Chemical Engineering Journal 442 (2022) 136013

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Accuracy of predictions made by machine learned models for biocrude yields obtained from hydrothermal liquefaction of organic wastes

Feng Cheng [a,1], Elizabeth R. Belden [a], Wenjing Li [b], Muntasir Shahabuddin [a], Randy C. Paffenroth [b,c], Michael T. Timko [a,*]

[a] Department of Chemical Engineering, Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA 01609, USA
[b] Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA 01609, USA
[c] Department of Computer Science, and Data Science Program, Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA 01609, USA

ARTICLE INFO

Keywords:
Hydrothermal liquefaction
Machine learning
Random Forest
eXtreme Gradient Boosting
Biocrude yield prediction

ABSTRACT

Hydrothermal liquefaction (HTL) has potential for converting abundant wet organic wastes into renewable fuels. Because HTL consists of a complex reaction network, deterministic, physics-based prediction of its biocrude yield is prohibitively difficult. Data-driven methods provide an alternative to the physics-based approach; however, rigorous testing must be performed to ensure the accuracy of predictions made by data-driven methods. To this end, a data set was assembled consisting of 570 data points appearing in the open literature. The data set was divided into training, validation, and test sub-sets and used for evaluating different machine learning regression approaches to predict biocrude yield. Among the tested algorithms, Random Forest and eXtreme Gradient Boosting (XGBoost) predicted biocrude yields in a test set that had not been used for training with the greatest accuracy, with root mean square errors (RMSE) of 8.34 and 8.57, respectively. Further refinement of the Random Forest model reduced its RMSE to 8.07. In comparison, predictions of a series of literature models resulted in RMSE ranging from 9.16 in the most accurate case to 27.6 in the least accurate; most literature models yielded RMSE values > 10. Using biocrude yield predictions from the most accurate Random Forest model and a probabilistic economic analysis found that the model accuracy is sufficient to prioritize allocation of resources based on projected minimum fuel selling price. The models and analysis presented here represent a major advance in the ability to use readily available data to predict biocrude yields on new feedstocks that have not previously been studied.

1. Introduction

An increasing number of nations have set aggressive goals to reach carbon neutrality within the next four decades in an effort to avert the most damaging impacts of global climate change [1]. Achieving the ambitious carbon neutrality goals requires abandoning nearly all fossil-based energy sources and substituting with carbon-free energy sources including biomass, nuclear, solar, wind, geothermal, and hydro energies [2–4]. Among these options, thermochemical conversion of nonedible biomass sources to produce liquid fuels has potential for decarbonizing the transportation sector due to biofuel compatibility with current infrastructure and the abundance of biomass [5,6].

Hydrothermal liquefaction (HTL) of algae, biomass, and wet organic waste streams has attracted attention in recent years as an efficient method for producing an energy-dense biocrude that can be upgraded into liquid transportation fuels [7,8]. In a near sub- or supercritical state, water acts as a reactant, catalyst, and reaction solvent to effectively decompose biomass components into smaller organic molecules that serve as precursors to form biocrude [9].

Maximizing biocrude yield and minimizing feedstock cost is one pathway to economically viable HTL-based production of liquid fuels [10,11]. Biocrude yield is mainly determined by feedstock properties [12,13], meaning that maximizing yield requires understanding of the relationship between feedstock properties and biocrude yield. Predictive methods that utilize readily available data are especially needed to help prioritize feeds for commercial development.

A classical approach for maximizing reaction yield is to optimize reaction conditions using a system of chemical pathways, which can

* Corresponding author.
E-mail address: mttimko@wpi.edu (M.T. Timko).
[1] Currently at Department of Chemical and Biological Engineering, University of Wisconsin–Madison, Madison, WI 53706, USA.

https://doi.org/10.1016/j.cej.2022.136013
Received 6 January 2022; Received in revised form 16 March 2022; Accepted 22 March 2022
Available online 26 March 2022

either be elementary or phenomenological [14]. However, waste feeds consist of many dozens of chemically distinct components, and HTL itself proceeds via a vast manifold of incompletely understood reactions consisting of hundreds or thousands of reactants, intermediates, and products [15]. Modeling this complex reaction system using physics-based models is clearly challenging. The common alternative used for making yield predictions is instead to measure yields for a handful of representative model compounds and use these as the basis for what can be termed multi-term or multi-component linear additivity models [16,17].

In multi-component linear additivity models, yields obtained for a family of model feeds are used to establish empirical values of coefficients that are then applied to generalize for other feeds with different compositions. In principle, terms can be added for any component that is suitable for HTL and well defined feeds with multiple components can be used to assign empirical coefficients to cross terms, intended to capture component-component interactions [18–21].

The advantage of the multi-component linear additivity model approach is that it maximizes the use of a limited data set for prediction of a wider range of feedstocks. The limitation is potential over reliance on empirical data obtained from a handful of feedstocks and subsequent overfitting of the empirical parameters that detract from the accuracy of predictions made for feedstocks that were not included in model development. The result is models that cannot be generalized to new feedstocks with confidence [22]; these models may retain accuracy for feedstocks included in the model development stage, but without testing them their accuracy cannot be guaranteed for new feeds.

Recent advances in machine learning can be harnessed for development of new types of data-driven models that relate feedstock properties to yields obtained by HTL conversion of biomass [23–25]. Unlike multi-component linear additivity models, machine learned models must be trained on much larger data sets consisting of hundreds, thousands, or even millions of data points, validated for robustness to avoid over fitting, and tested for predictability of data not included in the original data set [26]. Assuming that sufficient data are available, the resulting models can avoid the problem of over fitting, thereby permitting them to be used in a truly predictive manner – in other words achieving accuracy for feedstocks not explicitly considered during model development. Here, a distinction is drawn between model accuracy for data used in model regression (usually termed training data) and accuracy of predictions for conditions that were not explicitly included in the regression. A model that accurately fits data provided to it can be useful for many purposes; however, truly predictive models can be used for new situations that were not included in the regression and are therefore preferred in these cases.

Unfortunately, most modern machine learned models require millions of data points for training without overfitting [27,28]. Overfitting a machine learned, data-driven model detracts from the accuracy of its predictions, which defeats the purpose of the model [12,16]. Individual HTL experiments are labor intensive and generating a data set consisting of millions of biocrude yield data points is time and cost prohibitive. The need for machine learning methods that avoid overfitting and retain predictive accuracy without the requirement of millions of data points is a clear need for the chemical engineering community. Selecting an appropriate model type and then validating model performance to avoid overfitting becomes crucial in the low-data limit [29], yet the importance of this step is often over looked. As a result, machine learned models developed for <1000 data points are routinely over fit, thereby detracting from the accuracy for their predictions outside the original training data.

Encouragingly, some types of machine learning models have been proven to be retain predictive accuracy for regression of systems with hundreds of – rather than hundreds of thousands or more – data points [30,31]. That stated, machine learning in the low-data limit requires careful selection of the algorithm, as some are more prone to overfitting than others [32]; new strategies for selecting training data; generalizable methods for validating results [33]; and guidance for selection of independent variables that lead to accurate and reliable predictions [34].

To date, appropriate protocol for the aforementioned steps does not appear in the literature, despite reports on the use of data-driven models for predictions of HTL biocrude yields [35]. A recent study on machine learning predictions of biocrude yields implemented a validation step to minimize over fitting, but did not set aside data for testing [36], which means the model accuracy for fitting the training and validation is quantified but the predictions for feeds that did not appear in the training data set is not. Without comparison with a data test set, the accuracy of true model predictions cannot be ascertained.

The objective of this study was to evaluate the methodology for developing generalizable machine learning models to predict HTL biocrude yield in the low data limit (i.e., <1,000 data points). The study consisted of training eight different regression models, validating their predictions to determine the extent to which accuracy is influenced by random selection of training data; and testing them on a new subset of the data to determine accuracy when the models are used predictively. The most accurate model was then refined to predict biocrude yield based on new data that was not included in the training data. The accuracy of this model was compared with the accuracy of other literature models, especially multi-component linear additivity models [16–19,37]. Finally, the relationship between the accuracy of biocrude yields and economic performance was evaluated using Monte Carlo simulations to propagate biocrude uncertainty into uncertainties of projected minimum fuel selling price (MFSP).

2. Methodology

2.1. Overview

Fig. 1 is a schematic representation of the process that was followed for the study. Step 1 was assembly of a data set from studies present in the literature. Steps 2–4 are model development, which includes evaluating the effects and accuracy of different regression methods, different data handling protocols, and different ways to ensure accuracy of predictions made for conditions not included in initial model development. The end of Step 4 is down selection of the most promising models. Step 5 is the use of the most promising model for biocrude yield prediction, which was then used to determine the relative importance of different independent variables on performance and tested with several different modifications to the independent variables to investigate if model performance could be improved.

Following model development, the most accurate and generalizable model was then used to make biocrude yield predictions for a series of feeds that appear in the literature, but for which no HTL data are published. These yields were then used in an economic model [38] to evaluate the effect of yield prediction and uncertainty on projected economic performance. The final two steps, screening feedstocks and performing HTL experiments on them are recommendations of how the resulting regression models can be used.

2.2. Data collection, preparation, and curation

Development of a data-driven model requires careful selection and preparation of data so that it generates reliable results. By reviewing 190 publications appearing in the open literature on HTL of various feeds, 570 data points were selected for inclusion in the data set. Consistent criteria were applied for including a given data point in the data set: 1) adequate reporting of uncertainty and reproducibility, including reporting of at least two replicate runs as a measure of reproducibility; 2) thorough reporting of experimental conditions, including at a minimum biochemical composition of the feed and reactor conditions; 3) appearance in a peer-reviewed journal.

