

Article

Not peer-reviewed version

Biogas Prediction Enhancement of a Swine Farm Bio-Digester Using a Lag-Based Surrogate Machine Learning Model

M. E. Montes-Carmona , [I. A. Burgos-Castro](#) ^{*} , [R. de J. Portillo-Vélez](#) , [P. J. García-Ramírez](#) , [L. F. Marín-Urías](#) , [M. A. Hernández-Pérez](#)

Posted Date: 1 April 2026

doi: 10.20944/preprints202603.2505.v1

Keywords: biogas production; machine learning; surrogate model; digital twin; bio-digester; biogas prediction; methane production



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Biogas Prediction Enhancement of a Swine Farm Bio-Digester Using a Lag-Based Surrogate Machine Learning Model

M. E. Montes-Carmona ¹, I. A. Burgos-Castro ^{2,*}, R. de J. Portillo-Vélez ^{2,3},
P. J. García-Ramírez ¹, L. F. Marín-Urías ^{2,3} and M. A. Hernandez-Perez ¹

¹ Instituto de Ingeniería, Universidad Veracruzana, Boca del Río, Veracruz, México
² Facultad de Ingeniería de la Construcción y el Hábitat, Universidad Veracruzana, Boca del Río, Veracruz, México
³ Facultad de Ingeniería Eléctrica y Electrónica, Universidad Veracruzana, Boca del Río, Veracruz, México
* Correspondence: ian_498@hotmail.com

Abstract

Biogas production estimation has been one of the most important and challenging objectives for anaerobic digestion processes due to the complexity of its dynamics and the lack of high-quality open-access datasets. This study presents a hybrid modeling framework that combines a mechanistic model, based on ordinary differential equations (ODE), with a machine learning model. Rather than relying exclusively on experimental data, the proposed approach leverages physics-informed synthetic data generation, complemented by a lag-based feature engineering to capture inherent temporal dependencies in the process dynamics available in operational data of a bio-digester. Two configurations were evaluated: a baseline model and an enhanced version incorporating lag features and simplified temperature profile. While the improved model achieved high predictive performance ($R^2 = 0.97885$, $RMSE = 131.80[L/d]$), additional analyses reveal that this performance is partly driven by temporal memory and remains sensitive to noise and feature composition. Instead of presenting the model as a final solution, this work frames it as a step toward practical digital twin implementations, acknowledging the gap that still exists between simulation-based accuracy and real-world reliability.

Keywords: biogas production; machine learning; surrogate model; digital twin; bio-digester; biogas prediction; methane production

1. Introduction

The biogas generation process is a natural phenomenon that occurs in a variety of anaerobic environments, [1]. These include marine and freshwater environments, sediments, wastewater treatment plant sludge, etc. Interest in the process stems primarily from the following reasons. On the one hand, a high degree of organic matter reduction is achieved with a small increase—compared to aerobic processes—in bacterial biomass. On the other hand, biogas production can be used to generate various forms of energy (heat and electricity) or processed as automotive fuel. Nevertheless, biogas has a lower calorific value than natural gas, and in specific applications, such as automotive fuel, treatment is necessary to improve its quality, [2].

Energy production from the anaerobic digestion has been used worldwide for over 30 years, [3]. Its viability and profitability depend not only on the amount of biogas produced, the available technology, and the efficiency of the wastewater treatment operation, but also on external parameters such as the local cost of energy production and available energy resources, [4]. Besides the economic advantages, biogas has also yielded environmental benefits, [5]. Anaerobic digestion represents a mature yet still evolving technology for renewable energy production. Predicting methane yield remains challenging due strong nonlinearities and environmental coupling. Anaerobic digestion is a biological process where organic carbon is converted through subsequent oxidation and reduction

to its more oxidized state (CO_2) and its more reduced state (CH_4). A wide range of microorganisms catalyzes the process in the absence of oxygen. Nitrogen, hydrogen, ammonia, and hydrogen sulfide are also generated in smaller quantities (typically less than 1% of the total gas volume). The mixture of gaseous products is called biogas, and the anaerobic degradation process is often also referred to as the biogas process, [6]. Estimating biogas production in an anaerobic digester is a complicated problem because the system is not only chemical nor only biological, but a microbial ecosystem highly related to biochemical reactions, mass transfer and variable physicochemical conditions, [7]. Current society requirements demands high efficiency biogas production to be implemented in anaerobic digesters of wastewater treatment plants, [8]. Therefore, innovative solutions must be developed in the short term, [9].

Recent advances in machine learning, particularly deep learning architectures, have shown promising results in modeling temporal dynamics [10]. Nevertheless, some open problems are still challenging. In [11], the optimization of biogas production for treating domestic wastewater is studied using different machine learning models (XGBoost and PSO), pointing out that limitations on the volume of the training data influenced the performance of the model predictions. J. Schulz et. al. [12] investigated the features of Carbon to Nitrogen ratio of the substrates that can be used for long-term continuous Anaerobic co-digestion. In general, the study of sensors calibration and fouling, modeling and optimization approaches together with the lack of high quality standardized datasets are current trends in the literature, [13]. Some solutions using frontier technologies such as digital twins, [14] might help to shorten the implementation curve provided AI-based models are accurate and simple enough to be implemented in real-time biogas production processes.

Herein we aim to improve biogas estimation using lag-based training vectors. To avoid the lack of accurate and enough datasets, in this work, we propose a biogas estimation surrogate model based on machine learning. Unlike prior studies, our approach explicitly addresses data scarcity through synthetic data generation, temporal dependencies via lag-based modeling, as well as mechanistic consistency through ODE constraints. The model was trained using a set of data obtained from the solution of a set of differential equations with experimentally validated parameters. Our analysis indicates that biogas estimation can be achieved with more than 97% of accuracy. The rest of the work is structured as follows. Section 2 describes the proposed methodology while Section 3 depicts the corresponding results to validate our approach. Section 4 presents a brief discussion and, finally Section 5 closes the paper with conclusions and future work.

