Data mining of the Biomass Gasification using combination of Random Forest and the StaGen module

Kun Xie

January 4, 2024

Abstract

To uncover patterns within the Biomass Gasification dataset, this case study established multiple nonlinear machine learning models to attempt fitting the dataset and conducted comparisons. After considering various evaluation metrics, the most suitable foundational model for this task was identified as the Random Forest. Subsequently, a StaGen (Stabilizer and Generalizer) neural network module was trained on this basis to transform the predictive values of the foundational Random Forest, serving as the final prediction.This task primarily consisted of two phases. First, establishing and training multiple machine learning models, and identifying the best-performing one as the foundational model. 2) Collecting and establishing a StaGen dataset for training the StaGen module. The trained module was then integrated with the foundational model for testing.After comprehensive consideration and comparative analysis with the test set, the Random Forest model was chosen as the best machine learning algorithm for the foundational model. The integration with the StaGen module resulted in a significant improvement, enhancing stability and generalization capabilities, and reducing sensitivity to dataset partitioning.

1 Introduction

Considering the continuous increase in energy demands as the result of societal and technological advancement, the employment of renewable energy resources as the alternative to fossil fuels has become a promising initiative in response to extensive environmental concerns[1]. Among various conversion implementations, biomass gasification performs excellent efficiency, in which feedstocks are converted into a mixture of gases primarily composed of H2, CO, CO2 and CH4. Particularly, lignocellulosic biomass has been recognized as an ideal raw material due to its extensive presence and carbonaceous nature. Moreover, in terms of gasification reactor options, compared to fixed bed and moving bed, the fluidized bed gasifier shows a higher capability of mass production, better solids mixing and faster heat transfer, therefore has been identified as the most developed option. Several key factors are considered to impact the production and performance of the conversion process, particularly the lignocellulose composition, temperature, pressure, equivalence ratio, steam-to-biomass ratio, and superficial gas velocity[3]. On the other hand, several critical metrics are emphasized to measure the conversion performance, they are syngas composition, lower heating value, char yield and tar yield. In compliance with such experience, data is collected based on industrial production, containing 8 features and 7 labels corresponding to the previous factors [1].

With the advancement of data science and machine learning technologies, industrial data can be effectively mined and interpreted for underlying patterns using machine learning models. Particularly in dealing with large volumes of complex industrial data, machine learning models demonstrate their unique capability to efficiently identify valuable patterns and correlations within these datasets. This report will showcase the training results of several machine learning models, as well as the remarkable enhancement of the innovative StaGen module.

2 Literature Review

Existing work has already conducted numerous valuable explorations on the Biomass dataset. Kim [1] tested several machine learning models on this dataset, such as Random Forest, SVM, and neural networks, and introduced the Monte Carlo method for the analysis of experimental results. Additionally, in another article by Kim [2], the SHAP method was introduced to assess and interpret the model's predictive results. Kim's experimental outcomes have provided critical guidance for this case study and defined a baseline.

3 Methodology

3.1 Data analysis

The dataset contains 8 input features and 7 output targets. To model their potential relationships, the task is characterized as multi-objective supervised machine learning. The dataset comprises 336 samples, with no missing values in any input and output columns, and all are presented as continuous data types.

To check for outliers in the dataset, histograms and box plots were drawn. It was observed that there are a few outlier data points in the 'Steam to biomass mass ratio' and 'Superficial gas velocity' columns. After consideration, in the actual training process, sample rows with superficial gas velocity greater than 8 were removed.

Correlation Analysis: The Pearson correlation coefficient is a statistical measure that quantifies the linear correlation between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 implies no linear correlation between the variables. Through Pearson correlation analysis, significant correlations were identified among multiple variables in our dataset. The fact that most variable pairs have correlation coefficients far from zero provided insights into the selection of the foundational model and data processing strategies.

In predictive models, highly correlated features can lead to multicollinearity issues, thereby affecting the stability and interpretability of the model. To mitigate the nonlinear relationships between features, it might be necessary to use techniques like Principal Component Analysis (PCA) or feature engineering.

During experimentation, an attempt was made to use PCA to map the original feature space to a new low-dimensional independent feature space. However, the training results showed that the performance of several models actually decreased after the PCA transformation, possibly due to information loss during the PCA process. Since PCA did not aid in this task, I decided to abandon attempts to use PCA and feature engineering. I also excluded various machine learning models that lack nonlinear learning capabilities, such as linear regression and SVM.

