Feature-Driven Optimization of Syngas Production in Biomass Gasifiers using Machine Learning

Phrugsa Limbunlom
School of Computer Science and Electronic Engineering
University of Essex

Abstract—Biomass gasification is a significant thermal conversion method that uses fluidized bed reactors to generate syngas compositions with low heating values. Machine learning models were adopted to predict biomass composition and operating conditions. However, there has not yet been a comprehensive model developed through selective feature optimization. In this research, four regression machine-learning models were employed. The predictive capacity of syngas compositions and lower heating values (LHV) were assessed. The output products were derived from various lignocellulosic biomass feedstocks across a wide array of operating conditions. The four regression machine-learning algorithms are Decision Tree, Support Vector Machine (SVM), XGBoost, and Random Forest (RF), which were adopted to evaluate prediction performance after undergoing hyperparameter and feature selection optimization. Pearson correlation was applied to validate the correlation between input and output variables. XGBoost and RF established good performance results (XGBoost: R2 = 0.567–0.892, RMSE = 0.880–9.645; RF: R2 = 0.675–0.855, RMSE = 1.336–10.558). XGBoost provided low RMSE scores in CH4, LHV, and Tar yield (1.495, 0.880, and 9.645) and a high R2 score in LHV (0.892), whereas RF produced low RMSE scores in LHV and Tar yield (1.215 and 9.614). The XGBoost algorithm selected seven features after optimization, including cellulose, hemicellulose, lignin, temperature, pressure, equivalence ratio (ER), and steam-to-biomass ratio (SBR). In contrast, the RF algorithm selected all features, including cellulose, hemicellulose, lignin, temperature, pressure, equivalence ratio (ER), steam-to-biomass ratio (SBR), and superficial gas velocity.

1 Introduction

Due to the rapid growth in the economy and population and energy-saving awareness, the search for renewable sources is crucial to replace traditional fossil fuels and align with the global trend towards sustainable and eco-friendly energy solutions. Biomass is one of the crucial raw materials and alternative sources of fossil fuels. According to J.Y. Kim [1], lignocellulosic biomass—cellulose, hemicellulose, and lignin—serves as a renewable carbonaceous resource that can produce heat and electricity from combustion and yield liquid or gaseous biofuels through thermochemical conversion processes. Among the conversion approaches, biomass gasification is the most efficient process, which can convert biomass feedstocks to syngas. The primary components of syngas used in fuel production are H2, CO, CO2, and CH4. The types of gasification reactors, including fixed beds, fluidized beds, and entrained flow gasifiers, are also crucial technologies. Among these, the fluidized bed gasifier is the most mature reactor that has a high capability of efficient production at scale, thorough solid mixing, and rapid heat transfer. Three primary factors are required to manipulate a fluidized bed biomass gasifier: syngas composition, lower heating value, and char and tar yield. To address the measure of biomass gasifier components, the proportion of the lignocellulose composition and operation conditions are computed. Therefore, measuring those key factors is greatly important to designing a high level of efficiency for a fluidized bed gasifier.

A study conducted by Jun Young Kim et al. [2] harnesses machine learning approaches to predict the key performance metrics of syngas composition, lower heating value, and char and tar yields. Three different models after the optimization of hyperparameters were adopted: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). The input variables are lignocellulose composition (cellulose, hemicellulose, and lignin), temperature (T), pressure (P), steam-to-biomass ratio (SBR), and superficial gas velocity (Ug). Moreover, the study adopted Monte Carlo filtering (MCF) and integrated three machine learning models to predict the important features of the operating conditions and biomass composition.

Another study conducted by Jun Young Kim et al. [3] adopted automated machine learning (AutoML) to select the best machine learning algorithms and combined cooperative game theory (shapely additive explanation, SHAP) to develop an interpretable model. These approaches can identify the importance of each input feature.

From the previous background, the objective of this study is to investigate how varying numbers of input features—lignocellulosic biomass feedstocks and operating conditions—influence the accuracy of the prediction of targeted data—syngas compositions, lower heating values (LHV), and char and tar yields.

This research aims to determine whether the most effective features selected from the algorithm can improve the prediction performance of the previous models.

