Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

A Comparative Study of Aspen Plus and Machine Learning Models for Syngas Prediction in Biomass-Plastic Waste Co-gasification

Usman Khan Jadoon[a]*, Ismael Díaz[a], Manuel Rodríguez[a]

[a] *Departamento de Ingeniería Química Industrial y del Medioambiente, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Madrid, Spain*
* Corresponding Author: uk.jadoon@upm.es

ABSTRACT

The co-gasification of biomass and plastic waste offers a promising pathway for sustainable syngas production, necessitating precise prediction of its composition to optimize efficiency. This study compares the performance of Aspen Plus models, including the thermodynamic equilibrium model (TEM) and restricted thermodynamic equilibrium model (RTM), with machine learning (ML) techniques, focusing on the support vector regression (SVR) for syngas prediction during steam and air co-gasification. Aspen Plus simulations provided valuable mechanistic insights, while the ML model demonstrated superior predictive accuracy. The SVR, enhanced by principal component analysis (PCA), significantly improved performance, achieved R^2 values of 0.879 for H_2, 0.856 for CO, 0.859 for CO_2, and 0.744 for CH_4 on the testing dataset. It also outperformed other models in terms of RMSE, achieving exceptional precision for CH_4 (0.0087), CO (0.0193), and H_2 (0.0194). In contrast, RTM exhibited moderate accuracy with minimal deviations, while TEM showed the highest RMSE values across all predictions, indicating its limited reliability for real-world applications. This study advances gasification technologies by demonstrating the advantages of ML models, particularly SVR, in predicting syngas composition and supporting the global transition toward sustainable energy systems.

Keywords: Biomass, Aspen Plus, Modeling and Simulations, Plastic wastes, Syngas prediction

INTRODUCTION

The transition to sustainable energy production necessitates the optimization of thermochemical conversion processes, with co-gasification of biomass and plastic waste emerging as a promising route for syngas generation. Accurate prediction of syngas composition is crucial for efficient process design and operation. Traditional modeling approaches, primarily thermodynamic equilibrium and kinetic models implemented in process simulation software like Aspen Plus, have served as the foundation for gasification studies. These models leverage fundamental principles of thermodynamics and reaction kinetics to simulate the complex interactions within the gasifier. However, they often struggle to capture the non-linearities and intricate phenomena inherent in these processes, especially when dealing with mixed feedstocks [1].

Machine learning (ML) is increasingly being explored as a complementary or alternative approach, offering the potential to enhance predictive accuracy. ML techniques, such as random forest, artificial neural networks (ANNs), and support vector machines (SVMs) can learn complex relationships directly from data, bypassing some of the limitations of purely mechanistic models [2]. ML's capacity to learn from high-dimensional data and capture non-linear relationships has made them a popular choice for modeling complex systems. Significant research has focused on both Aspen Plus modeling and ML-based modeling of gasification. For example, Aspen Plus studies often focus on the effects of different operating conditions such as temperature, air/steam to fuel ratio on syngas composition . ML studies, on the other hand, tend to concentrate on developing predictive models that are trained

using either experimental data or output from process simulation software [3]. The comparison of these two approaches and the exploration of their potential synergy, however, is less well-explored in the current literature.

While both Aspen Plus and ML techniques have individually contributed to understanding and predicting syngas production from gasification, the direct comparative analysis of these methods, particularly in the context of co-gasification, is lacking. This research directly addresses this gap by comparing the predictive performance of Aspen Plus models, TEM and RTM, and surrogate ML model, SVM, for the co-gasification of biomass and plastic waste, and exploring the advantages each technique bring. The novelty of this study lies in the integration of these approaches and offers a comprehensive comparison of their prediction accuracy in prediction of syngas compositions. While conventional gasification methods and simulation tools like Aspen Plus have been widely used to study biomass and plastic waste gasification, this study will also explore how ML model trained on experimental data can overcome limitations in mechanistic models for predicting syngas outputs [4]. This approach has the potential to contribute towards a more accurate and data-driven approach to design and optimization of co-gasification plants.

