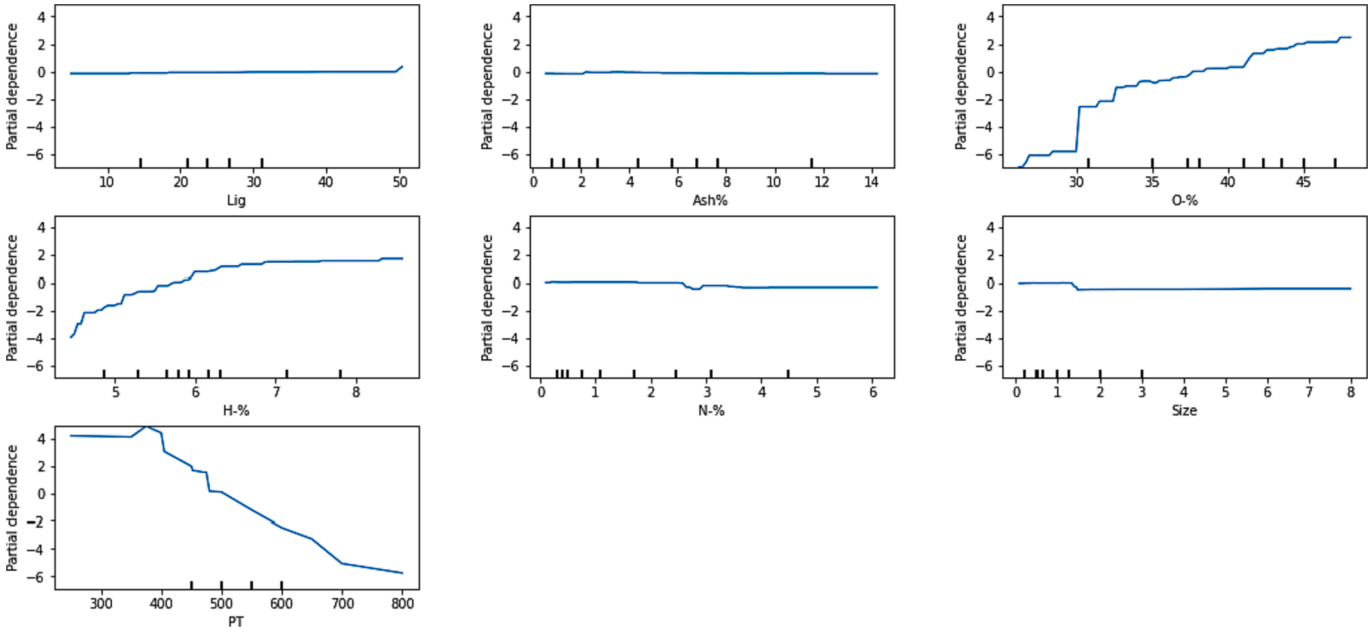


Fig. 15. Partial dependence plots for oil yield.

relationship between target and feature, including biomass composition and pyrolysis conditions. Therefore, the application of these methods in future research and industrial practice can lead to more accurate and efficient predictions of bioenergy yields, which can contribute to the development of sustainable energy production. Moreover, rapid prediction of pyrolysis yield and bio-oil properties plays a crucial role in decision-making related to the selection of feedstock or process conditions. This, in turn, contributes significantly to enhancing the economic and environmental assessment of biofuel production through pyrolysis.

While the study presents promising results in prediction, it is important to acknowledge several limitations. Firstly, most ML models require a substantial amount of data, which was not available in this study. Future research should consider evaluating the use of synthetic data generation methods, such as Generative Adversarial Networks (GAN) or Variational Autoencoders (VAE). These methods could be instrumental in augmenting the dataset and improving the model's predictive capabilities. A comparative analysis of these synthetic data generation techniques and their impact on the prediction accuracy of the model would be valuable.

Additionally, Physics-Informed Machine Learning (PIML) methods present another viable approach for modelling complex pyrolysis relationships. These methods integrate physical laws into the learning process, ensuring that the model not only learns from data but also adheres to the underlying mechanisms of pyrolysis. Future studies should explore the adoption of PIML methods to enhance the modelling of biomass pyrolysis, potentially leading to more accurate and reliable predictions.

