engineering proceedings

MDPI

Proceeding Paper

A Gradient Boosting Regression Model for Early-Stage Screening of Total Biogas Yield from Organic Substrates †

Zafar Turakulov [1,*], Jasmina Elmurodova [1], Jaloliddin Eshbobaev [1], Noilakhon Yakubova [2], Elyor Samadov [3], Golib Nazarov [4], Khasan Niyozov [4] and Mirolim Aripov [4]

[1] Department of Automation and Digital Control, Tashkent Institute of Chemical Technology, Tashkent 100011, Uzbekistan; eshbobayevjalol@gmail.com (J.E.)
[2] Department of Control System and Information Processing, Tashkent State Technical University, Tashkent 100095, Uzbekistan; yakubova.noila@gmail.com
[3] Department of Industrial Automation, Tashkent State Technical University, Tashkent 100095, Uzbekistan; elyor0689@gmail.com
[4] Department of Biotechnology, Tashkent Institute of Chemical Technology, Tashkent 100011, Uzbekistan; nazarovgolib05@gmail.com (G.N.); mirolim.oripov@yahoo.com (M.A.)
* Correspondence: webdastur@gmail.com
† Presented at the 3rd International Online Conference on Energies (IOCEN 2026), 12–15 May 2026; Available online: https://sciforum.net/event/IOCEN2026.

Abstract

Early-stage estimation of total biogas yield is useful for preliminary screening of anaerobic digestion substrates when detailed kinetic parameters, long-term reactor data and substrate-specific biochemical methane potential tests are not yet available. This study develops a compact Gradient Boosting Regression model for estimating total biogas yield from basic substrate and operating descriptors commonly reported in anaerobic digestion studies. A literature-informed reconstructed benchmark dataset containing 120 records from twelve organic substrate groups was used. The response variable was total biogas yield, expressed as mL g^{-1} volatile solids (VS) added; methane yield and methane content were not used as interchangeable target variables. Seven input variables were considered: volatile solids content, total solids content, carbon-to-nitrogen ratio, residence time descriptor, digestion temperature, pH and organic loading rate. The model was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), five-fold cross-validation and residual diagnostics. On the test subset, the model achieved R^2 = 0.9885, MAE = 13.40 mL g^{-1} VS, RMSE = 16.16 mL g^{-1} VS and MAPE = 3.73%. The mean five-fold cross-validation score was R^2 = 0.9720 ± 0.0180, indicating stable interpolation performance within the investigated data domain. Feature-importance analysis identified C/N ratio as the dominant predictor, followed by volatile solids and total solids. However, residual diagnostics showed heteroscedasticity, and the low apparent importance of pH and temperature was attributed to their narrow ranges in the compiled dataset, rather than to biological irrelevance. The proposed model is therefore intended as an early-stage screening tool for total biogas yield estimation, not as a universal methane yield or design-level prediction model.

Keywords: biogas production; anaerobic digestion; machine learning; Gradient Boosting Regression; organic waste; early-stage screening

Check for updates

Academic Editor: Anastassios Stamatelos

Published: 19 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Anaerobic digestion is one of the most widely applied biological processes for the conversion of organic waste into renewable energy [1]. It enables the production of biogas while contributing to waste stabilization, nutrient recovery and reduction in uncontrolled methane emissions from organic residues. The practical performance of anaerobic digestion depends on the combined effect of substrate composition, reactor operating conditions and microbial activity [2,3]. In particular, volatile solids content, total solids content, carbon-to-nitrogen ratio, organic loading rate, hydraulic retention time, temperature and pH are among the most frequently reported variables affecting biogas yield [2].

Reliable early-stage estimation of total biogas yield is especially important during the early stages of project development. At this stage, engineers and decision makers usually need to compare alternative feedstocks, evaluate preliminary feasibility and estimate expected energy recovery [4,5]. However, detailed kinetic parameters and long-term operational datasets are often unavailable. Conventional empirical correlations are simple and useful for approximate estimation, but they may not adequately represent nonlinear interactions among substrate and process variables. Mechanistic models, such as detailed anaerobic digestion models, provide better process interpretation but require extensive calibration data and are not always convenient for rapid preliminary assessment [6].

Machine learning offers a data-driven alternative for this type of early-stage estimation problem [6,7]. Regression-based machine-learning models can capture nonlinear relationships between commonly available input variables and biogas yield without requiring a complete mechanistic description of the anaerobic digestion process. Nevertheless, in bioprocess applications, model accuracy alone is not sufficient [8]. A model with high R^2 may still be limited by overfitting, small sample size, biased substrate representation or weak generalizability outside the training domain. Therefore, model diagnostics, cross-validation, residual analysis and feature-importance interpretation are required to assess whether the model is practically useful [6].

The objective of this study was to develop and evaluate a compact machine-learning regression model for early-stage screening of total biogas yield from organic substrates. The study focused on a limited set of input variables that are commonly available in the literature and preliminary feasibility studies. The model was evaluated using train–test performance, five-fold cross-validation, residual diagnostics and feature-importance analysis. Particular attention was given to the risk of overfitting and limited generalizability because small literature-informed reconstructed benchmark datasets may produce high apparent accuracy when similar substrate groups are represented in both training and testing subsets. Therefore, the model was interpreted primarily as an interpolation-based screening tool, rather than as a replacement for experimental or mechanistic anaerobic digestion assessment.

The novelty of this study is not the general use of machine learning in anaerobic digestion, since previous studies have already applied neural networks, soft sensors, fuzzy logic, IoT-based systems and other data-driven models for methane yield estimation, process monitoring and reactor optimization [4–7,9–12]. The specific contribution of the present work is the development and diagnostic evaluation of a compact Gradient Boosting Regression model for early-stage screening of total biogas yield using a limited number of descriptors that are commonly available during preliminary feasibility assessment. Unlike studies focused on methane yield, gas quality or online process control, this work focuses on total biogas yield estimation within a harmonized substrate dataset and explicitly evaluates model reliability through train–test analysis, five-fold cross-validation, residual normality, heteroscedasticity assessment and feature-importance interpretation. Recent studies on biochemical and fermentation-related processes have also shown that nonlinear

https://doi.org/10.3390/engproc2026147015

biological systems require structured modelling, control strategy selection and diagnostic interpretation [13,14]. However, those studies mainly address process control and dynamic fermentation modelling, rather than early-stage estimation of total biogas yield from substrate descriptors. Therefore, the present model is positioned as an interpolation-based screening tool, rather than a replacement for biochemical methane potential testing or mechanistic anaerobic digestion modelling.