2. Methodology

2.1. Experimental Design

To systematically evaluate the predictive capacity of the proposed surrogate model for the swine farm bio-digester, in this case methane is the produced biogas, the methodological framework was structured into two progressive experiments:

- **Base model:** 30,000 registers were generated using different stoichiometric and kinetic parameters based on the real bio-digester by means of the ODEs (1), (2), (3), (4) and features in terms of inflow and the organic substrate concentration S_1 to produce biogas calculated through the equation (5). An Extreme Gradient (XGBoost) model was trained and tested using a sequential 80/20 split, respectively.
- **Lag-based improved model:** 10,000 registers were generated incorporating a dynamic seasonal temperature factor to perturb the microbial kinetic rates and simulate real-world environmental exposure. To capture this physical complexity, integrating thermodynamic interaction variables and historical lags of the biogas production. The inclusion of this auto-regressive feature is valid given that this information is available in real world applications, it is calculated through equation (5). An XGBoost model was trained and tested using a sequential 80/20 split to evaluate the system's inertial memory and prevent data leakage.

2.2. Mathematical Modeling and Mass Balance

The physical and biochemical dynamics of the anaerobic digestion process (acidogenic and methanogenic phases) were simulated using a system of ODEs (1), (2), (3), (4). Given that the real rural bio-digester operates with a plug-flow hydrodynamic regime, the total reactor volume (V_{total}) was spatially discretized into a series of Continuous Stirred-Tank Reactors (CSTRs). For this study, the system was divided into $N = 3$ interconnected sub-reactors.

The general mass balance for any given state variable within the i -th sub-reactor is consistent with the fundamental principle of accumulation (Accumulation = Inflow - Outflow + Net Reaction) which is described by the following system of ODEs:

$$\frac{dX_{1,i}}{dt} = D_{sub}(X_{1,i-1} - X_{1,i}) + \mu_{1,i}X_{1,i} \quad (1)$$

$$\frac{dX_{2,i}}{dt} = D_{sub}(X_{2,i-1} - X_{2,i}) + \mu_{2,i}X_{2,i} \quad (2)$$

$$\frac{dS_{1,i}}{dt} = D_{sub}(S_{1,i-1} - S_{1,i}) - K_1\mu_{1,i}X_{1,i} \quad (3)$$

$$\frac{dS_{2,i}}{dt} = D_{sub}(S_{2,i-1} - S_{2,i}) + K_2\mu_{1,i}X_{1,i} - K_3\mu_{2,i}X_{2,i} \quad (4)$$

where $X_{1,i}$ and $X_{2,i}$ for $i = 1, 2, \dots, N$ denote the acidogenic and methanogenic biomass concentrations, respectively, for the i -th biodigester. S_1 represents the organic substrate concentration measured as Chemical Oxygen Demand (COD), and S_2 defines the Volatile Fatty Acids (VFA) concentration. D_{sub} is the local dilution rate for each sub-reactor, calculated as $D \times N$ (The global dilution D is defined as the flow rate Q_{in} divided by the total volume V_{total}). This formulation implies a reduction in the effective residence time within each sub-reactor, consistent with the representation of reactors in series.

For the first sub-reactor ($i=1$), the feed concentrations correspond to the raw inputs ($S_{1,0}=S_{1in}$ and $S_{2,0}=S_{2in}$), assuming a biomass-free feed stream ($X_{1,0}$ and $X_{2,0} = 0$).

Methane production is calculated using equation (5)

$$CH_4 = \sum_{i=1}^N K_4\mu_{2,i}X_{2,i}V_{sub} \quad (5)$$

2.3. Microbial Kinetics and Thermodynamic Perturbation

Microbial growth kinetics were modeled using the Monod equation for the acidogenic phase (6) and the Haldane inhibition model for methanogenesis (7). To improve the performance of the surrogate model, two scenarios were formulated. The base model analysis considered standard kinetic rates without environmental perturbations as expressed in equation (6) and equation (7).

$$\mu_{1,i} = \mu_{max1} \left(\frac{S_{1,i}}{K_{s1} + S_{1,i}} \right) \quad (6)$$

$$\mu_{2,i} = \mu_{max2} \left(\frac{S_{2,i}}{K_{s2} + S_{2,i} + \frac{S_{2,i}^2}{K_I}} \right) \quad (7)$$

where μ_{max} represents the maximum specific growth rates under standard conditions, K_{si} for $i = \{1, 2\}$ denotes the half-saturation constants, and K_I is the inhibition constant due to Volatile Fatty Acids (VFA) accumulation.

The lag-based improved model incorporated a seasonal temperature factor to simulate real-world environmental exposure and its direct impact on bacterial growth rates. It is important to note that, although the model is simplified, this representation captures the dominant seasonal dynamics while

providing a controlled and physically consistent perturbation framework for training and evaluating the surrogate model. The modified kinetic equations are shown in equation (8) and equation (9):

$$\mu_{1,i} = (\mu_{max1} \cdot T_{factor}) \left(\frac{S_{1,i}}{K_{s1} + S_{1,i}} \right) \quad (8)$$

$$\mu_{2,i} = (\mu_{max2} \cdot T_{factor}) \left(\frac{S_{2,i}}{K_{s2} + S_{2,i} + \frac{S_{2,i}^2}{K_I}} \right) \quad (9)$$

To model the annual environmental exposure of rural bio-digesters, the T_{factor} was defined by a cyclical model over a 365-day period as shown in equation (10):

$$T_{factor}(t) = 1.0 + 0.15 \sin\left(\frac{2\pi t}{365}\right) \quad (10)$$

Finally, the total daily methane production (CH_4), representing the system's target energy yield, was calculated using equation (5) where V_{sub} is the volume of each individual sub-reactor (V_{total}/N), and K_4 is the calibrated stoichiometric yield coefficient for methane.