Another notable limitation is the incorporation of biomass classes as

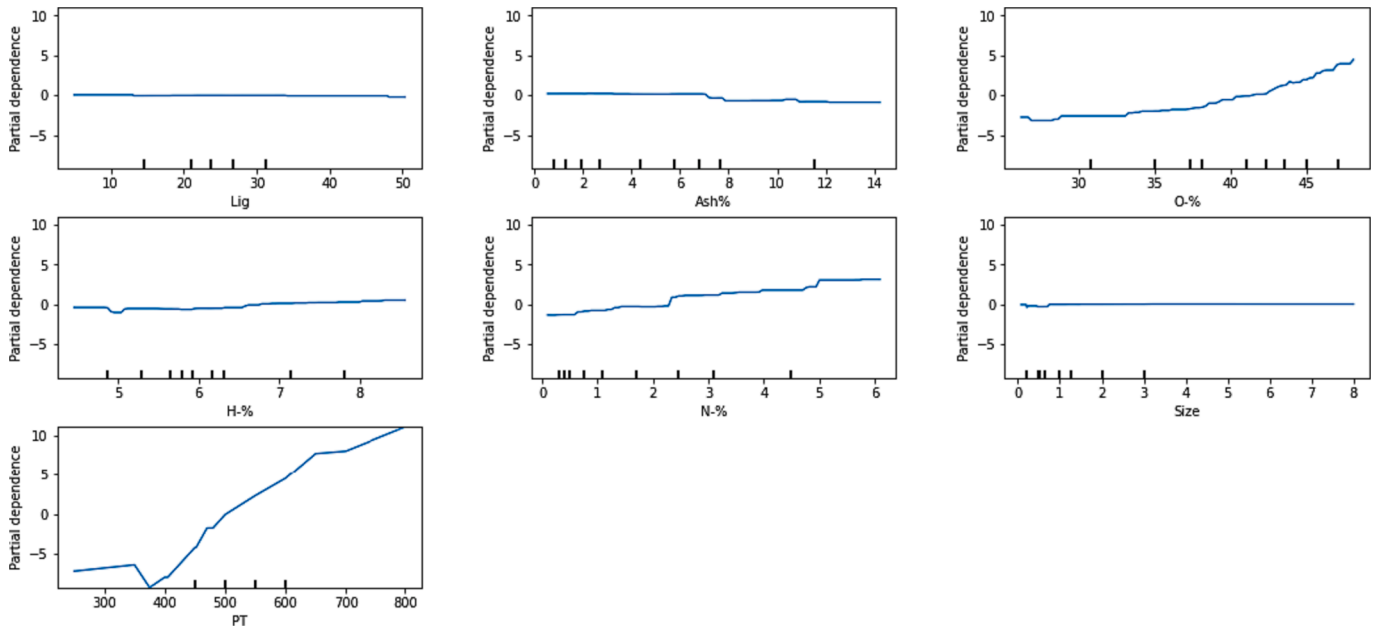


Fig. 16. Partial dependence plots for gas yield.

