Machine learning for hydrothermal liquefaction: prediction of bio-oil production

Nalluri Rishi Chaitanya Sri Prasad[1], Rohith Sai Rayapati[1], Prabakaran Ganeshan[2], Karthik Rajendran[2 *]

[1] Department of Computer Science and Engineering, School of Engineering and Sciences, SRM University-AP, Amaravati, Andhra Pradesh 522240, India.

[2] Department of Environmental Science, School of Engineering and Sciences, SRM University-AP, Amaravati, Andhra Pradesh 522240, India.

Corresponding author Email: rajendran.k@srmap.edu.in

Highlights

- Bio-oil yield, HHV, ERR are predicted by using machine learning models.
- Proximate analysis and operational conditions are considered for the prediction of oil yield.
- Oil yield was predicted with an accuracy of 89% using Random Forest algorithm.
- RF algorithm is the best model for the prediction of oil yield.

Abstract

Hydrothermal Liquefaction (HTL) is a sustainable approach to produce bio-oil from biomass (microalgae). But the identification and optimization of the process parameters by experimentation is not only a time-consuming process but also needs a huge workforce.

However, machine learning (ML) algorithms such as Linear regression (LR), Random Forest (RF), and Decision tree (DT) can be applied to identify the critical process parameters and bio-oil production rate. The input parameters for ML algorithms include proximate analysis, elemental composition, the biochemical composition of the feedstock, and operating conditions. More than 100 data were collected from the literature with a maximum of 18 input features. The dataset was spilt into four based on the feature importance and Pearson correlation matrix and it's used for all ML algorithms. Among the three algorithms, RF was found to be the best for bio-oil prediction using dataset 1 with 6 input features (R^2 = 0.84 and RMSE = 0.01). Meanwhile, the best-predicted model was validated with new data and the difference ranged between -1.3% to +7.8%. Datasets 1 & 4 were validated statistically, and then the P value was >0.05, which showed an insignificant difference for the prediction of oil yield. Feature importance implies that bio-oil production follows temperature, time, and pressure.

Keywords:

Bio- oil; Microalgae; Hydrothermal liquefaction; Prediction and optimization; Machine learning

1. Introduction

The global demand for crude oil increased by 5.2% in 2021 compared to 2020 levels (91 million barrels). The transport sector accounted for approximately 37% of the fuels (gasoline, jet fuels, petroleum, diesel, etc.) derived from crude oil in 2021, despite the impact of the COVID-19 pandemic on this sector. However, the carbon emissions generated during the usage of fuel derived from crude oil and the depleting fossil fuel resources have forced nations

worldwide to seek alternative energy sources capable of generating crude oil from renewable sources to enable sustainable development. In this context, biocrude oil generated from microalgae is expected to aid in overcoming the challenges due to its eco-friendliness. The ability of microalgae to grow in wastewater and saline water by utilizing the available nutrients prevents changes in land usage patterns. Additionally, microalgae also act as a carbon sink, thereby aiding in achieving global climate goals cost-effectively. Two methods are generally adopted to extract biocrude oil from microalgae, namely pyrolysis and the Hydrothermal Liquefaction (HTL) process (Gollakota et al., 2018). Among these, the HTL process, which occurs at high pressure (5-30 MPa) and temperature (180-400 °C), has been found to be efficient (Katongtung et al., 2022). The energy density of biocrude oil produced by this method is twice that of the pyrolysis method.

Table 1. Similar studies reported in literatures.

Algorithm used	Parameter considered	No. of parameters	Best Algorithm	Emphasis of the study	Result (R^2)	Author(s)
RF, SVM, DT, MLR	Proximate Analysis, Ultimate Analysis, Target variable	13	RF	Comparative study of ML methods for bio-oil yield prediction	0.98	(Ullah et al., 2021)
GBR, RF	Ultimate composition, atomic ratio, Biochemical composition, Operating conditions, bio-oil, bio-oil composition	20	GBR	Machine learning prediction and optimization of bio-oil production from HTL	0.88	(Zhang et al., 2021)
RF	Biomass characteristics, pyrolysis conditions, prediction targets	17	RF	Prediction of oil yield and oxygen content via biomass and pyrolysis conditions	0.92	(Yang et al., 2022)

RF	Biomass characteristics, pyrolysis conditions, Bio-oil characteristics	20	RF	ML prediction of bio-oil yield related to biomass and pyrolysis conditions	0.93	(Zhang et al., 2022)
XGB, RF, KRR, SVR	Feedstock characteristic, biological property, Elemental property, Operating condition, Output target	17	XGB	ML prediction of biocrude yields and HHV from HTL of wet biomass and wastes.	0.87	(Katongtung et al., 2022)

The HTL process involves a simple numerical analysis to address all the variables. Machine learning, being a statistical analysis model, is well-suited for predicting the biocrude oil yield, as conducting real-time experiments would be costly. Determining the optimum values for predicting higher oil yield is challenging; however, machine learning offers a solution. By collecting real-time experimental data under different conditions and considering all relevant features, a model can be developed using machine learning algorithms, saving both time and money. The machine learning approach can predict the yield approximately equivalent to the experimental yield.

Several similar studies have been conducted by other researchers (Ullah et al., 2021). employed a genetic-based algorithm for predicting bio-oil yield. Among the algorithms used in their study (RF, SVM, DT, MLR), RF demonstrated superior performance with an $R^2 = 0.98$. The pyrolysis method was used, and a Graphical User Interface (GUI) was developed for the model, facilitating better understanding. The test size was 0.2, with 12 input features and one output feature (Zhang et al., 2021). developed the HTL process from microalgae feedstock, identifying the Gradient Boost Regression (GBR) algorithm as the best for predicting oil yield, with an R^2 =

0.88. Four datasets were considered, and dataset 1 showed the best model performance for predicting oil yield, with 19 input features and one output feature (Yang et al., 2022). used the pyrolysis method, considering the oxygen content of bio-oil. RF was the only algorithm used, achieving an $R^2 = 0.92$ with 16 input features and one output feature (Zhang et al., 2022) discussed bio-oil yield related to biomass using pyrolysis conditions, with 19 total features and one output feature for predicting oil yield. RF was used, achieving an $R^2 = 0.93$ (Katongtung et al., 2022). Predicted yield and HHV of biocrude from the HTL process, using the Xgboost algorithm with an $R^2 = 0.87$ and 17 input features. Biomass characteristics accounted for over 60% of the model predictions, with three cases of input features.

Each study considered either the pyrolysis or HTL process, collecting data from different sources and using various feedstocks for oil yield production. Machine learning was consistently employed for its efficiency and accuracy in predicting oil yield compared to physical experiments. Advanced algorithms such as RF, DT, SVM, MLR, XGB, and GBR were used, along with hyperparameter tuning for optimization. No published work has yet focused on developing an ML model for high accuracy using only proximate analysis and operational conditions, as elemental composition, although effective, is expensive and requires large equipment. Using only proximate analysis and operational conditions with the HTL process can yield results approximately equivalent to those obtained with other input features.

Therefore, the present work focuses on predicting oil yield using the HTL process with microalgae feedstock. The ML algorithms used for model development include Linear Regression, Random Forest, and Decision Trees. The original dataset comprises 18 input features, including biochemical composition, elemental composition, proximate analysis, and

operational conditions, with three output features: HHV, ERR, and oil yield. Hyperparameter tuning is also performed to identify the best values for increasing the learning process of oil yield. The input features are divided into four datasets by feature importance, and the prediction of oil yield from the HTL process is explained and visualized using libraries such as seaborn in Python for better graph understanding. The entire process of this study is clearly illustrated in Fig. 3.

