International Journal on Computational Engineering

Journal Homepage www.comien.org/index.php/comien

Estimating the Potential of Biogas Yield from Anaerobic Co-digestion of Organic Waste with Ensemble Machine learning

Thi Minh Phuong Le [a], Van Huong Dong [b,*]

[a] School of Mechanical Engineering, Vietnam Maritime University, Haiphong, Vietnam
[b] Institute of Mechanical Engineering, Ho Chi Minh City University of Transport, Ho Chi Minh City, Viet Nam.
Corresponding author: *huong_dv@ut.edu.vn

***Abstract*— The biogas has the potential to serve as a substitute as fossil derived fuel. The present study investigates anaerobic co-digestion of organic waste for predictive modelling. Several co-digestion studies were performed with different pH, solid concentration, temperature, and co-digestion ratios. A water displacement apparatus was used to test biogas yield, and data was collected meticulously. Linear regression (LR) and Random Forest (RF) based models were built with Python-based tools and tested using statistical measures for the prediction of biogas yield. The LR showed a strong linear association, with R and R^2 values of 0.9892 and 0.9785, respectively. However, RF surpassed LR, with higher R and R^2 values of 0.9919 and 0.9826, respectively. Furthermore, RF had lower MSE and MAE values, indicating higher prediction accuracy and precision. RF consistently scored well in tests, demonstrating its ability to capture complicated relationships while minimizing prediction mistakes. Taylor's diagrams further demonstrated RF's excellent performance during both the training and testing periods. Overall, RF emerges as the optimum model for reliably estimating biogas output in anaerobic co-digestion systems, with important implications for waste-to-energy processes.**

***Keywords*— Anaerobic digestion; biogas; alternative fuel; random forest; machine learning.**

Manuscript received 22 Oct. 2023; revised 29 Dec. 2023; accepted 12 Feb. 2024. Date of publication 31 Mar. 2024.

I. Introduction

In response to the rising prices of petroleum products, the growing worries about the environment, and the diminishing availability of fossil fuels, a significant amount of research has been conducted into alternate and sustainable energy sources. Consequently, researchers are concentrating their efforts on the discovery of alternative energy sources and the use of these sources in order to reduce the adverse effects of their use [1]. The majority of the recent research that has been conducted on renewable energy sources has focused on a variety of waste products, including food waste, animal manure, organic waste, and municipal solid waste [2], [3]. Through the process of transforming these waste products into fuel, it is possible to lessen the many adverse impacts that these waste materials have on the environment and on living species, including people [4].

The waste products that are accessible include animal manure as well as food wastages, both of which comprises a considerable quantity of organic matter which is capable of being fermented without the presence of oxygen. Anaerobic digestion (AD) has been shown to be an effective approach for minimizing the amount of organic waste produced while simultaneously the recovery of valuable byproducts such as digestate and biogas [5]. There is a considerable relationship between the kind of waste that is utilized as a feedstock in the AD process and the performance of the process. Co-digestion, which includes employing a variety of organic waste for feedstock, is becoming popular owing to its capacity to boost the generation of biogas and methane in comparison to AD techniques that only use a single feedstock (mono-digestion). The carbon-to-nitrogen (C/N) ratio may be effectively balanced by the use of co-digestion, which also helps to reduce the inhibitory effects of ammonia and overcome the obstacles that are associated with mono-digestion for example [6].

Biogas is one of the several biofuels that are both renewable and cost-effective [7], [8]. The main and secondary gases make up the majority of its composition. Methane accounts for almost concentration of 60%–60% of primary gases, whereas carbon dioxide accounts for almost 40%–50% of primary gases. For example, hydrogen sulfide, hydrogen, vapor of water, and siloxane are examples of trace gases that are found in secondary gases, which are found in trace

amounts [9]. The majority of industrialized nations have improved their technological capabilities and infrastructure, which has led to the production of biogas on a massive scale from local facilities such as farms, food processing industries (FPI), effluent treatment plants (ETP), and other similar establishments [10]. The incorporation of such biofuel into the gas grid or as fuel for vehicles over an extended period of time has the potential to significantly contribute to the alleviation of concerns about greenhouse gas emissions (GHG), climate change, and the worsening of public health standards [11], [12]. Over the course of the last several years, the amount of food waste (FW) that is produced by food processing industries and residential facilities has reached 1.6 gigatons of FW annually. Everyday industry is the source of the majority of processed FW. On a global scale, the food supply chain is responsible for the generation of 160–295 kg/year/person of food waste [13]. This primarily encompasses activities such as production, processing, consumption, post-handling, and distribution [1]. The FW is characterized by it having a greater saline level, volatile solid, and moisture content. The FW that are thrown away are the primary contributor to greenhouse gas emissions and a foul odor. Because of its straightforward and consistent nature, the sources of FW creation in FPI may be simply recognized and recycled once they have been generated [1].