The impact of feedstock composition on model performance is the

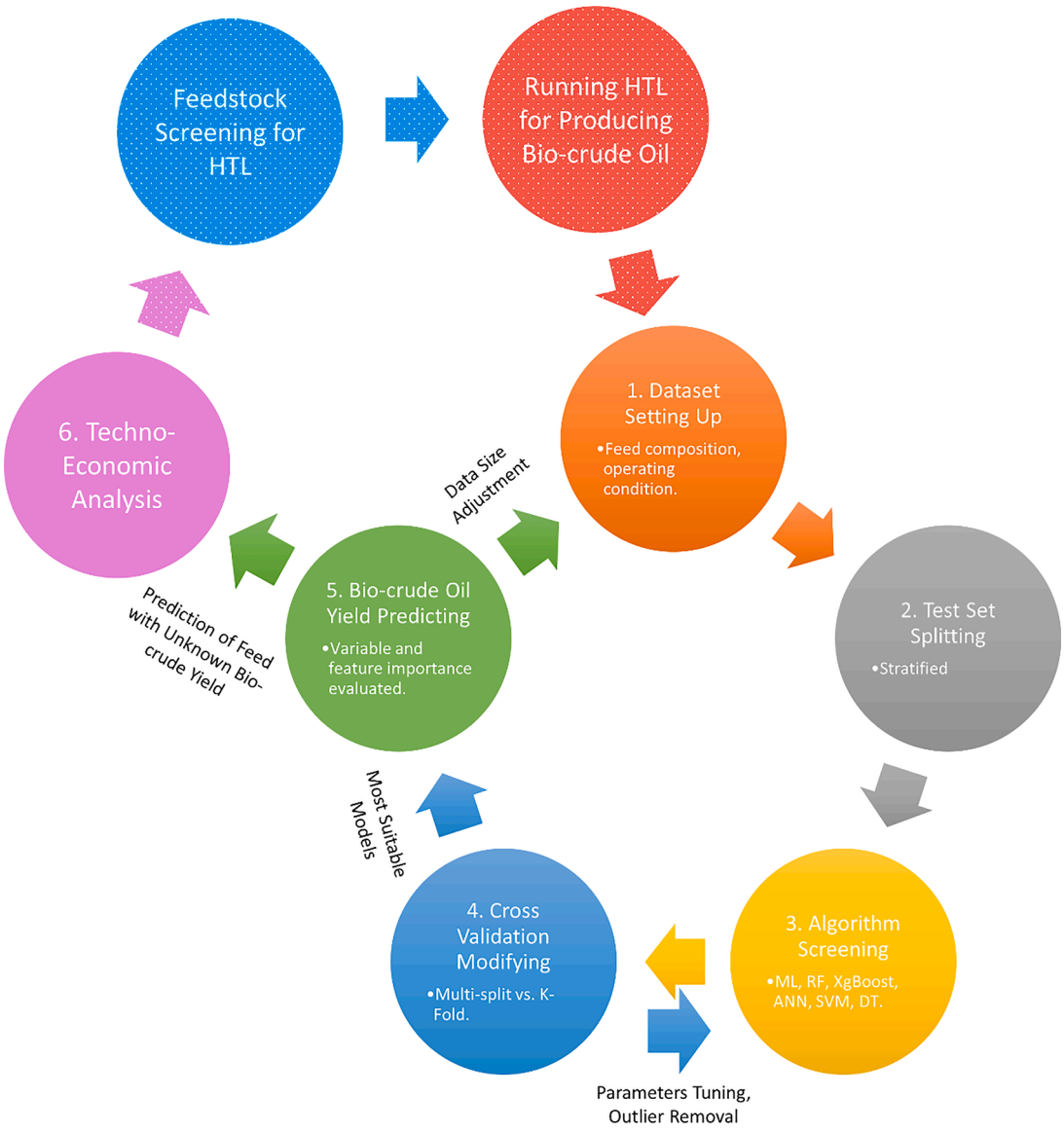

Fig. 1. Schematic diagram of the development and application of machine learning models in this study. HTL: hydrothermal liquefaction. RF: random forest. ANN: artificial neural network. SVM: support vector machine. DT: decision tree. The pure-color circle block means the work conducted in this study, and the white-dot-containing circle block means the work out of the scope of this study.

emphasis of this study. Therefore, to avoid feedstock overrepresentation by inclusion of all of the data from studies reporting biocrude yield for a single feedstock at many different conditions (as shown in Tables S.1 and S.2), only data reported at the "optimal" condition, i.e., conditions at which the maximum biocrude yield was observed, are included here from sources that report yields at many different reaction conditions.

Fig. 2 is a mosaic plot representation of the resulting data set divided into different feed categories, where the number of data points in a particular category and the number of sources used to extract data points were provided for a given feed category. Algae, lignocellulosic biomass, and model compounds (including fatty acids, proteins/amino acids, cellulose, glucan, glucose, hemicellulose, xylan, xylose, extracted lignin, etc.) are the most highly represented feeds. The rest of the data set includes food waste, manure, sludge, bioethanol residue, municipal solid waste, and seed plants.

Complete data tables are provided in the Supporting Information. Table 1 provides several representative entries. Here, the feedstock, sample type, and extractant are strings. All other independent and dependent variables are integers.

No studies were intentionally excluded that met the three aforementioned criteria; however, as the appearance of new publications on HTL is increasing rapidly the study makes no guarantee of including all published data. Instead, the methods used here guarantee a representative sampling of reliable data that can be extended as new data are published.

2.3. Selection of independent and dependent variables

Selecting independent and dependent variables is a key step in development of a data-driven model. Biocrude yield was selected as the dependent variable, as this is a key parameter determining economic viability of an HTL process [47]. Based on their importance in determining biocrude yield and general availability in published data, the independent variables included in the study are feedstock type, biochemical composition, solids loading (3–30 wt%), reaction temperature (220–370 °C), reaction time (0–120 min), heating rate (3–990 °C/min), organic biocrude extraction solvent, reactor type, reactor size (1.3–2000 mL), and yields of char, gas, and aqueous phase. Of these, the biochemical composition was described using seven composition categories: lipid, lignin, cellulose, hemicellulose, carbohydrates (e.g.,

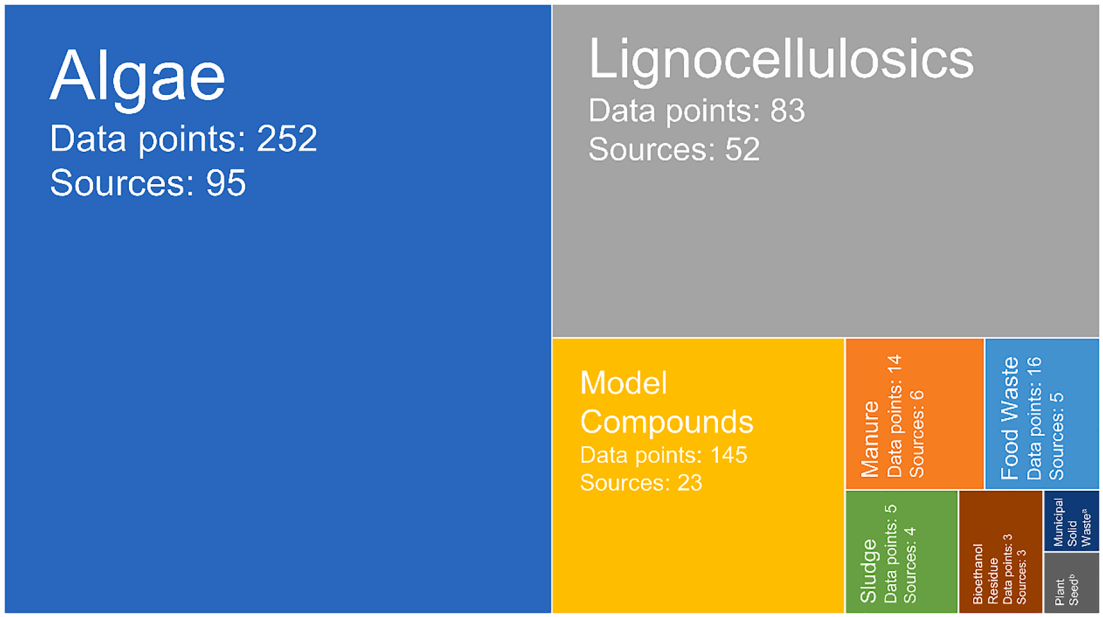

Fig. 2. Conceptual schematic diagram of the data set used in this study. [a] Municipal solid waste includes two data points from one literature source. [b] Plant seeds includes one data point from a single literature source.

glucose or starch), protein, and ash. Tables S.1 and S.2 in the Supporting Information summarize these variables.

Some publications do not report all values for all of the aforementioned independent variables, resulting in data gaps. For example, not all studies report heating rate. Missing data complicates comparisons since model accuracy depends on both the independent variables and the number of data points used in the regression. When the data related to a certain independent variable was missing, the entire entry was removed from the data set for that analysis (i.e., that row of the data table was entirely removed). Table S.11 lists the number of data points corresponding to different missing independent variables. To provide even footing and because the statistical methods used here depend on the number of data points, the impact of every individual independent variable on model performance was evaluated by generation of two data sets with the same number of data points, of which one included values for the independent variable to be studied and one lacked it. The difference in the predictions of these models was used to infer the impact of that variable.

2.4. Criteria for model evaluation

The criteria of model evaluation used in this study include mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (r^2), and mean relative error (MRE). The Supporting Information provides mathematical definitions of all four of these metrics. For all practical purposes, MAE, RMSE, MRE, and r^2 all respond similarly to changes in the model and/or modelled data set. Of these various methods to quantify model accuracy, RMSE is most sensitive to a small number of highly inaccurate predictions. For practical applications, highly inaccurate predictions are especially troublesome and so this work adopts RMSE as its primary way to quantify accuracy. For the current data set, RMSE is always greater than or equal to MAE, making RMSE a more conservative estimate of model accuracy.

2.5. Machine learning algorithms

After generating the data set, eight machine learning regression methods were evaluated for their performance in predicting biocrude yields. Table 2 summarizes these models as: (1) multiple linear regression, (2) Ridge regression, (3) Lasso regression, and (4) support vector machine regression (SVM); or nonlinear (5) decision tree regression, (6) multilayer perceptron (a form of "artificial neural network" or "ANN"), (7) random forest regression, and (8) eXtreme Gradient Boost (XGBoost) regression.