2. Methodology

2.1 Data Collection and pre-processing

The dataset was collected from the literature and articles from Scopus and keywords are bio-oil, microalgae, HTL (Hydrothermal Liquefaction), and Machine learning. The data were collected from the experimental study published by the researchers in journals. Following which these data values were used to form a dataset (Ayesha Jasmin et al., 2022; Biller et al., 2012; Li et al., 2021; López Barreiro et al., 2013; Shakya et al., 2017; Sharma et al., 2021; Yang et al., 2004). This dataset has a total of 100 experimental data. In detail, the input features are as follows elemental composition, biochemical composition, operational conditions, and proximate analysis. The output features are as follows properties of bio-oil (i.e., Oil yield, HHV, ERR). The key features for the prediction of oil yield are as follows elemental composition (i.e., C, H, N, O, S), proximate analysis (i.e., moisture, volatile, ash), and operational conditions (i.e., time, temperature, pressure). The other two output features are HHV and ERR. This HHV is calculated by using a formula that has components like (C, H, N, and O) for the evaluation [8]. The HHV is calculated by using equation (1) The ERR is calculated by using equation (2).

$$HHV(MJ/kg) = 0.3516 \times C + 1.16225 \times H - 0.1109 \times O + 0.0628 \times N \quad (1)$$

$$ERR\ oil\ (\%) = (HHV_{oil} * yield\ oil / HHV_{biomass}) \quad (2)$$

From the original dataset, three datasets were divided based on feature importance. In these four datasets, the original data is dataset 4 which includes all features for the prediction of bio-oil yield. Initially, dataset 4 has divided into dataset 3 by removing the feature of proximate analysis (i.e., moisture, volatile, ash) in the same way the biochemical composition and proximate analysis were removed in dataset 2 from dataset 4 finally, in dataset 1 elemental composition and biochemical composition were removed because elemental composition features are directly correlated to the yield then it is obvious that when this feature is present then Oil-yield % production is more. The minimum and maximum values of each feature in every dataset are clearly shown in Table 2.

After completing the datasets then the next step is to send the dataset to ML training models. This dataset was split into the training set and testing test by importing the train_test_split package in the scikit_learn library with python programming. The libraries used for the graphs are Matplotlib and seaborn. These libraries give an exceptionally good understanding of the data in the form of graphs and statistical analysis of the data. The training and testing ratio was divided into an 80:20 ratio. 80% for the training set and 20% for the testing set. In each dataset used for the final evaluation of the trained ML models with best hyperparameters. The algorithms used for these datasets are Random Forest, Decision Trees, and Linear Regression. All these models are regression problems that are used for predicting continuous variables in the datasets.

2.2 Machine Learning Algorithms

i. Linear Regression:

Linear regression is the basic algorithm in machine learning. Linear regression is used when the dataset is having at least one independent variable and one dependent variable. Independent variable is referred that the variable used to predict the other variable. The dependent variable is referred to as the variable you want to predict. In linear regression, there are two regressions i.e., simple linear regression and multi-linear regression. Simple linear regression means that it has one independent and one dependent. Multi-Linear regression means that it can have many independent variables and at least one output variable. The basic equation for the linear regression for both simple and multi-linear regression (3) is.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \qquad (3)$$

For each observation i=1,2……., n

This equation can also be given as (4):

$$y = \theta_1 + \theta_2 x \qquad (4)$$

θ_1=intercept, θ_2= Coefficient of x

This equation is used because suppose we have taken the data points in the graph with X and Y with some variable. Then by this equation, we can draw a prediction line in the graph that we can calculate the predicted values concerning the line it will fit that prediction line in such a way that the distance between the points and the predicted line must be minimum. It is especially important to update the and values, to reach the best value

minimize the error difference between the predicted and actual values. The cost function equation

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_0(x)^i - y^i\right)^2 \quad (5)$$

The cost function of linear regression is the RMSE (Root Mean Square Error) between predicted and actual values. The learning rate of linear regression is 0.01 because to have the best-fit line the error between actual and predicted should be less to search for the less error it has to reach the Global minima in the gradient descent of convex graph then, we can have the best model for the prediction.

ii. Decision Trees:

A Decision Tree is an algorithm that comes under supervised learning in machine learning. This algorithm, the Decision tree can be used in both solving classification and regression problems. The main aim of using decision trees is to train a model that can be used to predict the output variable. Decision trees have many algorithms to decide to split the trees into two or many trees. The purity of a node increases concerning the target variable. This entropy ranges from 0 to 1. If the split was all yes or no, then we can say that the split is pure split or else it is an impure split. It calculates the entropy(H) and information gain (IG). The purity test is done by entropy. Calculate the entropy(H) the equation is:

$$H(s) = -P_+ \, log_2 \, P_+ - P_- \; log_2 \, P_- \quad (6)$$

P_+ = Probability of yes, P_- = Probability of No, H(S) = Entropy

Gini impurity is the function shows that how well a decision tree was split. Gini impurity value ranges from 0 to 0.5. If the dataset has more rows in it, then we use Gini impurity and if it is fewer data then we use entropy. The equation of the Gini impurity is:

$$G.I = 1 - \sum_{i=1}^{n} (P)^2 \qquad (7)$$

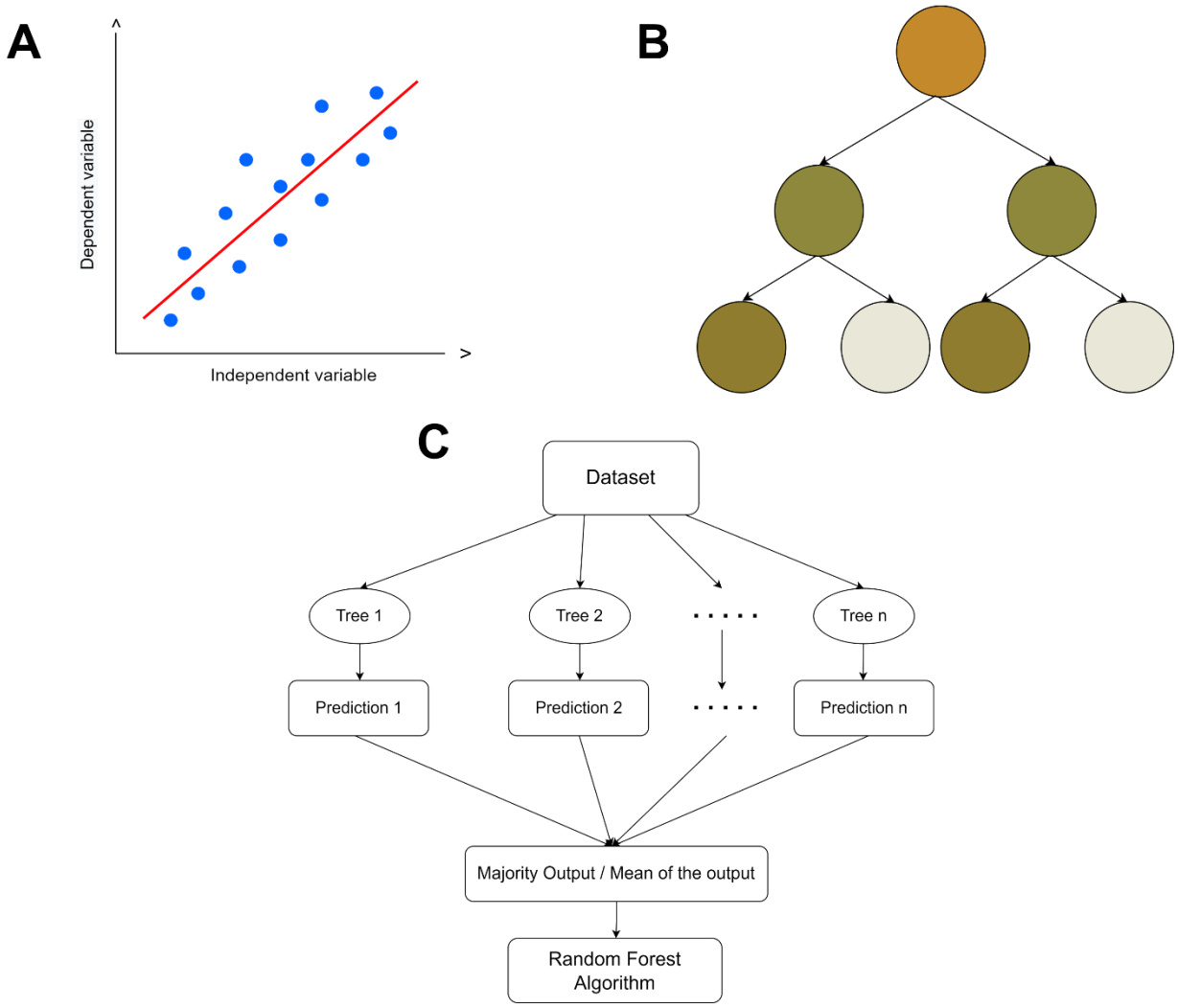
$$G.I = 1 - ((P_+)^2 + (P_-)^2) \qquad (8)$$

