**Journal of Information Systems Engineering and Management**
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Optimizing Biogas Production with Machine Learning: A Comparative Study of Predictive Models.

Omar J. Sinayobye [1] [+] , Richard Musabe[1], Alexander Ngenzi[1], Eric Hitimana[1], Alfred Uwitonze[2], Martin Kuradusenge[1], Angelique Mukasine[1], Zubeda Ukundimana[3], Richard Ishimwe[1], Kevine Mugisha[1].

[1] *Department of Computer and Software Engineering, College of Science and Technology, University of Rwanda, Kigali, Rwanda.*
[2] *Ottawa Campus, Herzing College, Ottawa, Canada.*
[3] *Department of Chemistry, College of Science and Technology, University of Rwanda, Kigali, Rwanda.*

**ARTICLE INFO**

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

**ABSTRACT**

The prediction of biogas production is essential for optimizing operational conditions, enhancing process efficiency, and supporting sustainable energy systems. Traditional biogas yield prediction methods struggle to capture the nonlinear and complex interactions among influential factors such as feedstock composition, temperature, pH, and retention time. Machine learning (ML) models provide a promising alternative by analyzing patterns in historical data to make accurate, data-driven predictions. This study evaluates the effectiveness of six ML models Linear Regression (LR), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), k-nearest Neighbors (k-NN), and Artificial Neural Networks (ANNs) for predicting biogas production on dataset of an experiment performed in 5 years from January 1, 2019, to October 30, 2024. Each model's performance was assessed using common evaluation metrics for regression analysis, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Squared ($R^2$) Score, to compare their accuracy, robustness, and suitability for biogas data, which often involves nonlinear relationships and multivariate interactions. The findings demonstrate that DT and RF outperform simpler approaches in terms of accuracy with of 0.999 and 0.998 respectively, making them ideal for complex biogas prediction tasks. This study underscores the potential of ML models in optimizing biogas production systems and contributes to developing efficient, scalable solutions for renewable energy management.

**Keywords:** Biogas production, Machine learning, Prediction models, Anaerobic digestion, Regression analysis.

## 1. INTRODUCTION

A well-functioning biogas system offers environmental and resource conservation benefits. Biogas is produced through anaerobic digestion (AD) of organic waste, yielding methane ($CH_4$) and carbon dioxide ($CO_2$), and is utilized for electricity, heat, or upgraded to biomethane [2]. AD is one of the oldest methods for industrial waste treatment and sludge stabilization [1]. Precise AD process control is crucial to maximize biogas yield, though production is complex and affected by factors like feedstock properties, temperature, pH, microbial activity, and hydraulic retention time. Predicting biogas output accurately is challenging, as traditional models often miss nonlinear interactions [3].

Anaerobic digestion (AD) is a widely adopted method for organic waste treatment, offering advantages such as biogas production, low sludge output, pathogen removal, and the creation of organic fertilizers [4]. As demand for sustainable energy grows, efforts to enhance biogas yield and improve AD energy efficiency have intensified [5]. Biogas, primarily composed of methane (55-70%) and carbon dioxide (30-40%), serves as a renewable energy source that can replace environmentally harmful and rapidly depleting fossil fuels [6][7]. However, biogas production is a complex, microorganism-driven process influenced by factors such as pH, temperature, and the carbon-to-nitrogen ratio, and it often faces stability challenges that affect efficiency [8]. Proper monitoring, process control, and modeling of the anaerobic process are crucial to predicting performance indicators like methane yield, enabling more stable and efficient plant operations [9].

[+] Corresponding author. Tel.: +250788439499
*E-mail address*: j.sinayobye@ur.ac.rw / sijaom2@gmail.com

ML models are powerful tools for addressing the complexities of biogas production, offering the ability to analyze historical data, identify patterns, and make reliable predictions [10]. Unlike traditional linear methods, ML captures both linear and nonlinear relationships, optimizing operational parameters in AD to enhance efficiency and stability [11,12]. Models like LR, DTs, RFs, SVM, k-NN, and ANNs have unique strengths, from baseline modeling to capturing complex nonlinear patterns [15–17]. Metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used to evaluate these models, with advanced models like RF and ANNs often outperforming simpler ones in predicting biogas production [18–23]. This study evaluates these models using real-world data to identify the most suitable approaches for accurate and interpretable biogas predictions, supporting improved sustainability in renewable energy management.

## 2. LITERATURE REVIEW

### 2.1 Related Work

The ML models have gained significant attention for predicting biogas production due to their ability to capture nonlinear relationships and complex interactions in AD processes. Traditional statistical methods, which rely on linear assumptions, struggle to predict biogas yields accurately when parameters like substrate composition, temperature, pH, and retention time exhibit nonlinear dependencies [24]. While LR is often used as a baseline model for biogas prediction, it performs well only with linear data or limited input features [25]. As data complexity increases, more advanced ML approaches are explored to improve prediction accuracy.

