

Machine-learning-based predictive control of anaerobic digestion considering meteorological conditions

Patryk Organiściak^{1*} , Arkadiusz Stęchły¹ , Paweł Kuraś² ,
Adam Maślona³ , Anna Turek⁴ 

¹ Department of Complex Systems, Faculty of Electrical and Computer Engineering, Rzeszow University of Technology, ul. M. C. Skłodowskiej 8, 35-036 Rzeszów, Poland

² Institute of Telecommunications and Cybersecurity, AGH University of Krakow, ul. Czarnowiejska 74, 30-059 Kraków, Poland

³ Department of Environmental Engineering and Chemistry, Faculty of Civil and Environmental Engineering and Architecture, Rzeszow University of Technology, ul. Powstańców Warszawy 6, 35-959 Rzeszów, Poland

⁴ Faculty of Mathematics and Applied Physics, Rzeszow University of Technology, Al. Powstańców Warszawy 8, 35-029 Rzeszów, Poland

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

ABSTRACT

Efficient control of the anaerobic digestion process in municipal wastewater treatment plants poses an engineering challenge due to the high inertia of the system and the impact of uncontrolled external disturbances. This paper presented a novel approach to optimising feedstock composition (the ratio of primary to excess sludge) to maximise biogas production. Unlike classical control methods, the proposed solution accounts for the impact of exogenous factors ambient temperature and precipitation which determine the physicochemical parameters of raw sludge. To build an effective model on a limited dataset, a spatial disaggregation technique was applied, merging data from four parallel digestion chambers. A hybrid decision-making system was developed, integrating a Random Forest model with a genetic algorithm (GA). The resulting predictive model achieved accuracy ($R^2 = 0.77$, MAPE = 7.33) under LODO validation, indicating that incorporating time delays in weather data allows for precise prediction of efficiency drops. The application of the evolutionary algorithm in a one-day numerical case study enabled the determination of optimal operational setpoints, indicating the potential for increased energy yield under difficult environmental conditions. Rather than a real-time autonomous control system, the presented solution constitutes a predictive optimisation framework and a functional module for a decision support system (DSS), enabling a transition from reactive to prescriptive strategies.

Keywords: anaerobic digestion, genetic algorithms, predictive control, machine learning, process optimisation, biogas.

INTRODUCTION

Efficient water and wastewater management can be of great importance to the environment, and moreover, the wastewater treatment process itself constitutes a key step towards improving the quality of the aquatic environment [1]. The composition and quantity of wastewater also have a significant impact on other environmental factors,

including biogas production aimed at the energy self-sufficiency of the treatment plant [2,3].

In contrast to the authors' previous studies [2,3], which focused on yield prediction based on feedstock characteristics, this paper extended the methodology to include meteorological factors. The methodological novelty lies in the transition from predictive to prescriptive analytics through the integration of a machine learning model

(Random Forest) with an optimisation algorithm (Genetic Algorithm). This has enabled not only production forecasting but also the active determination of optimal control trajectories in response to environmental disturbances, which had not been the subject of prior analyses.

The energy context and the specific nature of the process

Research conducted in national treatment systems identifies methane fermentation as the foundation for achieving energy self-sufficiency. Under Polish conditions, efficient cogeneration allows for covering from 40% up to even 70–80% of a facility's total electricity demand, as confirmed by energy intensity analyses for the south-eastern region [4]. The diversification of power sources through the production of on-site biogas is not merely a matter of cost optimisation, but a key element of the energy security of critical infrastructure [5]. Becoming independent from external energy supplies increases the operational stability of the plant in emergency situations. The evolution of the technological layout over the years has altered the conditions of sludge management described in previous works [6].

Despite the modernisations carried out at the examined facility, the non-linear nature of the fermentation process, which is sensitive to fluctuations in feedstock parameters, remains a fundamental challenge. To address this issue, recent literature has successfully applied advanced computational models, such as spiking neural networks (SNNs), to capture the complex multi-scale temporal dynamics and accurately predict long-term biogas yield [7,8]. This highlights the need for adaptive operator-support systems [9] that will ensure the stability of methane production regardless of the momentary characteristics of the influent.

Modern trends in environmental engineering strongly emphasise the digitalisation of technological processes. As Drewnowski *et al.* [9] pointed out, the integration of mathematical modelling with control systems (ICA) is crucial for improving the energy efficiency and reducing the carbon footprint of wastewater treatment plants. Nevertheless, the implementation of advanced solutions faces significant barriers. Andrade Cruz *et al.* [10] emphasised that the non-linearity of biological processes and the low interpretability

of “black-box” models hinder the transition from mere prediction to full prescriptive analytics.

Current research trends increasingly focus on omics technologies and advanced AI for comprehensive bioremediation and resource recovery [11]. However, due to the high level of complexity and the cost of data generation, these solutions remain out of reach for medium-sized facilities. This gap is exacerbated by the lack of proactive systems that – instead of passively collecting data – would integrate exogenous variables, such as weather forecasts, to generate ready-to-use control trajectories in real time [12,13]. Integrating external meteorological parameters into regression models effectively improves prediction accuracy in environmental engineering [14,15].