2. Methodology

2.1. Dataset Description, Harmonization and Response-Variable Definition

A literature-informed reconstructed benchmark dataset containing 120 records was used for model development and evaluation. The dataset was structured to support early-stage screening of total biogas yield from commonly reported substrate and operating descriptors. Published anaerobic digestion studies were used to define realistic substrate categories, variable ranges, operating condition ranges and total biogas yield ranges [1,2,9,15–22]. The individual numerical records were reconstructed within these literature-informed ranges and were not extracted verbatim as row-level experimental observations from specific publications. The final dataset included twelve substrate groups: food waste, cattle manure, sewage sludge, corn stover, wheat straw, co-digestion mixtures, algae, fruit and vegetable waste, organic fraction of municipal solid waste, paper waste, restaurant waste and dairy wastewater. The numerical records should be interpreted as literature-informed benchmark records, rather than as verbatim row-level observations extracted from individual publications.

The response variable in this study was total biogas yield, expressed as mL g^{-1} volatile solids (VS) added. Methane yield, methane concentration and biochemical methane potential values were not treated as interchangeable target variables, because total biogas volume and methane production represent different performance indicators [17–19]. Therefore, the model output should be interpreted as total gas yield estimation, rather than direct methane energy prediction. If methane fraction or methane yield is required for energy assessment, additional gas composition data must be used.

The input variables were selected according to their availability in anaerobic digestion studies and their relevance for preliminary substrate assessment. Seven predictors were used: volatile solids content, total solids content, carbon-to-nitrogen ratio, residence time descriptor, digestion temperature, pH and organic loading rate. Volatile solids and total solids were expressed as percentages, C/N ratio was dimensionless, temperature was expressed in °C, pH was dimensionless and organic loading rate was expressed as g VS L^{-1} d^{-1}. Total biogas yield was harmonized to mL g^{-1} VS added. Where reported yields were given as L kg^{-1} VS, they were converted using the equivalent relationship 1 L kg^{-1} VS = 1 mL g^{-1} VS.

Because residence time terminology differs among anaerobic digestion systems, the time-related variable was treated as a residence time descriptor, rather than as a strictly equivalent hydraulic retention time for all observations. For continuous and semi-continuous systems, the reported hydraulic retention time was used. For batch systems, the reported digestion or incubation time was recorded as the residence time descriptor. This treatment follows the practical distinction between batch biochemical methane potential testing and continuous or semi-continuous reactor operation reported in anaerobic digestion testing and modelling studies [17–19,22]. Therefore, this variable was retained for early-stage screening, but it should not be interpreted as a fully mechanistic residence time parameter across all reactor configurations. Reactor or operation mode is reported in Supplementary Table S1 where available.

https://doi.org/10.3390/engproc2026147015

The number of ten records per substrate group was a methodological choice intended to construct a balanced benchmark dataset. For each substrate category, benchmark values were defined within the literature-informed ranges of substrate properties, operating conditions and total biogas yield. This design was used to provide comparable representation of high-, medium- and low-yield substrate classes and to prevent any single substrate category from dominating model development. The resulting records should therefore be interpreted as reconstructed benchmark cases, rather than as individual experimental observations extracted directly from the cited publications.

To introduce limited within-group variability and avoid perfectly deterministic benchmark patterns, controlled stochastic perturbation was applied to selected numerical variables during dataset generation. A fixed random seed of 42 was used to ensure reproducibility. The resulting values therefore represent literature-informed reconstructed benchmark records, rather than direct experimental measurements.

Gas volume reporting was not fully consistent across the underlying literature sources. Therefore, total biogas yield was harmonized to a common yield unit, but no additional dry-gas, wet-gas or standard temperature-and-pressure correction was imposed unless this information was explicitly available in the original source. Where gas volume basis was reported, it is indicated in Supplementary Table S1; otherwise, it is marked as not reported. This reporting inconsistency was treated as a dataset limitation and is one reason why the model is interpreted as an early-stage screening tool, rather than a design-level prediction model. The need to report test conditions, gas basis, substrate characterization and assay conditions is consistent with published recommendations for biochemical methane potential and biomethane potential test standardization [17–19].

After harmonization, the final modelling table contained no missing values in the seven predictors or in the target variable. Table 1 presents descriptive statistics calculated by the authors from the final harmonized dataset after screening, unit conversion and benchmark balancing.

Table 1. Descriptive statistics calculated by the authors from the final harmonized dataset used for model development.

Variable	Unit	Minimum	Maximum	Mean	Standard Deviation
VS content	%	67.13	95.82	84.99	7.36
TS content	%	2.30	93.50	31.57	34.25
C/N ratio	—	5.16	143.84	34.88	37.83
Residence time descriptor	days	15.00	50.00	26.29	8.43
Temperature	°C	35.00	38.00	36.22	1.26
pH	—	6.80	7.40	7.10	0.15
OLR	g VS L^{-1} d^{-1}	0.94	2.98	1.95	0.44
Total biogas yield	mL g^{-1} VS	122.02	580.48	359.97	133.46

Table 2 summarizes the substrate-level distribution of total biogas yield in the final harmonized benchmark dataset. As shown in Figure 1, total biogas yield varied substantially among substrate groups. Restaurant waste and food waste showed the highest mean yields, while lignocellulosic substrates such as wheat straw, corn stover and paper waste showed lower yields. This separation is consistent with differences in biodegradability, organic composition and nutrient balance among substrate classes.

Table 2. Substrate-level distribution of total biogas yield in the harmonized dataset.

