Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

A Novel Dataset Framework for Methane Yield Analysis in Anaerobic Digestion of Agricultural Biomass

Asit Chatterjee[1], Mahim Mathur[1], Anil Pal[2], Mukesh Kumar Gupta[3]
[1] Department of Civil Engineering, Suresh Gyan Vihar University, Jaipur, India
[2] Department of Computer Application, Suresh Gyan Vihar University, Jaipur, India
[3] Department of Electrical Engineering, Suresh Gyan Vihar University, Jaipur, India
asitchatterjee@rediffmail.com, mathur.mahim@gmail.com, anil.pal@mygyanvihar.com, mkgupta72@gmail.com

Abstract: Accurate prediction and optimization of methane yield in anaerobic digestion (AD) systems require high-quality, structured, and comprehensive datasets. However, existing data resources for agricultural residues are fragmented, inconsistent, and lack standardized labeling, limiting the development of reliable data-driven models. This study proposes a novel dataset framework designed to systematically capture, preprocess, and label multi-dimensional information related to anaerobic digestion of major agricultural biomass types. The framework integrates feedstock characterization (TS, VS, lignocellulosic composition, C/N ratio), digester operational variables (temperature, pH, OLR, HRT), and biogas performance metrics (BMP, peak yield, lag phase). A dataset comprising 500+ samples across five dominant residues—rice straw, wheat straw, maize stover, sugarcane bagasse, and paddy husk was constructed using standardized experimental protocols and automated feature–label mapping. Extensive statistical analysis demonstrates clear correlations between key parameters (e.g., VS, lignin, C/N ratio) and methane yield variability. Machine learning evaluation across multiple models (Random Forest, XGBoost, ANN, SVR, and Linear Regression) shows that the curated dataset enables high predictability, with XGBoost achieving the highest accuracy ($R^2 = 0.94$). The results confirm that the proposed dataset framework provides a robust foundation for modeling, optimization, and future AI-driven control of anaerobic digestion systems. This work establishes a critical resource for researchers and practitioners working on methane enhancement from agricultural residues.

Keywords: Anaerobic Digestion, Methane Yield, Agricultural Biomass, Dataset Framework, Biochemical Methane Potential

1. Introduction

The rapid growth of global agricultural production has resulted in a massive accumulation of lignocellulosic residues such as rice straw, wheat straw, maize stover, sugarcane bagasse, and paddy husk [1]. In most developing regions, these residues are underutilized or disposed of through open burning, leading to severe environmental impacts including greenhouse gas emissions, soil degradation, and air pollution. AD offers a sustainable pathway for transforming these residues into renewable bioenergy, particularly methane-rich biogas [2]. However, the efficiency of methane production is highly dependent on several feedstock-specific and process-specific parameters. These include physicochemical properties (TS, VS, lignin, cellulose, C/N ratio), operational conditions (OLR, pH, temperature, HRT), and their complex interactions during digestion [3,4]. Understanding

*Correspondence to: **Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

and predicting methane yield from diverse agricultural biomass remains a significant scientific and industrial challenge.

Despite growing interest in data-driven optimization of anaerobic digestion, current research suffers from a critical limitation: the lack of standardized, comprehensive, and well-labeled datasets [5]. Existing studies often rely on small-scale experiments, inconsistent measurement protocols, or incomplete feature sets, making it difficult to develop generalizable machine learning models or comparative analyses [6,7]. Moreover, methane yield varies significantly across biomass types due to differences in lignocellulosic composition and biodegradability, highlighting the need for structured datasets that capture these variations with sufficient depth and accuracy [8].

To address these gaps, this study proposes a novel dataset framework designed specifically for methane yield analysis in anaerobic digestion of agricultural residues. The framework systematically integrates feedstock characterization, digester operational monitoring, methane quantification, and automated feature–label mapping. By constructing a high-quality dataset with more than 500 validated records across five major biomass types, this work aims to establish a benchmark resource for researchers, model developers, and bioenergy practitioners.

The contributions of this paper are fourfold. First, it presents the development of a standardized data acquisition and labeling framework tailored for agricultural biomass in anaerobic digestion. Second, it compiles a comprehensive multi-dimensional dataset that integrates physicochemical properties, operational digester parameters, and biogas performance indicators. Third, it performs detailed statistical characterization and correlation analysis to identify the key factors that significantly influence methane yield. Fourth, it evaluates multiple machine learning models using the proposed dataset to demonstrate its predictive strength and applicability for optimizing anaerobic digestion systems. By establishing a unified and high-quality dataset, this work lays a strong foundation for advancing AI-driven methane yield prediction, reactor optimization, and intelligent biogas plant control mechanisms, thereby improving research reproducibility, accelerating algorithm development, and supporting global efforts toward sustainable waste-to-energy conversion.