All regression methods were programmed, implemented, and optimized using Python 3.6.9. The Supporting Information provides additional descriptions of each of these models. Each of these types of models includes one or more parameters that can be optimized to improve model performance, e.g., the number of trees included in a Random Forest regression. Model parameters were carefully tuned to achieve optimal validation model performance based on their error metrics arising from regression of the complete data set (including RMSE, MAE, r^2, and MRE). The resulting optimized values of model parameters were then used for all subsequent implementations of that regression method and optimized values are provided in the Supporting Information.

2.6. Model training, validation, and testing

For development of regression models for each of the eight selected methods, the data set was divided into test and training subsets with a test to training ratio of 1:9. The training data were further split into data used explicitly for training and data used for internal or cross validation. Optimization to the training data resulted in a regressed model, which was then used for predicting biocrude yields for the 10% of the data initially set aside for testing. All performance metrics shown here are based on this test data, unless otherwise noted. Fig. 3 is a schematic of this process, showing the split between testing and training and the further split for internal validation.

Two approaches were used for splitting the data into testing and training/validation sets: 1) completely randomized sampling and 2) stratified sampling. Stratified sampling avoids the potential for random oversampling of a particular subset of the data, e.g., oversampling of high yield data during training, that results in a poor fit of the test data. Oversampling is especially problematic for small data sets, which are the subject of this study. On the other hand, data stratification can inadvertently introduce artifacts into the regression, since all regression models are based on the concept that sampling is totally random.

Analysis was performed first without stratification and then a second time using stratified data. The result of this comparison was the finding that data stratification is a beneficial technique for reducing RMSE,

Table 1
Description of representative data points withdrawn from Tables S.1 and S.2.

REF	Feedstock	Sample Type	Solid Loading (wt%)	Temp (°C)	Time (min)	Extractant	Reactor Size (mL)	Lipid[c]	Protein[c]	Cellulose[c]	Hemicellulose[c]	Carbohydrate[a,c]	Lignin[c]	Ash[c]	Bio-crude Yield (wt.%) [c]
[39]	swine manure	Manure	25.0	300	60	Toluene	100	20.3	24.5	3.8	27.3	0.0	3.6	16.3	39.2
[40]	Spirulina platensis	Algae	9.1	315	35	DCM	50	3.8	70.0	0.0	0.0	26.2	0.0	0.0	21.0
[37]	cornstarch	Model Compound	15.0	300	20	DCM	4	0.0	0.0	0.0	0.0	100.0	0.0	0.0	5.9
[41]	rice husk	Lignocellulosics	9.1	300	30	DCM	50	0.0	0.0	31.3	24.3	0.0	14.3	11.4	43.1
[42]	Pinewood	Lignocellulosics	16.7	350	20	Acetone	200	0.0	0.0	37.0	38.0	0.0	22.0	0.3	21.5
[43]	DDGS[b]	Bioethanol Residue	80.0	340	16	DCM	20	22.4	42.2	0.0	0.0	35.0	2.8	5.4	34.6
[44]	Sewage Sludge	Sludge	26.7	350	15	Diethyl Ether	41	3.2	28.5	0.0	0.0	34.2	0.0	34.0	16.7
[45]	Litsea cubeba seed	Plant Seed	N.A.	290	60	DCM	25	41.0	35.0	0.0	0.0	12.0	0.0	6.5	56.9
[46]	Streaky pork	Food Waste	10.0	320	60	DCM	7	58.7	39.6	0.0	0.0	0.0	0.0	1.6	55.6

[a] Water-soluble sugars. [b] Dried distillers grains with solubles. [c] Dry basis. DCM: Dichloromethane.

Table 2
The regression algorithms in this study.

Regression Methods		Method Description	Sources
Linear Regression	Multiple Linear	The most common-used and simplest linear regression method	[48]
	Ridge Lasso	The linear regression method modified by introducing a penalty term to inhibit overfitting.	[49,50]
	Support Vector Machine	A supervised learning method to fit data by finding hyperplane and defining acceptable error.	[51]
Non-Linear Regression	Decision Tree	A simple supervised learning method based on collecting data as roots and nodes, following nodes that meet required decision, and reaching leaf node as outcome.	[52]
	Multilayer Perceptron	An artificial neural network, consisting of an input layer, a non-linear hidden layer, and an output layer.	[53]
	Random Forest	An ensemble learning method that uses multiple decision tress and bootstrap aggregation to improve accuracy.	[54]
	XGBoost	A decision tree-based algorithm that has been reported to be most accurate for small to medium-sized structured data sets.	[55]

without making the model susceptible to erroneous predictions of test data due to biasing. Accordingly, results presented here utilize stratified data splitting.

The training data set was subjected to validation, a necessary step required to ensure that the model does not benefit from fortuitous selection of training data. The cross-validation scheme used here entails training the model multiple times on different sub-populations of the data set to obtain average values of the various error metrics and their corresponding standard deviations.

Internal validation was performed using a machine learning literature method termed the K-fold cross validation [56]. The K-fold method divides the data set into k equal sized sub-datasets, which are then assigned as training data or validation data (after test set has been split) in a training to validation ratio of 1:K-1. The K-fold validation process is then repeated k times, until all k subsamples have been used once as the validation set. Subsequently, the results of the k distinct regression are used to generate average values and standard deviations of RMSE that are reported here.

Typically, the K-fold method utilizes 5 to 10 splits. For the small data set, using 10 splits was found to result in a large standard deviation of validation set RMSE, arising from over sampling within the small-size dataset. Accordingly, a new modified K-fold validation method was developed, where the K-fold splitting was repeated between 10 and 1,000 times to determine average values of error metrics and their corresponding standard deviations.

The modified K-fold method was applied to the most accurate models, with representative results shown in Fig. S.3. Fig. S.3 indicates that splitting the dataset with >100 cycles reduces the standard deviation of RMSE predictions to <1%, a value that provides confidence that model performance is not impacted by fortuitous splitting of the data set and that the corresponding model can be used in a predictive manner.

Comparison of predictions with the test data provides a measure of model accuracy, and the dependence of model accuracy on how the data are split provides the truest measure of model predictability that can be obtained. A model lacking predictability can appear to be accurate by randomly fortuitous selection of the training data; however, the models which are less predictable offer no guarantee of performance for data not included in the training set. Overfitting is one of the main reasons for loss of predictability.

Some regression methods quantify the importance of the

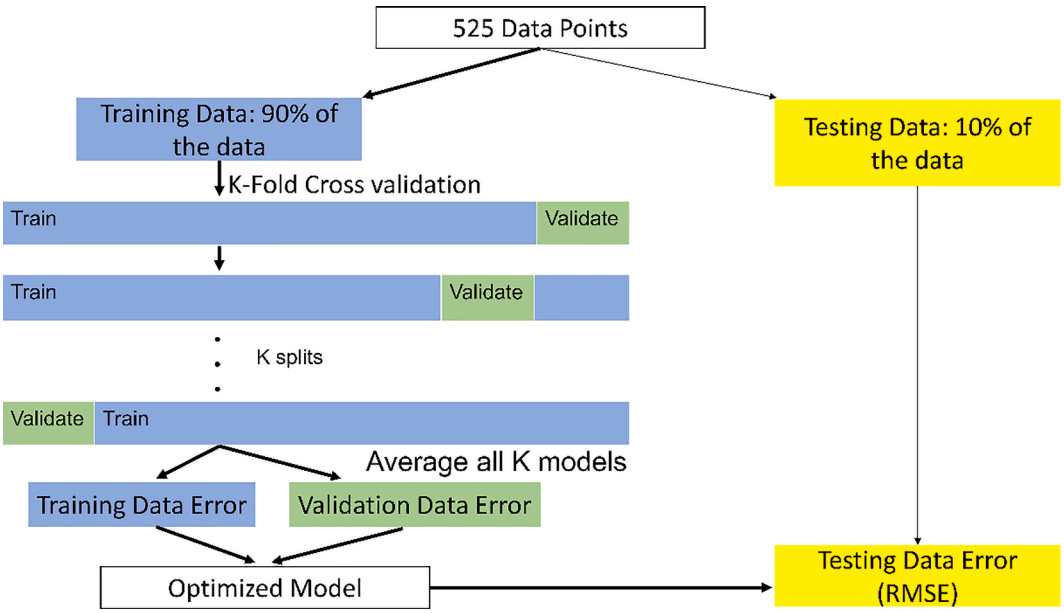

Fig. 3. Schematic diagram of the training, validation, and testing of the HTL data set. Training, validation, and testing all produce values of RMSE. Training and validation produce families of values, corresponding to an average value and standard deviation. The value of RMSE arising from the test data is the truest measure of model predictability. Here, RMSE values presented have been generated from the test data, unless otherwise noted.

independent variables appearing in their corresponding regression models. The metric is termed as feature importance, and it is a measure of how useful a given independent variable is for making a prediction. Values of feature importance were estimated using the built-in feature functions in Random Forest and XGBoost packages in Python.

2.7. Techno-economic analysis

A techno-economic model was built using the flowsheet published by Pacific Northwest National Lab (PNNL) for analysis of HTL conversion of sewage sludge to biocrude [38]. A discounted cash flow model, quantifying capital investment (equipment and initial costs), fixed operating costs (labor, insurance, and maintenance costs), and variable operating costs (raw material and utility costs) was built, with process parameters such as product yield and higher heating value (HHV) set as fixed parameters (as shown in Table 3). More details for each of these cash streams can be found in Tables S.8-S.10.

A set of eight feed streams was selected from the literature for economic modeling [57–62]. These feed streams were selected to represent a range of different sources, as summarized in Table S.6. The selected feed streams were characterized by distinct biochemical composition from one another, but they had never been treated using HTL and therefore represent feed streams that are completely unknown by the model.

The Random Forest regression model was trained using totally 570 data points obtained from Tables S.1 and S.2 and used to predict biocrude yield for the feed streams, as shown in Table S.6, along with a root mean square error (RMSE) as a measurement of model yield uncertainty, estimated based on the earlier dataset (including 570 data points). To calculate distributions of biocrude selling prices, Monte Carlo simulations were run, using the biocrude yield as the uncertain variable [63]. All other factors were held constant in this analysis to isolate the effect of biocrude yield uncertainty on minimum fuel selling price (MFSP), a key metric of economic performance.

