BioEnergy Research
https://doi.org/10.1007/s12155-024-10773-0

Machine Learning Approach for Predicting Hydrothermal Liquefaction of Lignocellulosic Biomass

Tossapon Katongtung[1,2] · Sanphawat Phromphithak[1,2] · Thossaporn Onsree[3] · Nakorn Tippayawong[2]

Received: 12 February 2024 / Accepted: 18 May 2024

Abstract
Hydrothermal liquefaction (HTL) of lignocellulosic biomass has gained attention as a promising technology for the production of biofuels and other value-added products. HTL process optimization is complex and involves various parameters such as reaction time, temperature, and pressure. In recent years, machine learning (ML) approaches have been adopted as a tool to optimize and predict the HTL process performance. The purposes of this study were to investigate the ML-based prediction of bio-crude yield (BCY) and their higher heating values (HHVs) from HTL of lignocellulosic biomass and to elucidate the relationship of features affecting the output of interest. Pre-processing and normalization were applied to a dataset of 215 data points with 17 input features. Feature selection using the Shapley value method identified key predictors. ML models including multilayer perceptron, kernel ridge regression, random forest, and extreme gradient boosting (XGB) were trained and evaluated. XGB algorithm shows superior performance in predicting the yields and their calorific values to within 5–8% of experimental values. Temperature was the most influential feature for both BCY and HHV prediction accounting for about 30%, followed by other feedstock and operational characteristics. In addition, a user interface was presented for ease of use in the ML modeling.

Keywords Biomass conversion · Biofuels · Clean energy · Data driven engineering · Hydrothermal processing

Introduction

Global warming, the gradual increase in the Earth's surface temperature due to the emission of greenhouse gases, has become one of the most pressing issues facing humanity [1, 2]. In recent years, scientists have been exploring various methods to reduce greenhouse gas emissions and slow down the pace of global warming [3–5]. One promising area of research is the use of biomass hydrothermal liquefaction (HTL) as a means of converting organic matter into a renewable energy source [6, 7]. Biomass HTL is a process that involves subjecting biomass, such as agricultural and forestry residues, to high temperatures and pressure in the presence of water [8, 9]. The process results in the conversion of the biomass into a liquid bio-oil, which can be used as a substitute for fossil fuels. Biomass HTL has several advantages over other forms of bioenergy production, including its ability to produce a high-quality fuel that is compatible with existing infrastructure [10].

Current research on biomass HTL is focused on improving the efficiency of the process, increasing the yield of bio-oil, and optimizing the quality of the final product. Researchers are also exploring the potential of using different types of biomass and developing new catalysts to enhance the reaction. Additionally, there is a growing interest in the use of biomass HTL as a means of reducing greenhouse gas emissions from various industries, such as agriculture and forestry. The use of biomass HTL shows great promise as a sustainable and renewable energy source that could help mitigate the effects of global warming. As research in this area continues, it is hoped that the technology will become increasingly efficient and economically viable, paving the

✉ Nakorn Tippayawong
nakorn.t@cmu.ac.th

1 Graduate PhD Degree Program in Energy Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand

2 Department of Mechanical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

3 Department of Chemical Engineering, College of Engineering and Computing, University of South Carolina, Columbia, SC, USA

Published online: 24 May 2024

way for a cleaner, more sustainable future [11, 12]. HTL is a promising technology for converting biomass into liquid fuels and chemicals. However, the complex and non-linear nature of the HTL process makes it challenging to model and optimize using traditional methods.

Machine learning (ML) has emerged as a powerful tool for modeling and optimization of complex processes, and recent research has shown that machine learning can be applied to HTL to improve its efficiency and sustainability [13]. ML models can learn from large datasets of HTL process variables and performance metrics to accurately predict process behavior and optimize operating conditions [14, 15]. Current research in ML for HTL is focused on developing models that can accurately predict HTL process performance and optimize process parameters, such as temperature, pressure, and residence time. This involves developing ML algorithms that can handle the large and complex datasets generated during HTL experiments and simulations and that can account for the non-linear behavior of the HTL process. In addition, ML models can be used to predict the chemical and physical properties of HTL products, which can be used to optimize downstream processing and identify potential applications for the resulting fuels and chemicals. ML can also be used to predict the environmental impacts of the HTL process and to identify strategies for reducing its carbon footprint. The application of ML to HTL modeling is an active area of research and has the potential to significantly improve the efficiency and sustainability of the process. As the technology matures, it is likely that ML will play an increasingly important role in the development and optimization of HTL and other biomass conversion technologies.

Previous studies have been conducted by various researchers on the application of ML in the HTL process. For example, Zhang [16] studied ML prediction and optimization of bio-oil production from HTL of algae. Shafizadeh [17] adopted ML to predict and optimize the HTL of biomass. Katongtung et al. [18] predicted bio-crude yields and higher heating values of products from HTL of wet biomass and wastes. The aforementioned studies share common feature inputs but differ in the number and type of data sets utilized. These studies demonstrate that ML models exhibit high predictive accuracy and can effectively capture the complex relationships among the variables. Moreover, the models can identify various factors that mutually affect each other.

Hence, the objective of this study is to investigate the effectiveness of ML in predicting the bio-crude yield and calorific value of the high-throughput screening process, utilizing unique biological property input features rarely previously utilized in other studies. Specifically, the input features in this study consist of cellulose, hemicellulose, and lignin, and their impact on the output is analyzed using the Shapley value method.

Materials and Methods

Data Collection and Pre-processing

The process of data collection and pre-processing involved several stages. Initially, raw data was collected from the HTL process, and the relevant input features, such as cellulose, hemicellulose, and lignin, were identified and extracted. Subsequently, the data underwent pre-processing procedures to identify and eliminate any inconsistencies, outliers, or missing values. The pre-processed data was then normalized and split into training and testing sets to ensure model accuracy and prevent overfitting. Finally, feature selection techniques were employed to identify the most significant input features, using the Shapley value method, for predicting the bio-crude yield and calorific value.

The dataset utilized in this study consists of lignocellulosic biomass, with a total of 215 data points, comprising both dependent variables (bio-crude yield and calorific value) and their corresponding independent features. The data was collected from existing publications and is further detailed in supplementary material Table S1. Table 1 presents a comprehensive list of 17 input features, derived from feedstock characteristics with dry basis and operating conditions. The table provides the name of each feature, along with its associated numerical and statistical values.

Machine learning (ML) algorithms often require input data to be standardized and normally distributed for optimal performance. To address this issue in the present study, standardization of all 17 input features was conducted using Yeo-Johnson's transformation [19], as shown in Eq. 1:

$$\hat{x}_{f,g} = \begin{cases} \frac{[(\tilde{x}_{f,g}+1)^{\lambda}-1]}{\theta}, if\theta \neq 0, \tilde{x}_{f,g} \geq 0 \\ ln(\tilde{x}_{f,g}+1), if\theta = 0, \tilde{x}_{f,g} \geq 0 \\ -\frac{[(-\tilde{x}_{f,g}+1)^{2-\theta}-1]}{(2-\theta)}, if\theta \neq 2, \tilde{x}_{f,g} < 0 \\ -ln(-\tilde{x}_{f,g}+1), if\theta = 2, \tilde{x}_{f,g} < 0 \end{cases} \quad (1)$$

Here, $\hat{x}$ represents the transformed data obtained from the original data $\tilde{x}$ for a particular feature g, with θ being the transformation factor derived from maximum likelihood estimation. The subscript f ranges from 1 to N, representing the individual data points for each feature. Following this transformation, the standardized data was further normalized using Eq. 2:

$$x_{f,g} = \hat{x}_{f,g} - \bar{x}_g/\sigma_g \quad (2)$$

Here, x represents the normalized data obtained from $\hat{x}$, while $\bar{x}$ and σ represent the mean and standard deviation values of each input feature j, respectively [20].