2.4. Synthetic Data Generation

To train the XGBoost model, extended operational periods were simulated by numerically solving the ODE's system using the `scipy.integrate.odeint` library in Python. To reflect the real operation, information from the records of the real bio-digester were considered using bounded normal distributions (empirical operation ranges), the numerical integration required defining an initial condition vector ($t = 0$) for each sub-reactor, representing the starting biomass and substrate concentrations inside the digester. Across all sub-reactors, the initial state was set to: $X_1 = 1.8$ g/L, $X_2 = 0.8$ g/L, $S_1 = 1500.0$ mg/L, and $S_2 = 10.0$ mmol/L.

- **Input flow rate (Q_{in}):** Modeled as $\mathcal{N}(4.5, 1.0)$ and physically constrained to the range [1.0, 8.0] m³/d.
- **Organic substrate loading (S_{1in}):** Modeled as $\mathcal{N}(2500, 500)$ and physically constrained to the range [1000, 4000] mg/L.

Through this computational approach, two distinct datasets were generated and exported for the machine learning pipeline:

- **Base model:** A dataset with 30,000 records with the exported feature space consisted of Simulation Time (Time), Input Flow Rate ($Q_{in_m3_d}$), Organic Substrate Loading ($S1_in_mg_L$), and the target variable Methane Biogas Production ($CH4_Prod_L_d$).
- **Lag-based improved model:** A dataset with 10,000 records with the exported feature space included the thermodynamic perturbation, consisting of Simulation Time (Time), Input Flow Rate ($Q_{in_m3_d}$), Organic Substrate Loading ($S1_in_mg_L$), the Seasonal Temperature Factor ($Temp_Factor$), and the target variable methane biogas Production ($CH4_Prod_L_d$).

2.5. Feature Engineering and Machine Learning Architecture

To ensure model stability and eliminate the influence of the initial unsteady-state mathematical start-up, a transient period was discarded from the synthetic datasets prior to the feature engineering phase (the first 50 days for the base model and the first 100 days for the lag-based improved model). To capture the system's temporal dynamics and provide the predictive models with historical memory, specific systematic lags vectors were engineered consistent with the available information in operational bio-digester. Lag intervals of $\tau = \{1, 5, 10, 15, 20, 25, 30\}$ days and $\tau = \{1, 2, 3, 5, 10, 15, 20\}$ days were incorporated into the feature space for the base model and lag-based improved model, respectively.

For the base model scenario, the input feature vector $\mathbf{x}_t$ constructed for any given day t resulted in a 16-dimensional array, organized as shown in equation (11):

$$\mathbf{x}_t = [S_{1in}(t), Q_{in}(t), S_{1in}(t - \tau_1), \dots, S_{1in}(t - \tau_7), Q_{in}(t - \tau_1), \dots, Q_{in}(t - \tau_7)]^T \quad (11)$$

where the lags set is defined as $\tau \in \{1, 5, 10, 15, 20, 25, 30\}$ days.

However, for the lag-based improved model, the T_{factor} and its physical interactions with the mass flows were considered. Additionally, the records of historical biogas production (CH_4) were considered. The inclusion of lagged methane production variables is consistent with real-world operational scenarios, where historical biogas measurements are continuously monitored and readily available. In this context, these variables provide valuable temporal information that enhances short-term predictive performance. Therefore, this information allows the model to exploit inherent temporal dependencies of the process, enhancing predictive capability within a realistic digital twin framework.

The thermodynamic interaction variables were defined as $I_{S1}(t) = T_{factor}(t) \cdot S_{1in}(t)$ and $I_Q(t) = T_{factor}(t) \cdot Q_{in}(t)$. Based on the refined lags set $\tau \in \{1, 2, 3, 5, 10, 15, 20\}$ days, the augmented input feature vector $\mathbf{x}_t$ expanded into a 40-dimensional array, organized as shown in equation (12):

$$\begin{aligned} \mathbf{x}_t = [& S_{1in}(t), Q_{in}(t), T_{factor}(t), I_{S1}(t), I_Q(t), \\ & S_{1in}(t - \tau_1), \dots, S_{1in}(t - \tau_7), \\ & Q_{in}(t - \tau_1), \dots, Q_{in}(t - \tau_7), \\ & T_{factor}(t - \tau_1), \dots, T_{factor}(t - \tau_7), \\ & I_{S1}(t - \tau_1), \dots, I_{S1}(t - \tau_7), \\ & CH_4(t - \tau_1), \dots, CH_4(t - \tau_7)]^T \end{aligned} \quad (12)$$

The XGBoost model was selected because of its robustness against overfitting and its capacity to handle non-linear interactions without explicit scaling.

2.6. Model Training and Evaluation Metrics

To prevent data leakage and rigorously evaluate predictive performance, a sequential chronological split was implemented for both experiments. The first 80% of the temporal registers were used for model training, while the remaining 20% were reserved as an unseen testing set.

The hyperparameters of the XGBoost regressor were tuned for each experiment to balance learning capacity and generalization:

- **Base model configuration:** The model used 1000 boosting rounds (estimators), a learning rate of 0.05, and a maximum tree depth of 5. Subsampling and column sampling by tree were both set to 0.80. Early stopping was triggered if the validation loss did not improve for 50 consecutive rounds.
- **Lag-based improved model configuration:** The model used 5000 estimators with a learning rate of 0.01 and a maximum tree depth of 12. Subsampling and column samplings were set to 0.70, with an early stopping patience of 100 rounds.

The performance of the model was assessed using the coefficient of determination (R^2), the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE), which are defined in equations (13), (14), and (15), respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (14)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{15}$$

where n represents the total number of observations in the testing set, y_i is the ground-truth biogas production generated by the phenomenological model, $\hat{y}_i$ is the predicted biogas production by the XGBoost model, and $\bar{y}$ is the mean of the actual observed values.