Aim of the study

The main objective of this study was the development and numerical verification of a predictive optimisation framework (serving as a decision support tool) for the anaerobic digestion process, utilising machine learning algorithms integrated with a genetic algorithm (GA). This objective was achieved by: (1) developing a predictive model that accounts for feedstock parameters and meteorological factors, (2) applying a data augmentation technique (Spatial Disaggregation) to enhance the model's generalisation capability, and (3) optimising the daily sludge dosages to maximize renewable energy production.

The proposed solution enables the utilisation of untapped energy reserves of the wastewater treatment plant without requiring capital expenditures for new infrastructure. From a scientific perspective, this work contributes to the methodology of processing small industrial datasets by demonstrating the effectiveness of combining data from parallel anaerobic Digesters into a single global model.

MATERIALS AND METHODS

The technological layout of the municipal wastewater treatment plant in Rzeszów underwent significant changes because of a series of modernisation investments implemented in stages during the years 1992–1998, 2001, 2003–2004, and 2013–2014. The introduced infrastructural modifications necessitate on-going research

to verify the current operational parameters of the facility. The plant is situated in the northern part of the agglomeration, at ul. Ciepłownicza 2, along the right bank of the Wisłok River [16].

The research was conducted based on operational data from the period between January 1st and August 31st, 2024 (244 operational days). The dataset was subjected to a detailed continuity analysis.

It was detected that 11.1% of the days (27 cases) lacked complete readings of production parameters. An analysis of the missing data structure revealed the occurrence of four separate incidents of data loss with lengths of 1, 4, 5, and 17 days, respectively. Such data collection gaps may result from various reasons, ranging from human resource shortages to common environmental telemetry disturbances in SCADA and IoT systems [17,18]. Due to the diverse characteristics of the interruptions, a hybrid data imputation strategy was adopted, justified by the specific nature of the fermentation process (large time constant and biological inertia). The missing data statistics are presented in Table 1.

The total average biogas production for the installation (sum of 4 chambers) during the studied period was approximately 6600 m³/d; however, for modelling purposes, parameters for individual anaerobic digesters were analysed (Table 2).

Meteorological data retrieved from the Open-Meteo [19] service were added to the process data. Ambient temperature, total precipitation, and relative humidity were taken into account. Raw hourly data were aggregated to daily values (average for temperature and humidity, sum for precipitation) in order to match the temporal resolution of the data from the treatment plant.

Short-term imputation (1-day gap)

Linear interpolation was applied. The literature indicates that this method is effective and sufficient for small data gaps where the process dynamics are preserved. A literature review conducted by Chaoui *et al.* [20] confirms that linear methods perform well in the case of minor missing data [21], and their high effectiveness is observed for a maximum of two consecutive missing values [22].

Medium- and long-term imputation (gaps of 4, 5, and 17 days)

Linear interpolation was avoided for longer sequences of missing data. Studies indicate that simple imputation methods fail to capture non-linear relationships and are generally unsuitable for long segments of missing data [23]; they can also lead to errors because they do not account for variance and the relationships between variables [24]. Instead, for the longest, 17-day gap, a method of averaging data from reference periods was applied (the average from the 7-day period preceding and following the incident). This allowed the statistical characteristics of the process to be preserved without introducing artificial linear trends.

The decision to employ imputation, rather than delete records, was driven by the necessity to maintain time series continuity for the learning algorithms, as well as by the stability of the biological system, which minimizes the risk of abrupt fluctuations during the analysed periods.

$$y_{imp}(t) = \begin{cases} y(t_{start}) + (t - t_{start}) \frac{y(t_{end}) - y(t_{start})}{t_{end} - t_{start}} \\ \frac{1}{2} \left(\frac{1}{N_{pre}} \sum_{i=1}^{N_{pre}} y(t_{start} - i) + \frac{1}{N_{post}} \sum_{j=1}^{N_{post}} y(t_{end} + j) \right) \end{cases} \quad (1)$$

where: Δt is the length of the measurement gap, N_{pre} is the averaging window (days).

Table 1. Statistics on missing data and the imputation methods used

Gap type	Number of occurrences	Total number of days	Method
Short-term (1 day)	1	1	Linear interpolation
Medium-term (4–5 days)	2	9	Surrounding mean (7 days)
Long-term (17 days)	1	17	Surrounding mean (7 days)
Total	4	27 (11.1%)	-

Data augmentation through spatial disaggregation

A challenge in modelling anaerobic digestion processes using machine learning (ML) methods is the limited availability of high-quality datasets. As indicated by Cruz *et al.* [10], an insufficient number of training samples is one of the main barriers to building models with high generalisation capabilities, often leading to the phenomenon of overfitting. During the analysed research period, 244 daily vectors were collected for the entire facility, constituting a dataset that is too small for complex regression algorithms such as Random Forest.

To address this issue, a data augmentation technique based on the physical architecture of the facility, known as spatial disaggregation, was applied. This approach relies on the concept of soft sensors and global modelling for parallel units, which significantly enhances the generalisation capabilities [25]. Detailed statistical characteristics of the process and meteorological variables for the dataset prepared in this manner are presented in Table 2.