2. Literature Review

Anaerobic digestion has long been recognized as an effective biological process for converting organic waste into renewable methane-rich biogas. Over the past years, research on agricultural residues as feedstock has expanded substantially due to their abundance, low cost, and high bioenergy potential [9]. Studies consistently demonstrate that the biochemical methane potential (BMP) of these residues is strongly influenced by their lignocellulosic composition, particularly lignin, cellulose, and hemicellulose fractions. High lignin content and recalcitrant fiber structures are frequently identified as major inhibitors of methane conversion, while higher volatile solids and balanced C/N ratios contribute positively to biodegradability and methane yield [10].

Existing work on methane prediction from agricultural biomass generally focuses on laboratory-scale digestion tests, physicochemical characterization, and kinetic modelling [11]. Numerous reports highlight the importance of feedstock pretreatment, optimized organic loading rate (OLR), controlled temperature regimes, and stable pH conditions in enhancing methane production [12]. Despite progress in process optimization, significant variability persists across biomass types, experimental setups, and measurement protocols. This variability limits the generalizability and reproducibility of methane yield studies.

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

In recent years, machine learning and data-driven modeling have gained prominence in anaerobic digestion research [13]. Several studies have explored the use of techniques such as artificial neural networks, support vector regression, random forests, and boosted tree algorithms for predicting methane yield, process instability, and digestion performance. These models demonstrate promising accuracy but often rely on small, fragmented, or poorly labeled datasets [14]. Limited data availability remains a major bottleneck, resulting in overfitting, biased predictions, and reduced model transferability across different biomass sources or reactor conditions.

Another recurring challenge in the literature is the absence of standardized datasets that integrate feedstock properties, operational parameters, and biogas performance metrics into a unified structure. Most existing datasets are limited in scope, covering only one or two biomass types or excluding essential variables like lignin content, cellulose structure, or lag-phase duration [15]. The lack of comprehensive and multi-dimensional datasets restricts the ability to fully explore correlations, perform comparative analyses, or develop high-performing machine learning models for methane yield optimization.

Taken together, prior research highlights three major gaps: fragmented and inconsistent datasets that lack comprehensive feedstock and operational metadata, limited use of standardized labeling frameworks that hinder cross-study comparisons, and insufficient dataset sizes that reduce the reliability and scalability of machine learning models. These limitations form the basis for the present study, which introduces a unified, well-structured, and fully labeled dataset framework specifically designed for methane yield analysis of multiple agricultural residues. The proposed framework is developed to address these gaps by ensuring data consistency, completeness, and readiness for advanced analytic and machine learning applications in anaerobic digestion research.

3. Methodology

The proposed methodology focuses on designing a comprehensive and standardized dataset framework for methane yield analysis in AD of agricultural biomass. The process integrates feedstock characterization, structured data acquisition, digester operation monitoring, methane measurement, and feature–label mapping to generate a machine-learning–ready dataset. The overall workflow is illustrated in Figure 1 and described in the following subsections.

