ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

Methane Yield Prediction for Anaerobic Digestion using Machine Learning Models

Sharan Aditya

Chemical Engineering, Valdel Engineers, Bengaluru, Karnataka, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150100097

Received: 06 February 2026; Accepted: 11 February 2026; Published: 17 February 2026

ABSTRACT

Anaerobic digestion (AD) systems exhibit complex, nonlinear interactions between operational parameters, making methane yield prediction challenging using traditional analytical methods. This study evaluates multiple machine learning (ML) techniques to model methane production from an industrial reactor dataset containing operational and environmental variables. A structured workflow incorporating correlation analysis, dimensionality reduction, clustering, and supervised learning was implemented to identify key predictors and assess model performance. Reactor temperature emerged as the dominant factor influencing methane yield, while most other variables showed weak direct correlations. Principal component analysis and K-means clustering revealed distinct operational regimes associated with different performance levels. Baseline linear regression models achieved moderate predictive accuracy ($R^2 \approx 0.5$), while nonlinear models—including ensemble methods and artificial neural networks—provided only marginal improvements, suggesting dataset limitations rather than algorithmic constraints. The results highlight the importance of richer biochemical and temporal data for improving predictive modelling and supporting data-driven optimization of industrial AD processes.

Keywords: Machine Learning, Anaerobic Digestion, Methane Yield, ANN, K-means cluster, Regression

INTRODUCTION

The UN Paris Agreement is an accord adopted in 2015 and ratified by 195 countries worldwide. Its goal is to decarbonise the planet, mitigate climate change, and reduce emissions intensity by 45% before 2030. To this aim, countries have been incentivising the adoption of programs that contribute to the effort of going green in energy generation and utilisation, curbing the use of carbon fuels, and cutting carbon emissions on a global scale. Energy businesses are offered major benefits to shift towards using clean fuel and farms for clean energy generation. Biofuels are a major component of available clean fuels in the market, accounting for 25% of the global energy supply in 2024, with ever-increasing demand [1]. Biofuel offers advantages over solar, wind, and other renewable energy resources: easy physical storage, a fixed, reliable rate of production and utilisation; an option for waste treatment and recycling, including limiting methane release as a greenhouse gas; and is low-cost to build.

Large-scale industrial processes to optimise the production of biogas from waste treatment plants have been researched extensively. A popular method for extracting methane is Anaerobic Digestion (AD). A great way to naturally process waste matter, it does not require high temperatures or large forms of energy to produce methane, is easily stored, and is a highly economical and high-yielding industrial process for methane extraction [2]. Extensive research in the field has accumulated a vast amount of data, and the scope of using machine learning models to optimise the operating conditions of this industrial process is in its early stages. Preliminary models using regression techniques, ANNs, and SVM have been used to evaluate and optimise operating conditions.

However, many concerns arise in ascertaining the efficacy of these models owing to overfitting, inconsistent feature engineering, and insufficient modelling [3]. Along with the difficulty of obtaining data, the practical application of these datasets is often overlooked because OEMs who prepare the reactor tanks that host the AD

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

process often optimise their tanks based on theoretical information and past data. The tanks cannot account for live tracking and optimisation of their performance. Since standardisation of feed composition is difficult, the yield percentage of methane cannot be standardised without actively studying live data from the reactor, and tank operators often only run the tank according to the manufacturer's manuals, leading to inconsistent results. This paper aims to focus solely on the model training aspect of the various studies conducted, including feature engineering, selection of relevant input variables, handling of multicollinearity, normalization and scaling strategies, cross-validation techniques, and hyperparameter optimisation. The discussion will emphasise how these methodological choices influence model generalizability and predictive performance, and will propose best practices for developing reliable and consistent data-driven tools to optimise the yield percentage of methane from AD.

Anaerobic Digestion: Principles, Benefits, and Process Overview

Anaerobic digestion (AD) is a well-established biotechnological process in which microorganisms convert organic matter into biogas in oxygen-deficient conditions. The biogas produced consists primarily of methane (CH_4) and carbon dioxide (CO_2), and can be used for heating, converted into electricity via combined heat and power (CHP) systems, or upgraded to biomethane for grid injection and transport fuel applications.

