Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 18(8), 1–15
https://doi.org/10.12913/22998624/192936
ISSN 2299-8624, License CC-BY 4.0

Received: 2024.06.16
Accepted: 2024.09.03
Published: 2024.11.01

Machine Learning – Based Prediction of Biogas Production from Sludge Characteristics in Four Anaerobic Digesters – Development of the AD2Biogas Prediction Tool

Patryk Organiściak[1], Adam A. Masłoń[2], Bartosz Kowal[1], Paweł Kuraś[1*], Bartosz Wadiak[3], Sylwia Sikorska-Czupryna[4], Veronika Vanivska[1]

[1] Department of Complex Systems, The Faculty of Electrical and Computer Engineering, Rzeszow University of Technology, ul. MC Skłodowskiej 8, 35-036 Rzeszów

[2] Department of Environmental Engineering and Chemistry, Faculty of Civil and Environmental Engineering and Architecture, Rzeszow University of Technology, Powstańców Warszawy 6, 35-959 Rzeszow, Poland

[3] Miejskie Przedsiębiorstwo Wodociągów i Kanalizacji Sp. z o.o. w Rzeszowie, ul. Naruszewicza 18, 35-055 Rzeszów, Poland

[4] Department of Computerization and Robotization of Industrial Processes, Faculty of Mechanical and Technological Engineering, Rzeszow University of Technology, Kwiatkowskiego 4, 37-450 Stalowa Wola, Poland

* Corresponding author's e-mail: p.kuras@prz.edu.pl

ABSTRACT

One of the alternative ways to obtain low-cost energy is to use biogas generated by the digestion process from sewage sludge. This paper presents an analysis of the processes in four anaerobic digesters (AD) – A, B, C and D. The study analyzed the amount of biogas produced in each digester tank and compared them with each other. Using data sets consisting of parameters relating to the pre-sludge and surplus sludge diverted to each tank, the effect of the proportion of these parameters on biogas production efficiency was studied. Based on this data, several models using different machine learning techniques were built and compared, which can be used to support the biogas production optimization process. A free convenient web tool written in Python language – AD2Biogas Predictor Tool - was also given away for sewage treatment plants to conveniently estimate the predicted amount of biogas produced on a given day using the implemented models. The main objective of the study is to understand how the studied parameters affect the efficiency of the process and identify potential optimization strategies, as well as to propose a model for biogas yield prediction based on sludge characteristics. The result of the study is to contribute to increasing the efficiency of sludge management in wastewater treatment plants and increasing biogas production, both in the form of developed models and a software tool.

Keywords: sewage sludge, biogas, anaerobic fermentation, machine learning, python, sewage plant, software tool.

INTRODUCTION

Biogas production is a well-explored technology used to generate renewable energy and manage organic waste through anaerobic processes. Nowadays, the biogas sector is growing rapidly, promoting a circular economy through the recycling of nutrients, reduction of greenhouse gas emissions and biorefinery goals, and a review of the current state and future prospects reveals the potential for optimizing the digestion process [1, 2]. Biogas is a gas produced from biomass, extracted from the anaerobic decomposition of organic compounds by the process of methane fermentation. It is a multi-step biochemical process that has a direct relationship to the chemical composition of the starting substrate, i.e., the organic compounds entering the digester and the bacteria involved in the process.

The substrate for fermentation can be any biodegradable compound, but they differ in their decomposition rates and methane yields [3, 4].

Nowadays, biogas is a commonly used resource obtained as a result of waste digestion. It may be used to generate electricity for the internal needs of wastewater treatment plants, but its excess can also be sold [5–8]. In addition, the combustion of biogas in boilers or combined heat and power (CHP) units allows for the generation of heat [9]. Another potential way to manage biogas is to inject it into the gas network or use it as vehicle fuel [10]. Combustion of biogas in CHP units enables economic energy production (1 m³ of biogas allows the generation of about 2.1 kW of electricity and 2.9 kW of heat energy) and provides higher efficiency of the system, compared to separate heat and power generation [11]. The treatment of wastewater does not only result in biogas. The process also produces sewage sludge, which can be utilized in various industries. The results of studies on the mechanical properties of concrete mixtures modified with post-fermentation sludge have demonstrated the potential for using these residues in concrete mixes [12]. Another area where sewage sludge can be used is in the production of unconventional material for road construction, as demonstrated in the studies by Wojcik (2018) [13], which used dewatered sewage sludge with a moisture content of 62%, glass powder, and quartz sand. Sewage sludge residue can be repurposed as a building material in the form of pellets [14]. The referenced study details the process of converting sewage sludge into pellets, which have potential applications in construction, particularly as thermal insulation or as a substitute for conventional building materials.

A reasonable approach to treating sewage sludge is to use it as a substrate for biogas production in the methane fermentation process. Biogas is a mixture of gases, mainly methane and carbon dioxide, which can be used to produce electricity and heat. In addition, the process of anaerobic digestion leading to the formation of biogas is beneficial for dehydrating sludge and reducing its volume. Given the numerous benefits of biogas production, the energy potential of sewage sludge is steadily increasing [15–17]. This approach can lead to an improvement in the energy efficiency of the entire treatment plant and even allow it to become self-sufficient [18–20]. Analyzing the amounts of sludge processed in digestion processes, it may be concluded that the energy potential inherent in sludge is utilized in less than 40% [21–24]. For comparison, in Croatia, about 30% of sewage sludge is processed through digestion into biogas [25]. It is estimated that up to 170 installations for biogas production from sewage sludge could still be built in wastewater treatment plants in Poland [26], and every district in Poland has a sufficient amount of substrates for biogas production [27]. Sludge treatment for biogas production consists of sludge pretreatment (thickening, conditioning), anaerobic decomposition of organic matter in methane digestion, biogas treatment (desulfurization) and its conversion to electricity or heat, and post-treatment of digested sludge (dehydration, drying). A simplified diagram of the biogas production process is presented in Figure 1.

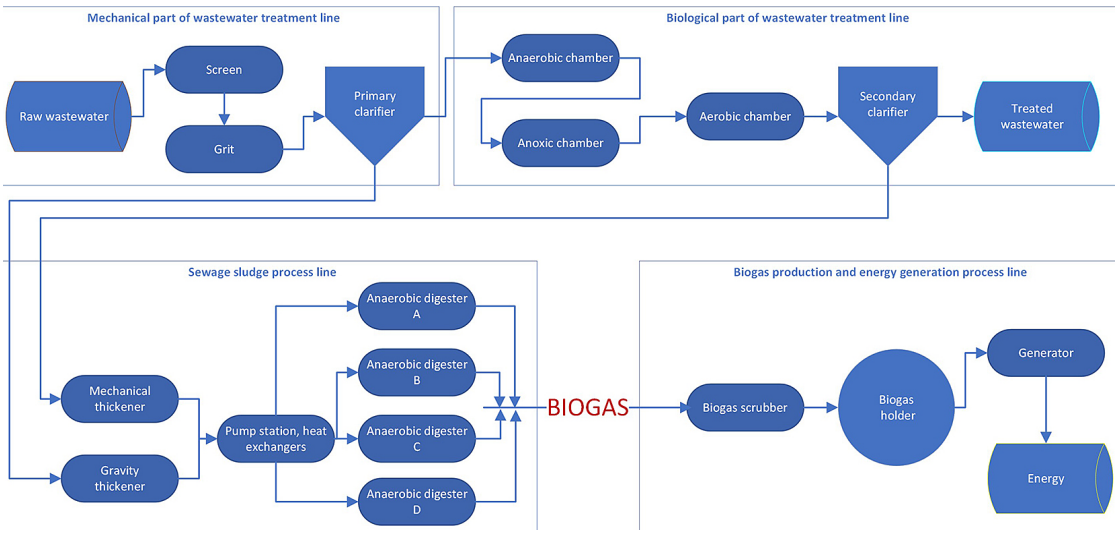

Figure 1. Process of producing biogas from wastewater with four anaerobic digesters