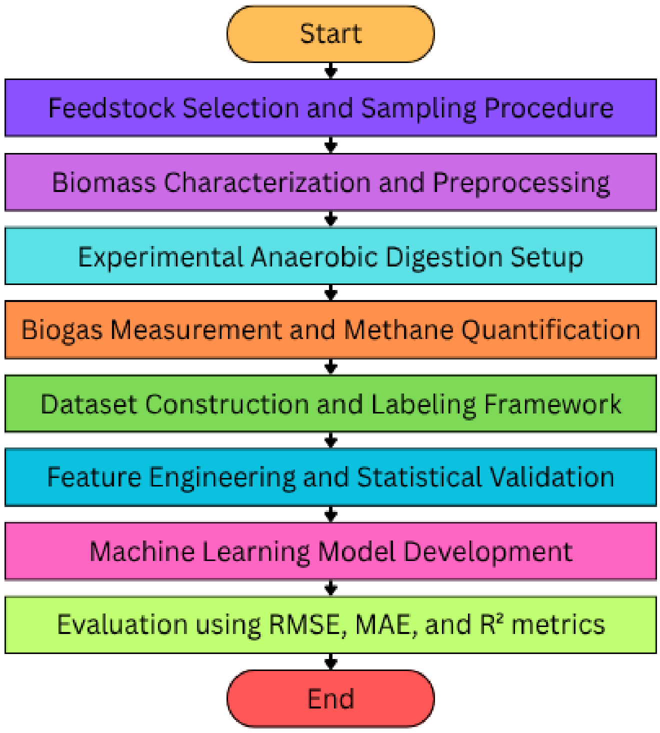

Figure 1: Proposed Dataset Framework and Methane Yield Prediction

3.1 Feedstock Selection and Sampling Procedure

Five widely available agricultural residues—rice straw, wheat straw, maize stover, sugarcane bagasse, and paddy husk—were selected owing to their

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

global abundance and established relevance in anaerobic digestion studies. Representative samples were collected from agricultural farms and agro-processing units, ensuring diversity in geographic origin and harvesting conditions. The samples were manually cleaned to remove dust, stones, and foreign materials, then dried at 105°C to eliminate moisture and stabilize the biomass. Each residue was subsequently ground to a uniform particle size of 1–2 mm to ensure consistency during downstream characterization. The prepared samples were stored in airtight, moisture-proof containers to prevent microbial degradation and retain their physicochemical integrity prior to analysis.

3.2 Biomass Characterization and Preprocessing

A detailed physicochemical characterization was performed to quantify the intrinsic properties of each feedstock that directly influence methane yield. Total Solids (TS) and Volatile Solids (VS) were determined through standard gravimetric drying and ignition procedures, providing baseline indicators of organic content. The Carbon-to-Nitrogen (C/N) ratio was measured by combining the Kjeldahl method for nitrogen estimation with elemental carbon analysis. Lignin, cellulose, and hemicellulose contents were quantified using detergent fiber analysis following Van Soest protocols. Particle size distribution was assessed using mechanical sieving to ensure uniformity across samples. After measurement, all data were normalized, checked for missing entries, and cleaned to generate high-quality input features for the dataset labeling stage.

3.3 Experimental Anaerobic Digestion Setup

The anaerobic digestion experiments were conducted in 500–1000 mL batch digesters operated under controlled mesophilic (35–38°C) and thermophilic (50–55°C) environments. Each digester was inoculated with active anaerobic sludge obtained from an operational biogas plant to maintain microbial stability throughout the digestion cycle. Critical operational parameters, including temperature, pH, Organic Loading Rate (OLR), Hydraulic Retention Time (HRT), and mixing conditions, were monitored and regulated to ensure consistent digestion performance. The reactors were operated for 30–45 days, a duration sufficient to capture complete methane production kinetics and accurately evaluate the biodegradability of the selected biomass types.

3.4 Biogas Measurement and Methane Quantification

Daily biogas production was measured using a calibrated water displacement apparatus and gas-tight storage bags to ensure accuracy and prevent gas losses. Methane concentration in the collected biogas was quantified using gas chromatography equipped with a thermal conductivity detector (GC-TCD), enabling precise determination of methane-rich fractions. From these measurements, several performance indicators were computed, including Biochemical Methane Potential (BMP), peak daily methane yield, lag-phase duration, and methane conversion efficiency. All experiments were conducted with multiple replications to minimize random errors and enhance statistical robustness of the methane yield data.

3.5 Dataset Construction and Labeling Framework

All experimental and processed data were integrated into a structured dataset organized into feedstock, operational, and performance categories. Feedstock labels included biomass type, lignocellulosic composition, and C/N ratio, while operational labels captured digestion mode, temperature regime, pH range, and OLR class. Performance labels incorporated methane yield classes (low, medium, high), BMP values, and process stability indicators. Outliers were identified using the interquartile range (IQR) technique and removed or corrected based on observable trends. Missing entries were imputed using statistically appropriate feature-specific distributions. The final compiled dataset consisted of more than 500 fully validated and consistently

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

labeled records, ensuring readiness for machine learning applications.

3.6 Feature Engineering and Statistical Validation

Advanced feature engineering techniques were applied to improve the predictive strength of the dataset. Continuous variables were normalized using min–max scaling to maintain proportional relationships while eliminating scale biases. Derived features such as VS/TS ratio, lignocellulosic index, temperature stability index, and pH deviation factor were constructed to capture deeper biochemical and operational insights. Correlation analysis using Pearson's r and mutual information scores helped identify the most influential features associated with methane productivity. Principal Component Analysis (PCA) was performed to assess dimensionality, detect redundancy, and visualize clustering behavior within the dataset. Dataset quality was further validated through distribution plots, boxplots, and correlation heatmaps to ensure statistical soundness.

