



Modeling and Prediction of Biogas Production in the South Tehran Wastewater Treatment Plant Using Machine Learning

Somayeh Alizadeh Attar¹, Seyedeh Mahtab Hashemi¹, Mohammad Amin Fatemimehr¹,
Khashayar Zare², Ahmad Hosseinnjad^{1*}

1. Department of Environmental Engineering, University of Tehran, Tehran, Iran

2. Department of Agricultural Machinery Engineering, Faculty of Agricultural, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

Abstract

This study investigates the application of machine learning techniques to predict biogas and methane production from the anaerobic digestion process at the South Tehran Wastewater Treatment Plant, which serves approximately 450,000 residents. The main objective is to develop and evaluate predictive models that can optimize energy recovery and contribute to climate change mitigation. Data collected during 2023–2025, including influent and effluent flow rates and polymer consumption, were analyzed using correlation methods to identify key parameters affecting biogas output. Various machine learning models such as Random Forest, Support Vector Regression, Gradient Boosting, and ExtraTreesRegressor were developed and optimized through grid search and cross-validation. The ExtraTreesRegressor model achieved the highest accuracy with an R^2 of 0.8422 and a mean absolute percentage error of 11.8%, demonstrating robust predictive capability. By applying a methane content factor of 0.65 to predicted biogas volumes, methane production estimates were obtained. shapley Additive explanations analysis highlighted polymer consumption as the most influential factor, followed by influent flow rate. The findings confirm that machine learning models can reliably predict biogas production, supporting optimized operation of anaerobic digesters. This approach offers significant potential to improve energy recovery from wastewater treatment processes, reduce reliance on fossil fuels, and mitigate greenhouse gas emissions, contributing to sustainable energy management and climate change mitigation.

Keywords

Machine leaning, Wastewater treatment plant, Wastewater gas consumption, CH₄ consumption, Anaerobic digestion, Biogas prediction

***Corresponding authors name and email:**

Ahmad Hosseinnjad, Email: ah.hosseinnjad@ut.ac.ir



1. Introduction

The current irrational use of fossil fuels and the impact of greenhouse gases on the environment are driving research into renewable energy production from organic resources and waste. The global energy demand is high, and most of this energy is produced from fossil resources [1]. The global warming potentials (GWPs) of greenhouse production using different heating systems are inadequately compared in the literature, which is the research gap addressed herein. Additionally, potential emission reductions of greenhouse production with industrial waste heat are yet to be assessed via lifecycle assessment (LCA) [2]. The use of anaerobic processes is widely practiced in the treatment of waste, since anaerobic microorganisms degrade the organic matter, which generates biogas and biofertilizer as final products [3]. Naturally, replacing fossil fuels with bioenergy and renewables significantly decreases the GWP. Biogas can be used for electricity and heat generation, and therefore by industrial or rural properties, bringing economic and environmental gains. The burning of biogas is beneficial for reducing the global warming potential [1]. A wide range of waste types can be used as substrates for biogas production using AD technology. The most typical forms of waste used in the European energy industry are: animal manure and slurry, sewage sludge, municipal solid waste, and food waste [1].

Biomethane (upgraded biogas) can also substitute for natural gas in chemicals production. Recent evaluations indicate that biogas produced via anaerobic digestion (AD) provides significant advantages over other forms of bioenergy because AD is an energy-efficient and environmentally friendly. Recent studies report that anaerobic digestion (AD) is an efficient alternative technology that combines biofuel production with sustainable waste management, and various technological trends exist in the biogas industry that enhance the production and quality of biogas [1]. According to the environmental conditions, the performance of the digestion process may be delayed or unsuccessful due to the accumulation of fat and suspended solids in the digester, since a decrease in pH, particularly, causes a reduction in the methanogenic activity and biomass. For this reason, several studies have recommended the previous separation of constituents with a high fat content, despite their higher heat capacity [3]. Further investments in AD are expected to meet with increasing success due to the low cost of available feedstocks and the wide range of uses for biogas (i.e., for heating, electricity, and fuel). Biogas production is growing in the European energy market and offers an economical alternative for bioenergy production [1].