Substrate Type	*n*	Minimum	Median	Maximum	Mean
Restaurant waste	10	474.3	533.8	580.5	529.2
Food waste	10	495.6	522.7	564.6	529.3
Algae	10	419.7	482.6	543.0	484.8
Fruit and vegetable waste	10	399.3	458.0	477.4	454.0
Co-digestion	10	388.5	431.9	494.3	431.5
Dairy wastewater	10	372.3	409.4	480.3	417.0
OFMSW	10	292.5	348.0	402.1	349.5
Sewage sludge	10	261.3	332.8	361.3	319.2
Cattle manure	10	243.1	275.2	302.3	274.5
Corn stover	10	169.7	210.7	253.7	212.4
Wheat straw	10	131.6	165.0	186.7	161.6
Paper waste	10	122.0	156.1	190.4	156.7

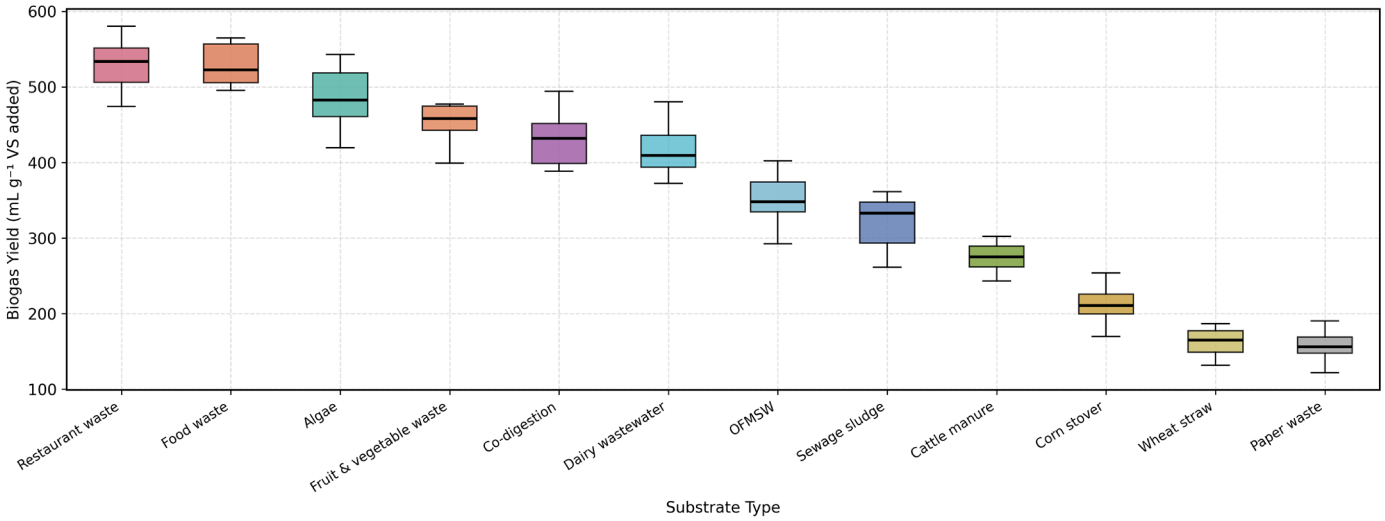


Figure 1. Distribution of total biogas yield by substrate type. Boxplot representation of total biogas yield values for twelve organic substrate groups included in the harmonized benchmark dataset. The figure shows clear differences between food-based high-yield substrates, intermediate-yield organic wastes and lower-yield lignocellulosic substrates.

The distribution shown in Figure 1 also has methodological implications for model interpretation. The clear separation among substrate groups improves the model's ability to learn substrate yield patterns, but it may also increase the risk that the model partially learns substrate-class structure, rather than only general physicochemical relationships. Therefore, the equal number of observations per substrate group should be interpreted as a benchmark design feature, rather than as a reflection of the natural distribution of anaerobic digestion records in the literature.

The dataset intentionally uses a limited set of descriptors because this study targets preliminary feasibility assessment, rather than full mechanistic simulation. More detailed variables such as lipid, protein, carbohydrate, cellulose, hemicellulose, lignin, ash content, inoculum-to-substrate ratio, pretreatment conditions, reactor geometry and gas composition data were not included because they were not consistently available across the literature basis used for this benchmark. The absence of these variables limits the model's ability to distinguish between substrates with similar VS, TS and C/N values but different biochemical compositions. This limitation is particularly relevant for high-yield substrates such as restaurant waste, food waste and algae, where lipid and protein contents can strongly influence total gas production [9,20,21].

2.2. Model Development

A Gradient Boosting Regression model was selected for this study. This algorithm was chosen because it can model nonlinear relationships and interaction effects between input variables while still allowing model interpretation through feature-importance analysis [23]. The model hyperparameters were selected to balance prediction accuracy and model complexity. In particular, the number of estimators, learning rate, maximum tree depth, minimum samples per split, minimum samples per leaf and subsample ratio were chosen to reduce excessive model variance while maintaining sufficient nonlinear fitting capacity. The final configuration is summarized in Table 3.

Table 3. Model configuration.

Parameter	Value
Algorithm	Gradient Boosting Regression
Number of estimators	200
Maximum tree depth	4
Learning rate	0.08
Minimum samples split	4
Minimum samples leaf	2
Subsample ratio	0.85
Train–test split	80:20
Training samples	96
Test samples	24
Cross-validation	5-fold
Random state	42

All computations were performed in Python 3.13 using NumPy 2.4.6, pandas 3.0.4, scikit-learn 1.8.0, and Matplotlib 3.11.1. Data handling and descriptive statistics were performed using pandas and NumPy. The GradientBoostingRegressor implementation from scikit-learn was used for model development, and Matplotlib was used for visualization. The model configuration was fixed before final evaluation to ensure reproducibility. The final model used 200 estimators, a maximum tree depth of 4, a learning rate of 0.08, minimum samples split of 4, minimum samples leaf of 2, a subsample ratio of 0.85 and random state = 42.

Before model training, preprocessing was performed using only the training subset in order to avoid information leakage from the test data. Although Gradient Boosting Regression is a tree-based algorithm and does not strictly require feature scaling, the same preprocessing workflow was retained for consistency and reproducibility. The dataset was divided into training and testing subsets using an 80:20 train–test split with a fixed random state. The training subset contained 96 samples, while the test subset contained 24 samples. Since random splitting can overestimate generalization performance when observations from the same substrate group appear in both subsets, the resulting test performance was interpreted as an estimate of interpolation ability within the compiled dataset domain.

