Multivariate analysis and soft computing-based prediction of energy potential in heterogenous waste streams

Oluwatobi Adeleke[1*], *Tien-Chien Jen*[1*]

[1]Mechanical Engineering Sciences, University of Johannesburg, Johannesburg, South Africa.

Abstract. This study presents a data-driven framework for characterizing waste-derived biomass for energy recovery. Utilizing a dataset comprising higher heating value (HHV), elemental composition, and proximate properties of diverse waste streams, correlation analysis and feature importance analysis (FIA) using Random Forest (RF)'s importance metrics were conducted to identify key parameters influencing HHV prediction. Carbon and and Hydrogen were identified as the most significant contributors, accounting for 75–80% of the model's predictive strength. Principal Component Analysis (PCA) was applied to cluster waste types based on compositional and energetic similarities, aiding in the classification of waste for optimized waste-to-energy (WtE) strategies. Dimensionality was effectively reduced with over 90-95% of variance captured in the first four principal components. The predictive performance of three machine learning models—Artificial Neural Network (ANN), Support Vector Machine (SVM). The RF model demonstrated superior performance during training with RMSE, MAE, MAD, and rMBE values of 0.8606, 0.5945, 0.3864, and 0.0895, respectively. This integration of statistical techniques and machine learning provides a robust tool for waste classification and HHV estimation, promoting data-informed decisions in sustainable waste management and energy.

1 Introduction

The overexploitation and usage of fossil resources result in significant harm to the environment and human health [1]. However, the growing demand for energy and the attendant environmental consequences of fossil fuel consumption has drawn significant attention to renewable energy resources [2]. Among these are waste derived biomass such as sludge, agricultural waste, with huge potential as alternative energy sources for the future [3]. Recovery of energy from waste is a suitable remedy to the problem of handling the enormous volume of waste produced since it keeps a significant amount of waste out of landfills which is believed to be among the biggest anthropogenic activity in the world [4]. A key element in ensuring the circular economy is the waste to energy, Waste has an energetic value of 0.87 GJ/tonne, which can be harnessed worldwide, and 1.25 GJ/tonne is expected by 2021 [5]. The estimated global energy consumption was 520 quadrillion BTUs in 2010, and by 2040, a rise of 56% of the current level is expected [6]. However, the efficiency of energy recovery from waste is significantly influenced by heterogenous nature of the waste rendering optimization a challenging challenge. Moreover, the laboratory bomb calorimeters approach for estimating waste biomass heating value (HV) is costly and time-consuming due to the pretreatment requirement [7]. Hence, a simple, less expensive approaches are necessitated. The advent of artificial intelligence has caused a paradigm shift from classical to intelligent prediction techniques. Consequently, data-driven systems have demonstrated efficacy in intelligent and real-time decision-making in waste-to-energy (WTE) system.

A variety of heating value predictive model based on proximate and ultimate analyses have been created, encompassing both white box (with explicit equations) and black box models [3]. Previous research has demonstrated the superior capacity of HV model based on ultimate analysis than those based on proximate analysis, particularly for the white box models, such as the linear regression (LR) model [8]. Mateus et al. [9] developed a LR model based on ultimate analysis for predicting the HV of fuel obtained from waste. A similar study by Amen et al. (2021) used ultimate and proximate properties to create a LR equation for the combustion enthalpy of waste-derived biomass for WTE assessment [10]. Using soft computing techniques like ANN, ANFIS, SVM, and MLP, the study by Mondal and Rafizul [11] assessed the prediction of the CV of MSW. With the greatest R^2 (0.9979) and lowest error criteria, ANFIS showed remarkable prediction accuracy above other models. Adeleke et al. [12] developed ANN and ANFIS models to predict the HV of waste. The results confirm ANFIS with grid partitioning (GP)'s accuracy with in localized waste HV prediction. Using four nonlinear models, You et al. [13] suggested a data-driven, expert-informed approach for real-time prediction of LHV in waste incineration. Among these, the ANFIS surpassed others in accuracy. The method allows automated and effective control of circulating fluidized bed incinerators.