ML models offer diverse capabilities for biogas prediction, each with unique strengths and limitations. DTs effectively handle nonlinearity and categorical variables, making them interpretable and useful for understanding biogas yield drivers, though prone to overfitting without proper hyperparameter tuning [26]. Ensemble models like RFs improve accuracy by reducing variance and overfitting but require significant computational resources for large datasets [27]. SVMs excel in capturing complex relationships within smaller datasets, though they require precise parameter tuning for optimal performance [28]. Simpler models like k-NN are useful for pattern recognition but face challenges with computational expense and sensitivity to neighbor selection in large datasets [29,30]. Advanced models like ANNs capture nonlinear interactions effectively, offering high prediction accuracy, but their high computational demands and lack of interpretability have spurred hybrid approaches combining ANNs with simpler methods for improved robustness and transparency [31–35].

ML enables computers to uncover hidden insights by learning from data using algorithms. In a study [36], five ML algorithms (XGBoost, SVM, ANN, RF, and LR) were used to forecast biogas production at an industrial-scale plant processing food waste. The Random Forest (RF) model performed best with an $R^2$ of 0.74 when all standard monitoring indicators were included. De Clerc et al. found that RF and XGBoost outperformed Elastic Net in predicting biomethane production, with $R^2$ values ranging from 0.80 to 0.88 across time horizons [37][38]. Long et al. showed that increasing data and features could improve prediction accuracy [39].

Existing research on biogas production largely focuses on lab or pilot-scale reactors, leaving a gap in applying AI-based models to full-scale sludge digestion in biological treatment plants. Advancements in ML and hybrid models have enhanced prediction accuracy, interpretability, and scalability. This study aims to predict biogas production rates using AI and regression models, evaluating their performance with various statistical indicators.

### 2.2 Machine Learning Techniques

In this research work, the 6 most common ML-regression algorithms, such as LR, DT, RF, SVM, KNN, and ANN were adopted to develop and compare ML models for predicting biogas production. The following section briefly describes these models. The Table 1 gives a clear and compact summary of the most widely used machine learning algorithms. They are described in terms of what sorts of problems and datasets they are best suited for. References to relevant studies and applications are included as much to support the explanations here, as to point interested readers toward further exploration.

Table 1: Machine Learning Techniques Commonly Used.

| Algorithms | Descriptions | References |
|---|---|---|
| Artificial Neural Network (ANN) | Mimics the human brain's neurons, excelling in nonlinear, complex problems. Commonly used in AD and environmental processes. Notable for prediction accuracy but lacks interpretability. | [40, 41, 59, 44] |

| | | |
|---|---|---|
| Random Forest (RF) | Ensemble method using multiple decision trees to improve predictions by considering all available attributes and reducing overfitting. Effective for high-dimensional data. | [45, 58, 46, 47] |
| Support Vector Machine (SVM) | Maps data into a higher-dimensional space to make it linearly separable, ideal for regression and handling non-linear relationships. Robust to outliers but less effective with noisy data. | [48, 49] |
| K-Nearest Neighbor (KNN) | Simple method predicting based on the k nearest neighbors, using distance metrics like Manhattan distance. Sensitive to the choice of 'k'. | [51, 42, 50] |
| Linear Regression (LR) | Predicts a dependent variable from multiple independent variables, assuming a linear relationship. Limited by its assumption of linearity and independence. | [52, 53] |
| Decision Tree (DT) | Splits data into nodes to make predictions, minimizing errors by selecting optimal splits. Flexible, interpretable, and handles non-linear relationships. | [53, 54, 60] |

## 3. METHODOLOGY

The methodology comprises different key stages: data collection, wrangling and preprocessing, feature selection, model training and tuning, performance evaluation, and model comparison. Each step is designed to maximize prediction accuracy and identify the most effective ML models for biogas production prediction.

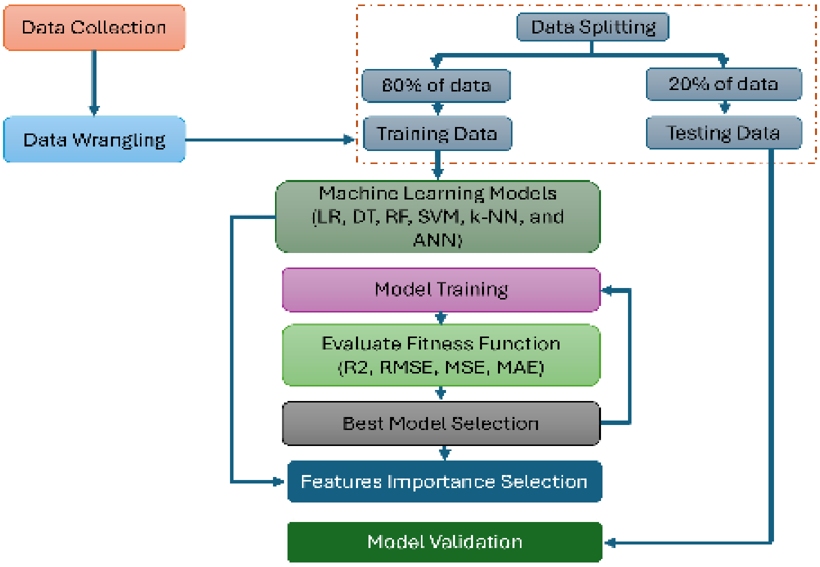

Fig. 1: Methodology for Machine Learning-based Models Development.

