Process modelling integrated with interpretable machine learning analysis for predicting hydrogen yield and char production during chemical looping gasification

Arnold E. Sison[1], [1]Sydney A. Etchieson, Fatih Güleç [2,3], and Jude A. Okolie[4]

[1]School of Chemical, Biological and Materials Engineering, Gallogly College of Engineering, University of Oklahoma, Norman United States

[2]Low Carbon Energy and Resources Technologies Research Group, Faculty of Engineering, University of Nottingham, Triumph Road, Nottingham NG7 2TU, the United Kingdom.

[3] Advanced Materials Research Group, Faculty of Engineering, University of Nottingham, Nottingham, NG7 2RD, the United Kingdom.

Engineering pathways, Gallogly College of Engineering, University of Oklahoma, Norman United States

Corresponding author

Jude A. Okolie

Assistant Professor, Engineering pathways

Gallogly College of Engineering, University of Oklahoma

Norman United States

Jude.okolie@ou.edu

Abstract

Chemical looping gasification (CLG) is a promising thermochemical process for the production of H_2. CLG process is mainly based on oxygen transfer from an air reactor to a gasification reactor using solid metal oxides (also called oxygen carriers) as oxidants. The unique oxygen separation system of CLG makes it an advanced process with a smaller carbon footprint compared to the conventional gasification process. The other advantages of CLG includes increased efficiency, reduced greenhouse gas emissions, and improved process stability compared to conventional biomass gasification. Despite the advantages of CLG, the relationship between biomass properties and experimental conditions is still unclear. This could be attributed to rigorous experimental requirements. To address the limitations, the present study proposes a process simulation is combined with experimental studies to generate large dataset used for interpretable machine learning (ML) studies. Three different ML models including support vector machine (SVM), random forest (RF), and gradient boost regression (GBR) were used to develop models for predicting the H_2 and char yield during CLG. The GBR outperformed other models for the prediction of H_2 and char yield during CLG with R^2 value > 0.9. Among the experimental conditions, the temperature (T) and steam to biomass ratio (SBR) were the most relevant parameters influence H_2 and char production. Biomass ash, C, VM and H also influenced H2 and char formation. Biomass ash could act as catalyst during CLG thereby promoting char accumulation. Overall, a combination of SHAP and partial dependence plot helped address the black box challenges of ML models.

Keywords: Gasification; chemical looping; temperature; machine learning; regression

1. Introduction

In modern times, the increasing energy demand brought about by population growth and the rapid pace of industrialization and urbanization creates a range of challenges globally. Currently, the primary energy source is derived from fossil fuels, but this approach is faced with various limitations, including, the finite nature of fossil resources, the greenhouse gases emissions, global warming, and pollution resulting from their extraction, refinement, and usage (Nanda et al., 2014). It is, therefore, important to explore alternative energy sources that are cost-effective, sustainable, dependable, and eco-friendly, like biofuels (e.g., bio-oil, biodiesel, bioethanol, biobutanol, biomethane, and biohydrogen). These sources should also have the ability to supplement fossil fuels and help address the environmental problems caused by their use (Okolie, 2022; Okolie et al., 2019).

Renewable energy sources such as wind, solar, and geothermal have been utilized for energy production, but they are not suitable for the production of liquid transportation fuels [Ref]. Additionally, these energy sources present further challenges related to intermittency, or the fluctuations in their output that can occur due to changes in weather or other factors. Biomass, on the other hand, is defined as an alternative source of renewable energy thanks to their abundance and ability to be converted to both liquid and gaseous transportation fuels. (Goel et al., 2022). The Intergovernmental Panel on Climate Change (IPCC) fifth assessment report highlighted that integrating biomass energy with effective carbon capture technologies could aid in reaching long-term climate goals for a negative carbon economy (Lamers et al., 2015).

Biomass can be converted into green fuels and chemicals through various methods such as thermochemical processes (e.g., pyrolysis, liquefaction, and gasification), biological processes (anaerobic digestion and fermentation), or integrated processes (Okolie et al., 2022a). Of these, thermochemical processes are preferred due to their high biofuel yields, shorter processing times, and improved economic viability. Gasification is a thermochemical process used to produce hydrogen-rich syngas from waste biomass using gasifying agents such as air, steam, CO_2, and supercritical water (hydrothermal

gasification). Compared to other thermochemical processes, gasification has advantages in terms of feedstock versatility, reduced greenhouse gas emissions, and increased energy efficiency (P.Basu, 2016). However, challenges such as the high cost of oxygen production, N_2 dilution, and emission of pollutants, as well as the strong endothermic reaction with steam gasification agents still exists (Huang et al., 2016). To overcome these challenges, the chemical looping gasification (CLG) process has been proposed as a promising technology for converting waste biomass into hydrogen-rich syngas.

The CLG process is mainly based on oxygen transfer from an air reactor to a gasification reactor using solid metal oxides (also called oxygen carriers) as oxidants. In the gasification reactor, biomass feedstocks are partially oxidized and gasify to high quality syngas (CO and H_2) by the oxygen provided by solid metal oxides and gasification agent (Mohamed et al., 2021; Wang et al., 2015). During this reaction, metal oxide is reduced to a lower oxidation state. In a second reactor (an air reactor), the reduced oxygen carrier is oxidised by air or steam stream, which is an exothermic reaction and supply a significant amount of heat source, which could be used in the endothermic gasification reactor and eliminate the requirement of external carbon combustion as in the traditional gasification processes. The re-oxidized oxygen carrier is ready for a new cycle between these two interconnected reactors in a solid circulatory loop while the gas flows in these reactors are isolated using gas seals between the reactors. Circulating oxygen carriers can provide a continuous oxygen source for biomass gasification, eliminating the need for pure oxygen production and reducing costs (Huang et al., 2016).

The unique oxygen separation system of CLG makes it an advanced process with a smaller carbon footprint compared to the conventional gasification process. The other advantages of CLG includes increased efficiency, reduced greenhouse gas emissions, and improved process stability compared to conventional biomass gasification (Goel et al., 2022). (Huang et al., 2016) Certain oxygen carriers, such as Fe- and Ni-based, and their reduction products, can effectively be used as catalyst to crack tar during the gasification process (Liu et al., 2012). The CLG process also eliminates the need for N_2 dilution, as the fuel is not directly exposed to air, leading to the production of high-quality synthesis gas with a high

calorific value and low tar content, at a lower cost (Goel et al., 2022). An overview of the CLG including details of the oxygen carrier and potential gas-solid interactions are presented in figure 1.

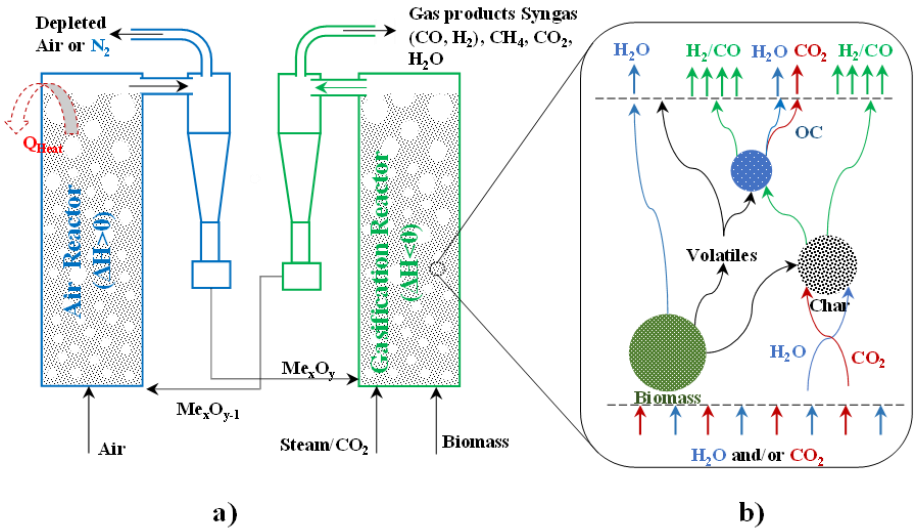

Figure 1: a) Overview of CLG of biomass for syngas production and b) potential reactions in the gasification reactor

Although CLG is a promising technology, it still faces several challenges such as high capital cost, oxygen carrier durability, complex reaction mechanism and process operations, scalability issues and increased CO_2 capture cost. Some of these challenges can be addressed understanding the impact of various process conditions on hydrogen yield and char formation during CLG. Char produced during CLG can accumulate in the reactor and reduce its efficiency. It can also clog the flow of gases through the reactor and reduce the quality of the syngas produced. It is, therefore, imperative to understand the influence of process parameters on char formation during CLG.