The AD process occurs through four sequential biochemical stages: hydrolysis, acidogenesis, acetogenesis, and methanogenesis. During hydrolysis, extracellular enzymes degrade complex macromolecules such as proteins, lipids, and polysaccharides into soluble monomers, including amino acids, fatty acids, and sugars. Acidogenesis converts these monomers into volatile fatty acids (VFAs) and CO_2, which are subsequently transformed into acetate and hydrogen during acetogenesis. In the final stage, methanogenesis, acetate and CO_2 (in the presence of hydrogen) are converted into CH_4. Among these stages, hydrolysis is frequently identified as the rate-limiting step in biogas production from recalcitrant biomass, thereby constraining the overall efficiency of the AD process [2].

The biochemical composition of food waste (FW) plays a critical role in determining methane yields and production kinetics in anaerobic digestion. FW can be broadly categorized into high protein and lipid content food waste (HPLFW) and high carbohydrate content food waste (HCFW). HPLFW - such as meat, dairy, and mixed food wastes—typically achieve significantly higher methane yields than carbohydrate-dominated feeds. Meat-dominant FW exhibits yield reaching approximately 337 mL CH_4/gCOD and conversion efficiencies above 96%, indicating high biodegradability. Dairy-dominant FW yields around 307 mL CH_4/gCOD with conversion efficiencies close to 88%, while mixed FW achieves approximately 297 mL CH_4/gCOD with conversion rates near 86%. These high yields are associated with balanced fermentation and methanogenesis rates. Lipids possess a high energy content and a theoretical methane potential exceeding 1,000 mL CH_4/gVS. However, excessive concentrations of long-chain fatty acids (LCFAs) can inhibit methanogenesis. In contrast, HCFW—such as fruit, vegetable, and grain wastes—generally produce lower methane yields. Vegetable-dominant FW yields approximately 238 mL CH_4/gCOD with conversion efficiencies around 68%, while grain-dominant FW yields only about 171 mL CH_4/gCOD with conversion efficiencies below 50%. The reduced performance is attributed to the inherently lower methane potential of carbohydrates and the slower degradation rates of fibrous carbohydrates [4]

Two primary factors lead to higher efficiencies in methane yield: FW pretreatment strategies and a high concentration of Anaerobic bacterial sludge maintained in the processing tanks. At least 10kg/m^3/day of concentrated sludge is required to meet the upkeep requirements for methane generation. Pretreatment strategies generally involve increasing the lipid content in FW to increase methane output generation at the hydrolysis stage. These methods include adding protein-rich substrates that possess a high theoretical methane potential of approximately 0.5 Nm^3/kg, but are susceptible to ammonia inhibition during degradation; Lipid-rich substrates can achieve higher biogas production with a theoretical methane potential of about 1.0 Nm^3/kg, but the high LCFA concentration can inhibit microbial activity; Lignocellulosic=rich substrates are often abundantly available but show variable degradability due to different fibre compositions, which hinder fatty acid generation and molecular breakdown during acidogenesis. Mixed substrates can exhibit vastly varying results due to their

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

highly variable composition, depending on diet, season, and cultural factors, with no clear benefit to the process[4]

Temperature management is a key operational factor in anaerobic digestion, with controlled systems typically operating in psychrophilic (10–20 °C), mesophilic (20–40 °C), or thermophilic (50–60 °C) ranges. Methane production rates generally increase with temperature, reaching optimal levels around 35–37 °C for mesophilic microorganisms and 55 °C or higher for thermophiles. Temperature shifts influence microbial community composition and metabolic activity, with increases from 15 °C to 35 °C often resulting in enhanced biochemical performance. While fluctuations can disrupt stability, anaerobic systems can recover following an acclimation period, allowing for controlled changes in operating regimes without irreversible loss of performance [2].

A promising alternative is the integration of electrodialysis into anaerobic digestion (ADED). This approach enables in-situ removal of ammonium ions while simultaneously enhancing process stability and methane production. ADED systems have demonstrated the ability to reduce ammonium concentrations from influent levels of up to 10,000 mg/L to below 2,000 mg/L. At moderate ammonia levels, methane production can increase by more than 40% compared to conventional digestion, and at extreme concentrations, ADED can prevent complete inhibition. The process also offers competitive energy efficiency, with consumption for ammonium recovery comparable to industrial ammonia synthesis. From an environmental standpoint, electrodialysis integration supports nutrient recovery, reduces the global warming potential of the process, sometimes to net-negative levels, and lowers acidification and eutrophication impacts, contributing to more sustainable waste management practices [5].