Table 1 Set of input features or variables

Variable	Abbreviations used in modeling	Value	Mean	Standard deviation
Input feature				
Feedstock characteristic (on dry basis)				
Biological property				
Cellulose (% w/w)	Cel	6.4–73.6	28.9	15.3
Hemicellulose (% w/w)	Hemi_Cel	4.6–37.5	19.4	8.9
Lignin (% w/w)	lig	6.1–38.1	23.1	9.6
Elemental property				
Carbon content (% w/w)	C	35.3–50.7	44.1	3.5
Hydrogen content (% w/w)	H	4–7.5	5.4	1.2
Nitrogen content (% w/w)	N	0–5.9	2.2	2.2
Oxygen content (% w/w)	O	24–59	44	8.2
H-to-C atomic ratio (-)	H/C	1–2.3	1.5	0.4
O-to-C atomic ratio (-)	O/C	0.4–1.3	0.8	0.1
Ash (% w/w)	Ash	0.17–27.9	14.3	8.9
Operating condition				
Residence time (min)	Time	0–240	59.8	53
Temperature (°C)	T	200–420	317.7	45.8
(Initial) pressure (MPa)	P	0.1–20	1.1	3.6
Reactor size (mL)	RS	8.8–2000	235.4	345.3
Biomass loading (g)	BL	0.88–50	13.3	14.1
Water (mL)	W	7.9–700	88.4	114.1
Solvent-to-BL ratio (-)	S/BL	2.3–20	7.8	2.9
Output target				
Bio-crude yield (% w/w)	BCY	7.8–45.4	25.2	8.5
Higher heating value (MJ/kg)	HHV	15.7–40.9	32.5	5.4

The definition of bio-crude may vary in the literature, leading to different methods of calculating bio-crude yields. In this study, bio-crude is defined as the organic fraction extracted/separated from the aqueous-phase (water-soluble) fraction using organic solvents, such as acetone or dichloromethane. To ensure consistency in calculations, the bio-crude yield ($\widetilde{y}_{\mathrm{BCY}}$) was recalculated using Eq. 3:

$$\widetilde{y}_{\mathrm{BCY}} = m_{\mathrm{BCY}}/M \tag{3}$$

Here, m_{BCY} represents the mass of bio-crude, while M represents the initial mass of feedstock on a dry basis. Additionally, the higher heating value (HHV) of the bio-crude was recalculated using Eq. 4 [18]:

$$\mathrm{HHV} = 0.338\mathrm{C} + 1.428(\mathrm{H} - \mathrm{O}/8) \tag{4}$$

The variables C, H, and O represent the carbon, hydrogen, and oxygen contents of the bio-crude (in w/w). Subsequently, the 2-target output, $\tilde{y}$, was scaled and normalized using their respective maximum values, denoted as $\acute{y}$.

$$y_{t,h} = \widetilde{y}_{t,h}/\acute{y}_h \tag{5}$$

Here, y represents the normalized data of an output h, which refers to the bio-crude yield (BCY) and higher heating value (HHV) of the bio-crude.

ML Model Development

This study utilized advanced ML algorithms, including multilayer perceptron (MLP), kernel ridge regression (KRR), random forest (RF), and extreme gradient boosting (XGB), which were implemented using the scikit-learn libraries [21] in the Python environment. The XGB library was obtained from open-source code developed by Chen and Guestrin [22]. All codes were run on a computer with a MacBook processor running at 1.1 GHz Dual-Core Intel Core m3. To evaluate the performance of these algorithms on a dataset consisting of 17 input features and 2 target outputs, the k-fold cross-validation method was employed. This involved randomly dividing the entire dataset into k groups or folds, where each fold was used as a test dataset, while the remaining k-1 folds were used as a training dataset. This process was repeated k times to ensure that each fold was used as a test dataset once. In this study, 10 folds (i.e., $k = 10$) were used to ensure that each fold was

large enough to represent the statistical properties of the entire dataset [23]. By using the *k*-fold cross-validation method, the prediction accuracy of each algorithm was obtained in an unbiased manner since the entire dataset was used for training and testing. This allowed for a fair comparison of the performance of the algorithms, and the results obtained were used to identify the algorithm that performed best on the given dataset. This study employed advanced ML algorithms implemented using the scikit-learn and XGB libraries in Python. The *k*-fold cross-validation method was used to evaluate the performance of these algorithms on a dataset with 17 input features and 2 target outputs. The results obtained were unbiased and were used to identify the best-performing algorithm for this particular dataset.

To evaluate the accuracy of the model's predictions, two metrics were used: the coefficient of determination (R^2) and the root-mean-square error (RMSE). R^2 is a statistical measure used in regression models to determine the proportion of variance in the output parameters that can be explained by the input parameters [24]. On the other hand, RMSE is a measure of the differences between the real values and their corresponding predicted values. In general, a higher R^2 and a lower RMSE indicate better predictive performance [25, 26].

The equations for calculating R^2 and RMSE are provided below as Eqs. (6) and (7), respectively, following the work of [19]:

$$R_a^2 = 1 - \sum_{1}^{n} (y_{a,b} - f_{a,b}(x))^2 / \sum_{1}^{n} (y_{a,b} - \bar{y}_b)^2 \tag{6}$$

$$\mathrm{RMSE}_a = \left(\frac{1}{n} \times \sum_{1}^{n} (y_{a,b} - f_{a,b}(x))^2\right)^{1/2} \tag{7}$$

where y is the observed value, $f(x)$ is the predicted value, $\bar{y}$ is the mean of the observed values, and N is the number of observations in the dataset. In addition to R^2 and RMSE, the normalized root-mean-square error (NRMSE) was also calculated in this study to account for any differences in the scale of the test datasets used in each fold. NRMSE is defined as the ratio of RMSE to the mean value of the observed data, and it is calculated using Eq. 8:

$$\mathrm{NRMSE} = \mathrm{RMSE}/\bar{y} \tag{8}$$

where $\bar{y}$ is the mean value of the observed data. The use of NRMSE allows for a more accurate comparison of the performance of the different models, as it takes into account the variability in the scale of the test datasets.

The equation for calculating NRMSE was derived from the work of [19] and was used in this study as an additional metric for assessing the performance of the models.

Selected ML Algorithm

In this study, four advanced ML algorithms, namely, MLP, KRR, RF, and XGB, were used to analyze the HTL datasets. Each algorithm has its own mathematical principles and parameters, which were optimized using a full-factor grid search technique in combination with a nested 10-fold cross-validation method [27, 28]. The optimized hyperparameters of all the algorithms are summarized in supplementary material Table S2.