METHODOLOGY

Experimental data on steam-O_2 biomass plastic gasification, based on a mixture of rice husk (RH), polyethylene PE, polyethylene terephthalate (PET), and tyres, was collected from the published articles by Brachi et al. (2014), Pinto et al. (2002), and Pinto et al. (2016) [5-7]. Experiments were performed on a bench scale with varying temperature and steam-to-biomass ratio. To conduct the ML model training, the eleven input features selected were ash, moisture, volatile matter, and fixed carbon from the proximate analysis. From the ultimate analysis, carbon, hydrogen, nitrogen, and oxygen were selected, and from the process variables, reactor temperature, steam-to-biomass ratio (S/B), and equivalence ratio (ER) were included. The dependent features are the molar fraction of the syngas composition H_2, CO, CO_2 and CH_4. The dataset consists of 40 rows with eleven input and four output features, so PCA was applied to reduce dimensionality before training the ML model to avoid possible overfitting. The data was then standard scaled, and 10-fold cross-validation was performed, splitting the data into an 80/20 ratio for training and testing purposes.

Equilibrium Modeling

This study employed Aspen Plus V.14.0 to model steam co-gasification and predict syngas composition under various conditions, utilizing both equilibrium and kinetic modeling strategies. Biomass, PE, and char were defined as non-conventional components using proximate and ultimate analyses from the ULTANAL and PROXANAL models, with enthalpy and density data provided by HCOALGEN and DCOALIGT models, respectively, based on biomass combustion and composition. Ash was treated as inert, and the Peng-Robinson thermodynamic package with the Boston-Mathias function was selected for high-temperature gasification simulations. Two equilibrium models were used: the TEM, which minimizes Gibbs free energy to predict product composition but neglects reaction kinetics and often overestimates or underestimates syngas, and the RTM, which incorporates temperature-dependent constraints to better reflect real-world gasification by minimizing Gibbs free energy based on elemental composition without requiring detailed kinetic data.

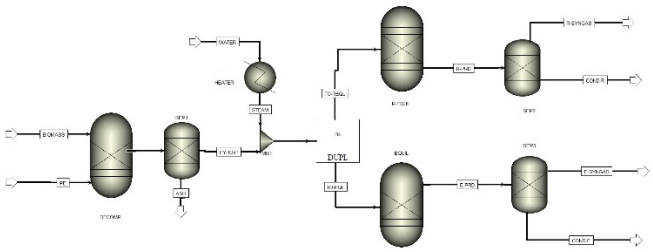

Figure 1: Aspen Plus modeling of TEM and TRM

In this study the Aspen Plus model structure developed by Acar et al., (2019) for gasification, was adopted, using both TEM and RTM [8]. Biomass and plastic waste were decomposed in a RYIELD reactor (DECOMP) into constituent elements based on ultimate and proximate analyses, with ash separated using a separator (SEP1). The components were mixed with steam and fed into an RGIBBS reactor, where the gasifier (EQIL) simulated the TEM approach and another reactor (R-GASI) modeled RTM. Syngas was separated from byproducts using separators (SEP2 for R-GASI and SEP3 for EQIL), as illustrated in Figure 1. Following Acar et al. (2019), four key reactions (Table 1) were used to restrict chemical equilibrium in RTM, and sensitivity analysis was conducted for each of the 40 experimental instances by varying the approach temperature of individual reactions, generating 22,500 cases per instance. The absolute difference between the molar fractions of the predicted and experimental syngas composition was calculated using Equation 1, where ME represents the minimum error, *e* is the experimental value, *p* is the predicted value, and *i* denotes the *i*-th case in the sensitivity analysis [9].

$$(ME)_i = |H_2^e - H_2^p|_i + |CO^e - CO^p|_i + |CO_2^e - CO_2^p|_i + |CH_4^e - CH_4^p|_i \quad (1)$$

Table 1: Reactions used in RTM

No.	Reaction	ΔHr (kJ/mol)
R1	$C + H_2O \rightarrow CO + H_2$	131
R2	$C + 0.5O_2 \rightarrow CO$	-111
R3	$CO + H_2O \rightarrow CO_2 + H_2$	-41.2
R4	$CH_4 + H_2O \rightarrow CO + 3H_2$	206

Dimensionality Reduction

In order to address the curse of dimensionality and possible multicollinearity among the input variables and also to represent independent variables with minimal redundancy, PCA has been used, which can find feature components that have low and high variance in a given data set [10]. PCA is a multivariate technique that allows one to transform a series of correlated variables into a set of uncorrelated variables known as principal components (PCs). The PCs of a data set *x* are obtained by solving an eigenvalue–eigenvector problem. In this regard, the covariance matrix can be computed by equation 2. Where *n* is total number of samples in the data set and $\overline{x_i} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$M_c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_i})(x_i - \overline{x_i})^T \quad (2)$$