* Corresponding author: thobyadeleke@gmail.com, tjen@uj.ac.za

This study investigates a combined data-driven approach to predict the energy potential of heterogeneous waste derived biomass. Understanding the complexity of waste's varied composition, the study develops a nexus of statistical analysis and machine learning technologies for energy recovery purposes. Specifically, correlation analysis was utilized to study interdependencies across physicochemical characteristics, Principal Component Analysis (PCA) was used to lower data dimensionality and reveal latent patterns in waste composition. Random Forest (RF) based Feature Importance Analysis (FIA) helped to find key variables influencing heating value forecasts. Machine-learning based predictive modeling employing ANN, SVM, and RF to forecast the Higher Heating Value (HHV) of waste streams further complemented these statistical methods. This multilayer integration of statistical and soft computing methodologies improves the dependability, efficiency, and scalability of waste-to-energy (WtE) assessments. The goal is to provide a strong analytical framework that enables strategic energy planning and waste categorization, hence maximizing the use of MSW as a renewable energy source.

2 Materials and methods

2.1 Data Collection and pre-processing

The data set for this study was gathered from extensive waste and biomass databases such as Phyllis2 [14] and Meraz et al [15] The data samples include a vast category of waste ranging from General waste, Paper and paper products, food waste, refuse-derived fuel (RDF) amongst others. The statistical summary of the waste variables is shown in Table 1.

Table 1. Statistical summary of the waste biomass data

Properties	Max	Min	*Mean*	*St. D*
H2O (%)	78.7	0	14.63	22.05
C (%)	87.1	0.5	42.75	18.96
H (%)	14.18	0.08	5.55	2.66
O (%)	47.84	0	25.75	15.68
N (%)	10	0	1.16	1.761
S (%)	4.08	0	0.29	0.47
Ash (%)	98.9	0	24.54	30.47
HHV (MJ/kg)	45.88	0.14	15.53	9.75

2.2 Statistical tool and machine learning framework

2.2.1 Correlation analysis

The linear interrelationship amongst the waste biomass variables were analyzed using a Pearson correlation matrix and visualized using the correlation heat-map. This expresses the potential co-linearity amongst the variable as well as the positive and negative correlation of the input parameters to the heating value.

2.2.2 Principal component Analysis

Principal component analysis (PCA) is a statistical approach used for reduction of the dimensionality in dataset while retaining the variance withing the data. This approach represents the original-matrix through an array of new uncorrelated variables known as principal components (PC) which retains most variance in the waste dataset. PCA was used in this study to retain critical variance (most significant information) in the waste biomass dataset.

2.2.3 Artificial neural network

Artificial neural network (ANN) is a framework for addressing intricate pattern-oriented challenges in both categorization and time-series research. The feed-forward neural network (FFNN) is a specific type of ANN where input layer data is transmitted directly to the output layer without any feedback. Neurons in ANN have three layers: the input layer, the hidden layer(s), and the output layer. The input layer receives inputs $x_j: (j = 1,2 \dots n)$, the hidden layer(s) consist of neurons $n_j: (j = 1,2 \dots n)$, and the output layer produces outputs $o_j: (j = 1,2 \dots n)$. They represent neuron output in the first hidden layer of a two-hidden neural network. First hidden layer has m_1 neurons, second has m_2 neurons. Weights linking the first hidden layer to the input layer are labelled ${w_{il}}^1$ and those connecting the second hidden layer to the first are labelled ${w_{ij}}^2$ and expressed in equations 1 and 2. Activation function for neurons in the first hidden layer is ϕ_i, and for neurons in the second layer, ψ_j [16], [17].

$$\xi_i = \psi_j\left(\sum_{i=0}^{p} {w_{il}}^1 U_l\right), \quad u_0 \; and \; \psi_j(\cdot) = 1 \qquad (1)$$

$$y = \sum_{i=0}^{m_2} w_1 \phi_i\left(\sum_{j=0}^{m_1} {w_{ij}}^2 \xi_j\right), \quad \phi_0\,(\cdot) = 1 \qquad (2)$$

2.2.4 Support vector machine

SVM is a supervised learning technique employed for classification and regression tasks. The primary advantage of SVM is its simplicity, computational efficiency, and capability to be trained with a little number of samples. Nonetheless, identifying the ideal kernel and its parameters presents the greatest obstacle. The basic idea of SVM is to maximize the geometric margin between two datasets while simultaneously minimizing the empirical classification error [18]. SVMs seek to identify the optimal hyperplane that delineates the data points of distinct classes. In high-dimensional feature spaces, this boundary is referred to as a hyperplane. The objective is to optimize the margin, defined as the distance between the hyperplane and the nearest data points of each category, hence facilitating the differentiation of data classes [19].