P_+ = Probability of yes, P_- = Probability of No, n = no.of outputs, G.I = Gini impurity

We have two methods to optimize decision trees post pruning and pre-pruning. This must be done because when the dataset is large then it leads to overfitting and a small tree may not capture all the important features of the dataset. In post pruning supposes, let us think there are 7 yes and 2 No then we can say that maximum is yes then there is no need of dividing it into subtrees then post pruning can be done. Pre-pruning or early stopping involves stopping the tree before it has completed in this, we will take the hyperparameters in this method like Max_depth and Max_leaf and so on. These values are given randomly by using a library in scikit_learn from importing GridsearchCV in Python. Then it gives the best parameters. The main disadvantages of the decision trees are to lead to overfitting I.e., low bias and high variance. If the train data runs well then it is low bias and if the test data doesn't run well then it is high variance. This decision tree has this issue that can be solved by the Random Forest algorithm.

iii. Random Forest:

Random forest is an advanced machine learning algorithm. Random forest uses an ensemble method for both classification and regression problems. This algorithm divides into many decision trees and uses an ensemble technique called bagging. Bagging is the technique that divides into multiple trees and the output of all the trees is taken and the majority voting of the output is taken as the final output. Bagging is also called bootstrap aggregation. This random forest solves two problems classification and regression. For this dataset we have used this as a continuous variable so, a random forest regressor is used to solve this dataset. As we know, a decision tree leads to overfitting (Low bias and High variance) using this random forest algorithm this should be converted to Low bias and Low variance then it becomes the generalized model for the prediction. In this Random Forest, the original dataset is divided into multiple decision trees. In this method, it doesn't lead to overfitting i.e., low bias and low variance. If the dataset is a classification problem, then this technique will be choosing output as the majority voting among the subtree's outputs. If the dataset is a regressor problem, then the output is given by taking the mean of all the subtree outputs. Control the process of the model performance some hyperparameters can be used to control the model learning process to achieve the best accuracy of the model. The main hyperparameters are max_depth and n_estimators. Max_depth means to decide how much depth the tree should be. N_estimators is diving the number of trees we want for the majority voting or the mean of all outputs. So, that we can get the best parameters and the best accuracy score for the model.



(a) - Linear Regression, (b) - Decision Trees, (C) - Random Forest algorithms

Fig 1. ML algorithms graphical representation

2.3 Model Evaluation

1. *RMSE:*

Mean square error is used to find the squared difference between actual and predicted values of the output variable. If the difference between actual and predicted

values is higher the root means the square value is higher. It helps to find the difference between actual and predicted. The RMSE helps to find how the actual data points are closer to the predicted data points. If the RMSE is lower as much as possible then we can say that the difference between true and predicted values is very less. The RMSE equation (9) is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (9)$$

y_i = Actual values, $\hat{y}_i$ = Predicted values, n = Number of data points.

2. *MAE:*

MAE (Mean Absolute Error) is the mean of the absolute difference between the true value and the predicted value. Mean square error increases with an increase in the difference between true and predicted. MAE uses the loss function for regression problems. MAE is the metric that shows how exactly the predictions are shown and what is the deviation from the true values. The MAE equation (10) is as follows:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|^2}{n} \quad (10)$$

y_i = Actual values, $\hat{y}_i$ = Predicted values, n = Number of data points

3. *R^2:*

The R2 is an especially important metric that is used to perform regression problems in machine learning. R2 is calculated by the difference between the actual and predicted values. If the highest R2 value is 1 then we can say the model is perfect. The R2 value

can be negative based on the difference between the actual and predicted value. when the R2 value is as close as 1 then we can say it has the best-fit model. The R2 equation (11) is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_{avg})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (11)$$

y_i = True value, y_{avg} = Predicted value, $\hat{y}_i$= mean value, n = Number of Data points

2.4 Hyperparameter Optimization:

Hyperparameter tuning is the method in machine learning, and hyperparameter optimization has a method of choosing a set of hyperparameters for learning the algorithm. A hyperparameter is a value used to control the learning process of the ML model. As we know from these datasets, we have used three machine learning algorithms based on train test split. 80% of our data is sent to training and 20% of the data is sent to testing. The max_depth of decision trees was tuned, and the average R2 was increased with the increase of max_depth of a certain limit. In random forest, the hyperparameter used is (max_depth, n_estimators). In the Decision tree, we have considered (max_depth, min_samples_split, and min_samples_leaf) as a hyperparameter. The hyperparameters are given in such a way that the ranges are given so that it can choose the best parameters for the hyperparameter tuning. This is clearly shown in Table 3. These values are given in the results and discussion part i.e., Hyperparameter Tuning.

2.5 Outliers of the features:

Outliers are the data points that are above the maximum or below the minimum in percentiles. which are not related to the other data points. These outlier data

points are visualized using the boxplot function in the Pandas library. These will be shown like black circles above or below the box plot (Fig 2). Outliers will cause problems during fitting the model mainly in linear models and gives errors in metrics i.e., Model evaluation (R2, RMSE, MAE) as well as the weights will be higher like in boosting techniques will calculate weights to become strong learner by the learning the patterns from the previous classifiers. These outliers are the main problem in creating an ML model. So, we must overcome this problem by removing these outliers and replacing them with NULL values. These outliers will be for only the features which have only continuous variables, not for the categorical variables. To overcome these outliers the first step is to calculate the first and third quartile i.e., Q1 and Q3 (25% and 75%) of the data. Then evaluate the IQR (Interquartile range) i.e., (12)