Variables		Ranges
Input variables		
Lignocellulose composition [wt.%]	Cellulose (Cell.)	0.20–0.58
	Hemicellulose (Hem.)	0.08–0.63
	Lignin	0.10–0.49
Temperature (T)		600–900
Pressure (P) [abar]		1–10
Equivalence ratio (ER)		0–0.86
Steam to biomass ratio (SBR)		0–8.03
Superficial gas velocity (Ug) [m/s]		0.02–9.59
Output variables		
Syngas composition [vol]	H2	5.39–66.03
	CO	5.01–55.44
	CO2	6.78–62.56
	CH4	1.31–20.1
Lower heating value (LHV) [MJ/Nm3]		1.74–15.0
Char yield [wt.%]		0–45
Tar yield [g/Nm3]		0-134.1

TABLE 1
Input and output variables of the constructed database

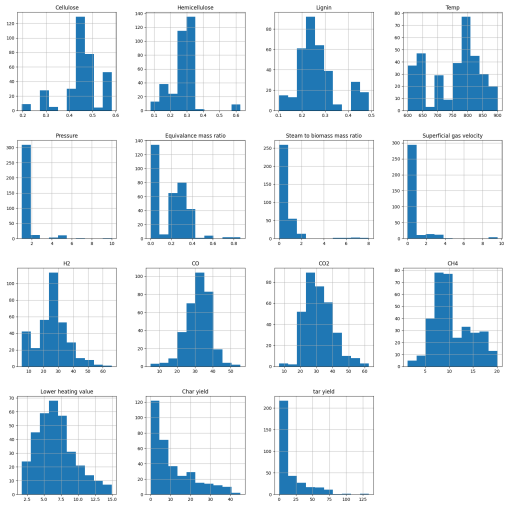

Fig. 1. The graph representation of input and output variables distribution

2 Background

Jun Young Kim et al. studied three machine learning models, RF, SVM, and ANN, to predict the syngas products, LHV and tar/char yield with lignocellulosic information and operating conditions from 336 literature data points. The result shows that the prediction accuracy of RF and ANN provided high R2 and low RMSE. Moreover, the integration of Monte Carlo Filtering with machine learning algorithms results indicates that steam to biomass ratio, cellulose and lignin are three important features of RF integrated with MCF, whereas ANN integrated with MCF are cellulose and hemicellulose [1].

Another study by Jun Young Kim et al. indicated that CatBoost and WeightedEnsemble_L2 provided the high R2 and low RMSE for the average of all input features. Moreover, the feature importance results from the SHAP value showed that cellulose and lignin are two input features, which equally affect the performance of both CatBoost and WeightedEnsemble_L2 models [2].

A study by Luca Parisi et al. adopted a feature-driven approach to improve the classification of Parkinson's patients. The result indicated that the feature-driven algorithm identified by a Multi-Layer Perceptron (MLP) to reduce an initial 27 features to 20 selected features before input to a Lagrangian Support Vector Machine (LSVM) can improve the overall classification accuracy [3].

3 Methodology

3.1 Dataset collection

The dataset was collected from previous reliable studies on the results of the fluidized beds experiment [1]. There are 336 samples containing lignocellulosic compositions, operating conditions (temperature, pressure and superficial gas velocity), equivalence ratio, and steam to biomass ratio as input variables, and syngas compositions (H2, CO, CO2 and CH4), lower heating value (LHV), char and tar yield as the output variables [1] shown in Table 1.

3.2 Pre-processing

Input and output variables were derived from the initial table. Following data validation, no null values were found in the dataset. However, it was necessary to convert the data from the object to the float type. Subsequently, a graphical representation was generated by plotting the data to assess its distribution, as shown in Figure 1.

According to the graph, CO, CO2, and CH4 have normal distributions, whereas the distributions of other variables are skewed.

To address the distribution imbalance, the min-max normalization method was applied to both input and output variables before their incorporation into the training process of the model.

3.3 Dataset split

The dataset was split into training, validation, and testing sets. There were 20% of the datasets for testing, and 5-fold cross-validation was applied.

3.4 Decision tree

A decision tree is a type of non-parametric supervised learning technique employed for both classification and regression tasks. The objective is to construct a predictive model for the target variable by extracting straightforward decision rules from the features in the provided data. The hierarchical structure of the tree consists of a root node, branches, decision nodes, and leaf nodes. [4].

3.4.1 Hyperparameters tuning

- `max_depth`: This parameter was tuned to determine the maximum depth of the decision tree, controlling the extent to which the tree is split to prevent overfitting [4].
- `min_samples_split`: Tuning this parameter was essential to defining the minimum number of samples required to split the nodes in the decision tree [4].
- `min_samples_leaf`: This parameter underwent tuning to identify the minimum number of samples required to form a leaf node in the decision tree [4].

