Robust Modeling of Syngas Composition, Energy ratio and Exergy Efficiency of Biomass Gasifiers Using Machine Learning Technics

Virginia Samca Montaño[a], Brunno F. Santos[b] and Florian Pradelle[a]

[a] Department of Mechanical Engineering (DEM) and Institute for Mobility and Sustainable Energy (IMES), Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, samca.virginia100@gmail.com; pradelle@puc-rio.br

[b] Department of Chemical and Materials Engineering (DEQM), Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, bsantos@puc-rio.br

Abstract:

Gasification is a thermochemical process for converting biomass into sustainable bioenergy, applicable in internal combustion engines, gas turbines, and the production of low-carbon hydrogen. Traditional thermodynamic modeling approaches, based on kinetics or equilibrium, are complex and often exhibit limited generalization capacity in input–output relationships, restricting their applicability under variable operating conditions. In contrast, machine learning (ML) provides an efficient and flexible alternative, capable of handling large datasets and generating predictive models with greater robustness. In this study, three ML techniques were implemented: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), with and without data balancing using the SMOGN algorithm. ANN models were developed in Scikit-learn, testing architectures with 1–3 hidden layers (10–40 neurons) and activation functions ReLU and tanh. SVM models were trained with the Radial Basis Function (RBF) kernel, while RF employed MultiOutputRegressor optimized through RandomizedSearchCV. These models were applied to predict syngas composition (H_2, CO, CO_2, and CH_4) using 581 experimental data points collected from the literature, covering diverse biomass types, gasifier agents, gasifiers, operating conditions, and physicochemical parameters. Results showed consistently high training performance across all models, with R^2 values between 0.94–0.97 for H_2 and CO_2, 0.83–0.95 for CO, and 0.81–0.86 for CH_4. In validation, RF emerged as the most robust model, achieving R^2 values of 0.89 for H_2, 0.72 for CO, 0.68 for CH_4, and 0.87 for CO_2, with low mean absolute errors. Data balancing with SMOGN did not substantially improve accuracy in ANN and SVM, though it provided greater stability in RF. Finally, energy ratio and exergy efficiency of gasification were evaluated through a residual analysis. The predictive performance of the RF model demonstrated, achieving a coefficient of determination R^2 of 0.978 and MAE of 1.505 for the energy ratio. Similarly, the prediction of exergy efficiency with an R^2 of 0.994 and a minimal MAE of 0.344. These low deviation metrics confirmed the robustness of ML methods in accurately mapping complex thermodynamic performance indicators and highlighted the potential of the developed model to satisfactorily predict the composition and conversion efficiency of this strategic vector under a large scope of gasifier types and operational conditions.

Keywords:

Artificial Neural Network (ANN); Gasification; Prediction; Random Forest (RF); Support Vector Machine (SVM).

1. Introduction

Biomass, as a renewable energy source, has the potential to contribute to achieve near-zero carbon dioxide emissions, contributing to the control of global temperature rise within the 1.5 °C limit established by the Paris Agreement [1]. Biomass conversion into biofuels —such as biogas, syngas, biodiesel, and biohydrogen [2] — represent key alternatives in the energy transition. Among them, gasification transforms carbonaceous materials into syngas, a gaseous mixture mostly composed CO, H_2, CH_4, H_2O and CO_2 [3–4]. However, accurately modeling this type of technology remains a challenge. Conventional thermodynamic equilibrium models often fail capture the kinetic constraints of high-temperature reactions (800–1200 °C), typically leading to a systematic overestimation of hydrogen and an underestimation of methane [4-5]. In response, data

science and machine learning (ML) provide tools capable of overcoming these limitations [6–7]. ML models identify complex patterns in data and predict syngas composition without requiring full knowledge of the underlying physical mechanisms [6,8]. Several studies have validated their application: Puig Arnavat et al. [6] developed artificial neural network (ANN) architectures to predict producer gas composition (CO, CO_2, H_2 and CH_4), and gas yield in fluidized bed gasifiers achieving coefficient of determination (R^2) higher than 0.97; Safarian et al. [9] reached R^2 higher than 0.999 in downdraft gasifiers by modeling net electrical power output for 86 types of biomass types; George et al. [8] obtained R^2 higher than 0.90 by assessing syngas composition in air -blown bubbling fluidized beds fed with wood residues. Furthermore, Pandey et al. [10] achieved R^2 higher than 0.99 by predicting the lower heating values (LHV) of the syngas and byproducts (tars and char), alongside syngas yield from municipal solid waste. These findings confirmed that ANNs are particularly effective in modeling thermochemical processes with complex nonlinear interactions [6,8]. The diversification of algorithms has further expanded these predictive capabilities. Random Forest (RF), based on ensembles of decision trees, reduces overfitting and has been applied in bioenergy to estimate biochar mass yield and carbon content, integrating life-cycle analyses [11–14]. Mutlu and Yucel [11] compared RF and SVM in downdraft reactors to classify discrete levels of syngas composition and its higher heating value (HHV), achieving accuracies between 89–96%. Support Vector Machines (SVM) have also been used to predict bio-oil yield and its heating value [15].

Beyond syngas composition prediction, sustainability indicators must be assessed. Energy analysis, based on the first and second laws of thermodynamics, measures cold gas efficiency (CGE), while exergy analysis is providing a general measure of the (potential) usefulness of any disequilibrium (apparent or latent) in nature. [16–17]. Recent studies have extensively quantified these efficiencies using diverse modeling frameworks. For instance, Vargas [18] utilized ANNs to predict chemical exergy and energy conversion for various Brazilian biomasses, reporting CGE of up to 82.21% for municipal waste and 80.66% for orange residues. Similarly, while Lewin [19], optimized energy efficiency (37.66%) during the co-gasification of municipal residues and sugarcane bagasse. Furthermore, Shahbeig et al. [20] provided a comprehensive review and numerical assessment of exergetic sustainability indices across different reactor configurations and feedstocks. In this context, the aim of the present study is to develop and compare robust predictive models for syngas composition and thermodynamic efficiencies, implementing three competitive techniques (ANN, SVM, and RF) and integrating the SMOGN algorithm for data balancing. This approach seeks to develop a comprehensive tool that can be used for the optimization of sustainable bioenergy systems.

2. Methodology

The overall methodological framework applied in this study is illustrated in Figure 1.