The structured machine learning workflow is illustrated in the Fig. 1, beginning with the gathering and preparation of raw data for analysis. The data are split into two subsets for model training and testing. The training subset comprises 80% of the complete dataset. With this, various models; Linear Regression, Decision Tree, Random Forest, SVM, KNN, and ANN are trained. Upon training in parallel, the model types are evaluated as to their respective performances. They are tallied using appropriate metrics: $R^2$, RMSE, MSE, and MAE. The model type evaluation and selection process are shown at the bottom left of the figure.

### 3.1 Data Collection and Pre-processing

The study employed a systematic sampling approach, collecting biogas production data at regular intervals to ensure a representative dataset. The dataset obtained is from an experiment performed in 5 years from January 1, 2019, to October 30, 2024. It was divided into categories namely, influent flow rate of the feed sludge, total solids content, total volatile solids content, alkalinity, volatile fatty acids, and total biogas production. Experiments were conducted under controlled mesophilic (35–40°C) and thermophilic (50–55°C) conditions to replicate real-world anaerobic digestion processes.

Data preprocessing steps included handling missing values, normalizing or standardizing features, and encoding categorical variables where applicable. Missing values were addressed through imputation techniques such as mean substitution for continuous variables and mode substitution for categorical ones. Features were normalized to bring them within a comparable scale, reducing the potential for skewed model training and ensuring all parameters contributed equally to the learning process. Categorical data, such as substrate types, were transformed using one-hot encoding to facilitate compatibility with ML models.

### 3.2 Model Training and Performance Evaluation

Six ML models were trained and optimized using an 80/20 split of a feature-selected dataset, with 80% for training and 20% for testing to evaluate generalizability. After training the model, it is important to measure the accuracy of prediction. The model accuracy was evaluated using three metrics: the determination coefficient ($R^2$), mean squared error (MSE), and mean absolute error (MAE). These are well-suited for regression problems aimed to predict continuous outcomes, and these metrics quantify the accuracy of predictions by evaluating the closeness of predicted values to actual ones and provide complementary viewpoints regarding how well the model performs.

$R^2$ measures the model's overall fit in terms of the explained variance. MSE is a more fine-grained, average assessment of how well the model's predicted values match the actual values. In contrast to MSE, MAE offers a more direct and straightforward interpretation of the model's error without any wild swings that might occur if some predictions are particularly far from the actual values. Together these three metrics ensure a comprehensive and nuanced picture of regression accuracy. They can be mathematically expressed by the following formulas.

Coefficient of determination ($R^2$ or R-squared): $R^2 = 1 - \frac{\sum_{i-1}^{m}(X_i - Y_i)^2}{\sum_{i-1}^{m}(\bar{Y} - Y_i)^2}$ (1)

Where, $X_i$ is the predicted $i^{th}$ value, the $Y_i$ element is the actual $i^{th}$ value and $\bar{Y}$ is the meaning of the true values constant.

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i \quad (2)$$

The coefficient of determination $R^2$ quantifies the proportion of variance in the predicted variable explained by the model's input parameters. A higher $R^2$ value indicates that the model incorporates significant input parameters and is well-trained to predict experimental values within the dataset [19, 55]. $R^2$ values range from 0 to 1, with values closer to 1 signifying better model performance [56].

Mean square error (MSE): $MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2$ (3)

MSE evaluates the average squared difference between observed and predicted values, serving as a measure of error in statistical models. An ideal model with no error has an MSE of zero, while higher MSE values indicate greater error [57]. Model selection prioritizes maximizing the coefficient of determination $R^2$ and minimizing MSE during both the testing and validation phases to ensure accurate fitting and prediction.

Mean absolute error (MAE): $MAE = \frac{1}{m}\sum_{i-1}^{m} |X_i - Y_i|$ (4)

MAE is useful when outliers represent corrupted data, as it does not heavily penalize training outliers, offering a bounded performance measure for models. However, if the test set contains numerous outliers, model performance may still degrade. MAE ranges from a best value of 0 (no error) to +∞ (worst performance).

### 3.3 Model Comparison and Analysis

A comparative analysis evaluated the strengths and weaknesses of various ML models for biogas prediction. Models were ranked based on performance across three evaluation metrics, while also considering interpretability, computational efficiency, and scalability. This multi-model assessment highlights the most suitable techniques for biogas production prediction, providing a foundation for future research and practical applications in biogas systems.

## 4. RESULTS INTERPRETATION AND DISCUSSION

### 4.1 Model Comparison and Analysis

The implementation begins with importing libraries for data preprocessing and model building, including Keras for deep learning. A Keras regressor wrapper integrates an ANN into the machine learning pipeline for evaluation alongside traditional models like LR, DT, RF, SVM, and k-NN. The dataset is cleaned by excluding non-numeric columns and rows with missing target values, then split into training and testing sets. Models are trained and evaluated using R² scores, which are visualized for comparison. The best model is identified based on R², with an example prediction demonstrating real-world utility.