3. Results

3.1. Synthetic Data Generation

Dataset generation described in section 2 was consistent with kinetic and stoichiometric parameters of the real swine bio-digester based on the previous mathematical approximation developed by Cardona [15]. Table 1 summarizes the parameters utilized for the simulation, highlighting the specific adjustments made to the baseline parameters to prevent computational instability and accurately reflect real-world biogas production which is often necessary according to the state of the art [13], [16].

Table 1. Kinetic, stoichiometric, and operational parameters utilized for the ODE numerical simulation.

Parameter	Description	Value	Reference
K_1	Organic matter consumption yield	1.30 g/g	[15]
K_2	VFA generation yield	1.06 mmol/g	[15]
K_3	VFA consumption yield	6.30 mmol/g	[15]
μ_{max1}	Maximum specific growth rate (acidogenic)	1.70 d ⁻¹	[15]
μ_{max2}	Maximum specific growth rate (methanogenic)	0.84 d ⁻¹	[15]
K_{s2}	Saturation constant for VFA	12.0 mmol/L	[15]
V_{total}	Total bio-digester volume	61.0 m ³	[15]
$N_{reactors}$	Number of CSTRs in series	3	[15]
K_4	Methane yield coefficient	0.18 mmol/g	Calibrated ¹
K_{s1}	Saturation constant for organic matter	6000 mg/L	Calibrated ²
K_I	Haldane inhibition constant	150 mmol/L	Calibrated ³

¹ Increased from the baseline 0.11 to reflect the real-world operational average of 4600 L/d. ² Adjusted from 20,000 mg/L to balance bacterial consumption and prevent rapid system acidification. ³ Increased from 50 mmol/L to enhance methanogenic resilience against VFA toxicity ensuring stable oscillatory production.

3.2. Performance of the Base Model

The first XGBoost model was trained using the 16 features described in equation (11) which contains current and lagged values of Q_{in} and S_{1in} . After training and testing procedures, the model achieved a Coefficient of Determination (R^2) of 0.6875, a Root Mean Square Error (RMSE) of 480.02 L/d, and a Mean Absolute Error (MAE) of 381.19 L/d.

The developed surrogate model approximates the real synthetic biogas production, oscillating stably between 4,000 and 10,000 L/d. This behavior closely matches the real physical system which reported averages of 4.6 m³/d with peaks from 6 to 8 m³/d as shown in Figures 1 and 2.

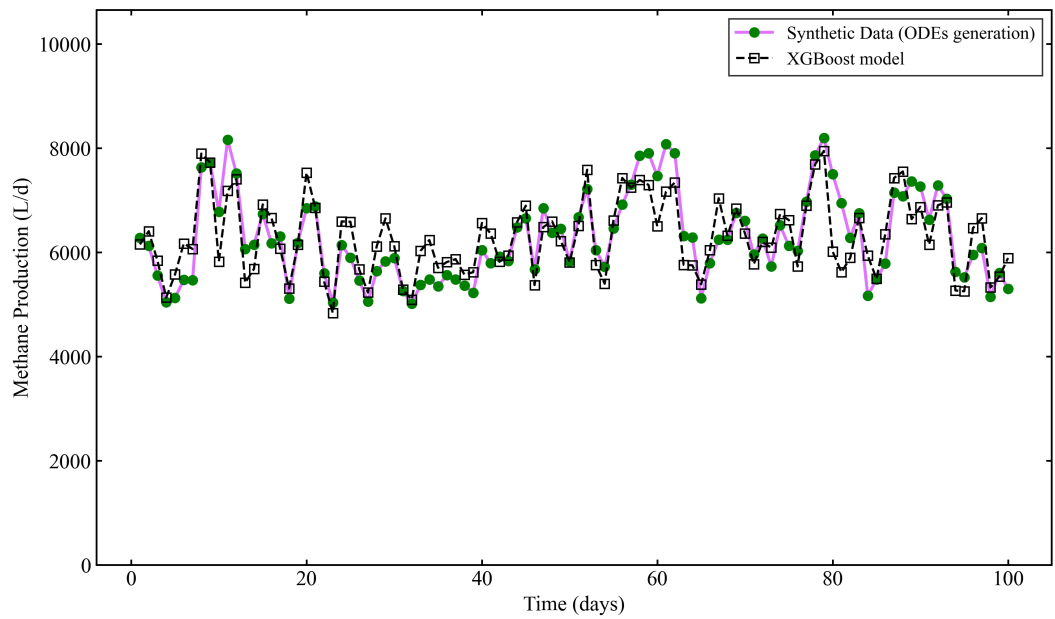


Figure 1. Performance of the base XGBoost model for 100 days. Comparison between the synthetic data generated by the ODEs and the surrogate model’s prediction ($R^2 = 0.6875$).