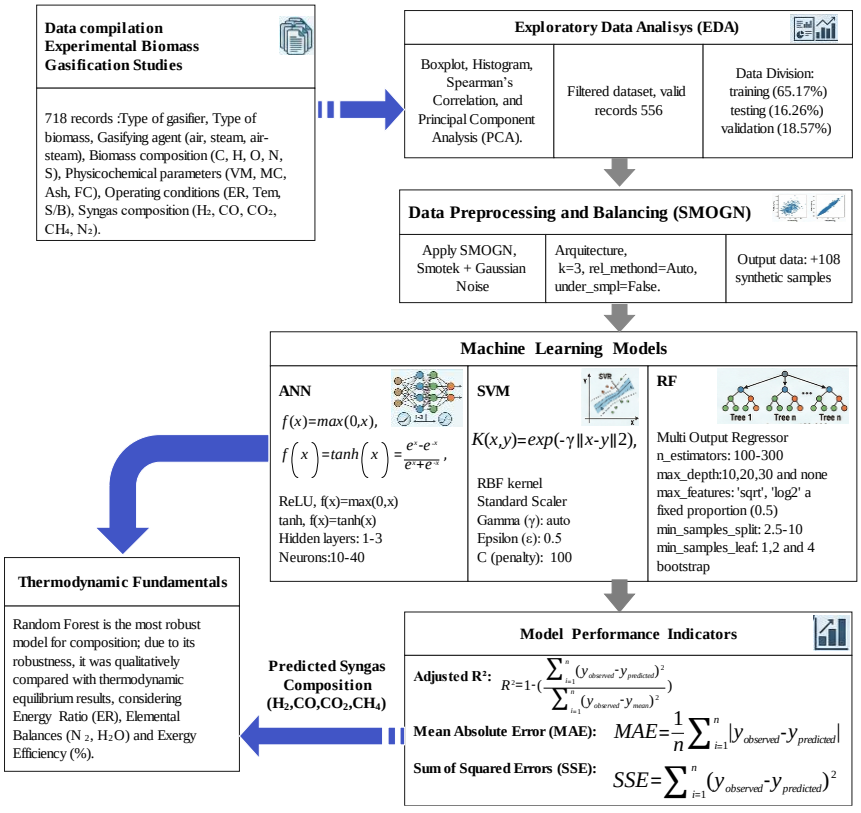

Figure. 1. *Methodological Workflow of the Study*

It integrates data science techniques (EDA, SMOGN, and machine learning models) with thermodynamic fundamentals to evaluate biomass gasification performance and considers the gaseous mixture behave as ideal gas.

2.1. Dataset Compilation and Preprocessing (SMOGN)

The dataset employed in this study was compiled by Pimentel et al. [21], focusing on experimental biomass gasification studies, and consisted of 718 records. The factors considered include: type of gasifier, type of biomass, gasifying agent, biomass composition, physicochemical parameters, operating conditions and syngas composition. The correlation heatmap illustrates the impact of biomass properties on syngas composition. Fixed carbon and volatile matter strongly drive CO, H_2, and CH_4 formation, whereas oxygen, moisture, and ash contents reduce efficiency and favor CO_2 and N_2 production.

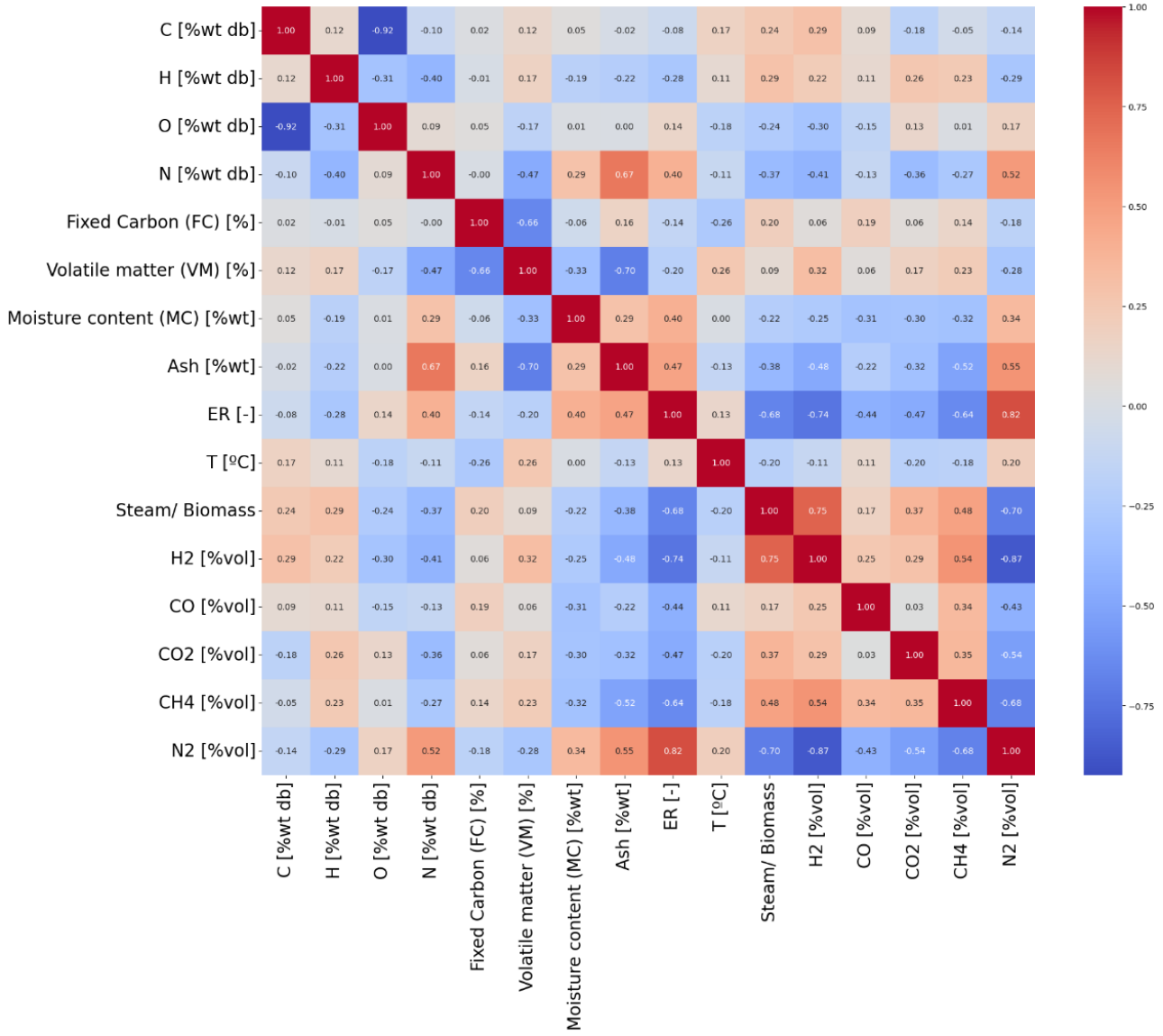

Figure. 2. *Correlation spearman showing the influence of biomass properties on syngas composition. Fixed carbon (FC) and volatile matter (VM) strongly enhance CO, H_2, and CH_4 yields, while oxygen, moisture, and ash contents limit efficiency and shift production toward CO_2 and N_2.*

To ensure data quality and consistency, an Exploratory Data Analysis (EDA) was conducted to remove outliers. After filtering, the final dataset consisted of 556 valid records, of which 81.43% were allocated to training/testing (80/20 split) and 18.57% to independent validation. Due to the imbalance in the target variables, the SMOGN algorithm (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) was applied. This hybrid method balances regression data by combining undersampling in dense regions with synthetic generation (SmoteR + Gaussian noise) in critical minority regions, ensuring a more equitable representation of the target variable [22].