For visual interpretation of prediction errors, tolerance bands were added to the predicted-versus-experimental and residual plots. In the test set predicted-versus-experimental plot, the shaded band represents $\pm 15\%$ relative deviation from the 1:1 reference line. In the residual plot, the shaded band represents $\pm 10\%$ relative residual error around zero. These shaded regions are visual engineering-tolerance guides and should not be interpreted as statistical confidence intervals.

https://doi.org/10.3390/engproc2026147015

2.3. Evaluation Metrics

The model was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE).

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

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

where $y_i, \hat{y}_i, \overline{y}$, and n denote the experimental value, predicted value, mean experimental value and number of observations, respectively.

R^2 measures the proportion of variance in the experimental biogas yield explained by the model. MAE represents the average absolute prediction error. RMSE gives higher weight to larger errors and is useful for identifying stronger deviations. MAPE expresses the prediction error as a percentage of the experimental value.

In addition to the train–test evaluation, five-fold cross-validation was performed to assess model stability. Residual diagnostics were used to examine prediction bias, residual normality and heteroscedasticity. Feature-importance analysis was applied to identify the relative contribution of each input variable. The absence of a broad external benchmark dataset was considered a methodological limitation. Therefore, the reported metrics were interpreted as internal validation results, rather than evidence of universal predictive performance for all anaerobic digestion substrates.

3. Results

3.1. Model Prediction Performance

The model showed strong agreement between experimental and predicted biogas yield values within the compiled dataset domain. For the training set, the model achieved R^2 = 1.0000, MAE = 0.56 mL g^{-1} VS, RMSE = 0.71 mL g^{-1} VS and MAPE = 0.17%. For the test set, the model achieved R^2 = 0.9885, MAE = 13.40 mL g^{-1} VS, RMSE = 16.16 mL g^{-1} VS and MAPE = 3.73%.

The test set results indicate that the model was able to predict test observations from the same compiled dataset with low average error. The MAPE value of 3.73% suggests that the model can provide practically useful preliminary estimates within the investigated range of substrate types and operating conditions. However, because the test subset was obtained from the same harmonized literature-informed benchmark dataset, these results should be interpreted as evidence of strong interpolation performance, rather than complete external generalization.

The corresponding training and test performance metrics are summarized in Table 4. The agreement between experimental and predicted total biogas yields, together with the residual pattern and feature-importance profile, is illustrated in Figure 2.

Table 4. Model performance metrics.

Metric	Training Set	Test Set
R^2	1.0000	0.9885
MAE, mL g^{-1} VS	0.56	13.40
RMSE, mL g^{-1} VS	0.71	16.16
MAPE, %	0.17	3.73
n	96	24

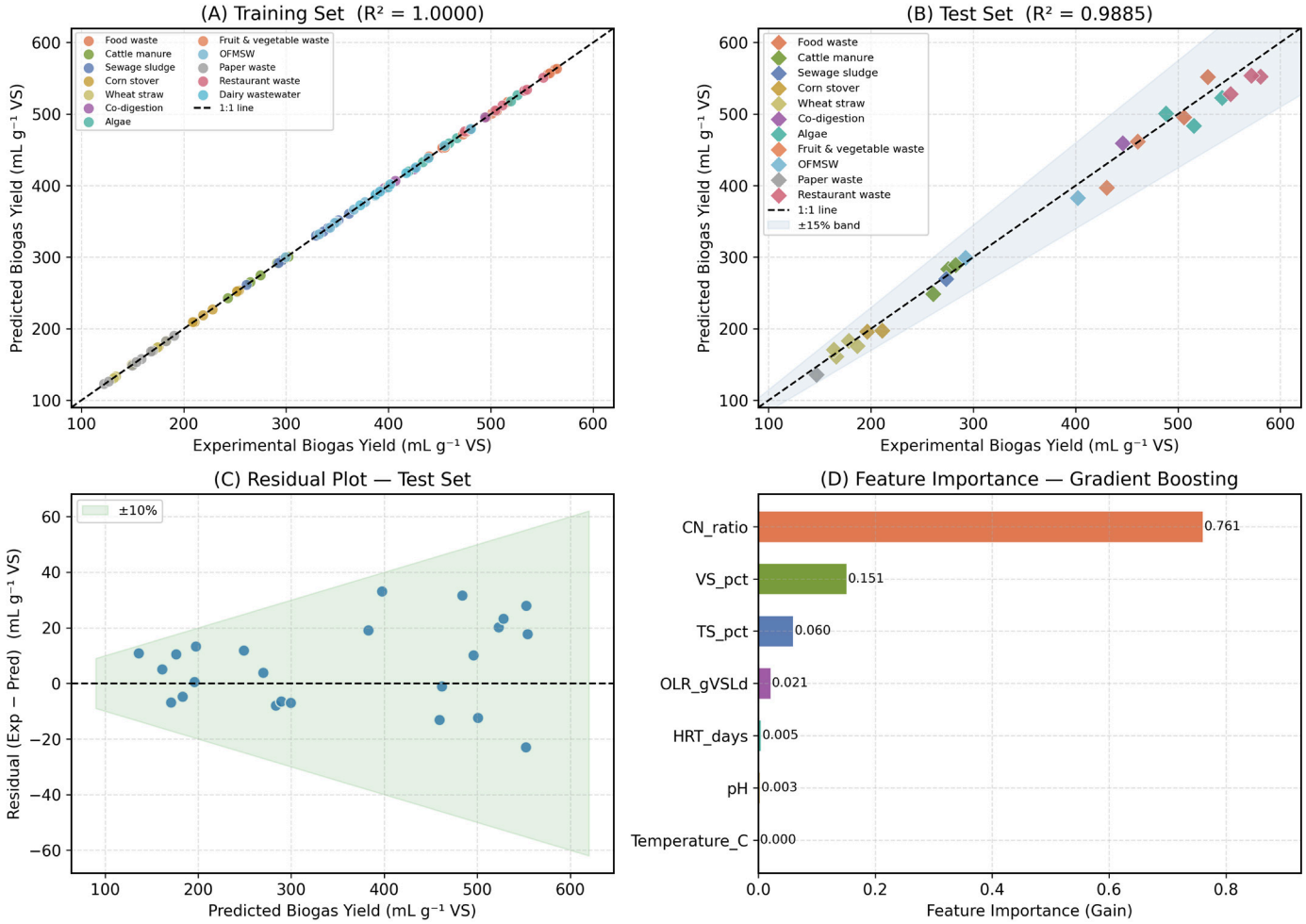


Figure 2. Predicted versus experimental total biogas yield and feature-importance analysis. Panel (**A**) shows the training set prediction, and panel (**B**) shows the test set prediction against the 1:1 reference line. The shaded band in panel B represents ±15% relative deviation from the 1:1 line. Panel (**C**) shows test set residuals, where the shaded band represents ±10% relative residual error around zero. Panel (**D**) shows the feature importance of the Gradient Boosting Regression model. The shaded regions are visual tolerance guides and do not represent statistical confidence intervals.