An MLP is a type of fully connected, feedforward artificial neural network (ANN) that has found widespread use in various applications due to its ability to learn complex relationships between input and output data. While the term MLP is sometimes used interchangeably with the more general term feedforward ANN, it more strictly refers to ANNs composed of multiple layers of perceptrons with threshold activation. The MLP architecture consists of at least three layers of nodes, including an input layer, one or more hidden layers, and an output layer. The nodes in each layer, except for the input layer, are modeled as neurons that use a non-linear activation function. The multiple layers and non-linear activation functions employed by MLPs enable them to learn complex and non-linear relationships in data, making them more powerful than linear perceptrons in handling data that is not linearly separable [29].

The backpropagation algorithm is a supervised learning technique that is commonly used to train MLPs. During training, the algorithm adjusts the weights between nodes in the network by iteratively computing the gradient of the loss function with respect to the weights and then updating the weights in the opposite direction of the gradient. This process is repeated until the network converges to a satisfactory solution. MLP is a powerful and widely used class of ANN that utilizes multiple layers and non-linear activation functions to enable the learning of complex relationships in data. Their ability to handle non-linearly separable data has made them particularly useful in various applications, and the backpropagation algorithm is commonly used to train them [30].

KRR is a machine learning algorithm that combines the principles of ridge regression and classification with the kernel trick. Specifically, KRR performs linear least squares with L2 norm regularization in the space induced by a given kernel and the input data. The kernel trick allows KRR to implicitly transform the input data into a higher-dimensional feature space, where a linear function can be used to model the relationship between the input features and the output variable. This is achieved by using a kernel function to compute the similarity between pairs of data points in the input space and then projecting these similarities into the higher-dimensional feature space. By using a non-linear kernel function, KRR is able to model non-linear relationships

between the input features and output variables. This corresponds to a non-linear function in the original input space. The regularization parameter in KRR controls the balance between model complexity and generalization performance and can be tuned to optimize the model for a specific dataset. KRR is a powerful machine learning algorithm that combines the strengths of ridge regression, classification, and the kernel trick to learn a non-linear function that can model complex relationships between the input features and output variables. The choice of kernel function and regularization parameter can significantly impact the performance of the model and must be carefully tuned to optimize performance on a given dataset [31, 32].

RF regression is a widely used supervised learning algorithm that leverages an ensemble learning method for regression tasks. Ensemble learning is a powerful technique that involves combining the predictions of multiple machine learning algorithms to improve the accuracy and generalization performance of the resulting model. RF algorithm constructs a collection of decision trees, where each tree is built using a random subset of the input data and a random subset of the input features. During prediction, the algorithm generates predictions from each decision tree and then aggregates them to produce a final prediction that is more accurate and less prone to overfitting than a single decision tree. The benefits of the RF algorithm include its ability to handle high-dimensional and complex data, its resilience to overfitting, and its ability to provide information on feature importance, making it an effective tool for feature selection. Moreover, the RF algorithm can handle both regression and classification tasks [33].

RF is a powerful machine learning algorithm that utilizes an ensemble learning method to combine the predictions of multiple decision trees, enabling it to model complex relationships between input features and output variables with high accuracy and robustness. The algorithm's ability to handle high-dimensional data, its resilience to overfitting, and its ability to provide information on feature importance make it a valuable tool for a wide range of regression tasks [34].

XGB is a popular supervised learning algorithm that uses an ensemble learning method to improve the accuracy of predictions. XGB is an extension of the gradient boosting algorithm, which iteratively trains a sequence of weak models and combines their predictions to make a final prediction. The XGB algorithm incorporates several advanced techniques to improve performance and scalability. It uses a technique called gradient boosting to iteratively add weak models to the ensemble and a loss function to optimize the model during training. The algorithm also employs regularization techniques, such as L1 and L2 regularization, to prevent overfitting. One key feature of XGB is its ability to handle missing data and categorical variables by encoding them in a unique way. This allows the algorithm to handle a wide range of data types and to extract meaningful features from complex datasets [35].

XGB has been used successfully in a variety of applications, including natural language processing, image recognition, and time series forecasting. It is also highly scalable, making it well suited for large-scale datasets. XGB is a powerful machine learning algorithm that uses an ensemble learning method to improve the accuracy of predictions. The algorithm incorporates several advanced techniques, including gradient boosting, regularization, and unique encoding of missing and categorical data. XGB has shown strong performance across a range of applications and is highly scalable, making it a valuable tool for a wide range of machine learning tasks [36].

Feature Evaluation and Interpretation

The ML models were optimized by tuning their hyperparameters using a full grid search methodology in conjunction with 10-fold cross-validation. To determine the relative importance of features in predicting the HHV and BCY, a built-in function of each algorithm was utilized to extract the relevant features. These features were subsequently ranked based on their significance in accurately predicting the desired outcomes. The Shapley method was employed to investigate the impact of input features on the output. In the context of machine learning, it is used to measure the relative importance of input features in predicting the output. The Shapley value for each feature represents the average marginal contribution of that feature across all possible feature combinations, encapsulating its relative importance within the model's decision-making process. By meticulously analyzing the Shapley values attributed to each input feature, researchers and practitioners can glean valuable insights into the nuanced impact and significance of individual features on shaping the final output of a machine learning model. This comprehensive understanding enables informed feature selection, model refinement, and deeper interpretability, thereby facilitating enhanced performance and transparency in machine learning systems [37]. Therefore, in this study, the Shapley method was employed as a tool for exploring and understanding the impact of input features on the output [38].

Results and Discussion

Model Selection and Accuracy

Table 2 presents the predictive accuracy results of the XGB, RF, MLP, and KRR models, each optimized using their respective best hyperparameters. The accuracy of each

Table 2 Prediction accuracy in terms of R^2, RMSE, and NRMSE from 10-fold cross-validation of XGB, RF, MLP, and KRR algorithms for all three cases of input features

Model	BCY			HHV		
	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE
XGB	0.8861 (0.0662)	1.9936 (0.5606)	0.0822 (0.0313)	0.8286 (0.1581)	1.6586 (0.8304)	0.0565 (0.0250)
RF	0.8306 (0.1136)	2.3986 (0.7545)	0.0985 (0.0393)	0.7049 (0.2877)	2.2456 (1.2261)	0.0768 (0.0376)
MLP	0.8103 (0.1449)	2.7617 (1.1743)	0.1115 (0.0446)	0.7681 (0.1573)	2.4247 (1.0241)	0.0840 (0.0331)
KRR	0.7887 (0.2421)	3.0816 (1.6608)	0.1261 (0.0694)	0.8014 (0.2461)	2.0068 (1.4818)	0.0684 (0.0465)

Note: The values reported herein were obtained by averaging the results obtained from a ten-fold cross-validation analysis. The number presented in parentheses corresponds to the standard deviation of the calculated values

model was evaluated using appropriate performance metrics, such as NRMSE, RMSE, and R^2. The presented results provide insights into the performance of these models in predicting the desired outcome and can help in selecting the best model for a given application.