2.2. Machine Learning Models

Three architectures were implemented in Python to predict syngas composition (H_2, CO, CO_2, CH_4): artificial neural networks (ANN), configured with Rectified Linear Unit (ReLU) and hyperbolic tangent (tanh) activation functions; support vector machines (SVM), which employ linear and nonlinear kernels to determine the optimal hyperplane that best fits the data [20]. In this study, the radial basis function (RBF) kernel was applied to measure similarity between vectors as a function of Euclidean distance; and random forests (RF), implemented through a MultiOutputRegressor optimized using RandomizedSearchCV [23]. This approach efficiently explored a space of 100–300 trees with variable depths, reducing computational cost compared to traditional grid search.

The robustness of the models was quantified using three fundamental metrics:

- **Coefficient of determination (R^2):** measures the proportion of variance in the dependent variable explained by the model. A value close to 1 indicates a perfect fit, while values near 0 reflect poor performance [24-25].

$$\boldsymbol{R^2} = 1 - \left(\frac{\sum_{i=1}^{n}(y_{observed} - y_{predicted})^2}{\sum_{i=1}^{n}(y_{observed} - y_{mean})^2}\right), \qquad (1)$$

  where $y_{observed}$ represents the actual experimental values obtained, $y_{predicted}$ refers to the values estimated by the model of ML, y_{mean} is the average of the observed experimental values, and n is the total number of data points.

- **Mean Absolute Error (MAE):** measures the average magnitude of absolute errors between predicted and observed values [26]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_{observed} - y_{predicted}), \qquad (2)$$

- **Sum of Squared Errors (SSE):** captures the total squared discrepancy between predicted and observed values, providing a measure of overall prediction error and serving as an essential component in the calculation of $R^2$and other error-based metrics [23].

$$SSE = \sum_{i=1}^{n}(y_{observed} - y_{predicted})^2, \qquad (3)$$

2.3. Thermodynamic Fundamentals and Efficiencies

The following section presents the detailed methodology applied in the energy and exergy analysis of the gasification process. In the formula (4), the exergy rate balance depends on the exergy rates of biomass ($Ex_{Biomass}$), agent of air or steam (Ex_{agent}), external heat ($Ex_{heat\ ext}$), moisture ($Ex_{moisture}$) and gas produced (Ex_{syngas}).The term Ex_{loss} represents the potential not recovered in the produced syngas, corresponding to the exergy rate of tar (Ex_{tar}), unreacted carbon (Ex_{uc}) and destruction ($Ex_{destruction}$), which are the internal irreversibilities of the gasification system, respectively [27].

$$Ex_{Biomass} + Ex_{agent} + Ex_{heat\ ext} + Ex_{moisture} = Ex_{syngas} + Ex_{loss} \qquad (4)$$

In this study the gasification system is macroscopically static, so the kinetic (Ex^{ke}) and potential (Ex^{pe}) exergies are neglected [27]. Thus, when the system is in equilibrium, its exergy consists of physical exergy (Ex_{ph}) and chemical exergy (Ex_{ch}).

$$Ex = Ex^{ph} + Ex^{ch}, \qquad (5)$$

Physical exergy reflects the ability to perform work due to pressure and temperature differences with the environment, while chemical exergy depends on the composition and imbalance of the components under standard conditions ($T_0 = 273\ K$, $P_0 = 1\ atm$).

2.3.1. Energy Ratio (ER)

Classical thermodynamic analysis is based on the first law of thermodynamics, which deals with the principle of energy conservation. Energy Ratio is quantified through the ratio between the lower heating value (LHV) in the syngas and the sum of the LHV of the dry biomass and provided heat, and this value is used for the evaluation and comparison of thermodynamic systems [28]

$$\%ER_{syngas} = \frac{n_{total\ dry\ syngas} * LHV_{dry\ syngas}}{LHV_{dry\ biomass} + Q_{Total}}, \qquad (6)$$

The Higher Heating Value (HHV) of dry biomass is calculated with equation 7, where $[C]$, $[H]$, $[O]$, and $[N]$ denote the mass fractions of carbon, hydrogen, oxygen, and nitrogen in dry biomass (%wt db) and Lower Heating Value (LHV), considering an enthalpy of vaporization of 2.4423 MJ/kg [28]. Q_{Total}, which includes the sensible heat of all reactants (biomass, air, steam and moisture) at reaction temperature:

$$HHV_{dry\ biomass}(MJ/kg) = 0.3491\ [C] + 1.1783[H] - 0.1034[O] - 0.0151[N], \quad (7)$$

$$LHV_{dry\ biomass}\ (MJ/kg) = HHV_{dry\ biomass} - 9 * [H/100] * h_{fg}, \quad (8)$$

$$LHV_{dry\ biomass}(\text{kJ/kmol}) = 1000 * LHV_{dry\ biomass}(MJ/kg) * PM_{dry\ biomass}\ (kg/kmol), \quad (9)$$

The LHV of syngas was calculated on a molar basis (kJ/kmol) from the molar fractions (x_i) of its combustible components and their respective enthalpies of combustion, listed in table 1. [29].

$$LHV_{dry\ syngas}(\text{kJ/kmol}) = \sum x_i * LHV_i\ , \quad (10)$$

To determine the syngas composition, elemental balances were applied to calculate the nitrogen and water, based on the biomass composition and the experimental values of H_2, CO, CO_2 and CH_4 in the mixture. The nitrogen quantity represents the sum of the nitrogen introduced with the air and the nitrogen inherently present in the biomass, the latter being considered inert. The water fraction, in turn, is obtained by applying the hydrogen balance to the system.

$$n_{N_2}\left(\frac{kmol\ of\ N_2}{kmol\ of\ biomass}\right) = (0.79 * \text{n}_{\text{air}}) + \frac{z}{2} \quad (11)$$

$$n_{H_2O}\left(\frac{kmol\ of\ H_2O}{kmol\ of\ biomass}\right) = \frac{x+2n_{MC}+2n_{\text{steam}}-2n_{H_2}-4n_{CH4}}{2} \quad (12)$$