3.2. Cross-Validation Stability

Five-fold cross-validation was used to evaluate the robustness of the model. The fold-wise R^2 values were 0.976, 0.979, 0.982, 0.986 and 0.937. The mean cross-validation score was $R^2 = 0.9720 \pm 0.0180$. The fold-wise R^2 values and the corresponding mean ± standard deviation are summarized in Table 5.

Table 5. Five-fold cross-validation results.

Fold	R^2
Fold 1	0.976
Fold 2	0.979
Fold 3	0.982
Fold 4	0.986
Fold 5	0.937
Mean ± SD	0.9720 ± 0.0180

The first four folds showed very similar performance, while the fifth fold produced a lower R^2 value. However, the fifth fold still remained above 0.90, suggesting that the model retained acceptable prediction capability across validation subsets. The cross-validation results support the conclusion that the model is stable for interpolation within the data domain used in this study. Nevertheless, five-fold cross-validation based on random sample partitioning cannot fully evaluate the model's ability to predict completely unseen substrate classes. For this reason, future work should include group-wise validation, such as leave-one-substrate-group-out testing, to determine whether the model can generalize beyond substrate groups represented during training.

The figure shows the five-fold cross-validation R^2 values and the summary of training and test performance metrics. The dashed horizontal line indicates the mean cross-validation score.

3.3. Residual Diagnostics

Residual analysis was conducted to evaluate the reliability of the model beyond conventional accuracy metrics. The residuals were calculated as the difference between experimental and predicted biogas yield values. The residual distribution was approximately normal, and the Shapiro test gave $p = 0.771$. This indicates that no strong deviation from normality was observed in the test residuals.

However, the residuals were not fully homoscedastic. The residuals-versus-predicted plot showed increasing residual spread at higher predicted biogas yield values. The heteroscedasticity indicator was $r = 0.609$ with $p = 0.002$, suggesting that the prediction uncertainty increased for substrates with elevated biogas yields. This behaviour was most relevant for high-yield substrates such as restaurant waste, food waste, algae and fruit and vegetable waste.

The mean residual was approximately +6.5 mL g^{-1} VS, indicating a slight tendency toward underprediction on the test set. This bias was small relative to the total biogas yield range of the dataset, but it should still be considered when the model is used for high-yield substrate evaluation.

The complete diagnostic assessment, including the learning curve, train–test–cross-validation comparison, residual distribution, heteroscedasticity behaviour, fold-wise stability and Q-Q plot, is summarized in Figure 3.