Despite its potential, there are still knowledge gaps in this process, including the optimization of reaction conditions and the development of cost-effective and durable materials for the chemical looping reactors. Addressing these gaps is crucial for the commercialization of this process and its widespread adoption as

a solution to reduce greenhouse gas emissions while meeting the growing demand for liquid fuels. Although some researchers have explored the process analysis and techno-economic analysis (TEA) of integrating CLG with FT synthesis (Roshan Kumar et al., 2022b, 2022a). To the best of the authors knowledge there are still relatively few study that have adopted interpretable machine learning (ML) approach for understanding the relationship between process parameters, hydrogen yield, and char formation during CLG. To address the knowledge gaps the present study develop and systematically compared three ML models; random forest (RF), support vector machine (SVM) and gradient boost regression (GBR) for predicting hydrogen and char yield during CLG. Another novelty is the integration of Aspen plus process models with experimental data for obtaining robust CLG covering a wide range of operating conditions including proximate and ultimate analysis of biomass.

2. Materials and methods

2.1 Process simulation description

Process modelling for CLG was developed in Aspen plus V12 licensed by the University of Oklahoma. Details of the process model is shown in figure 2. The feed which is at atmospheric conditions was decomposed at 700°C into various constituent elements. The decomposed product was then separated into gas and solid phases. The gas mixture was subsequently compressed to 23bar before it was sent into the stoichiometric reactor (RSTOIC) and continuous stirred tank reactor (RCSTR) for gasification (oxidation and reduction reactions occurred here). Fe-based oxygen carrier was used for the oxidation process. The products from the reactors were subsequently sent into separators that separated hydrogen gas and CO_2 from other reactor products.

The model consists of a yield reactor that decomposed the feedstock into its various constituents, an RSTOIC reactor for the volatilization process, an RSTOIC reactor for the oxidation processes

and an RCSTR reactor for the reduction gasification processes, a cyclone that separated solids (char from gas components), two heaters that provided heat for the raw feed and steam feed, a cooler to reduce the temperature of reactor products, 2 mixers, 2 compressors (that compressed the reactor feed as well as the hydrogen gas produced) and 3 separators that separate hydrogen from other reactor products and 2 storage tanks for CO_2 and H_2 storage. Summary of the model assumptions and equations are presented in Table 1.

The selection of a suitable oxygen carrier is important for the efficient operating of a CLG systems. A plethora of metal oxides of Fe, Ni, Co, Mn and Cu, as well as their blends, have been comprehensively evaluated as promising oxygen carriers (de Diego et al., 2004; Huijun et al., 2015). However, they have several limitations, such as the oxide of Cu suffers from agglomeration due to its low melting point (Ge et al., 2016). Mn and Co oxides ae prone to a poor issue while Ni oxides are characterized by low toxicity. Oxides of Fe are one of the most promising carriers and was the selected choice for the present study. The results of Aspen plus model were compared against experimental investigations and presented in figure 3. The experimental hydrogen yield for is close to the simulation results for two different experimental studies with similar oxygen carriers (FeO) and biomass at varying temperature ranges of 700 – 900 °C.

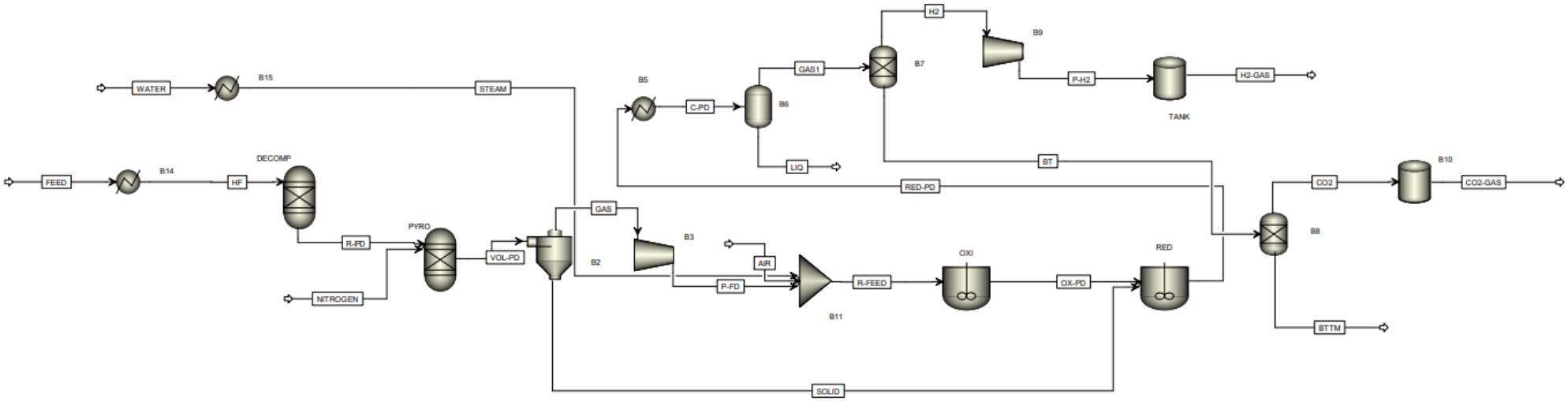

Figure 2: Schematics of the chemical looping gasification process designed in Aspen Plus

Table 1: An overview of the model assumptions and relevant equations

Component	**Enthalpy**	**Density**
Biomass (Non- conventional)	HCOALGEN	DCOALIGT
Ash	HCOALGEN	DCOALIGT
Reaction Kinetic Parameters Data		
REACTION	A/k	Ea. (KJ/Kmol)
R1	200	49900
R2	300000	125000
R3	2.78	12600
R4	1.05e10	135000
REDUCTION: • $C + H_2O \rightarrow CO + H_2$ • $CH_4 + H_2O \rightarrow CO + 3H_2$ • $CO + H_2O \rightarrow CO_2 + H_2$ • $C + CO_2 \rightarrow 2CO$ **FUEL COMBUSTION REACTION** **C + Fe_2O_3 → 2Fe + 3CO** **IRON OXIDATION REACTION** • **2Fe + H2O → Fe_2O_3 + H_2** • **Fe + CO_2 → FeO + CO** **OVERALL REACTION** • **C + Fe_2O_3 → 2Fe + 3CO** • **Fe + CO_2 → FeO + CO** • **FeO + C → Fe + CO**		

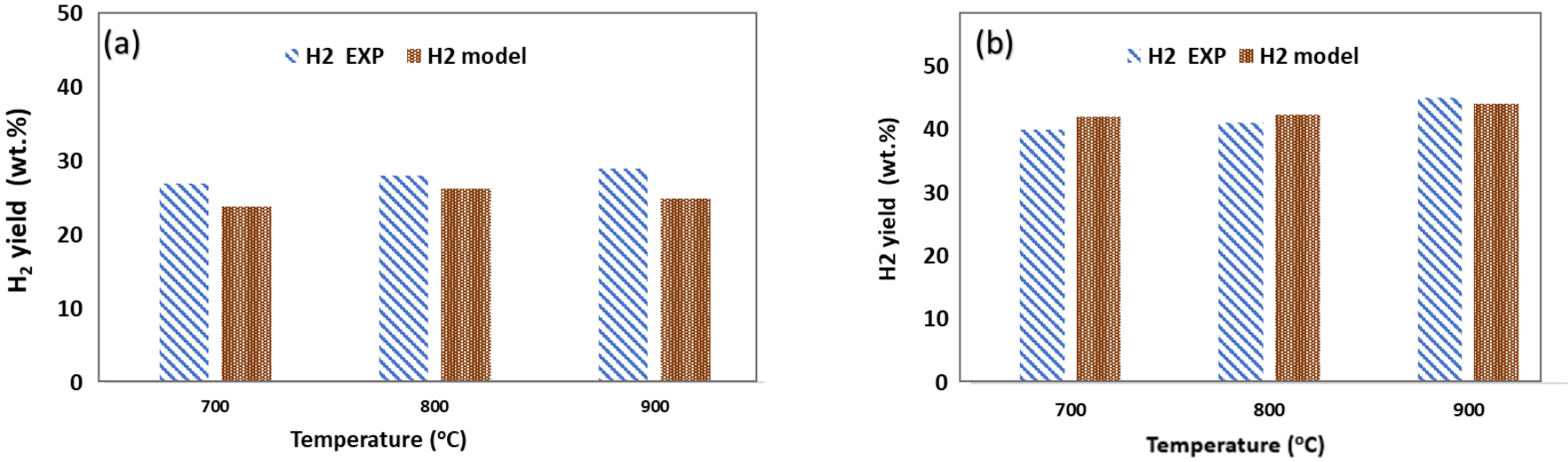

Figure 3: Comparison between Aspen plus model results and experimental hydrogen yield (a) Experimental investigations from Al-Quadri et al. (Al-Qadri et al., 2022) (b) Experimental investigations from Ge et al. (Ge et al., 2016)