2.4.2. Exergy Efficiency

Exergy analysis is providing a general measure of the (potential) usefulness of any disequilibrium (apparent or latent) in nature and identifies irreversibilities through the calculation of exergy destruction. Input exergy is composed of three main contributions: biomass exergy, gasification agent exergy (air and steam), exergy heat external and exergy moisture [28]:

$$\%\eta_{Ex\ syngas} = \left(\frac{Ex_{wet\ syngas}}{Ex_{dry\ biomass}+Ex_{steam}+Ex_{air}+Ex_{heat\ ext}+Ex_{moisture}}\right) * 100, \quad (13)$$

For biomass exergy ($Ex_{Biomass}$), the physical exergy is assumed to be zero ($Ex^{ph}_{Biomass} = 0$) since it enters under ambient conditions (25 °C, 1 atm). Thus, total biomass exergy equals its chemical exergy, calculated using the correlation factor β [29]:

$$Ex_{dry\ biomass}(kJ/kmol) = Ex^{ch}_{Biomass} = \beta *\ LHV_{dry\ biomass}, \quad (14)$$

The correlation factor β is determined as a function of the oxygen-to-carbon (O/C) ratio of the biomass. In this study, the O/C ratio falls within the range of values greater than 0.667 and less than 2.67; therefore, Equation (10) was applied [30]. In this equation, x, y, and z represent the stoichiometric coefficients of hydrogen, oxygen, and nitrogen, respectively, in the empirical formula of the dry biomass ($CH_xO_yN_z$), obtained from the elemental analysis used in the LHV calculation.

$$\beta = \frac{1.0438+0.1882\ x-0.2509(1+0.7256\ x)+0.0383\ z}{1-0.3035\ y} \quad (15)$$

Air exergy is purely chemical, since its physical exergy is null at 25ºC and 1 atm. Following the methodology of Kotas [30], dry air is considered an ideal mixture of O_2 (21%) and N_2 (79%). The standard chemical exergy of air is 128.6 kJ/kmol, according to equation (16).

$$Ex^{ch,0}_{Air} = \left(x_{N2} * Ex^{ch,0}_{N2} + x_{O2} * Ex^{ch,0}_{O2}\right) +\ RT_0\ (x_{N2} * ln(x_{N2}) + x_{O2} * ln(x_{O2})) \quad (16)$$

$$Ex^{agent}_{air}(kJ/kmol) = n_{air}\ * Ex^{ch,0}_{Air}, \quad (17)$$

Steam exergy includes both chemical and physical components. The chemical part corresponds, $Ex^{ch,0}_{H2O(g)}$ to the standard chemical exergy of water vapor, while physical exergy is evaluated considering its gaseous state from the vaporization temperature ($T_{vap} = 100°C$), and includes specific heat at constant pressure of steam $(C_{p,steam})$, system temperature (T), environmental reference temperature (T_0) and denotes the number of moles of steam (n_{steam}).

$$Ex_{steam}^{agent}(kJ/kmol) = n_{steam} * Ex_{H2O(g)}^{ch,0} + n_{steam} * Cp_{steam} * \left((T - T_{vap}) - T_0 ln\left(\frac{T}{T_{vap}}\right)\right), \quad (18)$$

The exergy of biomass moisture is calculated following standard thermodynamic methodology; in Equation (19), %wt MC represents the mass fraction of moisture content in biomass, %wt VM the mass fraction of volatile matter, $PM_{H2O}$the molecular weight of water, and $PM_{CH_xO_yN_z}$the molecular weight of dry biomass and $Ex_{H2O(l)}^{ch,0}$ the standard chemical exergy of liquid water.

$$Ex_{moisture}(kJ/kmol) = \left(\frac{\%_{wt,MC}/PM(H_2O)}{\%_{wt,VM}/PM(CH_xO_yN_z)}\right) * Ex_{H2O(l)}^{ch,0}, \quad (19)$$

External heat exergy (Ex_{heat_ext}) is calculated from the sensible heat required to raise the temperature of biomass, air, steam and moisture to reaction temperature, multiplied by the Carnot quality factor assuming that the process temperature is equal to the average temperature for each heat rate:

$$Ex_{heat\ ext}(kJ/kmol) = Q_{biomass} * \left(1 - \frac{T_0}{(T+T_0)/2}\right) + Q_{air} * \left(1 - \frac{T_0}{(T+T_0)/2}\right) + Q_{steam} * \left(1 - \frac{T_0}{(T+T_{vap})/2}\right) + Q_{moisture*}\left(1 - \frac{T_0}{(T+T_0)/2}\right), \quad (20)$$

The Syngas exergy (Ex_{syngas}) is obtained as the sum of its physical and chemical contributions. The chemical part is calculated from the standard chemical exergy values of each component ($Ex_i^{ch,0}$) [28,30], while the physical part is derived from the average molar heat capacities (Cp_i) using the polynomial correlation of the equation 15 [29,31] expressed as:

$$Ex_{syngas} = \sum_{i=1}^{n} x_i Ex_i^{ch,0} + \sum_{i=1}^{n} x_i Cp_i\,(T)\left((T - T_0) - T_0 ln\left(\frac{T}{T_0}\right)\right) + RT_0 ln\left(\frac{P}{P_0}\right), \quad (21)$$

$$Cp_i = R\left(A + BT + CT^2 + DT^3 + \frac{E}{T^2}\right), \quad (22)$$

With constants listed in Table 1, which also includes standard chemical exergies and lower heating values at reference conditions ($T_0 = 298.15\,K$, $P_0 = 1\,atm$, $R = 8.314\,kJ/(kmol \cdot K)$). Since the reactor pressure is assumed equal to ambient pressure ($P = P_0$), the pressure-dependent entropy term cancels out.

Table 1 *Specific heat capacity constants, standard chemical exergy, and lower heating values*

Chemical Species	A	B	C	D	E	$T_{maximun}$ (K)	$Ex_i^{ch,0}$ (Kj/kmol)	Lower heating value (kJ/kmol × 10^6)
CH_4	1.702	9.08E-03	-2.16E-06	-	-	1500	836510	0.8026210
H_2	3.249	4.22E-04	-	-	8.30E-02	3000	238490	0.24182
CO	3.376	5.57E-04	-	-	-3.10E-02	2500	275430	0.283
CO_2	5.457	1.05E-03	-	-	-1.157E+00	2000	20140	-
N_2	3.280	5.93E-04	-	-	4.00E-02	2000	720	-
H_2O (g)	3.470	1.45E+00	-	-	1.21E-01	2000	11710	-
H_2O (l)	8.712	1.25E-03	-1.80E-07	-	-	100°C	3120	-
O_2	3.639	5.06E-04		-	-2.27E-01	1500	3970	-
Air	-	-	-	-	-		128.6	-

3. Results and Discussion

3.1. Model performance in syngas component prediction

The best-performing models in syngas component prediction are presented, with numerical results detailed in Table 2. The validation metrics provide a comparative summary of model performance across all prediction targets. Hydrogen exhibited the most outstanding performance, particularly with the Random Forest (RF) Original model, achieving a test R^2 of 0.992 and a validation of 0.980. This result surpasses the values reported by Pimentel et al. [21], who obtained a test $R^2$of 0.912 using ANN, and is higher than the $R^2$of 0.94 reported by Sakheta et al. [5] with XGBoost. The high accuracy in predicting H_2 is consistent with the literature, which attributes this reliability to the strong thermodynamic correlation of hydrogen with temperature and the steam-to-biomass ratio [32]. Notably, the application of SMOGN to ANN increased its test $R^2$from 0.788 to 0.913, this improvement suggests that the model successfully overcame the bias toward the most frequent biomass compositions in the training set, allowing the architecture to capture the non-linear gradients of H2 production even in extreme operational conditions that were originally underrepresented. Only high-complexity models or those trained with low-noise simulated datasets, such as Vargas [33] ($R^2 = 0.993$) or Zhao et al. [12] ($R^2 = 0.9782$ in Supercritical Water Gasification (SCWG), exhibit levels of accuracy comparable to those presented in this study.