The XGB model exhibited the best overall performance, with an R^2 value of 0.8861 for BCY and approximately 0.8286 for HHV. The RMSE was 1.9936 for BCY and 1.6586 for HHV. In contrast, the RF model achieved an R^2 value of 0.8103 for BCY and 0.7049 for HHV, with an RMSE of 2.3986 for BCY and 2.2456 for HHV. The MLP model yielded an R^2 value of 0.8103 for BCY and 0.7681 for HHV, with an RMSE of 2.7617 for BCY and 2.4247 for HHV. Furthermore, the KRR model exhibited the lowest accuracy among the models tested in this study, with an R^2 value of 0.7887 for BCY and 0.7681 for HHV. The RMSE was 3.0816 for BCY and 2.0068 for HHV. It should be noted that the XGB model demonstrated the highest level of accuracy and the lowest error rate in this study.

To provide further insight into the prediction performance of the XGB model, Fig. 1 illustrates the scatter plots of predicted values versus actual (test) values for both BCY and HHV in both the training and testing phases. The black trend line indicates the positions where the predicted values are equivalent to the test values. Meanwhile, the green band represents a 10% error range, and the blue band denotes a 20% error range. This visualization enables a comparison of the performance of each model in accurately predicting both BCY and HHV of bio-crude. This figure allows for a visual comparison of the predicted values to the actual

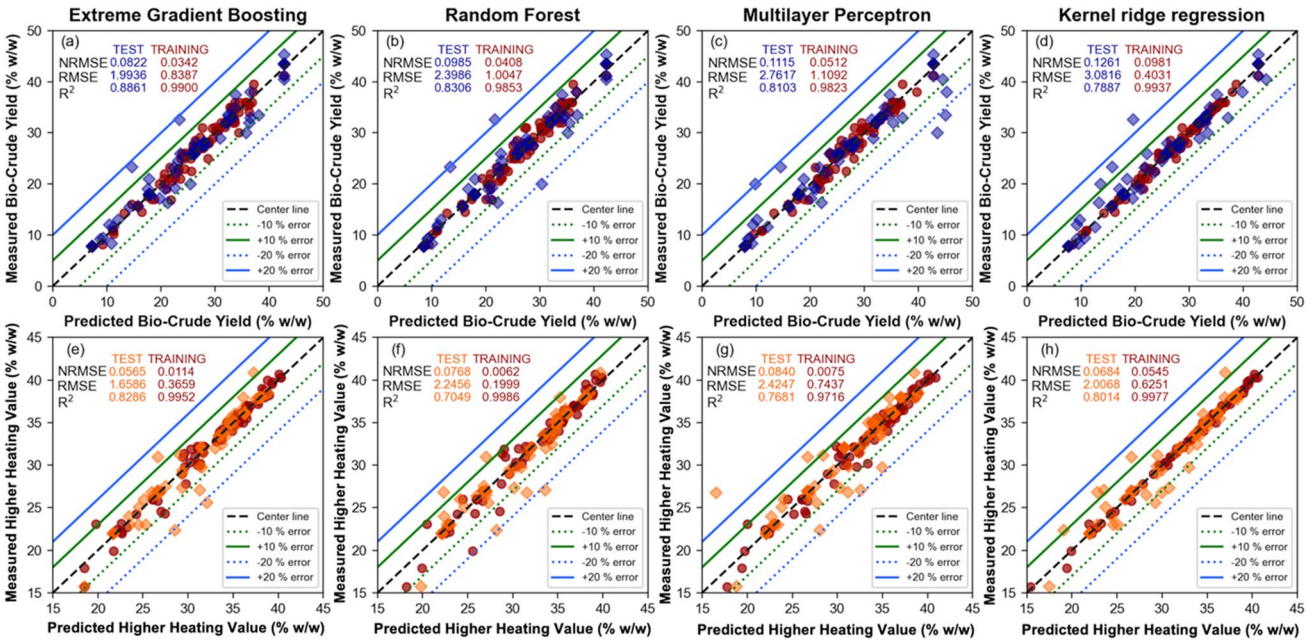

Fig. 1 depicts the distribution of prediction data against the test dataset in ten-fold cross-validation of both BCY and HHV using four different machine learning models, namely, XGB (**a**, **e**), random forest (RF) (**b**, **f**), multilayer perceptron (MLP) (**c**, **g**), and kernel ridge regression (KRR) (**d**, **h**)

values, providing a more intuitive understanding of the model's performance. It is evident that the prediction accuracy of the model was higher in the training phase compared to the testing phase. For BCY, the training case had R^2 values within 0.98–0.99 and NRMSE between 0.03 and 0.09, while the test case provided an R^2 at 0.78–0.88 and an NRMSE of 0.08–0.12. For HHV, the training case offered an R^2 within 0.97–0.99 and an NRMSE between 0.00 and 0.05. The test case had an R^2 of 0.70–0.82 and an NRMSE of 0.05–0.08. The XGB algorithm demonstrates promising effectiveness in facilitating the development of predictive models for the yield and HHV of bio-crudes obtained from HTL of lignocellulosic biomass, utilizing input variables derived from feedstock characteristics, such as biological and elemental properties, as well as operating conditions. The level of precision achieved by the XGB model in predicting the yield and HHV of bio-crudes from HTL of lignocellulosic biomass is comparable to that of models specifically developed for certain biomass types or other ML models applied to different biomass conversion techniques. Therefore, the XGB algorithm is a valuable tool for facilitating model development and optimizing the conversion of lignocellulosic biomass into bio-crude.

SHAP Summary Plot

SHAP summary plot is a function in the SHAP (SHapley Additive exPlanations) library used for visualizing the summary of Shapley values for a set of features in a machine learning model. The plot displays the importance and impact of each feature in the model predictions and how they contribute to the final outcome. The function generates a bee swarm where features are ranked based on their importance, and the magnitude and direction of the Shapley values are shown using colored points. Positive Shapley values indicate that the corresponding feature increases the prediction, while negative values imply the opposite. The plot can help identify the most relevant features and understand the relationship between them and the model's output.

Figure 2 illustrates the sequence of the impact of input variables on outcomes. Specifically, Fig. 2a displays the order of effect of input variables on BCY, highlighting that the input feature with the most significant impact on BCY is T, followed by H, N, Lig, and W. This order of effect is primarily attributed to the characteristics of the feedstock. On the other hand, Fig. 2b exhibits the order of effect of input variables on HHV. In this case, the feature with the greatest influence on HHV is T, followed by ash, time, Hemi_cel, and Cel. These variables are primarily affected by the operating conditions. The importance of the five features that have the greatest impact on both BY and HHV can be described as follows, while other less important features are shown in Supplementary Material Table S3.

Temperature

Temperature is one of the most important variables that affect the BCY and HHV in the HTL process of lignocellulosic biomass. The effects of temperature on the HTL process can be summarized as follows:

Bio-crude yield (BCY): The BCY increases with increasing temperature. This is because the higher temperatures promote the breakdown of complex biomass molecules into simpler compounds, which increases the bio-crude yield. However, excessive heating can also cause degradation of the bio-crude, resulting in a lower yield. Generally, the optimal temperature range for maximum bio-crude yield is between 250 and 350 °C.