Variable distributions were defined using the base-case yield for a given feedstock as the expected value, and the base-case yield +/- the RMSE as the upper and lower bounds of a triangular distribution, used to represent an “expert opinion” [64]. Monte Carlo simulations were run 10,000 times, with each iteration resulting in an estimated value of the MFSP [65,66]. The mean of the MFSPs was used as the expected MFSP for a given feedstock [67]. The resulting upper and lower bounds of the simulation are used as uncertainty bounds for the MFSP. The result is an estimated biocrude yield and corresponding uncertainty from the regression model, and a projected MFSP with corresponding uncertainty estimated from the techno-economic analysis.

Table 3
Techno-economic analysis key parameters.

Variable	Value*
Fixed Capital Investment ($)	34,118,811
Variable Operating Costs ($/yr)	1,545,000
Fixed Operating Costs ($/yr)	2,725,000
Internal Rate of Return	10%
Income Tax Rate	21%
Biocrude HHV (MJ/kg)	36.1
Plant Scale (DTPD)	110
Feedstock Cost ($/Dry Ton)	0

* Key economic parameters mirror the values used by Snowden-Swan et al. 2017 [38].

3. Results and discussions

The objective of this work is development and evaluation of regression models for prediction of biocrude yields obtained by HTL of different feed streams. A particular emphasis was placed on the accuracy of the machine learned models for test data, that is, data which had been withheld during training and validation, compared with that observed for the multi-component linear additivity models common in the literature [16–19,37]. Accuracy for test data is the truest indication of model predictability, which is a frequent goal of engineering models – whether data driven or physics-based.

The structure of the Results follows the steps shown previously in Fig. 1, beginning with selection and initial data evaluation and continuing with evaluation of the accuracy of eight distinct, representative regression methods [68]. The most accurate method was then refined to improve its accuracy in the small data limit [69,70], and its accuracy for predicting test data not involved in any other step of model development was compared with the accuracy of several literature

models. Finally, biocrude yields predicted by this same model were used to project economic performance for a series of feeds that had not previously been evaluated for HTL.

3.1. Evaluation of Machine learning regression models

Tables S.1 and S.2 in the Supporting Information provide the biocrude yield, reactor condition, and biochemical composition data used in this study. Figs. S.1. and S.2 present Pearson correlation coefficients determined for the data set, showing that the single strongest correlation exists between lipid content and biocrude yield (0.79). While this correlation is not predictive, it does foreshadow a prominent role for lipid content in any accurate predictive model of biocrude yield.

The data in Tables S.1 and S.2 were divided into training and testing subsets, as shown in Fig. 3, and used for development of a series of regression models. Unfortunately, no theorem exists for selection of the most appropriate regression method for a given data set. For example, methods based on neural networks can be highly accurate, yet they are prone to instability for small data sets and are easily misled by inclusion of superfluous independent variables [71,72]. Accordingly, a careful study must consider multiple options.

Many different types of regression methods are available for modeling engineering data [68]. Based on these considerations, eight popular and well-developed machine learning regression methods were selected for modeling HTL biocrude yields. Of these, four of the regression models were linear and four were non-linear (as shown previously in Table 2). Each of these models was trained, cross validated using the K-fold method (with k = 10),[73] and then used to predict test data, following the schematic shown in Fig. 3 in the Methods.

Physically, the various biochemical components can reasonably be expected to interact with one another [18], and polynomial terms have been proposed to capture these interactions [17,18,37]. Accordingly, polynomial terms were included in the regression analysis to determine the impact of interaction on biocrude yield predictions. Because new variables can be added to decision tree based machine learning models (a family that includes Random Forest and XGBoost) without risking over fitting, both binary (21) and ternary (35) interaction terms were included as polynomial terms. This contrasts with the situation encountered when regressing multi-component linear additivity models, which can include interaction terms but at the risk of over fitting. Accordingly, including both binary and ternary terms in the current study is a conservative and comprehensive approach for capturing interactions between feedstock constituents that goes beyond what has already been evaluated in the literature. Table S.4. in the methods were provided in the Supporting Information is a list of all the interaction terms.

Fig. 4 summarizes the results of the initial implementation of the eight regression methods. RMSE values are shown for training, validation, and test data and for models both with and without polynomial interaction terms. Each of these sets of RMSE values is valuable and each will be considered separately.

RMSE values obtained from training and validation provide an indication of how well each model fits data input to it. Taken collectively, the RMSE values obtained from training range from 9.7% for the multi-linear regression (ML) to 2.9% for the XGBoost model with polynomial interaction terms. Accordingly, XGBoost clearly is most able to capture data that is supplied to it. In all cases, addition of interaction terms decreases the value of the RMSE obtained from model training, meaning that polynomial terms improve the fit of data fed to the model.

RMSE values obtained from validation provide a sense of how much fortuitous data splitting impacts model accuracy. As expected, RMSE values obtained from validation are always greater than or equal to those observed for training. The reason why the validation RMSE is greater than the training RMSE is that the model is not fit directly to this data, and the effect of validation is to de-tune the model to minimize over fitting. Validation set RMSE values vary from 9.7 (again for multi-linear regression) to 8.3 (again for the XGBoost method with polynomial terms). As with the training data, addition of polynomial terms reduces validation set RMSE values.

Standard deviations of the RMSE values obtained from validation range from 1 to 4%, with the largest value observed for the support vector machine (SVM) method with polynomial terms. These values give a sense of how important fortuitous splitting is to model performance. Here, SVM is very sensitive to how the data are split and the large standard deviation of its validation RMSE value recommends against using this method. In comparison, the standard deviations of RMSE

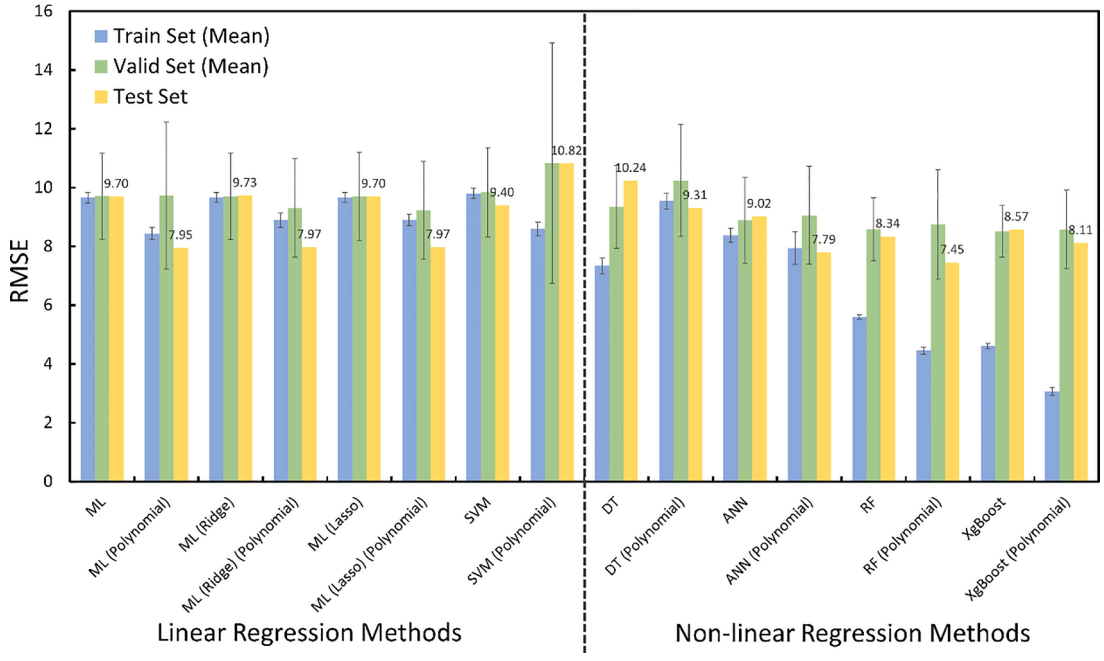

Fig. 4. The root-mean-square error (RMSE) of eight linear and non-linear regression models with/without the introduction of polynomial biochemical component terms. The model parameters for all optimal regression methods were provided in the Supporting Information.

values obtained for XGBoost and especially the Random Forest model are less than those observed for other methods, without sacrificing accuracy.

Interestingly, inclusion of polynomial terms always increases the standard deviations of the RMSE obtained during validation, an indication of overfitting and sensitivity to fortuitous data selection that detracts from predictability. Accordingly, while naïve reliance on training set RMSE values recommends inclusion of polynomial terms, the performance of polynomial-based models is sensitive to fortuitous splitting, recommending that they not be used in predictive models. For this reason, models with polynomial terms should be used with caution.

Finally, Fig. 4 provides values of the test set RMSE. Testing is done only one time, so the test set RMSE has no standard deviation – unlike the training and validation RMSE values, which are calculated for each K-split. As expected, based on the methodology used in this study, test set RMSE values obtained for each method always fall within the range determined by the corresponding mean value and standard deviation of the RMSE value obtained from validation. This observation indicates that the validation method properly captures sensitivity to data splitting. Test set RMSE values range from 10.82 (SVM with polynomial terms) to 7.45 (Random Forest with polynomial terms).

Fig. 4 contains all of the information required to select the model that is expected to be the most accurate for making predictions on feeds that have not been included in training. That decision comes down to a combination of test-set RMSE and the standard deviation of the validation set RMSE. Based solely on test set RMSE, the Random Forest, XGBoost, and artificial neural network (ANN) models have clear advantages over the other models considered here. Of these, the test set RMSE values obtained for Random Forest and XGBoost are less than those observed for the ANN model; the standard deviations of the validation set RMSE are also less for Random Forest and XGBoost than ANN. Accordingly, Random Forest and XGBoost are the preferred methods for predictability of new data.