Table 2. *Evaluation of ANN, SVR and RF of metrics in Train/Test/Validation.*

Model			R^2 Train	R^2 Test	R^2 Validation	MAE Train	MAE Test	MAE Validation	SSE Train	SSE Test	SSE Validation
CO	ANN	Original	0.926	0.507	0.548	1.421	3.602	3.255	2243.960	4428.770	3672.440
		SMOGN	0.858	0.385	0.043	1.043	2.698	3.268	2280.920	1530.470	3733.080
	SVM	Original	0.681	0.537	0.349	1.770	2.554	2.574	3421.896	1313.158	2839.317
		SMONG	0.675	0.394	0.208	1.810	2.868	3.113	3486.984	1717.125	3456.867
	RF	Original	0.981	0.886	0.805	0.456	1.108	1.390	213.025	254.694	786.400
		SMOGN	0.862	0.642	0.448	1.168	2.266	2.545	2278.277	1014.425	2407.598
CO_2	ANN	Original	0.958	0.769	0.715	1.239	2.537	2.868	1688.130	2093.500	3044.110
		SMOGN	0.946	0.687	0.769	0.818	2.048	1.847	1265.950	1392.000	966.785
	SVM	Original	0.907	0.753	0.824	1.439	2.189	1.998	2152.012	1251.071	826.343
		SMOGN	0.902	0.712	0.766	1.497	2.364	2.219	2275.187	1462.822	1102.187
	RF	Original	0.993	0.965	0.944	0.351	0.908	1.086	126.388	179.376	312.897
		SMOGN	0.958	0.840	0.847	0.809	1.882	1.659	1009.915	809.479	718.480
H_2	ANN	Original	0.967	0.788	0.859	1.800	4.064	3.225	3612.74	5916.780	3930.840
		SMOGN	0.981	0.913	0.807	1.159	2.646	3.179	2020.82	2244.260	4888.230
	SVM	Original	0.958	0.915	0.832	1.774	2.893	3.565	4173.448	2524.410	4773.516
		SMOGN	0.958	0.921	0.813	1.779	2.813	3.730	4153.119	2335.396	5331.288
	RF	Original	0.998	0.992	0.980	0.396	0.842	1.163	171.566	177.020	517.122
		SMOGN	0.981	0.928	0.876	1.118	2.534	3.064	2043.068	2122.397	3518.334
CH_4	ANN	Original	0.831	0.629	0.808	0.954	1.458	0.957	1105.850	514.127	261.220
		SMONG	0.943	0.461	0.565	0.536	1.305	1.183	401.913	565.255	492.466
	SVM	Original	0.911	0.733	0.758	0.872	1.240	1.224	568.566	319.445	306.006
		SMOGN	0.909	0.682	0.602	0.894	1.281	1.459	578.791	379.669	504.614
	RF	Original	0.991	0.934	0.969	0.198	0.537	0.373	46.185	97.472	41.654
		SMOGN	0.963	0.625	0.750	0.400	1.095	0.975	266.002	448.368	316.468

Carbon monoxide (CO) systematically represented the greatest predictive challenge, a trend widely documented in the literature [5]. Nevertheless, our RF Original model achieved a test $R^2$of 0.886, significantly surpassing the 0.700 obtained by Pimentel et al. [21], and showing practically the same performance as the 0.88 reported by Sakheta et al. [5]. The inherent difficulty with CO lies in the fact that its linear fits often deviate from the ideal line $Y = T$, due to the complexity of partial oxidation reactions. The poor performance of ANN with SMOGN in validation ($R^2 = 0.043$) suggests that, for this component, the oversampling algorithm may have generated synthetic cases outside physical reality, exacerbating the latent imbalance in input variables such as carbon and oxygen [21].

For CO_2, the RF Original model demonstrated superior stability, achieving a test $R^2$of 0.965 and a validation of 0.944. These values are highly competitive compared to the $R^2$of 0.813 reported by Pimentel et al. [21], and the 0.94 obtained by Ascher et al. [34]. The robustness of tree-based models (such as RF) over ANN for this component can be explained by their ability to handle moderately sized datasets without incurring in overfitting, which often affects neural network architectures when data density is limited. [12]

The prediction of methane (CH_4) content in syngas represented one of the main achievements of this study. The Random Forest (RF) Original model reached a test $R^2$of 0.934 and validation of 0.969, demonstrating superior predictive capacity. These values significantly surpass those reported by Pimentel et al. [21], ($R^2 = 0.858$ with ANN), Sakheta et al. [5] ($R^2 = 0.84$ with XGBoost), and Ascher et al. [34]. ($R^2 = 0.88$ with RF). Gil et al. [32] reported an R^2 equal to 0.978 for this component, our result is similar and evidences robustness against biomass variability. In contrast, ANN exhibited marked instability, with a drop in the R^2 for the test to 0.629 in its original version and even lower values when applying data augmentation with SMOGN ($R^2 = 0.461$). This behavior is consistent with the findings of Vargas [33], who noted that methane may present greater convergence challenges and error fluctuations during validation due to its typically low concentration range and sensitivity to experimental inaccuracies.

As a consequence, the RF model demonstrated superior generalization, consistently maintaining the highest R^2values and the lowest errors (MAE and SSE) across training, testing, and validation phases for all components. This performance is highly competitive with benchmarks cited in the literature; for example, it aligns with the R^2 higher than 0.97 reported by Puig Arnavat et al [6], for syngas composition and the R^2 higher than 0.90 achieved by George et al.[8]. While Safarian et al. [9], reported slightly higher values (R^2 higher than > 0.99) for power output, our RF model achieves comparable precision for multi-target gas yields.