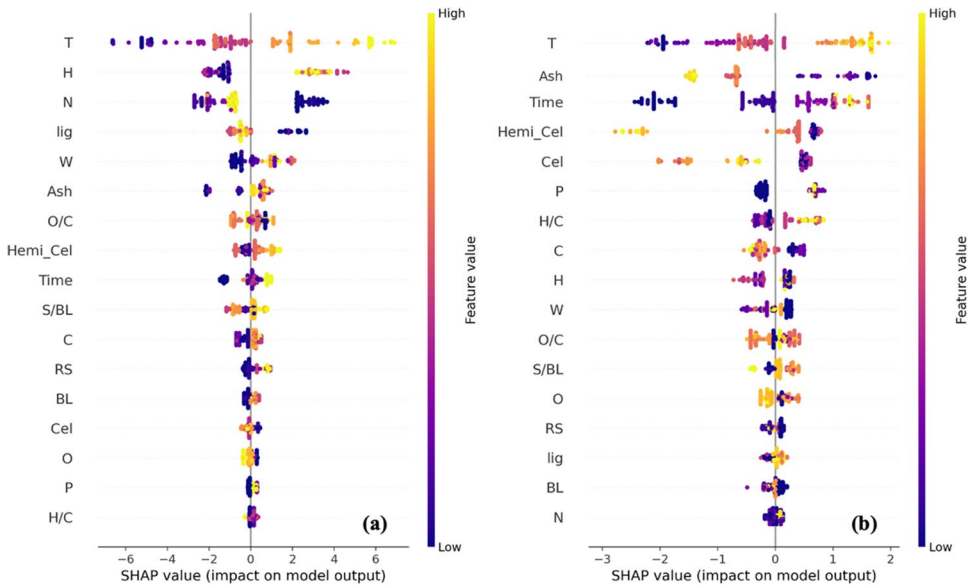

Fig. 2 Feature importance by SHAP value for **a** BCY and **b** HHV

Higher heating value (HHV): The HHV of the bio-crude also increases with increasing temperature. This is because the higher temperatures lead to the production of more stable and energy-dense compounds, such as aromatics, which increases the HHV. However, excessive heating can also cause the formation of less stable compounds, resulting in a lower HHV. Generally, the optimal temperature range for maximum HHV is between 300 and 400 °C.

The temperature variables are therefore crucial in the HTL process as they can impact the bio-crude yield and HHV, which are important factors for the economic viability of the process. Proper control of temperature variables can also improve the quality of the bio-crude, making it more suitable for use in a variety of applications, including fuel and chemical production [39–41].

Ash Content

Ash content is an important variable to consider in the HTL process of lignocellulosic biomass as it can affect the BCY and HHV in several ways. The effects of ash variables on the HTL process can be summarized as follows:

BCY: The presence of ash in the biomass feedstock can reduce the BCY by interfering with the thermal decomposition of biomass compounds. This is because ash can act as a heat sink, which reduces the temperature in the reaction zone and inhibits the formation of bio-crude. In addition, ash can catalyze unwanted reactions that lead to the formation of tars and other undesirable compounds, further reducing the BCY.

HHV: The presence of ash in the bio-crude can reduce its HHV by diluting the energy content of the bio-crude. This is because ash is mostly composed of inorganic materials that do not contribute to the energy content of the bio-crude. In addition, ash can also cause fouling and corrosion of equipment, which can increase maintenance costs and reduce the overall efficiency of the HTL process.

Therefore, it is important to minimize the ash content in the biomass feedstock to maximize the bio-crude yield and HHV in the HTL process. This can be achieved by using clean biomass feedstocks with low ash content or by pretreating the biomass to remove or reduce the ash content before HTL. Proper monitoring and control of ash variables can also improve the overall efficiency and economic viability of the HTL process [42].

Time

Reaction time is a significant consideration when undertaking the HTL process of lignocellulosic biomass as it can affect the BCY and HHV in several ways. The effects of time variables on the HTL process can be summarized as follows:

BCY: The BCY increases with increasing reaction time up to a certain point, beyond which it starts to plateau or decrease. This is because the longer reaction time allows for a more complete breakdown of the complex biomass molecules into simpler compounds, which increases the BCY. However, excessive reaction time can also cause degradation of the bio-crude, resulting in a lower yield. Generally, the optimal reaction time for maximum BCY is between 30 and 60 min.

HHV: The HHV of the bio-crude also increases with increasing reaction time up to a certain point beyond which it starts to plateau or decrease. This is because the longer reaction time allows for a more complete conversion of the biomass into energy-dense compounds, such as aromatics, which increases the HHV. However, excessive reaction time can also cause the formation of less stable compounds, resulting in a lower HHV. Generally, the optimal reaction time for maximum HHV is between 30 and 60 min.

The time variables are therefore crucial in the HTL process as they can impact the bio-crude yield and HHV, which are important factors for the economic viability of the process. Proper control of time variables can also improve the quality of the bio-crude, making it more suitable for use in a variety of applications, including fuel and chemical production. In addition, shorter reaction times can also increase the overall efficiency of the HTL process, reducing the capital and operating costs associated with longer reaction times [43].

Hydrogen Content

The variable of hydrogen content is an important consideration in the HTL process of lignocellulosic biomass as it can affect the BCY and HHV in several ways. The effects of hydrogen content variables on the HTL process can be summarized as follows:

BCY: The BCY increases with increasing hydrogen content up to a certain point, beyond which it starts to plateau or decrease. This is because the higher hydrogen content promotes the hydrogenation of the biomass compounds, which increases the BCY. However, excessive hydrogenation can also cause degradation of the bio-crude, resulting in a lower yield. Generally, the optimal hydrogen content for maximum BCY is between 10 and 20 wt%.

HHV: The HHV of the bio-crude also increases with increasing hydrogen content up to a certain point beyond which it starts to plateau or decrease. This is because the higher hydrogen content leads to the production of more stable and energy-dense compounds, such as aromatics, which increases the HHV. However, excessive hydrogenation can also cause the formation of less stable compounds, resulting in a lower HHV. Generally, the optimal hydrogen content for maximum HHV is between 10 and 20 wt%.

The hydrogen content variables are therefore crucial in the HTL process as they can impact the bio-crude yield and

HHV, which are important factors for the economic viability of the process. Proper control of hydrogen content variables can also improve the quality of the bio-crude, making it more suitable for use in a variety of applications, including fuel and chemical production. In addition, hydrogenation can also increase the overall efficiency of the HTL process, reducing the capital and operating costs associated with longer reaction times or higher temperatures [44].

Nitrogen Content

The variable of nitrogen content is important to assess in the HTL process of lignocellulosic biomass, as it can affect BCY and HHV in various ways. The effects of nitrogen content variables on the HTL process can be summarized as follows:

BCY: The BCY decreases with increasing nitrogen content. This is because nitrogen compounds are less reactive than other biomass components and can act as catalyst poisons, reducing the effectiveness of the HTL process. Additionally, nitrogen can also lead to the formation of tar-like substances, which can clog the reactor and decrease BCY. Therefore, proper control of nitrogen content is important for maximizing BCY.

HHV: The HHV of the bio-crude decreases with increasing nitrogen content. This is because nitrogen-containing compounds have a lower HHV than other biomass components, such as lignin and cellulose. Therefore, higher nitrogen content in the bio-crude can decrease its overall energy density and decrease its usefulness as a fuel. However, the effect of nitrogen content on HHV can be mitigated by separating nitrogen-rich compounds from the bio-crude.