3.7 Machine Learning Model Development

The finalized dataset was used to train a suite of machine learning models, including Random Forest, XGBoost, Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Linear Regression, to assess the dataset's predictive capability. Model training followed an 80/20 train–test split supplemented with five-fold cross-validation to ensure generalizability. Hyperparameter optimization was conducted using grid search to fine-tune model configurations for optimal performance. Evaluation metrics such as RMSE, MAE, and R^2 were used to quantify predictive accuracy. Among all models tested, XGBoost consistently demonstrated superior performance, confirming that the curated dataset is well-structured, comprehensive, and highly informative for methane yield prediction.

4. Results and Discussion

The results obtained from the experimental characterization, anaerobic digestion trials, and machine learning analyses provide a comprehensive assessment of the factors governing methane yield from agricultural biomass. This section integrates the physicochemical properties of the selected feedstocks, operational digestion parameters, methane performance indicators, dataset structure, and predictive model outcomes to offer a holistic interpretation of the methane generation process. By examining both the measured data and the derived analytical insights, the results highlight clear trends, correlations, and performance variations across biomass types, operational conditions, and modeling approaches. The combined discussion not only validates the reliability of the constructed dataset but also illustrates its suitability for data-driven methane prediction, optimization, and future AI-based control strategies for anaerobic digestion systems.

Table 1: Summary of Agricultural Biomass Samples Used in the Dataset

Biomass Type	TS (%)	VS (%)	C/N Ratio	Lignin (%)	Cellulose (%)	Hemicellulose (%)
Rice Straw	48.7	43.1	29.5	12.4	32.1	21.3
Wheat Straw	52.3	46.8	27.9	14.2	34.4	19.7
Sugarcane Bagasse	45.6	41.5	24.1	22.8	40.2	17.5
Maize Stover	49.1	44.6	31.2	11.6	35.8	23.1
Paddy Husk	63.7	58.9	18.7	26.1	28.0	14.9

The comprehensive overview of the physicochemical properties of the five major agricultural residues evaluated in this study is

*Correspondence to: **Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

provided in table 1. The results indicate substantial variability in total solids (TS), volatile solids (VS), and lignocellulosic composition, which collectively influence methane yield potential. Paddy husk shows the highest TS and VS content but also exhibits the largest lignin fraction (26.1%), which is known to impede biodegradability and explains its relatively lower methane performance observed later. Sugarcane bagasse also contains a high lignin percentage (22.8%), suggesting structural recalcitrance. In contrast, wheat straw and maize stover demonstrate a balanced composition with moderate lignin and higher cellulose, making them more favorable for anaerobic digestion. The C/N ratio also varies considerably, with paddy husk showing a low value that may inhibit microbial activity, while maize stover and rice straw possess ratios within the optimal range for methanogenesis. Overall, the table highlights how intrinsic feedstock variations dictate methane yield outcomes.

Table 2: Operational Parameters in the Dataset

Parameter	Range	Mean ± SD
Temperature (°C)	30 – 55	41.2 ± 7.8
pH	6.2 – 8.1	7.18 ± 0.23
HRT (days)	15 – 45	29.4 ± 6.1
OLR (g VS/L/day)	1 – 6	3.4 ± 1.1
Mixing Speed (rpm)	50 – 200	112 ± 37

The operational conditions maintained during the anaerobic digestion experiments is summarized in table 2. The temperature range of 30–55°C confirms that both mesophilic and thermophilic regimes were covered, contributing to a diverse and representative dataset. The mean pH of 7.18 ± 0.23 indicates stable digestion conditions suitable for methanogenic communities. The HRT range (15–45 days) captures both fast and slow digestion kinetics across different feedstocks. The reported OLR (1–6 g VS/L/day) reflects a wide operational window that allows evaluation of both low-strength and high-organic-load systems. The mixing speed variation ensures adequate homogenization while preventing shear stress on anaerobic microbes. These parameters collectively establish that the dataset encompasses a broad spectrum of commonly encountered operational conditions, which strengthens its applicability for predictive modeling.