The process of biogas generation through this process is complex and is influenced by several control factors. In these circumstances it becomes difficult to model the process through conventional numerical methods. The objective has been to develop a précised biogas yield prediction model in order to conduct an analysis of the working system of anaerobic co-digestion. Data-driven modeling that makes use of machine learning methods is one method that may be used to do this [14]. To facilitate the training process, a priori databases are necessary. To validate the models, test data comprised of observed test cases is employed. To predict the quantity of biogas generated through anaerobic co-digestion processes, several models utilizing machine learning techniques have been devised. These models can be developed using different ML based methods like random forests (RF), fuzzy logic, artificial neural networks (ANN), gradient boosting, gene expression programming, and extreme gradient boosting (XGBoost). According to the findings of Pei et al. [15], the extreme learning machine (ELM) model offered the most accurate forecast for the production of biogas. The model had a a coefficient of determination (R^2) as 0.95 and mean absolute error (MAE) of 0.67. Through the use of the random forest (RF) model, they were also able to determine that acetic acid, butyric acid, and pH are key parameters that influence the yield of biogas production [13]. Also, Karichappan et al. [16] used a statistical approach (Box-Behnken) in RSM to conduct an analysis cum optimization of biogas production. This analysis took into account several factors, like reactor temperature, pH, process alkalinity, and feedstock retention metrics. It was concluded that variation in pH and temperature had a substantial impact on the average amount of biogas produced, the cumulative amount of biogas production, and the concentration of methane.

The literature reveals that biogas generation through co-digestion of organic waste is a complex and non-linear process and specially challenging for modelling through numerical methods. In this study an ensemble method of ML is explored for this purpose. It will be compared with linear regression for comparison as baseline process. The data collected from testing phase will be utilized for this purpose.

II. Material and Methods

A. Anaerobic co-digestion

The test setup employed in the study comprised of lab-scale anaerobic batch reactors. The glass reactor's overall capacity was 3 liters, with a useful operational capacity of 2.45 liters. The reactors were equipped with a bath of water for temperature control and a magnetic stirring device for agitation. Co-digestion experiments were carried out at different operation setting of: pH, solid concentration (Solid Conc.%), temperature (T, °C), and Co-digestion. For the measurement of biogas yield a water displacement type apparatus was employed. The data was carefully collected for subsequent use in prediction modelling.

B. Machine learning

Linear Regression (LR) and Random Forest (RF) are two common ML methods used for regression applications that require predicting continuous data. Let us go down each of these strategies in simple words. Assume one have a series of data points on a graph that essentially follow a straight line. Linear regression is equivalent to drawing the best-fitting line between the points. It aids in understanding the link between two variables by identifying the line that minimizes the distance between the actual data points and the line itself. For example, if we want to forecast someone's height based on their age, linear regression can tell us how much height grows with each year of life. A typical flow chart for implementation of LR is depicted in Figure 1 [17].

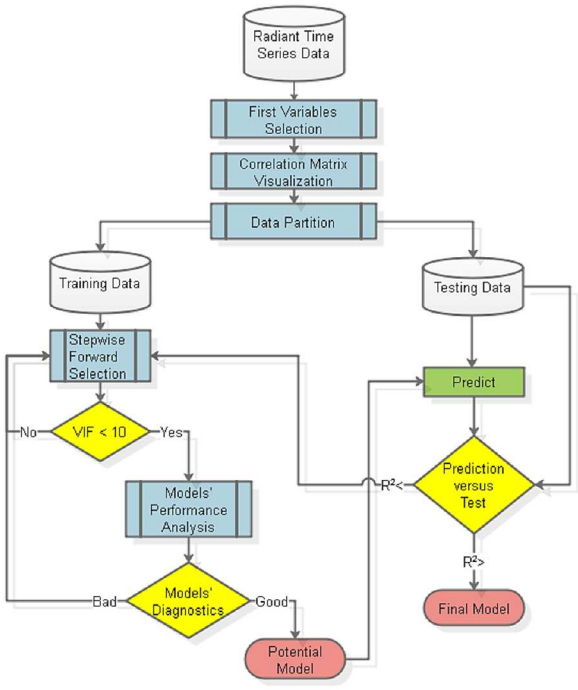


Fig. 1 LR Flow chart [17]

Random Forest: Assume you have a forest with several trees, each with its own unique method of forecasting things.