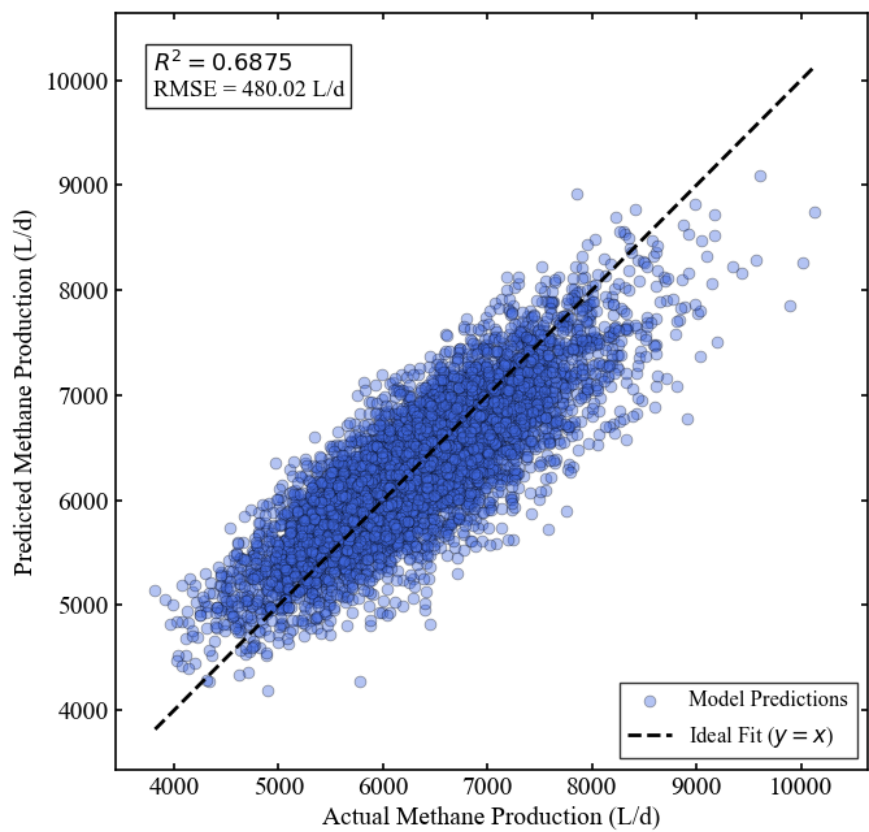


Figure 2. Actual Methane Production vs Predicted value (Base Model). Comparison between current methane production vs model’s prediction ($R^2 = 0.6875$).

3.3. Performance of the Lag-based Improved Model

To improve the performance of the prediction, the lag-based model considered the biogas registers of the target variable CH_4 as detailed in section 2 included within the 40 features. After training and testing, the model showed a significant improvement in predictive accuracy achieving an R^2 score of

0.9788, an RMSE of 131.80 L/d and a MAE of 85.48 L/d, the performance of the model in 100 days is shown in Figures 3 and 4.

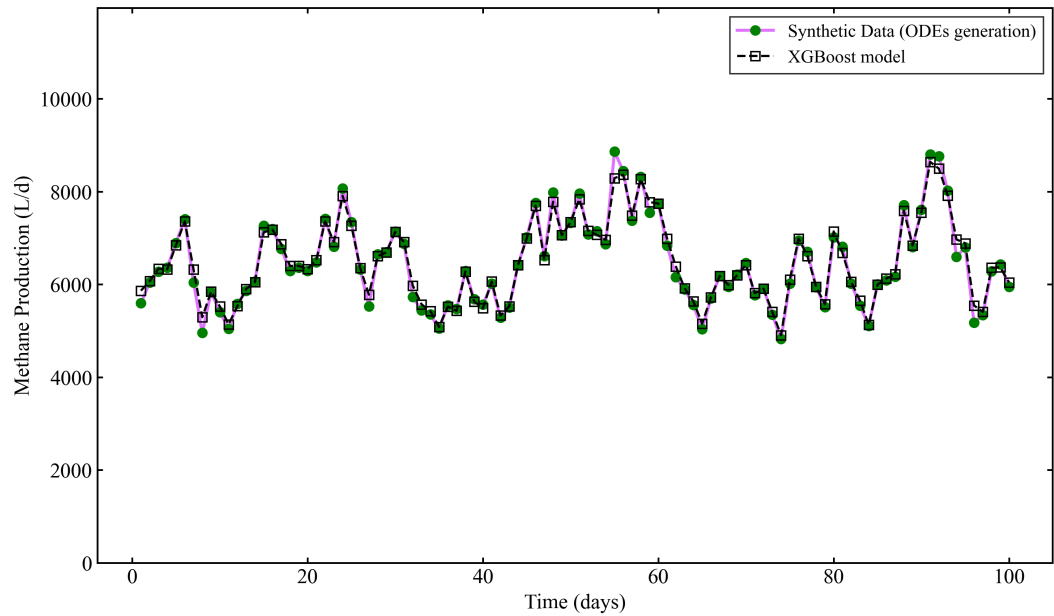


Figure 3. Performance of XGBoost model (Lag-base Improved Model). Comparison between the synthetic data generated by the ODEs and the surrogate model’s prediction ($R^2 = 0.9788$).