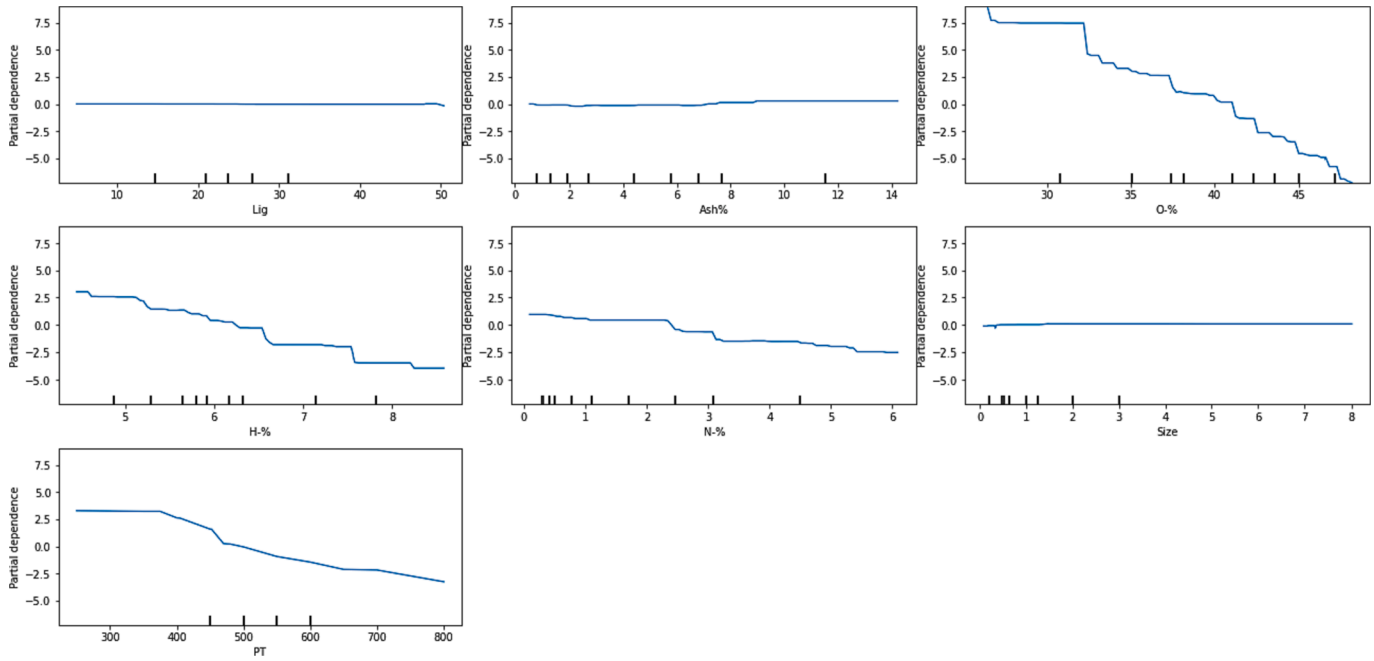


Fig. 17. Partial dependence plots for char yield.

part of the input features in the model. The selection of a diverse range of biomass classes, such as woody biomass, energy crops, and algae, could significantly improve the machine-learning model's performance. By incorporating a broader spectrum of biomass types, the model can learn from a more varied dataset, which could lead to better generalization and predictive accuracy in diverse pyrolysis scenarios. This expansion in biomass class variety would likely contribute to a more robust and versatile model, better suited for practical applications in biofuel production.

https://github.com/DouglasDivine/Thesis.

Note that the Aspen plus process simulation file is not included due to proprietary reasons however, it is available on request by the readers.

CRediT authorship contribution statement

Douglas Chinenye Divine: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Stell Hubert:** Supervision, Project administration, Funding acquisition, Data curation, Conceptualization. **Emmanuel I. Epelle:** Writing – original draft, Visualization, Validation. **Alaba U. Ojo:** Writing – original draft, Visualization, Validation. **Adekunle A. Adeleke:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision. **Chukwuma C. Ogbaga:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision. **Olugbenga Akande:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation. **Patrick U. Okoye:** Writing –

review & editing, Writing – original draft, Visualization, Validation, Supervision. **Adewale Giwa:** Writing – review & editing, Writing – original draft, Visualization, Validation. **Jude A. Okolie:** Writing – original draft, Supervision, Project administration, Data curation, Conceptualization.

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.

Data availability

The link to the code and dataset are in github repository in the data section of the manuscript.

Acknowledgement

The authors would like to thank everyone who contributed to the manuscript's success, including the editors and reviewers. The authors would like to appreciate the Petroleum Technology Development Fund (PTDF) Nigeria for providing support for Douglas Divine graduate study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fuel.2024.131346.