$$IQR = Q3 - Q1 \qquad (12)$$

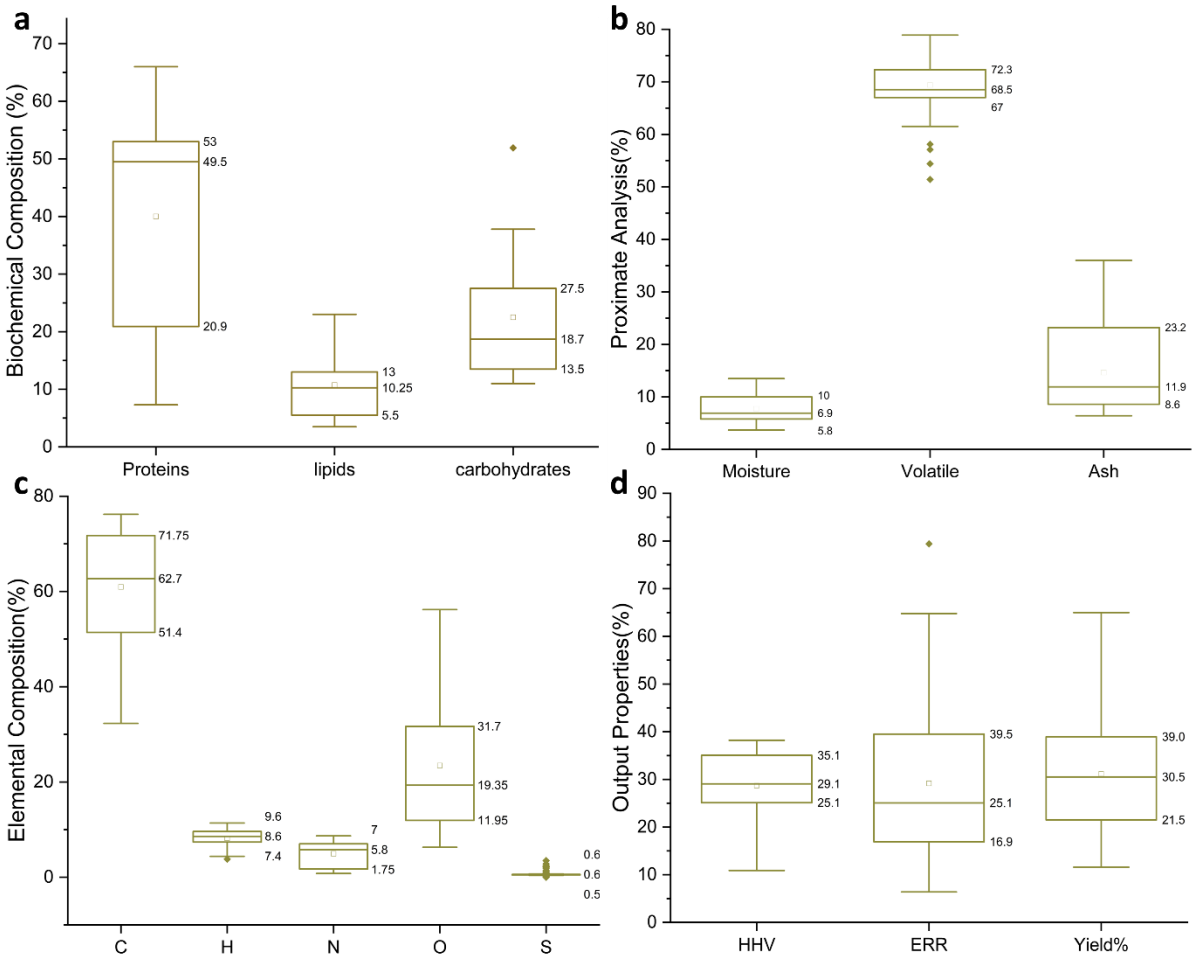
Then, we can find the outliers by the lower Bound (L.B) and upper bound (U.B). If the value exceeds this range, then it is shown that the data point is an outlier. The lower bound and upper bound equations are shown in (13) and (14):

$$L.B = Q1 - 1.5(IQR) \qquad (13)$$

$$U.B = Q3 + 1.5(IQR) \qquad (14)$$

By this, the outliers are found then we can replace the outliers with NULL values. In our original dataset, the outliers are clearly shown in Fig 2. But there are very few that will not affect that much during training the model.

281



282 **Fig 2.** Outliers of the features (Box plot), a – Biochemical Composition, b -Proximate Analysis, c -
283 Elemental Composition, d – Output Variables

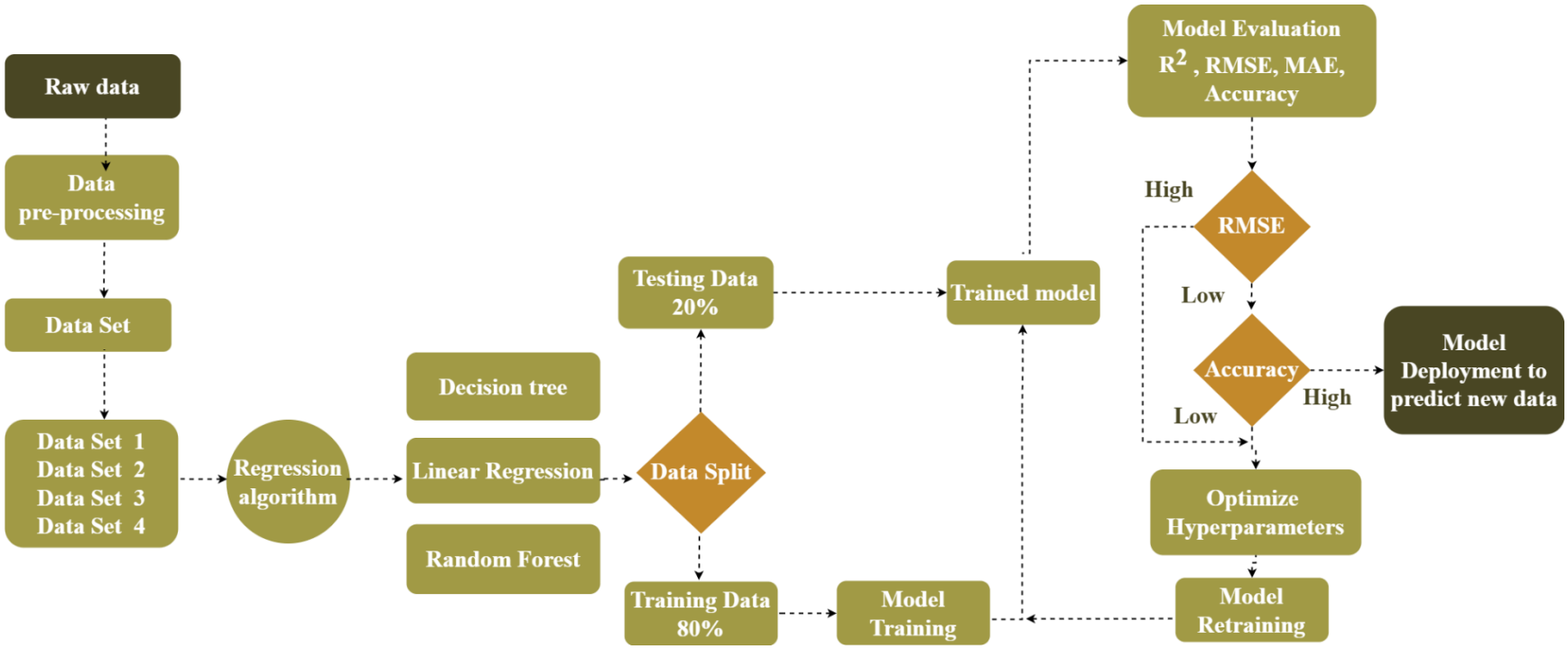


Fig 3. Methodology Process

Table 2. Detailed Input features and the ranges (min-max) of each feature in different datasets.

Items		Dataset 1	Dataset 2	Dataset 3	Dataset 4
Input features					
Elemental Composition	Carbon (%)	-	32.3-76.2	32.3-76.2	32.3-76.2
	Hydrogen (%)	-	3.80-11.4	3.80-11.4	3.80-11.4
	Nitrogen (%)	-	0.80-8.70	0.80-8.70	0.80-8.70
	Oxygen (%)	-	6.30-56.2	6.30-56.2	6.30-56.2
	Sulfur (%)	-	0.00-3.50	0.00-3.50	0.00-3.50
Biochemical composition	Protein (%)	-	-	7.30-66.0	7.30-66.0
	Lipid (%)	-	-	3.50-23.0	3.50-23.0
	Carbohydrate (%)	-	-	11.00-51.9	11.00-51.9
Operational conditions	Time (min)	30.0-60.0	30.0-60.0	30.0-60.0	30.0-60.0
	Temperature (°C)	150-420	150-420	150-420	150-420
	Pressure (bar)	50.0-221	50.0-221	50.0-221	50.0-221
Proximate Analysis	Moisture (%)	3.70-13.50	-	-	3.70-13.50
	Volatile (%)	51.40-81.40	-	-	51.40-81.40
	Ash (%)	6.40-36.00	-	-	6.40-36.00
Output Response					
Properties	Oil yield (%)	11.60-65.00	11.60-65.00	11.60-65.00	11.60-65.00
	HHV (MJ/kg)	10.90-38.20	10.90-38.20	10.90-38.20	10.90-38.20
	ERR (%)	6.40-79.40	6.40-79.40	6.40-79.40	6.40-79.40

HHV -High Heating Value; ERR – Energy Recovery Ratio.