Methane, right after carbon dioxide, is considered the most harmful greenhouse gas to the climate. Almost 59% of the world's emissions of this gas come from anthropogenic sources, of which the largest share (40–53%) is agriculture, especially intensive production [4].

The reduction of methane emissions slows down negative climate changes and improves air quality. Most of the legal acts concerning climate policy pay mostly attention to limiting carbon dioxide emissions. However, more and more attention is paid to the reduction of methane emissions [4].



Greenhouses protect plants from atmospheric impacts and modify the microclimate for the crops, leading to improved quality and higher yields of production, while lengthening the market availability of domestic crops. Due to their high economic efficiency, greenhouses are an attractive business for agricultural entrepreneurs. As a downside, greenhouse production consumes high quantities of energy, requires multiple production inputs, and generates large amounts of waste, which can lead to environmental concerns [5]. Organic farming is presented as a solution to secure sustainable food production. However, the low yields of organic farming have in some cases increased the environmental impact of greenhouses due to higher energy demand [2].

Therefore, the use of biomethane produced in anaerobic digesters as a source of energy supply for greenhouses leads to both stopping the release of methane from the treatment plant into the air and the consequent reduction of the greenhouse effect, and makes greenhouses unnecessary from the use of fossil fuels. This is called the food-energy nexus, which leads to a sustainable food supply and reduces the impact of global warming.

In this study, the wastewater treatment plant in southern Tehran was studied. The treatment plant covers a population of about 450,000 people. Its processes are divided into two parts: liquid and sludge section. The sludge section receives its input from two primary and secondary sedimentation settling ponds, which itself consists of sludge thickening, anaerobic digestion, and dewatering sections. Depending on the input flow rate, wastewater quality and pretreatment processes, a significant amount of biogas is produced in the anaerobic digesters. The biogas produced is a mixture of several gases, about 0.65 [6] of which is methane. Machine learning-based algorithms help us easily predict the methane produced in the sludge section with a series of treatment plant data, such as the average daily flow rate of the input wastewater, and use it optimally to reduce energy consumption in the required sectors.

2. Material and Methods

In this study, the data were collected from the South Tehran Wastewater Treatment Plant, located in the northern part of Shahr-e Rey, between Tehran and Varamin, just south of Shahid Avini Highway. The facility is situated approximately one kilometer from residential areas of Shahr-e Rey. Using Pearson and Spearman correlation analyses, key operational parameters influencing biogas production namely inlet flow rate, outlet flow rate, and polymer consumption were identified. Data were gathered during the Iranian calendar years 1402 and 1403 (equivalent to 2023–2025), and after preprocessing and data cleaning, a total of 495 data points were retained for analysis.

To model biogas production, a grid search strategy was employed to tune hyperparameters, and a variety of machine learning models were developed, including Random Forest, Decision Tree, SVR, XGBoost, Gradient Boosting, MLP, Linear Regression, Polynomial Regression, Ridge, Lasso, ElasticNet, and ExtraTreesRegressor. These models were trained using supervised learning algorithms using 85% of all data as the training and 15% as test data, and their performance was optimized and systematic hyperparameter tuning (Figure 1). Ultimately,

the top three models were combined into a voting ensemble, and model interpretability was enhanced using the SHAP (SHapley Additive exPlanations) framework. Model performance was evaluated based on R^2 and MSE metrics. Among all tested models, ExtraTreesRegressor delivered the most accurate predictions and was selected as the best-performing model. The results demonstrate that machine learning models have strong potential for predicting biogas output based on operational inputs—an insight that can significantly support the optimization of treatment processes, efficient waste utilization, and sustainable, cost-effective exploitation of biogas technology.