Having provided guidance on the type of model to select (Random Forest or XGBoost), the next question was whether to include polynomial terms in the regression. On the one hand, inclusion of polynomial interactions further decreases test set RMSE for both Random Forest and XGBoost. On the other hand, including polynomial terms increases the standard deviation of the validation set RMSE, detracting from the confidence of using the polynomial methods in a predictive manner. No formal theory exists for balancing the merits of test set RMSE and standard deviation of the validation set RMSE and selecting either Random Forest or XGBoost with or without polynomial terms can be justified. In fact, statistical analysis (summarized in the Supporting Information, especially Table S.4) indicates that the only interaction terms with significant correlation with biocrude yield were lipid × lignin and protein × ash. Due to an abundance of caution for this small data set and to be conservative to guard against over fitting and fortuitous selection of test data, the Random Forest model without polynomial terms was selected for further refinement.

Various approaches to improving the accuracy of the Random Forest model were considered, including: extension of the K-fold method to > 10 cycles; inclusion of additional independent variables, such as reactor temperature, reaction time, and reactor volume; and consideration of co-product yields such as char, gas, and aqueous phases as independent variables Complete details are provided in the Supporting Information, especially Figs. S.3-S.13 and Table S.5.

Some of the aforementioned refinements resulted in modest improvements in model performance. Modifying the K-fold validation step to >100 cycles (instead of the customary 10) yet still keeping the training-validation ratio fixed at 8:2, reduced the standard deviation of the validation set RMSE from 0.4 to 0.05 and – because the resulting model is more robust – reduced the corresponding test set RMSE from 8.43 to 8.07. Full details are provided in the Supporting Information, especially Fig. S.4. Future work on machine learning regression of small data sets should adopt the modified K-fold method proposed in this work, in which the dataset was split with ≥100 cycles.

Treating char yield as an independent variable reduced test set RMSE from 8.43 to 7.43. On the other hand, including gas or aqueous phase yields did not reduce RMSE, indicating that these are not statistically related to biocrude yield. To be conservative, the RMSE calculated without using char as a regression variable is used for all comparisons reported later in this study.

Aside from the modified K-fold method and treating char yield as an independent variable, none of the other refinements evaluated here resulted in model improvement, either in terms of test set RMSE or the standard deviation of the validation set RMSE. More details are provided in the Supporting Information. The lack of improvement observed for inclusion of reaction temperature is likely due to the fact that biocrude yield is only weakly sensitive to reaction temperature near the optimal value (approximately 300 °C) and the published data are biased to reporting in this range. For this reason, when a study reported yields as a function of temperature, only data at or near the optimum were included in the data collection (Table S.1). Naturally, performing HTL at temperatures much less than or greater than the optimum will negatively impact yield.

The optimized version of the Random Forest model achieves test set RMSE of 8.07, which should be regarded as approaching the practical limit of predictive accuracy. Typical values of reported experimental uncertainty are on the order of 5% [74]. The value of RMSE reported here (8.07) is only slightly greater than this average value of experimental uncertainty, and the accuracy of model predictions is not expected to be greater than the reported uncertainty of the data being modeled. Accordingly, a test set RMSE of approximately 5% is a realistic lower limit on the accuracy of a data-driven biocrude yield model. Any reported value less than this should be treated with skepticism.

3.2. Comparison of biocrude yield predictions with literature models

The premise of this study was to understand predictability of biocrude regression models. A Random Forest model achieved an RMSE of 8.07, a value which includes use of an improved K-fold method. As mentioned in the introduction, numerous other biocrude yield prediction models appear in the literature [16,17]. The predictive capabilities of these models is difficult to ascertain, as they are nearly uniformly developed based on yields observed for a handful of model feeds and then tested over a small sub-set of real feeds. Models of this type can be termed “multi-component linear additivity models”. The other type of model is a non-linear regression, which uses a training set usually consisting of a few dozen data points to regress a family of parameters to fit the training data. Again, the predictive power of a non-linear regression cannot be determined solely from its ability to fit a limited training set.

The current study provides an opportunity to assess the predictive capability of the multi-component linear additivity models and nonlinear regression models, using the same test set as was used to select the Random Forest model as the most accurate available method. Accordingly, the various literature models were used to predict the biocrude yields in the test set, with subsequent calculation of the RMSE. Tables 4 and S.7 summarizes the results of this exercise.

Interestingly, values of the test set RMSE calculated for the various literature models were always greater than found for the Random Forest model. This comparison is completely fair since none of the literature models nor the Random Forest model were developed for the test data. Interestingly, the nonlinear regression models (27.6 and 12.01 RMSE) are two of the least accurate models, despite their reported values of r^2 > 0.98. In all likelihood, the poor predictive performance of the nonlinear regressions is a result of overfitting to their respective training data. The RMSE values corresponding to the two nonlinear regression shows that r^2 calculated for a training set is not a good indicator of predictive capability. Models of this type can instead be used for other purposes as they are highly accurate for capturing data provided for training.

Values of test set RMSE found for the multi-component linear additivity models vary over a wide range, from 9.16 (Li et al. [76]) to 17.1 (Deniel et al. [75]). This finding indicates that the basic form of the multi-component linear additivity model can be nearly as accurate as the Random Forest. In fact, using a published multi-component linear additivity model is simpler and more convenient than using a published Random Forest model, so for preliminary estimates the model of Li et al. [76] will often be suitable. That stated, even the most accurate multi-component linear additivity model results in errors greater than 10% 15 out of 53 times (i.e., 74% of the time the error was less than 10%). By comparison, predictions made by the Random Forest model are more accurate than 10% for 81% of the test set data points. Accordingly, the Random Forest model appears to be more effective than the multi-component linear additivity model at avoiding errors greater than 10%.

The analysis to this point indicates that Random Forest is the most predictive available model type of those considered. Some versions of the multi-component linear additivity model nearly duplicate the predictive capability of the Random Forest model. On the other hand, some of the multi-component linear additivity models are much less accurate than the Random Forest model. Moreover, no correlation exists between the number of fitting parameters and the predictive capability of the multi-component linear additivity models, implying that any benefit in representing the training data is offset by over fitting and further obscuring selection of an accurate model not guided by the analysis provided here. As a consequence, only a rigorous study on test data that the model had not used for training and as presented here can be used to identify predictive forms of the multi-component linear additivity model type.

Interestingly, the RMSE values of many of the multi-component linear additivity models cluster between 11 and 13. The reason for this clustering arises from the fact that lipid content is the single most important factor determining biocrude yield. Fig. 5 is a plot of "feature importance", as determined for the Random Forest, XGBoost, and Decision Tree algorithms. Feature importance plays a role similar to a correlation constant in a linear regression, with its value increasing as predictions become more sensitive to the values of a particular independent variable, or feature [79].

In all three cases shown in Fig. 5, lipid content is the most important feature for predicting biocrude yield. For the Decision Tree algorithm, the simplest and least accurate of the three models shown in Fig. 5, lipid content accounts for 89% of the variability observed in biocrude, a remarkable agreement with the observation that many of the RMSE values of multi-component linear additivity models cluster around 11–13.

The multi-component linear additivity models do not have a feature importance metric. However, the magnitude of the coefficients in the model plays a similar role as feature importance. Not surprisingly, the lipid coefficient in the multi-component linear additivity models is always the greatest, regardless of the number of terms present in the model. Similarly, a simple linear regression of the current data set to lipid content as the sole independent variable (training), followed by evaluating predictive accuracy using the test data, resulted in an RMSE of 11.7. Any model that accurately captures the effect of lipid content on biocrude yield can be expected to have an RMSE in the range from 11 to 13, which coincides exactly with the most frequent accuracy observed here for multi-component linear additivity models.

Fig. 5 also shows that the feature importance of the lipid terms in the more sophisticated algorithms, i.e., Random Forest and XGBoost, is

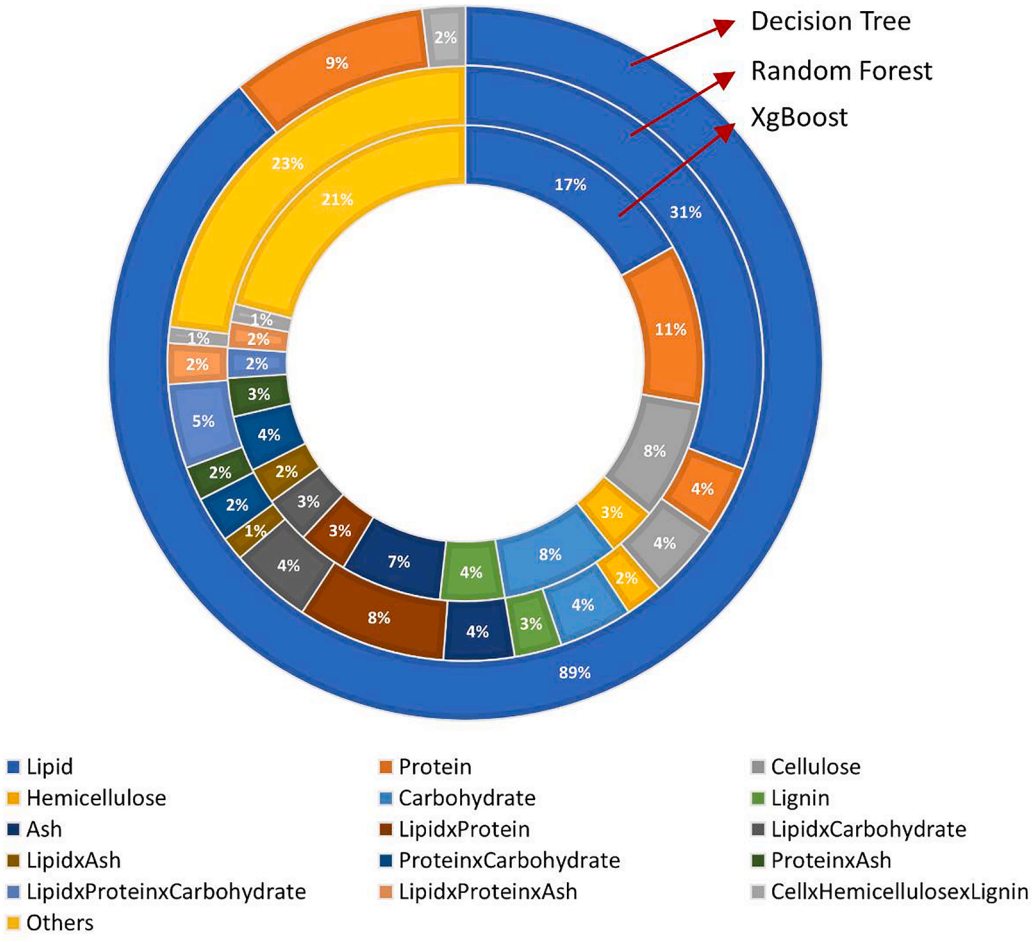

Fig. 5. Feature importance of Random Forest (RF), XGBoost, and decision tree (DT).