The classic arrangement for biogas generation in a wastewater treatment plant is one in which the sludge is initially thickened, and then pumped into the separated or enclosed digesters, where the mesophilic fermentation process takes place in the temperature range of 35–39 °C. The holding time is about 20 days, and this is the period during which biogas is produced with the participation of microorganisms [3]. The most common in biogas production is mesophilic fermentation carried out at a temperature of 35 °C. In such a system, organic substances decompose up to 40% within a retention time of 30 to 40 days. It is crucial that the thermal conditions remain stable; otherwise, there may be a development of non-methanogenic bacteria populations, which would significantly reduce the methane content in the biogas [28]. According to literature, raw sludge can produce about 315–400 Nm³ of methane per ton of dry organic matter, while for excess sludge this figure is lower, at 190–240 Nm³ of methane [29]. However, in terms of the amount of wastewater, 1000 m^3 of municipal wastewater can yield 100–200 Nm^3 of biogas. It is also estimated that 1 m^3 of sludge (with a dry matter content of 4–5%) can produce 10–20 m^3 of biogas with a 60% methane content [3].

The amount of biogas produced largely depends on process conditions and the characteristics of the wastewater load. Flexible (demand-driven) biogas production is possible both in conventional anaero- bic digestion (AD) and in advanced AD using thermal hydrolysis process (THP), yielding better results compared to using a steady operational regime [30] [31]. Biogas production is often analyzed using advanced mathematical models and artificial intelligence, emphasizing the importance of process controllability, including substrate selection. Anaerobic co-fermentation of activated sludge with wheat straw (a mixture with a 7% ratio) increased biogas production 15-fold compared to monofermentation of the sludge. This was confirmed by mathematical models (TDMMs) and ANN [32]. In another study, it was demonstrated using artificial neural net- work (ANN) that organic loading rates (OLRs) obtained at optimal levels have a significant impact on increasing biogas production [33]. In the study by Qdais et al. operational parameters of the fermentation chamber, such as temperature (T), total solids content (TS), total volatile solids content (TVS), and pH, were used. Using these data and the ANN algorithm, the effectiveness of the model in predicting methane production was demonstrated, with a correlation coefficient of 0.87 [34]. A variety of substrates can be used to improve biogas yields, such as maize silage, slurry, and distillery waste [15], food wastes [35], sorghum [36], and brewery sludge [37]. The optimal selection and proportion of different substrates, supported by advanced mathematical models and artificial intelligence, allow for the maximization of biogas yield and improvement of the overall energy efficiency of the process [38].

In the study by Wang et al. [39], the k-nearest neighbors (KNN) algorithm showed the highest prediction accuracy in regression models, achieving a root mean square error (RMSE) of 26.6, with the dataset values ranging from 259.0 to 573.8, after excluding extreme outliers from the validation set. The study [40] develops a three-layer ANN and nonlinear regression models to predict biogas production rates from an anaerobic hybrid reactor (AHR) under various conditions. The results demonstrate that both models, particularly the ANN (RMSE 217.4, range: 1510–8084), accurately predict biogas production, providing valuable insights for optimizing reactor performance and improving economic and environmental sustainability in biogas production. The research [41] implemented ANN models to predict biogas production rates. The performance of these models was enhanced by selecting significant process variables using GA and ACO optimization techniques. GA-ACO-Optimized ANN Model achieved the best results with RMSE of 6.24% for test data, The review [42] discusses the application of machine learning (ML) in AD pro- cesses, highlighting its potential for optimizing, predicting, and stabilizing biogas production. Various models such as ANN, adaptive neuro-fuzzy inference system (ANFIS), and support vector machine (SVM) have been effectively used to address the complex and nonlinear nature of AD systems. The review also identifies challenges such as the need for large data sets and the ”black box” nature of some ML models, suggesting that future research should focus on improving model transparency and integrating different algorithms for better performance. Of all the methods, ANN were the most commonly used, accounting for 33% of the cases studied.

The aim of this study is to analyze sludge and biogas management in four anaerobic digesters at the wastewater treatment plant in Rzeszow using

data from 2024. The study examined the impact of input data on the amount of biogas produced in four independent digesters: A, B, C, and D. The calculations aim to determine whether the digesters operate in a similar manner and, consequently, whether prioritizing a specific digester has economic justification due to varying biogas production efficiencies in the digesters. Another objective of the study is to determine the effectiveness of multiple (22) ML algorithms in predicting the amount of biogas produced based on substrate proportions, in order to enable more flexible energy management at the wastewater treatment plant. The wastewater treatment plant in Rzeszow is a municipal facility that primarily handles domestic sewage, which distinguishes it from other studies that often focus on industrial or agricultural wastewater treatment plants. This difference in the type of wastewater treated has a significant impact on the biogas production process, as the characteristics of the sewage sludge and the resulting biogas yield can vary greatly depending on the source of the wastewater. The study utilized algorithms that are less commonly used in this type of research, compared to methods like ANN and ANFIS [42]. By focusing on a municipal plant with its specific challenges and conditions, this research provides insights that are directly applicable to similar facilities, offering a valuable contribution to the field. Additionally, approaches were applied where data from multiple tanks were analyzed independently or dependently, which is a novelty, as previous studies typically examined a single tank or treated them collectively. The final element of this work is the design of a public tool for predicting the amount of gas based on input data to support the operation of the analyzed wastewater treatment plant. Using the online tool, it is possible to estimate the amount of product based on the data for each tank individually.

MATERIALS AND METHODS

Dataset

The study utilized a dataset from the wastewater treatment plant in Rzeszow. The results were obtained for the periods from 01.01 to 01.02 and from 19.02 to 18.04.2024, comprising a total of N = 91 results. The measurements are not continuous. The dataset includes a set of operational parameters for four independent anaerobic digesters, A, B, C, and D. These parameters include information about the digesters (such as their size) and daily operation metrics, such as the amount of primary sludge (PS), the amount of excess sludge (ES), dry mass (Load A), organic dry mass (Load A'), the amount of biogas produced, and biogas yield per cubic meter of sludge. Daily measurements of the wastewater treatment plant's operation indicate that the PS to ES ratios for individual anaerobic digesters are different, which, in turn, should have an impact on the different amounts of biogas production. The characteristics of these data are provided on Figure 2. To evaluate the ML models and feature correlation, the data presented in Table 1 were used as input. There were 108 days during which data were collected. Each data entry contains separate information for the