Machine learning

SVR is a powerful supervised learning algorithm based on the principles of Support Vector Machines (SVMs). Unlike traditional regression models, SVR aims to find a function that not only minimizes error but also avoids overfitting by balancing complexity and predictive performance. SVR achieves this by defining a margin of tolerance within which predictions are considered acceptable deviations from the actual data points. In SVR, the objective is to find a hyperplane that best fits the data points, allowing for some deviations. The SVR function can be expressed in equation 3, where w is weight vector, x is input feature vector, and b is bias term.

$$f(x) = [w, x] + b \quad (3)$$

Performance Evaluation

The accuracy and reliability of the surrogate models were evaluated using Coefficient of Determination (R^2), RMSE, and Mean Absolute Error (MAE) for both training and testing datasets. R^2, as defined by Equation 4 measures how well the model explains variance in the target variable. RMSE, defined by Equation 5, quantifies the average prediction error with greater sensitivity to outliers. In addition, MAE, as described in Equation 6, provides a straightforward measure of overall error magnitude. These evaluation metrics were specifically applied to the molar fractions of syngas components—H_2, CO, CO_2, and CH_4—where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\overline{y}_i$ is the mean of the observed values, and *n* is the number of data points.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

RESULTS AND DISCUSSION

SVR with a radial basis function kernel was employed to develop a predictive model based on the eleven input features, utilizing MultiOutputRegressor to handle multiple output variables. The model was configured with an RBF kernel, a regularization parameter of 1.0, and an epsilon value of 0.1, which defines the margin of tolerance for error. The model's accuracy was assessed using R^2, RMSE, and MAE for both training and testing datasets. The results are summarized in **Table 2**, values in parentheses representing the test dataset. The results indicate that the SVR model performed well in capturing the variance in the training set, with R^2 values exceeding 0.78 across all outputs. However, a decrease in R^2 for the testing set shows overfitting.

Table 2: Performance matrix of train(test) based on real input features

	R^2	RMSE	MAE
H_2	0.894(0.652)	0.022(0.040)	0.016(0.030)
CO	0.966(0.862)	0.015(0.031)	0.012(0.025)
CO_2	0.936(0.779)	0.023(0.043)	0.015(0.031)
CH_4	0.784(0.577)	0.013(0.019)	0.008(0.014)

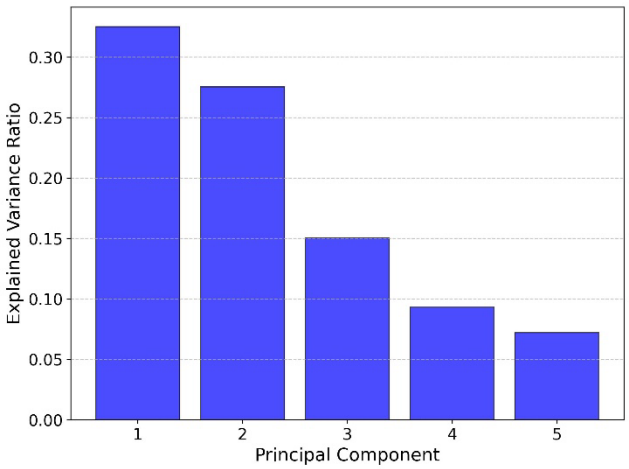

Figure 2: Explained variance for every PC

To address potential overfitting and improve model performance, PCA was applied to reduce the dimensionality of the input features while retaining essential information. The dataset, sourced from published studies, was first standardized using the Standard Scaler to

ensure uniform scaling across variables. PCA was then performed, and cumulative explained variance analysis showed that five principal components captured over 90% of the total variance. This transformation effectively simplified the data structure while preserving key patterns, allowing for a refined input representation for subsequent modeling steps. The cumulative explained variance, shown in **Figure 2**, further highlights the efficiency of PCA in retaining maximum information. The bars in the graph indicate that the first PC accounts for over 30% of the variance, the second PC captures approximately 28%, and the third PC retains around 15%, with the first three PCs collectively preserving more than 75% of the total variance. This finding provides a solid foundation for further analysis using the reduced dataset, ensuring computational efficiency without compromising data integrity. Elmaz et al., (2020) applied PCA to reduce input dimensionality in a downdraft biomass gasification study before training SVR, DTR, and MLP models for prediction [11].