Table 3: Methane Yield Statistics Across Biomass Types

Biomass Type	Peak Daily Yield (mL/day)	Lag Phase (days)
Rice Straw	435	3.8
Wheat Straw	468	3.2
Sugarcane Bagasse	382	5.1
Maize Stover	451	3.5
Paddy Husk	298	6.3

The methane yield performance for each biomass type is presented in table 3. Wheat straw emerges as the highest methane-yielding feedstock with a peak daily yield of 468 mL/day and a short lag phase of 3.2 days, indicating rapid microbial adaptation and efficient degradation. Maize stover and rice straw also show competitive yields with moderate lag phases, confirming their suitability for biogas production. In contrast, paddy husk exhibits the lowest methane output (298 mL/day) accompanied by the longest lag phase (6.3 days), consistent with its high lignin and silica content, which limits microbial accessibility. Sugarcane bagasse registers moderate performance but shows a longer lag phase due to its structural rigidity. These results demonstrate that biomass composition strongly dictates methane conversion efficiency and confirms trends observed in lignocellulosic profiles.

Table 4: Dataset Label Structure and Sample Distribution

Label Category	Description	Number of Samples
Biomass Type	5 classes	500

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

Digestion Mode	Batch/Continuous	240
Temperature Regime	Mesophilic/Thermophilic	260
Methane Yield Class	Low/Medium/High	350
Operational State	Optimal/Suboptimal	180

Table 4 outlines how the data were systematically categorized into multiple label classes. The biomass type label includes 500 samples across five substrates, ensuring balanced representation. Digestion mode and temperature regime labels provide essential process categorization, and the sample numbers indicate adequate coverage of both mesophilic and thermophilic conditions. Methane yield class distribution (350 samples) enables effective classification and regression modeling, reflecting diverse performance outcomes rather than biased clustering. The operational state label (optimal/suboptimal) captures process stability features that are crucial for predicting system failures or low-yield events. The structured and evenly-distributed labeling ensures that the dataset is suitable for supervised machine learning tasks without imbalance-related biases.

Table 5. Performance of ML Models Using the Proposed Dataset

Model	RMSE	MAE
Random Forest	16.2	12.7
XGBoost	14.8	11.3
ANN (3-layer)	18.5	13.9
SVR (RBF kernel)	21.2	16.5
Linear Regression	28.4	22.1

The performance of five machine learning models trained on the developed dataset is compared in table 5. XGBoost achieves the best predictive accuracy with RMSE of 14.8 and MAE of 11.3, demonstrating its ability to capture nonlinear relationships between feedstock features and methane yield. Random Forest follows closely, also delivering strong performance due to its ensemble-based nature. The ANN model provides moderate accuracy, reflecting the need for larger datasets to fully exploit deep learning architectures. SVR and Linear Regression perform less effectively, indicating limited suitability for high-dimensional nonlinear datasets like anaerobic digestion. The performance hierarchy confirms that tree-based ensemble models are highly effective for methane prediction and validates the robustness of the proposed dataset.

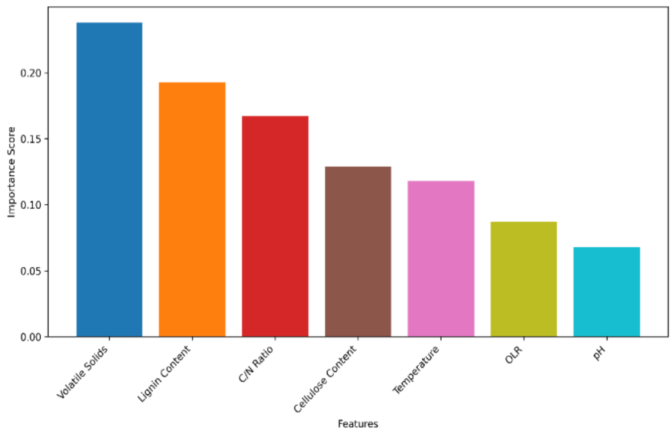

Figure 2: Importance ranking for methane yield prediction

The relative importance of key physicochemical and operational features in predicting methane yield using machine learning models is illustrated in figure 2. Volatile Solids (VS) emerge as the most influential predictor, highlighting the strong dependence of methane generation on the amount of organic matter available for microbial conversion. Lignin content appears as the second most important parameter due to its inhibitory effect on hydrolysis efficiency. The C/N ratio also plays a significant role, reflecting the importance of nutritional balance for microbial growth. Cellulose content and temperature contribute moderately, indicating their relevance for enzymatic breakdown and microbial activity. OLR and pH show relatively lower importance but still influence system performance. This ranking confirms that methane production is

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

driven primarily by feedstock composition, supported by operational stability variables.