Random forest is analogous to having a collection of these trees, and when we want to make a forecast, we ask each tree for its opinion. Then we integrate all of these views to make a final projection [18], [19]. Each tree in the forest is trained on a random selection of data and generates its own forecast. Random forest excels at managing more complicated interactions between variables and may detect nonlinear patterns in data [20], [21]. When comparing the two approaches, linear regression is more straightforward and easier to comprehend. It works best when the connection between variables is linear, or can be represented by a straight line. However, it may underperform when the connection is more complicated or nonlinear. A typical RF flow chart is depicted in Figure 2 [22].

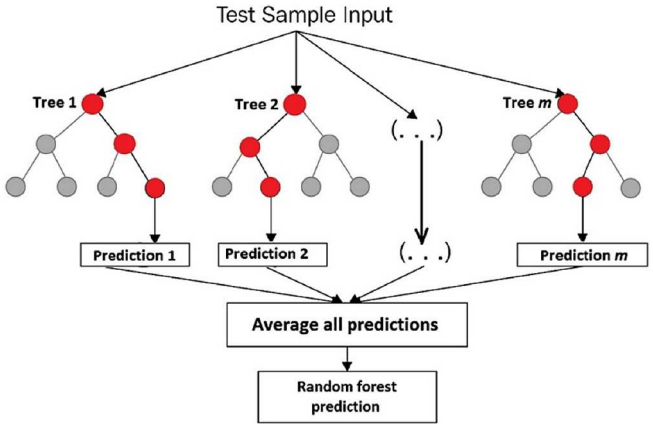


Fig. 2 Flow chart of RF [22]

Random forest, on the other hand, is more adaptable and more suited for complicated interactions. It's like having a group of experts (trees) collaborate to get a resolution. Random forest is more accurate and resilient, particularly when there are several variables or interactions between them. To summarize, linear regression is a basic tool that works well for simple situations with linear connections, but random forest is a sophisticated tool that can handle more complicated problems and detect nonlinear patterns in data. Both strategies have advantages and disadvantages, and the decision is based on the particular situation and data at hand [22], [23].

III. Result and Discussion

The data collected from the anaerobic co-digestion of waster organic matter was used for model development. LR was used as a baseline ML while ensemble ML technique RF was also used to explore this highly efficient ML in this case. The data was randomly split in two parts. One larger chunk (70%) was employed for model training and remaining was employed for model testing.

A. Data pre-processing

The data was used for development of correlation heatmap and calculating the correlation matrix. The correlation matrix is listed as Table 1 while the correlation heatmap is depicted in Figure 3.

The correlation matrix in Table 1 gives useful information on the links between the factors involved in anaerobic co-digestion of organic waste. It was shown that the association between biomethane solid concentration (%) and yield (mL) is -0.2534. That association suggests that when solid content rises, biomethane production falls, and vice versa. Moving on to the undisturbed data, we discovered correlations between two variables. For example, the correlation coefficient between pH and temperature (T) is 0.1459, showing a positive relationship between the two variables. Similarly, co-digestion efficiency (%) and yield (mL) have a high positive correlation of 0.945441, showing that as co-digestion efficiency grows, so does biogas output.

Overall, the correlation matrix gives a thorough description of the correlations between the factors involved in anaerobic co-digestion, allowing researchers to uncover patterns and possible insights that may aid in future study and decision-making processes.