2.2.5 Random Forest

Random forest (RF) is an ensemble learning technique which finds application in classification and regression problems operating by generating numerous decision trees during the training process. The outcome of the RF is the class chosen by the majority of trees during classification, while in regression tasks, the result is the mean of the predictions from the trees. It includes a tree

structure $\{(h(x,(k)k = 1,2 \ldots \ldots)\}$, a distinctive independent vector $\{\theta(k)\}$, and input for the most influential x class [20]. RF constructs numerous decision trees that are aggregated to enhance predictive capability. For regression task, the ensemble computes the mean of the outputs from all individual trees. The RF was also utilized for the feature importance assessment using the Gini importance (GI) metrics. The GI quantifies the proportional contribution of each feature to the model's predictions.

2.2.6 Model hyper-parameter settings

Table 2 defines some of the key parameters of the four models developed for the prediction of the corrosion rate based on relevant predictors

Table 2. hyper-parameters of the developed model

Models	Hyper-parameters	Values
ANN (7-10-1)	Hidden Layer neurons	10
	Activation Function	ReLU
	Training Algorithm	LM
	Max Iterations	500
	Random State	42
SVM	Kernel	RBF
	c	1.0
	Epsilon	0.1
	Gamma	0.1
Random forest	Number of Estimators	100
	Criterion	Squared Error
	Bootstrap	True

3 Results and discussion

3.1 Correlation assessment of waste properties and the heating value

To enhance the combustion properties and efficiency of waste-derived biomass, it is essential to comprehend the relationship between the waste features and its heating value. The Pearson correlation heap in figure 1 shows the correlation between various physico-chemical properties of waste biomass, including elemental composition, moisture, ash, and Higher Heating Value (HHV). A very strong positive correlation (0.84) was observed between the C (%) and HHV. Since carbon is the primary driver of combustion energy, more carbon content raises the HHV of biomass. Because of the role of hydrogen in calorific value through formation of water vapor and heat during combustion, it is observed to be positively correlated (0.81) with HHV. By a value of -0.58, ash (%) strongly negatively correlates with HHV. Ash is inert and does not contribute to combustion. High ash content reduces the HHV by diluting flammable material. MC lowers HHV as energy is used evaporating water prior to ignite. A correlation value of -0.40 explains its moderate negative correlation. High MC biomass contains less useable energy. The greatest of all the variables, a very strong positive correlation (0.93), was found between Carbon and the HHV. Similar organic sources like cellulose, lignin in biomass most likely explains this. On the other hand, Carbon and ash have an inverse correlation of -0.87. Biomass high in organic material (carbon) often has less inorganic residue (ash), hence A comparable pattern was seen between hydrogen and ash since hydrogen-rich materials are usually less mineralized.