2.2 Data preparation

In total 236 datasets related to CLG of various biomass feedstock were obtained from literature and process simulation results. The biomass feedstock selected for this study includes agricultural residues, energy crops, municipal waste, algae biomass, and poultry waste. The dataset includes twelve input variables (carbon (C), hydrogen (H), Nitrogen (N), oxygen (O), Sulphur (S), Moisture content, Volatile matter (VM), Ash content (Ash), Fixed carbon (FC), gasification temperature (T), oxygen carrier quantity measured as mass ratio to biomass (OC) and Steam to biomass ratio (SBR)). The output variables are H_2 and char yield. CLG is a relatively new process with limited experimental results on various feedstock and oxygen carriers therefore with the validated process simulation data were used to supplement experimental findings. In addition, CLG process conditions such as OC, SBR and T were selected as part of the input variable due to their significant influence on hydrogen yield (Al-Qadri et al., 2022; Ge et al., 2016). In this study 236 number of data were used to train and test the ML models (The entire dataset can be found in GitHub link of the supplementary information). The dataset was split into training and test data in 80/20 ratios, this is performed in google colab in python. The training data were used to train the ML models while the testing data set were implemented in the verification of the model accuracy and generality of the trained ML models.

2.3 Model development and evaluation

2.3.1 Data pre-processing

Data cleansing was performed before developing the ML model. The implementation of data cleansing ensures optimal ML model prediction performance. Data cleansing involves sequential steps of handling missing data, outliers and standardization. All missing data were replaced using the median approach. This is performed by replacing the missing values with the mean, median, or mode of the available data. This was implemented by using the “fillna “function in python

Standardization is a technique used to transform data from varying scales into a common scale for easy comparison. This is achieved by normalizing the data with a mean of 0 and a standard deviation of 1 through the subtraction of the mean and division by the standard deviation. This approach is advantageous since several machine learning algorithms presume normally distributed data and features on the same scale. By scaling down outliers along with the rest of the data, standardization also helps to minimize their impact. The sklearn.preprocessing library's StandardScaler, MinMaxScaler, and RobustScaler were utilized to assess model performance. Of these, the MinMaxScaler was found to be the most effective and thus chosen. MinMaxScaler is a machine learning pre-processing technique that scales features in a dataset by changing them to a specified range, frequently 0 to 1, by subtracting each data point's minimum value and dividing by the range. This method normalizes the data and is particularly beneficial for algorithms that are sensitive to input feature scaling.

It is crucial to remove outliers before building machine learning models for several reasons. Outliers can skew a model's parameters, leading to poor performance on new data, and increase model complexity, making it harder to understand and use. They can also cause overfitting, resulting in poor performance on new data, and affect data distribution, causing issues with certain types of models (Ascher et al., 2021). A box plot was prepared and used to remove outliers from the dataset (Figure S1 of the supplementary information).

2.3.2 Model development

Three different ML algorithms; the support vector machine (SVM), random forest (RF), and gradient boost regression (GBR) were used to develop models for predicting the H_2 and char yield during CLG. The optimal hyper-parameter were selected through the grid search method. It involves specifying a discrete set of values for each hyperparameter and evaluating the performance of the model for all possible combinations of these values. The combination that results in the best performance metric is then chosen as the optimal set of hyperparameters. The ML algorithms were selected based on their effectiveness in handling multiple datasets related to several biomass composition and operating

conditions for thermochemical conversion processes. Details of each ML algorithm has been meticulously described elsewhere (Afolabi et al., 2022; Okolie et al., 2022b).

The performance of each ML algorithm towards predicting the output was evaluated with the mean absolute error (MAE), regression coefficient (R^2), and root-mean-square error (RMSE). These parameters have been described elsewhere (Umenweke et al., 2022) and are calculated from equations 1 – 3.

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

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

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

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. N represents the total number of data points.

2.4 Interpretable analysis

Due to the complexity of their internal mechanisms, many machine learning models are often regarded as "black boxes". To improve interpretability, one approach is to use inherently interpretable models, which can be categorized as either model-specific or model-agnostic (Onsree et al., 2022). Model-specific techniques are limited to particular model types, such as interpreting regression weights in a linear model. In contrast, model-agnostic methods can be utilized with any machine learning model and applied post-training (Onsree et al., 2022). As a result, model-agnostic methods are generally more adaptable and all-encompassing, while also providing a consistent standard for interpreting a range of machine learning models. To provide a comprehensive and global interpretation of the relationship between the input and outputs in datasets, this research employs the widely used model-agnostic method known as the SHAP feature importance method.

SHAP represents a promising method that enables both individual prediction explanation and global interpretation (Ascher et al., 2022). Grounded in game theory, SHAP estimates feature importance by distributing optimal credits based on Shapley values. SHAP force plots visually depict how various features influence an individual prediction. This method is advantageous for global interpretation because it determines the importance of each feature and its relationship with the output. Moreover, SHAP ensures that its predictions are equitably distributed across feature values, which is essential for establishing trust in the method (Pintelas et al., 2020). To gain insights into the model's local interpretation, a partial dependence plot (PDP) will be utilized. By holding other features constant, a PDP illustrates the relationship between a feature and a model's predicted outcome. It aids in identifying non-linear relationships or interactions between features and provides insights into how a specific feature influences a model's predictions for a specific value range.

3. Results and discussion

3.1 Dataset description and statistical analysis

The statistical analysis of the final preprocessed dataset were quantified through the standard deviation and corresponding mean. These values are presented in Table 2. The input dataset were classified as ultimate analysis (C, H, N, O, S), proximate analysis (Ash, FC, moisture, VM) and operating conditions (T, OC, SBR). In contrast, there are two output variables (H_2 yield and char yield). The C, H, N, O, S dataset were in the range of 40 – 60.5%, 4 – 13.7%, 0.1 – 8.2%, 24.7-59.2% and 0 – 2.3% respectively. In contrast, the Ash, FC, moisture, VM values ranged between 0.4 – 42%, 3.4 – 26.6%, 0.001 – 0.002% and 49.4% - 94.2%. It should be mentioned that the proximate and ultimate analysis were all on dry basis therefore the moisture content used in the dataset is close to zero. The dry basis was selected because CLG requires pre-drying of feedstock to minimize the moisture content. In addition, the process simulation had a feedstock drying step before the actual CLG simulation. The output dataset ranged between 11.5 wt.% - 73.3 wt.% for H_2 yield and 0.6 – 68.2 wt.% for the char yield.

The large range of distribution between the proximate and ultimate analysis could be attributed to the various categories of biomass selected for the simulation. In order to evaluate the impact of biomass type and composition on CLG different range of biomass types were selected including agricultural residues, industrial waste, woody biomass, energy crops and municipal solid waste.

The operating conditions ranged between 500 – 1300 ^{o}C for T, 0.5 – 5 for SBR and 10 – 60 for OC. It should be mentioned that the range of values for the operating conditions were selected based on optimization results from previous experimental studies (Al-Qadri et al., 2022) (Ge et al., 2016).