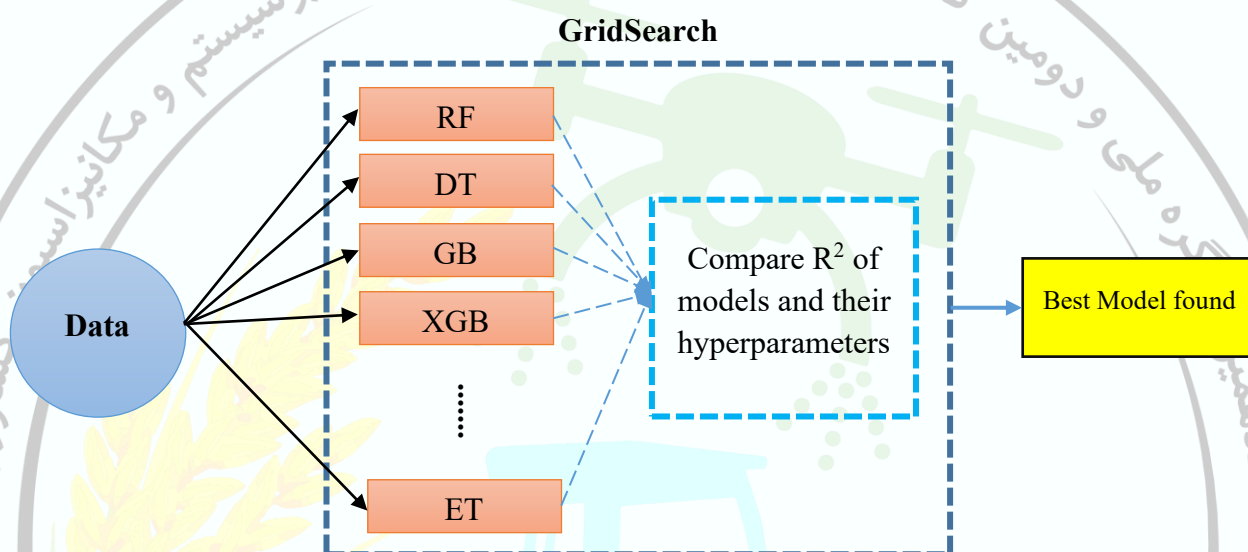


Figure 1. Finding biogas production model

To evaluate the agreement between predicted outputs and actual values, the coefficient of determination (R^2) was used. This metric indicates the proportion of variance in the actual data that is explained by the model. The R^2 value ranges between zero and one, with values closer to one indicating better model performance. This metric was calculated using the built-in `r2_score` function from the scikit-learn library, and its value is derived from Equation (eq.1), where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of actual values, and n is the number of samples.

To assess the prediction accuracy of the model, the Mean Squared Error (MSE) was employed. This metric measures the average squared difference between actual and predicted values, with lower values indicating higher model accuracy. The MSE was computed using the `mean_squared_error` function from the scikit-learn library, and its value is obtained from Equation (eq.2).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{eq. 1})$$



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

Additionally, the Mean Absolute Percentage Error (MAPE), a widely used metric for evaluating the accuracy of predictive models, was employed. MAPE quantifies the relative error of the model as a percentage by calculating the average absolute ratio of the prediction error to the actual value, as defined in Equation (eq.3):

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (\text{eq. 3})$$

Here, y_i represents the actual value, $\hat{y}_i$ denotes the predicted value, and N is the number of samples. Since MAPE expresses error as a percentage relative to the actual value, it simplifies interpretation for analysts and end-users, while also enabling straightforward comparison of model performance.

3. Results and Discussion

In this study, various machine learning models were developed, trained, and tested to predict biogas production based on real operational data collected from the South Tehran Wastewater Treatment Plant. The primary input variables included the daily influent flow, effluent flow, and polymer dosage, parameters that are routinely measured in large-scale treatment facilities. The influent flow averaged 222,663 cubic meters per day, while the effluent flow reached 214,784 cubic meters per day. The average polymer dosage applied in the treatment process was 35 kilograms per day.

After extensive testing and optimization through a grid search approach, the ExtraTreesRegressor emerged as the most effective model for this application. This model demonstrated strong predictive accuracy, achieving a coefficient of determination (R^2) of 0.8422, a mean squared error (MSE) of 296,655.1, and a mean absolute percentage error (MAPE) of 11.8%. These results highlight the model's capability in capturing the nonlinear relationships inherent in the anaerobic digestion process, while maintaining a reasonable balance between precision and generalizability. Given that approximately 65% of the biogas produced at this facility is methane, the predicted biogas values can be easily translated into methane yield estimates by applying a simple conversion factor, further increasing the model's practical utility. The results of the other models for the optimal parameters are presented in Table 1.