References

[1] Raja RB, Sarathi R, Vinu R. Selective Production of Hydrogen and Solid Carbon via Methane Pyrolysis Using a Swirl-Induced Point-Plane Non-thermal Plasma Reactor. Energy Fuel 2022;36:826–36. https://doi.org/10.1021/acs.energyfuels.1c03383.
[2] Okolie JA, Nanda S, Dalai AK, Kozinski JA. Chemistry and Specialty Industrial Applications of Lignocellulosic Biomass. Waste Biomass Valorization 2021;12:2145–69. https://doi.org/10.1007/S12649-020-01123-0/FIGURES/6.
[3] Bhaskar T, Pandey A. Advances in Thermochemical Conversion of Biomass-Introduction. Recent Advances in Thermochemical Conversion of Biomass 2015:3–30. https://doi.org/10.1016/B978-0-444-63289-0.00001-6.
[4] Okolie JA, Epelle EI, Tabat ME, Orivri U, Amenaghawon AN, Okoye PU, et al. Waste biomass valorization for the production of biofuels and value-added products: A comprehensive review of thermochemical, biological and integrated processes. Process Saf Environ Prot 2022;159:323–44. https://doi.org/10.1016/J.PSEP.2021.12.049.
[5] Shen Y, Jarboe L, Brown R, Wen Z. A thermochemical-biochemical hybrid processing of lignocellulosic biomass for producing fuels and chemicals. Biotechnol Adv 2015;33:1799–813. https://doi.org/10.1016/j.biotechadv.2015.10.006.
[6] Mohan D, Pittman CU, Steele PH. Pyrolysis of wood/biomass for bio-oil: A critical review. Energy Fuel 2006;20:848–89. https://doi.org/10.1021/ef0502397.
[7] Sorunmu Y, Billen P, Spatari S. A review of thermochemical upgrading of pyrolysis bio-oil: Techno-economic analysis, life cycle assessment, and technology readiness. GCB Bioenergy 2020;12:4–18. https://doi.org/10.1111/gcbb.12658.
[8] Tanger P, Field JL, Jahn CE, Defoort MW, Leach JE. Biomass for thermochemical conversion: targets and challenges. Front Plant Sci 2013;4:218. https://doi.org/10.3389/fpls.2013.00218.
[9] Sánchez-Borrego FJ, Barea de Hoyos-Limón TJ, García-Martín JF, Álvarez-Mateos P. Production of Bio-Oils and Biochars from Olive Stones: Application of Biochars to the Esterification of Oleic Acid. Plants 2022, Vol 11, Page 70 2021;11:70. https://doi.org/10.3390/PLANTS11010070.
[10] Akubo K, Nahil MA, Williams PT. Pyrolysis-catalytic steam reforming of agricultural biomass wastes and biomass components for production of hydrogen/syngas. J Energy Inst 2019;92:1987–96. https://doi.org/10.1016/j.joei.2018.10.013.
[11] 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.
[12] Wang C, Luo Z, Diao R, Zhu X. Study on the effect of condensing temperature of walnut shells pyrolysis vapors on the composition and properties of bio-oil. Bioresour Technol 2019;285:121370. https://doi.org/10.1016/J.BIORTECH.2019.121370.
[13] Trubetskaya A, Timko MT, Umeki K. Prediction of fast pyrolysis products yields using lignocellulosic compounds and ash contents. Appl Energy 2020;257:113897. https://doi.org/10.1016/J.APENERGY.2019.113897.