Table 2: Overview of the statistical summary of the pre-processed dataset including the input and output features.

	count	mean	std	min	25%	50%	75%	max
C (%)	236	49.9	4	40	47.6	48.9	52.8	60.5
H (%)	236	6.4	1.3	4	6	6.2	6.4	13.7
N (%)	236	1.4	1.5	0.1	0.5	0.9	1.6	8.2
O (%)	236	42.2	4.7	27.4	39.4	44.1	45.6	52.9
S (%)	236	0.2	0.4	0	0	0	0.3	2.3
VM (%)	236	75.3	7.6	49.4	71.1	76.1	80.7	94.2
Ash (%)	236	7.4	7.9	0.4	1.9	4.5	10	42
FC (%)	236	17.3	5.3	3.4	13.2	17.2	20.9	26.6
Moisture (%)	236	0	0	0	0	0	0	0
T (°C)	236	902.2	259.5	500	700	900	1100	1300
OC (%)	236	33.4	17.4	10	20	30	50	60
SBR	236	2.6	1.4	0.5	1.5	2.5	3.5	5
H2 (wt.%)	236	29.4	13.2	11.5	20.6	25.9	35.8	73.3
Char yield (wt.%)	236	13.2	13.7	0.6	3.6	7.4	17.6	68.2

Figure 4 shows the Pearson coefficients heat map for the input and output variables. The figure helps to understand the correlation between related variables in the ML model and it is based on the Spearman correlation coefficient (SCC) as shown in equation (4). Understanding the correlation between different sets of variables is important especially when it is required to develop trends between them.

$$SCC = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (4)$$

Where SCC = Spearman's rank correlation coefficient,

d_i = difference between the two ranks of each observation,

n = number of observations

SCC ≈ 0 signifies that the variables are weakly correlated, whereas SCC ≈ ±1 suggests the highest correlation strength. It should be mentioned that most of the features demonstrated non-linear relationships with each other, which was favorable for establishing ML models. The figure reveals that the char yield is strongly affected by the proximate and ultimate analysis (except the moisture content) of the feedstock as well as the T and SBR. In contrast, H_2 yield is influenced by the char yield, T, SBR, ash, S, N and H content of the feedstock. The FC content represents the amount of solid combustible residue that is left behind after the biomass is heated and the volatile matter is expelled. While the ash content is the number of solid residues remaining after complete burning of biomass (Cai et al., 2017). Since both the FC and ash contents are solid residues left behind, they would influence the char formation.

SBR (0.51), ash (0.22), S (0.21), N (0.13) and H (0.14) have strong positive correlation with H_2 yield while T (-0.72), moisture (-0.13), C (-0.19) and VM (-0.16) have strong negative correlation. On the contrary, S (0.63), N (0.62), H (0.44) has a strong positive correlation with char yield while VM (-0.72), FC (-0.37), moisture (-0.37) and C (-0.18) are negatively correlated.

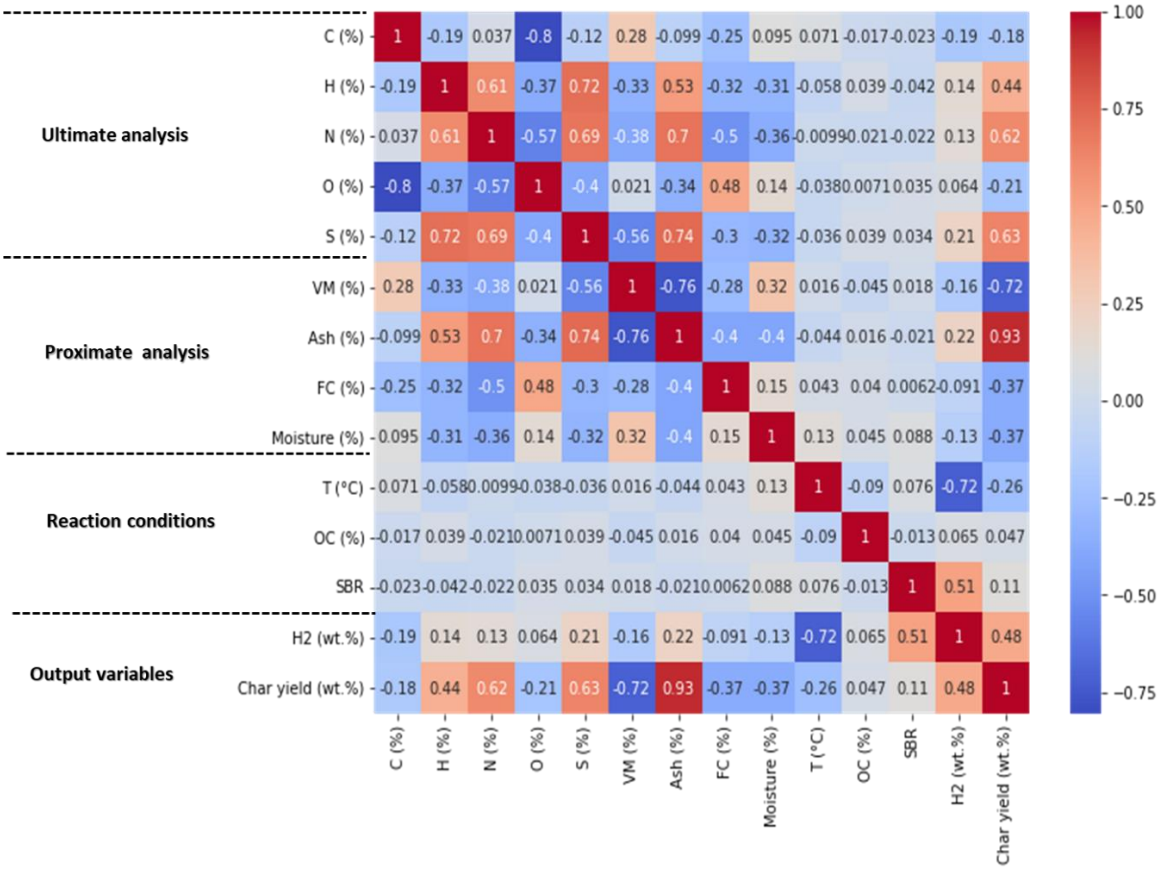

Figure 4: Pearson coefficient heat map between any two variables of interest

3.2 **Model optimization and evaluation of model performance**

Tables 3 and 4 exhibit the outcomes of model performance before and after eliminating outliers and conducting hyperparameter optimization, comparing the optimized models with default models using effective preprocessing methods such as data normalization and outlier elimination. Among the three ML models, only GBR showed slight improvement as a result of the hyperparameter optimization and data cleansing, whereas SVM and RF did not exhibit any improvement. For instance, the R^2 values for the test dataset of SVM H_2 declined from 0.9417 (Table 3) to 0.93819 (Table 4), and the R^2 values for the char yield of SVM also decreased from 0.9941 (Table 3) to 0.9381 (Table 4). Similar patterns were observed for the R^2 values of the test dataset of RF. On the other hand, GBR's R^2 value for the test dataset increased after removing outliers and conducting hyperparameter optimization for the H_2 yield alone. R^2

value for the test dataset of GBR rose slightly from 0.9449 to 0.9453 for H_2 yield. In contrast, the value declined slightly for GBR char yield (0.9657 to 0.9453). While the RMSE and MAE values for GBR test dataset decreases after removing outliers and conducting hyperparameter optimization (Except for the RMSE char yield). In general, hyperparameter optimization can be useful in obtaining the optimal model structure when computational resources during model training are not a constraint.

The model evaluation results in Tables 3 and 4 shows that the R^2 values of all the ML models were > 0.9 for H_2 and char yield. This indicates a satisfactory fitting effect of all the three models. However, by comparing the performance of the models the GBR appears to be the most promising model in terms of R^2, RMSE and MAE. This was further confirmed by the relative error distribution plot for the three ML models presented in figures 5. The relative error in this case can be defined as the ratio of difference between the actual and predicted values against the actual values. From the plot in figure 5, the GBR had the lowest relative error followed by RF and SVM. It should be mentioned that the relative error distribution range for all the three models were less than 50%. Although, GBR had the lowest relative error distribution (< 5% for H_2 yield and 35% for char yield).

The GBR and RF are more advantageous than the SVM model for predicting the H_2 and char yields during CLG because they are ensemble methods. These methods implement multiple models to improve the accuracy and robustness of predictions. The idea behind ensemble methods is that by combining the predictions of multiple models, we can reduce the variance and bias of individual models and achieve better overall performance. Ensemble methods are generally considered to be better than individual models because they can improve the accuracy, robustness, and generalization of predictions, and can also help to identify important features and patterns in the data (Ascher et al., 2022). A key difference between the RF and GBR is that multiple decision trees are trained independently on different subsets of the training data in RF. In contrast, a sequence of decision trees are trained in a way that each new tree is optimized to correct the errors of the previous ones for the GBR. Both methods also differs in the wat that handle their corresponding features.