3.2 Potential candidates of the foundational model

Based on the data analysis conducted, the following machine learning models were selected as potential candidates for the foundational model. After testing, the Random Forest was chosen as the best foundational model for establishing the StaGen dataset and integrating the StaGen module as the final model for this case study.

Random Forest In this study, the Random Forest model was selected as one of the candidates for the foundational model, owing to its exceptional performance in handling complex datasets. Random Forest is an ensemble learning method based on decision trees. It builds multiple decision trees and combines their predictions to enhance the model's accuracy and robustness. This method is particularly suitable for our research in the following aspects:

Handling Complex Relationships Among Features: As our dataset involves complex interactions among multiple features, Random Forest can effectively capture these nonlinear relationships. Each tree learns different parts of the data, and this diversity allows the model as a whole to understand the data more comprehensively.

Reducing Overfitting Risks: Compared to a single decision tree, Random Forest reduces the risk of overfitting by integrating predictions from multiple trees. This is because a single tree may be overly sensitive to specific training samples, whereas an ensemble method can balance this effect and provide more generalized predictions.

Robustness and Flexibility: Random Forest demonstrates strong robustness to missing values and imbalanced datasets, making it a reliable choice under various data conditions. Considering these advantages, Random Forest emerged as an ideal analytical tool in this study. It not only enhanced the model's predictive capabilities but also provided a deeper understanding of the data, which is crucial for addressing our research questions.

XGBoost XGBoost (eXtreme Gradient Boosting) is an advanced implementation of the gradient boosting algorithm, particularly suited for complex regression and classification problems. The reasons for choosing XGBoost are primarily based on the following aspects: Regularization Features: XGBoost incorporates L1 and L2 regularization on top of traditional gradient boosting, which helps in reducing model overfitting and enhances its generalization ability. This is especially important to avoid overlearning on complex datasets. Tree Pruning and Parallel Processing: Unlike traditional gradient boosting methods, XGBoost employs a more efficient tree pruning strategy and can perform parallel computations across multiple cores on a single machine, significantly improving training efficiency. Given these advantages, XGBoost was selected as a key analytical tool in this study. Its robust modeling capabilities and flexibility make it an ideal choice for addressing the complex data analysis challenges we face.

Neural Network Fully Connected Neural Network, foundational in the field of deep learning, are renowned for their robust feature learning and pattern recognition capabilities. The main reasons for choosing FCNNs include: Powerful Feature Extraction Ability: FCNNs, through multiple layers of nonlinear transformations, can automatically learn and extract complex feature representations from raw data. This deep feature learning mechanism enables FCNNs to excel in handling datasets with high-dimensional feature spaces. Handling Nonlinear Relationships: FCNNs are adept at capturing nonlinear relationships in data, which is crucial for complex data pattern recognition involved in this study. Generalization Capability: Properly trained FCNNs possess excellent generalization abilities, allowing them to make accurate predictions on new, unseen data. High Flexibility and Customizability: The architecture of FCNNs can be adjusted according to specific data and task requirements. By altering the number of layers, neurons, and activation functions, networks can be constructed to accommodate various complexity needs.

3.3 StaGen (Stabilizer & Generalizer)

StaGen–RF The StaGen-RF (Random Forest with Stabilizer and Generalizer Neural Network) model is the core and innovative aspect of this report. Its purpose is to further reduce the prediction loss of the Random Forest and enhance its stability and generalization capabilities. As mentioned in the data analysis section, the dataset for this case study is not very large. The test set results obtained during model training are significantly influenced by the dataset partitioning. Even the high-performing Random Forest model exhibits a certain degree of instability. In the test set results, sometimes very low loss is achieved, which may seem like a favorable outcome, but it is largely due to the dataset partitioning. This does not necessarily indicate successful model training but rather a coincidental split of the dataset, where simpler samples end up in the test set. This actually reflects a lack of successful learning and inadequate generalization capability in the model. Therefore, this report introduces the StaGen module, aimed at optimizing the predictions of the Random Forest.

StaGen module

As illustrated in the Figure 1, the StaGen module is essentially a Fully Connected Neural Network (FCNN) that receives the predictive values from the Random Forest as input. The process involves a series of transformations: linear transformation, ReLU activation function, BatchNormalization layer, and finally an output layer to yield the ultimate prediction. The dimensions of both the input received from the Random Forest predictions and the final output prediction are 7.