[14] Lv P, Chang J, Wang T, Wu C, Tsubaki N. A kinetic study on biomass fast catalytic pyrolysis. Energy Fuel 2004;18:1865–9. https://doi.org/10.1021/EF0400262/ASSET/IMAGES/LARGE/EF0400262F00005.JPEG.
[15] Wong KI, Wong PK, Cheung CS, Vong CM. Modelling of diesel engine performance using advanced machine learning methods under scarce and exponential data set. Appl Soft Comput 2013;13:4428–41. https://doi.org/10.1016/J.ASOC.2013.06.006.
[16] Aghbashlo M, Peng W, Tabatabaei M, Kalogirou SA, Soltanian S, Hosseinzadeh-Bandbafha H, et al. Machine learning technology in biodiesel research: A review. Prog Energy Combust Sci 2021;85:100904. https://doi.org/10.1016/J.PECS.2021.100904.
[17] Shafizadeh A, Shahbeig H, Nadian MH, Mobli H, Dowlati M, Gupta VK, et al. Machine learning predicts and optimizes hydrothermal liquefaction of biomass. Chem Eng J 2022;445:136579. https://doi.org/10.1016/J.CEJ.2022.136579.
[18] Ascher S, Wang X, Watson I, Sloan W, You S. Interpretable machine learning to model biomass and waste gasification. Bioresour Technol 2022;364:128062. https://doi.org/10.1016/J.BIORTECH.2022.128062.
[19] 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.
[20] Xing J, Luo K, Wang H, Fan J. Estimating biomass major chemical constituents from ultimate analysis using a random forest model. Bioresour Technol 2019;288:121541. https://doi.org/10.1016/J.BIORTECH.2019.121541.
[21] Khan M, Ullah Z, Mašek O, Raza Naqvi S, Khan NA, M.. Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms. Bioresour Technol 2022;355:127215. https://doi.org/10.1016/J.BIORTECH.2022.127215.
[22] Yang Y, Shahbeik H, Shafizadeh A, Masoudnia N, Rafiee S, Zhang Y, et al. Biomass microwave pyrolysis characterization by machine learning for sustainable rural biorefineries. Renew Energy 2022;201:70–86. https://doi.org/10.1016/J.RENENE.2022.11.028.
[23] Li H, Chen J, Zhang W, Zhan H, He C, Yang Z, et al. Machine-learning-aided thermochemical treatment of biomass: a review. Biofuel Res J 2023;10:1786–809. https://doi.org/10.18331/BRJ2023.10.1.4.
[24] Shahbeik H, Rafiee S, Shafizadeh A, Jeddi D, Jafary T, Lam SS, et al. Characterizing sludge pyrolysis by machine learning: Towards sustainable bioenergy production from wastes. Renew Energy 2022;199:1078–92. https://doi.org/10.1016/J.RENENE.2022.09.022.
[25] Wei H, Luo K, Xing J, Fan J. Predicting co-pyrolysis of coal and biomass using machine learning approaches. Fuel 2022;310:122248. https://doi.org/10.1016/J.FUEL.2021.122248.
[26] Akinpelu DA, Adekoya OA, Oladoye PO, Ogbaga CC, Okolie JA. Machine learning applications in biomass pyrolysis: From biorefinery to end-of-life product management. Digital Chemical Engineering 2023;8:100103. https://doi.org/10.1016/J.DCHE.2023.100103.