Table 2: Machine Learning Techniques Commonly Used.

| SN | Models | MAE | MSE | R² Score |
|---|---|---|---|---|
| 1 | LR | 1996.3064 | 6867818.00 | 0.574689 |
| 2 | DT | 5.023474 | 7310.16 | 0.999547 |
| 3 | RF | 70.383118 | 27754.34 | 0.998281 |
| 4 | SVM | 3103.4048 | 15499340.00 | 0.040156 |
| 5 | k-NN | 468.70141 | 782832.40 | 0.951521 |
| 6 | ANN | 1746.6639 | 5767541.00 | 0.642828 |

From Table II, the DT and RF achieved the best results with very low error and a high R² score. Linear Regression and ANN showed moderate R² scores, with errors indicating they captured some but not all patterns in the data. SVM and k-NN have the lowest R².

### 4.2 Model Comparison and Analysis

The visualization from Fig. 2 provides a quick comparison of each model's effectiveness, highlighting which ones are better suited for each target parameter. It illustrates the R² scores of six different models applied across various target parameters. Each target parameter is displayed on the y-axis, while the R² score is shown on the x-axis, allowing for direct comparison of model performance on each specific parameter. The RF and DT models tend to have higher R² scores, indicating strong predictive capabilities, while Linear Regression and Support Vector Machines generally show lower scores for these tasks.

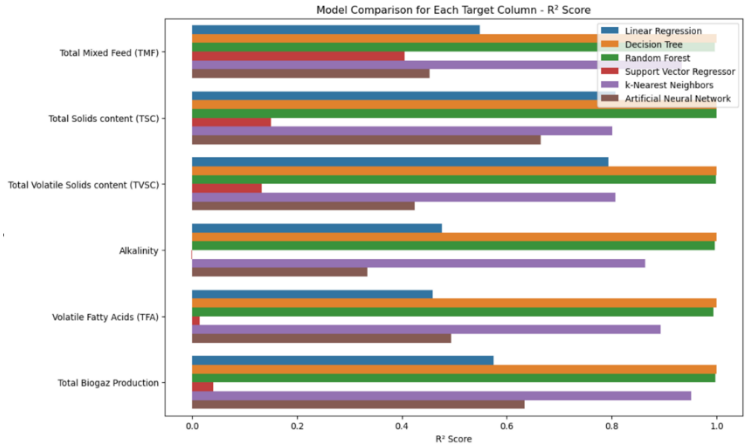

Fig. 2: Model Comparison for each targeted parameter on R² score.

### 4.3 Prediction of biogas production by the Best Performer ML model.

The table shows the performance of a DT model in predicting various target parameters with exceptionally high accuracy.

Table 3: Best Performer ML Models.

| SN | Targeted parameter | Models | MAE | MSE | R² Score |
|---|---|---|---|---|---|
| 1 | Alkalinity | DT | 4.67E-15 | 6.14378E-28 | 1.00000 |
| 2 | Total Biogas Production | DT | 5.02 | 7310.16 | 0.999547 |

| | | | | | |
|---|---|---|---|---|---|
| 3 | Total Mixed Feed (TMF) | DT | 0.000000 | 0.000000 | 1.000000 |
| 4 | Total Solids content (TSC) | DT | 1.07E-15 | 9.66632E-30 | 1.000000 |
| 5 | Total Volatile Solids content (TVSC) | DT | 4.59E-16 | 1.84E-30 | 1.000000 |
| 6 | Volatile Fatty Acids (TFA) | DT | 5.77E-15 | 3.68E-28 | 1.000000 |

For most parameters (Alkalinity, Total Mixed Feed, Total Solids Content, Total Volatile Solids Content, and Volatile Fatty Acids), the model achieves an $R^2$ Score of 1.000, indicating perfect predictive accuracy. Both MAE and MSE are extremely low, often close to zero, suggesting minimal prediction error. The only parameter with a slightly lower performance is Total Biogas Production, where the model achieved an $R^2$ of 0.999547 with a small MAE of 5.02 and MSE of 7310.16, still indicating a very high level of accuracy.

### 4.4 Developed ML models Vs Previous Works.

The table provides a comparison of the recent 4 papers related to biogas prediction, along with the algorithms used and the best model performance in terms of $R^2$ accuracy. In common, RF achieved the highest $R^2$ accuracy. Compared with our paper both DT and RF demonstrated very high accuracy in predicting biogas production, with DT reaching $R^2$ of 0.999 and RF at 0.998. This overview highlights RF as a frequently effective algorithm in this field.