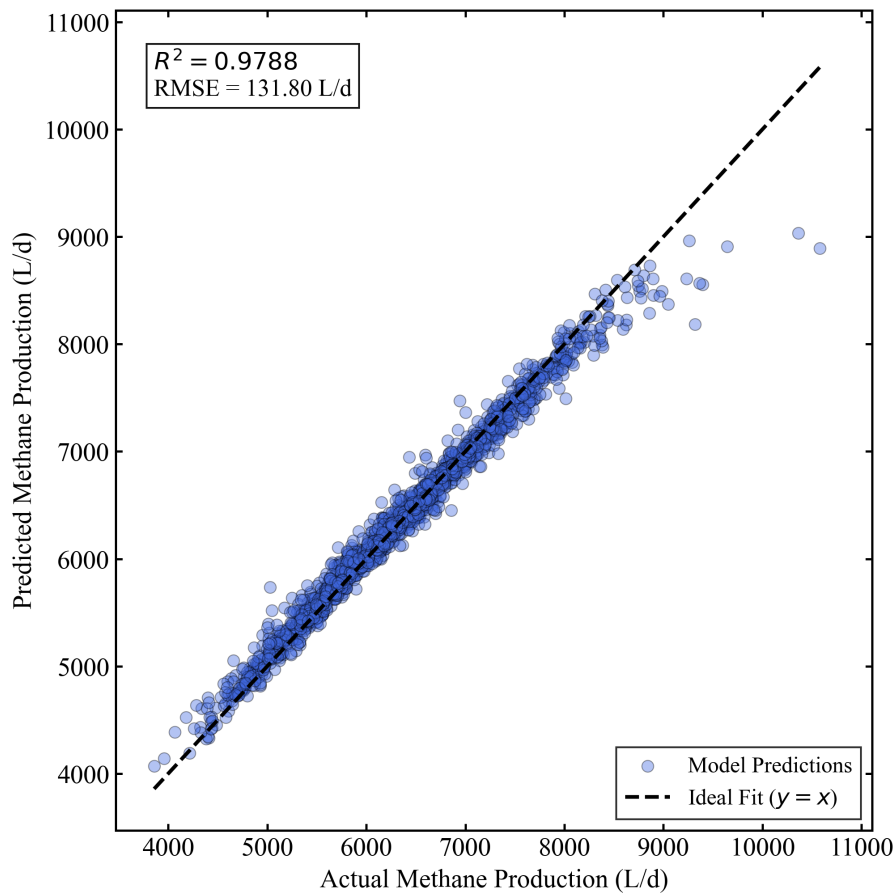


Figure 4. Actual Methane Production vs Predicted value (Lag-base Improved Model). Comparison between current methane production vs model’s prediction ($R^2 = 0.9788$).

The model’s ability to maintain high fidelity tracking of methane production under dynamic seasonal perturbations demonstrates that the lag components effectively act as a "microbial memory" making it highly suitable for biogas prediction in rural anaerobic digestion processes. A comprehensive comparison of the predictive performance metrics for both models (base model and lag-based improved model) is summarized in Table 2.

Table 2. Comparison of the evaluation metrics for the machine learning models developed in base model analysis and lag-based improved model.

Analysis	Model	R^2	RMSE (L/d)	MAE (L/d)
Base Model	XGBoost (16 features)	0.6875	480.02	381.19
Lag-based Model	XGBoost (40 features)	0.9788	131.80	85.48

3.3.1. Feature Importance and Microbial Memory

To understand the contribution of the every predictor variable, a feature importance analysis was conducted as shown in Figure 5 (Only top 15 feature were plotted for clarity). The influence of each variable was evaluated using the Normalized Gain metric, which quantifies the relative contribution of each feature to the reduction of the model’s prediction error across all decision trees in the XGBoost.

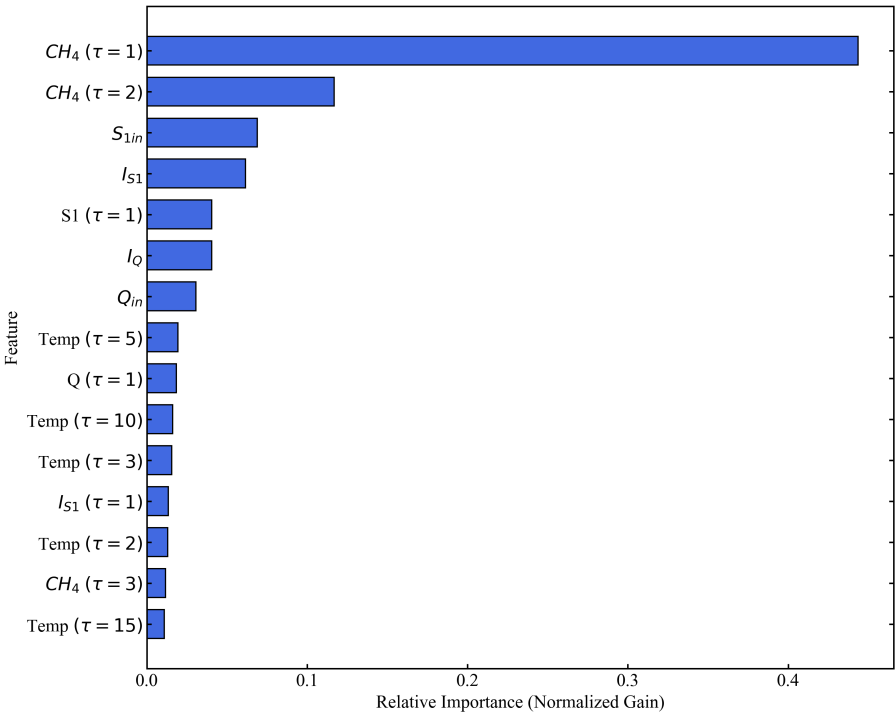


Figure 5. Top 15 Feature Importance for the Lag-based Improved Model (XGBoost) evaluated by Normalized Gain.

The results revealed the relative contribution and the top 6 features values were expressed in the vector **G** for clearly visualization, directly mapped to the its associated feature in the vector x_t . Equation (16) synthesized the information.

$$\mathbf{G} = \begin{bmatrix} 44.36 \\ 11.68 \\ 6.90 \\ 6.15 \\ 4.04 \\ 4.04 \end{bmatrix} \% \quad \text{corresponding to} \quad \mathbf{x}_t = \begin{bmatrix} CH_4(t-1) \\ CH_4(t-2) \\ S_{1in}(t) \\ I_{S1}(t) \\ S_{1in}(t-1) \\ I_Q(t) \end{bmatrix} \tag{16}$$

3.3.2. Five-Fold Cross-Validation

To rigorously validate the lag-based improved model, a 5-fold cross-validation adapted for time series [17] was evaluated confirming the high predictive accuracy with an average $R^2 = 0.9740$ and an RMSE = 140.03 L/d, consistent with the standard 80/20 split validation ($R^2 = 0.9788$ and RMSE = 131.80 L/d). Table 3 details the predictive performance across each fold.

Table 3. Time Series Cross-Validation (5 Folds) performance metrics for the Lag-based Improved Model.