[27] Zhang T, Cao D, Feng X, Zhu J, Lu X, Mu L, et al. 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.
[28] Sluiter JB, Ruiz RO, Scarlata CJ, Sluiter AD, Templeton DW. Compositional analysis of lignocellulosic feedstocks. 1. Review and description of methods. J Agric Food Chem 2010;58:9043–53. https://doi.org/10.1021/JF1008023.
[29] Alvarez J, Kumagai S, Wu C, Yoshioka T, Bilbao J, Olazar M, et al. Hydrogen production from biomass and plastic mixtures by pyrolysis-gasification. Int J Hydrogen Energy 2014;39:10883–91. https://doi.org/10.1016/j.ijhydene.2014.04.189.
[30] Phyllis2. Phyllis2 - Database for the physico-chemical composition of (treated) lignocellulosic biomass, micro- and macroalgae, various feedstocks for biogas production and biochar. 2022 n.d. https://phyllis.nl/ (accessed March 31, 2023).
[31] Liu YC, Wang WC. Process design and evaluations for producing pyrolytic jet fuel. Biofuels Bioprod Biorefin 2020;14:249–64. https://doi.org/10.1002/BBB.2061.
[32] Okolie JA, Nanda S, Dalai AK, Kozinski JA. Hydrothermal gasification of soybean straw and flax straw for hydrogen-rich syngas production: Experimental and thermodynamic modeling. Energy Convers Manag 2020;208:112545. https://doi.org/10.1016/J.ENCONMAN.2020.112545.
[33] Ge Y, Tao J, Wang Z, Chen C, Mu L, Ruan H, et al. Modification of anaerobic digestion model No.1 with Machine learning models towards applicable and accurate simulation of biomass anaerobic digestion. Chemical Engineering Journal 2023;454:140369. https://doi.org/10.1016/j.cej.2022.140369.
[34] Wei P, Lu Z, Song J. Variable importance analysis: A comprehensive review. Reliab Eng Syst Saf 2015;142:399–432. https://doi.org/10.1016/j.ress.2015.05.018.
[35] Thara TDK, Prema PS, Xiong F. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognit Lett 2019;128:544–50. https://doi.org/10.1016/j.patrec.2019.10.029.
[36] Adhianto L, Banerjee S, Fagan M, Krentel M, Marin G, Mellor-Crummey J, et al. HPCTOOLKIT: Tools for performance analysis of optimized parallel programs. Concurr Comput 2010;22:685–701. https://doi.org/10.1002/cpe.
[37] Okolie JA. Can biomass structural composition be predicted from a small dataset using a hybrid deep learning approach? Ind Crops Prod 2023;203:117191. https://doi.org/10.1016/J.INDCROP.2023.117191.
[38] Haq ZU, Ullah H, Khan MNA, Naqvi SR, Ahsan M. Hydrogen production optimization from sewage sludge supercritical gasification process using machine learning methods integrated with genetic algorithm. Chem Eng Res Des 2022;184:614–26. https://doi.org/10.1016/J.CHERD.2022.06.020.
[39] Pannakkong W, Thiwa-Anont K, Singthong K, Parthanadee P, Buddhakulsomsiri J. Hyperparameter Tuning of Machine Learning Algorithms Using Response Surface Methodology: A Case Study of ANN, SVM, and DBN. Math Probl Eng 2022;2022:8513719. https://doi.org/10.1155/2022/8513719.