Table 4: Developed ML Models Vs Previous Published Works.

| SN | Scientific Task | Algorithm | Best Accuracy | References |
|---|---|---|---|---|
| 1 | Prediction of gaseous products | RF, SVM | RF with $R^2$=0.87 | [59] |
| 2 | Prediction of biochar yield and carbon contents | RF | RF with $R^2$=0.8548 | [60] |
| 3 | Biogas Prediction for Industrial-scale Digestor | RF, XGBoost | XGBoost with $R^2$=0.88 | [38] |
| 4 | Biogas prediction accuracy | SVM, ANN, RF, KNN | RF with $R^2$=0.620 | [58] |
| 5 | Predicting Biogas Production | LR, DT, SVM, ANN, RF, KNN | RF with $R^2$=0.998, DT with $R^2$=0.999 | Our paper |

While both Decision Trees (DT) and Random Forests (RF) achieved high accuracy in predicting biogas production, RF is generally preferred due to its robustness, stability, and better generalization to unseen data. Unlike DT, which is prone to overfitting, RF reduces variability by averaging multiple trees, ensuring consistent predictions. It also handles high-dimensional data effectively and provides reliable feature importance insights. Though DTs are simpler and more interpretable, RF offers superior accuracy, scalability, and resilience to missing data, making it the preferred choice for complex, data-intensive applications like biogas production optimization.

## 5. CONCLUSION

This study highlights the potential of ML in biogas production, a sustainable alternative to fossil fuels. By evaluating six ML models (LR, DT, RF, SVM, k-NN, and ANN), it identified Decision Trees (DT) and Random Forests (RF) as the most effective. With RF achieving an $R^2$ of 0.998 and DT at 0.999, the research highlights their superior predictive power over simpler regression models. Using a five-year dataset, it ensures robustness and real-world applicability. Additionally, DT and RF provide insights into key influencing factors, guiding optimization efforts. By comparing results with existing studies, these results suggest RF as a powerful tool for maximizing biogas yield and minimizing environmental impact. Future work recommends the use of AI-driven advancements which could focus on integrating real-time data, exploring hybrid ML models for improved accuracy, and developing user-friendly tools for broader adoption across biogas systems.

## REFRENCES

This work is acknowledged by the Rwanda Environment Management Authority (REMA) for executing the project titled “Building the capacity of Rwanda’s government to advance the National Adaptation Planning (NAP) process projects”.

## REFRENCES

[1] Moses Jeremiah Barasa Kabeyi, Oludolapo Akanni Olanrewaju, Joseph Akpan, Biogas Production and Process Control Improvements, From Biomass to Biobased Products. *Journal of Energy*,10.5772/intechopen.113061, (2024). https://doi.org/10.1155/2022/8750221.

[2] Singh AK, Pal P, Rathore SS, Sahoo UK, Sarangi PK, Prus P, Dziekański P. Sustainable Utilization of Biowaste Resources for Biogas Production to Meet Rural Bioenergy Requirements. *Energies*. 2023; 16(14):5409. https://doi.org/10.3390/en16145409.

[3] Mohammed Khaleel Jameel, Mohammed Ahmed Mustafa, Hassan Safi Ahmed, Amira jassim Mohammed, Hameed Ghazy, Maha Noori Shakir, Amran Mezher Lawas, Saad khudhur Mohammed, Ameer Hassan Idan, Zaid H. Mahmoud, Hamidreza Sayadi, Ehsan Kianfar, Biogas: Production, properties, applications, economic and challenges: A review,*Results in Chemistry*, Volume 7, 2024, 101549, ISSN 2211-7156, https://doi.org/10.1016/j.rechem.2024.101549.

[4] J. Ward, P. J. Hobbs, P. J. Holliman, and D. L. Jones, 'Optimisation of the anaerobic digestion of agricultural resources', *Bioresource Technology*, vol. 99, no. 17, pp. 7928–7940, Nov. 2008. Https://doi.org/10.1016/j.biortech.2008.02.044.

[5] G. Choi, H. Kim, and C. Lee, 'Long-term monitoring of a thermal hydrolysis-anaerobic co-digestion plant treating high-strength organic wastes: Process performance and microbial community dynamics', *Bioresource Technology*, vol. 319, p. 124138, Jan. 2021, https://doi.org/10.1016/j.biortech.2020.124138.

[6] Yadvika, Santosh, T. R. Sreekrishnan, S. Kohli, and V. Rana, 'Enhancement of biogas production from solid substrates using different techniques––a review', *Bioresource Technology*, vol. 95, no. 1, pp. 1–10, Oct. 2004, https://doi.org/10.1016/j.biortech.2004.02.010.

[7] Y. Qian, S. Sun, D. Ju, X. Shan, and X. Lu, 'Review of the state-of-the-art of biogas combustion mechanisms and applications in internal combustion engines', *Renewable and Sustainable Energy Reviews*, vol. 69, pp. 50–58, Mar. 2017, https://doi.org/10.1016/j.rser.2016.11.059.

[8] Andrade Cruz et al., 'Application of machine learning in anaerobic digestion: Perspectives and challenges', *Bioresource Technology*, vol. 345, p. 126433, Feb. 2022, https://doi.org/10.1016/j.biortech.2021.126433.

[9] P. Ghofrani-Isfahani, B. Valverde-Pérez, M. Alvarado-Morales, M. Shahrokhi, M. Vossoughi, and I. Angelidaki, 'Supervisory control of an anaerobic digester subject to drastic substrate changes', *Chemical Engineering Journal,* vol. 391, p. 123502, Jul. 2020, https://doi.org/10.1016/j.cej.2019.123502.