Fold	R^2	RMSE (L/d)
1	0.9550	190.90
2	0.9774	130.79
3	0.9797	122.00
4	0.9796	121.08
5	0.9784	135.36
Average	0.9740 ± 0.0095	140.03 ± 26.00

4. Discussion

The presented results demonstrate the efficacy of a machine learning-based surrogate model for predicting biogas production in anaerobic digesters. In the first analysis, the predictive performance of the base model in terms of R^2 was limited even by increasing the number of registers to train the model (100,000 registers, $R^2=0.6949$) as expressed in the sensitivity analysis in Figure 6, which confirms the higher performance of the lag-base model (with only 10,000 registers achieved an $R^2=0.9788$), in addition, when implemented with 100,000 samples the improved model produced an $R^2=0.9939$.

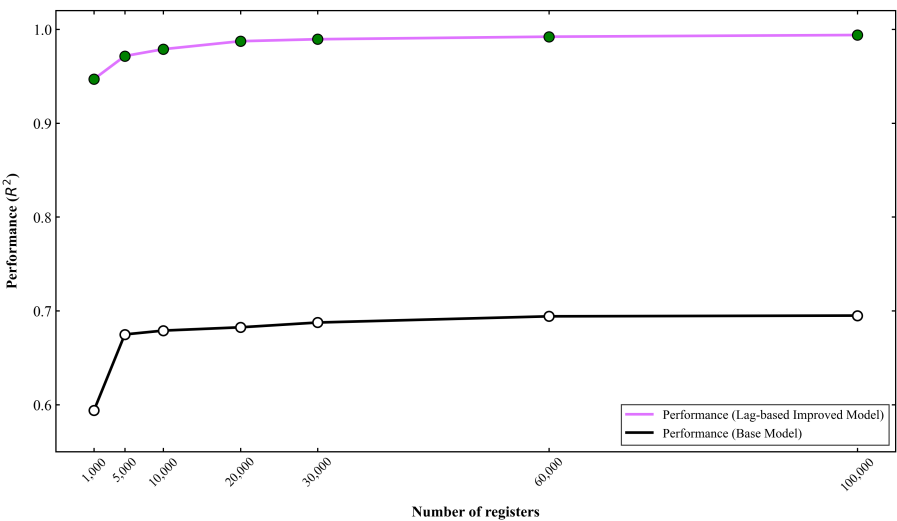


Figure 6. Performance of the models (R^2) vs Number of registers

To evaluate the model’s reliability under unpredictable real-world weather, stochastic noise was added to the temperature profile. As shown in Table 4, even with severe 20% daily temperature variations, the lag-based model maintained high predictive accuracy ($R^2 = 0.9599$). This minimal

drop in performance demonstrates that the model is highly robust and avoids overfitting to idealized conditions. Ultimately, it proves that the incorporated lag features successfully capture the system’s historical memory, reflecting the natural thermal inertia of real biogas plants.

Table 4. Evaluation of the lag-based model’s robustness under stochastic temperature variations.

Temperature Noise Level	R^2	RMSE (L/d)
0% (Deterministic control)	0.9788	131.80
10% Stochastic noise	0.9682	169.50
20% Stochastic noise	0.9599	180.90

This contribution is highly valuable as it addresses a current challenge in the application of artificial intelligence to anaerobic digestion: the lack of open-access training databases, as recently highlighted in the state of the art [13]. By generating and utilizing the two synthetic datasets described in this study, our methodological approach effectively addressed this limitation.

Furthermore, this methodology has the potential to optimize biogas production by allowing the computational tuning of operational parameters prior to physical implementation, which translates into significant cost reductions. Another key contribution is the model’s applicability to physically implemented bio-digesters. By integrating real-world sensor data to create a surrogate model (Digital Twin), operators can evaluate the system’s response to dynamic variations such as fluctuating organic load inputs or substrate concentrations without risking the biological stability of the physical system. Limitations of the proposed approach are identified as the capability to estimate more complex biological dynamics.

Finally, while the initial base model performed acceptably, the lag-based architecture demonstrated superior reliability and predictive accuracy by incorporating historical biogas production data, a variable that is typically monitored and readily available in field operations. This was confirmed by the importance of the feature analysis where the immediately preceding biogas production day is alone responsible for 44.36% of the model’s overall predictive accuracy, the remaining percentage of the predictive precision is distributed among the other 39 model’s predictors. This result highlights the relevance of recent historical measurements in capturing the temporal dynamics of anaerobic digestion processes, particularly for short-term forecasting scenarios in operational environment. The additional validation scheme through time-series cross-validation (folds = 5) confirmed the high predictive accuracy maintaining the $R^2 = 0.97$ presented in the 80/20 split validation.

5. Conclusions

On the one hand, biogas estimation is a challenging problem due to the complex models involved in the dynamics of the biochemical process. On the other hand, the lack of, quality and quantity enough, data availability for biogas generation processes makes it difficult to train and validate machine learning models to be useful in most applications. In the first stage of our proposal, the methodology renders an option to consider differential equation models to generate synthetic data considering the kinetic, stoichiometric and operational parameters of a state of the art real-world implemented bio-digester to address the lack of high quality open-access datasets. Moreover, in the second stage, two machine learning-based models that reflect the operational behavior of the physical system (according to the biogas production) were obtained. While the base model performed acceptably, a 29% performance improvement was achieved by properly including a historical memory by designing specific lag vectors with the available information from the bio-digester operational data representing the main contribution of this paper. Sensitivity analysis and uncertainty analysis under data quantity and noise in temperature profile was performed, respectively, validating the robustness of our findings. Both of the machine learning models were developed in terms of operational parameters such as the organic substrate loading (S_{1in}) and the input flow rate (Q_{in}). This approach facilitates the implementation of a digital twin, allowing operators to troubleshoot the system virtually before applying changes to the

physical biodigester, which represents our primary direction for future work as well as considering temperature perturbations in a more realistic strategy rather than the sinusoidal wave with noise, and using the approach for other biogas production processes in wastewater treatment plants to validate its implementation feasibility in such scenarios or considering modeling more complex models, as microorganisms' biological dynamics.