The economic feasibility of AD projects depends on multiple revenue and cost factors. For large-scale centralised plants, gate fees for waste intake remain an important income source, while the sale of electricity, heat, and co-products such as fertilizer enhances profitability. Compared to composting, anaerobic digestion offers a competitive advantage as a net energy producer rather than an energy consumer. The viability of projects is strongly influenced by supportive policies, including feed-in tariffs for renewable energy and restrictions on landfilling organic waste [6].

All these parameters controlling the output and efficiency of methane production have given rise to many machine learning models being used to study operating conditions, as their popularity has grown over the last decade. Automation and process monitoring have become increasingly important for optimising plant performance. With the increasing complexity of datasets required to measure the various input parameters of production units for methane yield, traditional methods fall short in improving process efficiency. ML models have the added advantage of better recognizing and predicting non-linear correlations between inputs, for example, finding links between operating temperatures and pressures with the FW composition, among other inputs. In essence, the more detailed the input data is from methane production units, the more efficient ML models become in predicting their behaviour. ML models are also much more efficient when processing datasets collated at an industrial scale, with tens of thousands of data points requiring much less computing power than traditional methods.

METHODOLOGY

The dataset used in this study contains a comprehensive set of process variables relevant to anaerobic digestion in a continuous stirred-tank reactor (CSTR). Key parameters include feed composition, nutrient solution usage, reactor and biogas temperature, system pressure variables, and waste vapor characteristics. Methane yield was designated as the target variable, representing the key performance indicator for process efficiency. The first stage of analysis involved a complete inventory of available features, followed by statistical correlation analyses to identify the most influential predictors of methane yield. The Pearson/Spearman correlation coefficient was employed to measure linear dependencies, and the mutual information was used to capture nonlinear relationships. Pearson correlations quantify the general covariance to assess the strength and direction of linear relationships between two continuous random variables, which can explain the target's dependencies on

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

individual weights, if present. The Spearman correlation measures the strength and direction of a monotonic relationship, whether linear or non-linear, giving a ranking to the Pearson correlation.

These methods ensured that both direct and complex dependencies between variables and methane yield are quantified. Additionally, exploratory data analysis (EDA) was conducted to examine the distribution, variance, and potential interactions among the input features [7].

Preprocessing of the dataset was performed to improve model performance. Continuous features were normalised using standard scaling to account for differences in measurement units. Feature engineering strategies were used to find improved correlations using a theoretical basis, such as creating derived ratios (e.g., methane yield per unit of substrate input) and interaction terms between the input variables[8], to improve the richness of the input space.

Machine learning model development was structured hierarchically. Baseline regression models, including linear regression and support vector regression (SVR), were initially applied to establish reference performance. [9]Subsequently, more advanced nonlinear models were implemented, including random forests and gradient boosting algorithms (XGBoost), to capture higher-order interactions and improve prediction accuracy. For potential scalability, artificial neural networks (ANNs) were also considered for modeling the complex relationships inherent in biochemical processes.

To ensure robustness, the random forest regressor was trained and evaluated using k-fold cross-validation. Model performance was assessed using multiple error metrics [7], including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), to provide a comprehensive view of predictive accuracy. Hyperparameter optimisation was performed using grid search and Bayesian optimization techniques, while regularization strategies and feature selection were employed to mitigate overfitting and enhance model generalizability.

The methodological framework developed in this study provides a systematic approach to replacing traditional empirical methods with machine learning-based predictive models. By integrating statistical analysis, feature engineering, and multi-model evaluation, this methodology is designed to maximise the efficiency of anaerobic digestion and provide actionable insights for process optimization in methane production.

Correlation Analysis

To investigate the relationships between process parameters and methane yield, Pearson's correlation coefficient (r) and Spearman's rank correlation coefficient (ρ) were computed between all numeric features and the target variable, Methane Standard Volume.

The results consistently highlight Reactor Temperature [°C] as the strongest predictor of methane yield, with a Pearson r of 0.704 and a Spearman ρ of 0.706, indicating a strong positive correlation that is both linear and monotonic. This suggests that higher reactor temperatures are strongly associated with increased methane production. We also see a strong negative correlation with the organic loading rate (OLR), calculated as (standard volume/reactor temperature), with a Pearson r of -0.577 and a Spearman ρ of -0.562.