[10] Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, Christian Janiesch, Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability, *International Journal of Information Management*, Volume 69, 2023, 102538, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2022.102538.

[11] Long Chen, Pinjing He, Hua Zhang, Wei Peng, Junjie Qiu, Fan Lü, Applications of machine learning tools for biological treatment of organic wastes: Perspectives and challenges, *Circular Economy*, Volume 3, Issue 2, 2024, 100088, ISSN 2773-1677, https://doi.org/10.1016/j.cec.2024.100088.

[12] Zhang, Pengshuai, Tengyu Zhang, Jingxin Zhang, Huaiyou Liu, Cristhian Chicaiza-Ortiz, Jonathan TE Lee, Yiliang He, Yanjun Dai, and Yen Wah Tong. "A machine learning assisted prediction of potential biochar and its applications in anaerobic digestion for valuable chemicals and energy recovery from organic waste." *Carb Neutrality* 3, 2 (2024). https://doi.org/10.1007/s43979-023-00078-0.

[13] Portugal, P. Alencar, and D. Cowan, 'The use of machine learning algorithms in recommender systems: A systematic review', *Expert Systems with Applications,* vol. 97, pp. 205–227, May 2018, https://doi.org/10.1016/j.eswa.2017.12.020.

[14] G. S. Fanourgakis, K. Gkagkas, E. Tylianakis, and G. Froudakis, 'A Generic Machine Learning Algorithm for the Prediction of Gas Adsorption in Nanoporous Materials', *J. Phys. Chem. C*, vol. 124, no. 13, pp. 7117–7126, Apr. 2020, https://doi.org/10.1021/acs.jpcc.9b10766.

[15] Rossi E, Pecorini I, Iannelli R. Multilinear Regression Model for Biogas Production Prediction from Dry Anaerobic Digestion of OFMSW. *Sustainability*. 2022; 14(8):4393. https://doi.org/10.3390/su14084393.

[16] Wang, Zhengxin, Xinggan Peng, Ao Xia, Akeel A. Shah, Huchao Yan, Yun Huang, Xianqing Zhu, Xun Zhu, and Qiang Liao. "Comparison of machine learning methods for predicting the methane production from anaerobic digestion of lignocellulosic biomass." *Energy*, Volume 263, Part D, 2023, 125883, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2022.125883.

[17] Hunter W. Schroer and Craig L. Just. Feature Engineering and Supervised Machine Learning to Forecast Biogas Production during Municipal Anaerobic Co-Digestion. *ACS ES&T Engineering* 2024 4 (3), 660-672. Https://doi.org/10.1021/acsestengg.3c00435.

[18] Hannay K. 2020. Everything is a regression: in search of unifying paradigms in statistics. Available at https://towardsdatascience.com/everything-is-just-a-regression-5a3bf22c459c (accessed 12 Janvier 2025).

[19] Chicco Davide, Warrens Matthijs J., Jurman Giuseppe 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. *Peer J Computer science*., 2021, https://doi.org/10.7717/peerj-cs.623.

[20] Elicia Yee Ting Gan, Yi Jing Chan, Yoke Kin Wan, Timm Joyce Tiong, Woon Chan Chong, Jun Wei Lim. Examining the synergistic effects through machine learning prediction and optimization in the anaerobic Co-digestion (ACoD) of palm oil mill effluent (POME) and decanter cake (DC) with economic analysis, *Journal of Cleaner Production*, Volume 437, 2024, 140666, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2024.140666.

[21] Yang, Yong & Zheng, Shuaishuai & Ai, Zhilu & Molla Jafari, Mohammad Mahdi. On the Prediction of Biogas Production from Vegetables, Fruits, and Food Wastes by ANFIS- and LSSVM-Based Models. *BioMed Research International*. 2021. 1-8. Https://doi.org/10.1155/2021/9202127.

[22] Olatunji, Kehinde Oladoke & Ahmed, Noor & Madyira, Daniel & Adebayo, Ademola & Ogunkunle, Oyetola & Adeleke, Oluwatobi. Performance evaluation of ANFIS and RSM modeling in predicting biogas and methane yields from Arachis hypogea shells pretreated with size reduction. *Renewable Energy*, 189(1), 2022. Https://doi.org/10.1016/j.renene.2022.02.088.

[23] Daniel Jia Sheng Chong, Yi Jing Chan, Senthil Kumar Arumugasamy, Sara Kazemi Yazdi, Jun Wei Lim. Optimization and performance evaluation of response surface methodology (RSM), artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) in the prediction of biogas production from palm oil mill effluent (POME), *Energy,* Volume 266, 2023, 126449, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2022.126449.

[24] Gupta, R., Ouderji, Z.H., Uzma et al. Machine learning for sustainable organic waste treatment: a critical review. *npj Mater. Sustain*. 2, 5 (2024). https://doi.org/10.1038/s44296-024-00009-9.