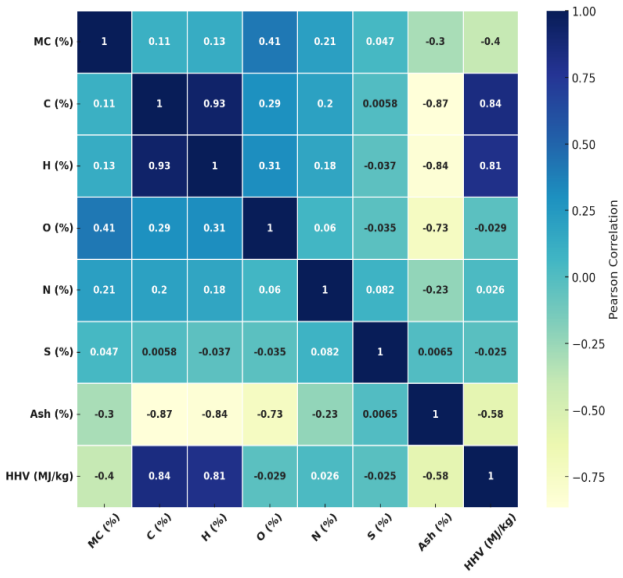

Fig. 1. Pearson correlation heat map of the waste data

3.2 Feature ranking of biomass parameters

This feature ranking facilitates the optimization of waste biomass parameters for optimum energy recovery by focusing on the most significant biomass properties. The feature analysis evaluates the impact of each waste property on the prediction of HHV. Features with large GI values substantially affect HHV prediction, whereas those with lower values contribute a little or negligibly. Using a Random Forest model, Figure 2 shows the feature importance of several waste variables for predicting the HHV of waste biomass. With around 50% contribution to the model, carbon has the most important influence on HHV prediction. It fits exactly with physical fuel chemistry that carbon is the main contributor to energy content in biomass. Next to carbon in importance is the Hydrogen Content, which is roughly 25–30%. By means of the generation of water vapor and energy release, it directly impacts fuel combustion. Emphasizing their importance in modeling HHV, Carbon and Hydrogen collectively make around 75–80% of the predictive capacity of the model. Moisture Content accounts for 10–12% of the overall significance since it is inversely related to HHV. Ash content is equally as important as moisture content. Being non-combustible and inert, it is a non-energy-adding material that even lowers it by displacing combustibles. With less than 2% importance, the combination of oxygen, nitrogen, and sulfur combined has a very little effect on HHV prediction. Although chemically important for emissions, these constituents have no direct effect on energy yield, hence the model learns to disregard them.

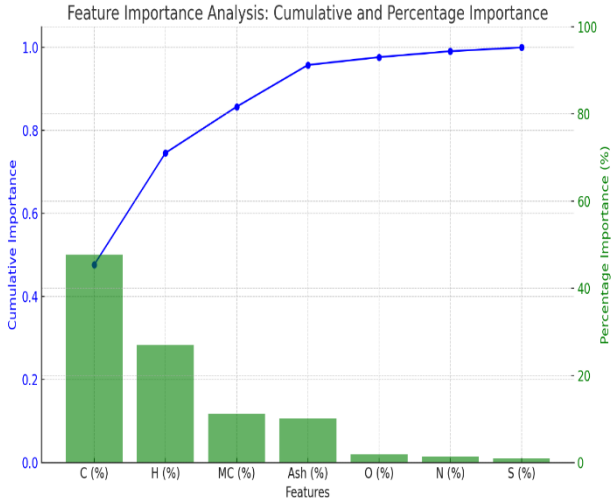

Fig. 2. Cumulative and percentage importance of the waste biomass

3.3 Dimensionality reduction of biomass features

Another multivariate analysis was carried out on the waste biomass properties to understand the underlying structure and dimensionality of the dataset. The principal

Component Analysis (PCA) was used to convert the original set of connected variables into a smaller number of uncorrelated main components, each of which captures a part of the total variance in the data. Figure 3 provides the explained and variance as well as the eigen value related to each principal component (PCs). PC1 alone accounts for roughly 45–47% of the variance, while PC2 contributes an extra 20%, bringing the total variation to 67%. Generally regarded as enough for dimensionality reduction, PC4 accounts for roughly 90–95% of the overall variance. This implies that only the first 3–4 components carry most of the meaningful variation in waste dataset. Without losing much information, the waste biomass features have been reduced from 7 original variables to 3–4 PCs. This is good for clustering, visualization, or input into machine learning models. The scree plot reveals a moderate fall through PC4 and a sharp drop-off after PC1. While the PC3 is at the borderline, PC1 and PC2 fulfill the requirement of eigen value =1 threshold. From PC5 onward, eigenvalues drop considerably below 1, thereby adding very little additional knowledge. Based on previous conversations, PC1 probably reflects a strong axis of variance ruled by the most influential characteristics, mostl likely Carbon, Hydrogen, Ash, and moisture content. The decline in eigenvalues indicates the redundancy among input features for example, strong correlation between Carbon and Hydrogen.