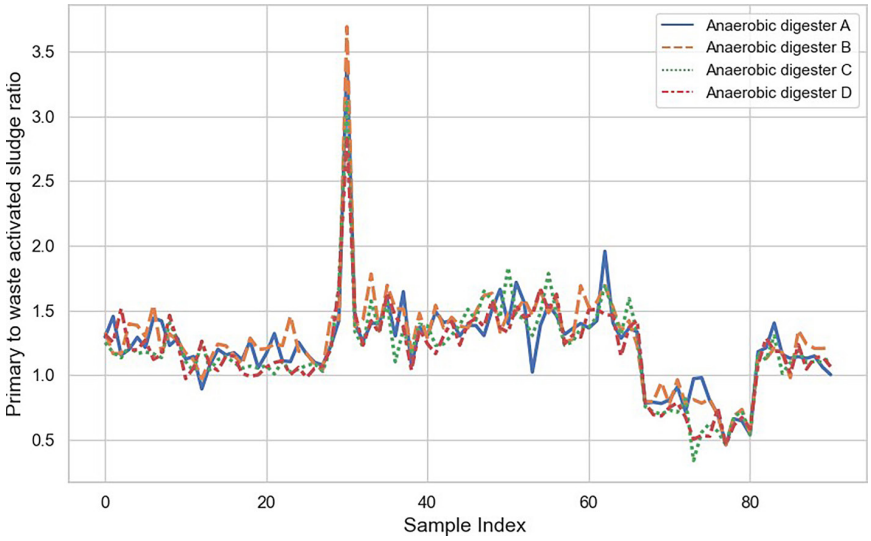

Figure 2. The ratio of primary sludge to excess sludge for 4 anaerobic digesters

Table 1. List of regression models used in each of three proposed variants

No.	Model	No.	Model
1	Linear regression	12	Extra tree regressor
2	Ridge regression	13	Random forest regressor
3	Lasso regression	14	Gradient boosting regressor
4	ElasticNet regression	15	Adaboost Regressor
5	Bayesian ridge regression	16	Bagging regressor
6	SGD regressor	17	Extra trees regressor
7	Huber regressor	18	Support vector regressor
8	Passive aggressive regressor	19	Linear SVR
9	Theil-Sen regressor	20	Kernel ridge regressor
10	MLP regressor	21	Gaussian process regressor
11	Decision tree regressor	22	K-Neighbors regressor

four AD tanks. The PS+ES and PS/ES columns, although they appear to be redundant, were not removed from the data frame. In scenarios where features from all digesters were mixed or validation was performed on the combined features, the feature set also included a column indicating the specific digester as a char (A, B, C or D).

Methods

In this study, three data validation variants were employed to evaluate the effectiveness of biogas production prediction. Calculations were performed for all variants, and the best-fitting algorithm was selected. In addition, each variant was carefully analyzed, and the predictions in each case were presented on diagrams. To evaluate the ML models, the MSE was adopted as the performance metric.

First variant (I): Each anaerobic chamber was analyzed individually, and biogas production was pre- dicted independently for each chamber. The leave-one-out (LOO) validation method was used. This variant aimed to determine whether the regression performance depends on the specific anaerobic chamber and whether the same algorithms would provide similar results in different chambers.

Second variant (II): The leave-one-subject-out (LOSO) validation method was applied, where the regression performance was tested on an anaerobic chamber that was not included in the training process. This approach aimed to assess whether the characteristics of the chambers were similar and if the model could generalize well to an unseen chamber. Third variant (III): All data from the four anaerobic chambers were combined, and the Leave-One-Out validation method was used. This approach allowed for a fourfold increase in the number of samples and aimed to mitigate the potential variability in chamber characteristics, providing a more robust assessment of the model's predictive performance.

MSE has been chosen as the evaluation metric for our models due to its widespread recognition in the literature, which is intuitively understandable and easy to interpret, facilitating the presentation and comprehension of results. MSE is sensitive to large errors, enabling the identification of models that minimize significant deviations from actual values. Furthermore, it can be easily converted to RMSE, providing an additional perspective on model accuracy. In our study, we are aware that the data may not be fully precise and that accuracy may vary between digesters. Therefore, it is more important to limit large errors, even at the cost of increasing the average error. In this study, various Python libraries were utilized to support the data analysis, machine learning model development, and evaluation processes. Specifically, pandas was used for data manipulation and analysis, while scikit-learn provided tools for implementing and evaluating multiple machine learning models, including linear regression, neural networks, decision trees, random forests, and support vector machines. Additionally, numpy was employed for numerical operations, and matplotlib was used for visualizing the results. The study also leveraged advanced cross-validation techniques, such as leave-one-out, to rigorously assess model performance.

For each anaerobic digester, correlations were examined and the significant components for each specific digester were identified. Results included in Table 2 indicate that the parameters studied have a significant, but not always the same, effect

on specific recipients. This information forms the basis for considering whether all anaerobic digesters operate in the same way and whether it is worth considering the proposed calculation variants.

In Table 1, 22 machine learning algorithms are presented, which were initially used for the evaluation of the digesters. Each algorithm was applied to analyze the performance of individual digesters based on various parameters. The results of these evaluations help identify the most effective models for predicting biogas production. Each method was run with the default settings of the sklearn library, with only the n estimators and max iter parameters changed to 1000. Random forest regressor is an algorithm that creates multiple decision trees during training and derives predictions by averaging the results of each tree (in the case of regression) [43]. For the random forest regressor, the n-estimator's value means that the random forest consists of 1000 decision trees. This is a relatively large number, which usually leads to more stable and accurate predictions because a higher number of trees better averages the predictions and reduces model variance. The parameter was tested in configurations of 10, 100, 500 and 1,000, and the best one was chosen. Similarly, in the Extra Trees Regressor, this value represents the number of decision trees in the model. Extra trees differs from Random Forest in that it randomly selects the split threshold at each node of the tree, leading to greater randomness. [44] In the Bagging Regressor, n estimators also refers to the number of base models (by default, decision trees in sklearn) that are trained on different bootstrap samples of the data. A higher number of estimators increases accuracy. The final result is an average of the results of all models [45].

RESULTS

Regression for each anaerobic digesters independently (Variant I)

The first variant, where each anaerobic chamber was tested independently, aims to determine whether the same methods are suitable for each chamber and whether their effectiveness will be similar. Figures 3, 4, 5, and 6 contain the

Table 2. Pearson correlation coefficient for selected features of anaerobic digesters dataset

Parameter	AD A	AD B	AD C	AD D
Biogas [m³/d]	1.000000	1.000000	1.000000	1.000000
PS [m³/d]	0.066277	0.053039	0.340254	0.341549
ES [m³/d]	-0.008963	0.030069	0.284818	0.308429
PS+ES [m³/d]	0.043022	0.054525	0.381637	0.384922
Load A [kg/m³]	-0.233100	-0.234764	0.129208	0.152170
Load A' [kg/m³]	-0.224789	-0.211233	0.143224	0.175614
PS / ES	0.196363	0.091932	0.174754	0.132134

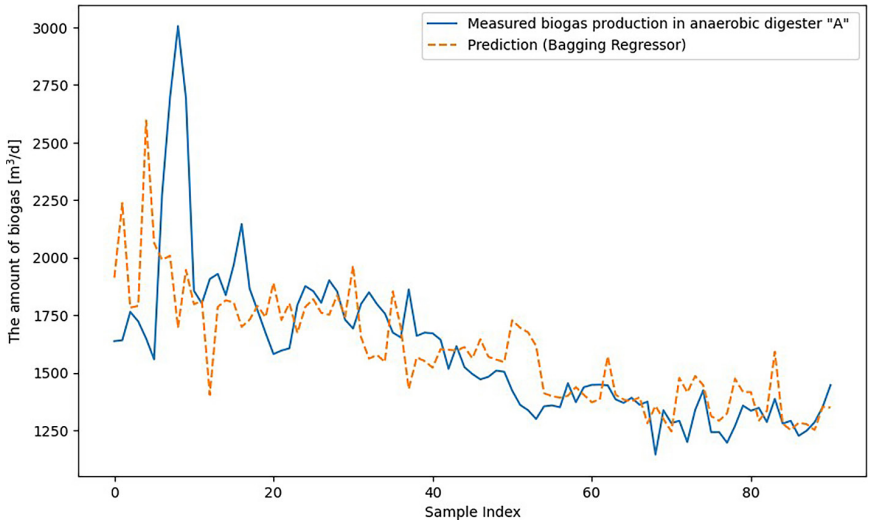

Figure 3. Prediction for biogas production in AC "A" using bagging regressor, MSE: 71473