[25] Duong, Cuong Manh, and Teng-Teeh Lim. "Use of regression models for development of a simple and effective biogas decision-support tool." *Scientific Reports* vol. 13,1 4933. 27 Mar. 2023, https://doi.org/10.1038/s41598-023-32121-6.

[26] Kyoungok Kim, Jung-sik Hong, A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis, *Pattern Recognition Letters*, Volume 98, 2017, Pages 39-45, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2017.08.011.

[27] Peng, Shurong, Lijuan Guo, Yuanshu Li, Haoyu Huang, Jiayi Peng, and Xiaoxu Liu. "Biogas Production Prediction Based on Feature Selection and Ensemble Learning." *Applied Sciences (2076-3417), Vol* 14, no. 2, p90, 2024, https://doi.org/10.3390/app14020901.

[28] David Akorede Akinpelu, Oluwaseun A. Adekoya, Peter Olusakin Oladoye, Chukwuma C. Ogbaga, Jude A. Okolie, Machine learning applications in biomass pyrolysis: From biorefinery to end-of-life product management, *Digital Chemical Engineering*, Volume 8, 2023, 100103, ISSN 2772-5081, https://doi.org/10.1016/j.dche.2023.100103.

[29] Abdul Hai, G. Bharath, Muhamad Fazly Abdul Patah, Wan Mohd Ashri Wan Daud, Rambabu K., PauLoke Show, Fawzi Banat, Machine learning models for the prediction of total yield and specific surface area of biochar derived from agricultural biomass by pyrolysis, *Environmental Technology & Innovation,* Volume 30, 2023, 103071, ISSN 2352-1864, https://doi.org/10.1016/j.eti.2023.103071.

[30] Halder, R.K., Uddin, M.N., Uddin, M.A. et al. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *J Big Data* 11, 113 (2024). https://doi.org/10.1186/s40537-024-00973-y.

[31] Chen W-Y, Chan YJ, Lim JW, Liew CS, Mohamad M, Ho C-D, Usman A, Lisak G, Hara H, Tan W-N. Artificial Neural Network (ANN) Modelling for Biogas Production in Pre-Commercialized Integrated Anaerobic-Aerobic Bioreactors (IAAB). *Water*. 2022; 14(9):1410. https://doi.org/10.3390/w14091410.

[32] Sarah M. Hunter, Edgar Blanco, Adiuan Borrion, predicting total biogas potential of food waste using the initial output of biogas potential tests as input data to train an artificial neural network, *Bioresource Technology Reports,* Volume 26, 2024, 101845, ISSN 2589-014X, https://doi.org/10.1016/j.biteb.2024.101845.

[33] Mukasine A, Sibomana L, Jayavel K, Nkurikiyeyezu K, Hitimana E. Maximizing Biogas Yield Using an Optimized Stacking Ensemble Machine Learning Approach. *Energies*. 2024; 17(2):364. https://doi.org/10.3390/en17020364.

[34] Organiściak, Patryk et al. "Machine Learning-Based Prediction of Biogas Production from Sludge Characteristics in Four Anaerobic Digesters: Development of the AD2Biogas Prediction Tool." *Advances in Science and Technology Research Journal,* vol. 18, no. 8, 2024, pp. 1-15. https://doi.org/10.12913/22998624/192936.

[35] Peng S, Guo L, Li Y, Huang H, Peng J, Liu X. Biogas Production Prediction Based on Feature Selection and Ensemble Learning. *Applied Sciences*. 2024; 14(2):901. https://doi.org/10.3390/app14020901.

[36] L. Wang, F. Long, W. Liao, and H. Liu, 'Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms', *Bioresource Technology*, vol. 298, p. 122495, Feb. 2020, https://doi.org/10.1016/j.biortech.2019.122495.

[37] C. Li, P. He, W. Peng, F. Lü, R. Du, and H. Zhang, 'Exploring available input variables for machine learning models to predict biogas production in industrial-scale biogas plants treating food waste', *Journal of Cleaner Production,* vol. 380, p. 135074, Dec.2022, https://doi.org/10.1016/j.jclepro.2022.135074.

[38] D. De Clerc, Z. Wen, F. Fei, L. Caicedo, K. Yuan, and R. Shang, 'Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion', *Science of The Total Environment,* vol. 712, p. 134574, Apr. 2020, https://doi.org/10.1016/j.scitotenv.2019.134574.

[39] F. Long, L. Wang, W. Cai, K. Lesnik, and H. Liu, 'Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data', *Water Research,* vol. 199, p. 117182, Jul. 2021, https://doi.org/10.1016/j.watres.2021.117182.

[40] H. Guo, S. Wu, Y. Tian, J. Zhang, and H. Liu, 'Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: A review', *Bioresource Technology*, vol. 319, p. 124114, Jan. 2021, https://doi.org/10.1016/j.biortech.2020.124114.

[41] 'Accurate prediction of chemical exergy of technical lignin's for exergy-based assessment on sustainable utilization processes | *Elsevier Enhanced Reader'*.https://reader.elsevier.com/reader/sd/pii/S0360544221032904?token=B2F5FF20AFDBFECF72548FA53DDEA1

F0C692 559EEAB441F8828BA941FF01C4C73AC400217DEED4D7BF62A6D5755D7A07&originRegion=eu-west-1&originCreation=20230209 160436 (accessed January 09, 2025).