3.5 Support Vector Machine

Support Vector Machine is a supervised machine learning technique that finds the best separable hyperplane to separate two classes. The algorithm will find the maximum distances between the two classes to draw the margin between hyperplanes [5].

3.5.1 Hyperparameters tuning

- `C`: This is a regularization parameter. It was tuned to figure out the trade-off between having a smooth decision boundary and classifying the training points correctly [5].
- `kernel`: It is a function that takes data as input and transforms it into a high-dimensional space. It was tuned to find a non-linear relationship in the data [5].
- `gamma`: This hyperparameter is specific to the Radial Basis Function (RBF) kernel. It was tuned to define how far the influence of a single training example reaches, which could determine the smoothness of the decision boundary [5].

3.6 XGBoost

XGBoost operates as an ensemble learning technique, acknowledging that the efficacy of a single machine-learning model may be insufficient. Ensemble learning provides a systematic approach to harnessing the predictive capabilities of multiple learners. The outcome is a unified model that consolidates the outputs of several models into a collective prediction [6].

3.6.1 Hyperparameters tuning

- `n_estimators`: This parameter was tuned to find the number of trees in the model [7].
- `max_depth`: This parameter was tuned to find the depth of each tree [7].
- `learning_rate`: This parameter was tuned to prevent overfitting since the learning rate is used in the gradient boosting model [8].

3.7 Random Forest

Random Forest is a machine-learning algorithm that amalgamates the outcomes of numerous decision trees to produce a unified result [9].

3.7.1 Hyperparameters tuning

- `n_estimators`: This parameter was tuned to denote the number of trees the algorithm constructs before combining their predictions [9].
- `max_depth`: This parameter was tuned for the same purpose in the decision tree.
- `min_samples_leaf`: This parameter was tuned for the same purpose in the decision tree.
- `min_samples_split`: This parameter was tuned for the same purpose in the decision tree.

3.8 Feature Selection

In addition to tuning the hyperparameters of the machine learning algorithms mentioned above, an individual model underwent feature tuning based on the input variables through the implementation of a recursive feature elimination (RFE) approach. The selective features applied to the models ranged from 1 to 8 features.

3.9 Evaluation

The models were evaluated using regression coefficient (R2) and root mean square error (RMSE) scores, which are the same evaluation methods mentioned in the previous experiment [1]. R2 was used to validate the correlation between two variables and RMSE was used to verify the error between the predicted data and actual target data.

4 Results

4.1 Pearson correlation

The Pearson correlation is a correlation coefficient frequently utilized in linear regression. The correlation values range from -1 to 1, where -1 denotes a strong negative relationship, 0 represents no relationship, and 1 indicates a strong negative relationship.
The correlation finding aligned with the correlation observed in the previous study [1], as shown in Figure 2. According to the map, SBR exhibits the most pronounced positive relationship with H2, while ER demonstrates the most significant negative relationship. SBR is negatively correlated with CO, while ER shows a positive correlation with CO2. Temperature and ER stand out as the two notable factors, with the strongest reverse correlation with CH4. Similarly, ER exhibits a negative correlation with LHV. Lignin displays a positive correlation with char yield. Conversely, both temperature and ER have a negative correlation with tar yield.

4.2 Training performance of different machine learning approaches

R2 and RMSE scores were used to evaluate the performance of training models. The scores were obtained from the average score of a 5-fold cross-validation. The result of the training performance of each machine-learning model, along with its feature selection, is shown in Table 2.

4.3 The prediction performance of different machine learning approaches

Following the hyperparameter optimization for each training model, predictions for multi-output variables were made on the test dataset. The predicted data was then juxtaposed with the actual data obtained from the labels in the test dataset. The comparison for each output variable from each machine learning algorithm is shown in Table 3, which is compared to the evaluation matrix from the previous study [1].