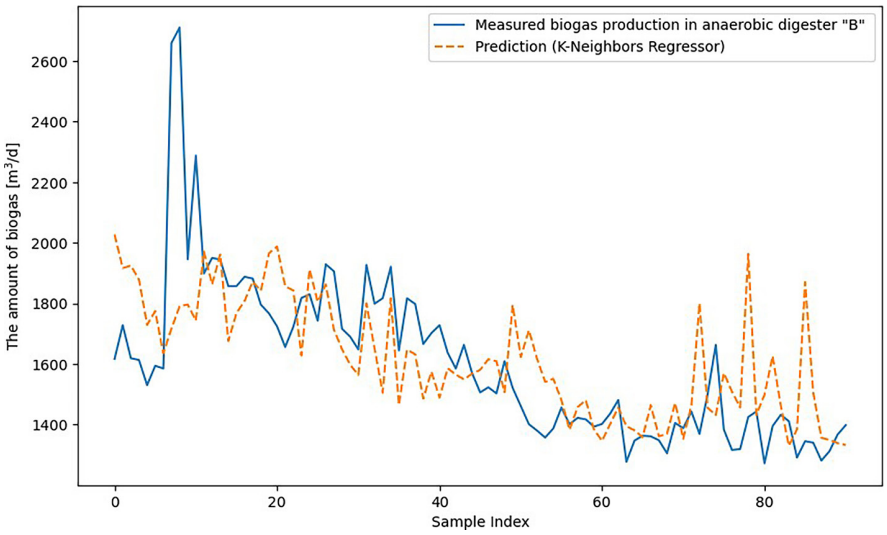

Figure 4. Prediction for biogas production in AC ”B” using k-neighbors regressor, MSE: 51071

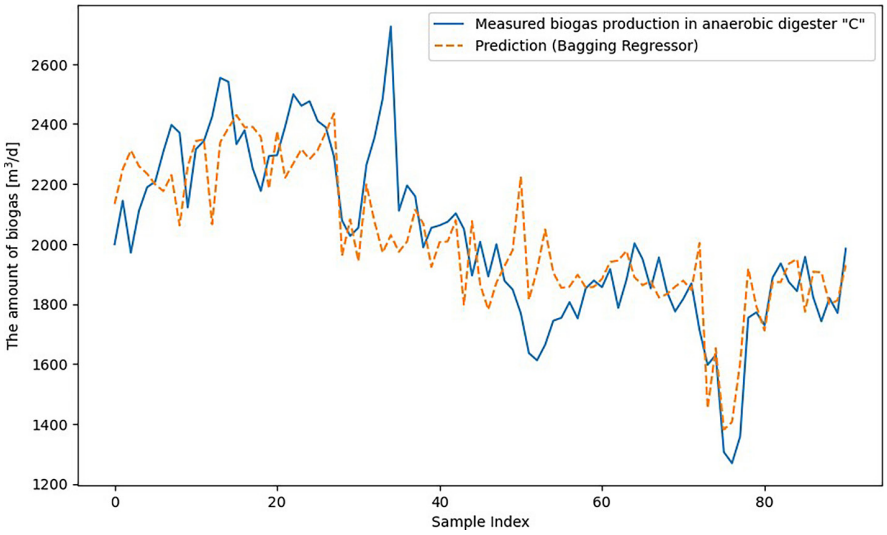

Figure 5. Prediction for biogas production in AC ”C” using bagging regressor MSE: 31040

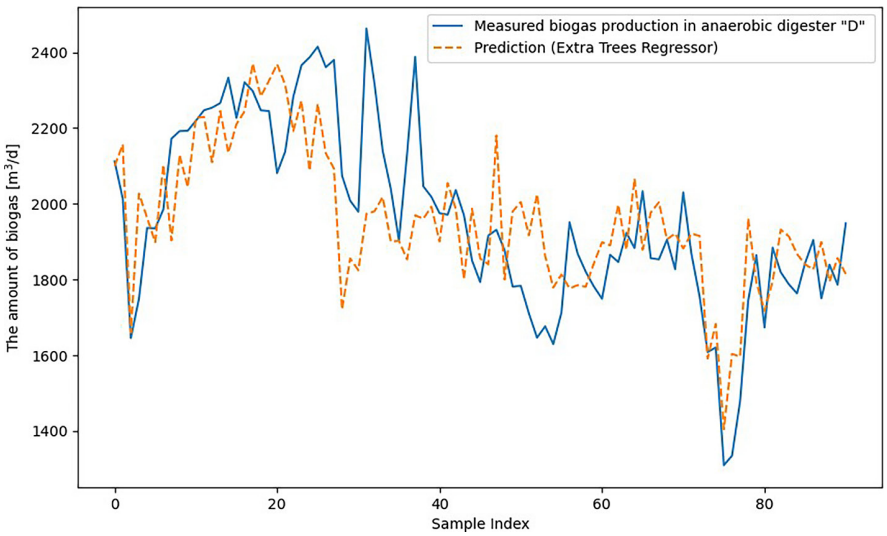

Figure 6. Prediction for biogas production in AC ”D” using extra trees regressor MSE: 27133

Table 3. Average MSE for each regression method across all chambers

Method	Average MSE
Bagging regressor	46136.10
Random forest regressor	46126.20
Gradient boosting regressor	48297.91
K-Neighbors regressor	62380.02
Extra trees regressor	49108.96
AdaBoost regressor	39923.97

predicted values for the individual datasets. To determine which method was best overall, the average MSE values for each method from all four tanks and compare these averages was calculated. Average MSE is presented as a summary in Table 3. The results indicate that while certain methods, such as Random Forest Regressor and Bagging Regressor, consistently perform well across different chambers, there are variations in their performance. For example, the Random Forest Regressor had the lowest average MSE of 46126.20, suggesting it was the best method overall. Bagging Regressor was very close with a score of 46136.10. AdaBoost Regressor demonstrated the lowest MSE in the variants where it was applied, with a value of 39923.97, but it was not used across all chambers, so this comparison may not be fully representative. These findings suggest that while some algorithms may generally perform well, their performance can still vary depending on the specific characteristics of each chamber. This underscores the importance of tailored approaches when optimizing biogas production predictions for different anaerobic chambers.