Author Contributions: Conceptualization, M.E.M.C. and R.J.P.V.; methodology, I.A.B.C. and R.J.P.V.; software, I.A.B.C.; validation, I.A.B.C, M.E.M.C. and L.F.M.U.; investigation, R.J.P.V., M.A.H.P. and M.E.M.C.; resources, P.J.G.R. and M.E.M.C.; data curation, M.E.M.C. M.A.H.P. and I.A.B.C.; writing—original draft preparation, I.A.B.C., M.E.M.C, R.J.P.V. L.F.M.U. M.A.H.P. and P.J.G.R. ; writing—review and editing, P.J.G.R., R.J.P.V., M.A.H.P. and L.F.M.U.; visualization, I.A.B.C. and L.F.M.U.; supervision, P.J.G.R, M.E.M.C and R.J.P.V.; project administration, M.E.M.C, M.A.H.P. and P.J.G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: Data generated in this study is available at <https://github.com/mecatronico-consultor/biogas-prediction>

Acknowledgments: Second author acknowledges support from SECIHTI-Mexico through scholarship number 4030171.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tabatabaei, M.; Ghanavati, H. Biogas. *Springer International Publishing* **2018**.
2. Ashokkumar, V.; Kumar, G.; Lakshmanan, H.; Chandramughi, V.; Flora, G.; Kothari, R.; Piechota, G. A critical review of biogas production and upgrading from organic wastes: Recent advances, challenges and opportunities. *Biomass and Bioenergy* **2025**, *194*, 107566.
3. Nayeri, D.; Mohammadi, P.; Bashardoust, P.; Eshtiaghi, N. A comprehensive review on the recent development of anaerobic sludge digestions: Performance, mechanism, operational factors, and future challenges. *Results in Engineering* **2024**, *22*, 102292. <https://doi.org/10.1016/j.rineng.2024.102292>.
4. Wang, Z.; Liu, Y.; Zhang, A.; Liu, Z.; Gai, H. A review of process development, mechanistic insights, and enhancement technologies for anaerobic digestion in industrial wastewater treatment. *Journal of Environmental Chemical Engineering* **2025**, *13*, 118217. <https://doi.org/10.1016/j.jece.2025.118217>.
5. Simeonov, I.; Chorukova, E.; Kabaivanova, L. Two-stage anaerobic digestion for green energy production: A review. *Processes* **2025**, *13*, 294.
6. Gavala, H.N.; Angelidaki, I.; Ahring, B.K. Kinetics and modeling of anaerobic digestion process. *Biomethanation I* **2003**, pp. 57–93.
7. Farid, M.U.; Olbert, I.A.; Bück, A.; Ghafoor, A.; Wu, G. CFD modelling and simulation of anaerobic digestion reactors for energy generation from organic wastes: A comprehensive review. *Heliyon* **2025**, *11*.
8. Lucas, D.; Oliveira, P.; Bessa, A.; Marcondes, F.S.; Rodrigues, M. Towards Efficient Biogas Production: Deep Learning-Based Methane Forecasting in Anaerobic Digesters of Wastewater Treatment Plants. In *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems*. Springer, 2025, pp. 154–165.
9. Shamshad, J.; Rehman, R.U. Innovative approaches to sustainable wastewater treatment: a comprehensive exploration of conventional and emerging technologies. *Environmental Science: Advances* **2025**, *4*, 189–222.
10. Ling, J.Y.X.; Chan, Y.J.; Chen, J.W.; Chong, D.J.S.; Tan, A.L.L.; Arumugasamy, S.K.; Lau, P.L. Machine learning methods for the modelling and optimisation of biogas production from anaerobic digestion: a review. *Environmental Science and Pollution Research* **2024**, *31*, 19085–19104.
11. Kumar, S.; Kumar, S.; Kumar, D.R.; Sharma, D.; Wipulanusat, W. Machine learning-based optimization of biogas and methane yields in UASB reactors for treating domestic wastewater. *Biodegradation* **2025**, *36*, 55.
12. Schultz, J.; Scherzinger, M.; Elbanhawy, A.Y.; Kaltschmitt, M. Long-term continuous anaerobic co-digestion of residual biomass—model validation and model-based Investigation of different carbon-to-nitrogen ratios. *BioEnergy Research* **2025**, *18*, 58.
13. Marycz, M.; Turowska, I.; Glazik, S.; Jasiński, P. Artificial Intelligence in Anaerobic Digestion: A Review of Sensors, Modeling Approaches, and Optimization Strategies. *Sensors* **2025**, *25*, 6961.

14. Kim, M.; Ghobadi, F.; Tayerani Charmchi, A.S.; Lee, M.; Lee, J. Digital Twins for Clean Energy Systems: A State-of-the-Art Review of Applications, Integrated Technologies, and Key Challenges. *Sustainability* **2025**, *18*, 43.
15. Cardona Acuña, L.D. Development of an approximate mathematical model for rural biodigesters (*Desarrollo de un modelo matemático aproximado para biodigestores rurales*). Master thesis in spanish, Universidad de Ibagué, Colombia. Available at: <https://hdl.handle.net/20.500.12313/3983>, 2021.
16. Cheng, M.; Zhao, X.; Dhimish, M.; Qiu, W.; Niu, S. A Review of Data-Driven Surrogate Models for Design Optimization of Electric Motors. *IEEE Transactions on Transportation Electrification* **2024**, *10*, 8413–8431. <https://doi.org/10.1109/TTE.2024.3366417>.
17. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis* **2018**, *120*, 70–83. <https://doi.org/10.1016/j.csda.2017.11.003>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.