Table 4
Comparisons of the accuracy of biocrude yield predicted by the Random Forest regression model developed in the present study with some literature models.

Ref.	Model Type	Reported Error	# Parameters	RMSE on Current Test Set	% of Test Set predicted with <10% accuracy
Aierzhati et al. 2019 [12]	Nonlinear Regression	$R^2 = 0.983$	9	27.6	37
Deniel et al. 2017 [75]	Multi-Component Component Additivity Model	$R^2 = 0.998$	10	17.12	42
Sheng et al. 2018 [16]	Nonlinear Regression	$R^2 = 0.981$	6	12.01	62
Li et al. 2017 [76]	Linear Component Additivity Model	$R^2 = 0.884$	3	9.16	72
Leow et al. 2015 [77]	Linear Component Additivity Model	$R^2 = 0.463$	3	9.38	74
Biller et al. 2011 [78]	Linear Component Additivity Model	MAE = 1.7%	3	16.1	40
Yang et al. 2018 [17]	Multi-Component Linear Additivity Model	$R^2 = 0.9562$	7	13.03	51
Lu et al. 2018 [18]	Multi-Component Linear Additivity Model	SSE = 471	15	11.7	57
Subramanya and Savage. 2021 [19][a]	Multi-Component Linear Additivity Model	MAE = 7.84%	10	11.7	58
Teri et al. 2014 [37]	Multi-Component Linear Additivity Model	Not Reported	6	11.95	60
Teri et al. 2014 [37]	Linear Component Additivity Model	MAE = 3.2%	3	12	60
This study	Random Forest	N/A	7	8.07	81

[a] For the 326–400 Celsius degree model.

much less than observed for Decision Tree. Accordingly, Fig. 5 indicates that more accurate capturing of secondary factors, including especially protein and cellulose, reduces RMSE from 11 to 13 to roughly 8. Similarly, the more accurate predictions afforded by some of the multi-component linear additivity models [19] can be attributed to more accurate capturing of similar effects. The upshot is that, as a rule, simpler models with fewer parameters and that emphasize lipid content are preferred for predictive purposes, and that refinements should then focus on cautious addition of secondary factors to improve accuracy without overfitting.

In addition to the multi-component linear additivity models shown in Table 4, Random Forest models of biocrude yields appear in the literature [35]. Table 4 does not include predictions from previously reported Random Forest regression models [25,31,35,80]. This is because unlike the multi-component linear additivity models, a Random Forest model is not a closed form equation, which means direct inter-comparison is difficult as variance of model outcomes among different studies depends on the natures of original data set (e.g., data size and types of independent variables), the way to pre-process data set (e.g., stratification, as used here), as well as the way to split data set (e.g., if the databased include a test set that never used for training model).

While a direct comparison of the current Random Forest model without considering the prerequisite may not be entirely appropriate, a qualitative comparison is nonetheless instructive. One of the Random Forest models previously appearing in the literature reports an RMSE of 6.42 [35], an apparent improvement over the value of 8.07 reported here for a similar model. That stated, the previously reported model did not include a testing step [35], where the true accuracy of predictions for data was not included in model regression was ascertained. As a result, comparing the two RMSE values to one another is not appropriate. In fact, few bioenergy studies report machine learning performance using accuracy of test set predictions [23,30,81]. The benefit of the current study is to establish the predictive accuracy of the Random Forest method as corresponding to an RMSE value of approximately 8 (8.07, to be precise, as reported here).

3.3. Evaluating the limits of accuracy for economic projections

Regression analysis and model refinement results in predictions with accuracy of 8.07 (RMSE) (as shown in Table 4). In other words, for an actual biocrude yield of 50%, the most accurate models developed here would predict a value between 42% and 58%. The question becomes: is this level of accuracy sufficient for practical applications? Of course, the answer to this question depends on the application. A common situation, prediction of minimum fuel selling price (MFSP) using model predicted values of biocrude yields, was used as a case study. MFSP is highly dependent on biocrude yield [38,63,82,83], making the projection of MFSP and especially its corresponding uncertainty based on predicted biocrude yields and their corresponding uncertainties a practical and discerning test. Naturally, economic projections are sensitive to many factors, particularly scale, and feedstock costs [84–86]. Accordingly, all other factors were held constant during this analysis, so that the impact of the uncertainty of biocrude yield predictions could be isolated from other factors of obvious importance in a full economic analysis.

For a blind test, data for several viable HTL feeds were obtained from the literature and used as the starting point for economic analysis [57–62]. These feeds had never been used for HTL, meaning that the model was used in a predictive fashion. These feeds are shown in Table S.6 in the Supporting Information. As shown in Table S.6, biocrude yield was then predicted using the RF model, after its refinement using the modified K-fold method to reduce its RMSE to 8.07.

This economic analysis serves to demonstrate how uncertainty in modeled biocrude yields propagates in practical usage of the model – in this case, through calculation of MFSP. Biocrude yield predictions with sufficient accuracy will permit discernment between feedstock options, assuming that all other factors are held constant. In a full analysis, these other factors will not be constant, and so the outcomes presented here are limited to understanding the relationship between the accuracy of yield predictions and estimated MFSP. Further analysis, which takes into account other key factors [87,88], can then be applied for final allocation of finite resources.

To place the analysis on a common basis, a previously published economic analysis was used for estimating all costs [47]. Detailed cash flows are provided in Tables S.8-S.10, in the literature [38]. Similarly, the scale was held constant at 110 dry tons per day (DTPD) of feedstock processed. Uncertainty in the biocrude yield, as estimated by model RMSE, was then propagated through the economic analysis using a Monte Carlo simulation method, consisting of a triangular distribution around the predicted yield value with bounds +/- the RF RMSE. Maximum and minimum values of the MFSP estimated using this Monte Carlo method are used as the limiting values expected for a given feed.

The results of the accuracy analysis are summarized in Fig. 6 as a plot of estimated MFSP in $ per gallon of gasoline equivalent (GGE) as a function of predicted biocrude yield. As expected, MFSP decreases monotonically with increased biocrude yield [89]. The effect is dramatic, with the "worst" feed (in this case a sewage sludge) resulting in a MFSP more than twice that of the "best" feed (here, a type of pig manure). The horizontal error bars in Fig. 6 represent the RMSE value determined by regression analysis (±8.07) in terms of biocrude yield. The vertical error bars represent the corresponding uncertainty in MFSP

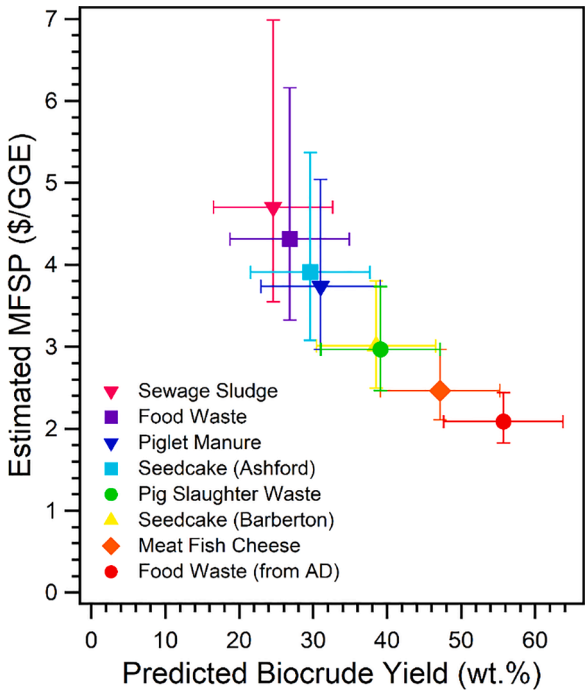

Fig. 6. The relation between estimated MFSP and biocrude yield predicted from the Random Forest regression model. None of these feed streams appear in the original data set. Biocrude yield error bar represents the RMSE obtained from analysis of the original test set (8.07). MFSP error bar represents propagation of the biocrude yield and uncertainty through an economic model.

determined via the Monte Carlo analysis.

For biocrude yields less than 35%, the uncertainty in MFSP is too great to differentiate feeds from one another. In other words, for biocrude yield less than 35%, more accurate predictions are required than afforded by the Random Forest model presented here to differentiate one feed from another based on projected MFSP. On the other hand, as predicted biocrude yield increases to values greater than 35%, the range of projected MFSPs becomes increasingly compressed, a consequence of the natural sensitivity of MSFP on biocrude yield [90]. Similar feeds – e.g., meat/fish/cheese and food waste from an anaerobic digester (AD) – cannot be differentiated solely based on predicted biocrude yields; however, the current level of accuracy is sufficient to provide a rough prediction of which feeds will be most promising for HTL. As a result, resources can be properly allocated to generate further information for only the most promising feeds, which signals the usefulness of the current model.

To provide a common basis of comparison, the economic predictions shown in Fig. 6 are based only on differences in biocrude yield. In actual situations, factors such as feedstock abundance, and hence scale, feedstock cost, tipping fees, and other techno-economic factors such as presence of impurities or foreign objects that detract from processibility, should be included in a comprehensive analysis. The models presented here allow for rudimentary understanding of yield impacts on techno-economic outcomes without performing expensive experiments, thereby allowing resources to be allocated optimally. Future work can refine the model approach by inclusion of new data and by testing it against data not used in the training, validation, or testing of the models presented here.