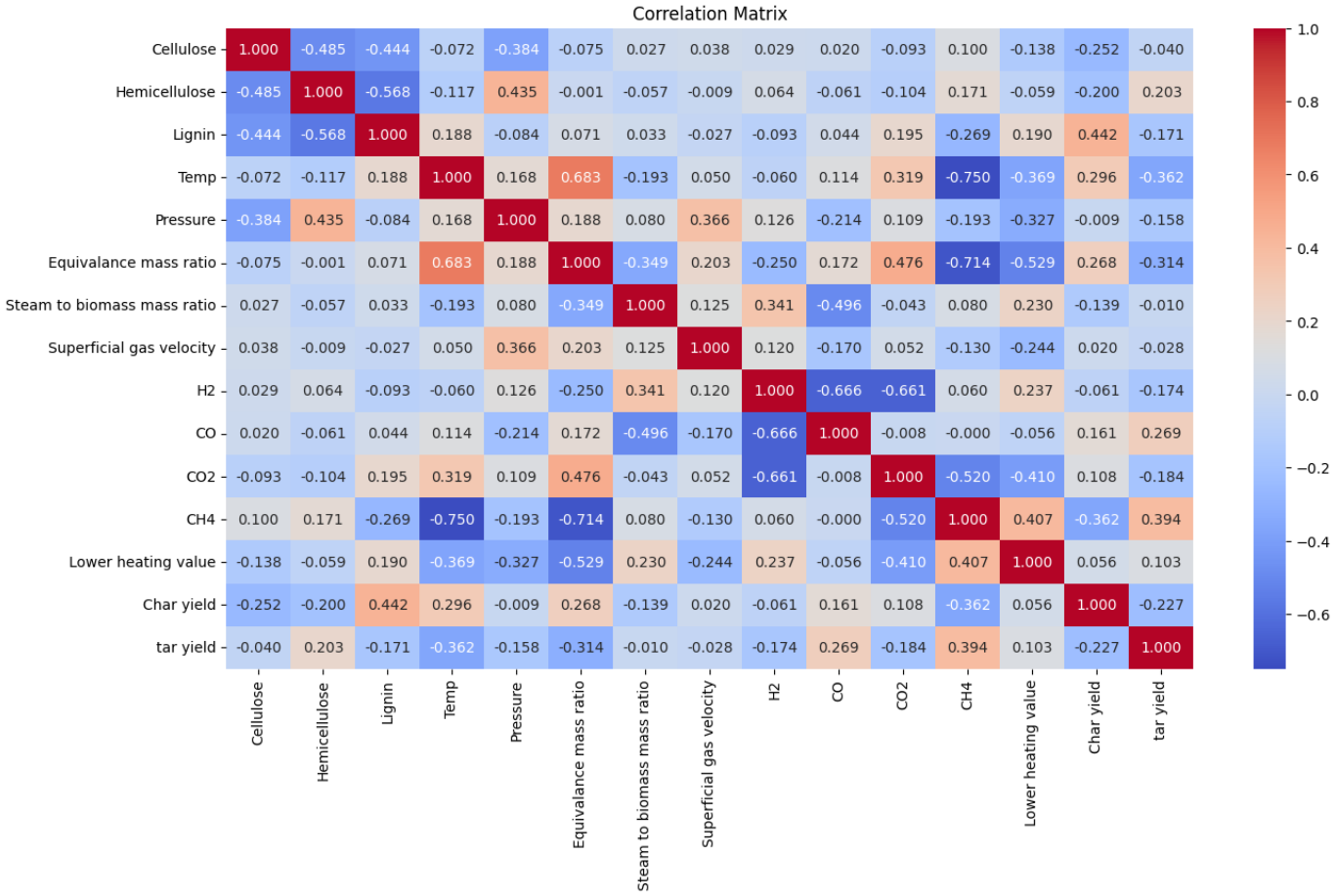

Fig. 2. The figure shows the correlation between input and output variables using the heatmap technique.

Model	Statistical Indicator		Features	Features Names
Desicion Tree	R2	0.705	2	Hemicellulose
	RMSE	7.524		Superficial gas velocity
SVM	R2	0.752	8	Cellulose
	RMSE	7.925		Hemicellulose
				Lignin
				Temp
				Pressure
				ER
				SBR
				Superficial gas velocity
XGBoost	R2	0.794	7	Cellulose
	RMSE	6.679		Hemicellulose
				Lignin
				Temp
				Pressure
				ER
				SBR
Random Forest	R2	0.780	8	Cellulose
	RMSE	7.236		Hemicellulose
				Lignin
				Temp
				Pressure
				ER
				SBR
				Superficial gas velocity