Prediction using the leave-one-subject-out (LOSO) for 4 anaerobic digesters (Variant II)

Using the LOSO, The extra trees regressor algorithm outperformed other mod- els with the lowest mean MSE, making it the most suitable model for predicting outcomes in unseen anaerobic chambers. In Table 4, the top 5 algorithms with the lowest average MSE errors are presented. The individual columns contain the MSE error calculations for cases where the specific set was the test set and the average error value for the given algorithm. In Figures 7, 8, 9 and 10, the predicted values for each of the anaerobic digesters and the algorithm with the lowest MSE error

Table 4. MSE values of different regression models for the leave-one-subject-out (LOSO)

Model	AD A	AD B	AD C	AD D	Mean
Extra trees regressor	59532.49	83049.72	80335.66	32974.38	63973.06
Bagging regressor	61025.31	48411.45	199939.32	27686.68	84265.69
Random forest regressor	60703.95	48717.10	200236.83	27470.19	84282.02
AdaBoost regressor	93803.80	67780.00	169001.98	31800.44	90596.56
Lasso regression	95798.70	86715.74	104989.74	84865.07	93092.31

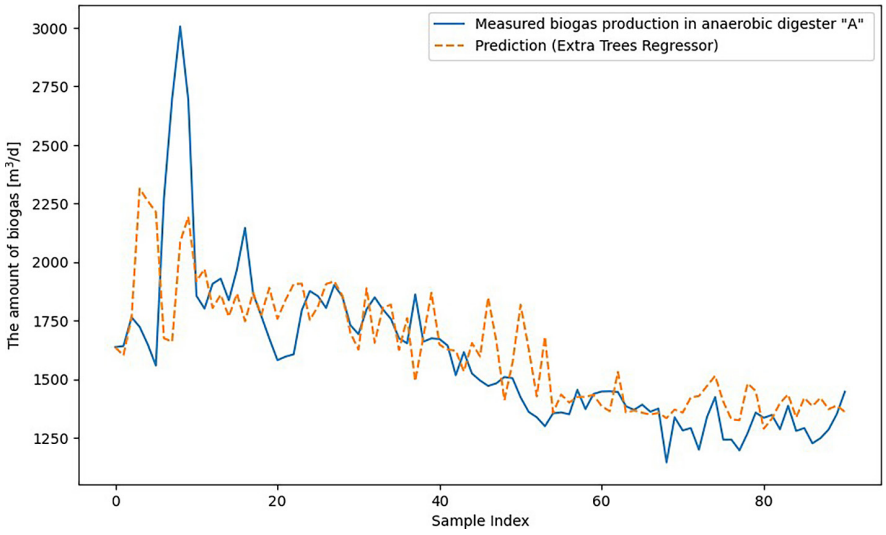

Figure 7. Prediction for biogas production in AD ”A” using Extra Trees Regressor, MSE: 59532.49

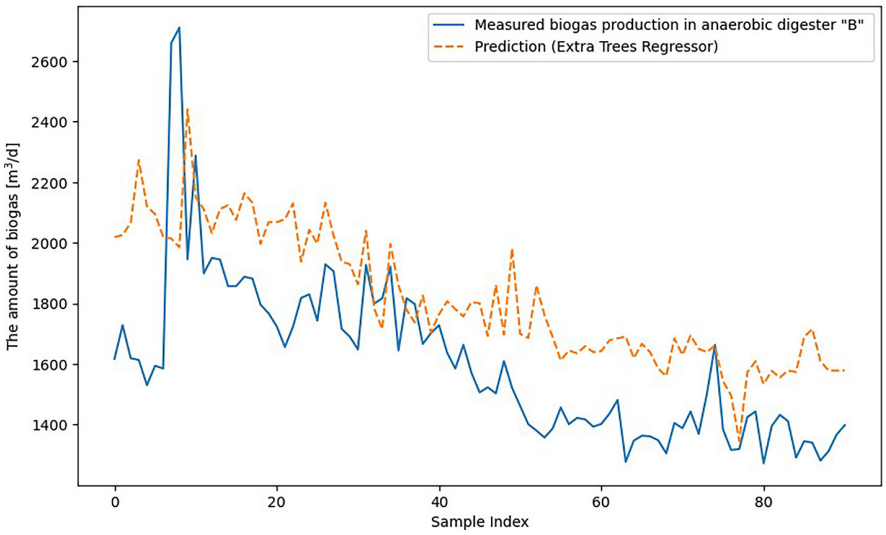

Figure 8. Prediction for biogas production in AD ”B” using extra trees regressor, MSE: 83049.72

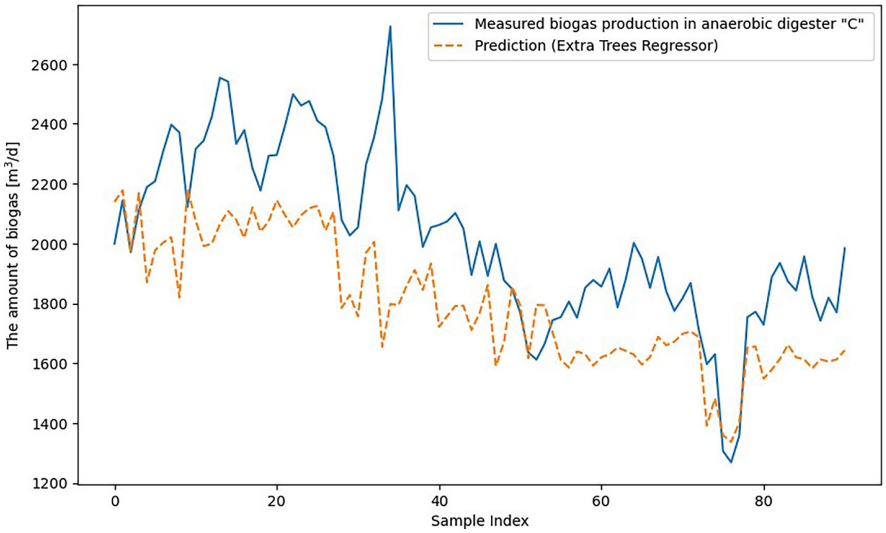

Figure 9. Prediction for biogas production in AD ”C” using extra trees regressor, MSE: 80335.66

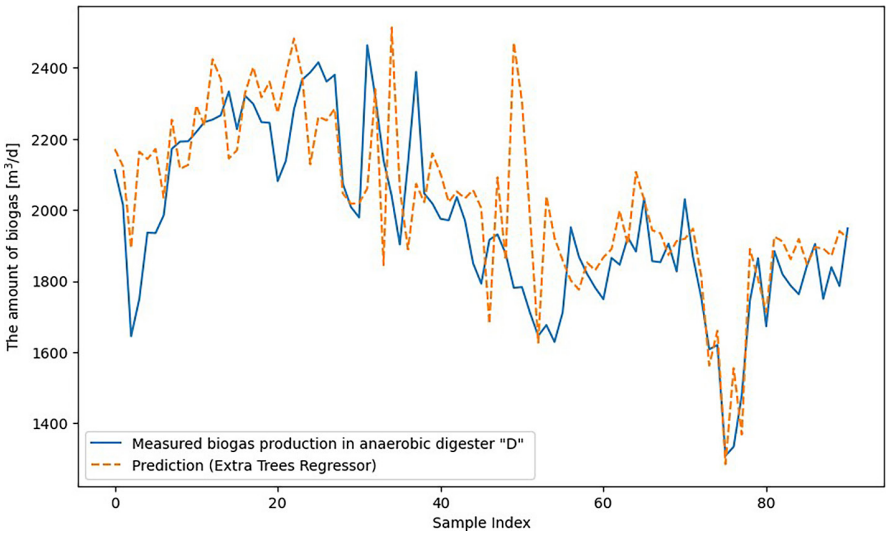

Figure 10. Prediction for biogas production in AD ”D” using extra trees regressor, MSE: 32974.38

are shown. Fact that the results obtained using the LOSO method are superior in one of the datasets indicates several potential aspects. The specific anaerobic chamber used as the test set may have characteristics more similar to the training chambers, enabling better generalization and prediction accuracy. Additionally, the model might exhibit robustness in handling the variability and nuances of that particular test chamber, suggesting effective learning of underlying patterns consistent across similar chambers. The data from this test chamber could also be of higher quality and consistency, contributing to improved prediction accuracy, especially if it is less noisy or more representative of the training data. However, significantly better performance on one test set raises concerns about potential overfitting to certain data patterns, underscoring the necessity for the model to generalize well to new, unseen data. Furthermore, the experimental conditions for data collection in that chamber might be more controlled or consistent, reducing variability and enhancing model performance. These superior results highlight the im- portance of assessing model performance across diverse and varied test sets to ensure robust generalization and avoid overfitting to specific data patterns.