3. Results and discussion

3.1 Data statistics before modelling:

Table 2 shows the detailed information of each data set that the maximum and minimum values of every feature in the dataset. Based on the data analysis related to the elemental composition C and O are the major elements for the prediction of bio yield. C ranges from 32.3-76.2 wt% and O ranges from 6.30-56.2 wt%. Another essential feature category was the biochemical composition of biomass. The temperature of HTL was 150-420 and the time was from 30-60 min, and other features like operational conditional, and proximate analysis are shown in Table 2. Now the output variables in the dataset are bio-yield, HHV, and ERR. Oil yield ranges from 11.6-65.0 wt%. HHV ranges from 10.9-38.2 wt% ERR ranges from 6.4-79.4 wt%. The datasets were in such a way that higher to lower order, all input features in dataset 4 and only six input features in dataset 1 because even without the elemental composition and biochemical composition the yield should be predicted for creating the best model for predicting the oil yield.

3.2 Input feature determination for bio-yield prediction

3.2.1 Hyperparameter tuning:

Equivalently for all datasets, the same hyperparameters are used. The maximum value for the hyperparameter of n_estimators in the random forest is 25 (Table 3), after this limit, there is no increase in the R2 and the max_depth maximum value is 10. In the same way in decision trees max_depth value is 10, the min_samples_split is 60, min_samples_leaf is 10 (Table 3). In linear regression, the hyperparameter learning rate is 0.01

which is very less because the error between the actual and predicted value should reach the local minima of the gradient descent to minimize the cost function. This hyperparameter tuning is done with the module of GridsearchCV in python that it searches like it forms a grid in the data and searches for the best parameters for the hyperparameter optimization. This was fully summarised in Table 3.

Table 3. Hyperparameter Tuning

Algorithm	*Hyperparameters*	*Range*	*Optimized value*
Random forest	Method		Bagging
	n_estimators	1 - 100	53
	max_depth	1 - 30	21
Decision Trees	max_depth	1 - 30	10
	min_samples_split	1 - 80	60
	min_samples_leaf	1 - 80	10
Linear regression	Learning Rate	0.01 - 1	0.01

3.2.2 Comparison of Machine Learning algorithms:

As we know, the datasets have 100 data in them. Based on the feature importance technique the dataset is divided into four datasets. These datasets are trained and tested by some of the machine learning algorithms. From the model evaluation, we know that the lesser the distance between the actual and predicted values, it is confirmed that it is the best model. For these datasets, the machine learning algorithms used are Random Forest, Decision Trees, and Linear Regression. The ensemble technique has bagging and boosting but for these kinds of datasets only the bagging method gives accurate prediction because these datasets have unstructured data so, the boosting technique does not perform well. Boosting techniques like Adaboost, and Xgboost. Finally, the best machine learning

algorithms suited for these datasets are Random Forest, Decision Trees, and Linear Regression. The train R2 and test R2 values and the model evaluation values are clearly shown in Table 4.

Table 4. Detailed values of Optimal model, R^2, RMSE, MAE for all datasets

Data set	No. of features	Target	Algorithm	Train			Test		
				R^2	RMSE	MAE	R^2	RMSE	MAE
1	6	Oil yield	RF	0.93	0.007	1.69	0.89	0.22	2.12
			DT	0.81	10.6	6.2	0.60	17.1	10.4
			LR	0.13	8.33	6.14	0.10	0.51	4.71
		HHV	RF	0.92	0.03	0.87	0.87	0.29	1.65
			DT	-0.33	146.0	12.1	-0.55	146.0	12.2
			LR	-0.18	1.66	4.17	-0.33	0.22	4.22
		ERR	RF	0.83	0.51	2.81	0.75	0.41	3.23
			DT	0.86	1.86	5.53	0.48	0.28	11.0
			LR	0.35	2.72	5.73	-0.52	2.01	7.53
2	10	Oil yield	RF	0.96	0.0003	1.86	0.82	3.30	2.57
			DT	0.82	4.54	6.73	0.61	0.001	9.52
			LR	0.40	3.75	5.47	-0.28	0.07	7.81
		HHV	RF	0.99	0.01	0.28	0.99	0.007	0.25
			DT	-0.12	136.1	12.07	-0.45	104.9	12.1
			LR	0.99	6.74	0.26	0.99	0.02	0.22
		ERR	RF	0.88	0.05	2.37	0.79	1.30	3.72
			DT	0.91	0.75	4.45	0.68	1.14	7.44
			LR	0.30	4.26	6.21	0.24	3.11	7.18
3	13	Oil yield	RF	0.91	0.0001	1.53	0.78	1.55	2.92
			DT	0.82	1.64	6.99	0.63	0.53	9.57
			LR	0.30	1.98	5.77	0.29	0.72	5.95
		HHV	RF	0.99	0.0001	0.21	0.98	2.16	0.45
			DT	-0.15	135.7	12.2	-0.21	90.5	12.7
			LR	0.99	1.38	0.09	0.97	0.06	0.38
		ERR	RF	0.88	0.03	2.17	0.74	0.19	3.46

			DT	0.88	0.04	5.10	0.79	0.14	6.25
			LR	0.40	7.74	6.43	0.29	3.66	7.36
4	17	Oil yield	RF	0.95	0.02	1.55	0.86	3.46	2.44
			DT	0.79	10.1	7.32	0.66	2.67	6.17
			LR	0.56	1.17	4.34	0.47	0.50	6.11
		HHV	RF	0.99	0.003	0.24	0.99	0.01	0.37
			DT	-0.17	131.1	11.7	-0.63	190.7	14.1
			LR	0.99	3.86	0.21	0.99	0.001	0.31
		ERR	RF	0.91	0.22	2.23	0.79	1.22	2.22
			DT	0.91	0.97	4.50	0.82	1.94	6.83
			LR	0.62	4.06	5.13	0.32	6.66	7.48

HHV -High Heating Value; ERR – Energy Recovery Ratio; RF – Random Forest; DT – Decision Tree; LR – Linear Regression; RMSE – Root Mean Square Error; MAE – Mean Absolute Error.

3.2.3 ML based feature importance analysis:

Feature importance is a method that shows the most important features from the dataset. The least important feature in the dataset is negligibly considered as the feature in the dataset. Based on this feature importance the original dataset is divided into three datasets among all these four it is so that fewer input features to high input features from dataset 1 to dataset 4. In dataset 4 every feature has been considered to perform the ML models. After training and testing, the R2 was ~0.80, RMSE~0.24, and MAE ~2.77. These values are given in Table 4. After the feature importance technique was done on dataset 4, then the most important features for the prediction of oil yield are operational conditions, proximate analysis, and elemental composition (Fig 4).

Elemental composition is the most important feature for the prediction of oil yield. But we must predict the oil yield without taking the elemental composition as the feature i.e., dataset 1. Even without taking elemental composition if the yield is predicted accurately same

concerning all features, then we can say that the best model of predicting the oil yield. Without any elemental composition or biochemical composition if the model gives an accurate value, then it is the best and most efficient method to produce the oil yield. Dataset 1 has R2 ~0.84, RMSE ~0.01, and MAE ~2.89 (Table 4). Dataset 2 and dataset 3 have a similar R2~0.80. But input features are less for dataset 1 even though it gives the R2 ~0.84 and the dataset which has all input features is dataset 4 its R2 ~0.80 (Table 4). Then we can confirm that the comparison between these two datasets, dataset 1 is the efficient process of predicting the oil yield. Therefore, it is considered that the proximate analysis and operational conditions play a vital role in predicting the oil yield.