Biogas Temperature [°C] also exhibited a positive association with methane yield, though weaker in magnitude ($r = 0.188$; $\rho = 0.296$), implying that, while it has an effect, it is less influential than reactor temperature. Other process variables, including Biogas Pressure (BP), Nutrient Solution Usage, and Feed, showed very weak correlations with methane yield, suggesting a limited direct role in influencing methane generation under the studied conditions. Variables such as BP, Waste Vapor Pressure, and Standard Volume were either weakly negative or near zero, further confirming their negligible direct effect on methane production.

The heatmap visualization of the Pearson correlation matrix (Figure X) further reinforces these findings, showing a strong correlation between methane yield and reactor temperature, while other parameters remain largely uncorrelated. To explore the influence of operating parameters on methane yield, scatter plots were generated

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

comparing methane yield against the remaining input parameters. The resulting visualizations reveal distinct patterns that inform the selection of predictive features for subsequent modelling. Figure 03 illustrates the pairwise relationships between methane standard volume and the other input variables. Across all scatter plots, methane yield exhibits a pronounced banded structure, indicating the presence of discrete operational regimes rather than a continuous response, with low, medium, and high yield states consistently observed. Variables such as BP, nutrient solution usage, standard volume, biogas pressure, and waste vapor pressure show weak monotonic trends and large within-band dispersion, suggesting limited direct influence on methane yield when considered independently. In contrast, reactor temperature and, to a lesser extent, biogas temperature show clearer structuring of the data, with higher temperature regimes more frequently associated with elevated methane production, consistent with established anaerobic digestion kinetics[10]. The appearance of distinct vertical alignments in several plots arises from discretised or fixed operating setpoints (e.g., reactor temperature, feed type, pressure), reflecting controlled experimental or synthetic design conditions. These discrete inputs reinforce the need for non-linear and tree-based machine learning models capable of capturing regime transitions and interaction effects rather than relying on purely linear formulations.

Reactor temperature was found to vary discretely across the dataset, with values centred around approximately 28°C, 42°C, and 55°C. Substantial variation in yield was observed within each band, indicating that reactor temperature alone does not fully determine methane output. This highlights the need to consider additional process variables in predictive modelling. Alternatively, biogas temperature shows a continuous distribution and a clearer monotonic relationship with methane yield.

As the biogas temperature increased, methane yield generally rose as well, though the relationship was not strictly linear [10]. The scatter plot suggests that the yield approaches a plateau at higher biogas temperatures, implying nonlinear or threshold effects. This variable, therefore, emerges as a critical determinant of yield, likely reflecting its influence on microbial activity and fermentation kinetics.

Biogas pressure showed only marginal variation, with values concentrated in a narrow band around 1.01–1.03 bar. Other parameters also show a similar lack of correlations with methane yield, suggesting that these variables are not significant in this process under the studied conditions. Reactor temperature as a categorical operating regime and biogas as a continuous non-linear driver will primarily influence the model behaviour in a non-linear study.

To study the distinct operational regimes seen in the scatter plot further, an unsupervised clustering approach was applied to the input features. Principal Component Analysis (PCA) was first used to reduce the high-dimensional operating data into two principal components (PC1 and PC2) that capture the majority of the variance. The transformed dataset was then subjected to K-means clustering with k = 3, based on prior evidence in the literature[7] suggesting the existence of three typical operating regimes in biogas reactors.

The clustering results (Fig 04) revealed three distinct groups of operating conditions. Cluster membership was subsequently compared against methane yield to assess whether specific clusters were associated with more favourable outcomes. A boxplot of methane yield distribution across the clusters (Figure Y) demonstrated clear

differences between groups. Cluster 0 exhibited an average methane yield of 1496 NmL d^{-1}, with a wide distribution spanning from low to moderately high values. Cluster 1 was characterized by the lowest methane yield, averaging 1125 NmL d^{-1}, and displayed a comparatively narrow range. Cluster 2 consistently showed the highest methane yield, with a mean of 1803 NmL d^{-1} and tighter variance, suggesting more stable operation under these conditions.