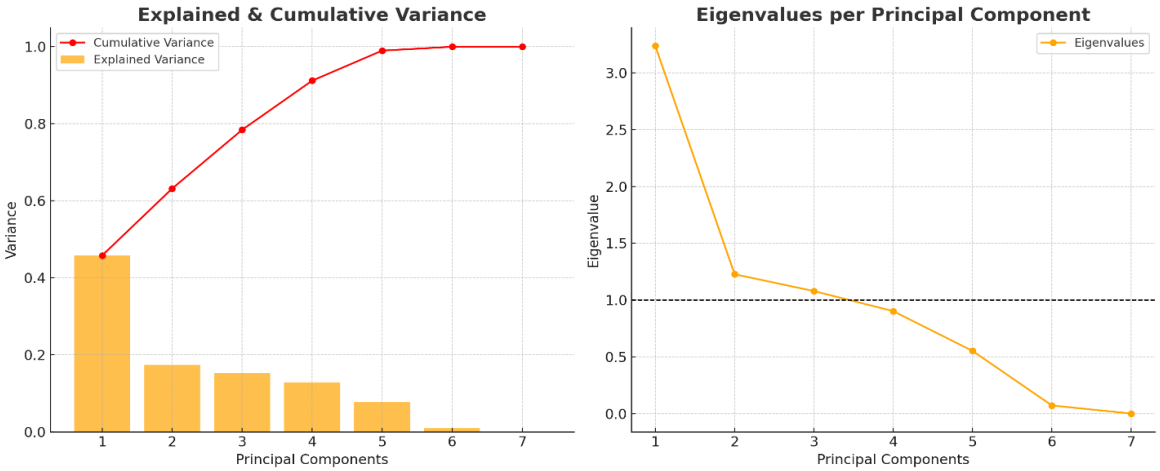

Fig. 3. Scree plot showing (a) explained and cumulative variance (b) eigen values relating to each principal component

3.4 Performance assessment of the machine learning models

Table 3 provides the performance of the developed models for HHV prediction based on the key performance metrics value during the training and testing phase. Across all training metrics, Random Forest outperformed both ANN and SVM. Its better accuracy in capturing HHV patterns is indicated by the lowest RMSE and MAE. The smaller MAD values confirms the lesser number of outliers, while the rMBE is low as well, suggesting less bias and correct training data learning. Random forest learns the training data quite effectively with no indications of overfitting. Random Forest keeps excellent performance even on unseen test data. Though error levels increased a little (as expected), they are still far lower than those of ANN and SVM. RF generalizes well with minimal MAE and RMSE. Though not significant, rMBE of 0.1554 points to a slight positive bias. Though ANN marginally beats RF in rMBE, its total predictive power is better, particularly considering its always lowest error margins. RF shows to be quite generalizable, avoiding overfitting and performing consistently across data splits

Table 3. Statistical metric result of the developed model

	Model	Performance metrics			
		RMSE	MAE	MAD	rMBE
Train	ANN	2.0249	1.4315	0.8703	0.0208
	SVM	5.8905	3.0949	1.3484	1.7577
	RF	0.8606	0.5945	0.3864	0.0895
Test	ANN	2.0824	1.5735	1.1437	0.0501
	SVM	3.5711	1.8407	0.9671	0.7568
	RF	1.4004	1.1798	1.0693	0.1554

Figures 4 and 5 shows radial charts depicting the trend of the performance metrics, RMSE, MAE, MAD and rMBE of the three models developed for HHV predictions at the training and testing phase respectively. From figure 5 (training phase), random forest has the smallest and most compact polygon, closely engaging the center. This shows consistent performance with low error and high accuracy across all metrics. The almost perfect symmetry suggests strong training. In particular, RMSE and MAE polygon in ANN is moderately sized and uneven, indicating moderate error values. Slight elongation in the MAE and MAD axes suggests these were weaker areas in ANN training performance. The SVM features a broad, spiky polygon, stretching far from the center. This reveals significant error values on all metrics. SVM has the least compact polygon, indicating poor training performance. Moreover, it Shows problems with overfitting or failure to capture intricate patterns in the data throughout training. Likewise during the testing stage (Figure 6), random forest has a tiny and tight polygon although a little more extended than in training, It performs well on unseen data and keeps compactness, particularly on RMSE and MAE axes. Though quite symmetrical, ANN has a polygon larger than the random forest. This points to a larger size than training recommends a bit larger testing error. Furthermore, ANN is uniform and fair across criteria, suggesting that although not the greatest, it's a reliable model. At the testing stage, SVM still boasts the largest and most extended polygon size. Though a little smaller than in training, it nevertheless reveals notable errors. Visibly distant from the center, the polygon shows poor predictive power on RMSE and MAE.