Table 3 presents the performance metrics (R^2, RMSE, and MAE) for each syngas component, with values in parentheses representing the testing dataset. These results are based on PCA-reduced input features, which significantly improved model performance. H_2 exhibits the highest R^2 values for both training (0.977) and testing (0.853), along with favorable RMSE and MAE values, indicating superior predictive accuracy. CO demonstrates comparable training performance (R^2 = 0.950) but surpasses H_2 in testing (R^2 = 0.878), suggesting improved generalization. While CO's training RMSE (0.0191) is lower than that of H_2 (0.0240), its other error metrics remain slightly below H_2. CO_2 and CH_4 show lower R^2 values in both training and testing datasets, with CH_4 exhibiting the lowest R^2 (0.744) for testing. However, CH_4 also has the lowest RMSE and MAE, indicating relatively small errors despite lower variance capture. The improvements observed in these results highlight the effectiveness of PCA in reducing dimensionality while preserving essential information, thereby enhancing the SVR model's predictive capabilities. The performance variations across syngas components may stem from differences in their formation mechanisms or disparities in the quality and quantity of available training data. Further investigation into these factors could refine model accuracy even further.

Table 3: Performance matrix of train(test) based on PCA reduced features

	R^2	RMSE	MAE
H_2	0.977(0.853)	0.0240(0.038)	0.016(0.029)
CO	0.950(0.878)	0.0191(0.035)	0.014(0.027)
CO_2	0.942(0.805)	0.0223(0.040)	0.015(0.028)
CH_4	0.944(0.744)	0.0135(0.018)	0.008(0.013)

*Performance matrix for SVR is given as train(test)

Table 4 presents the predicted values from all models, clearly showing that SVR consistently outperformed the other techniques across all syngas components by achieving the lowest RMSE values. Among the predictions, the most accurate prediction by SVR was for CH_4, with an exceptionally low RMSE of 0.0137, signifying high precision. Similarly, SVR also demonstrated strong predictive performance for CO and CO_2, with RMSE values of 0.0197 and 0.0238, respectively. Although the RMSE for H_2 (0.0242) was slightly higher than that of RTM (0.0238), SVR still outperformed the remaining models by a considerable margin.

Table 4: RMSE values of different modeling techniques

	RTM	TEM	SVR
H_2	0.0238	0.1971	0.0242
CO	0.0391	0.0911	0.0197
CO_2	0.0539	0.1038	0.0238
CH_4	0.0488	0.1125	0.0137

In comparison, the RTM showed commendable prediction accuracy, though not as precise as SVR. It achieved moderately low RMSE values for H_2 (0.0238), CO (0.0391), CO_2 (0.0539), and CH_4 (0.0488), making it a viable alternative in some scenarios. However, its performance for the H_2/CO ratio, with an RMSE of 0.2569, was less accurate compared to SVR. Conversely, TEM performed the weakest among the three techniques, with significantly higher RMSE values across all predictions. For example, the RMSE for H_2 was 0.1971, almost ten times higher than that of SVR, while the RMSE values for CO (0.0911), CO_2 (0.1038), and CH_4 (0.1125) further highlighted its lack of precision.

In this Study the predictions obtained from different models were analyzed and plotted to evaluate their performance. **Figure 3** illustrates the comparison between experimental and predicted H_2 values across various feedstock compositions and process variables. Among the models, the SVR demonstrates exceptional predictive accuracy, with green data points closely aligning with the experimental values for all the experimental values. The minimal error observed in SVR predictions underscores its reliability and effectiveness. In contrast, the RTM also performed well for the H_2 concentration above 35%. The TEM model, on the other hand, consistently overpredicts H_2 values across all feedstock compositions, likely due to limitations in Gibbs free energy estimations for syngas constituents, as highlighted by Ajorloo et al. (2022) [1].

The predictions of all models for CO against the experimental values are given in the **Figure 4**. Just like the predictions of H2 the SVR also has excelled in predicting the CO in this study. The SVR model consistently demonstrates superior predictive performance, with the predicted values (green points) closely matching the experimental results. The RTM model performs relatively well

in predicting CO values at lower concentrations but exhibits slight deviations at higher concentrations. Conversely, the predictions by TEM model are quite diverse as shown by the red points, reflecting potential inaccuracies in the equilibrium modeling of CO. This is also evident by the RMSE values from the **Table 4** where TEM has highest RMSE among other models.