The technical advantage of RF lies in its structural simplicity and ease of implementation, as it averages multiple independent decision trees, effectively filtering experimental noise and reducing variance without requiring extensive hyperparameter tuning or large-scale data density through the bagging method (bootstrap aggregating) [34]. Unlike ANN, which demand a rigorous trial-and-error process to define the optimal number of hidden layers, neurons, and activation functions in order to avoid the “black box” problem [33]. In this study, the use of SMOGN was essential to improve ANN performance by mitigating training biases. On the other hand SVM, which showed intermediate performance, this model depends heavily on the selection of complex kernel functions and regularization parameters, since its ability to capture syngas variability is often limited unless a highly specific kernel is optimized for the dataset.

3.2. Evaluation of Energy Ratio and Exergy Efficiency (Thermodynamic Validation)

Model performance for energy ratio and exergy efficiencies is detailed in Table 3, validating the high accuracy achieved by the RF architecture in training, testing, and external validation.

Table 3. *Performance Metrics of RF Model for Energy Ratio and Exergy Efficiency for Train, Test and Validation, respectively*

RF	R^2 Train	R^2 Test	R^2 Validation	MAE Train	MAE Test	MAE Validation	SSE Train	SSE Test	SSE Validation
% Energy Ratio	0.992	0.963	0.959	0.994	2.643	2.285	1327.095	1714.668	1654.784
% Exergy efficiency	0.998	0.993	0.987	0.229	0.481	0.619	63.585	59.801	116.308

The predictive performance of the RF model reached a coefficient of determination (R^2) of 0.993 in the test set and 0.987 in the validation for exergy efficiency (Table 3). In turn, the energy ratio showed an R^2 of 0.963 in the test set. These results are fully consistent with the findings reported in [18, 35], where Artificial Neural Network (ANN) architectures were employed to characterize gasification processes in Brazilian biomasses. In Vargas et al. [18], focused on predicting chemical exergy and syngas yield, ANN models systematically achieved R^2 values above 0.980 for all evaluated cases. Specifically, they reported a training R^2 of 0.998 for chemical exergy conversion and an exceptional validation fit for hydrogen (H_2), with an R^2 of 0.991. When comparing these values with our RF model (validation equal to 0.987), a concordance in the order of magnitude of accuracy is observed, validating the use of ensemble algorithms as a competitive alternative to deep neural networks for mapping complex thermochemical properties. Likewise, in a previous work from the same authors [35] oriented to electricity generation from residues, these authors obtained outstanding accuracies both in training and testing, reporting an R^2 greater than 0.993 for energy conversion prediction. The accuracy of 0.963 for the energy ratio, although slightly lower in absolute value, remains within the high-fidelity standards defined in those works for models applied to more specific Brazilian bioenergy systems.

The table 4 and the figures 3 and 4 compares the predicted and experimental values and the distribution of residuals for the energy ratio and exergy efficiency, respectively.

Table 4. *Statistical summary of the global performance of the RF model for energy ration and exergy efficiency*

	R^2	MAE	SSE	Mean Residual (Bias)	Max Positive Deviation	Max Negative Deviation
% Energy Ratio	0.978	1.5050	4696.547	-0.027	22.086	-14.335
% Exergy efficiency	0.994	0.344	239.694	0.013	.5.756	-4.032

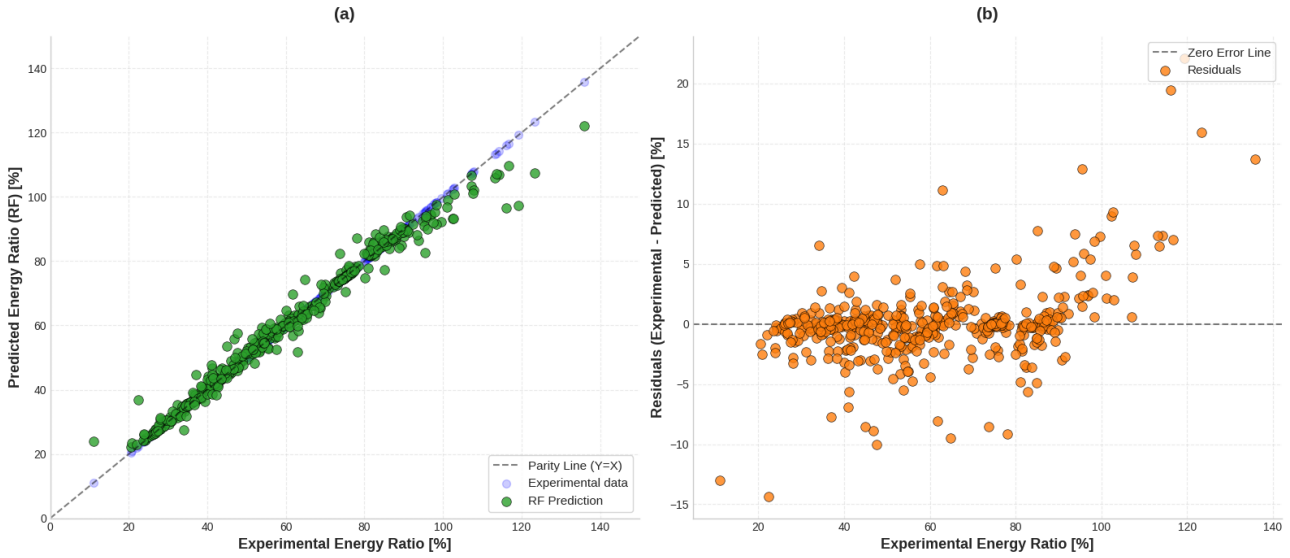

Figure. 3. *(a) Predicted vs. experimental values (b) Distribution of residuals for the energy ratio*