3.4 Create StaGen dataset

It is a critical procedure to create a appropriate dataset for training the StaGen module. For data collection, the original Biomass dataset was used to train the Random Forest 40 times. Each training iteration involved the following steps: Dataset Splitting: First, the Biomass dataset is divided into 80% training samples and 20% test samples. Random Forest Training: A Random Forest model is

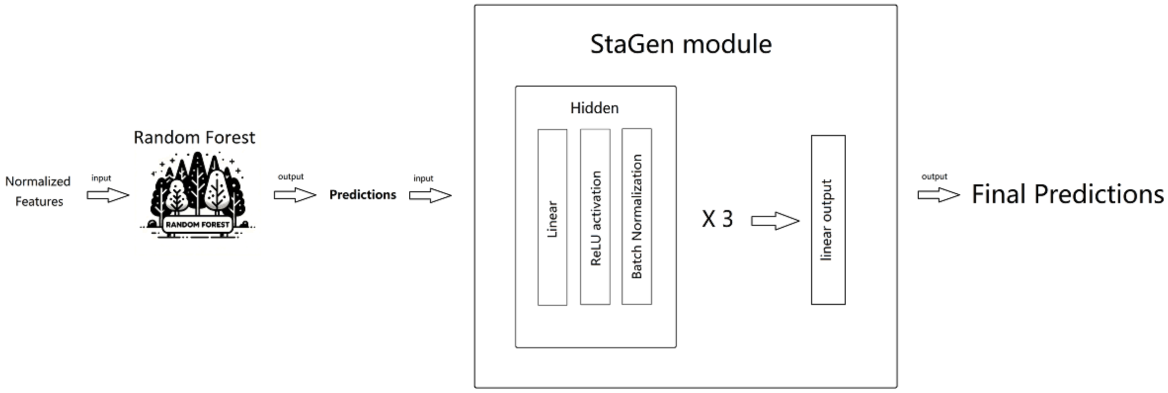

Figure 1: Structure of StaGen-RF

then trained using the training samples. Prediction Generation: The features of the test set samples are fed into the trained Random Forest model to generate predictive values for the test set. Dataset Formation for StaGen: These predictive values are used as features, and the actual values of the test set are used as labels. This combination forms a new dataset. This process is repeated 40 times, with a new split of the Biomass dataset in each iteration. Finally, the StaGen dataset was collected and contains 2680 rows (336 * 0.2 *40) and 14 columns (

7 features: RFpredictions of H2, CO, CO2, CH4, Lower heating value, Char yield, tar yield

7 labels: H2, CO, CO2, CH4, Lower heating value, Char yield, tar yield

). This dataset embodies the corresponding relationships between the predictions made by 40 Random Forest models and their actual values. The role of the StaGen module is to learn this relationship and acquire the ability to optimize the predictive values of the Random Forest.

3.5 Evaluation Metrics

In this project, three evaluation metrics will be used to assess the predictive and generalization capabilities of the model based on its performance on the test set.

Root Mean Squared Error (RMSE)

To evaluate the precision of the model's predictions, we have chosen Root Mean Square Error (RMSE) as the primary metric for performance assessment.

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

RMSE is the square root of the average of the squared differences between the actual observed values and the model's predicted values. It provides an intuitive measure of the model's prediction error, reflecting the variability of the predictions and their deviation from the actual values. The use of RMSE offers several key advantages that make it suitable for this study:

Sensitivity: RMSE is particularly sensitive to larger errors, as it squares the errors during calculation, amplifying the impact of outliers. This is beneficial in many practical applications because larger prediction errors are typically more undesirable than smaller ones and should be given more attention.

Interpretability: Since RMSE is on the same scale as the original data, it provides an intuitive assessment of model performance. Researchers and practitioners can directly compare the RMSE value with the actual range and context of the data, making it easier to interpret its meaning.

Consistency: As a commonly used standard, RMSE facilitates comparison with results from other studies or different models. This consistency is crucial for establishing the reliability and validity of the research.