Table 3: Performance of the model before removing outliers and hyper parameter optimization

Models	H_2 yield						Char yield					
	R^2_train	**R^2_test**	**RMSE_train**	**RMSE_test**	**MAE_train**	**MAE_test**	**R^2_train**	**R^2_test**	**RMSE_train**	**RMSE_test**	**MAE_train**	**MAE_test**
SVM	0.9946	0.9417	0.9825	2.9149	0.3627	1.7319	0.9999	0.9941	0.1171	0.9922	0.0976	0.634
RF	0.9853	0.9229	1.6219	3.3529	1.0681	2.417	0.9904	0.9472	1.3544	2.9621	0.778	1.7628
GBR	0.9984	0.9449	0.5345	1.4854	0.3733	0.9824	0.9992	0.9657	0.3963	2.3878	1.1591	0.2818

Table 4: Performance of the model after removing outliers and hyper parameter optimization

Models	H_2 yield						Char yield					
	R^2_train	R^2_test	RMSE_train	RMSE_test	MAE_train	MAE_test	R^2_train	R^2_test	RMSE_train	RMSE_test	MAE_train	MAE_test
SVM	0.9464	0.93819	3.0963	3.0024	1.4084	1.8558	0.9464	0.9381	3.0963	3.0024	1.4084	1.8558
RF	0.9906	0.9438	1.2930	2.8622	0.8562	2.0771	0.9906	0.9438	1.2930	2.8622	0.8562	2.0771
GBR	0.9999	0.9453	0.0753	2.8228	0.0500	1.7440	0.9999	0.9453	0.0753	2.8228	0.0500	1.7440

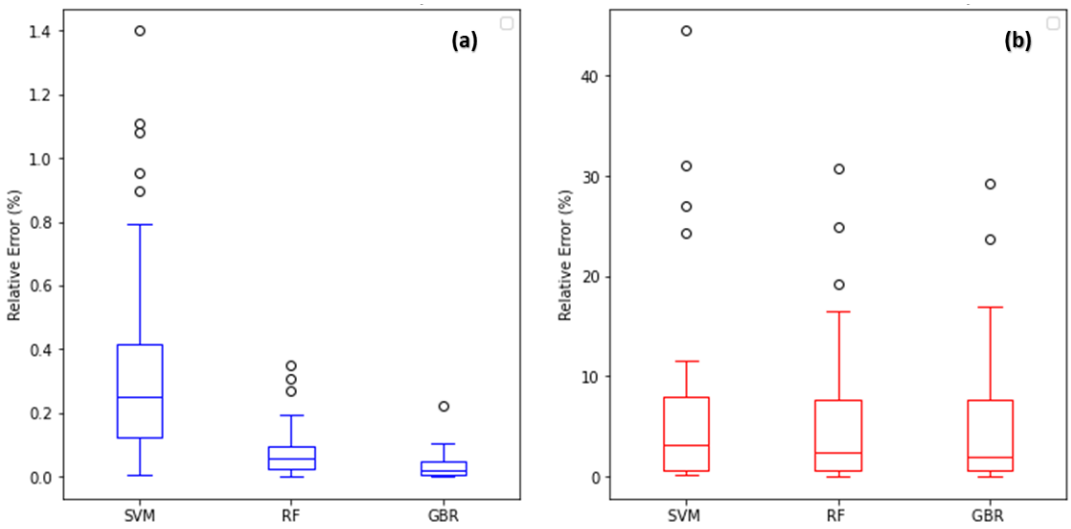

Figure 5: Relative error distribution of the three ML models for (a) H_2 yield (b) char yield

Several studies related to the application of ML for understanding and optimization of biofuels production system have demonstrated the promising capability of the ensemble methods. Zhao et al. (Zhao et al., 2021) compared the performance of GBR, RF, ANN and SVM for the prediction of H_2 yield during supercritical water gasification. The authors reported that RF performed better than all the other models. In another study, GBR performed better than the RF and decision regression tree (DRT) for the prediction of bio-oil properties including the yield, nitrogen content, higher heating values and energy recovery during the hydrothermal liquefaction of waste biomass (Li et al., 2021).

After evaluating the different models in the present study, the optimized GBR was chosen for further analysis, prediction, and interpretation of the input data. A plot of the predicted output values against the actual (measured) output values was presented in Figure 6, which provides a detailed understanding of the GBR's performance in predicting H_2 and char yield. It should be noted that a higher number of cluster points around the 45° line indicates that the ML model provided optimal performance. As shown in Figure 6, both the measured and predicted H_2 and char yield cluster closely around the straight line, indicating that the GBR model is well-suited for making accurate predictions.

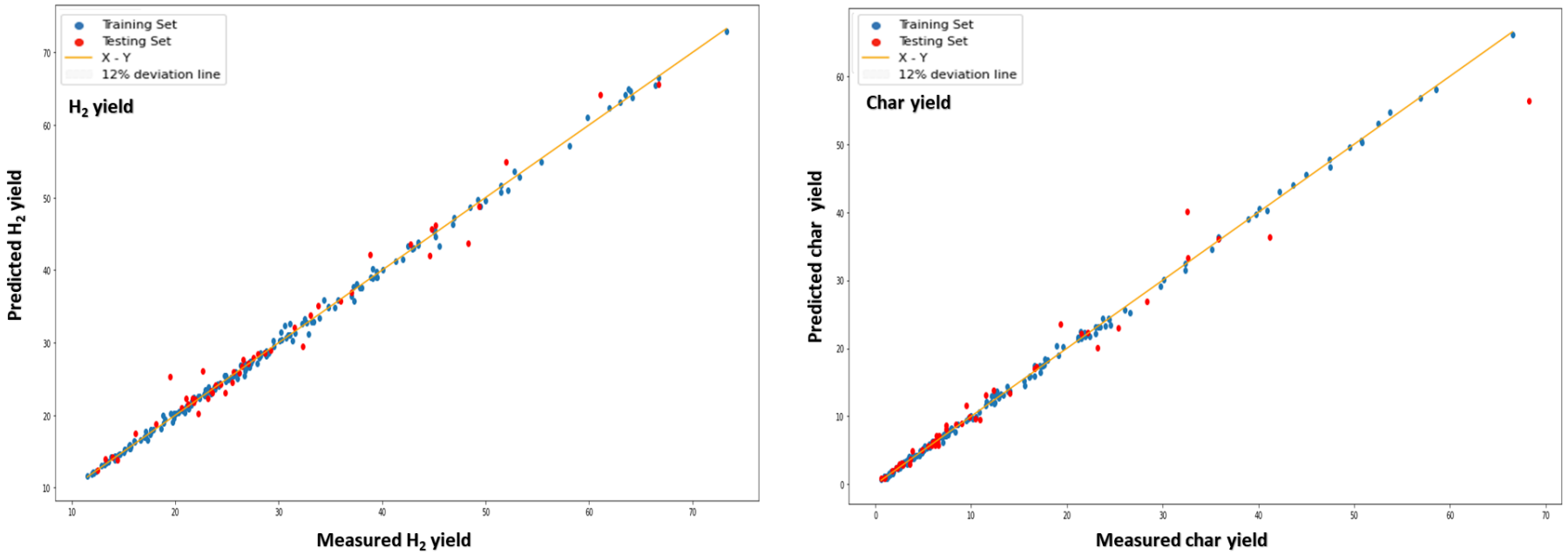

Figure 6: Comparison of the predictions of the GBR model for each output fitted from the training and testing set

Overall, the GBR model's performance evaluation results and the scatter plot analysis suggest that the GBR model can accurately predict the H_2 and char yield values, and it can be used for further analysis and prediction of the input data. The presented results and analysis provide useful insights into the machine learning model's performance, helping researchers to better understand the relationship between the input and output variables and improve the model's accuracy in future studies.

Although the present study concluded that GBR was the most effective model, previous research has identified a variety of models that are also suitable. No single model type appears to be dominant in the literature. Afolabi et al. identified RF as the most promising model for predicting the HHV of different classes of lignocellulosic data based on 237 biomass datasets (Afolabi et al., 2022). Elmaz et al. (Elmaz et al., 2020) conducted a study in which they trained polynomial regression, SVM, ANN, and decision tree models using data from an in-house gasifier that was fed with pinecones and wood pellets. They found that decision trees and ANN were the preferred model types for their data set. However, despite their data set being more uniform than the one used in the present study, the decision tree models they studied achieved a lower test performance, with R2 values ranging from 0.81 to 0.94. In contrast to the proposed model optimization approach in our study, Sun et al. (Sun et al., 2022) employed particle swarm optimization to develop an ANN model for predicting syngas yield, gas species concentrations, and char yield. The model achieved an excellent test performance with an R^2 value of 0.97. However, the authors pointed out that the model was only trained on data from pine wood gasification, and they highlighted the importance of expanding the data set by incorporating a broader range of feedstocks and gasification conditions to enhance the model's applicability. Tang et al.(Tang et al., 2021) aimed to predict pyrolytic gas yield and compositions based on pyrolysis conditions and biomass characteristics using ML algorithms combined with feature reduction. RF and SVM were compared, and the results indicated that six features were adequate to accurately forecast the yield while the compositions only required three.