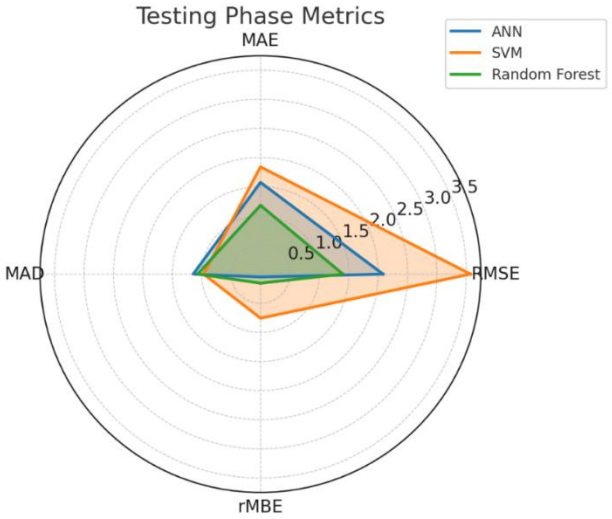

Fig. 4. Radar chart depicting all model's performance at the training

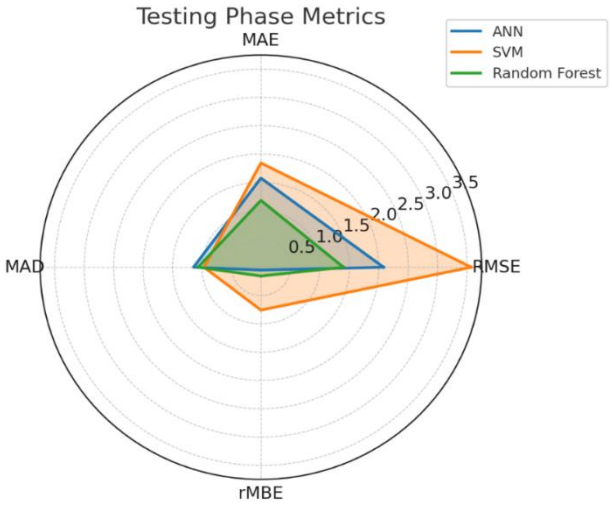

Fig. 5. Radar chart depicting all model's performance at the testing

Figures 6 and 7 compare the experimental and predicted HHV at the traiinng and testing stages. The figures clearly demonstrates that the random forest model achieves excellent precision in training and exhibits robust generalization on testing data. Low bias is suggested by the error distribution's center around zero at the training phase in figure 7. Its shape closely matches a normal distribution curve, which indicates a well-balanced model. The errors are mostly between **-2** and +2 MJ/kg, suggesting close prediction intervals. At the testing phase, the predicted values still follow the actual trend quite closely, though minor deviations are visible in certain samples. The alignment is good, but some wider gaps occur, especially in samples with extreme HHV values (outliers). The error histogram is centered near zero, meaning the model retains low bias even on unseen data. The histogram is slightly wider and more uneven compared to training, indicating higher variance in prediction errors. Errors mostly fall within -2.5 to +3