Table 1. best R^2 of each model with its best hyperparameters

Model	R^2 for the best hyperparameter
RandomForestRegressor	0.7884
DecisionTree	0.6800



SVR	0.0189
XGBoost	0.5824
GradientBoostingRegressor	0.6897
MLP	0.2030
LinearRegression	0.2056
PolynomialRegression	0.2382
Ridge	0.2056
Lasso	0.2056
ElasticNet	0.2055
ExtraTreesRegressor	0.8422

Beyond the statistical results, the interpretability of the model was further enhanced through SHAP (SHapley Additive exPlanations) analysis. This technique was used to understand the relative importance of each input feature and its contribution to the predicted output. As shown in Figure 2, the SHAP analysis reveals that polymer dosage is the most influential factor affecting biogas production, followed by influent flow rate. Effluent flow had minimal contribution in comparison. Additionally, the overall ranking and distribution of feature effects are visualized in Figure 3, further confirming the dominance of polymer usage in driving the model's predictions. These findings are consistent with prior research, such as studies by Sun et al. (2023) and Laycock et al. (2024), which also emphasized the critical role of chemical inputs and organic load in biogas generation processes [7], [8] (Figure 2).

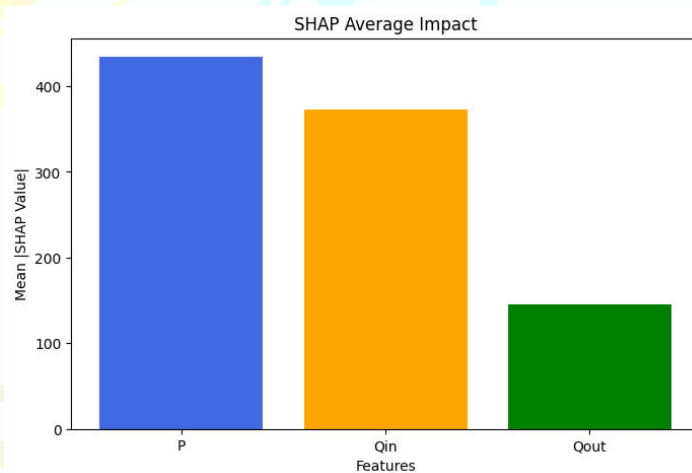


Figure 2. SHAP Analysis of Feature Impact on Biogas Prediction

The SHAP analysis revealed that polymer consumption had the strongest influence on biogas production predictions, followed by inlet flow rate, while outlet flow rate showed the least impact. These findings highlight the model's ability not only to predict biogas output with high accuracy, but also to clearly interpret the relative importance of operational parameters in the performance of anaerobic digesters (Figure 3).

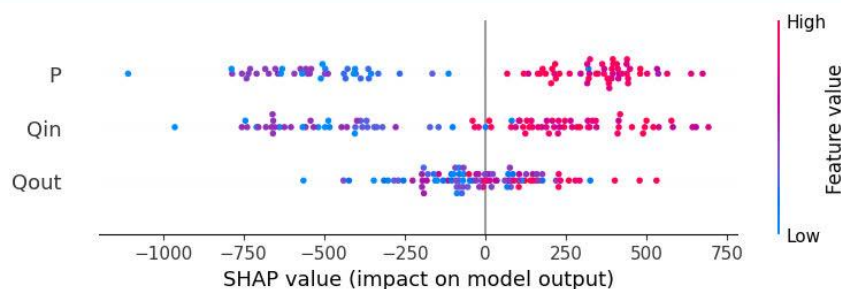


Figure 3. SHAP Plot of Feature Importance for Biogas Prediction

The SHAP scatter plot results revealed that among the input features, polymer consumption had the highest influence on biogas production predictions. This was followed by inlet flow rate, while outlet flow rate had the least effect. The spread and intensity of SHAP values indicate that the developed model is not only accurate but also highly interpretable. This allows for a deeper understanding of how each operational parameter contributes to the behavior of the anaerobic digestion process.