The nitrogen content variables are therefore crucial in the HTL process as they can impact the BCY and HHV, which are important factors for the economic viability of the process. Proper control of nitrogen content variables can also improve the quality of the bio-crude, making it more suitable for use in a variety of applications, including fuel and chemical production. In addition, minimizing nitrogen content can increase the overall efficiency of the HTL process, reducing the capital and operating costs associated with longer reaction times or higher temperatures [44, 45].

Water

The amount of water used plays a pivotal role in the HTL process of lignocellulosic biomass, and it can have a substantial impact on both BCY and HHV in various ways. The effects of water variables on the HTL process can be summarized as follows:

BCY: The BCY generally increases with increasing S/BL up to a certain point, beyond which the yield may begin to decrease. This is because water acts as a solvent and catalyst in the HTL process, facilitating the liquefaction of biomass and promoting the formation of bio-crude. However, excessive water can also lead to reduced reaction rates and increased energy consumption, as well as dilution of the bio-crude, which can reduce its yield. Therefore, the optimal S/BL ratio is important for maximizing bio-crude yield.

HHV: The HHV of the bio-crude generally decreases with increasing the S/BL ratio. This is because increasing the water content can lead to more extensive degradation of the bio-crude, resulting in lower energy density. Additionally, higher S/BL ratios can also increase the concentration of inorganic elements in the bio-crude, which can reduce its HHV. Therefore, smaller S/BL ratios may be advantageous in maintaining the quality of the bio-crude and maximizing its energy density.

The water variables are therefore crucial in the HTL process as they can impact the BCY and HHV, which are important factors for the economic viability of the process. Proper control of water variables can also improve the quality of the bio-crude, making it more suitable for use in a variety of applications, including fuel and chemical production. In addition, an optimal S/BL ratio can lead to higher productivity and lower operating costs, making the process more competitive. However, the optimal S/BL ratio may vary depending on the specific biomass feedstock and process conditions and should be determined through experimental optimization [43, 46].

Cellulose

Cellulose is a major component of lignocellulosic biomass and plays a significant role in determining the BCY and HHV in the HTL process. The effects and importance of cellulose variables on BCY and HHV can be summarized as follows:

BCY: The presence and amount of cellulose in the biomass feedstock can significantly affect the bio-crude yield in the HTL process. Cellulose is a complex polysaccharide that requires hydrolysis and solubilization to be liquefied into bio-crude. Therefore, biomass with a higher cellulose content is expected to have a higher bio-crude yield. However, excessive cellulose content can lead to the formation of solid residues, reducing the yield of bio-crude. Therefore, the optimal cellulose content in the biomass feedstock should be determined experimentally to maximize BCY.

HHV: The heating value of the bio-crude is also influenced by the cellulose content of the biomass feedstock. Cellulose is a high-energy component of biomass, and its presence can increase the energy density of the bio-crude. However, the degree of cellulose degradation and solubilization during the HTL process can also impact the HHV of the bio-crude. Excessive degradation of cellulose can lead to the formation of low-energy compounds, reducing the HHV of

the bio-crude. Therefore, the optimal cellulose content and processing conditions should be determined to maximize the HHV of the bio-crude.

In summary, cellulose content is an important variable that can significantly impact the BCY and HHV in the HTL process. Higher cellulose content can increase the BCY and energy density, but excessive cellulose can lead to the formation of solid residues and low-energy compounds. Therefore, the optimal cellulose content and processing conditions should be determined experimentally for each biomass feedstock to achieve the best results [47, 48].

Hemicellulose

Hemicellulose is another major component of lignocellulosic biomass and plays an important role in determining the BCY and HHV in the HTL process. The effects and importance of hemicellulose variables on BCY and HHV can be summarized as follows:

BCY: Hemicellulose is a complex polysaccharide that can be easily hydrolyzed and solubilized during the HTL process, making it an important contributor to BCY. Biomass with a higher hemicellulose content is expected to have a higher BCY due to its ease of solubilization. However, excessive hemicellulose content can lead to the formation of solid residues, reducing the yield of bio-crude. Therefore, the optimal hemicellulose content in the biomass feedstock should be determined experimentally to maximize BCY.

HHV: The heating value of the bio-crude is also influenced by the hemicellulose content of the biomass feedstock. Hemicellulose is a high-energy component of biomass, and its presence can increase the energy density of the bio-crude. However, the degree of hemicellulose degradation and solubilization during the HTL process can also impact the HHV of the bio-crude. Excessive degradation of hemicellulose can lead to the formation of low-energy compounds, reducing the HHV of the bio-crude. Therefore, the optimal hemicellulose content and processing conditions should be determined to maximize the HHV of the bio-crude.

In summary, hemicellulose content is an important variable that can significantly impact the bio-crude yield and HHV in the HTL process. Higher hemicellulose content can increase the BCY and energy density, but excessive hemicellulose can lead to the formation of solid residues and low-energy compounds. Therefore, the optimal hemicellulose content and processing conditions should be determined experimentally for each biomass feedstock to achieve the best results [48].

Lignin

Lignin is a major component of lignocellulosic biomass and plays an important role in determining the BCY and HHV in the HTL process. The effects and importance of lignin variables on BCY and HHV can be summarized as follows:

BCY: Lignin is a complex polymer that is relatively resistant to hydrolysis and solubilization during the HTL process. As a result, lignin content in the biomass feedstock has a negative impact on bio-crude yield. A higher lignin content can lead to the formation of solid residues, reducing the yield of bio-crude. Therefore, it is important to consider the lignin content of the biomass feedstock when selecting the feedstock for HTL process.

HHV: The heating value of the bio-crude is also influenced by the lignin content of the biomass feedstock. Lignin is a high-energy component of biomass, and its presence can increase the energy density of the bio-crude. However, the degree of lignin degradation and solubilization during the HTL process can also impact the HHV of the bio-crude. Excessive degradation of lignin can lead to the formation of low-energy compounds, reducing the HHV of the bio-crude. Therefore, the optimal lignin content and processing conditions should be determined to maximize the HHV of the bio-crude.

In summary, lignin content is an important variable that can significantly impact the bio-crude yield and HHV in the HTL process. Higher lignin content can reduce the bio-crude yield but increase the energy density, while excessive degradation of lignin can lead to low-energy compounds and reduced HHV. Therefore, the optimal lignin content and processing conditions should be determined experimentally for each biomass feedstock to achieve the best results [48].

User Interface

The user interface (UI) serves as the bridge between the user and the product, facilitating interaction. It primarily emphasizes appearance, design, and user-friendliness, aiming to ensure ease of use and avoid complexity. In this research, UI plays a crucial role in enhancing user convenience when utilizing the researcher's developed ML model. Users are not required to possess programming knowledge for effective model utilization. Figure 3 displays the UI screen for predicting bio-crude oil yield and higher heating value in the HTL process, which can incorporate biomass data for value prediction.

Figure 3 illustrates that in order to predict the values of bio-crude oil yield or higher heating value, it is necessary to input complete data for both operating conditions and elemental properties. For the biological properties, users have the option to enter values selectively within the cellulose, hemicellulose, and lignin groups or input all values for comprehensive predictions.