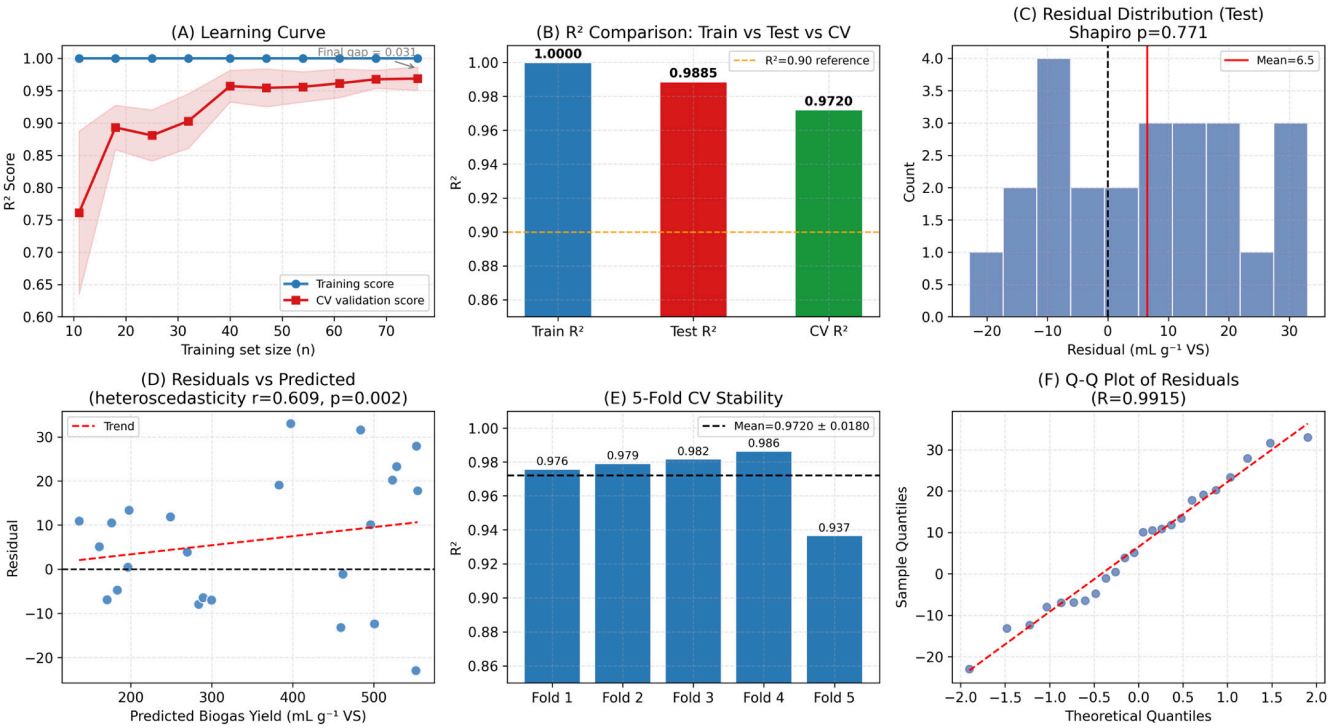


Figure 3. Model diagnostics and validation behaviour. (**A**) Learning curve; (**B**) comparison of training, test and cross-validation R^2, where the orange dashed line indicates the R^2 = 0.90 reference level; (**C**) test-set residual distribution, where the black dashed line indicates zero residual and the red line indicates the mean residual; (**D**) residuals versus predicted values, where the black dashed line represents zero residual and the red dashed line represents the fitted residual trend; (**E**) five-fold cross-validation stability, where the black dashed line indicates the mean cross-validation R^2; and (**F**) Q-Q plot of residuals, where the red dashed diagonal line represents the theoretical normal-reference line. The diagnostics show strong interpolation performance within the investigated dataset domain, but also reveal heteroscedastic residual behaviour at higher total biogas yields.

3.4. Feature-Importance Analysis

Feature-importance analysis showed that the C/N ratio was the dominant input variable, with an importance value of 0.761. The next most important variables were volatile solids content and total solids content, with importance values of 0.151 and 0.060, respectively. Organic loading rate had a smaller contribution, while hydraulic retention time or residence time descriptors, pH and temperature showed limited importance in the fitted model. The complete feature-importance ranking is reported in Table 6.

Table 6. Feature importance of input variables.

Input Variable	Feature Importance
C/N ratio	0.761
VS content	0.151
TS content	0.060
OLR	0.021
Residence time descriptor	0.005
pH	0.003
Temperature	0.000

The dominance of C/N ratio is physically reasonable because the nutrient balance of the substrate strongly affects microbial activity and anaerobic digestion performance. However, this result should not be interpreted as direct causal proof that C/N ratio alone

controls biogas yield. In the present dataset, C/N ratio may also act as an indirect descriptor of substrate class. For example, lignocellulosic substrates generally had higher C/N ratios and lower yields, whereas food-based substrates had more favourable C/N ranges and higher yields. Therefore, the feature-importance result should be interpreted as a model-based association, rather than a standalone biochemical conclusion.

4. Discussion

The obtained results demonstrate that a compact Gradient Boosting Regression model can estimate total biogas yield with high accuracy when applied within the substrate and operating condition domain represented in the harmonized benchmark dataset. The term "estimate" is more appropriate than "predict" in this context, because this study is based on a limited literature-informed dataset and does not include independent external experimental validation. The high test set R^2 and low MAPE indicate that the model captured the main nonlinear associations between substrate descriptors, operating variables and total biogas yield.

Compared with previously published data-driven approaches in anaerobic digestion, the present study has a narrower, but practically oriented, objective. Earlier machine-learning studies have often focused on methane yield prediction, reactor monitoring, soft sensing, temperature control, volatile fatty acid estimation, process optimization or high-solids anaerobic digestion performance [13–19]. In contrast, the present work focuses on early-stage screening of total biogas yield using a compact set of commonly reported substrate and operating descriptors. Therefore, the results should not be interpreted as direct numerical superiority over previous models, because published studies differ in target variable, dataset structure, substrate coverage, input descriptors and validation strategy. The main contribution of this work is the combination of a simple Gradient Boosting Regression model with explicit diagnostic analysis, including train–test performance, cross-validation stability, residual distribution, heteroscedasticity and feature-importance interpretation.

At the same time, the nearly perfect training set performance requires careful interpretation. A training R^2 of 1.0000 may indicate that the model has fitted the training data very closely. In many cases, such behaviour can be associated with overfitting. However, the high test set performance and stable cross-validation results suggest that the model retained strong interpolation capability for test samples drawn from the same dataset domain. This does not eliminate the possibility of overfitting to substrate-specific patterns, especially because substrate groups may be indirectly represented through variables such as C/N ratio, VS and TS. Therefore, the model should not be described as broadly generalizable until it is tested on independent datasets or substrate groups excluded from training.

The residual diagnostics provide additional insight into the model's limitations. The approximately normal residual distribution supports the statistical consistency of the prediction errors. However, the presence of heteroscedasticity indicates that prediction uncertainty is not constant across the yield range. In particular, higher-yield substrates showed larger residual variation. This may be related to the greater compositional complexity of food-based and lipid-rich wastes, where differences in carbohydrate, protein and lipid fractions can significantly affect biogas production even when general indicators such as VS, TS and C/N ratio are similar.

The strong dominance of C/N ratio in the feature-importance ranking also requires cautious interpretation. C/N ratio is an important parameter in anaerobic digestion because it influences nutrient balance, ammonia formation and microbial stability. However, in the present dataset, C/N ratio may also reflect substrate class. For example, paper waste, wheat straw and corn stover have high C/N ratios and low biodegradability, while restaurant waste and food waste have lower C/N ratios and higher biogas potential. Therefore, the

model may partly learn substrate-type patterns, rather than only fundamental physicochemical relationships. This limitation is common in the literature-derived datasets where substrate groups are well separated but within-substrate variability is limited.

The very low feature-importance values of temperature and pH should be interpreted as dataset-specific statistical results, rather than evidence that these variables are biologically or operationally unimportant. Temperature and pH are known to be key factors affecting microbial activity, process stability and inhibition in anaerobic digestion [2,5,6]. However, the present dataset covered only a narrow temperature range of 35–38 °C and a narrow pH range of 6.80–7.40. These values mainly represent stable mesophilic operating conditions and do not include broad psychrophilic, thermophilic or inhibitory pH regimes. Consequently, the model had limited statistical variation from which to learn the wider effects of temperature and pH. In datasets covering wider operating ranges, these variables would be expected to show stronger predictive importance.

From a practical point of view, the proposed model is useful for early-stage screening. It can support preliminary comparison of feedstocks, rapid feasibility assessment and initial estimation of expected biogas yield when only basic substrate and operating data are available. For example, the model may help rank candidate substrates before conducting detailed biochemical methane potential tests. However, it should not be used as the sole basis for reactor sizing, investment decisions or guaranteed energy yield estimation, because total biogas yield does not directly represent methane yield, gas quality or long-term reactor stability.