Table 5. Comparison of different regression models and their mean squared error (MSE)

Model	MSE
Random forest regressor	44519.72
Bagging regressor	44597.86
Extra trees regressor	47911.41
Gradient boosting regressor	51887.12
AdaBoost regressor	52291.93

Regression for combined data from four anaerobic digester

In the Third Variant, where all data from the four chambers are combined, leave-one-out cross-validation was used as the validation method for each machine learning algorithm, providing a more robust assessment by leveraging the increased sample size and mitigating potential variability between chambers. Table 5 summarizes the performance of different evaluated regression models. The Random Forest Regressor achieved the lowest MSE among the considered models. Predicted values are shown in Figure 11 using sample origins.

Discussion

The study addressed the prediction of biogas production based on parameters from a wastewater treatment plant in Rzeszow (Table 6). For the calculations, three variants were used: each anaerobic chamber was analyzed independently using LOO cross-validation (First Variant). The LOSO validation method was applied, where the regression performance was tested on an anaerobic chamber that was not included in the training process (Second Variant). Finally, all data from the four anaerobic chambers were combined, and leave-one-out cross-validation was applied (Third Variant). Each of these validation strategies has its own merits and limitations. The First

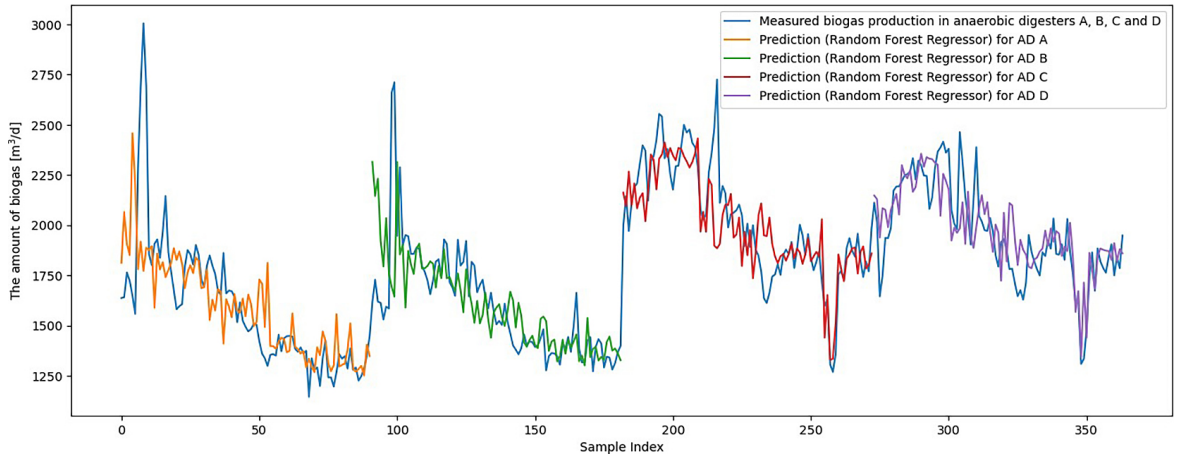

Figure 11. Prediction for biogas production using random forest regressor where samples from four anaerobic digester are merged, MSE: 44519.72

Table 6. Statictics of data provided by treatment plants in Rzeszow collected in 2024

Parameter	Count	Mean	Std	Min	25%	50%	75%	Max
Biogas [m^3/d]	364.0	1793.255	345.583	1145.0	1503.75	1793.5	1984.25	3005.0
PS [m^3/d]	364.0	76.630	9.073	35.1	73.1	78.1	80.1	97.3
ES [m^3/d]	364.0	46.654	6.631	25.5	41.8	47.6	51.9	61.3
PS + ES [m^3/d]	364.0	123.284	12.770	61.0	118.875	125.4	131.2	147.0
Load A [kg/m^3]	364.0	2.371	0.383	0.98	2.13	2.38	2.64	3.12
Load A' [kg/m^3]	364.0	1.828	0.289	0.77	1.64	1.86	2.03	2.44
PS / ES	364.0	1.226	0.356	0.342	1.075	1.231	1.402	3.691
Anaerobic digester	364.0	2.5	1.120	1.0	1.75	2.5	3.25	4.0

Table 7. Top 5 regression methods for anaerobic chambers A, B, C and D based on MSE

Rank	Chamber A	Chamber B
1	Bagging regressor (71472.89)	K-Neighbors regressor (51070.89)
2	Random forest regressor (71652.81)	AdaBoost regressor (53264.62)
3	Gradient boosting regressor (73055.35)	Random forest regressor (53905.84)
4	K-Neighbors regressor (73689.14)	Bagging regressor (54140.13)
5	Extra trees regressor (73737.89)	Extra trees regressor (61993.55)
Rank	Chamber C	Chamber D
1	Bagging regressor (31039.58)	Extra trees regressor (27132.87)
2	Random forest regressor (31113.47)	Random forest regressor (27834.69)
3	Extra trees regressor (33569.52)	Bagging regressor (27893.80)
4	AdaBoost regressor (37322.93)	AdaBoost Regressor (29182.37)
5	Gradient boosting regressor (41371.68)	Gradient boosting regressor (30466.70)

Variant, while offering insights into the performance of models on individual chambers, may suffer from limited data, leading to overfitting and reduced generalizability. The Second Variant, using LOSO, provides a more stringent test of model robustness by ensuring that one chamber is entirely unseen during training. However, it assumes a degree of similarity between chambers that may not exist, potentially leading to lower predictive accuracy if the chambers differ significantly in their operational characteristics. The Third Variant, which combines all data, maximizes the training data available to the model, offering potentially higher stability and generalization, but at the cost of possibly averaging out unique chamber-specific characteristics, which might be crucial for accurate predictions.

When comparing our results to those in the literature, several observations emerge. Wang et al. [39] reported an RMSE of 8.45% using the K-nearest neighbors (KNN) algorithm. This method's success highlights the importance of local data structures in predicting biogas production. Tufaner [40], using an ANN, achieved a significantly lower RMSE of 3.31%, indicating the superior capability of ANN models in capturing the complex nonlinear relationships inherent in biogas production processes. Beltramo et al. [41] further improved upon these results, recording an RMSE of 6.24% by employing an ANN optimized through genetic algorithms (GA) and ant colony optimization (ACO), showcasing the potential of hybrid optimization techniques in enhancing model performance.