TABLE 2
Performance of training models and selective features after optimization

Model	Statistical Indiciator	H2	CO	CO2	CH4	LHV	Char	Tar
RF	R2	0.946	0.896	0.923	0.906	0.809	0.884	0.857
	RMSE	3.85	3.54	4.38	1.50	1.39	4.70	11.54
SVM	R2	0.814	0.635	0.671	0.840	0.575	0.697	0.723
	RMSE	4.94	4.77	5.65	1.74	1.97	6.25	13.07
ANN	R2	0.924	0.775	0.921	0.876	0.565	0.837	0.771
	RMSE	3.55	3.88	3.69	1.61	1.46	5.41	10.56
Decision Tree	R2	0.838	0.686	0.646	0.840	0.685	0.443	0.242
	RMSE	4.389	4.187	5.387	1.615	1.502	7.927	17.554
SVM	R2	0.699	0.449	0.626	0.790	0.662	0.437	0.513
	RMSE	5.991	5.543	5.540	1.853	1.556	7.965	14.068
XGBoost	R2	0.877	0.670	0.798	0.863	**0.892**	0.567	0.771
	RMSE	3.829	4.292	4.068	**1.495**	**0.880**	6.988	**9.645**
Random Forest	R2	0.763	0.675	0.692	0.855	0.751	0.690	0.726
	RMSE	5.312	4.256	5.025	1.541	**1.336**	5.913	**10.558**

TABLE 3
Performance of predicted multiple outputs from each machine learning algorithm

5 Discussion

According to the findings in Table 2, the training performance of the four machine-learning algorithms was observed. XGBoost exhibited the most favourable results, with the lowest RMSE (6.679) and highest R2 (0.794) scores compared to the other three models. Following feature optimization, seven features—cellulose, hemicellulose, lignin, temperature, pressure, equivalence mass ratio, and steam-to-biomass ratio—were selected.

Random Forest demonstrated the second-lowest RMSE (7.236) and the second-highest R2 (0.780) scores. All features underwent optimization for both the Decision Tree and Random Forest, while SVM specifically chose two features, as outlined in Table 2.

Regarding the prediction performance, XGBoost yields satisfactory results with lower RMSE on CH4, LHV, and Tar yield (1.495, 0.880, and 9.645, respectively). These scores were compared to those of the three models from the previous study, as outlined in Table 3. Notably, XGBoost achieved a higher R2 score for LHV (0.892) compared to the models in the previous study.

In contrast to the three machine-learning models from the previous research [1], Random Forest demonstrated lower RMSE scores for LHV and Tar yield (1.336 and 10.558, respectively).

6 Conclusion

Decision tree, SVM, XGBoost, and Random Forest are four machine-learning algorithms employed for the prediction of syngas compositions, lower heating values (LHV), and char/tar yield. This prediction utilizes data derived from 336 points in the literature review, incorporating lignocellulose composition and operating conditions (temperatures, pressure, equivalent mass ratio, and superficial gas velocity). Pearson correlation was used to find the correlation of each input feature and output variable. Moreover, recursive feature elimination was applied to refine the selection of features during model training. R2 and RMSE were used to evaluate the model performance to identify the optimal model. Additionally, the models were employed to predict multiple output data points from the testing set, which constituted 20% of the original dataset. Subsequently, each output prediction was used to compute the R2 and RMSE scores with the corresponding targets from the test set.

In terms of training performance, XGBoost outperformed other models with the highest R2 score and the lowest RMSE, followed by Random Forest. Regarding the prediction performance, XGBoost exhibited lower RMSE scores in prediction CH4, LHV, and Tar yield and a higher R2 score in LHV, whereas Random Forest produced lower RMSE scores in LHV and Tar yield.

References

[1] Jun Young Kim, Dongjae Kim, Zezhong John Li e, Claudio Dariva f, Yankai Cao g, Naoko Ellis. Predicting and optimizing syngas production from fluidized bed biomass gasifiers: A machine learning approach. https://doi.org/10.1016/j.energy.2022.125900

[2] Jun Young Kim, Ui Hyeon Shin, Kwangsu Kim. Predicting biomass composition and operating conditions in fluidized bed biomass gasifiers: an automated machine learning approach combined with cooperative game theory. https://doi.org/10.1016/j.energy.2023.128138

[3] Luca Parisi, Narrendar RaviChandran, Marianne Lyne Manaog. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. https://doi.org/10.1016/j.eswa.2018.06.003

[4] Anshul Saini. Decision Tree Algorithm – A Complete Guide. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[5] Anshul Saini. Guide on Support Vector Machine (SVM) Algorithm. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

[6] Introduction to XGBoost Algorithm in Machine Learning. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[7] Jason Brownlee PhD. How to Tune the Number and Size of Decision Trees with XGBoost in Python. https://machinelearningmastery.com/tune-number-size-decision-trees-xgboost-python/

[8] Jason Brownlee PhD. Tune Learning Rate for Gradient Boosting with XGBoost in Python. https://machinelearningmastery.com/tune-learning-rate-for-gradient-boosting-with-xgboost-in-python/

[9] Sruthi E R. Understand Random Forest Algorithms With Examples (Updated 2024).https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/