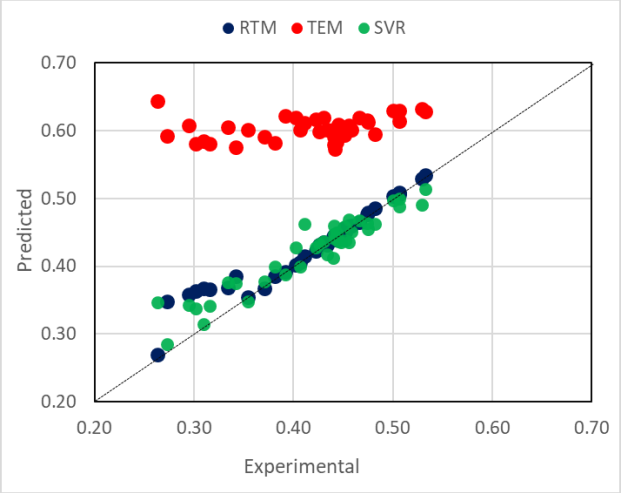

Figure 3: Experimental and predicted mole fraction of H_2

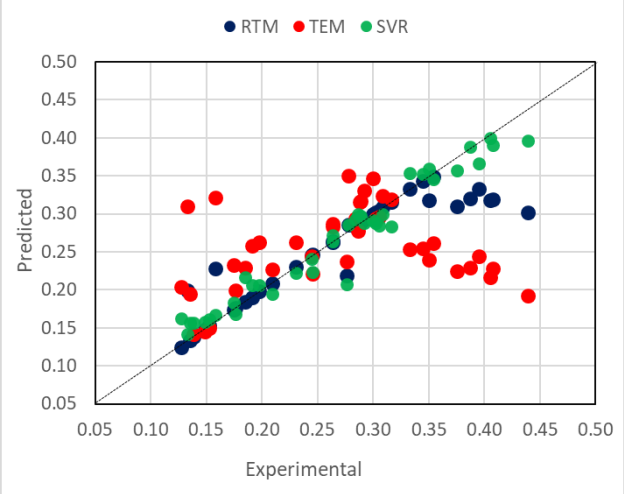

Figure 4: Experimental and predicted mole fraction of CO

Figure 5 and Figure 6 shows the predicted values by the models for CO2 and CH4 respectively. Similar to the predictions of H_2 and CO, the SVR model demonstrates excellent agreement with experimental data, as the predicted values align closely with the experimental results. On the other hand The RTM model shows reasonable accuracy for higher CO_2 concentrations but fails to maintain the same precision for predicting CH_4. The TEM model, however, systematically underpredict CO_2 values in Figure 5 and predictions for the CH_4 are quite lower below 1% even sometimes the predicted values for CH_4 are like 10^{-5} which raises serious question mark on the predictive accuracy and use of TEM for the real world scenario. SVR remained most accurate predicting model which shows that data based modeling outperforms the Aspen Plus, TEM and RTM, modeling strategy.