[42] Z. Wang et al., 'Comparison of machine learning methods for predicting the methane production from anaerobic digestion of lignocellulosic biomass', *Energy,* vol. 263, p.125883, Jan. 2023, https://doi.org/10.1016/j.energy.2022.125883.

[43] S. Dreiseitl and L. Ohno-Machado, 'Logistic regression and artificial neural network classification models: a methodology review', *Journal of Biomedical Informatics,* vol. 35, no. 5, pp. 352–359, Oct. 2002, https://doi.org/10.1016/S1532-0464(03)00034-0.

[44] M. Kannangara, R. Dua, L. Ahmadi, and F. Bensebaa, 'Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches', *Waste Management*, vol. 74, pp. 3–15, Apr. 2018, https://doi.org/10.1016/j.wasman.2017.11.057.

[45] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, 'Random Forest for big data classification in the internet of things using optimal features', *Int. J. Mach. Learn. & Cyber*., vol. 10, no. 10, pp. 2609–2618, Oct. 2019, https://doi.org/10.1007/s13042-018-00916-z.

[46] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[47] S. S. Matin and S. C. Chelgani, 'Estimation of coal gross calorific value based on various analyses by random forest method', *Fuel*, vol. 177, pp. 274–278, Aug. 2016, https://doi.org/10.1016/j.fuel.2016.03.031.

[48] C. Cortes and V. Vapnik, 'Support-vector networks', *Mach Learn*, vol. 20, no. 3, pp.273–297, Sep. 1995, https://doi.org/10.1007/BF00994018.

[49] L. Liu and Y. Lei, 'An accurate ecological footprint analysis and prediction for Beijing based on SVM model', *Ecological Informatics*, vol. 44, pp. 33–42, Mar. 2018, https://doi.org/10.1016/j.ecoinf.2018.01.003.

[50] D. Subramanian, 'A Simple Introduction to K-Nearest Neighbors Algorithm', *Medium*, Jul. 12, 2021. https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e (accessed January 10, 2025).

[51] Chomboon, Kittipong, Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak Kerdprasop, and Nittaya Kerdprasop. "An empirical study of distance metrics for k-nearest neighbor algorithm." *In Proceedings of the 3rd International Conference on Industrial Application Engineering*, vol. 2, p. 4. 2015. https://doi.org/10.12792/iciae2015.051.

[52] Abdulhafedh, A. (2022) Comparison between Common Statistical Modeling Techniques Used in Research, Including Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. *Open Access Library Journal*, 9, 1-19. https://doi.org/10.4236/oalib.1108414.

[53] Abdul Hai, G. Bharath, Muhamad Fazly Abdul Patah, Wan Mohd Ashri Wan Daud, Rambabu K., PauLoke Show, Fawzi Banat, Machine learning models for the prediction of total yield and specific surface area of biochar derived from agricultural biomass by pyrolysis, *Environmental Technology & Innovation,* Volume 30, 2023, 103071, ISSN 2352-1864, https://doi.org/10.1016/j.eti.2023.103071.

[54] Huynh-Cam T-T, Chen L-S, Le H. Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. *Algorithms*. 2021; 14(11):318. https://doi.org/10.3390/a14110318.

[55] N. H. Khashaba, R. S. Ettouney, M. M. Abdelaal, F. H. Ashour, and M. A. El-Rifai, 'Artificial neural network modeling of biochar enhanced anaerobic sewage sludge digestion', *Journal of Environmental Chemical Engineering*, vol. 10, no. 4, p. 107988, Aug. 2022, https://doi.org/10.1016/j.jece.2022.107988.

[56] 'Coefficient of Determination', *Corporate Finance Institute*. https://corporatefinanceinstitute.com/resources/data-science/coefficient-of-determination. (accessed November 12, 2024).

[57] J. Frost, 'Mean Squared Error (MSE)', *Statistics By Jim*, Nov. 12, 2021. https://statisticsbyjim.com/regression/mean-squared-error-mse/ (accessed November 13, 2024).

[58] Ranjan Gaida, Gamunu L. Samarakoon Arachchige, Zahir Barahmand, Carlos Dimnarca. "Application of Machine Learning in Biogas Process", *FMH606 Master's Thesis,* 2023. https://openarchive.usn.no/usnxmlui/bitstream/handle/11250/3076266/no.usn%3Awiseflow%3A6838201%3A54569109.pdf?sequence=1&isAllowed=y. (accessed 12 November 2024).

[59] Q. Tang *et al.*, 'Machine learning prediction of pyrolytic gas yield and compositions with feature reduction methods: Effects of pyrolysis conditions and biomass characteristics', *Bioresource Technology*, vol.339, p. 125581, Nov. 2021, https://doi.org/10.1016/j.biortech.2021.125581.

[60] X. Zhu, Y. Li, and X. Wang, 'Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions', *Bioresource Technology*, vol. 288, p. 121527, Sep. 2019, https://doi.org/10.1016/j.biortech.2019.121527.