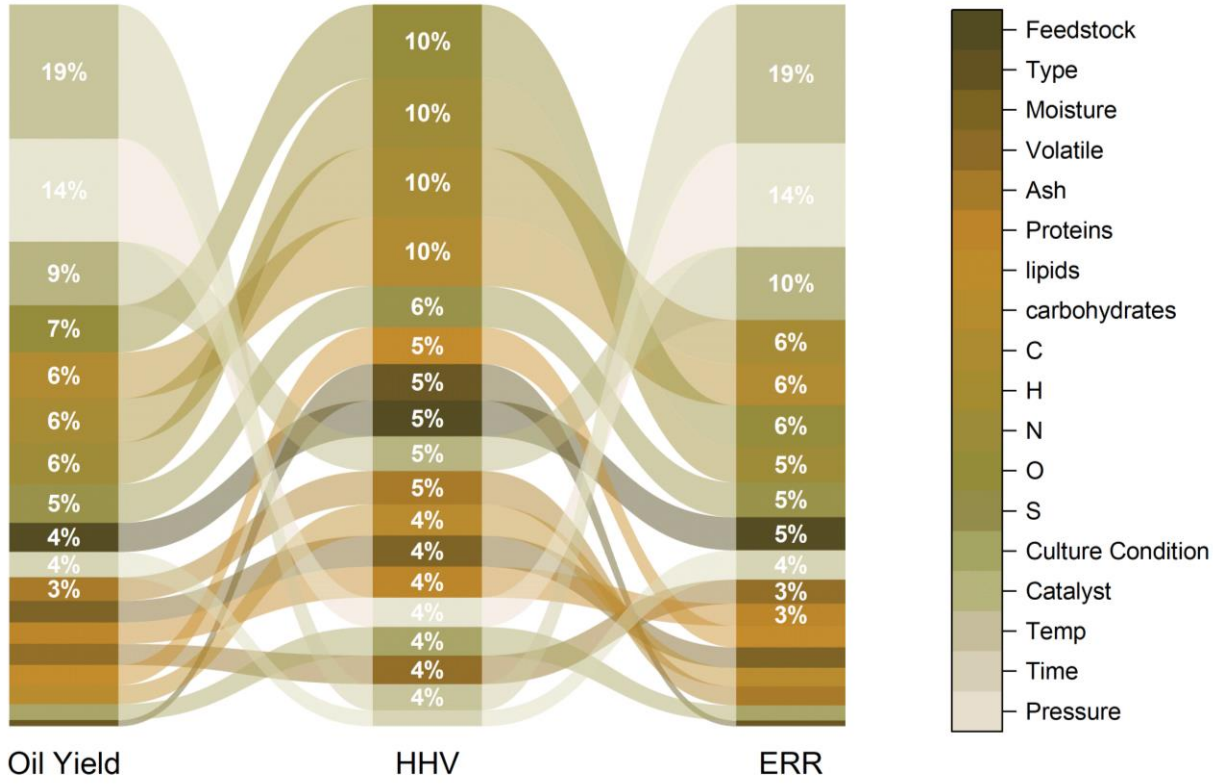


Fig 4. Feature importance graphs

3.3 Predicting analysis of ML model.

3.3.1 Actual vs predicted analysis:

According to the above section, there are three output variables for the prediction. In that oil, yield is considered as the important output among them. Oil yield is predicted well by developing ML models. The other two main targets HHV and ERR were determined to supply more information on bio-oil quality and the energy recovery ability of HTL. There are four features (operational conditions, elemental and biochemical composition, and proximate analysis). Were all considered for the multitask prediction model development. The values of model evaluation for these output variables are seen in Table 4. After hyperparameter tuning

for Decision Trees, Random Forest. The best models with their own tuned hyperparameters were achieved, as seen in Table 3.

Three ML models were used for the predictive analysis. In these three, RF is the most efficient ML model for the prediction of oil Yield. The important feature of the overall three targets. It can be found that operational conditions and proximate analysis are more important features. Biochemical composition is not an important feature for the prediction of oil yield. In dataset 4 all input features are considered for the prediction of oil yield. It has train R2~0.94 and test R2~0.80. These are not notable features then; the dataset may generate the noise to intercept the model fitting. It has the chance to get a less R2 value. After removing these unimportant features from the dataset, the testing performance was slightly increased from 0.80-0.84 that is in dataset 1 has only 6 input features i.e., proximate analysis and operational conditions the oil yield Train R2 ~0.90 and Test R2 ~0.84 in Random Forest algorithm. In each dataset for all the output variables, the actual vs predicted graph was given in the Random Forest algorithm because this algorithm is more efficient and gives an accurate prediction of oil yield by considering a straight line i.e., y=x line within the graph. This line is a reference line for both train and tests data points i.e., this graph shows a clear understanding of actual and predicted values (Fig 5).

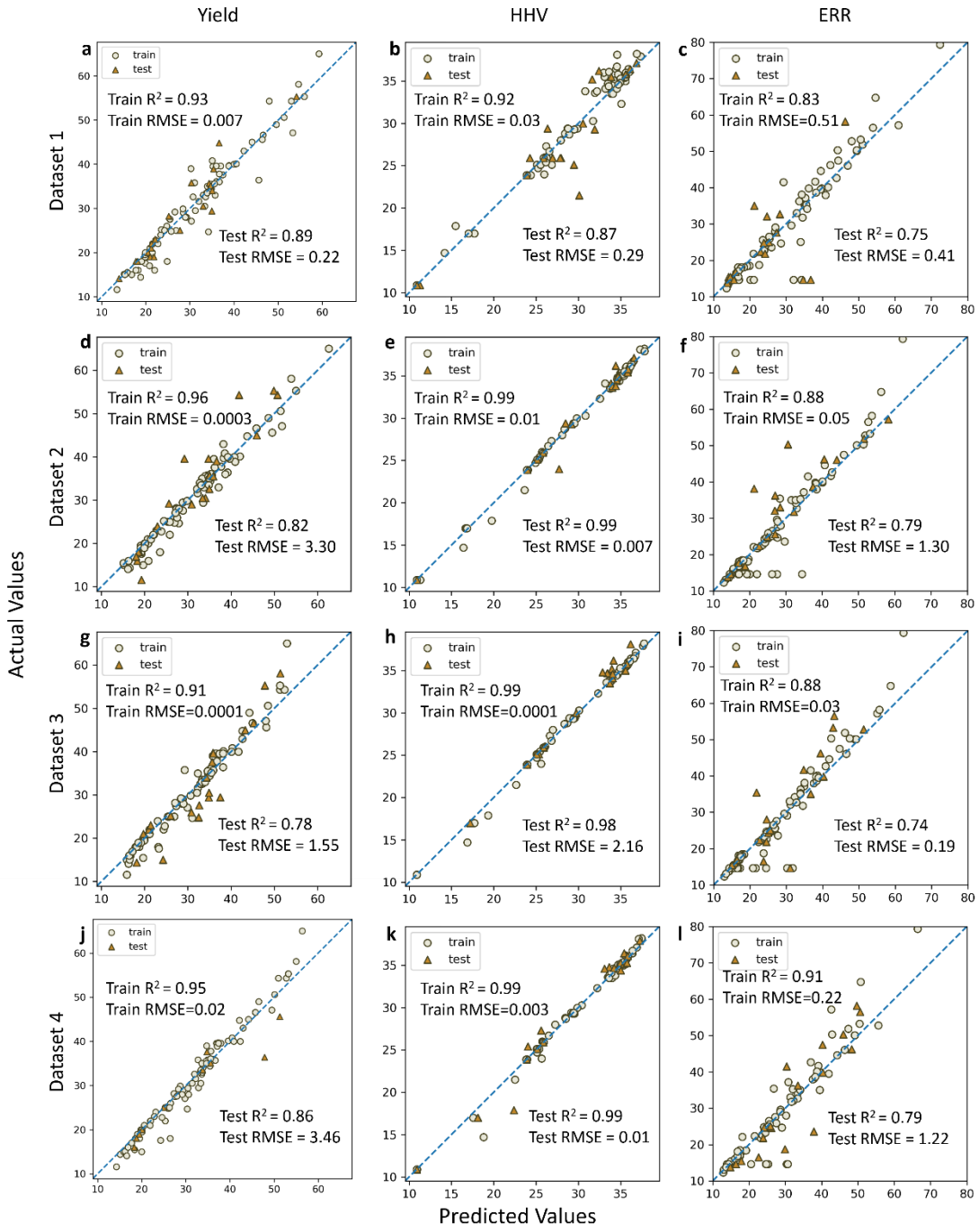


Fig 5. Yield prediction plots (train and test) of best RF developed from dataset #1 (a-c), Dataset #2 (d-f) and dataset #3 (g-i) and dataset #4 (j-l).