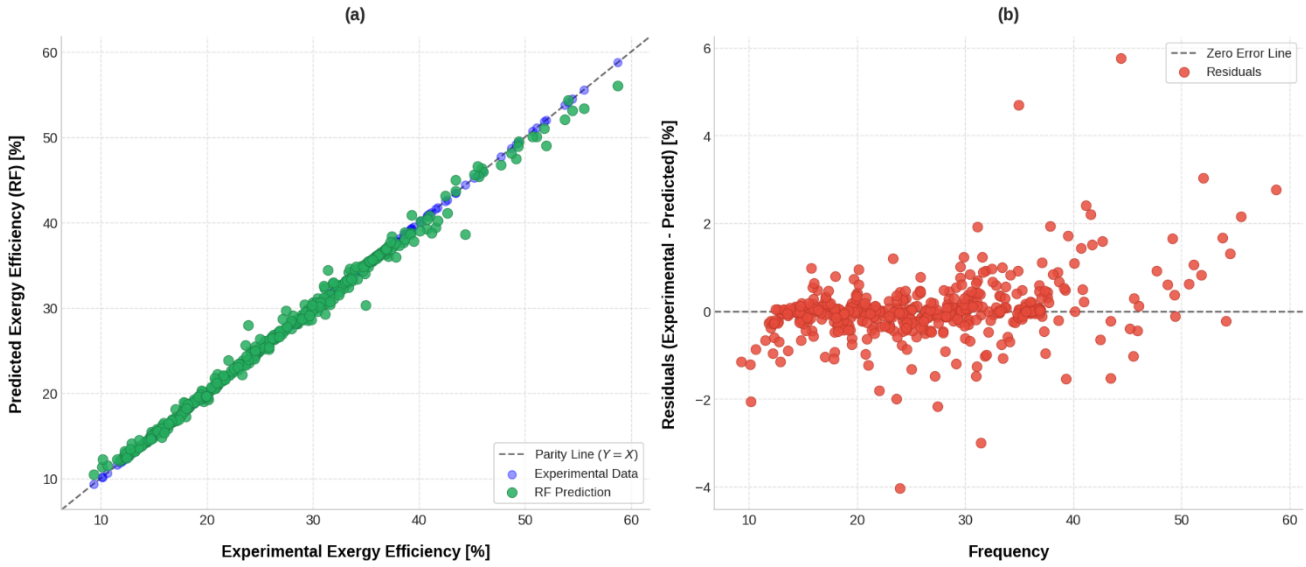

Figure. 4. *(a) Predicted vs. experimental values (b) Distribution of residuals for the exergy efficiency*

According to figure 3 (a) and Table 3, the model prediction (60.81%) almost exactly replicated the experimental mean (58.62%). This value falls within an acceptable range; it is higher than the 37.66% reported in [19], but lower than the optimized scenarios in [33], which reached 82.21%. A critical finding is that the RF model reduced the standard deviation from 21.62% to 20.63%. Although the Random Forest model achieved high overall accuracy ($R^2 = 0.978$), the residual analysis figure 3 (b) reveals larger deviations in regions of higher exergy efficiency (>100%). This indicates that, while the model adequately captures the traditional operating range (residue lower than 5%), its predictive capacity decreases at extreme values, where residuals tend to be higher.

On the other hand, in predicting exergy efficiency, the RF model showed outstanding performance, with very high accuracy ($R^2 = 0.994$) and minimal errors (MAE = 0.344). Figure 4(a) demonstrates the close correspondence between experimental values (average 25.94%) and predicted values (average 27.15%), a result highly consistent with the findings of [33], which reported exergy efficiencies ranging from 24.42% to 42.57%. The stability of the predicted limits (minimum of 10.46% and maximum of 55.97%) confirms that the algorithm respects the boundaries imposed by the destruction of chemical and physical exergy due to irreversibilities. Finally, the residual analysis (Figure 4 b) shows a balanced distribution around zero, with slightly higher deviations (more than 2%) in the region of higher efficiency.

4. Conclusions

This study demonstrates that Machine Learning, led by the Random Forest model, surpasses the limitations of deterministic approaches in predicting biomass gasification. With outstanding performance in estimating composition syngas, the algorithm effectively captured the complexity of the process. In parallel, the Support Vector Machine (SVM) exhibited greater technical competitiveness compared to Artificial Neural Networks (ANN), handling nonlinear relationships with enhanced robustness and reduced risk of overfitting in moderately sized datasets. Thermodynamic validation corroborated the physical consistency of the system, with the predicted energy ratio (R^2 = 0.978, MAE = 1.505, SSE = 4696.547) and exergy efficiency (R^2 = 0.994, MAE = 0.344, SSE = 239.694) closely aligned with experimental values.

These results provide a solid foundation for future research, considering heterogeneous biomass, different types of gasifiers, co-gasification blends, and additional factors such as catalyst type, reactor geometry, and residence time, with the aim of increasing the overall robustness of the model. On the algorithmic front, the adoption of hybrid models, such as Physics-Informed Neural Network (PINN), incorporating thermodynamic constraints and oriented toward multi-objective optimization (gas quality, exergy, and energy) can expand predictive capacity and prevent physical inconsistencies.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. The authors would like to thank the CNPq/MCTIC for the financial support to the Department of Mechanical Engineering (DEM) and the Department of Chemical and Materials Engineering (DEQM) at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), the CNPq for the Bolsa de Produtividade em Pesquisa (PQ; grant number: 308627/2025-3) and the FAPERJ for Jovem Cientista do Nosso Estado (JCNE; grant number: E-26/204.488/2025) grants awarded to Florian Pradelle.

Nomenclature

Letter symbols

$\boldsymbol{ANN}$ Artificial Neural Network

$\boldsymbol{Cp}$ heat capacity at constant pressure, kJ/(kmol.K)

Ex specific exergy, kJ/kmol

ER energy ratio, %

FC fixed carbon, %

HHV higher heating value, MJ/kg

LHV lower heating value, MJ/kg

$\boldsymbol{MAE}$ mean absolute error,

$\boldsymbol{MC}$ moisture content, %

$\boldsymbol{n}$ molar flow rate, kmol/kmol biomass

Q sensible heat, kJ/kmol

$\boldsymbol{R}$ universal gas constant, kJ/(kmol.K)

R^2 coefficient of determination

$\boldsymbol{RBF}$ Radial Basis Function

$ReLU$ Radial Basis Function

$\boldsymbol{RF}$ Ramdon Forest

$\boldsymbol{SSE}$ sum of squared errors,

$\boldsymbol{S/B}$ steam to biomass ratio,

$\boldsymbol{SVM}$ Support Vector Machines

T temperature, K

T_0 reference temperature, K

VM volatile matter, %

x_i molar fraction of component, i

Greek symbols

η efficiency exergy, %

β correlation factor for biomass exergy

Subscripts and superscripts

ch chemical contribution

ph physical contribution

0 reference state (T_0 = 298.15 K, P_0 = 1 atm)

References

[1] Qianshi, S., Wei, Z., Xiaowei, W., Xiaohan, W., Haowen, L., Zixin, Y., Yue, Y. and Guangqian, L., 2023. "Comprehensive effects of different inorganic elements on initial biomass char-CO_2 gasification reactivity in micro fluidised bed reactor: theoretical modeling and experiment analysis". Energy, Vol. 262. https://doi.org/10.1016/j.energy.2022.125379

[2] Ozbas EE, et al. Hydrogen production via biomass gasification, and modeling by supervised machine learning algorithms. Int J Hydrog Energy. 2019;44(33):17260-8. https://doi.org/10.1016/j.ijhydene.2019.02.108

[3] Chigozie E, Nwankpa W, Ijomah A, Gachagan S, Marshall S. Activation functions: comparison of trends in practice and research for deep learning. Preprint. 2018.