These findings indicate that the operating features of the reactor naturally segregate into regimes that align with performance differences. In particular, the high-yield cluster (Cluster 2) may represent a set of operating conditions optimal for methane production. Conversely, Cluster 1 corresponds to a low-yield regime that may reflect suboptimal or stressed operating states.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

Importantly, cluster labels can now be incorporated as an additional feature in predictive modelling of methane yield, or used to stratify the dataset for regime-specific analyses. This provides a pathway to integrate both data-driven pattern recognition and mechanistic process understanding.

Machine Learning Models

Baseline Linear Models

To establish a baseline, Ordinary Least Squares (OLS) regression and its regularized variants (Ridge and Lasso) were applied to the dataset. Model performance was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), mean average error (MAE), and mean squared error (MSE).

Ordinary Least Squares (OLS), Ridge, and Lasso regression yield comparable coefficients of determination ($R^2 \approx 0.51$) and high RMSE values (~443 NmL d^{-1}), suggesting that only about half of the variance in methane production is explained by linear combinations of the other process variables. The near-identical performance of OLS and its counterparts implies that multicollinearity is not the primary limitation and that regularization does not improve generalization. Instead, these results point to an inherently weak linear relationship between the available operational features and methane yield, consistent with the non-linear and threshold-driven nature of anaerobic digestion. Consequently, linear models serve as a useful baseline but are insufficient for accurate prediction, reinforcing the need for non-linear and ensemble-based approaches to capture the complex process dynamics.

The MAE values remain relatively high and comparable between training and test sets (~354–364 NmL d^{-1}), indicating moderate average prediction error and reinforcing that linear models capture only a limited portion of the variability in methane production. Similarly, the large and consistent MSE values (~$1.85–1.97 \times 10^5$) suggest the presence of substantial residual variance and occasional larger prediction deviations, supporting the earlier observation that linear regression provides only a baseline representation of the process dynamics.

The close agreement across all three methods demonstrates that the underlying relationships between process parameters and methane yield are largely non-linear under the studied conditions [11]. Linear models fail to capture subtle non-linear interactions that are known to influence biogas production dynamics. As such, these results serve as a strong baseline against which more flexible non-linear models (e.g., tree-based ensembles or kernel methods) can be evaluated.

Decision Tree Model

The tree-based models demonstrate a moderate improvement over linear regression, indicating that non-linear relationships are present, but are not sufficiently strong or well-resolved to enable high predictive accuracy. The Decision Tree regressor shows reasonable training performance ($R^2 \approx 0.59$) but a noticeable drop on the test set ($R^2 \approx 0.48$), accompanied by an increase in RMSE. This divergence suggests limited generalisation and a tendency toward overfitting to local patterns in the training data.

Feature importance analysis reveals a dominant dependence on reactor temperature, which accounts for the majority of variance explained by the model, while all other operational variables contribute marginally. This concentration of importance indicates that the model relies heavily on a single controlling variable rather than capturing broader multivariate process interactions.

The Random Forest model improves robustness relative to the single-tree approach, achieving higher training performance ($R^2 \approx 0.76$) and slightly improved test performance ($R^2 \approx 0.51$). However, the persistence of a gap between training and test metrics, along with a cross-validation mean R^2 close to the test value, suggests that the ensemble primarily stabilizes variance rather than uncovering substantially new predictive structure. Similarly, the Gradient Boosting model exhibits strong training performance ($R^2 \approx 0.75$) but does not translate this advantage to the test set, where performance remains comparable to the Random Forest and only marginally

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

better than linear baselines. The higher training accuracy compared to testing again points to partial overfitting and limited data in the feature space.

Overall, while ensemble methods outperform linear and single-tree models, their gains are modest, indicating that the current feature set does not capture the dynamics of the governing methane production in its entirety [12]. The dominance of temperature-related variables across all models reinforces their central role in anaerobic digestion performance, but the inability of more sophisticated models to achieve substantial predictive improvement suggests that key biochemical, microbial, or temporal descriptors are missing. These results highlight the need for richer feature engineering and possibly dynamic or time-resolved inputs to meaningfully enhance data-driven optimization of anaerobic digestion systems.