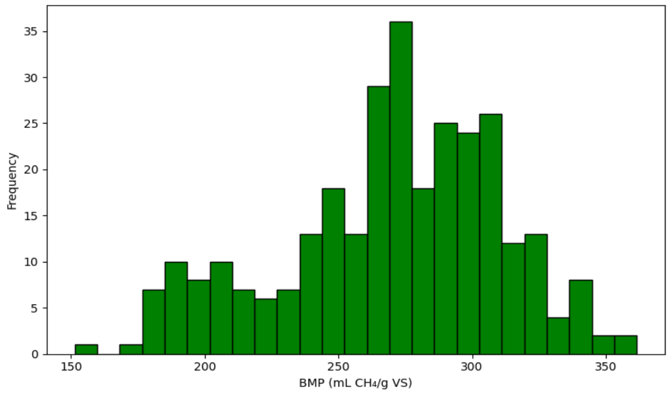

Figure 3: Distribution of methane yield (BMP)

The distribution of Biochemical Methane Potential (BMP) across all samples in the dataset is shown in figure 3. The distribution exhibits a clear multimodal pattern, representing the inherent variability in methane productivity among different agricultural residues. A significant proportion of samples cluster around moderate BMP values, corresponding to biomass types such as rice straw and maize stover. Higher BMP values indicate more degradable substrates like wheat straw, while lower values reflect recalcitrant feedstocks such as paddy husk and sugarcane bagasse. The distribution curve demonstrates that the dataset is well-balanced, capturing both low- and high-yield feedstocks. This wide variability strengthens the robustness of the machine learning models and enhances their ability to generalize across diverse biomass types.

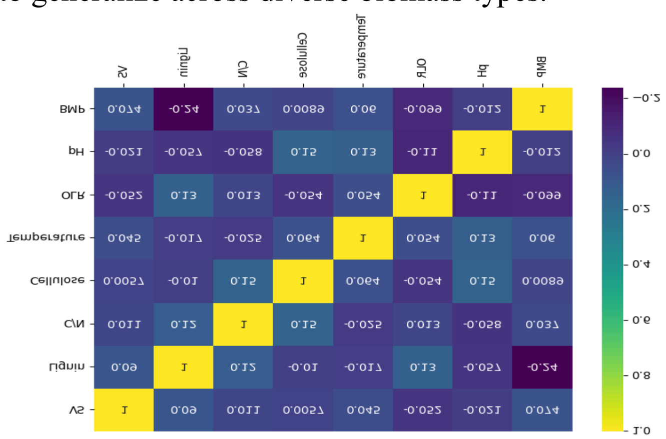

Figure 4: Correlation heatmap of key features

The correlation heatmap illustrating the relationships among major physicochemical, operational, and performance features is given in figure 4. A strong positive correlation is observed between volatile solids and BMP, confirming that higher organic content enhances methane yield. Lignin shows a significant negative correlation with BMP, reinforcing its well-known inhibitory effect on biodegradability. The C/N ratio exhibits moderate positive correlation with methane production, demonstrating the importance of nutrient balance. Temperature shows weak to moderate correlations with both pH and BMP, reflecting its role in microbial kinetics but also suggesting that feedstock properties play a greater role than operational parameters. These correlations validate the feature selection approach and highlight the multidimensional interactions governing anaerobic digestion performance.

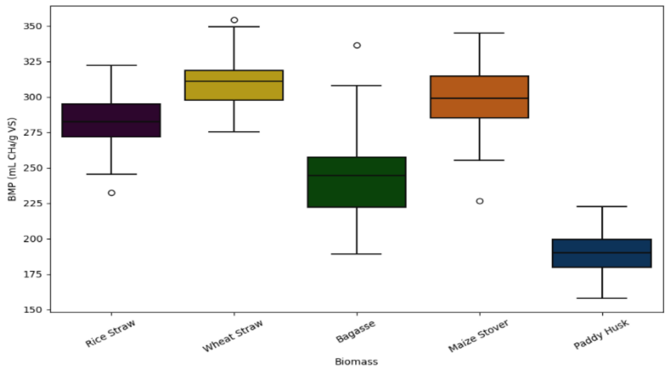

Figure 5: BMP across different biomass types

The BMP distribution across the five major biomass types included in the dataset is compared in figure 5. Wheat straw exhibits the highest median BMP, reflecting its favorable cellulose-to-lignin ratio and higher biodegradability. Maize stover and rice straw also show relatively strong methane production, indicating balanced compositional characteristics. Sugarcane bagasse shows greater variability due to its fibrous structure and heterogeneous composition. Paddy husk exhibits the lowest BMP values across

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

all samples, which is consistent with its high lignin and silica content that restrict microbial access during anaerobic digestion. Overall, this figure highlights the clear differences in methane potential among agricultural residues and emphasizes the need for feedstock-specific process optimization strategies.