The consistency between the model predictions and actual biogas production was further verified through visual trend comparison. As illustrated in Figure 4, the alignment between actual and predicted values combined with fitted linear trendlines demonstrates the model's ability to reliably follow the real behavior of the digestion process, even in the presence of natural fluctuations or measurement noise. This strengthens the case for using the model in real-time monitoring and forecasting applications in operational environments.

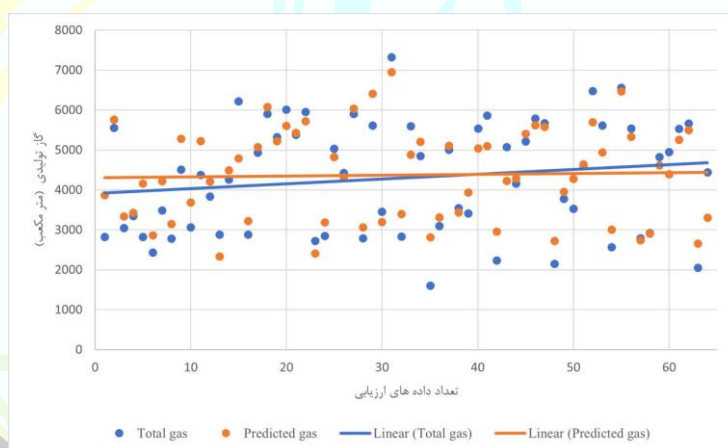


Figure 4. Actual vs. Predicted Biogas Production with Linear Trendlines

The comparison between actual and predicted gas production showed that the machine learning model effectively captured the overall trend in the data. The close alignment between the linear trendlines of the real and predicted values demonstrates the model's ability to reconstruct the system's dynamic behavior. While some deviations were observed, the predictions overall were consistent with real-world performance, suggesting the model's reliability for practical applications in energy management and process optimization.



A comparison with earlier works also helps place the current results into context. For instance, Kusiak and Wei (2011) reported R^2 values above 0.9 using ANFIS and neural networks [9], although their dataset was smaller and relied on fewer process variables. Similarly, Isenkul et al. (2025) obtained $R^2 = 0.898$ using a support vector regression model with six input features [10]. While these models show slightly better R^2 scores, they often require more complex configurations and data that are not always available in routine operations. In contrast, the current model achieves respectable accuracy with just three readily accessible inputs, which makes it more suitable for practical integration into real-time monitoring and control systems.

Further supporting this point, Yetilmezsoy et al. (2021) developed an artificial neural network for a full-scale sludge digestion system and reported $R^2 = 0.813$ with $MSE = 0.4019$ [11], and a fuzzy logic model that improved R^2 to 0.88, though both used five input variables. In a separate study, Yetilmezsoy et al. (2013) designed ANN-based models with R^2 values of 0.935 and 0.924 for biogas and methane predictions, respectively, but required eight process parameters [12]. Compared to these models, our approach provides a favorable trade-off between accuracy and simplicity.

Additionally, experimental studies such as that by Sunada et al. (2012) revealed that shorter hydraulic retention times (7 to 14 days) in anaerobic digestion of poultry effluent could lead to high methane production rates, up to $0.41 \text{ m}^3 \text{ CH}_4/\text{kg COD}$ added. Although their work was empirical and did not report metrics like R^2 or MSE , their findings reinforce the notion that process design and input characteristics significantly influence gas yield—factors that our model captures through key operational parameters.

Finally, in a comprehensive review, Swami et al. (2023) emphasized the increasing role of artificial intelligence in optimizing biogas plant operations [13]. Among the AI techniques examined, ensemble methods were identified as highly effective in handling complex, nonlinear systems. This observation aligns closely with our findings, further underscoring the suitability of ensemble models like ExtraTreesRegressor for wastewater treatment applications.