[4] Azzone E, Morini M, Pinelli M. Development of an equilibrium model for the simulation of thermochemical gasification and application to agricultural residues. Renew Energy. 2012; 46:248-54. https://doi.org/10.1016/j.renene.2012.03.017

[5] Sakheta, A., Ramirez, J., Raj, T., Nayak, R., & O'Hara, I. Improved prediction of biomass gasification models through machine learning. Computers & Chemical Engineering, 2024, 108834. https://doi.org/10.1016/j.compchemeng.2024.108834

[6] Puig-Arnavat M, Hernández JA, Bruno JC, Coronas A. Artificial neural network models for biomass gasification in fluidized bed gasifiers. Biomass Bioenergy. 2013;49:279-89. https://doi.org/10.1016/j.biombioe.2012.12.012

[7] Domingos P. A few useful things to know about machine learning. Commun ACM. 2012; 55(10):78-87. https://doi.org/10.1145/2347736.2347755

[8] George J, Arun P, Muraleedharan C. Assessment of producer gas composition in air gasification of biomass using artificial neural network model. Int J Hydrog Energy. 2018;43(20):9558-68. https://doi.org/10.1016/j.ijhydene.2018.04.007

[9] Safarian S, et al. Artificial neural network integrated with thermodynamic equilibrium modeling of downdraft biomass gasification-power production plant. Energy. 2020;213:118800. https://doi.org/10.1016/j.energy.2020.118800

[10] Pandey DS, Das S, Pan I, Leahy JJ, Kwapinski W. Artificial neural network based modelling approach for municipal solid waste gasification in a fluidized bed reactor. Waste Management 2016;58:202–213.

[11] Mutlu AY, Yucel O. An artificial intelligence-based approach to predicting syngas composition for downdraft biomass gasification. Energy. 2018; 165:895-901. https://doi.org/10.1016/j.energy.2018.09.131

[12] Zhao S, Li J, Chen C, Yan B, Tao J, Chen G. Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. J Clean Prod. 2021; 316:128244. https://doi.org/10.1016/j.jclepro.2021.128244

[13] Kardani N, et al. Modelling of municipal solid waste gasification using an optimised ensemble soft computing model. Fuel. 2021; 289:119903. https://doi.org/10.1016/j.fuel.2020.119903

[14] Cheng F, Luo H, Colosi LM. Slow pyrolysis as a platform for negative emissions technology: An integration of machine learning models, life cycle assessment, and economic analysis. Energy Convers Manag. 2020; 223:113258. https://doi.org/10.1016/j.enconman.2020.113258

[15] Chen X, et al. Prediction of product distribution and bio-oil heating value of biomass fast pyrolysis. Chem Eng Process. 2018; 130:36-42. https://doi.org/10.1016/j.cep.2018.05.018

[16] Aghbashlo M, Rosen MA. Exergoeconoenvironmental analysis as a new concept for developing thermodynamically, economically, and environmentally sound energy conversion systems. J Clean Prod. 2018; 187:190-204

[17] Mahian O, Mirzaie MR, Kasaeian A, Mousavi SH. Exergy analysis in combined heat and power systems: A review. Energy Convers Manag. 2020; 226:113467. https://doi.org/10.1016/j.enconman.2020.113467

[18] Vargas, G. G.; Oliveira Jr, S. Approach to predict chemical exergy and syngas yield in Brazilian biomass waste gasification using an artificial neural network. In: Proceedings of the 37th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems (ECOS 2024), Rhodes, Greece, June 30–July 5, 2024. DOI: 10.52202/077185-0173.

[19] Lewin, C. S. Modelagem, simulação e otimização de um gaseificador de resíduos sólidos em operação cocorrente. Master's Thesis, Programa de Pós-Graduação em Engenharia Mecânica, Centro Técnico Científico da PUC-Rio, Rio de Janeiro, 2020.

[20] Shahbeig H, Shafizadeh A, Rosen MA, Sels BF. Exergy sustainability analysis of biomass gasification: a critical review. Renew Sustain Energy Rev. 2023

[21] Pimentel FS, Santos BF, Pradelle F. Investigation of artificial neural network topologies to predict biomass gasification and comparison with a thermodynamic equilibrium model. Energy. 2024; 308:132762. https://doi.org/10.1016/j.energy.2024.132762

[22] Branco PO, Torgó L, Ribeiro RP. SMOGN: a pre-processing approach for imbalanced regression. In: Learning with Imbalanced Domains: Theory and Applications. 2017. p.36-50

[23] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(Feb):281-305.

[24] Umenweke GC, et al. Machine learning methods for modeling conventional and hydrothermal gasification of waste biomass: A review. Bioresour Technol Rep. 2022; 17:100976. https://doi.org/10.1016/j.biteb.2022.100976

[25] Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 5th ed. Hoboken (NJ): Wiley; 2012.

[26] Willmott C, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005;30(1):79-82.

[27] Chen W, Deng Z, Xiang X, Wang B, Liu S, Wang H, Guo J, Li T, He C, Zhou X, Peng D. An integrated exergy efficiency and machine learning method for optimizing organic solid waste gasification process. Eng Appl Artif Intell. 2025;126:111805. Doi.org/10.1016/j.engappai.2025.111805

[28] Fiuza RVG. Modelagem termodinâmica da gaseificação de biomassa: otimização das condições operacionais na co-gaseificação de resíduos sólidos urbanos [dissertation]. Rio de Janeiro (Brazil): Pontifícia Universidade Católica do Rio de Janeiro; 2022.

[29] Perry RH, Green DW. Perry's Chemical Engineers' Handbook. 8th ed. Columbus (OH): McGraw-Hill; 2008.

[30] Kotas TJ. The Exergy Method of Thermal Plant Analysis. Malabar (FL): Krieger Publishing Company; 1995.

[31] Zhang Y, Zhao Y, Gao X, Li B, Huang J. Energy and exergy analyses of syngas produced from rice husk gasification in an entrained flow reactor. J Clean Prod. 2015;95:273-80.

[32] Gil, M. V., Jablonka, K. M., Smit, B., Garcia, S., & Pevida, C. *Biomass to energy: a machine learning model for optimum gasification pathways*. *Digital Discovery*, 2 (2023), Article D3DD00079F. https://doi.org/10.1039/d3dd00079f

[33] Vargas, G. G. *Integrated assessment of residual biomass gasification for hydrogen, ammonia, and syngas production: exergy analysis, neural networks, and environmental impact*. Doctoral Thesis, Polytechnic School of the University of São Paulo, Mechanical Engineering – Energy and Fluids, São Paulo, 2024

[34] Ascher, S., Wang, X., Watson, I., Sloan, W., & You, S. Interpretable machine learning to model biomass and waste gasification. Bioresource Technology, 364 (2022), 128062. https://doi.org/10.1016/j.biortech.2022.128062

[35] Vargas, G. G.; Oliveira Jr, S. Exergy assessment of electricity generation via biomass gasification by neural network algorithm. In: Proceedings of the 36th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems (ECOS 2023), Las Palmas de Gran Canaria, Spain, June 25–30, 2023. DOI: 10.52202/069564-0150.