Our study, with RMSE values of 11.34% using the Random Forest Regressor in the LOO validation scheme and 8.86% using the Extra Trees Regressor in the LOSO validation, did not achieve the lower error rates reported in the aforementioned studies. This discrepancy can be attributed to several factors. Firstly, the complexity and variability of the data across different anaerobic digesters in our study posed a significant challenge. Unlike the more homogeneous datasets used in other studies, our data varied significantly between digesters, leading to increased difficulty in model training

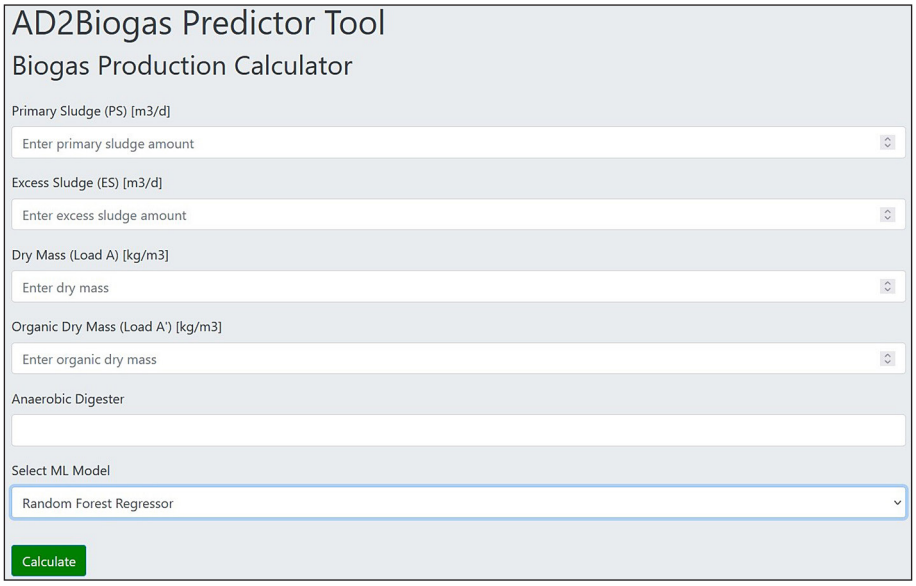

Figure 12. ML-based software tool for predicting biogas production - AD2Biogas predictor tool

and prediction. Additionally, our models, while robust, may not be as finely tuned as the ANN models optimized through GA and ACO. This suggests that while our approach is practical and easier to implement, particularly in environments where data may be limited or of varying quality, there is considerable potential for improvement by adopting more advanced techniques.

Furthermore, the differences in RMSE between our best-performing models and those in other studies underscore the trade-offs involved in model selection. While ANNs and other advanced methods like ANFIS (Adaptive Neuro-Fuzzy Inference System) can potentially offer lower RMSE, they also come with higher computational costs, increased complexity, and a greater risk of overfitting, particularly when data is sparse or noisy. In contrast, methods like Random Forest Regressor and Extra Trees Regressor, while potentially less precise in capturing complex interactions, offer robustness, easier implementation, and better handling of small and irregular datasets. This makes them particularly suitable for practical applications where data quality and availability are significant concerns.

The study also introduced the AD2Biogas Predictor Tool, a Python-based web software for predicting biogas production (Figure 12). This tool is especially valuable for the analyzed treatment plant, as it enables the selection of the anaerobic digester that produces the most biogas under similar conditions. This tool's applicability could be further enhanced by integrating advanced optimization techniques, such as those used by Beltramo et al. [41], to refine predictions and improve decision-making. Moreover, the tool's development, which adheres to strict safety standards [47], positions it as a reliable resource for plant operators seeking to optimize biogas production and manage resources more effectively. In the current version, this tool has been successfully implemented at the studied wastewater treatment plant, achieving an error margin of approximately 10%, which has been considered sufficiently accurate. Its deployment has provided operators with a valuable and efficient tool for swiftly estimating essential parameters, enhancing operational decision-making. The broader implications of this study lie in its demonstration of the importance of selecting the appropriate machine learning model and validation strategy for specific biogas production environments. The study's comparison of different validation approaches and models provides valuable insights into the trade-offs between accuracy, robustness, and computational efficiency. Future work should explore the integration of more advanced AI techniques, such as deep learning models or ensemble methods that combine multiple machine learning approaches, to further enhance prediction accuracy and generalizability. Additionally, expanding the dataset to include more diverse and continuous samples would likely improve model performance and offer more reliable insights for optimizing biogas production.

Given the performance results and the characteristics of the three variants, the Third Variant

with the Random Forest Regressor emerged as the most suitable choice from the 22 models tested for creating a prediction tool for biogas production in wastewater treatment plants. This variant's ability to leverage all available data ensures the development of a robust and generalizable model, making it a strong candidate for practical implementation.

CONCLUSIONS

The study focused on predicting biogas production using data from a municipal wastewater treatment plant in Rzeszow, where three validation strategies were employed: LOO cross-validation for individual chambers, LOSO cross-validation, and a combined dataset approach. Each approach has its strengths and limitations, with the combined data variant offering robust generalization but possibly averaging out unique chamber characteristics. The novelty of this research lies in its application to a municipal plant, where the variability in sludge composition and biogas yield presents distinct challenges compared to industrial or agricultural settings. This focus allows for insights that are directly applicable to similar facilities, filling a gap in existing literature. Our study's RMSE results were higher than those achieved in other research using more advanced models like ANNs or ANFIS, which can better capture complex, nonlinear relationships but come with greater computational costs and a higher risk of overfitting, especially with limited or noisy data. The AD2Biogas Predictor Tool, developed from this study, was successfully implemented at the Rzeszow plant, achieving an error margin of about 10%, which has been deemed sufficient for operational use. This tool has proven to be a valuable resource for operators, enabling efficient estimation of biogas production and better management of resources. By focusing on a real-world municipal setting, this study offers practical contributions to the field and demonstrates the importance of selecting appropriate machine learning models and validation strategies tailored to specific environments. Future work should explore advanced AI techniques and larger datasets to further enhance prediction accuracy and generalizability. Given the results, the Third Variant using the Random Forest Regressor emerged as the most suitable for creating a biogas prediction tool, providing a robust and generalizable model ideal for practical implementation in wastewater treatment plants.

REFERENCES

1. Wicki L, Naglis-Liepa K, Filipiak T, Parzonko A, Wicka A. Is the production of agricultural biogas environmentally friendly? Does the structure of consumption of first- and second-generation raw materials in Latvia and Poland matter? Energies. 2022; 15(15): 5623.
2. Kougias PG, Angelidaki I. Biogas and its opportunities—a review. Frontiers of Environmental Science & Engineering. 2018; 12:1–12.
3. Dabrowska S, Masłon A. The use of biogas from the anaerobic digestion of sewage sludge to improve the energy balance of wastewater treatment plants. Czasopismo Inżynierii Lądowej.
4. Bharathiraja B, Sudharsana T, Jayamuthunagai J, Praveenkumar R, Chozhavendhan S, Iyyappan J. Biogas production–a review on composition, fuel properties, feedstock and principles of anaerobic digestion. Renewable and Sustainable Energy Reviews. 2018; 90(April): 570–582.
5. Masłon A, Czarnota J, Szczyrba P, Szaja A, Szulżyk-Cieplak J, Łagód G. Assessment of energy self-sufficiency of wastewater treatment plants—a case study from Poland. Energies. 2024; 17(5): 1164.
6. Masłon A, Czarnota J, Szaja A, Szulżyk-Cieplak J, Łagód G. The enhancement of energy efficiency in a wastewater treatment plant through sustainable biogas use: Case study from Poland. Energies. 2020; 13(22): 6056.
7. Myszograj S, Sadecka Z, Bochenski D, Suchowska-Kisielewicz M. Green energy from biogas in a Polish-German sewage treatment plant. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects. 2013; 35(13): 1249–1255.
8. Szczyrba P, Masłon A, Czarnota J, Olszewski K. Analiza gospodarki osadowej i biogazowo-energetycznej w oczyszczalni ścieków w Opolu. Inżynieria Ekologiczna. 2020; 21(2): 26–34.
9. Masłon A. An analysis of sewage sludge and biogas production at the Zamość WWTP. In: Blikharskyy Z, Koszelnik P, Mesaros P, editors. Proceedings of CEE 2019. Cham: Springer International Publishing; 2020; 291–298.
10. Piechota G, Iglinski B. Biomethane in Poland—current status, potential, perspective and development. Energies. 2021; 14(6).
11. Owczuk M, Wardzińska D, Zamojska-Jaroszewicz A, Matuszewska A. Wykorzystanie odpadów biodegradowalnych do produkcji biogazu jako alternatywnego źródła energii odnawialnej. Studia Ecologiae

et Bioethicae. 2013; 11(3): 133–144.