MJ/kg, slightly wider than training but still well-controlled

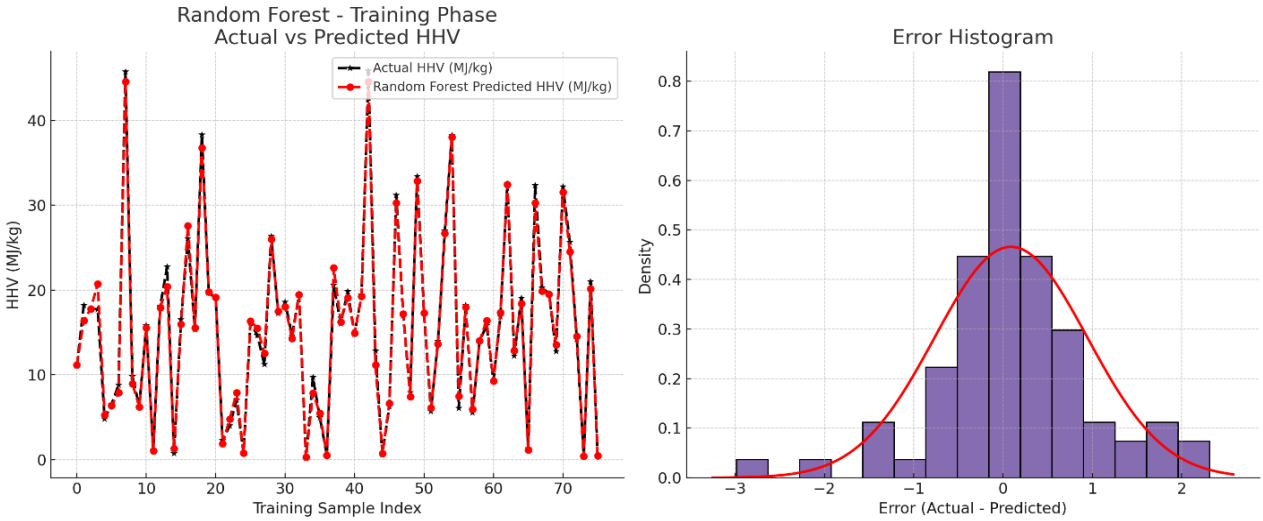

Fig. 6. Comparison trend plot of the actual and RF-prediction HHV values and error histogram at the training

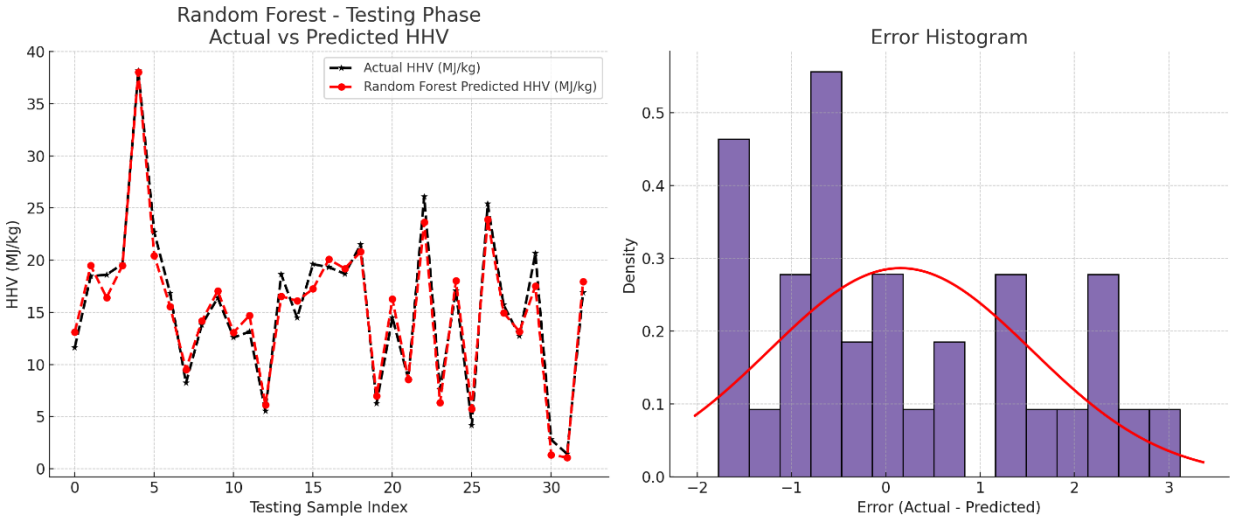

Fig. 7. Comparison trend plot of the actual and RF-prediction HHV values and error histogram at the training

4 Conclusion

This study shows how well a multivariate statistical and machine learning-driven method works for forecasting the energy potential of waste-derived biomass and defining it. Through a nexus of statistical and machine learning-based model such correlation analysis, feature ranking, dimensionality reduction, we provide data driven insights into the compositional characteristics of waste for energy recovery. The analyses of correlation and feature importance revealed Carbon and Hydrogen to be the main contributors to HHV, together comprising as much as 80% of model influence. By means of PCA, more than 90% of the variation was captured by the first four components, so enabling waste grouping by means of dimensionality reduction of the dataset. Of the models assessed, RF showed better performance on all metrics in both training and testing stages, hence highlighting its strength and generalizability. Combining soft computing methods with statistical tools creates a strong, data-driven framework for waste-to-energy optimization, hence enabling strategic decision-making in WTE system

The authors appreciate the Department of Mechanical engineering Science, University of Johannesburg for providing workspace e for this research.