The treatment plant operates four independent digestion chambers (A, B, C, D) with identical construction and similar process parameters. This allowed for the transformation of the data structure from a time-series format (aggregate for the entire plant) to a panel data format. This process involved splitting each measurement day into four parallel observations corresponding to individual reactors:

$$\mathcal{D}_{augmented} = \bigcup_{k \in A, B, C, D} \{(\mathbf{x}_{t,k}, y_{t,k}) \mid t \in 1, \dots, T\} \quad (2)$$

where: $\mathbf{x}_{t,k}$ is the vector of input parameters (load, hydraulics) for chamber k , and $y_{t,k}$ its

corresponding biogas production, and T is the total number of measurement days in the dataset.

It should be emphasised that the applied augmentation strategy does not artificially create independent observations. Rather, the original time-series dataset is restructured into a panel representation consisting of correlated measurements from parallel digesters operating under comparable technological conditions. Such an approach enables the model to learn shared process dynamics while increasing robustness against local reactor-specific disturbances.

Such an approach allows the model to learn the universal characteristics of the process, reducing the dependence of predictions on the aggregated operation of the entire plant, which significantly increases the model's resistance to disturbances [25]. This approach, known in the process engineering literature as Global Modelling for Parallel Units, is based on the assumption that the biochemical processes occurring in each of the chambers are subject to the same fundamental laws of fermentation kinetics. A model trained on such a set becomes more resistant to local disturbances (noise) of a single tank and learns to generalise the relationship between the load and anaerobic digestion, which improves transferability across anaerobic digesters with similar operational characteristics.

Control optimisation algorithm

To determine the optimal feedstock proportions, a hybrid approach was applied, combining a surrogate model based on the Random Forest algorithm with a GA as a global optimiser.

Table 2. Descriptive statistics for process and weather variables following augmentation (N=976)

Variable	Unit	Mean	Std. Dev.	Min.	Max.
Biogas production (Q_{bio})	m ³ /d	1652.31	324.84	962.00	3005.00
Primary sludge flow (Q_{PS})	m ³ /d	80.23	10.21	35.10	112.20
Excess sludge flow (Q_{ES})	m ³ /d	45.32	7.11	25.50	83.20
PS/ES volumetric ratio ($R_{PS/ES}$)	-	1.82	0.41	0.74	3.6
Total solids loading (L_{TS})	kg _{TS} /m ³	2.4	0.52	0.98	4.11
Volatile solids loading (L_{VS})	kg _{VS} /m ³	1.76	0.36	0.77	2.75
Ambient temperature (T_{amb})	°C	12.88	8.45	-14.70	26.90
Total precipitation (P_{sum})	mm	2.08	3.79	0.00	31.80
Air humidity (H_{avg})	%	74.31	10.37	47.00	95.00

Model quality assessment and validation

To verify the model's ability to generalise to real operational data, the specific nature of the augmented dataset was taken into account, in which each day is represented by four spatially correlated samples (chambers). The use of standard cross-validation carries the risk of overestimating results due to data leakage between parallel reactors. Consequently, a two-stage validation procedure was implemented:

Leave-one-out (LOO) method

A preliminary procedure in which the model is trained on the entire dataset, excluding a single test sample. This allows the model's internal consistency to be assessed, whilst permitting the use of weather data from neighbouring cells within the same 24-hour period.

Leave-one-day-out (LODO) method

A rigorous procedure used to evaluate actual predictive capability. It involves removing all samples belonging to the tested day from the training set. The training set $D_{train}^{(t)}$ in the group validation (LODO) iteration is defined as:

$$D_{train}^{(t)} = D \setminus \{(x_{t,k}, y_{t,k}) \mid k \in \{A, B, C, D\}\} \quad (3)$$

where: t denotes the day index and k is the chamber identifier.

This approach completely eliminates information leakage resulting from spatial correlation and identical meteorological conditions for a given day, simulating real prognostic conditions (lack of knowledge regarding the future for the entire facility).

Model evaluation metrics

The prediction quality for both methods was evaluated using standard performance indicators: root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2):

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{N_R} \sum_{i \in R} (Q_{bio,i} - \hat{Q}_{bio,i})^2} \\ \text{MAPE} &= \frac{1}{N_R} \sum_{i \in R} \left| \frac{Q_{bio,i} - \hat{Q}_{bio,i}}{Q_{bio,i}} \right| \cdot 100\% \end{aligned} \quad (4)$$

where: $Q_{bio,i}$ – actual biogas production for the i -th sample (measurement), $\hat{Q}_{bio,i}$ – predicted biogas production by the model, $\bar{Q}_{bio}$ – arithmetic mean of actual biogas production in the validation set, N_R – size of the validation set (number of actual samples).

It is important to emphasise that while the imputed records were retained in the overall training dataset to preserve time-series continuity for lagged meteorological variables, they were explicitly excluded from the test sets. During both LOO and LODO validations, the model was tested exclusively against the original, non-imputed operational data. Consequently, all reported performance indicators reflect the model's true predictive accuracy under real-world variance, without artificial smoothing from the imputation process.