TABLE I
Correlation Matrix

	Solid Conc.%	pH	T, °C	Co-dig., %	Yield, mL
Solid Conc.%	1	-8.23E-17	-1.21E-16	-7.11E-17	-0.2534
pH	-8.23E-17	1	0.14599	3.95E-17	0.14344
T, °C	-1.21E-16	0.14599	1	-1.31E-16	-0.0258
Co-dig., %	-7.11E-17	3.95E-17	-1.31E-16	1	0.94544
Yield, mL	-0.2534	0.14344	-0.0258	0.94544	1

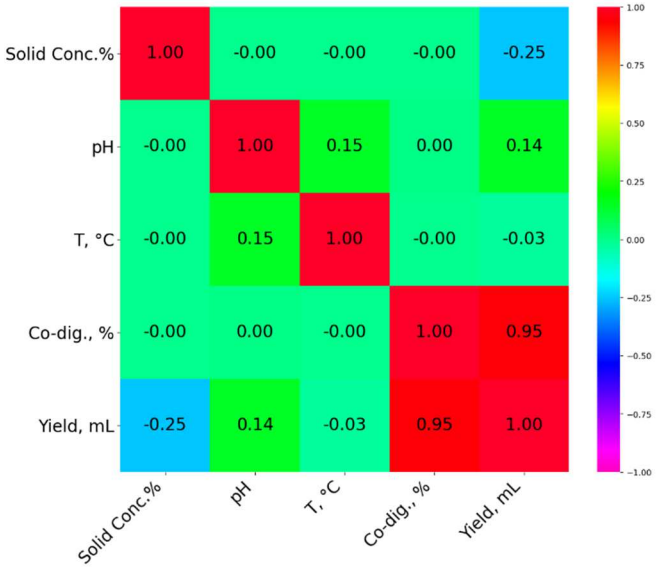


Fig. 3 Correlation heatmap

B. Model development and evaluation

The models were developed in python based open access libraries. Both LR and RF model were developed and used for making predictions. Then these predictions were tested on different statistical metrics a listed in table 2. The actual vs model predicted biogas yield using LR is plotted in Figure 4a for model training phase while it is shown in Figure 4b for RF

based model during model training phase. Similarly, for model testing phase the Figure 5a depicts for LR and Figure 5b for RF based models.

The statistical analysis shown in Table 2 as an exhaustive analysis of the LR and RF models that are used for the purpose of forecasting the biogas yield. The Pearson correlation coefficient (R), the coefficient of determination (R^2), the Kling-Gupta Efficiency (KGE), the mean squared error (MSE), and the mean absolute error (MAE) are the metrics that are used for assessment. The LR model acquired a R value of 0.9892 and an R2 value of 0.9785, which indicates a strong linear connection and a high degree of variance explained by the model. These values were obtained in the first set of data when the model was applied. The fact that the KGE value is 0.9847 is more evidence that the model is successful in understanding the variability that has been seen. The LR model, on the other hand, had MSE and MAE values that were substantially higher than average, coming in at 43.95 and 5.36, respectively, indicating that there was some degree of prediction inaccuracy.

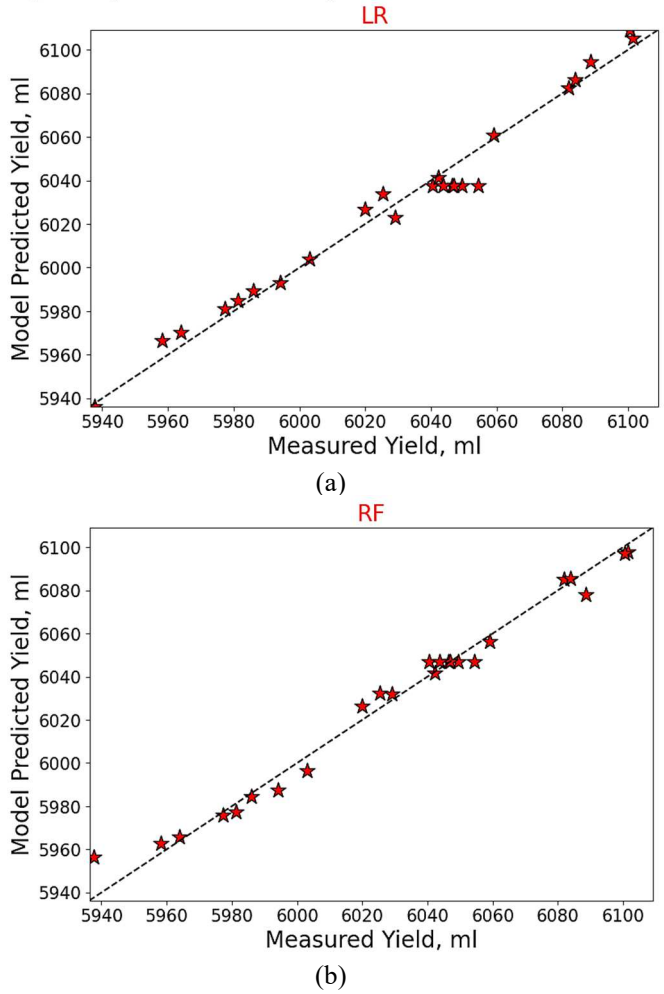


Fig. 4 Model training performance in terms of actual vs predicted in training phase for (a) LR (b) RF based biogas yield