5. Conclusions

A Gradient Boosting Regression model was developed for early-stage screening of total biogas yield using a harmonized literature-informed benchmark dataset of 120 observations from twelve organic substrate groups. The model used seven commonly available input variables: volatile solids content, total solids content, C/N ratio, residence time descriptor, digestion temperature, pH and organic loading rate.

The model achieved high internal test set accuracy, with $R^2 = 0.9885$, MAE = 13.40 mL g^{-1} VS, RMSE = 16.16 mL g^{-1} VS and MAPE = 3.73%. Five-fold cross-validation produced a mean R^2 of 0.9720 $\pm$ 0.0180, confirming stable interpolation performance within the investigated data domain. Feature-importance analysis identified C/N ratio as the most influential predictor, followed by volatile solids and total solids.

Despite the strong predictive performance, model diagnostics revealed important limitations. The nearly perfect training set fit suggests possible overfitting, although the test and cross-validation results remained strong. Residual analysis showed heteroscedasticity, meaning that prediction uncertainty increased for high-yield substrates. In addition, the dominance of C/N ratio suggests that the model may partially learn substrate-type patterns, rather than only general physicochemical relationships. The low feature importance of temperature and pH should be interpreted as a consequence of their narrow ranges in the compiled dataset, not as evidence that these variables are unimportant in anaerobic digestion.

Overall, the proposed model is suitable as a preliminary screening tool for early-stage comparison of organic substrates for total biogas production within the investigated dataset range. Its use outside this domain should be considered exploratory until external validation is performed. Future work should focus on expanding the dataset, increasing within-substrate variability, including more detailed compositional descriptors, reporting gas volume basis consistently and validating the model using independent experimental datasets and group-wise validation strategies.

https://doi.org/10.3390/engproc2026147015

6. Limitation Statement

The main limitation of this study is related to the size, structure and reporting consistency of the literature-informed benchmark dataset. Although the dataset included twelve substrate groups, each group contained only ten observations, which helped maintain balance, but did not fully represent the natural variability of anaerobic digestion experiments. The response variable was total biogas yield, not methane yield or methane content; therefore, the model output cannot be directly interpreted as methane energy potential without additional gas composition information. In addition, the original literature sources did not consistently report dry or wet gas basis, standard temperature-and-pressure normalization, inoculum-to-substrate ratio, detailed biochemical composition, pretreatment conditions or reactor configuration. The residence time descriptor should also be interpreted cautiously because residence time definitions differ among batch, continuous, and semi-continuous systems. Residual diagnostics showed heteroscedasticity, indicating higher prediction uncertainty for high-yield substrates. Moreover, the dominance of C/N ratio suggests that the model may partially learn substrate-class patterns, rather than only general physicochemical relationships. Finally, the low feature importance of temperature and pH reflects their narrow ranges in the compiled dataset and should not be interpreted as evidence that these variables are unimportant in anaerobic digestion. Future work should incorporate larger externally validated datasets, broader operating condition ranges, detailed compositional descriptors, consistent gas volume reporting and group-wise validation such as leave-one-substrate-group-out testing.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/engproc2026147015/s1, Table S1: Harmonized literature-informed benchmark dataset used for Gradient Boosting Regression modelling of total biogas yield. The supplementary dataset contains 120 observations covering twelve organic substrate groups and includes the substrate classification, total biogas yield target, volatile solids content, total solids content, C/N ratio, residence time descriptor, digestion temperature, pH, organic loading rate, reactor or operation mode where available, gas volume reporting information, data origin notes, and the train–test allocation used for model development. Additional notes and substrate-level summary statistics are also provided to facilitate interpretation and reproducibility.

Author Contributions: Conceptualization, Z.T. and G.N.; methodology, N.Y. and K.N.; validation, Z.T., G.N. and J.E. (Jaloliddin Eshbobaev); investigation, J.E. (Jasmina Elmurodova) and E.S.; resources, M.A. and K.N.; writing—original draft preparation, Z.T.; writing—review and editing, Z.T.; visualization, J.E. (Jaloliddin Eshbobaev); supervision, Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The complete dataset used in this study is provided as Supplementary Table S1. The table includes sample ID, substrate group, target indicator, total biogas yield, input variables, unit-normalized values, residence time descriptor, reactor or operation mode where available, gas volume basis where reported, and data origin notes. Values with incomplete gas-basis information are explicitly marked as not reported, rather than being corrected by assumption.

Acknowledgments: The authors acknowledge the support of the Department of Automation and Digital Control and the Department of Biotechnology of Tashkent Institute of Chemical Technology for their methodological discussions related to biogas process modelling and data interpretation. The authors also acknowledge Tashkent State Technical University for academic support during the preparation of the manuscript.

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

Abbreviations

C/N	Carbon-to-nitrogen ratio
CV	Cross-validation
HRT	Hydraulic retention time
IoT	Internet of Things
MAE	Mean absolute error
MAPE	Mean absolute percentage error
OFMSW	Organic fraction of municipal solid waste
OLR	Organic loading rate
Q-Q	Quantile–quantile
R^2	Coefficient of determination
RMSE	Root mean square error
SD	Standard deviation
TS	Total solids
VS	Volatile solids