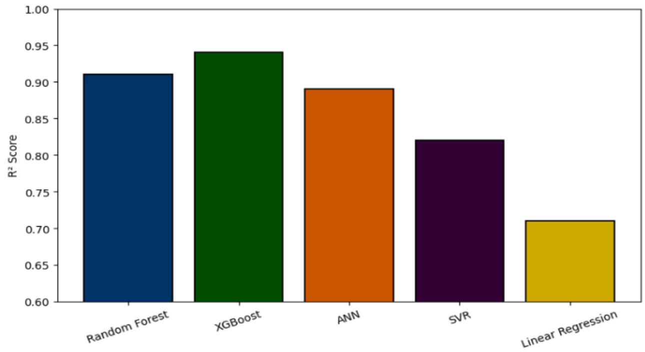

Figure 6: Performance comparison of ML models

The performance comparison of various machine learning models trained for methane yield prediction is depicted in figure 6. XGBoost achieves the highest predictive accuracy, demonstrating its strong capability in modeling nonlinear interactions and handling complex datasets. Random Forest also performs well, reflecting the effectiveness of ensemble-based learning for biological systems. ANN exhibits moderate performance, consistent with the limited dataset size, which restricts deep learning models from achieving optimal generalization. SVR and Linear Regression perform comparatively poorly, indicating that simpler models struggle to capture the nonlinear biochemical and operational relationships inherent in anaerobic digestion. This performance hierarchy confirms that advanced tree-based models are better suited for methane prediction tasks using multi-dimensional AD datasets.

Overall, the combined findings from the tables, figures, and model evaluations emphasize the strong influence of feedstock composition, operational stability, and engineered features on methane productivity. The ability of advanced machine learning models—particularly XGBoost and Random Forest—to accurately predict methane yield further demonstrates the robustness and analytical value of the proposed dataset framework. The broad variability observed across biomass types and digestion conditions reinforces the importance of standardized data acquisition and labeling practices, as implemented in this study. Collectively, the results confirm that the developed dataset offers a reliable and scalable foundation for methane prediction research, supports algorithm development for biogas optimization, and provides meaningful insights for designing intelligent and efficient anaerobic digestion systems.

5. Conclusion

This study presented a novel and systematic dataset framework for methane yield analysis in anaerobic digestion of major agricultural biomass residues. By integrating comprehensive feedstock characterization, operational digester parameters, and biogas performance indicators, the proposed framework addressed a critical gap in the availability of consistent, labeled, and machine-learning–ready datasets for anaerobic digestion research. The constructed dataset, comprising more than 500 samples across five commonly utilized agricultural residues, enabled statistically meaningful insights into the influence of VS, lignin content, C/N ratio, and process conditions on biomethane production. The predictive analyses demonstrated that advanced machine learning models—particularly XGBoost and Random Forest—achieve high accuracy when trained on the curated feature set, confirming the dataset's robustness and relevance for data-driven methane optimization. The results also validated the strong correlation between feedstock composition and methane yield, offering valuable direction for substrate selection and process tuning. Overall, the dataset framework developed in this work establishes an essential foundation for future

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

research aimed at optimizing anaerobic digestion, designing intelligent AD monitoring systems, and enabling AI-driven real-time control strategies. Future extensions may include expanding the dataset to additional biomass sources, incorporating microbial community data, and integrating continuous real-time digester monitoring to further enhance model generalizability and operational applicability.