Average Root Mean Squared Error (RMSE-Ave)

It is necessary to define a comprehensive evaluation of models, this report adopts the mean of 7 RMSE values to measure the performance of a certain model

Metric	RF	XGBoost	FCNN
Average RMSE	4.4	4.77	4.78
R2 Score	0.75	0.70	0.68
RMSE - H2	5.18	4.99	5.25
RMSE - CO	3.54	4.18	3.86
RMSE - CO2	3.77	4.19	4.50
RMSE - CH4	1.60	1.73	1.74
RMSE - LHV	1.08	1.15	1.39
RMSE - char yield	5.18	5.02	4.87
RMSE - tar yield	10.49	12.16	11.88

Table 1: Evaluations of models

4 Result

4.1 Phase 1: Selecting the best foundational model

The original dataset was divided into a training set and a test set, accounting for 80% and 20% of the total number of samples, respectively. It was found that the experimental results were considerably influenced by the division of the dataset, where the randomness introduced by the splitting into training and testing sets contributed to significant instability in the experimental outcomes. This could be attributed to the relatively small sample size of the dataset, which resulted in the training outcomes being highly sensitive to the dataset division. Additionally, certain features exhibited uneven distributions and even outliers, potentially increasing the instability of the dataset partitioning process.

Table 1 displays the performance of three foundational models on the test set.

Random Forest: The Random Forest model achieved commendable results on the test set and demonstrated relatively stable characteristics across multiple experiments. It was less affected by the randomness introduced by dataset splitting compared to other models. In most experimental results, it consistently yielded the lowest loss and the best generalization capabilities.

XGBoost: XGBoost also performed well, with several training rounds showing results close to, and occasionally surpassing, those of the Random Forest. However, considering of multiple training outcomes, XGBoost was still defeated by Random Forest in terms of fitting ability, generalization capability, and loss scores. Furthermore, compared to Random Forest, XGBoost showed less stability and was more sensitive to the randomness introduced by experimental procedures.

Fully Connected Neural Network (FCNN): The performance of the FCNN was moderate. Compared to Random Forest, its training results were more variable, likely due to the randomness introduced by network weight initialization. To mitigate this, the experiments attempted using Xavier normal initialization instead of PyTorch's default Xavier uniform, where weight parameters followed a normal distribution. After trying various optimizers, AdamW was selected as the best optimizer for training. With these configurations, the neural network's training results improved in terms of loss and stability but still did not surpass Random Forest, with an RMSE-avg loss approximately 0.6 higher. Several hypotheses could explain these observations: 1.The task's complexity is not high, and the difficulty of feature extraction is manageable, making traditional machine learning adequate. 2.The small sample size of the dataset may not support the full fitting and generalization enhancement of a more complex neural network. 3.The network structure has not been fully explored. There is potential for improvement in combinations of hyperparameters, network structure, neuron count, and hidden layer design.

After comprehensive consideration of various evaluation metrics such as RMSE-Avg and R2, along with other non-quantified criteria like stability, low sensitivity to random factors, model complexity, and training complexity, it is concluded that the Random Forest is the most suitable model for this case study among the evaluated machine learning models. Consequently, Random Forest has been chosen as the foundational model for creating the StaGen dataset and integrating the StaGen module as the final model for the study.