RESULTS AND DISCUSSION

Correlation analysis and the performance of the predictive model

A correlation analysis was carried out (Table 3). Both Pearson's linear correlation coefficient (r), as

Table 3. Correlation between process variables and selected weather indicators and biogas production

Input variable	Unit	Pearson (r)	Spearman (ρ)
Primary sludge flow (Q_{PS})	m ³ /d	-0,181	-0,134
Excess sludge flow (Q_{ES})	m ³ /d	0.049	0.038
PS/ES volumetric ratio ($R_{PS/ES}$)	-	-0,154	-0,12
Total solids loading (L_{TS})	kg _{TS} /m ³	-0,183	-0,141
Volatile solids loading (L_{VS})	kg _{VS} /m ³	-0,12	-0,06
Digester unit ID ($ID_{digester\ unit}$)	-	0.484	0.524
Digester effective volume (V_{eff})	m ³	0.558	0.595
Weather conditions (Most significant delay)			
Ambient temperature "Lag 1" (T_{amb-1})	°C	-0.557	-0.483
Total precipitation "Lag 2" (P_{sum-2})	mm	0.022	0.084
Air humidity "Lag 5" (H_{avg-5})	%	0.395	0.404

Table 4. Analysis of the long-term impact of weather conditions (weekly aggregate)

Period (Lag)	Pearson (r)	Spearman (ρ)
Ambient temperature (T_{amb})		
Avg. temp. (Week 1) [days 1–7]	-0.559	-0.497
Avg. temp. (Week 2) [days 8–14]	-0.511	-0.482
Avg. temp. (Week 3) [days 15–21]	-0.496	-0.466
Avg. temp. (Week 4) [days 22–28]	-0.497	-0.481
Total precipitation (P_{sum})		
Total precip. (Week 1) [days 1–7]	0.029	0.059
Total precip. (Week 2) [days 8–14]	0.025	0.039
Total precip. (Week 3) [days 15–21]	-0.036	-0.009
Total precip. (Week 4) [days 22–28]	-0.037	-0.010
Air humidity (H_{avg})		
Avg. humidity (Week 1) [days 1–7]	0.506	0.522
Avg. humidity (Week 2) [days 8–14]	0.451	0.461
Avg. humidity (Week 3) [days 15–21]	0.401	0.421
Avg. humidity (Week 4) [days 22–28]	0.424	0.445

well as Spearman's rank (ρ), which is more robust against outliers and process non-linearity [22].

The analysis revealed a strong, non-linear relationship between biogas production and selected weather indicators, in particular ambient temperature with a one-day lag (T_{amb-1} , $\rho = -0.483$). A statistically significant correlation with air humidity, with a 5-day lag, is also evident; this may be due to the long-term influence of atmospheric conditions on the physicochemical properties of the incoming wastewater. The use of Spearman's correlation coefficient confirmed the monotonic nature of these relationships and their higher significance compared to Pearson's linear correlation, which justifies the inclusion of lagged meteorological variables in the input vector of predictive models. Incorporating exogenous meteorological data into machine learning algorithms is proven to enhance the forecasting of complex environmental phenomena [14].

In order to investigate the time horizon of the impact of external factors, a correlation analysis was conducted for the data aggregated in weekly intervals (Table 4). The results suggest a long-term inertia of the system in relation to ambient temperature and humidity, where strong Spearman rank correlations (respectively $\rho \approx -0.50$ and $\rho \approx 0.45$) remain at a stable level throughout the full 28-day period. Atmospheric precipitation exhibits the opposite characteristic; its influence disappears with weekly aggregation, suggesting its purely impulsive, short-term nature

in disrupting process hydraulics. The analysis above justifies the use of a wide time window in machine learning models to capture the seasonal trends affecting fermentation kinetics.

Predictive model selection

The choice of the Random Forest (RF) algorithm as the simulation engine in this study is not incidental. In previous comparative research conducted by the authors at the same technological facility [2], the effectiveness of six machine learning algorithms (including KNN, Linear Regression, and Extra Trees) was tested. It was demonstrated that ensemble methods, and specifically Random Forest, are best suited to handle the non-linear nature of the fermentation process and the measurement noise characteristic of real-world industrial data. While other advanced artificial intelligence architectures, such as Spiking Neural Networks (SNNs), have also proven highly effective in modelling the non-linear temporal dynamics of anaerobic digestion [7,8], ensemble methods like Random Forest were selected here to serve as a robust, computationally efficient objective function estimator for the genetic algorithm. The impact of explanatory variables on the model's predictive performance is shown in Figure 1.