3.3 Interpretable machine learning analysis

As stated earlier, the GBR model showed superior performance, indicating its effectiveness in describing the complex correlation between input features and output (H_2 and char yield). Therefore, the GBR model was used to analyze the importance of features and the partial dependence of various inputs on the production of H_2 and char yield during CLG. Figure 7 shows the feature importance plot, which displays the relevance of proximate and ultimate analysis of biomass and the CLG operating conditions on H_2 and char yield. The analysis reveals that among the biomass properties, only FC, VM, ash, C, H, and S have a noticeable impact on the H_2 yield. Likewise, operating conditions such as SBR and T have a significant impact on H_2 yield. Although the effects of Ash, VM, S, and H are relatively small. With regards to the formation of char, biomass properties such as Ash, VM, S, H, and C have a significant impact. Additionally, operating conditions including SBR, and T also affect the formation of char. Based on the feature analysis findings in Figure 7, the order of relevance of process parameters in influencing H_2 and char yield is as follows: T > SBR > OC. When considering the impact of biomass properties, the order is as follows: FC > C > Ash > VM > S > H > moisture (for H_2 yield) and Ash > VM > S > C > H > FC (for char yield). It should be noted that the content of O, N, and moisture had no impact on the H_2 and char yield.

In theory, it is not surprising that biomass with a higher content of H can produce more H_2 during CLG. Two prominent parameters in evaluating H_2 and char yield during CLG are T and SBR, for several reasons. Some studies have shown that the presence of steam can minimize tar and char formation during gasification (Gao et al., 2009; Niu et al., 2018). Moreover, steam can reform tars and heavier hydrocarbons, while also reacting with carbon black trapped in the surface of ceramic pores, increasing hydrogen production (Gao et al., 2009). Because most of the reactions involved in CLG are either endothermic or exothermic, the impact of temperature is significant in controlling these reactions.

.

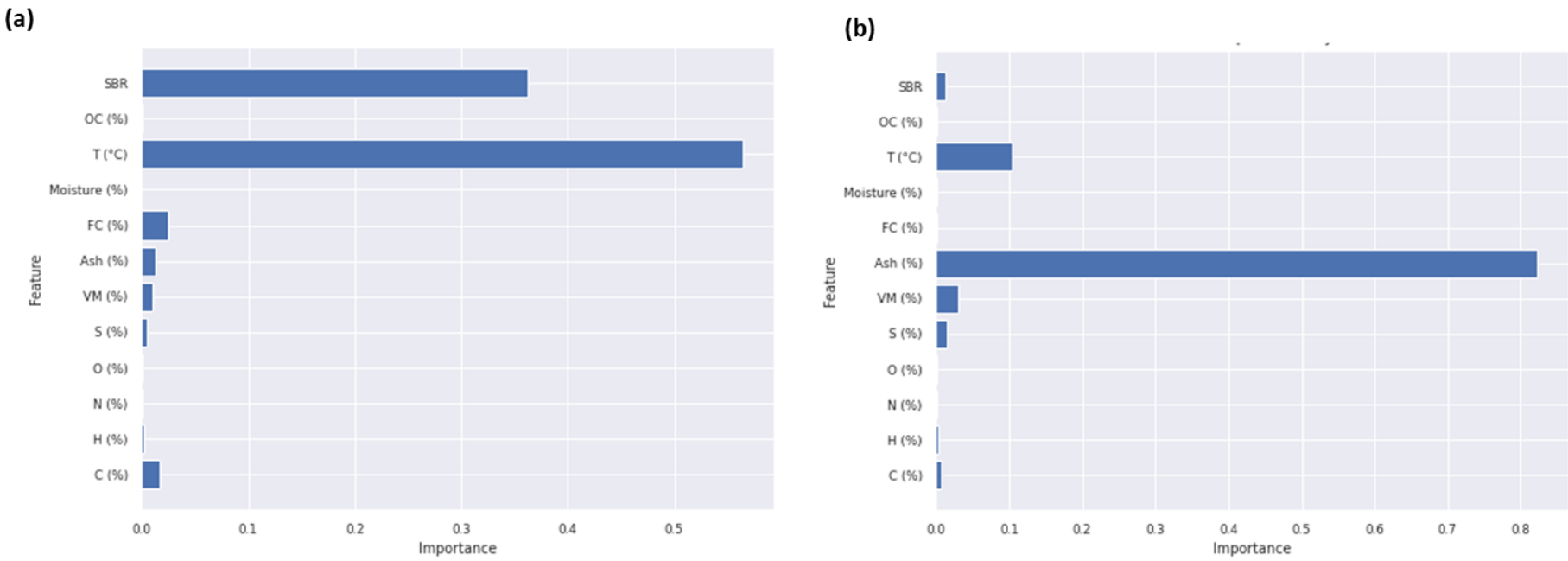

Figure 7: Feature importance analysis showing the impact of CLG operating conditions and biomass properties on (a) H_2 yield (b) char yield

A PDP of the key features were presented in figures 8 and 9 to help understand the marginal effect of one or two features on the predicted results of the GBR model for both H_2 and char yield. The prediction function was fixed at certain values of the selected features, and the other features are averaged, so as to separate the influence of changing the important feature value on H_2 and char yield. The results for H_2 yield showed that it is not affected by O, N and moisture content of the biomass. Also, the oxygen carrier quantity measured as mass ratio to biomass (OC) does not influence H_2 yield. In contrast, H_2 yield is slightly affected by a change in C, H, S, VM, Ash, and FC content of the biomass feedstock. H_2 yield is strongly impacted by T and SBR. An elevation in SBR led to a rapid increase in H_2 yield. Surprisingly, H2 yield increased up to a maximum of 700 ^oC then declined rapidly as the temperature rose to 1300 ^oC.

Regarding the char yield, the properties of biomass such as C, H, N, S, O, FC, and moisture contents almost had negligible effect on char formation. Similarly, OC had negligible effect on char formation. In contrast, biomass with higher ash and FC content are predicted to produce immense amount of char. Specifically, increasing the ash content of biomass feedstock led to a significant rise in char formation. Since char are undesirable during CLG, it is important to ensure that the ash content of biomass feedstock is minimal. The operating conditions including T and SBR also influenced the char yield. Increasing T led to a decline in char yield while higher SBR slightly increased char formation.

SBR plays a critical role in the CLG process as it influences the thermochemical conversion of biomass into gases. The amount of steam in the gasification process determines the extent of tar decomposition, while also playing a role in the char gasification and H_2 generation as seen in figures 8 and 9. Increasing the SBR ratio can lead to better H_2 production by facilitating the water-gas shift reaction, which increases the amount of H_2 produced. Additionally, high SBR ratios help in reducing the formation of tar and other heavy hydrocarbons by increasing the amount of steam available for the reactions, resulting in improved gas quality.