12. Ismail ZZ, Jasim HS. Biogas recovery from refinery oily sludge by co-digestion followed by sustainable approach for recycling the residual digestate in concrete mixes. Advances in Science and Technology Research Journal. 2022; 16(5): 178–191.
13. Wójcik M, Łukasz B, Stachowicz F. Unconventional materials from sewage sludge with a potential application in road construction. Advances in Science and Technology Research Journal. 2018; 12(4): 65–75.
14. Kosiński P, Kask B, Franus M, Piłat-Rożek M, Szulżyk-Cieplak J, Łagód G. The possibility of using sewage sludge pellets as thermal insulation. Advances in Science and Technology Research Journal. 2023; 17(2): 161–172.
15. Koryś KA, Latawiec AE, Grotkiewicz K, Kubon M. The review of biomass potential for agricultural biogas production in Poland. Sustainability. 2019; 11(22).
16. Marks S, Dach J, Fernandez Morales FJ, Mazurkiewicz J, Pochwatka P, Gierz Ł. New trends in substrates and biogas systems in Poland. Journal of Ecological Engineering. 2020; 21(4).
17. Bachmann N, la Cour Jansen J, Bochmann G, Montpart N. Sustainable biogas production in municipal wastewater treatment plants. IEA Bioenergy. 2015; 20.
18. Gu Y, Li Y, Li X, Luo P, Wang H, Wang X, Wu J, Li F. Energy self-sufficient wastewater treatment plants: Feasibilities and challenges. Energy Procedia. 2017; 105: 3741–3751.
19. Gu Y, Li Y, Li X, Luo P, Wang H, Robinson ZP, Wang X, Wu J, Li F. The feasibility and challenges of energy self-sufficient wastewater treatment plants. Applied Energy. 2017; 204: 1463–1475.
20. Sarpong G, Gude VG. Near future energy self-sufficient wastewater treatment schemes. International Journal of Environmental Research. 2020; 14(4): 479–488.
21. Neczaj E, Grosser A. Circular economy in wastewater treatment plant–challenges and barriers. Proceedings. 2018; 2: 614.
22. Piwowar A. Agricultural biogas—an important element in the circular and low-carbon development in Poland. Energies. 2020; 13(7): 1733.
23. Szymańska D, Lewandowska A. Biogas power plants in Poland—structure, capacity, and spatial distribution. Sustainability. 2015; 7(12): 16801–16819.
24. Ruszel M, Maslon A, Ogarek P. Analysis of biogas from sewage sludge digestion in terms of diversification in the natural gas production structure in Poland. Desalination Water Treat. 2021; 232: 298–307.
25. Durdević D, Blecich P, Jurić Z. Energy recovery from sewage sludge: The case study of Croatia. Energies. 2019; 12(10): 1927.
26. Chamber of Commerce "Polish Waterworks" (IGWP). Information materials. 2020. Available online: https://www.igwp.org.pl/ (accessed on 12 June 2020).
27. Szymańska D, Lewandowska A. Biogas power plants in Poland—structure, capacity, and spatial distribution. Sustainability. 2015; 7(12): 16801–16819.
28. Vindis P, Mursec B, Janzekovic M, Cus F. The impact of mesophilic and thermophilic anaerobic digestion on biogas production. Journal of Achievements in Materials and Manufacturing Engineering. 2009; 36(2): 192–198.
29. Gandiglio M, Lanzini A, Soto A, Leone P, Santarelli M. Enhancing the energy efficiency of wastewater treatment plants through co-digestion and fuel cell systems. Frontiers in Environmental Science. 2017; 5: 70.
30. Lafratta M, Thorpe RB, Ouki SK, Shana A, Germain E, Willcocks M, Lee J. Dynamic biogas production from anaerobic digestion of sewage sludge for on-demand electricity generation. Bioresource Technology. 2020; 310: 123415.
31. Lafratta M, Thorpe RB, Ouki SK, Shana A, Germain E, Willcocks M, Lee J. Demand-driven biogas production from anaerobic digestion of sewage sludge: Application in demonstration scale. Waste and Biomass Valorization. 2021; 12(12): 6767–6780.
32. Abdel daiem MM, Hatata A, Galal OH, Said N, Ahmed D. Prediction of biogas production from anaerobic co-digestion of waste activated sludge and wheat straw using two-dimensional mathematical models and an artificial neural network. Renewable Energy. 2021; 178: 226–240.
33. Kannah RY, Rohini KB, Gunasekaran M, Gokulakrishnan K, Kumar G, Banu JR. Prediction of effective substrate concentration and its impact on biogas production using artificial neural networks in hybrid upflow anaerobic sludge blanket reactor for treating landfill leachate. Fuel. 2022; 313: 122697.
34. Abu Qdais H, Bani Hani K, Shatnawi N. Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm. Resources, Conservation and Recycling. 2010; 54(6): 359–363.
35. Galdino de Oliveira LR, dos Santos Filho DA, Marques Fraga TJ, Thomé Jucá JF, da Motta Sobrinho MA. Kinetics assessment and modeling of biogas production by anaerobic digestion of food wastes and acclimated sewage sludge. Journal of Material Cycles and Waste Management. 2021; 23(4): 1646–1656.
36. Prask H, Fugol M, Dyjakon A, Głab L, Sowiński J, Whitaker A. The impact of sewage sludge-sweet sorghum blends on the biogas production for energy purposes. Energies. 2023; 16(5).
37. Babel S, Sae-Tang J, Pecharaply A. Anaerobic co-digestion of sewage and brewery sludge for biogas production and land application. International Journal of Environmental Science & Technology. 2009; 6: 131–140.
38. Abu Qdais H, Bani Hani K, Shatnawi N. Modeling

and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm. Resources, Conservation and Recycling. 2010; 54(6): 359–363.

39. Wang L, Long F, Liao W, Liu H. Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. Bioresource Technology. 2020; 298: 122495.
40. Tufaner F, Demirci Y. Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models. Clean Technologies and Environmental Policy. 2020; 22: 713–724.
41. Beltramo T, Klocke M, Hitzmann B. Prediction of the biogas production using GA and ACO input features selection method for ANN model. Information Processing in Agriculture. 2019; 6(3): 349–356.
42. Cruz IA, Chuenchart W, Long F, Surendra KC, Andrade LRS, Bilal M, Liu H, Figueiredo RT, Khanal SK, Ferreira LFR. Application of machine learning in anaerobic digestion: Perspectives and challenges. Bioresource Technology. 2022; 345: 126433.
43. Breiman L. Random forests. Machine Learning. 2001; 45: 5–32.
44. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006; 63: 3–42.
45. Breiman L. Bagging predictors. Machine Learning. 1996; 24: 123–140.
46. AD2Biogas Online Tool. https://ad2b.tools.prz.edu.pl. Accessed: 2024-06-16.
47. Gugała Ł, Łaba K., Dul M. Protecting web applications from authentication attacks. Advances in Web Development Journal. 2023; 1(1).