Feature importance analysis indicates strong predictive relevance of weather factors. Ambient temperature (T_{amb-1}) dominates the ranking (with an importance of 39.2%) which is consistent with the

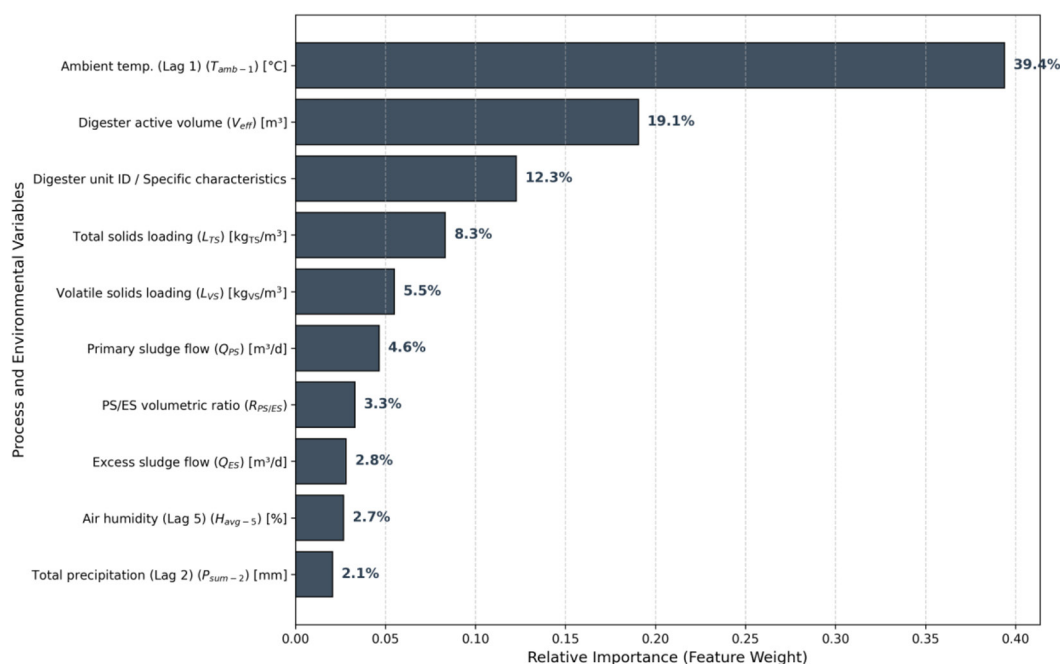


Figure 1. Feature importance ranking in the Random Forest final model. The greatest impact on biogas production is exhibited by the average ambient temperature (with a 1-day lag) and the specificity of the given fermentation chamber

thermophilic nature of the process and its sensitivity to cooling [24]. The significant contribution of the chamber identification variable ($ID_{digester\ unit}$) indicates structural or operational differences between the reactors that the model must compensate for.

Therefore, in the present approach, the re-validation of simple linear models was omitted, utilising Random Forest as a suggested objective function estimator for the genetic algorithm. The main extension over previous work is the expansion of the model's input vector to include meteorological parameters and its application in a feedback loop

Table 5. Results of the LOO validation of the predictive model (Random Forest)

Quality metric	Value
Coefficient of determination (R^2)	0.81
Root mean square error (RMSE)	137.92 m³/d
Mean absolute percentage error (MAPE)	6.24%

with an evolutionary optimiser. Detailed values of the model goodness-of-fit indicators (R^2 , RMSE, MAPE) are summarised in Tables 5 and 6 (LOO, LODO). A graphical interpretation of the LOO

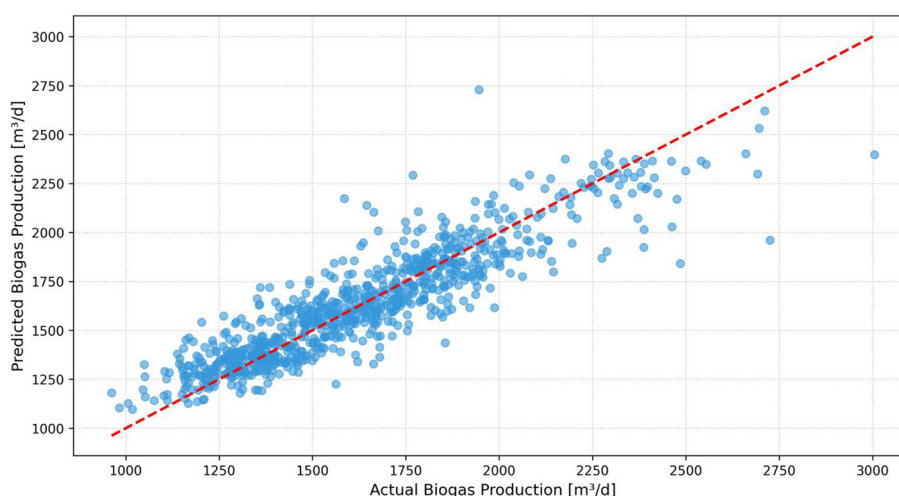


Figure 2. Model fit plot (LOO validation). The horizontal axis represents the actual biogas production, and the vertical axis represents the value predicted by the model. The red dashed line indicates a perfect fit

Table 6. Results of the LODO validation of the predictive model (Random Forest)

Quality Metric	Value
Coefficient of determination (R^2)	0.77
Root mean square error (RMSE)	154.98 m ³ /d
Mean absolute percentage error (MAPE)	7.33%

validation results is presented in Figure 2. The parity plot shows a strong concentration of points along the line of perfect fit ($y = \hat{y}$), which confirms the model's high generalisation capability.