References

1. Alengebawy, A.; Ran, Y.; Osman, A.I.; Jin, K.; Samer, M.; Ai, P. Anaerobic Digestion of Agricultural Waste for Biogas Production and Sustainable Bioenergy Recovery: A Review. *Environ. Chem. Lett.* **2024**, *22*, 2641–2668. [CrossRef]
2. Harirchi, S.; Wainaina, S.; Sar, T.; Nojoumi, S.A.; Parchami, M.; Parchami, M.; Varjani, S.; Khanal, S.K.; Wong, J.; Awasthi, M.K.; et al. Microbiological Insights into Anaerobic Digestion for Biogas, Hydrogen or Volatile Fatty Acids (VFAs): A Review. *Bioengineered* **2022**, *13*, 6521–6557. [CrossRef] [PubMed]
3. Jameel, M.K.; Mustafa, M.A.; Ahmed, H.S.; Mohammed, A.J.; Ghazy, H.; Shakir, M.N.; Lawas, A.M.; Mohammed, S.K.; Idan, A.H.; Mahmoud, Z.H.; et al. Biogas: Production, Properties, Applications, Economic and Challenges: A Review. *Results Chem.* **2024**, *7*, 101549. [CrossRef]
4. Anand, K.; Mittal, A.P.; Kumar, B. ANN-Based Sensorless Adaptive Temperature Control System to Improve Methane Yield in an Anaerobic Digester. *Biomass Conv. Bioref.* **2023**, *13*, 7265–7285. [CrossRef]
5. Paladino, O. Data Driven Modelling and Control Strategies to Improve Biogas Quality and Production from High Solids Anaerobic Digestion: A Mini Review. *Sustainability* **2022**, *14*, 16467. [CrossRef]
6. 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. *Environ. Sci. Pollut. Res.* **2024**, *31*, 19085–19104. [CrossRef] [PubMed]
7. Andrade Cruz, I.; Chuenchart, W.; Long, F.; Surendra, K.C.; Renata Santos Andrade, L.; Bilal, M.; Liu, H.; Tavares Figueiredo, R.; Khanal, S.K.; Fernando Romanholo Ferreira, L. Application of Machine Learning in Anaerobic Digestion: Perspectives and Challenges. *Bioresour. Technol.* **2022**, *345*, 126433. [CrossRef] [PubMed]
8. Appala, V.N.S.G.; Pandhare, N.N.; Bajpai, S. Mathematical Models for Optimization of Anaerobic Digestion and Biogas Production. In *Zero Waste Biorefinery*; Nandabalan, Y.K., Garg, V.K., Labhsetwar, N.K., Singh, A., Eds.; Energy, Environment, and Sustainability; Springer: Singapore, 2022; pp. 575–591.
9. Lohani, S.P. Anaerobic Co-Digestion of Food Waste, Goat and Chicken Manure for Sustainable Biogas Production. *Int. J. Energy Appl. Technol.* **2020**, *7*, 120–125. [CrossRef]
10. Kazemi, P.; Steyer, J.-P.; Bengoa, C.; Font, J.; Giralt, J. Robust Data-Driven Soft Sensors for Online Monitoring of Volatile Fatty Acids in Anaerobic Digestion Processes. *Processes* **2020**, *8*, 67. [CrossRef]
11. Pandian, I.; Begum, S.; Kumaravel, S.P. An Integrated IoT and Fuzzy Logic Controller System for Biogas Digester to Predict Methane Generation. *Environ. Dev. Sustain.* **2021**, *27*, 22517–22529. [CrossRef]
12. Ganeshan, P.; Bose, A.; Lee, J.; Barathi, S.; Rajendran, K. Machine Learning for High Solid Anaerobic Digestion: Performance Prediction and Optimization. *Bioresour. Technol.* **2024**, *400*, 130665. [CrossRef] [PubMed]
13. Yusupov, M.; Turakulov, Z.; Yusupbekov, A. Selection and Application of Control Principles in Beer Brewing Processes Based on MCDA Framework. *Braz. J. Biosyst. Eng.* **2026**, *19*, 1295. [CrossRef]
14. Yusupov, M.; Eshbobaev, J.; Turakulov, Z.; Usmanov, K.; Kadirova, D.; Yusupbekov, A. Modeling and Control of Nonlinear Fermentation Dynamics in Brewing Industry. *Eng. Proc.* **2026**, *117*, 67. [CrossRef]
15. Piadeh, F.; Offie, I.; Behzadian, K.; Rizzuto, J.P.; Bywater, A.; Córdoba-Pachón, J.-R.; Walker, M. A Critical Review for the Impact of Anaerobic Digestion on the Sustainable Development Goals. *J. Environ. Manag.* **2024**, *349*, 119458. [CrossRef] [PubMed]

16. Archana, K.; Visckram, A.S.; Senthil Kumar, P.; Manikandan, S.; Saravanan, A.; Natrayan, L. A Review on Recent Technological Breakthroughs in Anaerobic Digestion of Organic Biowaste for Biogas Generation: Challenges towards Sustainable Development Goals. *Fuel* **2024**, *358*, 130298. [CrossRef]
17. Filer, J.; Ding, H.H.; Chang, S. Biochemical Methane Potential (BMP) Assay Method for Anaerobic Digestion Research. *Water* **2019**, *11*, 921. [CrossRef]
18. Angelidaki, I.; Alves, M.; Bolzonella, D.; Borzacconi, L.; Campos, J.L.; Guwy, A.J.; Kalyuzhnyi, S.; Jenicek, P.; Van Lier, J.B. Defining the Biomethane Potential (BMP) of Solid Organic Wastes and Energy Crops: A Proposed Protocol for Batch Assays. *Water Sci. Technol.* **2009**, *59*, 927–934. [CrossRef] [PubMed]
19. Holliger, C.; Alves, M.; Andrade, D.; Angelidaki, I.; Astals, S.; Baier, U.; Bougrier, C.; Buffière, P.; Carballa, M.; De Wilde, V.; et al. Towards a Standardization of Biomethane Potential Tests. *Water Sci. Technol.* **2016**, *74*, 2515–2522. [CrossRef] [PubMed]
20. Kafle, G.K.; Chen, L. Comparison on Batch Anaerobic Digestion of Five Different Livestock Manures and Prediction of Biochemical Methane Potential (BMP) Using Different Statistical Models. *Waste Manag.* **2016**, *48*, 492–502. [CrossRef] [PubMed]
21. Zhang, C.; Su, H.; Baeyens, J.; Tan, T. Reviewing the Anaerobic Digestion of Food Waste for Biogas Production. *Renew. Sustain. Energy Rev.* **2014**, *38*, 383–392. [CrossRef]
22. Rocamora, I.; Wagland, S.T.; Villa, R.; Simpson, E.W.; Fernández, O.; Bajón-Fernández, Y. Dry Anaerobic Digestion of Organic Waste: A Review of Operational Parameters and Their Impact on Process Performance. *Bioresour. Technol.* **2020**, *299*, 122681. [CrossRef] [PubMed]
23. Cai, W.; Wei, R.; Xu, L.; Ding, X. A Method for Modelling Greenhouse Temperature Using Gradient Boost Decision Tree. *Inf. Process. Agric.* **2022**, *9*, 343–354. [CrossRef]

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.