References

1. Yin, Z., Zhou, S., Zhang, X., Li, X., Wang, Z., Wang, J., Cao, W., & Sun, C. (2023). A Novel Batched Four-Stage–Two-Phase Anaerobic Digestion System to Facilitate Methane Production from Rice Straw and Cow Manure with Low Inoculum/Substrate Ratios. Fermentation, 9(6), 565. https://doi.org/10.3390/fermentation9060565
2. Wang, J.; Ma, D.M.; Lou, Y.; Ma, J.; Xing, D.F. Optimization of biogas production from straw wastes by different pretreatments: Progress, challenges, and prospects. Sci. Total Environ. 2023, 905, 166992.
3. Archana Kasinath, Sylwia Fudala-Ksiazek, Malgorzata Szopinska, Hubert Bylinski, Wojciech Artichowicz, Anna Remiszewska-Skwarek, Aneta Luczkiewicz, Biomass in biogas production: Pretreatment and codigestion, Renewable and Sustainable Energy Reviews, Volume 150, 2021, 111509, https://doi.org/10.1016/j.rser.2021.111509.
4. Cremonez, P.A.; Teleken, J.G.; Weiser Meier, T.R.; Alves, H.J. Two-Stage anaerobic digestion in agroindustrial waste treatment: A review. J. Environ. Manag. 2021, 281, 111854.
5. Selormey, G.K., Barnes, B., Kemausuor, F., Darkwah, L. (2021). A review of anaerobic digestion of slaughterhouse waste: Effect of selected operational and environmental parameters on anaerobic biodegradability. Reviews in Environmental Science and Bio/Technology, 20(4): 1073-1086. https://doi.org/10.1007/s11157-021-09596-8.
6. Saif, I.; Thakur, N.; Zhang, P.; Zhang, L.H.; Xing, X.H.; Yue, J.W.; Song, Z.Z.; Nan, L.; Yujun, S.; Usman, M.; et al. Biochar assisted anaerobic digestion for biomethane production: Microbial symbiosis and electron transfer. J. Environ. Chem. Eng. 2022, 10, 107960.
7. Qian, S., Chen, L., Xu, S., Zeng, C., Lian, X., Xia, Z., & Zou, J. (2025). Research on Methane-Rich Biogas Production Technology by Anaerobic Digestion Under Carbon Neutrality: A Review. Sustainability, 17(4), 1425. https://doi.org/10.3390/su17041425
8. Calbry-Muzyka, A.; Madi, H.; Rüsch-Pfund, F.; Gandiglio, M.; Biollaz, S. Biogas composition from agricultural sources and organic fraction of municipal solid waste. Renew. Energy 2022, 181, 1000–1007.
9. Barahmand, Z., Samarakoon, G. (2023). Anaerobic digestion process modeling under uncertainty: A narrative review. International Journal of Energy Production and Management, Vol. 8, No. 1, pp. 41-54. https://doi.org/10.18280/ijepm.080106
10. Pal, D.B.; Tiwari, A.K.; Mohammad, A.; Prasad, N.; Srivastava, N.; Srivastava, K.R.; Singh, R.; Yoon, T.; Syed, A.; Bahkali, A.H.; et al. Enhanced biogas production potential analysis of rice straw: Biomass characterization, kinetics and anaerobic co-digestion investigations. Bioresour. Technol. 2022, 358, 127391.
11. Duan, Y.; Gao, Y.; Zhao, J.; Xue, Y.; Zhang, W.; Wu, W.; Jiang, H.; Cao, D. Agricultural methane emissions in China: Inventories, driving forces and mitigation strategies. Environ. Sci. Technol. 2023, 57, 13292–13303
12. Viktoria Wechselberger, Marlies Hrad, Marcel Bühler, Thomas Kupper, Bernhard Spangl, Anders Michael Fredenslund, Marion Huber-Humer, Charlotte Scheutz, Assessment of whole-site methane emissions from anaerobic digestion plants: Towards establishing emission factors for various plant configurations, Waste Management, Volume 191, 2025, Pages 253-

Correspondence to: ***Asit Chatterjee, Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 137-147

263, https://doi.org/10.1016/j.wasman.2024.11.021.

13. Pyykkönen V, Winquist E, Seppänen AM, Vainio M, Virkkunen E, Koppelmäki K, Rasi S. Anaerobic Digestion of Solid Agricultural Biomass in Leach-Bed Reactors. Bioengineering (Basel). 2023 Mar 29;10(4):433. doi: 10.3390/bioengineering10040433. PMID: 37106620; PMCID: PMC10135786.
14. Ankathi, S., Chaudhari, U., Handler, R., & Shonnard, D. R. (2024). Sustainability of Biogas Production from Anaerobic Digestion of Food Waste and Animal Manure. Applied Microbiology, 4(1), 418-438. http://doi.org/10.3390/applmicrobiol4010029
15. Salehi, R., Yuan, Q., & Chaiprapat, S. (2022). Development of Data-Driven Models to Predict Biogas Production from Spent Mushroom Compost. Agriculture, 12(8), 1090. https://doi.org/10.3390/agriculture12081090

Correspondence to: ***Asit Chatterjee*, *Department of Civil Engineering, Suresh Gyan Vihar University Jaipur***
Corresponding author. E-mail addresses: asitchatterjee@rediffmail.com