The even distribution of residuals on both sides of the diagonal indicates an absence of systematic model bias the algorithm shows no tendency to consistently underestimate or overestimate production within the typical operating range. Minor deviations are visible only for extreme values, which is typical for real-world data affected by process noise. A comparison of the validation results reveals a significant characteristic of the process. The model tested using the LOO method achieved $R^2 = 0.81$ (Figure 2), while the more rigorous LODO (leave-one-day-out) validation indicated $R^2 = 0.77$ (Figure 3). This difference stems from strong spatial correlation in LOO validation, the model could leverage information regarding the operation of adjacent chambers on the same day. The LODO result, though lower, is a more reliable indicator in the context of predictive control, as it reflects a scenario where future parameters for the entire facility are unknown. Despite the decrease in accuracy, the LODO result

still confirms the high utility of the model in control systems (MAPE = 7.33%).

Optimisation scenario using a genetic algorithm

The primary objective of the control system is to maximise biogas production Q_{bio} by selecting optimal control values: the flow rates of primary sludge Q_{PS} and excess sludge Q_{ES} , while accounting for current and forecasted weather conditions. Due to the non-linear nature of the process and the lack of an analytical form for the process model, a trained Random Forest estimator (M_{RF}) was used as the fitness function within the genetic algorithm. The optimisation problem at time t is formulated as follows:

$$\max_{u_t} J = M_{RF}(u_t, d_t, h_t) \quad (5)$$

where: $u_t = [Q_{PS}, Q_{ES}]^T$ – vector of decision variables (controls), $d_t = [T_{amb}, P_{sum}, H_{avg}]^T$ – vector of measurable weather disturbances (independent of control), h_t – vector of historical variables (lags) resulting from process dynamics.

The search space is limited by the technological capabilities of the facility and process requirements, defined as inequality constraints:

$$\begin{aligned} Q_{PS}^{min} &\leq Q_{PS} \leq Q_{PS}^{max} \\ Q_{ES}^{min} &\leq Q_{ES} \leq Q_{ES}^{max} \\ R_{PS/ES}^{min} &\leq \frac{Q_{PS}}{Q_{ES}} \leq R_{PS/ES}^{max} \\ L_{TS}^{min} &\leq L_{TS}(u_t) \leq L_{TS}^{max} \end{aligned} \quad (6)$$

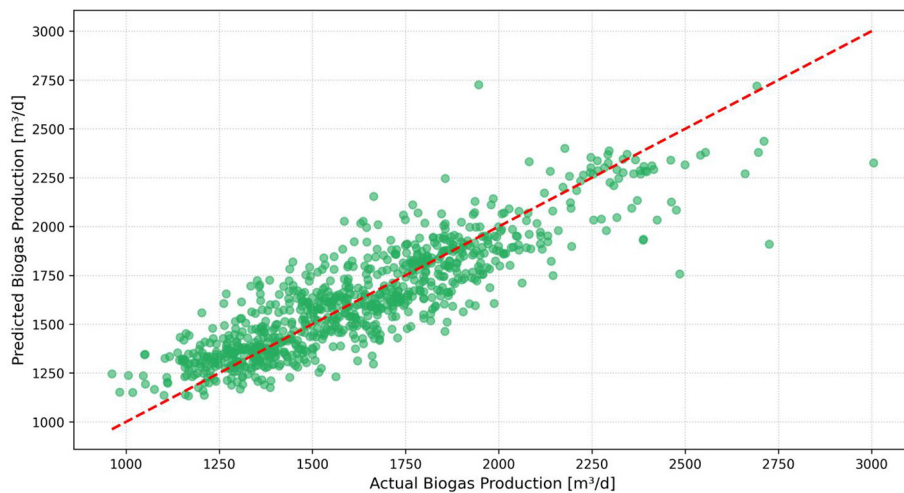


Figure 3. Model fit plot (LODO validation). The horizontal axis represents the actual biogas production, while the vertical axis represents the value predicted by the model. The red dashed line indicates the perfect fit

Table 7. Genetic algorithm parameters

Parameter	Value
Number of generation (N_{gen})	30
Population size (N_{pop})	15
Selection	Tournament selection
Crossover probability	0.8
Mutation probability	0.15

where: $R_{PS/ES}^{min}$ denote the permissible range of the sludge mixing ratio, and L_{TS} represents the resulting organic loading rate of the chamber.

To prevent model extrapolation, the search space bounds were strictly derived from the historical operational limits (Table 2).

The genetic algorithm operates on a population of N_{pop} individuals, where each chromosome represents a potential operating point of the plant ($\hat{Q}_{PS}, \hat{Q}_{ES}$). In each iteration of the algorithm:

- A new population of controls is generated (via tournament selection, crossover, and mutation).
- For each individual, an input feature vector is created, supplemented by weather parameters (d_i) and historical data (h_i) that remain constant for that specific step.
- The Random Forest model predicts the expected biogas production.
- The predicted value serves as the fitness score for the individual.

The procedure terminates upon reaching the maximum number of generations or when the improvement in the objective function between iterations falls below a specified tolerance threshold.