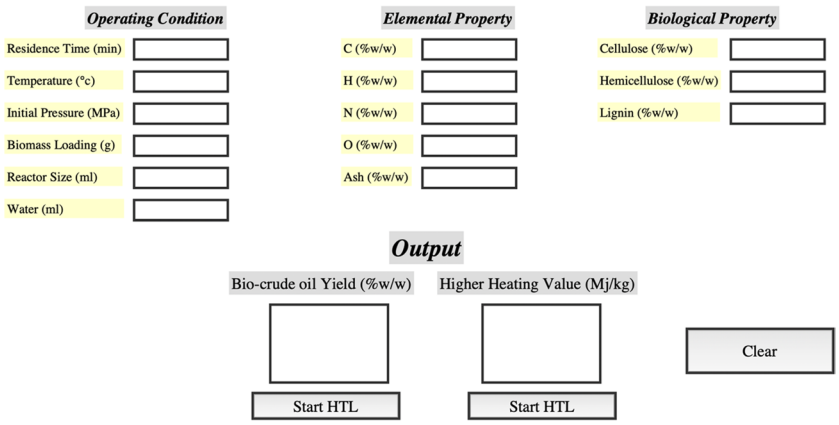

Fig. 3 User interface of the HTL process

Limitations of this Study

A limitation of this study is the relatively small number of datasets utilized in the model, which may result in lower accuracy compared to previous research by Djandja et al. [49] who investigated machine learning prediction of bio-oil yield during solvothermal liquefaction of lignocellulosic biowaste. However, in machine learning, varying input features can yield different results, such as the number and novelty of input features. Additionally, considerations such as data management and the ratio of training to testing data are crucial in machine learning tasks. Consequently, despite addressing similar topics, many machine learning tasks exhibit differences.

Another limitation of this study is that the authors describe the variables influencing the output solely in one dimension through the SHAP summary plot. While this method effectively reveals the importance and positive or negative effects of different variables on the output, the author recognizes the significance of elucidating the relationships among various variables influencing the output. This could be achieved by employing the partial dependence plot analysis or SHAP dependence plot in future studies to comprehensively explain the relationships among the variables affecting the output.

Conclusions

This study aimed to predict the resulting bio-crude oil yields and their calorific values from HTL of lignocellulosic biomass using a machine learning approach. Feature selection, employing the Shapley value method, identified significant input features from a dataset comprising 215 data points, 17 input features, and 2 target outputs. An extreme gradient boosting algorithm was demonstrated to provide the best performance, followed by random forest and multilayer perceptron. Conversely, kernel ridge regression exhibited lower accuracy. For the bio-crude yield, the temperature was found to have the most significant impact, followed by hydrogen, nitrogen, lignin contents, and the amount of water used. Meanwhile, for the calorific value, temperature also emerged as the most influential feature on model predictions, followed by ash content, reaction time, hemicellulose, and cellulose. The analysis of feature effects and interactions proved significant in the understanding of the HTL system.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12155-024-10773-0.

Acknowledgements This project is partially funded by the National Research Council of Thailand (NRCT), SCG Co., Ltd. and Fundamental Fund 2567: Chiang Mai University (CMU). The 1st author also wishes to thank the Teaching and Research Assistantships from the CMU Graduate School and the NRCT Research & Researchers for Industry PhD program.

Data Availability The datasets generated during and/or analyzed during the current study are available on reasonable request.

References

1. Barcena-Vazquez J, Caro K, Bermudez K et al (2023) Designing and evaluating Reto Global, a serious video game for supporting global warming awareness. Int J Hum Comput Stud 177:103080. https://doi.org/10.1016/j.ijhcs.2023.103080
2. Wang L, Wang L, Li Y et al (2023) A century-long analysis of global warming and earth temperature using a random walk with drift approach. Decis Analy J 7. https://doi.org/10.1016/j.dajour.2023.100237
3. Jiang L, Zhao Y, Yao Y et al (2023) Adding siderophores: a new strategy to reduce greenhouse gas emissions in composting. Bioresour Technol 384. https://doi.org/10.1016/j.biortech.2023.129319
4. Zhao J, Xie H, Liu D et al (2023) Climate-smart management for increasing crop yield and reducing greenhouse gas emission in Beijing-Tianjin-Hebei region, China. Agric For Meteorol 339. https://doi.org/10.1016/j.agrformet.2023.109569.
5. Wang W, Zhang H, Mo F et al (2022) Reducing greenhouse gas emissions and improving net ecosystem economic benefit through long-term conservation tillage in a wheat-maize