On the other hand, the RF model displayed greater performance, with a R value of 0.9919 and an R^2 value of 0.9826. These values indicate that the RF model provides a stronger explanation for correlation and variance in comparison to the LR model. In addition, the RF model produced MSE and MAE values that were lower, coming in at 35.44 and 4.47, respectively, which indicates that the prediction accuracy and precision of the model were improved. The second set of findings demonstrates that the performance of the LR model deteriorated throughout the testing period. This is shown by the fact that the R and R2 values declined, in addition to the MSE and MAE values increasing. The RF model, on the other hand, maintained a robust performance, exhibiting continuously high R, R^2, and KGE values while maintaining MSE and MAE values that were relatively low. These data demonstrate that RF is superior than LR when it comes to forecasting biogas output, especially when it comes to capturing complicated nonlinear interactions and reducing the number of mistakes that occur during prediction.

TABLE II
STATISTICAL EVALUATION OF LR AND RF BASED BIOGAS YIELD PREDICTION MODEL

Phase	Model	R	R^2	KGE	MSE	MAE
Model	LR	0.9892	0.9785	0.9847	43.95	5.36
training	RF	0.9919	0.9826	0.9551	35.44	4.47
Model	LR	0.924	0.826	0.747	356.1	16.01
testing	RF	0.9903	0.974	0.907	54.02	6.19

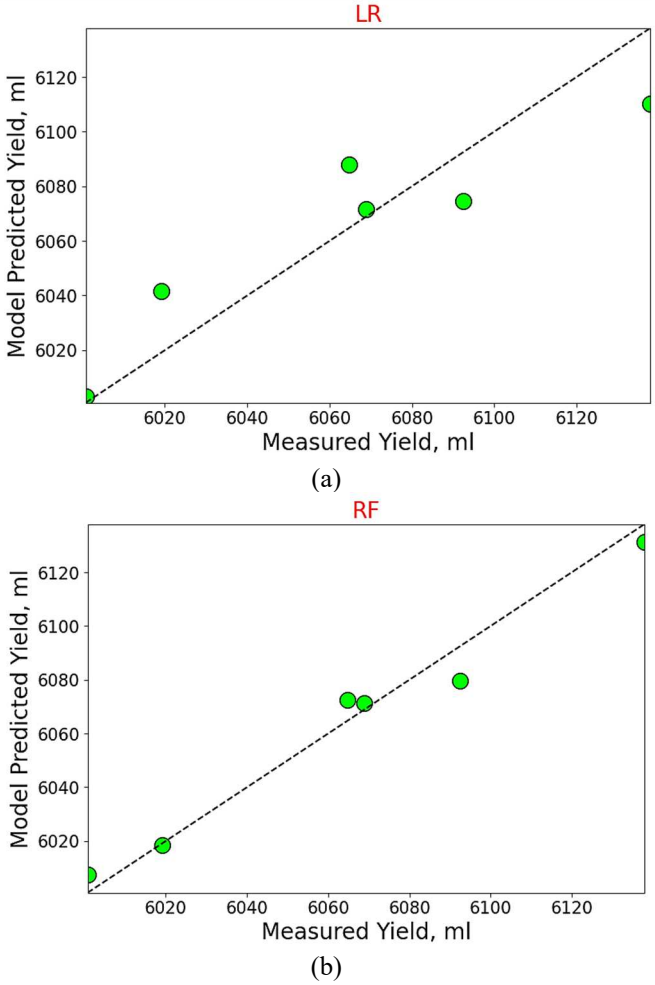
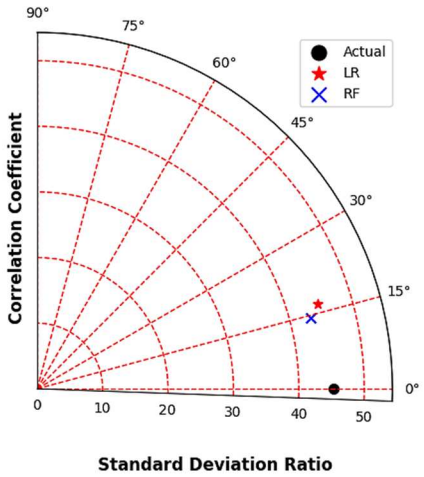