Method validation – a case study

To evaluate the effectiveness of the proposed predictive optimisation framework, a numerical simulation was conducted for representative operating conditions of the facility. A forecasting scenario was defined covering a time horizon of $T + 1$ (one day), with fixed environmental disturbance vectors: $T_{amb} = 15$ °C, $P_{sum} = 0$ mm, $H_{avg} = 60\%$.

The evolutionary parameters of the genetic algorithm were determined experimentally, aiming for a trade-off between convergence time and the exploration of the solution space (Table 7).

The optimisation process converged to a stable solution after 30 iterations. The operating point determined by the algorithm u^* , maximising the estimated biogas production, is:

$$u^* = \begin{bmatrix} Q_{PS} \\ Q_{ES} \end{bmatrix}^* \approx \begin{bmatrix} 85.43 \text{ m}^3/\text{d} \\ 40.78 \text{ m}^3/\text{d} \end{bmatrix} \quad (7)$$

Implementing the above control procedure, given the assumed sludge composition, results in the chambers being loaded to a level of $LTS \approx 3.77$ kg TS/m³ (total solids) and $L_{VS} \approx 2.06$ kg VS/m³ (volatile solids), which are considered within the safe operating range. For the determined setpoints, the Random Forest model predicts a production level of $Q_{bio}^{max} \approx 1848.51$ m³/d. The resulting sludge volume ratio ($R \approx 2.1$) remains in close alignment with the technological constraints of the fermentation process, suggesting the correct implementation of the penalty function in the genetic algorithm.

CONCLUSIONS

In this paper, a hybrid method for biogas production optimisation was proposed, integrating machine learning (Random Forest) with evolutionary algorithms. The conducted research allows for the formulation of the following conclusions.

The use of the Random Forest estimator enabled the effective approximation of the non-linear dynamics of the anaerobic digestion process, accounting for the lagged influence of meteorological factors. Feature importance analysis revealed that the one-day-lagged ambient temperature is the dominant predictor of biogas production. While technological variables (feed-stock parameters) are not the leading contributors according to the final model, they remain important. This justifies their selection as decision variables in the optimisation task, as they represent the key operational parameters that can be actively adjusted to compensate for external environmental disturbances.

Global optimisation – coupling a regression model with a genetic algorithm enables the determination of optimal control trajectories in a multidimensional space, which constitutes a significant advantage over classical control methods based on rigid expert rules.

The obtained numerical simulation results demonstrate the method's potential as a decision support framework aimed at maximising the

energy efficiency of the treatment plant whilst adhering to process constraints. However, it must be emphasised that the presented optimisation example is a numerical case study conducted for fixed environmental conditions and a short, one-day time horizon. Therefore, it illustrates predictive optimisation, rather than a fully implemented real-time control strategy. Further research involving real-world validation is necessary before deploying this framework as an autonomous, real-time predictive control system.

The limitations of the analysis include the lack of a full annual data set, including the absence of the autumn-winter transition period. This limits the ability to fully assess seasonality and to generalise the results to all weather conditions throughout the year. Further research will focus on validating the method under real-world conditions and extending the model to include adaptive estimation mechanisms that account for variations in the physicochemical composition of the sediment.

A comparative analysis of validation methods (LOO vs LODO) confirmed that data augmentation through spatial disaggregation is effective; however, it carries the risk of overestimating results if the temporal correlation between parallel reactors is not taken into account. The use of group validation (LODO) made it possible to eliminate this error and confirm the stability of the predictions at a level acceptable for industrial control systems.