multiple cropping system in the Loess Plateau, China. Eur J Agron 141. https://doi.org/10.1016/j.eja.2022.126619
6. Naaz F, Samuchiwal S, Dalvi V et al (2023) Hydrothermal liquefaction could be a sustainable approach for valorization of wastewater grown algal biomass into cleaner fuel. Energy Convers Manag 283. https://doi.org/10.1016/j.enconman.2023.116887
7. Zoppi G, Tito E, Bianco I et al (2023) Life cycle assessment of the biofuel production from lignocellulosic biomass in a hydrothermal liquefaction – aqueous phase reforming integrated biorefinery. Renew Energy 206:375–385. https://doi.org/10.1016/j.renene.2023.02.011
8. Kaliyan N, Morey RV, Tiffany DG (2011) Reducing life cycle greenhouse gas emissions of corn ethanol by integrating biomass to produce heat and power at ethanol plants. Biomass Bioenergy 35:1103–1113. https://doi.org/10.1016/j.biombioe.2010.11.035
9. Alola AA, Adebayo TS (2023) Analysing the waste management, industrial and agriculture greenhouse gas emissions of biomass, fossil fuel, and metallic ores utilization in Iceland. Sci Total Environ 887. https://doi.org/10.1016/j.scitotenv.2023.164115
10. Miranda AM, Sáez AA, Hoyos BS et al (2021) Improving microalgal biomass production with industrial CO2 for bio-oil obtention by hydrothermal liquefaction. Fuel 302. https://doi.org/10.1016/j.fuel.2021.121236
11. Wang H, Han X, Zeng Y et al (2023) Development of a global kinetic model based on chemical compositions of lignocellulosic biomass for predicting product yields from hydrothermal liquefaction. Renew Energy 215:118956. https://doi.org/10.1016/j.renene.2023.118956
12. Fan Q, Fu P, Song C et al (2023) Valorization of waste biomass through hydrothermal liquefaction: a review with focus on linking hydrothermal factors to products characteristics. Ind Crops Prod 191. https://doi.org/10.1016/j.indcrop.2022.116017
13. Peng W, Karimi Sadaghiani O (2023) Enhancement of quality and quantity of woody biomass produced in forests using machine learning algorithms. Biomass Bioenergy 175. https://doi.org/10.1016/j.biombioe.2023.106884
14. Sonwai A, Pholchan P, Tippayawong N et al (2023) Machine learning approach for determining and optimizing influential factors of biogas production from lignocellulosic biomass. Bioresour Technol 383. https://doi.org/10.1016/j.biortech.2023.129235
15. Haq ZU, Ullah U, Khan MNA et al (2022) Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction. Bioresour Technol 363. https://doi.org/10.1016/j.biortech.2022.128008
16. Zhang W (2021) Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. Bioresour Technol 342. https://doi.org/10.1016/j.biortech.2021.126011
17. Shafizadeh A (2022) Machine learning predicts and optimizes hydrothermal liquefaction of biomass. J Chem Eng 445. https://doi.org/10.1016/j.cej.2022.136579
18. Katongtung T, Onsree T, Tippayawong N (2022) Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. Bioresour Technol 344. https://doi.org/10.1016/j.biortech.2021.126278
19. Onsree T, Tippayawong N (2021) Machine learning application to predict yields of solid products from biomass torrefaction. Renew Energy 167:425–432. https://doi.org/10.1016/j.renene.2020.11.099
20. Phromphithak S, Onsree T, Tippayawong N (2021) Machine learning prediction of cellulose-rich materials from biomass pretreatment with ionic liquid solvents. Bioresour Technol 323. https://doi.org/10.1016/j.biortech.2020.124642
21. Pedregosa F (2011) Scikit-learn: machine learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. http://scikit-learn.sourceforge.net
22. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery 785–794. https://doi.org/10.1145/2939672.2939785
23. Elmaz F, Yücel Ö, Mutlu AY (2020) Predictive modeling of biomass gasification with machine learning-based regression methods. Energy 191. https://doi.org/10.1016/j.energy.2019.116541
24. Đukanović M, Kašćelan L, Vuković S et al (2023) A machine learning approach for time series forecasting with application to debt risk of the Montenegrin electricity industry. Energy Rep 9:362–369. https://doi.org/10.1016/j.egyr.2023.05.240
25. Wang H, Yang J, Chen G et al (2023) Machine learning applications on air temperature prediction in the urban canopy layer: a critical review of 2011–2022. Urban Climate 49. https://doi.org/10.1016/j.uclim.2023.101499.
26. Mabula MJ, Kisanga D, Pamba S (2023) Application of machine learning algorithms and Sentinel-2 satellite for improved bathymetry retrieval in Lake Victoria, Tanzania. Egypt J Remote Sens Space Sci 26:619–627. https://doi.org/10.1016/j.ejrs.2023.07.003
27. Malakouti SM (2023) Babysitting hyperparameter optimization and 10-fold-cross-validation to enhance the performance of ML methods in predicting wind speed and energy generation. Intel Syst Appl 19. https://doi.org/10.1016/j.iswa.2023.200248
28. Zhang X, Liu CA (2023) Model averaging prediction by K-fold cross-validation. J Econom 235:280–301. https://doi.org/10.1016/j.jeconom.2022.04.007
29. Sumayli A (2023) Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. Arab J Chem 16. https://doi.org/10.1016/j.arabjc.2023.104833
30. Semmad A, Bahoura M (2023) Scalable serial hardware architecture of multilayer perceptron neural network for automatic wheezing detection. Microprocess Microsyst 99. https://doi.org/10.1016/j.micpro.2023.104844.
31. Kumar PS, Laha SK, Kumaraswamidhas LA (2023) Assessment of rolling element bearing degradation based on dynamic time warping, kernel ridge regression and support vector regression. Appl Acoust 208. https://doi.org/10.1016/j.apacoust.2023.109389
32. Rezaei I, Amirshahi SH, Mahbadi AA (2023) Utilizing support vector and kernel ridge regression methods in spectral reconstruction. Results Opt 11: 100405. https://doi.org/10.1016/j.rio.2023.100405.
33. Ghosh A, Dey P (2021) Flood Severity assessment of the coastal tract situated between Muriganga and Saptamukhi estuaries of Sundarban delta of India using frequency ratio (FR), fuzzy logic (FL), logistic regression (LR) and random forest (RF) models. Reg Stud Mar Sci 42. https://doi.org/10.1016/j.rsma.2021.101624
34. Khajavi H, Rastgoo A (2023) Predicting the carbon dioxide emission caused by road transport using a random forest (RF) model combined by meta-heuristic algorithms. Sustain Cities Soc 93. https://doi.org/10.1016/j.scs.2023.104503
35. Le HA (2022) An extreme gradient boosting approach to estimate the shear strength of FRP reinforced concrete beams. Structures 45:1307–1321. https://doi.org/10.1016/j.istruc.2022.09.112
36. Jarajapu DC, Rathinasamy M, Agarwal A et al (2022) Design flood estimation using extreme gradient boosting-based on bayesian optimization. J Hydrol (Amst) 613. https://doi.org/10.1016/j.jhydrol.2022.128341
37. Dong L, Liu Z, Zhang K et al (2023) Affordable federated edge learning framework via efficient Shapley value estimation. Future Gen Comput Syst 147:339–349. https://doi.org/10.1016/j.future.2023.05.007

38. Louhichi M, Nesmaoui R, Mbarek M et al (2023) Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering. Proc Comput Sci 220:806–811. https://doi.org/10.1016/j.procs.2023.03.107
39. Sharma N (2021) Effect of catalyst and temperature on the quality and productivity of HTL bio-oil from microalgae: a review. Renew Energy 174:810–822. https://doi.org/10.1016/j.renene.2021.04.147
40. Reddy HK (2016) Temperature effect on hydrothermal liquefaction of Nannochloropsis gaditana and Chlorella sp. Appl Energy 165:943–951. https://doi.org/10.1016/j.apenergy.2015.11.067
41. Biswas B, Arun Kumar A, Bisht Y et al (2017) Effects of temperature and solvent on hydrothermal liquefaction of Sargassum tenerrimum algae. Bioresour Technol 242:344–350. https://doi.org/10.1016/j.biortech.2017.03.045
42. Chen WT (2017) Effect of ash on hydrothermal liquefaction of high-ash content algal biomass. Algal Res 25:297–306. https://doi.org/10.1016/j.algal.2017.05.010
43. Duan P, Chang Z, Xu Y (2013) Hydrothermal processing of duckweed: Effect of reaction conditions on product distribution and composition. Bioresour Technol 135:710–719. https://doi.org/10.1016/j.biortech.2012.08.106
44. Tian C (2015) Hydrothermal liquefaction of harvested high-ash low-lipid algal biomass from Dianchi Lake: effects of operational parameters and relations of products. Bioresour Technol 184:336–343. https://doi.org/10.1016/j.biortech.2014.10.093
45. Tang X, Zhang C, Yang X (2020) Optimizing process of hydrothermal liquefaction of microalgae via flash heating and isolating aqueous extract from bio-crude. J Clean Prod 258. https://doi.org/10.1016/j.jclepro.2020.120660
46. Yoo G, Park MS, Yang JW et al (2015) Lipid content in microalgae determines the quality of biocrude and energy return on investment of hydrothermal liquefaction. Appl Energy 156:354–361. https://doi.org/10.1016/j.apenergy.2015.07.020
47. Amar VS (2021) Hydrothermal liquefaction (HTL) processing of unhydrolyzed solids (UHS) for hydrochar and its use for asymmetric supercapacitors with mixed (Mn,Ti)-Perovskite oxides. Renew Energy 173:329–341. https://doi.org/10.1016/j.renene.2021.03.126
48. Yiin CL (2022) A review on potential of green solvents in hydrothermal liquefaction (HTL) of lignin. Bioresour Technol 364. https://doi.org/10.1016/j.biortech.2022.128075
49. Djandja OS, Salami AA, Yuan H et al (2023) Machine learning prediction of bio-oil yield during solvothermal liquefaction of lignocellulosic biowaste. J Anal Appl Pyrol 175. https://doi.org/10.1016/j.jaap.2023.106209