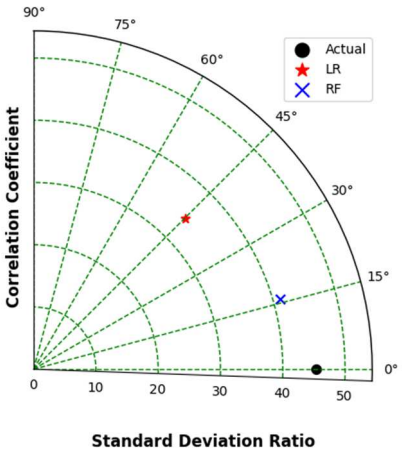
Fig. 4 Model training performance in terms of actual vs predicted in testing phase for (a) LR (b) RF based biogas yield

The models were also compared using Taylor's diagram to show case the model performance and also for their comparison. The Taylor's diagram for training as well as

testing phase of models is depicted in Figure 5a and Figure 5b respectively. It helped in easy identification of RF based model as better performing model both in training as well as testing phases.



(a)



(b)

Fig. 5 Taylor's plots for model (a) training (b) testing

IV. Conclusion

In the present study the data gathered from anaerobic co-digestion of organic waste matter was employed for the prognostic model developments. Co-digestion tests were carried out with various pH, solid concentration, temperature, and co-digestion ratios. A water displacement device was used to assess biogas yield, and thorough data collection was performed in preparation for predictive modeling. The LR and RF models were built using Python-based open-access modules and tested using a variety of statistical criteria. LR had a strong linear association with R and R^2 values of 0.9892 and 0.9785, respectively, but RF performed better with higher R and R^2 values of 0.9919 and 0.9826, respectively. Furthermore, RF produced lower MSE and MAE values, suggesting higher prediction accuracy and precision than LR. RF performed well throughout the testing phase, demonstrating its ability to capture complicated nonlinear interactions while reducing prediction errors. Taylor's illustrations demonstrated RF's better performance throughout both the training and testing periods. Overall, RF emerges as the most accurate model for predicting biogas generation in anaerobic co-digestion systems.

References

[1] D. Singh, M. Tembhare, N. Machhirake, and S. Kumar, "Biogas generation potential of discarded food waste residue from ultra-processing activities at food manufacturing and packaging industry," *Energy*, vol. 263, 2023, doi: 10.1016/j.energy.2022.126138.

[2] Z. Huiru, Y. Yunjun, F. Liberti, P. Bartocci, and F. Fantozzi, "Technical and economic feasibility analysis of an anaerobic digestion plant fed with canteen food waste," *Energy Convers Manag*, vol. 180, 2019, doi: 10.1016/j.enconman.2018.11.045.

[3] G. L. K. Srinivas, D. Singh, and S. Kumar, "Transition of Biofuels from the First to the Fourth Generation: The Journey So Far," in *Biofuels: Technologies, Policies, and Opportunities*, 2023. doi:10.1201/9781003197737-2.

[4] O. A. Aworanti *et al.*, "Enhancing and upgrading biogas and biomethane production in anaerobic digestion: a comprehensive review," *Frontiers in Energy Research*, vol. 11. 2023. doi:10.3389/fenrg.2023.1170133.

[5] Z. Hajabdollahi Ouderji *et al.*, "Integration of anaerobic digestion with heat Pump: Machine learning-based technical and environmental assessment," *Bioresour Technol*, vol. 369, 2023, doi:10.1016/j.biortech.2022.128485.

[6] R. Karki *et al.*, "Anaerobic co-digestion: Current status and perspectives," *Bioresource Technology*, vol. 330. 2021. doi:10.1016/j.biortech.2021.125001.

[7] G. Piechota and B. Igliński, "Biomethane in Poland—Current Status, Potential, Perspective and Development," *Energies (Basel)*, vol. 14, no. 6, 2021, doi: 10.3390/en14061517.