4. Conclusions

A data set of HTL biocrude yields consisting of 570 data points was assembled from the literature. The data set was divided into training data – used to optimize regression models – and test data – used to determine model accuracy. Then, eight different regression algorithms were evaluated for the accuracy of their biocrude yield predictions. The Random Forest and XGBoost models provided the most accurate predictions of test set data, with values of root mean square error of 8.34 and 8.67, respectively. Further refinement of the Random Forest model reduced its RMSE to 8.07, an improvement that was achieved by development of a K-fold validation method that minimized overfitting. In comparison, literature models for predicting biocrude yield were generally over fit, with corresponding values of RMSE ranging from 9.16 to 27.6. Further model analysis revealed that lipid content is the most important predictor of biocrude yield and that further improvements in accuracy are gained when secondary factors such as cellulose and protein content are accurately captured.

The absolute accuracy of the Random Forest model was evaluated by using it for making predictions of biocrude yield for a set of feeds that have never been used for HTL. These predictions were then used in a probabilistic economic model that projected minimum fuel selling price for the different feeds. All other factors were held constant in this analysis to isolate the dependence of the uncertainty of economic outcomes on the accuracy of biocrude yield predictions obtained from the Random Forest model. The accuracy of the Random Forest model was sufficient to prioritize resource allocation to development of HTL processes for different feeds based on predicted yields, with the greatest predictive capability found for the most economically viable feeds.

Declaration of Competing Interest

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

Acknowledgements

This work was funded by the DOE Bioenergy Technology Office (DE-EE0008513), the Massachusetts Clean Energy Center (MassCEC), and the U.S. National Science Foundation (#2021871). Dr. N. Aaron Deskins, Department of Chemical Engineering, WPI, and Mr. Jian, Jiamin, Department of Mathematical Sciences, WPI, provided helpful suggestions for conceptualization and methodology.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2022.136013.

References

[1] E. Newburger, Here's what countries pledged on climate change at Biden's global summit, in, Consumer News and Business Channel, Englewood Cliffs, NJ, U.S.A., 2021.
[2] DOE, DOE Announces $61.4 Million for Biofuels Research to Reduce Transportation Emissions, in, Department of Energy, Washington, D.C., U.S., 2021.
[3] S. Harman, How We're Moving to Net-Zero by 2050, in, Department of Energy, Washington, D.C., U.S., 2021.
[4] R. Pielke, Net-Zero Carbon Dioxide Emissions By 2050 Requires A New Nuclear Power Plant Every Day, in, Forbes, Jersey City, NJ, U.S., 2019.
[5] J. Zhang, X. Zhang, The thermochemical conversion of biomass into biofuels, in: Biomass Biopolymer-Based Mater. Bioenerg., Elsevier, 2019, pp. 327-368.
[6] S.Y. Lee, R. Sankaran, K.W. Chew, C.H. Tan, R. Krishnamoorthy, D.-T. Chu, P.-L. Show, Waste to bioenergy: a review on the recent conversion technologies, BMC Energy 1 (2019) 1–22.
[7] A. Dimitriadis, S. Bezergianni, Hydrothermal liquefaction of various biomass and waste feedstocks for biocrude production: a state of the art review, Renew. Sust. Energ. Rev. 68 (2017) 113–125.
[8] B. de Caprariis, P. De Filippis, A. Petrullo, M. Scarsella, Hydrothermal liquefaction of biomass: influence of temperature and biomass composition on the bio-oil production, Fuel 208 (2017) 618–625.
[9] S.S. Toor, L. Rosendahl, A. Rudolf, Hydrothermal liquefaction of biomass: a review of subcritical water technologies, Energy 36 (2011) 2328–2342.
[10] Y. Nie, X.T. Bi, Techno-economic assessment of transportation biofuels from hydrothermal liquefaction of forest residues in British Columbia, Energy 153 (2018) 464–475.

[11] T.H. Pedersen, N.H. Hansen, O.M. Pérez, D.E.V. Cabezas, L.A. Rosendahl, Renewable hydrocarbon fuels from hydrothermal liquefaction: a techno-economic analysis, Biofuels Bioprod. Biorefining 12 (2) (2018) 213–223.
[12] A. Aierzhati, M.J. Stablein, N.E. Wu, C.-T. Kuo, B. Si, X. Kang, Y. Zhang, Experimental and model enhancement of food waste hydrothermal liquefaction with combined effects of biochemical composition and reaction conditions, Bioresour. Technol. 284 (2019) 139–147.
[13] F. Cheng, Z. Cui, K. Mallick, N. Nirmalakhandan, C.E. Brewer, Hydrothermal liquefaction of high-and low-lipid algae: mass and energy balances, Bioresour. Technol. 258 (2018) 158–167.
[14] J.D. Adjaye, N. Bakhshi, Catalytic conversion of a biomass-derived oil to fuels and chemicals I: Model compound studies and reaction pathways, Biomass Bioenerg. 8 (1995) 131–149.
[15] S. He, J. Wang, Z. Cheng, H. Dong, B. Yan, G. Chen, Synergetic effect and primary reaction network of corn cob and cattle manure in single and mixed hydrothermal liquefaction, J. Anal. Appl. Pyrolysis 155 (2021), 105076.
[16] L. Sheng, X. Wang, X. Yang, Prediction model of biocrude yield and nitrogen heterocyclic compounds analysis by hydrothermal liquefaction of microalgae with model compounds, Bioresour. Technol. 247 (2018) 14–20.
[17] J. Yang, Q.(. He, H. Niu, K. Corscadden, T. Astatkie, Hydrothermal liquefaction of biomass model components for product yield prediction and reaction pathways exploration, Appl. Energy 228 (2018) 1618–1628.
[18] J. Lu, Z. Liu, Y. Zhang, P.E. Savage, Synergistic and antagonistic interactions during hydrothermal liquefaction of soybean oil, soy protein, cellulose, xylose, and lignin, ACS Sustain. Chem. Eng. 6 (11) (2018) 14501–14509.
[19] S. Mahadevan Subramanya, P.E. Savage, Identifying and modeling interactions between biomass components during hydrothermal liquefaction in sub-, near-, and supercritical water, ACS Sustain, Chem. Eng. 9 (41) (2021) 13874–13882.
[20] D.C. Hietala, P.E. Savage, A molecular, elemental, and multiphase kinetic model for the hydrothermal liquefaction of microalgae, Chem. Eng. J. 407 (2021) 127007.
[21] J.D. Sheehan, P.E. Savage, Modeling the effects of microalga biochemical content on the kinetics and biocrude yields from hydrothermal liquefaction, Bioresour. Technol. 239 (2017) 144–150.
[22] M.R. Forster, Key concepts in model selection: Performance and generalizability, J. Math. Psychol. 44 (1) (2000) 205–231.
[23] X. Zhu, Y. Li, X. Wang, Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions, Bioresour. Technol. 288 (2019), 121527.
[24] J. Li, L. Pan, M. Suvarna, Y.W. Tong, X. Wang, Fuel properties of hydrochar and pyrochar: prediction and exploration with machine learning, Appl. Energy 269 (2020), 115166.
[25] T. Zhang, D. Cao, X. Feng, J. Zhu, X. Lu, L. Mu, H. Qian, Machine learning prediction of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions, Fuel 312 (2022), 122812.
[26] A. Ghorbani, J. Zou, Data shapley: Equitable valuation of data for machine learning, in, International Conference on Machine Learning, PMLR (2019) 2242–2251.
[27] X. Dastile, T. Celik, M. Potsane, Statistical and machine learning models in credit scoring: a systematic literature survey, Appl. Soft Comput. J. 91 (2020) 106263.
[28] D.R. Stockwell, A.T. Peterson, Effects of sample size on accuracy of species distribution models, Ecol. Modell. 148 (2002) 1–13.
[29] H. Jabbar, R.Z. Khan, Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study), in: J. Stephen, H. Rohil, V. S (Eds.) Computer Science, Communication and Instrumentation Devices, Research Publishing, 2015, pp. 163-172.
[30] A. Pathy, S. Meher, P. Balasubramanian, Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods, Algal Res. 50 (2020), 102006.
[31] Q. Tang, Y. Chen, H. Yang, M. Liu, H. Xiao, Z. Wu, H. Chen, S.R. Naqvi, Prediction of bio-oil yield and hydrogen contents based on machine learning method: effect of biomass compositions and pyrolysis conditions, Energy Fuels 34 (2020) 11050–11060.
[32] E. Kaiser, J.N. Kutz, S.L. Brunton, Sparse identification of nonlinear dynamics for model predictive control in the low-data limit, Proc. R. Soc. A 474 (2018) 20180335.
[33] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, N. Khovanova, Machine learning for predictive modelling based on small data in biomedical engineering, IFAC-Pap. 48 (2015) 469–474.
[34] A. Tulsyan, C. Garvin, C. Ündey, Advances in industrial biopharmaceutical batch process monitoring: machine-learning methods for small data problems, Biotechnol. Bioeng. 115 (2018) 1915–1924.
[35] F. Cheng, M.D. Porter, L.M. Colosi, Is hydrothermal treatment coupled with carbon capture and storage an energy-producing negative emissions technology? Energy Convers. Manag. 203 (2020), 112252.
[36] T. Katongtung, T. Onsree, N. Tippayawong, Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes, Bioresour. Technol. 344 (2022), 126278.
[37] G. Teri, L. Luo, P.E. Savage, Hydrothermal treatment of protein, polysaccharide, and lipids alone and in mixtures, Energy Fuels 28 (2014) 7501–7509.
[38] L.J. Snowden-Swan, Y. Zhu, M.D. Bearden, T.E. Seiple, S.B. Jones, A.J. Schmidt, J. M. Billing, R.T. Hallen, T.R. Hart, J. Liu, Conceptual Biorefinery Design and Research Targeted for 2022: Hydrothermal Liquefacation Processing of Wet Waste to Fuels, in, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2017.
[39] W.-T. Chen, Y. Zhang, J. Zhang, L. Schideman, G. Yu, P. Zhang, M. Minarick, Co-liquefaction of swine manure and mixed-culture algal biomass from a wastewater treatment system to produce bio-crude oil, Appl. Energy 128 (2014) 209–216.
[40] B. Zhang, J. Chen, S. Kandasamy, Z. He, Hydrothermal liquefaction of fresh lemon-peel and Spirulina platensis blending-operation parameter and biocrude chemistry investigation, Energy 193 (2020), 116645.
[41] Y. Hu, S. Wang, J. Li, Q. Wang, Z. He, Y. Feng, A.-E.-F. Abomohra, S. Afonaa-Mensah, C. Hui, Co-pyrolysis and co-hydrothermal liquefaction of seaweeds and rice husk: comparative study towards enhanced biofuel production, J. Anal. Appl. Pyrolysis 129 (2018) 162–170.
[42] Z. Liu, F.-S. Zhang, Effects of various solvents on the liquefaction of biomass to produce fuels and chemical feedstocks, Energy Convers. Manag. 49 (2008) 3498–3504.
[43] P. Biller, R.B. Madsen, M. Klemmer, J. Becker, B.B. Iversen, M. Glasius, Effect of hydrothermal liquefaction aqueous phase recycling on bio-crude yields and composition, Bioresour. Technol. 220 (2016) 190–199.
[44] A.A. Shah, S.S. Toor, T.H. Seehar, R.S. Nielsen, A.H. Nielsen, T.H. Pedersen, L. A. Rosendahl, Bio-crude production through aqueous phase recycling of hydrothermal liquefaction of sewage sludge, Energies 13 (2020) 493.
[45] F. Wang, Z. Chang, P. Duan, W. Yan, Y. Xu, L. Zhang, J. Miao, Y. Fan, Hydrothermal liquefaction of Litsea cubeba seed to produce bio-oils, Bioresour. Technol. 149 (2013) 509–515.
[46] C. Yang, S. Wang, M. Ren, Y. Li, W. Song, Hydrothermal liquefaction of an animal carcass for biocrude oil, Energy Fuels 33 (2019) 11302–11309.
[47] Y. Zhu, S. Jones, D. Anderson, R. Hallen, A. Schmidt, K. Albrecht, D. Elliott, Techno-economic Analysis of Whole Algae Hydrothermal Liquefaction (HTL) and Upgrading System, Pacific Northwest National Laboratory. Richland, WA, USA, 2015.
[48] R.A. Bottenberg, J.H. Ward, Applied multiple linear regression, 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Lackland Air Force Base, 1963.
[49] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[50] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Series B Stat. Methodol. 58 (1996) 267–288.
[51] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[52] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[53] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, 2009.
[54] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (2002) 18–22.
[55] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[56] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, in: L. Liu, M.T. Özsu (Eds.) Encyclopedia of Database Systems, Springer, Boston, MA, 2009, pp. 532-538.
[57] S. Bayr, J. Rintala, Thermophilic anaerobic digestion of pulp and paper mill primary sludge and co-digestion of primary and secondary sludge, Water Res. 46 (2012) 4713–4720.
[58] Y. Li, H. Liu, K. Xiao, X. Liu, H. Hu, X. Li, H. Yao, Correlations between the physicochemical properties of hydrochar and specific components of waste lettuce: Influence of moisture, carbohydrates, proteins and lipids, Bioresour. Technol. 272 (2019) 482–488.
[59] L. Alibardi, R. Cossu, Effects of carbohydrate, protein and lipid content of organic waste on hydrogen production and fermentation products, Waste Manag. 47 (2016) 69–77.
[60] A. Fekria, A. Isam, O. Suha, E. Elfadil, Nutritional and functional characterization of defatted seed cake flour of two Sudanese groundnut (Arachis hypogaea) cultivars, Int. Food Res. J. 19 (2012).
[61] T.T.T. Cu, T.X. Nguyen, J.M. Triolo, L. Pedersen, V.D. Le, P.D. Le, S.G. Sommer, Biogas production from Vietnamese animal manure, plant residues and organic waste: influence of biomass composition on methane yield, Asian-Australas. J. Anim. Sci. 28 (2) (2015) 280–289.
[62] S. Xue, Y. Wang, X. Lyu, N. Zhao, J. Song, X. Wang, G. Yang, Interactive effects of carbohydrate, lipid, protein composition and carbon/nitrogen ratio on biogas production of different food wastes, Bioresour. Technol. 312 (2020), 123566.
[63] L. Ou, R. Thilakaratne, R.C. Brown, M.M. Wright, Techno-economic analysis of transportation fuels from defatted microalgae via hydrothermal liquefaction and hydroprocessing, Biomass Bioenerg. 72 (2015) 45–54.
[64] V. Molak, Fundamentals of Risk Analysis and Risk Management, CRC Press, 1996.
[65] L.-C. Ma, B. Castro-Dominguez, N.K. Kazantzis, Y.H. Ma, A cost assessment study for a large-scale water gas shift catalytic membrane reactor module in the presence of uncertainty, Sep. Purif. Technol. 166 (2016) 205–212.
[66] M.S. Peters, K.D. Timmerhaus, R.E. West, Plant Design and Economics for Chemical Engineers, McGraw-Hill New York, 2003.
[67] L.-C. Ma, B. Castro-Dominguez, N.K. Kazantzis, Y.H. Ma, Integration of membrane technology into hydrogen production plants with CO2 capture: an economic performance assessment study, Int. J. Greenh. Gas Control. 42 (2015) 424–438.
[68] M. Aghbashlo, W. Peng, M. Tabatabaei, S.A. Kalogirou, S. Soltanian, H. Hosseinzadeh-Bandbafha, O. Mahian, S.S. Lam, Machine learning technology in biodiesel research: a review, Prog. Energy Combust. Sci. 85 (2021) 100904.
[69] Y. Zhang, C. Ling, A strategy to apply machine learning to small datasets in materials science, Npj Comput. Mater. 4 (2018) 1–8.