3.3.2 Best model selection procedure for prediction:

So far, we have seen all the R2, RMSE, and MAE values and the feature importance analysis and hyperparameter tuning. The algorithms used are Random Forest, Decision Trees, and Linear Regression for predicting the target values by taking all the input features. With this module selection procedure, we can conclude which is the best dataset and machine learning algorithm for the prediction of oil yield. For this, in the first step, we must compare all R2 values of oil yield for each dataset. The least R2 values are eliminated then in this dataset three R2 values are filtered out. The second step is comparing RMSE values with the lowest RMSE value showing it is the best model for prediction Fig 6. RMSE values are all the least so every dataset RMSE value is filtered out to the next step which is comparing MAE values, as same as the RMSE, MAE value is also should be as less as possible. Then in the dataset, 1 MAE value is 2.89 here dataset 2 has less MAE than dataset 1, even though dataset 1 is considered the best prediction model for oil yield. Because considering R2 dataset 1 has more R2 than dataset 2 at last we can conclude that dataset 1 and the random forest algorithm is the best suited for predicting the oil yield.

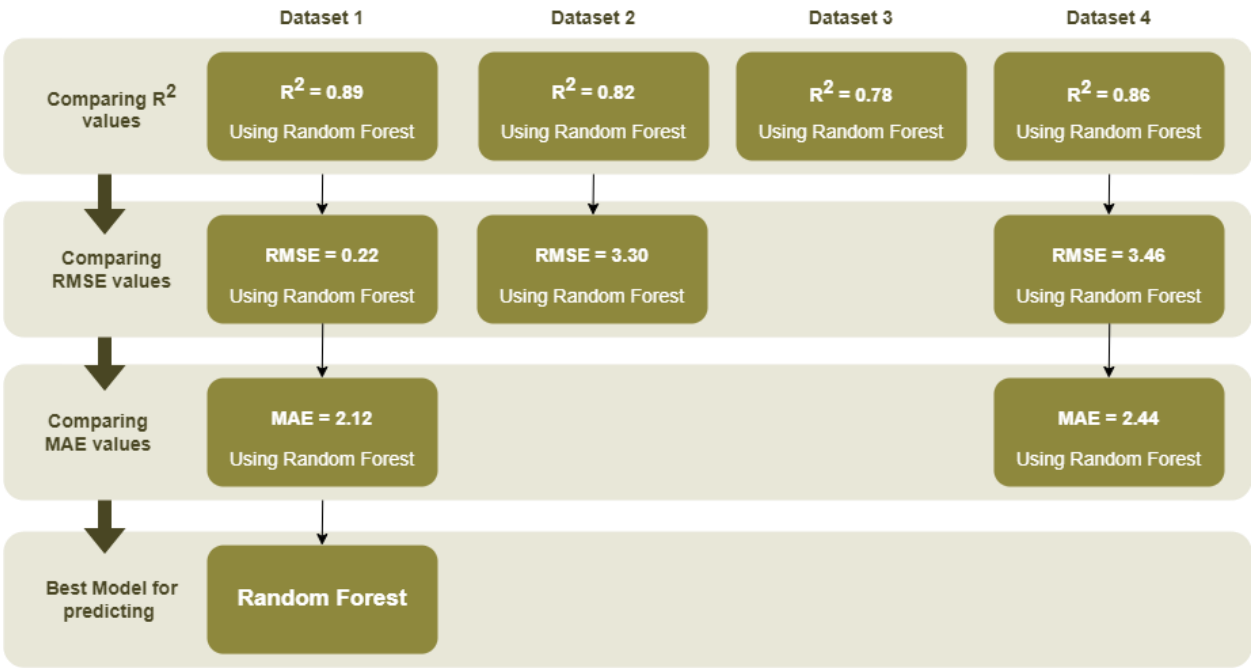


Fig 6. Algorithm comparison for prediction

3.3.3 Testing model against New Data:

Elemental composition, operational conditions, proximate analysis, and Biochemical composition are the features for predicting the oil yield. With all these features these are divided into four datasets by using feature importance. These are collected to develop Machine Learning models for the prediction and optimization of oil yield, HHV, and ERR. ML algorithms including Linear Regression, Decision Trees, and Random Forest were used to train the models for the prediction of oil yield, HHV, and ERR. Among all these algorithms Rf (Random Forest) is given as the best algorithm for the prediction. In all four datasets, Dataset 1 is given as the best model because out of all input features i.e., Dataset 4, six input features are enough for the prediction of the oil yield. Then, we can assume that with only 6 features the oil yield is produced as same as with all the input features. The average training R2 > 0.90 and test R2 ~

0.84, RMSE ~ 0.65 and MAE ~ 2.89. Table 4 is the comparison of Dataset 4 and Dataset 1 i.e., All input features Vs six input features. In Table 5, five new data are taken for each dataset and checked with true value and predicted value and calculated the difference.

Table 5. Predicting yield% with New Data

Dataset	True Value	Predicted Value	Difference (%)
1	28	32.12	-4.12
	27.1	34.85	-7.75
	22	22.82	-0.82
	34	34.22	-0.22
	38.5	33.51	4.99
4	28	32.59	-4.59
	27.1	29.92	-2.82
	22	21.03	0.97
	34	32.71	1.29
	38.5	33.11	5.39

3.3.4 Pearson Correlation:

The correlation measures the relationship between two variables. That it is positively or negatively correlated. This correlation coefficient is given in the range of 1 to -1. 1 indicates there is a perfect linear relationship between the variables, 0 is a non-linear relationship between the variables, and -1 indicates there is a negative linear relationship between the variables. The correlation of dataset 4 is given in Fig 7. This correlation heatmap gives some understanding of the one-on-one feature importance. Here, the main aim is used to predict the oil yield. The proximate analysis, elemental composition, operational conditions, and catalyst have a positive linear correlation. It shows that yield is easily predicted with proper adjustment of these positive linear correlated parameters. Biochemical composition, feedstock, and culture condition has a negative linear correlation, showing its reliability and significance.