SVR and KNN Performance

The Support Vector Regression (SVR) model exhibits limited predictive capability for methane yield, with training and testing R^2 values of approximately 0.48 and 0.43, respectively. The relatively close alignment between train, test, and cross-validation performance indicates stable generalisation but also suggests that the model is unable to extract a strong nonlinear structure from the available features. The high RMSE values further confirm that prediction errors remain substantial, implying that the kernel-based mapping does not sufficiently capture the underlying process complexity or that the feature space lacks the resolution required for effective margin-based regression.

In contrast, the k-Nearest Neighbours (KNN) model demonstrates extreme overfitting, achieving perfect training performance ($R^2 = 1.00$) with zero error, while generalisation performance deteriorates markedly on the test set ($R^2 \approx 0.45$). This behaviour is characteristic of instance-based learning methods in relatively noisy or sparsely informative datasets, where local neighbourhoods in the training data do not represent the broader input space. Overall, these results reinforce the conclusion that while non-linear models are necessary, models that rely heavily on local interpolation or fixed kernel mappings are insufficient, further emphasizing the need for richer process descriptors or hybrid modelling strategies.

ANN

The artificial neural network implemented as a multi-layer perceptron (MLP) demonstrates moderate predictive capability, achieving a training R^2 of 0.58 and a test R^2 of 0.51. The model was run with 128 hidden layers to improve performance and avoid overfitting, resulting in a relatively small gap between training and testing performance, particularly compared to more flexible models such as KNN. However, the magnitude of the RMSE on both sets indicates that prediction errors remain substantial, reflecting the limited explanatory power of the available input features.

These results indicate that while the ANN can capture some non-linear relationships in the anaerobic digestion process, its performance does not markedly exceed that of ensemble tree-based methods [13]. This points towards the dataset's coverage being rather limited as opposed to the model's inability to scope the dataset. In the absence of additional biochemical, microbial, or time-dependent features, even flexible learning architectures such as neural networks cannot fully resolve the complex dynamics governing methane production.

CONCLUSION

This study demonstrates the potential—and current limitations—of applying machine learning models to predict methane yield in anaerobic digestion systems (Table 01) using a static, process-level dataset. Across linear, tree-based, kernel-based, and neural network models, predictive performance consistently plateaued at a test R^2 of approximately 0.5, with RMSE values remaining relatively high. While non-linear and ensemble models modestly outperformed linear baselines, none achieved a substantial improvement in generalisation, indicating that model choice was not the primary bottleneck. Correlation analysis, clustering, and feature importance results consistently identified temperature-related variables as the dominant predictors, reinforcing established anaerobic digestion kinetics. However, the strong reliance on a small subset of features, coupled with the limited

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

contribution of other operational variables, suggests that the dataset does not adequately capture the full biochemical and microbial complexity governing methane production.

The dataset provided by OEMs to improve the efficiency of methane yield from their CSTR tanks sees a primary correlation only with reactor temperature. From an operational standpoint, OEMs should prioritise tight thermal management within the optimal mesophilic or thermophilic window observed in the dataset, ensuring minimal fluctuations and avoiding frequent regime switching [14]. Secondary process inputs such as nutrient dosing or feed variations should be maintained within stable baseline ranges rather than aggressively tuned unless supported by additional biochemical monitoring.

The clustered operational patterns further suggest that tanks may benefit from clearly defined operating regimes rather than broad exploratory ranges, with temperature serving as the primary control variable, and other parameters acting as stabilising factors. However, since the dataset lacks dynamic biological indicators and feedstock composition variability, these recommendations should be interpreted as operational guidance based on historical performance trends rather than universal process optimization rules.

ML Models	Training R^2	Test R^2	Test RMSE	Test MAE	Test MSE ($*10^5$)
OLS, ridge, Lasso Regressions	0.53	0.51	443	364.36	19.6
Decision Tree	0.58	0.47	460.4	369.55	21.2
Random Forest	0.75	0.50	446.9	357.56	19.9
Gradient Boosting	0.74	0.49	450.82	361.26	20.3
SVR	0.48	0.43	480.23	337.31	23.1
KNN	1.00	0.45	471.93	375.95	22.3
ANN	0.59	0.5	450.24	359.71	20.2

The principal limitation of this work lies in the lack of data richness rather than methodological rigor. The dataset is constrained by discretized operating regimes, the absence of time-resolved dynamics, limited representation of feedstock composition variability, and no direct descriptors of microbial population, inhibition effects, or transient process behavior. As a result, even highly expressive models such as ANNs fail to uncover additional structure beyond what is already encoded in temperature-driven regimes [15]. These findings underscore that meaningful gains in predictive accuracy will likely require integrating longitudinal data, real-time sensor measurements, and mechanistically informed features rather than further algorithmic complexity alone. Future work should focus on hybrid approaches that combine process knowledge with data-driven models, as well as on the generation and sharing of standardized, high-resolution industrial datasets, which are essential for translating machine learning from academic feasibility studies into practical tools for anaerobic digestion optimization.