[70] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, E. Hernandez-Lemus, Machine learning algorithm validation with a limited sample size, PloS one 14 (11) (2019) e0224365.
[71] G.J. Bowden, H.R. Maier, G.C. Dandy, Optimal division of data for neural network models in water resources applications, Water Resour. Res. 38 (2002) 2-1-2-11.
[72] T. Shaikhina, N.A. Khovanova, Handling limited datasets with neural networks in medical applications: a small-data approach, Artif. Intell. Med. 75 (2017) 51–63.
[73] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995, pp. 1137-1145.
[74] J. Yang, L. Yang, A review on hydrothermal co-liquefaction of biomass, Appl. Energy 250 (2019) 926–945.
[75] M. Déniel, G. Haarlemmer, A. Roubaud, E. Weiss-Hortala, J. Fages, Modelling and predictive study of hydrothermal liquefaction: application to food processing residues, Waste Biomass Valorization 8 (6) (2017) 2087–2107.
[76] Y. Li, S. Leow, A.C. Fedders, B.K. Sharma, J.S. Guest, T.J. Strathmann, Quantitative multiphase model for hydrothermal liquefaction of algal biomass, Green Chem. 19 (2017) 1163–1174.
[77] S. Leow, J.R. Witter, D.R. Vardon, B.K. Sharma, J.S. Guest, T.J. Strathmann, Prediction of microalgae hydrothermal liquefaction products from feedstock biochemical composition, Green Chem. 17 (2015) 3584–3599.
[78] P. Biller, A. Ross, Potential yields and properties of oil from the hydrothermal liquefaction of microalgae with different biochemical content, Bioresour. Technol. 102 (2011) 215–225.
[79] G. Casalicchio, C. Molnar, B. Bischl, Visualizing the feature importance for black box models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2018, pp. 655–670.
[80] J. Li, W. Zhang, T. Liu, L. Yang, H. Li, H. Peng, S. Jiang, X. Wang, L. Leng, Machine learning aided bio-oil production with high energy recovery and low nitrogen content from hydrothermal liquefaction of biomass with experiment verification, Chem. Eng. J. 425 (2021), 130649.
[81] P.J. García Nieto, E. Garcia-Gonzalo, J.P. Paredes-Sánchez, A. Bernardo Sánchez, M. Menendez Fernandez, Predictive modelling of the higher heating value in biomass torrefaction for the energy treatment process using machine-learning techniques, Neural. Comput. Appl. 31 (2019) 8823–8836.
[82] Y. Jiang, S.B. Jones, Y. Zhu, L. Snowden-Swan, A.J. Schmidt, J.M. Billing, D. Anderson, Techno-economic uncertainty quantification of algal-derived biocrude via hydrothermal liquefaction, Algal Res. 39 (2019), 101450.
[83] Y. Zhu, M.J. Biddy, S.B. Jones, D.C. Elliott, A.J. Schmidt, Techno-economic analysis of liquid fuel production from woody biomass via hydrothermal liquefaction (HTL) and upgrading, Appl. Energy 129 (2014) 384–394.
[84] M.J. Biddy, R. Davis, S.B. Jones, Y. Zhu, Whole algae hydrothermal liquefaction technology pathway, in, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2013.
[85] P. Ranganathan, S. Savithri, Techno-economic analysis of microalgae-based liquid fuels production from wastewater via hydrothermal liquefaction and hydroprocessing, Bioresour. Technol. 284 (2019) 256–265.
[86] M. Kumar, A.O. Oyedun, A. Kumar, A comparative Technoeconomic analysis of algal thermochemical conversion technologies for diluent production, Energy Technol. 8 (2020) 1900828.
[87] M.M. Wright, D.E. Daugaard, J.A. Satrio, R.C. Brown, Techno-economic analysis of biomass fast pyrolysis to transportation fuels, Fuel 89 (2010) S2–S10.
[88] L.Y. Batan, G.D. Graff, T.H. Bradley, Techno-economic and Monte Carlo probabilistic analysis of microalgae biofuel production system, Bioresour. Technol. 219 (2016) 45–52.
[89] J.R. Collett, J.M. Billing, P.A. Meyer, A.J. Schmidt, A.B. Remington, E.R. Hawley, B.A. Hofstad, E.A. Panisko, Z. Dai, T.R. Hart, Renewable diesel via hydrothermal liquefaction of oleaginous yeast and residual lignin from bioconversion of corn stover, Appl. Energy 233 (2019) 840–853.
[90] Y. Zhu, S.B. Jones, A.J. Schmidt, K.O. Albrecht, S.J. Edmundson, D.B. Anderson, Techno-economic analysis of alternative aqueous phase treatment methods for microalgae hydrothermal liquefaction and biocrude upgrading system, Algal Res. 39 (2019), 101467.