[40] Ascher S, Watson I, You S. Machine learning methods for modelling the gasification and pyrolysis of biomass and waste. Renew Sustain Energy Rev 2022;155:111902. https://doi.org/10.1016/j.rser.2021.111902.
[41] An G, Xing M, He B, Liao C, Huang X, Shang J, et al. Using machine learning for estimating rice chlorophyll content from in situ hyperspectral data. Remote Sens (Basel) 2020;12. https://doi.org/10.3390/RS12183104.
[42] Geng L, Che T, Ma M, Tan J, Wang H. Corn Biomass Estimation by Integrating Remote Sensing and Long-Term Observation Data Based on Machine Learning Techniques. Remote Sens (Basel) 2021;13. https://doi.org/10.3390/rs13122352.
[43] Alahy Ratul QE, Serra E, Cuzzocrea A. Evaluating Attribution Methods in Machine Learning Interpretability. IEEE International Conference on Big Data (Big Data) 2021;2021:5239–45. https://doi.org/10.1109/BigData52589.2021.9671501.
[44] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics (Switzerland) 2019;8:1–34. https://doi.org/10.3390/electronics8080832.
[45] Brito J, Proença H. A short survey on machine learning explainability: An application to periocular recognition. Electronics (Switzerland) 2021;10:1–11. https://doi.org/10.3390/electronics10151861.
[46] El Shawi R, Sherif Y, Al-Mallah M, Sakr S. Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); 2019. p. 275–80. https://doi.org/10.1109/CBMS.2019.00065.
[47] Tang Q, Chen Y, Yang H, Liu M, Xiao H, Wu Z, et al. Prediction of Bio-oil Yield and Hydrogen Contents Based on Machine Learning Method: Effect of Biomass Compositions and Pyrolysis Conditions. Energy Fuel 2020;34:11050–60. https://doi.org/10.1021/acs.energyfuels.0c01893.
[48] Gomes R, Denton A, Franzen D. Quantifying efficiency of sliding-window based aggregation technique by using predictive modeling on landform attributes derived from DEM and NDVI. ISPRS Int J Geoinf 2019;8. https://doi.org/10.3390/ijgi8040196.
[49] Wen X, Xie Y, Jiang L, Li Y, Ge T. On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development. Accid Anal Prev 2022;168:106617. https://doi.org/10.1016/j.aap.2022.106617.
[50] Sun X, Hao M, Wang Y, Wang Y, Li Z, Li Y. Reservoir Dynamic Interpretability for Time Series Prediction: A Permutation Entropy View. Entropy 2022;24:1–19. https://doi.org/10.3390/e24121709.
[51] Yang Y, Yuan Y, Han Z, Liu G. Interpretability analysis for thermal sensation machine learning models: An exploration based on the SHAP approach. Indoor Air 2022;32:1–24. https://doi.org/10.1111/ina.12984.
[52] Chan MC, Pai KC, Su SA, Wang MS, Wu CL, Chao WC. Explainable machine learning to predict long-term mortality in critically ill ventilated patients: a retrospective study in central Taiwan. BMC Med Inform Decis Mak 2022;22:1–11. https://doi.org/10.1186/s12911-022-01817-6.
[53] Baruah D, Baruah DC, Hazarika MK. Artificial neural network based modeling of biomass gasification in fixed bed downdraft gasifiers. Biomass Bioenergy 2017;98:264–71. https://doi.org/10.1016/j.biombioe.2017.01.029.
[54] Zhao S, Li J, Chen C, Yan B, Tao J, Chen G. Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. J Clean Prod 2021;316:128244. https://doi.org/10.1016/j.jclepro.2021.128244.
[55] Kardani N, Zhou A, Nazem M, Lin X. Modelling of municipal solid waste gasification using an optimised ensemble soft computing model. Fuel 2021;289:119903. https://doi.org/10.1016/j.fuel.2020.119903.
[56] Mei Y, Yang Q, Yang H, Li J, Zeng K, Chen Y, et al. Impact of cellulose deoxidization temperature on the composition of liquid products obtained by subsequent pyrolysis. Fuel Process Technol 2019;184:73–9. https://doi.org/10.1016/J.FUPROC.2018.11.003.
[57] Abu Bakar S, Ahmed A, Hussain M, Mo F, Ullah H, Zada N, et al. A Review on Catalytic Co-Pyrolysis of Biomass and Plastics Waste as a Thermochemical Conversion to Produce Valuable Products. Energies 2023, Vol 16, Page 5403 2023;16:5403. https://doi.org/10.3390/EN16145403.
[58] Chen D, Cen K, Zhuang X, Gan Z, Zhou J, Zhang Y, et al. 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. Combust Flame 2022;242:112142. https://doi.org/10.1016/J.COMBUSTFLAME.2022.112142.
[59] Fan L, Zhang Y, Liu S, Zhou N, Chen P, Cheng Y, et al. Bio-oil from fast pyrolysis of lignin: Effects of process and upgrading parameters. Bioresour Technol 2017;241:1118–26. https://doi.org/10.1016/J.BIORTECH.2017.05.129.
[60] Biswas B, Kumar A, Kaur R, Krishna BB, Bhaskar T. Co-hydrothermal Liquefaction of Lignin and Macroalgae: Effect of Process Parameters on Product Distribution. Bioenergy Res 2023;16:33–44. https://doi.org/10.1007/S12155-022-10437-X/TABLES/4.
[61] Klemetsrud B, Eatherton D, Shonnard D. Effects of Lignin Content and Temperature on the Properties of Hybrid Poplar Bio-Oil, Char, and Gas Obtained by Fast Pyrolysis. Energy Fuel 2017;31:2879–86. https://doi.org/10.1021/ACS.ENERGYFUELS.6B02836/SUPPL_FILE/EF6B02836_SI_001.PDF.
[62] Jahirul MI, Rasul MG, Chowdhury AA, Ashwath N, Jahirul MI, Rasul MG, et al. Biofuels Production through Biomass Pyrolysis —A Technological Review. Energies (Basel) 2012;5:1–50.