BIBLIOGRAPHY

1. “Statistical Review of World Energy - American Gas Association.” Accessed: Jan. 17, 2026. [Online]. Available: https://www.aga.org/research-policy/resource-library/statistical-review-of-world-energy/?utm_source=chatgpt.com
2. T. Z. D. De Mes, A. J. M. Stams, J. H. Reith, and G. Zeeman, “Methane production by anaerobic digestion of wastewater and solid wastes.”
3. H. Rutland, J. You, H. Liu, L. Bull, and D. Reynolds, “A Systematic Review of Machine-Learning Solutions in Anaerobic Digestion,” Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/bioengineering10121410.
4. T. M. Anacleto *et al.*, “Methane yield response to pretreatment is dependent on substrate chemical composition: a meta-analysis on anaerobic digestion systems,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-51603-9.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

5. J. Meng, L. Shi, Y. Hu, Z. Wang, Z. Hu, and X. Zhan, "Integration of anaerobic digestion and electrodialysis for methane yield promotion and in-situ ammonium recovery," *Bioresour. Technol.*, vol. 402, Jun. 2024, doi: 10.1016/j.biortech.2024.130770.
6. A. H. Bhatt and L. Tao, "Economic perspectives of biogas production via anaerobic digestion," *Bioengineering*, vol. 7, no. 3, pp. 1–19, Sep. 2020, doi: 10.3390/bioengineering7030074.
7. S. Ö. Cinar, S. Cinar, and K. Kuchta, "Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process," *Fermentation*, vol. 8, no. 2, Feb. 2022, doi: 10.3390/fermentation8020065.
8. R. Llanos-Lizcano, L. Senila, and O. C. Modoi, "Evaluation of Biochemical Methane Potential and Kinetics of Organic Waste Streams for Enhanced Biogas Production," *Agronomy*, vol. 14, no. 11, Nov. 2024, doi: 10.3390/agronomy14112546.
9. S. Mokraoui, A. Halilu, M. A. Hashim, and M. K. Hadj-Kali, "Modeling and simulation of biomass anaerobic digestion for high biogas yield and CO2 mineralization," *Mater. Renew. Sustain. Energy*, vol. 12, no. 2, pp. 105–116, Aug. 2023, doi: 10.1007/s40243-023-00233-8.
10. E. Piercy, X. Sun, P. R. Ellis, M. Taylor, and M. Guo, "Temporal Dynamics of Microbial Communities in Anaerobic Digestion: Influence of Temperature and Feedstock Composition on Reactor Performance and Stability."
11. M. Mohammadianroshanfekr, M. Pazoki, M. B. Pejman, R. Ghasemzadeh, and A. Pazoki, "Kinetic modeling and optimization of biogas production from food waste and cow manure co-digestion," *Results in Engineering*, vol. 24, Dec. 2024, doi: 10.1016/j.rineng.2024.103477.
12. O. J. Sinayobye *et al.*, "Optimizing Biogas Production with Machine Learning: A Comparative Study of Predictive Models," 2024. [Online]. Available: https://www.jisem-journal.com/
13. D. A. Samuel, B. I. Eziefula, G. Orkuma, and A. Usman, "Forecasting of Biogas and Biomethane Outputs from Anaerobic Co-digestion Using Multilayer Perceptron Artificial Neural Networks (MLP-ANN)," *International Journal of Multidisciplinary Approach Research and Science*, vol. 3, no. 02, pp. 561–568, May 2025, doi: 10.59653/ijmars.v3i02.1529.
14. T. Z. D. De Mes, A. J. M. Stams, J. H. Reith, and G. Zeeman, "Methane production by anaerobic digestion of wastewater and solid wastes."
15. "Artificial Neural networks for predicting methane content in biogas from livestock waste".