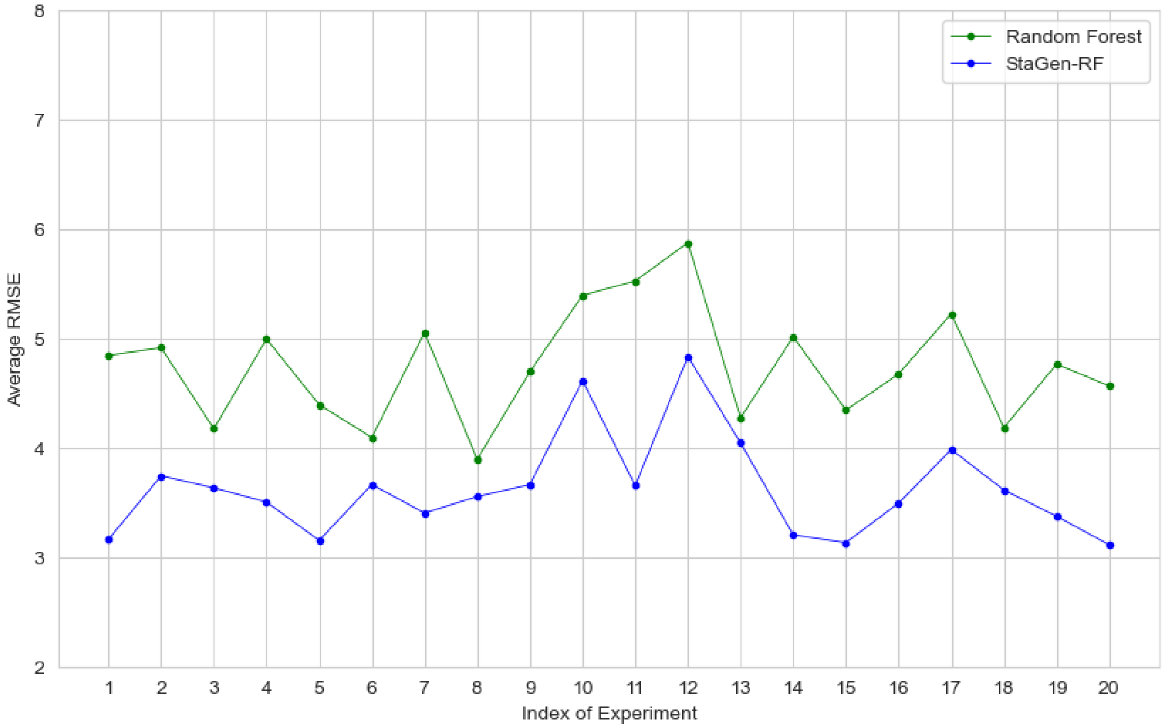

Figure 2: Comparison betwenn RF and StaGenRF

4.2 Phase 2: Combining Random Forest and StaGen Module

In order to assess the enhancement effect of the StaGen module, an additional 20 rounds of Random Forest training will be conducted in the second phase. The predictive values of the test set from the Random Forest model will be fed into the StaGen module to obtain the final predictions. The RMSE-Avg will be calculated for the predictive values both before and after the integration with the StaGen module. This process will result in two lists, each containing 20 test results.

The graph illustrates the performance on the test set of 20 Random Forest training iterations, showcasing both the direct outputs (pre-StaGen integration) and the outputs after integration with the StaGen module. For each training iteration, the Biomass dataset was re-partitioned.

The improvement brought by the StaGen module to the Random Forest is evident from the graph. It not only reduces the RMSE loss of the Random Forest but also enhances its stability and mitigates the volatility introduced by dataset partitioning.

By calculating the average and standard deviation of these two lists, the benefit brought by the StaGen module to the Random Forest can be measured quantitatively. After computing their averages, the original Random Forest's average loss is found to be 4.75, while the StaGen-RF's average loss is 3.63. This indicates a significant improvement, surpassing the gains typically achieved through modifying model structures or tuning hyperparameters.

On the other hand, calculating the standard deviation of these two arrays reveals that the standard deviation for the Random Forest is 0.51, while for StaGen-RF, it is 0.45. This reduction in standard deviation indicates that integrating the StaGen module results in a more stable model performance, less affected by dataset partitioning or other random factors present in the training process. It suggests an enhancement in the model's generalization capability, an important aspect for its robust performance on unseen data.

5 Conclusion

As the result shown in the Phase 1, the Random Forest performed exceptionally well in this task compared to other models. However, the integration of the StaGen module still managed to bring substantial improvements. Fundamentally, the core idea behind the StaGen module is an ensemble learning approach. The training of the StaGen module is based on the predictive results of the previous 40 Random Forest modelsTherefore, the study demonstrates that using a neural network to learn the characteristics and patterns of 40 Random Forest models is a promising strategy. This approach significantly enhances the performance of the foundational Random Forest model. However, it's important to acknowledge that the ultimate performance of the StaGen-RF model still largely depends on the performance of the underlying Random Forest. A poorly trained Random Forest cannot have its flaws simply reversed by connecting it to the StaGen module. The StaGen module can enhance but not completely compensate for fundamental weaknesses in the base model. This underscores the importance of ensuring a robust and well-performing foundational model before integrating additional layers or modules like StaGen.

References

[1] Jun Young Kim, Dongjae Kim, Zezhong John Li, Claudio Dariva, Yankai Cao, and Naoko Ellis. Predicting and optimizing syngas production from fluidized bed biomass gasifiers: A machine learning approach. *Energy*, 263:125900, 2023.

[2] Jun Young Kim, Ui Hyeon Shin, and Kwangsu Kim. Predicting biomass composition and operating conditions in fluidized bed biomass gasifiers: An automated machine learning approach combined with cooperative game theory. *Energy*, 280:128138, 2023.

[3] Antonio Molino, Vincenzo Larocca, Simeone Chianese, and Dino Musmarra. Biofuels production by biomass gasification: A review. *Energies*, 11(4), 2018.