REFERENCES

1. Czarnota J, Masłoń A. Evaluation of the effectiveness of a wastewater treatment plant with MBBR technology. *Rocznik Ochrona Środowiska* 2019;21:906–25.
2. Organiściak P, Masłoń AA, Kowal B, Kuraś P, Wadiak B, Sikorska-Czupryna S, Vanivska V. Machine learning-based prediction of biogas production from sludge characteristics in four anaerobic digesters: Development of the AD2Biogas prediction tool. *Advances in Science and Technology Research Journal* 2024;18:1–15. <https://doi.org/10.12913/22998624/192936>
3. Organiściak P, Turek A, Stęchły A, Rysz J, Vanivska V, Masłoń A, Wadiak B. Prediction of biogas yield from anaerobic digesters using meta-regression. *Zeszyty Naukowe SGSP* 2025;1:81–100. <https://doi.org/10.5604/01.3001.0055.0610>
4. Masłoń A. Analysis of energy consumption at the Rzeszów Wastewater Treatment Plant. *E3S Web of Conferences* 2017;22:00115. <https://doi.org/10.1051/e3sconf/20172200115>
5. Szczyrba P, Masłoń A, Czarnota J, Olszewski K. Analysis of sewage sludge and biogas-energy management at the Opole wastewater treatment plant. *Inżynieria Ekologiczna* 2020;21:26–34. <https://doi.org/10.12912/23920629/122657>
6. Masłoń A, Tendera K. Gospodarka osadami ściekowymi w oczyszczalni ścieków Rzeszów. *Forum Eksploatatora* 2017;88:38–43.
7. Capizzi G, Lo Sciuto G, Napoli C, Woźniak M, Susi G. A spiking neural network-based long-term prediction system for biogas production. *Neural Networks* 2020;129:271–9. <https://doi.org/10.1016/j.neunet.2020.06.001>
8. Lo Sciuto G, Susi G, Cammarata G, Capizzi G. A spiking neural network-based model for anaerobic digestion process. 2016 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), IEEE; 2016, p. 996–1003. <https://doi.org/10.1109/SPEEDAM.2016.7526003>
9. Drewnowski J, Szelać B, Sabba F, Piłat-Rożek M, Piotrowicz A, Łagód G. Innovations in wastewater treatment: harnessing mathematical modeling and computer simulations with cutting-edge technologies and advanced control systems. *Journal of Ecological Engineering* 2023;24:208–22. <https://doi.org/10.12911/22998993/173076>
10. Andrade Cruz I, Chuenchart W, Long F, Surendra KC, Renata Santos Andrade L, Bilal M, Liu H, Tavares Figueiredo R, Khanal SK, Fernando Romanholo Ferreira L. Application of machine learning in anaerobic digestion: Perspectives and challenges. *Bioresour Technol* 2022;345:126433. <https://doi.org/10.1016/j.biortech.2021.126433>
11. Renganathan P, Gaysina LA. Next-generation wastewater treatment: Omics and AI-driven microbial strategies for xenobiotic bioremediation and circular resource recovery. *Processes* 2025;13:3218. <https://doi.org/10.3390/pr13103218>
12. Nicolae A, Korodi A, Silea I. Weather-based prediction strategy inside the proactive historian with application in wastewater treatment plants. *Applied Sciences* 2020;10:3015. <https://doi.org/10.3390/app10093015>
13. Iratni A, Chang N-B. Advances in control technologies for wastewater treatment processes: status, challenges, and perspectives. *IEEE/CAA Journal of Automatica Sinica* 2019;6:337–63. <https://doi.org/10.1109/JAS.2019.1911372>
14. Susi G, Lo Sciuto G, Capizzi G, Famoso F. A novel regression method for pollutant prediction using Albedo, normalized difference vegetation index, and surface kinetic temperature metrics derived from ASTER satellite image. *Signal Image Video Process* 2026;20:149. <https://doi.org/10.1007/s11760-026-05263-z>
15. Kowal B, Organiściak P, Kuraś P, Masłoń A, Wadiak

- B. Machine Learning in forecasting the amount of sewage for the wastewater treatment plant of the city of Rzeszów Based on weather conditions. *Pomiary Automatyka Robotyka* 2025;29:81–90. https://doi.org/10.14313/PAR_255/81
16. Maj K, Masłoń A. Wieloparametryczna ocena efektywności oczyszczalni ścieków w Rzeszowie przed modernizacją. *Czasopismo Inżynierii Lądowej, Środowiska i Architektury* 2015;62:299–315.
17. Lo Sciuto G. Air pollution effects on the intensity of received signal in 3G/4G mobile terminal. *International Journal of Energy and Environmental Engineering* 2019;10:221–9. <https://doi.org/10.1007/s40095-019-0298-8>
18. Kuraś P, Organiściak P, Kowal B, Vanivska V, Smalara K, Demidowski K, Krawiec P. IOT communication security: Challenges and solutions. *Zeszyty Naukowe SGSP* 2024;2:69–90. <https://doi.org/10.5604/01.3001.0054.3829>
19. Zippenfenig P. Open-Meteo Weather API 2023.
20. Chaoui A, Rebija K, Chkaiti K, Laaouan M, Bourziza R, Sebari K, Elkhousi W. Applications of missing data imputation methods in wastewater treatment plants. *International Journal of Advanced Computer Science and Applications* 2023;14. <https://doi.org/10.14569/IJACSA.2023.0141049>
21. De Mulder C, Flameling T, Weijers S, Amerlinck Y, Nopens I. An open software package for data reconciliation and gap filling in preparation of water and resource recovery facility modeling. *Environmental Modelling & Software* 2018;107:186–98. <https://doi.org/10.1016/j.envsoft.2018.05.015>
22. Oliveira-Esquerre KP, Seborg DE, Bruns RE, Mori M. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill. *Chemical Engineering Journal* 2004;104:73–81. <https://doi.org/10.1016/j.cej.2004.05.011>
23. Kazadi Mbamba C, Keymer P, Alvi M, Topalian SON, Ud Din F, Batstone DJ. Enhancing data quality in wastewater processes: Missing data imputation with deep Variational Autoencoders and genetic algorithms. *Comput Chem Eng* 2025;199:109123. <https://doi.org/10.1016/j.compchemeng.2025.109123>
24. Han H, Sun M, Wu X, Li F. Double-cycle weighted imputation method for wastewater treatment process data with multiple missing patterns. *Sci China Technol Sci* 2022;65:2967–78. <https://doi.org/10.1007/s11431-022-2163-1>
25. Kadlec P, Gabrys B, Strandt S. Data-driven Soft Sensors in the process industry. *Comput Chem Eng* 2009;33:795–814. <https://doi.org/10.1016/j.compchemeng.2008.12.012>