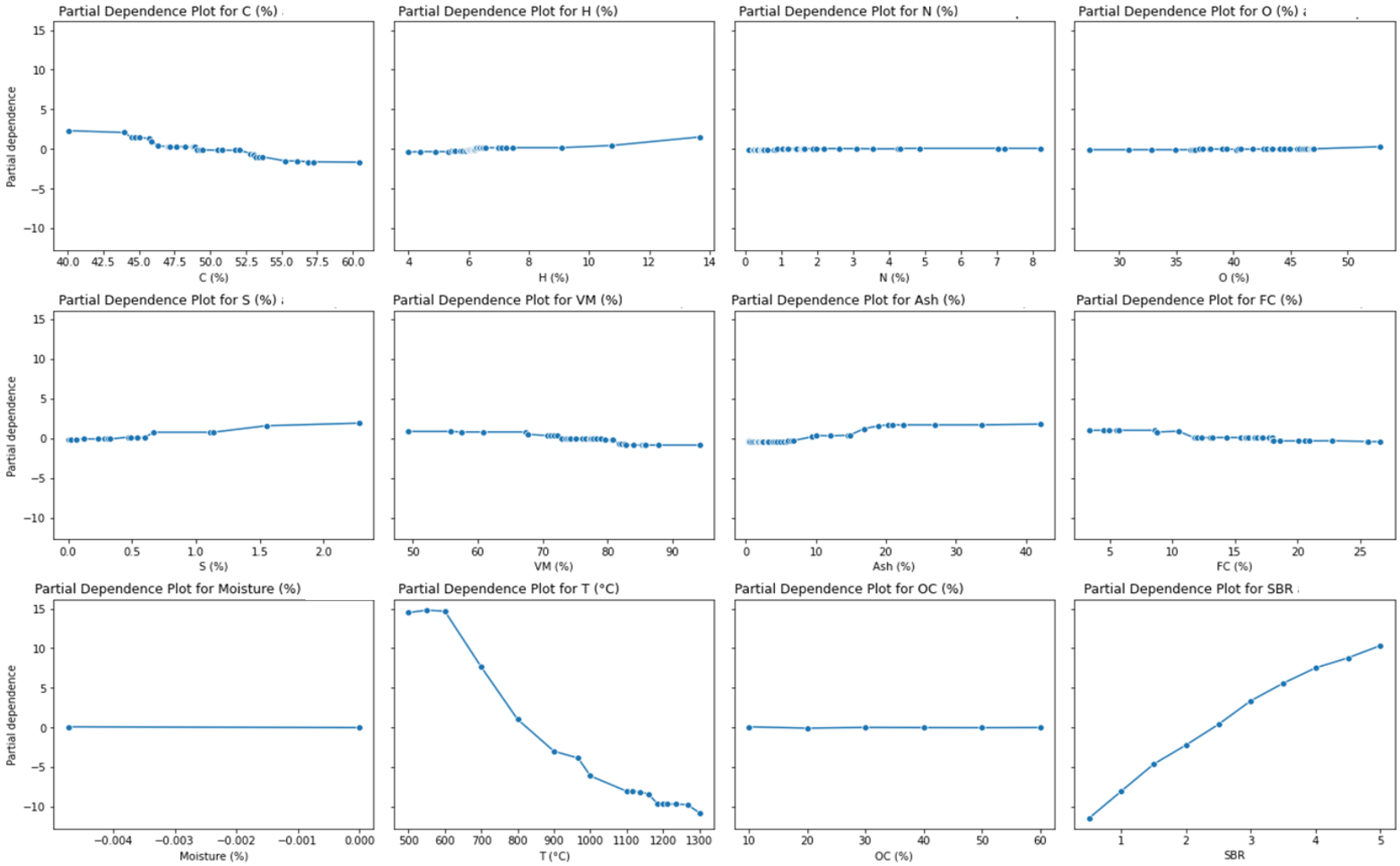

Figure 8: One variable partial dependence plot showing the impact of operating conditions and biomass properties on H_2 yield during CLG

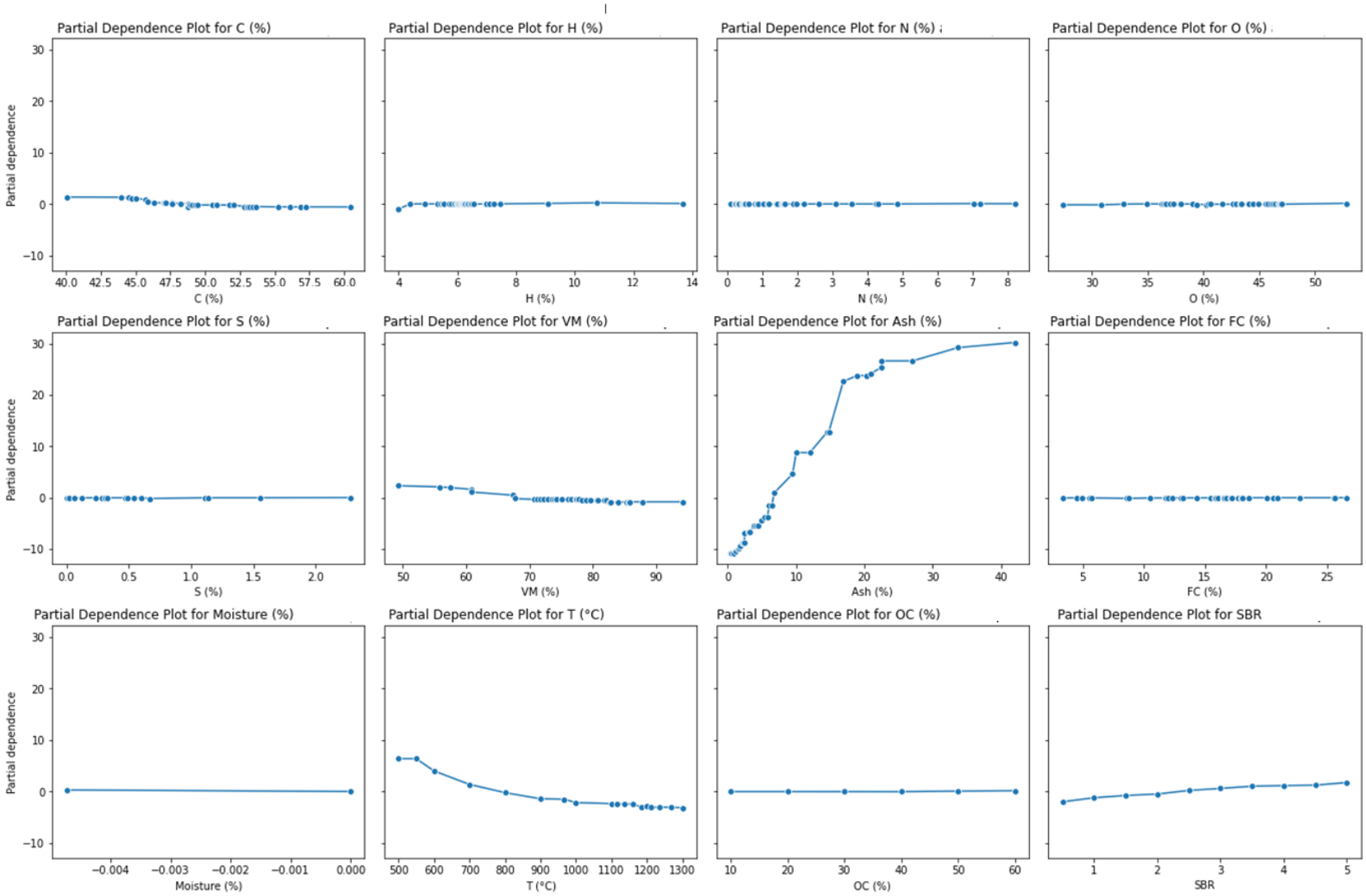

Figure 9: One variable partial dependence plot showing the impact of operating conditions and biomass properties on char yield during CLG

T is a crucial factor that influences the rate and extent of reactions involved in CLG. By increasing the temperature, the reaction rate is accelerated, and H_2 production is enhanced through the promotion of biomass gasification and cracking (Okolie et al., 2020). Furthermore, higher temperatures facilitate gasification reactions, reducing char formation. However, it is worth noting that elevated temperatures may also result in the production of undesirable by-products, such as soot, which can reduce process efficiency. This might explain the rapid decline in H_2 yield with increasing temperature (Figure 8) and the decreasing char yield with rising temperature (Figure 9).

The rise in char formation and H_2 yields with increasing ash content in biomass can be explained as follows. Ash content is known to promote char formation during gasification (Rodriguez Correa and Kruse, 2018). The presence of ash can catalyze gasification reactions, leading to an increase in char yield (Okolie et al., 2020). Minerals such as calcium, potassium, and magnesium in ash are known to enhance char formation during gasification. These minerals serve as sources of alkali and alkaline earth metals, which can catalyze gasification reactions by breaking down the biomass's hydrocarbons. Moreover, char formation during gasification results from incomplete biomass combustion, influenced by factors like temperature, residence time, and feedstock composition. The presence of ash can create a protective layer around char particles, preventing further gasification reactions and promoting char accumulation. Additionally, ash minerals can react with the biomass feedstock, resulting in the formation of solid carbonaceous residues.

The results from partial dependence plots presented in this study were closed aligned with literature values. For instance, Niu et al. (Niu et al., 2018) showed that the H_2 yield increased with the increase of temperature from 500 °C to 800 °C while the tar yield showed a contrary changing trend during CLG. However, at higher temperatures above 800 ^oC a decline in H_2 yield was observed. In the same study, H_2 yield elevated from 20.53% to 32.61% with the increase of SBR from 0 to 1.25. The authors noted some reactions responsible for the increase below:

$C + H_2O \rightarrow H_2 + CO \quad \Delta H = 131.7\ kJ/mol$ Equation 1

$CO + H_2O \rightarrow H_2 + CO_2 \quad \Delta H = -35.7\ kJ/mol$ Equation 2: water gas shift reaction

$CH_4 + H_2O \rightarrow 3H_2 + CO \quad \Delta H = 223.7\ kJ/mol$ Equation 3: Steam reforming reaction

$C + CO_2 \rightarrow 2CO \quad \Delta H = 171.3\ kJ/mol$ Equation 4: Boudouard reaction

Increasing the SBR would promote equations 1 – 3 leading to the formation of H_2. Also, increasing the temperature favours the endothermic reactions while it hinders the exothermic reactions.