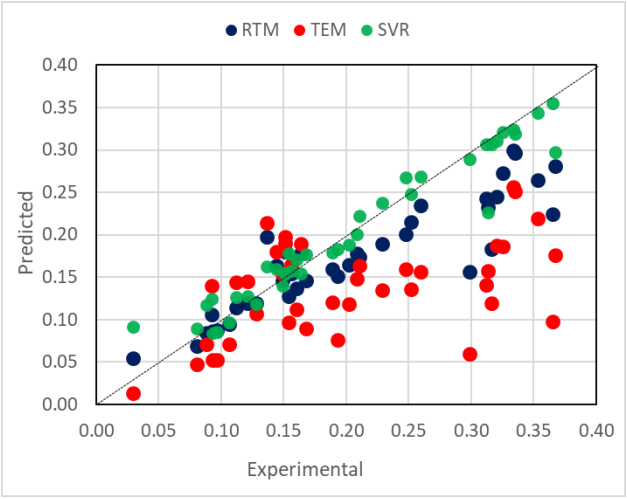

Figure 5: Experimental and predicted mole fraction of CO_2

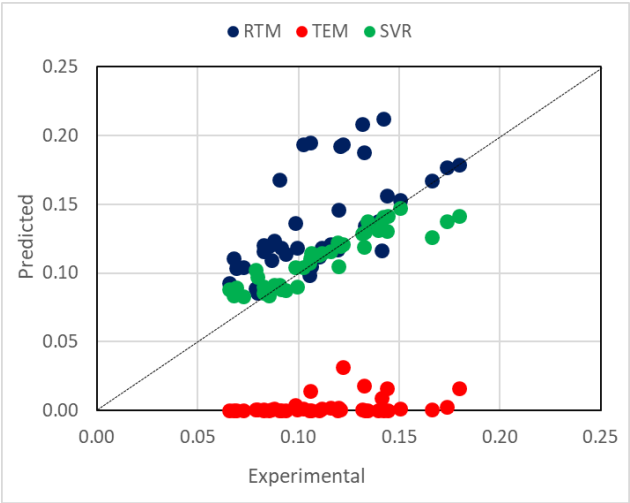

Figure 6: Experimental and predicted mole fraction of CH_4

CONCLUSIONS

In this study, a comparative analysis of Aspen Plus models and machine learning techniques was conducted to evaluate their performance in predicting syngas composition during the co-gasification of biomass and plastic wastes. The results clearly demonstrate that the SVR, in accurately capturing the non-linearities of the gasification process. Without the use of PCA, the SVR model achieved high performance in training dataset, particularly for syngas components like CO and CH_4, with RMSE values as low as 0.0197 and 0.0137, respectively, across training and testing datasets. However, the application of

PCA to reduce the dimensionality of the input features resulted in substantial improvements in model accuracy. For instance, after applying PCA, the RMSE for CH_4 further decreased to 0.0135, while the R^2 for H_2 increased from 0.894 to 0.977 for training and 0.652 to 0.853 for testing, indicating better generalization.

While the SVR model with PCA consistently outperformed the other techniques across all syngas components, achieving notably low RMSE values, the RTM showed commendable accuracy, although not as precise as SVR. The RTM model achieved moderately low RMSE values for H_2 (0.0238), CO (0.0391), CO_2 (0.0539), and CH_4 (0.0488), but fell behind SVR in terms of predictive precision. The TEM model, on the other hand, performed the weakest, with significantly higher RMSE values for all components. For example, the RMSE for H_2 was 0.1971, almost ten times higher than that of SVR, underscoring the strength of SVR in handling non-linear relationships and high-dimensional data.

These findings highlight the importance of dimensionality reduction through PCA when applying machine learning to predict complex systems such as syngas composition. The superior performance of the PCA-enhanced SVR model, coupled with its ability to generalize better across different feedstock compositions and operating conditions, emphasizes its potential as a robust predictive tool in renewable energy technologies. Furthermore, this study underscores the potential of ML models to complement traditional mechanistic models like Aspen Plus, providing valuable insights into syngas production processes, thus advancing the transition toward more sustainable energy systems.

ACKNOWLEDGEMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 945139

REFERENCES

1. Ajorloo, M., et al., Recent advances in thermodynamic analysis of biomass gasification: A review on numerical modelling and simulation. Journal of the Energy Institute, 2022. **102**: p. 395-419.
2. Wen, H.-T., J.-H. Lu, and M.-X. Phuc, Applying artificial intelligence to predict the composition of syngas using rice husks: A comparison of artificial neural networks and gradient boosting regression. Energies, 2021. **14**(10): p. 2932.
3. Mutlu, Ö.Ç. and T. Zeng, *Challenges and opportunities of modeling biomass gasification in Aspen Plus: A review.* Chemical Engineering & Technology, 2020. **43**(9): p. 1674-1689.
4. Xue, P., et al., Prediction of syngas properties of biomass steam gasification in fluidized bed based on machine learning method. International Journal of Hydrogen Energy, 2024. **49**: p. 356-370.
5. Brachi, P., et al., Fluidized bed co-gasification of biomass and polymeric wastes for a flexible end-use of the syngas: Focus on bio-methanol. Fuel, 2014. **128**: p. 88-98.
6. Pinto, F., et al., Effect of gasification agent on co-gasification of rice production wastes mixtures. Fuel, 2016. **180**: p. 407-416.
7. Pinto, F., et al., Co-gasification study of biomass mixed with plastic wastes. Fuel, 2002. **81**(3): p. 291-297.
8. Acar, M.C. and Y.E. Böke, Simulation of biomass gasification in a BFBG using chemical equilibrium model and restricted chemical equilibrium method. Biomass and Bioenergy, 2019. **125**: p. 131-138.
9. Jadoon, U.K., I. Díaz, and M. Rodríguez, Comparative analysis of aspen plus simulation strategies for woody biomass air gasification processes. Biomass and Bioenergy, 2025. **194**: p. 107626.
10. Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 2016. **374**(2065): p. 20150202.
11. Elmaz, F., Ö. Yücel, and A.Y. Mutlu, Predictive modeling of biomass gasification with machine learning-based regression methods. Energy, 2020. **191**: p. 116541.