References

1. O. Adeleke, O. O. Olatunji, T.-C. Jen, and I. Olawuyi, “Enhanced prediction of heating value of municipal solid waste using hybrid neuro-fuzzy model and decision tree-based feature importance

assessment," Green Energy and Resources, **3(1)** :100119 (2025).
2. F. Ardolino, F. Parrillo, and U. Arena, "Biowaste-to-biomethane or biowaste-to-energy? An LCA study on anaerobic digestion of organic waste," J Clean Prod, **174**:462–476 (2018)
3. P. Jiang *et al.*, "Establishing a generalized model for accurate prediction of higher heating values of substances with large ash fractions," Green Chemical Engineering, (2024)
4. S. Das, S.-H. Lee, P. Kumar, K.-H. Kim, S. S. Lee, and S. S. Bhattacharya, "Solid waste management : Scope and the challenge of sustainability," *J Clean Prod*, vol. 228, pp. 658–678, 2019, doi: 10.1016/j.jclepro.2019.04.323
5. M. K. Awasthi *et al.*, *Global Status of Waste-to-Energy Technology*. Elsevier B.V., 2019.
6. M. D. Khan *et al.*, "Bioelectrochemical conversion of waste to energy using microbial fuel cell technology," Process Biochemistry. **57**, 141–158, (2017)
7. K. C. R. Drudi, R. Drudi, G. Martins, G. Colato. Antonio, and J. Tofano. C. Leite, "Statistical model for heating value of municipal solid waste in Brazil based on gravimetric composition," Waste Management, **87** :782–790, (2019)
8. Z. Chen *et al.*, "Higher heating value prediction of high ash gasification-residues: Comparison of white, grey, and black box models," Energy, **288,** 129863 (2024)
9. M. M. Mateus, J. M. Bordado, and R. Galhano dos Santos, "Simplified multiple linear regression models for the estimation of heating values of refuse derived fuels," Fuel, **294** (2021).
10. R. Amen *et al.*, "Modelling the higher heating value of municipal solid waste for assessment of waste-to-energy potential: A sustainable case study," J Clean Prod, **287**, (2021)
11. S. Mondal and I. M. Rafizul, "Predicting calorific value through proximate analysis of municipal solid waste using soft computing system," Discover Applied Sciences**. 7(3).** 212, (2025).
12. O. Adeleke, S. Akinlabi, T.-C. Jen, and I. Dunmade, "Prediction of the heating value of municipal solid waste: a case study of the city of Johannesburg," Int J of Ambient Energy, **43(1)**, 3845-3856 (2020)
13. H. You *et al.*, "Comparison of ANN (MLP), ANFIS, SVM, and RF models for the online classification of heating value of burning municipal solid waste in circulating fluidized bed incinerators," Waste Management, **68**, 186–197, (2017)
14. ECN, ""Phyllis2, database for biomass and waste " https://phyllis.nl/Browse/Standard/ECN-Phyllis accessed on 10th August, (2018)
15. L. Meraz, A. Domínguez, I. Kornhauser, and F. Rojas, "A thermochemical concept-based equation to estimate waste combustion enthalpy from elemental composition," Fuel, **82(12)** :1499–1507, (2003).
16. M. Zanganeh, "Improvement of the ANFIS-based wave predictor models by the Particle Swarm Optimization," Journal of Ocean Engineering and Science, **5(1)**:84–99 (2020),
17. X. Tan, Y. Wang, L. Wu, Y. Yu, Q. Yu, and G. Sun, "An ANFIS-Based indirect control strategy for solar heating system: Exploring PMV approach," *Energy Build*, **309**:114056 (2024)
18. W. S. Lee, V. Alchanatis, C. Yang, M. Hirafuji, D. Moshou, and C. Li, "Sensing technologies for precision specialty crop production," Comput Electron Agric, **74(1),** 2–33 (2010).
19. J. Wang, Q. Chen, and Y. Chen, "RBF Kernel Based Support Vector Machine with Universal Approximation and Its Application," in *Advances in Neural Networks – ISNN 2004*, F.-L. Yin, J. Wang, and C. Guo, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 512–517.
20. P. Dutta, S. Paul, and A. Kumar, "Chapter 25 - Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19," in *Electronic Devices, Circuits, and Systems for Biomedical Applications*, S. L. Tripathi, V. E. Balas, S. K. Mohapatra, K. B. Prakash, and J. Nayak, Eds., Academic Press, 2021, pp. 521–540.