[8] D. Mignogna, P. Ceci, C. Cafaro, G. Corazzi, and P. Avino, "Production of Biogas and Biomethane as Renewable Energy Sources: A Review," *Applied Sciences (Switzerland)*, vol. 13, no. 18. 2023. doi: 10.3390/app131810219.

[9] J. C. Frigon and S. R. Guiot, "Biomethane production from starch and lignocellulosic crops: A comparative review," *Biofuels, Bioproducts and Biorefining*, vol. 4, no. 4. 2010. doi:10.1002/bbb.229.

[10] O. Eriksson, M. Bisaillon, M. Haraldsson, and J. Sundberg, "Enhancement of biogas production from food waste and sewage sludge - Environmental and economic life cycle performance," *J Environ Manage*, vol. 175, 2016, doi:10.1016/j.jenvman.2016.03.022.

[11] P. C. Slorach, H. K. Jeswani, R. Cuéllar-Franca, and A. Azapagic, "Environmental sustainability of anaerobic digestion of household food waste," *J Environ Manage*, vol. 236, 2019, doi:10.1016/j.jenvman.2019.02.001.

[12] Y. Ma and Y. Liu, "Turning food waste to energy and resources towards a great environmental and economic sustainability: An innovative integrated biological approach," *Biotechnology Advances*, vol. 37, no. 7. 2019. doi: 10.1016/j.biotechadv.2019.06.013.

[13] S. Baroutian, M. T. Munir, J. Sun, N. Eshtiaghi, and B. R. Young, "Rheological characterisation of biologically treated and non-treated putrescible food waste," *Waste Management*, vol. 71, 2018, doi:10.1016/j.wasman.2017.10.003.

[14] M. Alruqi and P. Sharma, "Biomethane Production from the Mixture of Sugarcane Vinasse, Solid Waste and Spent Tea Waste: A Bayesian Approach for Hyperparameter Optimization for Gaussian Process Regression," *Fermentation*, vol. 9, no. 2, p. 120, Jan. 2023, doi:10.3390/fermentation9020120.

[15] Z. Pei *et al.*, "Understanding of the interrelationship between methane production and microorganisms in high-solid anaerobic co-digestion using microbial analysis and machine learning," *J Clean Prod*, vol. 373, 2022, doi: 10.1016/j.jclepro.2022.133848.

[16] T. Karichappan, S. Venkatachalam, and P. M. Jeganathan, "Investigation on biogas production process from chicken processing industry wastewater using statistical analysis: Modelling and optimization," *Journal of Renewable and Sustainable Energy*, vol. 6, no. 4, 2014, doi: 10.1063/1.4892604.

[17] N. M.-A. Mutombo and B. P. Numbi, "Development of a Linear Regression Model Based on the Most Influential Predictors for a Research Office Cooling Load," *Energies (Basel)*, vol. 15, no. 14, p. 5097, Jul. 2022, doi: 10.3390/en15145097.

[18] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, 2001, doi: 10.1023/A:1010933404324.

[19] A. M. Walker *et al.*, "Evaluating the performance of random forest and iterative random forest based methods when applied to gene

expression data," *Comput Struct Biotechnol J*, vol. 20, 2022, doi:10.1016/j.csbj.2022.06.037.

[20] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, 2020, doi:10.1177/1536867X20909688.

[21] P. Chen, A. Niu, W. Jiang, D. Liu, B. Ma, and T. Bao, "Air pollutant prediction: Comparisons between LSTM, light GBM and random forests," *Journal of Environmental Protection and Ecology*, vol. 20, no. 3, 2019.

[22] H. A. Zeini, D. Al-Jeznawi, H. Imran, L. F. A. Bernardo, Z. Al-Khafaji, and K. A. Ostrowski, "Random Forest Algorithm for the Strength Prediction of Geopolymer Stabilized Clayey Soil," *Sustainability*, vol. 15, no. 2, p. 1408, Jan. 2023, doi:10.3390/su15021408.

[23] M. Gholizadeh, M. Jamei, I. Ahmadianfar, and R. Pourrajab, "Prediction of nanofluids viscosity using random forest (RF) approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 201, p. 104010, Jun. 2020, doi: 10.1016/j.chemolab.2020.104010.

[1]