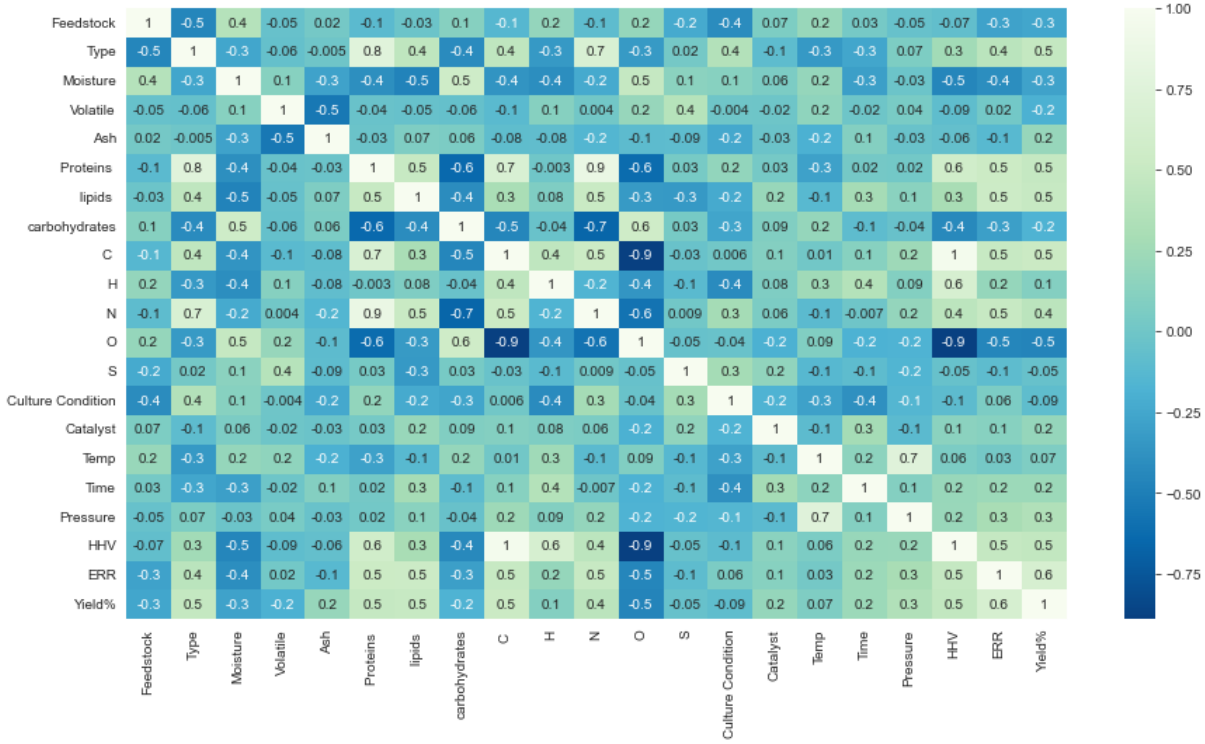


Fig 7. Data correlation heatmap

3.3.5 Statistical analysis of new Data:

The statistical analysis of the new data of dataset 4 and dataset 1 is compared in the graph using MATLAB (Fig 8). If the probability is > 0.05 then it is considered the best model then, the value should be less when compared the datasets. Dataset 1 < Dataset 4 it is given that dataset 1 is the best model for the prediction of oil yield.

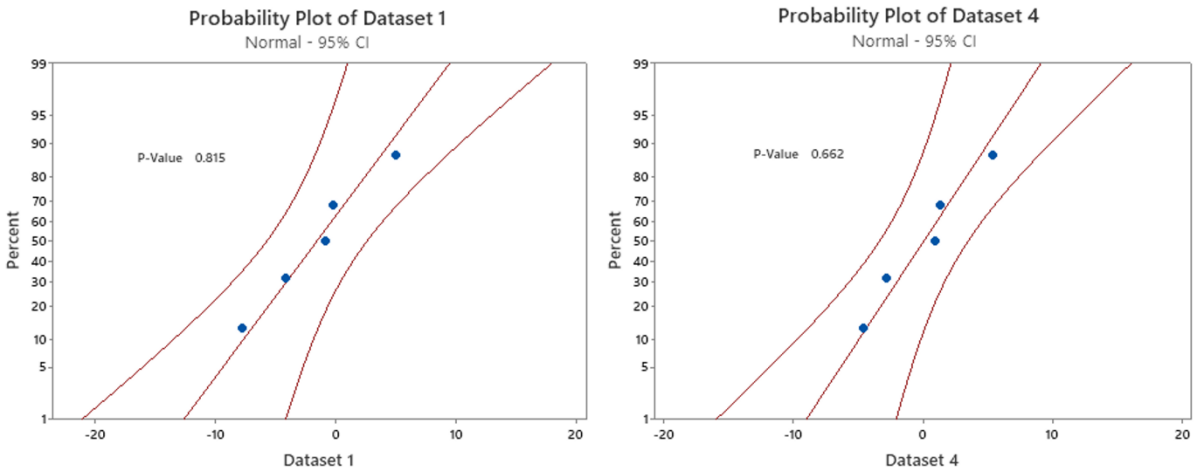


Fig 8. The probability plot of the different cases confirmed against the predicted result.

4. Conclusion:

Biocrude oil is the final product of this paper that is predicted with some of the input features. The data is collected using some of the articles and works of literature and by the feature importance. The whole dataset is divided into three sub-datasets. That is clearly shown in the methodology section. If experimenting physically costs more and takes more time. So, the solution to this is by using a machine learning approach. This reduces time and effort, and it gives approximately the same as the yield done in the experiment. There are mainly three algorithms used i.e., Linear Regression, Decision Trees, and Random Forest. This RF is the best model for the predicted oil yield. Then the hyperparameter tuning is the most important part of controlling the learning process of the model. Then it may have the chance of an increase in R2 of the model. The most important features are the proximate analysis and operational conditions for the prediction of the oil yield. Finally, the testing of the accuracy of yield was done physically and the yield with the Machine learning approach. We have taken five new data and by comparing the yield. ML model predicts the yield approximately the same as compared

to the actual data. Fig. 8 it is shown the statistical analysis of this new data from Dataset 1 and Dataset 4, This analysis shows that there is no significance between these two datasets. Dataset 1 is having more R2 compared to other datasets.

References

Ayesha Jasmin, S., Ramesh, P., Tanveer, M., 2022. An intelligent framework for prediction and forecasting of dissolved oxygen level and biofloc amount in a shrimp culture system using machine learning techniques. Expert Syst. Appl. 199. https://doi.org/10.1016/j.eswa.2022.117160

Biller, P., Ross, A.B., Skill, S.C., Lea-Langton, A., Balasundaram, B., Hall, C., Riley, R., Llewellyn, C.A., 2012. Nutrient recycling of aqueous phase for microalgae cultivation from the hydrothermal liquefaction process. Algal Res. 1, 70–76. https://doi.org/10.1016/j.algal.2012.02.002

Gollakota, A.R.K., Kishore, N., Gu, S., 2018. A review on hydrothermal liquefaction of biomass. Renew. Sustain. Energy Rev. 81, 1378–1392. https://doi.org/10.1016/j.rser.2017.05.178

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, 126278. https://doi.org/10.1016/j.biortech.2021.126278

Li, J., Zhang, W., Liu, T., Yang, L., Li, H., Peng, H., Jiang, S., Wang, X., Leng, L., 2021. 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. https://doi.org/10.1016/j.cej.2021.130649

López Barreiro, D., Zamalloa, C., Boon, N., Vyverman, W., Ronsse, F., Brilman, W., Prins, W., 2013. Influence of strain-specific parameters on hydrothermal liquefaction of microalgae. Bioresour. Technol. 146, 463–471. https://doi.org/10.1016/j.biortech.2013.07.123

Sánchez-Bayo, A., Megía Hervás, I., Rodríguez, R., Morales, V., Bautista, L.F., Vicente, G., 2021. Biocrude from nannochloropsis gaditana by hydrothermal liquefaction: An experimental design approach. Appl. Sci. 11. https://doi.org/10.3390/app11104337

Shakya, R., Adhikari, S., Mahadevan, R., Shanmugam, S.R., Nam, H., Hassan, E.B., Dempster, T.A., 2017. Influence of biochemical composition during hydrothermal liquefaction of algae on product yields and fuel properties. Bioresour. Technol. 243, 1112–1120. https://doi.org/10.1016/j.biortech.2017.07.046

Sharma, N., Jaiswal, K.K., Kumar, V., Vlaskin, M.S., Nanda, M., Rautela, I., Tomar, M.S., Ahmad, W., 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

Ullah, Z., khan, M., Raza Naqvi, S., Farooq, W., Yang, H., Wang, S., Vo, D.V.N., 2021. A comparative study of machine learning methods for bio-oil yield prediction – A genetic algorithm-based features selection. Bioresour. Technol. 335, 125292. https://doi.org/10.1016/j.biortech.2021.125292

Yang, K., Wu, K., Zhang, H., 2022. Machine learning prediction of the yield and oxygen content of bio-oil via biomass characteristics and pyrolysis conditions. Energy 254, 124320.

https://doi.org/10.1016/j.energy.2022.124320

Yang, Y.F., Feng, C.P., Inamori, Y., Maekawa, T., 2004. Analysis of energy conversion characteristics in liquefaction of algae. Resour. Conserv. Recycl. 43, 21–33. https://doi.org/10.1016/j.resconrec.2004.03.003

Zhang, T., Cao, D., Feng, X., Zhu, J., Lu, X., Mu, L., Qian, H., 2022. Machine learning prediction of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions. Fuel 312, 122812. https://doi.org/10.1016/j.fuel.2021.122812

Zhang, W., Li, J., Liu, T., Leng, S., Yang, L., Peng, H., Jiang, S., Zhou, W., Leng, L., Li, H., 2021. Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. Bioresour. Technol. 342, 126011. https://doi.org/10.1016/j.biortech.2021.126011