The connection between individual factors and the output can be further clarified using the SHAP method. Figure 10 (a) demonstrates the absolute significance of features for H_2 yield, whereas Figure 10 (b) displays the feature importance for H_2 yield along with their detailed impacts. Figures 10 (c) and (d) displays the significant of features for the char yield. The greater the distance of a point from the baseline SHAP value of zero, the more it influences the output. In this manner, the association between a feature and the SHAP value (and consequently, the predicted output) can be better comprehended.

From figure 10 (a) and (b) it is evident that increasing the temperature led to a rise in H2 yield at first after which there is a significant decrease. Also, elevating the SBR promotes H_2 formation. According to figure 11 (b), elevating the ash content led to an increase in char formation. This observation further confirms the findings from the partial dependence plot. Features found in the middle had a relatively lower impact on the output. Such features include OC, O, moisture, S and N content of the biomass.

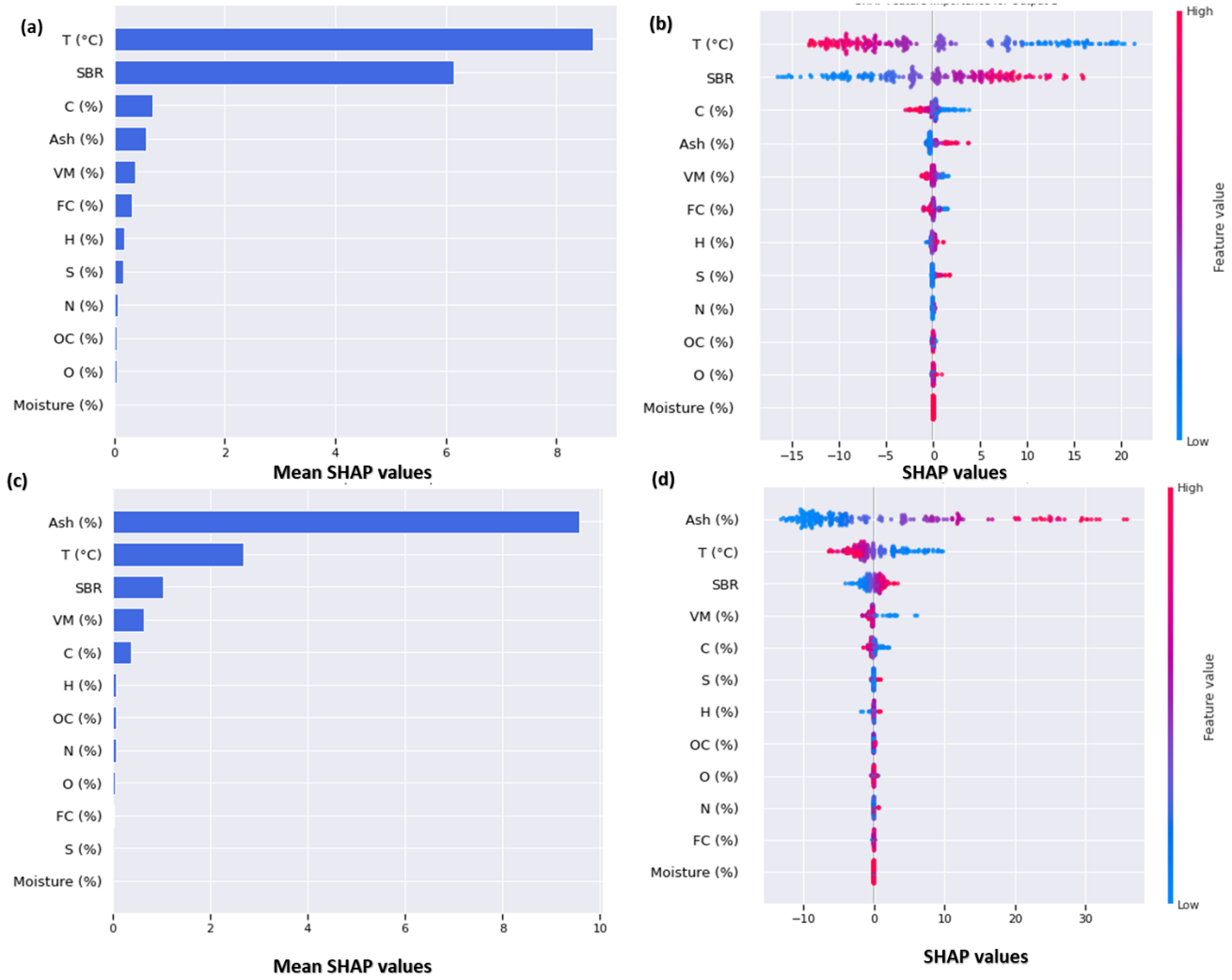

Figure 10: Representation of the feature importance analysis by the SHAP method fir the prediction of (a) and (b) for H_2 yield while (c) and (d) for char yield. Figures (a) and (c) shows the absolute importance of features while figure (c) and (d) displays the impact of individual predictions on the overall importance scores. High feature values are represented in red, while low feature values are depicted in blue.

3.4 Discussion and outlook

Although the present study applied the feature importance and SHAP analysis as a means of interpreting the ML behaviour and prediction of H_2 and char yield during CLG, there are still relatively few studies in this field. While ML methods have often been referred to as black box due to this inability to adequately explain the relationship between features and output, several interpretable methods have been implemented. For instance, some researchers have adopted the Feature importance and partial dependence plot to explain the impact biomass properties and operating conditions on H_2 yield during hydrothermal gasification (Zhao et al., 2021). It should be mentioned that the partial dependence plots are very useful and easy to implement especially when a straightforward relationship is required between individual feature and output. However, there might be some limitations especially when the data requires multiple input and multiple output such as the one implemented in the present study. Also, the partial dependence methods do not have the ability to capture confounding factors (Ascher et al., 2022).

Some researchers have also adopted the permutation feature importance due to the ease of implementation with different model types (Yuan et al., 2021). However, this method is limited by potential non- linear relationships, may overestimate, or underestimate the importance of individual features in the presence of multicollinearity and often sensitive to noise in dataset. In this study a combination of SHAP (global) and PDP (local) methods were used to provide a detailed overview of the ML model prediction capability. The SHAP method helps to identify the mean behaviour of the ML model as well as the trends inherent between the features and output. In contrast, the PDP provided an understanding of which factors might affect the H_2 and char yield. This could help researchers in identifying what factors to vary in optimizing H_2 and char formation. Importantly, the trained ML model could serve as valuable resources thereby saving a lot of experimental cost.

There are several limitations that should be addressed in subsequent study with interpretable ML models. The success of CLG is dependent on the oxygen carriers, therefore ML model could be adopted for the screening of the oxygen carriers. However, this would require a lot of experimental data. Incorporating lignocellulosic biomass composition in the model as input variable would help understand how biomass properties influence the H_2 and char yield. However, this should be part of future work where in depth experimental and analytical characterization would be used to collect reasonable data for performing the ML analysis. Optimizing the heat transfer between the oxidation and reduction reactor is also another area that requires further attention as part of future studies.

4. Conclusions

The present study developed a novel ML algorithm based on GBR, RF and SVM for the prediction of H_2 and char yield during CLG. While there are relatively few experimental studies related to CLG, a combination of process modelling in Aspen plus and experimental data in literature were used to generate input and output dataset. The dataset comprises of input features such as carbon (C), hydrogen (H), Nitrogen (N), oxygen (O), Sulphur (S), Moisture content, Volatile matter (VM), Ash content (Ash), Fixed carbon (FC), gasification temperature (T), oxygen carrier quantity measured as mass ratio to biomass (OC) and Steam to biomass ratio (SBR)). The output variables are H_2 and char yield. Among the ML algorithms, GBR was found to be the most promising in terms of the R^2, MAE and RMSE values. Additionally, GBR outperformed other models for the prediction of H_2 and char yield. PDP analysis showed that H_2 yield is not affected by O, N and moisture content of the biomass. Also, the oxygen carrier quantity measured as mass ratio to biomass (OC) does not influence H_2 yield. In contrast, H_2 yield is slightly affected by a change in C, H, S, VM, Ash, and FC content of the biomass feedstock. H_2 yield is strongly impacted by T and SBR. An increase in SBR led to a rapid increase in H_2 yield. H_2 yield increased up to a maximum of 700 ^oC then declined rapidly as the temperature rose to 1300 ^oC. On the contrary, biomass with higher ash and FC content are predicted to produce immense amount of char. The

findings presented in this study could help researchers in the optimization of CLG process and save experimental cost.

Data availability

All data can be found in the GitHub link***:*** ***https://github.com/judebebo32/ML_CLG***

CRediT authorship contribution statement