4. Conclusions

This study confirmed the potential of ensemble-based machine learning methods, with ExtraTreesRegressor achieving strong predictive accuracy ($R^2 = 0.8422$, $MSE = 296,655.1$, and $MAPE = 11.8\%$) for biogas production in the South Tehran Wastewater Treatment Plant. Remarkably, the model required only three operational variables, daily influent and effluent flow rates and polymer dosage, demonstrating that reliable predictions can be obtained without complex or costly data collection.

Compared with existing studies, this approach offers a meaningful balance between interpretability and performance. While some models have reported slightly higher accuracies, they often rely on larger input sets or opaque architectures. Here, SHAP analysis enhanced transparency by identifying polymer dosage as the dominant driver of methane production.



These findings highlight the broader value of data-driven methods in wastewater treatment management. By integrating such models into operational workflows, plant operators can stabilize digester performance, optimize methane yields, and reduce dependence on fossil fuels. Ultimately, this research demonstrates how combining machine learning with process engineering can support sustainable energy recovery and climate change mitigation.

Acknowledgements

The authors sincerely extend their gratitude to the colleagues at the Tehran Wastewater Engineering Company and the South Tehran Wastewater Treatment Plant, particularly to Engr. Norouzi, Director of Modules 5 and 6, for their kind cooperation and invaluable assistance in facilitating the collection of the necessary data for this study.

References

- [1] S. Achinas, V. Achinas, and G. J. W. Euverink, "A Technological Overview of Biogas Production from Biowaste," *Engineering*, vol. 3, no. 3, pp. 299–307, 2017, doi: 10.1016/J.ENG.2017.03.002.
- [2] M. P. Marttila, V. Uusitalo, L. Linnanen, and M. H. Mikkilä, "Agro-industrial symbiosis and alternative heating systems for decreasing the global warming potential of greenhouse production," *Sustain.*, vol. 13, no. 16, 2021, doi: 10.3390/su13169040.
- [3] S. Sunada *et al.*, "Potential of biogas and methane production from anaerobic digestion of poultry slaughterhouse effluent," *Rev. Bras. Zootec.*, vol. 2012, pp. 2379–2383, 2012.
- [4] P. Pochwatka and J. Dach, "Biogas Plant Exploitation in a Middle-Sized Dairy," no. December, pp. 0–17, 2020.
- [5] A. Hosseinnajad, Y. Saboohi, G. Zarei, and J. Shayegan, "Developing an integrated model for allocating resources and assessing technologies based on the watergy optimal point (water-energy nexus), case study: a greenhouse," *J. Sustain. Dev. Energy, Water Environ. Syst.*, vol. 11, no. 1, 2023, doi: 10.13044/j.sdewes.d10.0416.
- [6] E. Sihlangu, D. Luseba, T. Regnier, P. Magama, I. Chiyanzu, and K. A. Nephawe, "Investigating Methane, Carbon Dioxide, Ammonia, and Hydrogen Sulphide Content in Agricultural Waste during Biogas Production," *Sustain.*, vol. 16, no. 12, 2024, doi: 10.3390/su16125145.
- [7] J. Sun, Y. Xu, S. Nairat, J. Zhou, and Z. He, "Prediction of biogas production in anaerobic digestion of a full-scale wastewater treatment plant using ensembled machine learning models," *Water Environ. Res.*, vol. 95, no. 6, 2023, doi: 10.1002/wer.10893.
- [8] B. G. Laycock, C. M. Chan, and P. J. Halley, "A review of computational approaches used in the modelling, design, and manufacturing of biodegradable and biobased polymers," *Prog. Polym. Sci.*, vol. 157, 2024, doi: 10.1016/j.progpolymsci.2024.101874.
- [9] A. Kusiak and X. Wei, "Prediction of methane production in wastewater treatment facility: A data-mining approach," *Ann. Oper. Res.*, vol. 216, no. 1, pp. 71–81, 2014, doi: 10.1007/s10479-011-1037-6.



- [10] M. E. Isenkul *et al.*, “Predicting biogas production in real scale anaerobic digester under dynamic conditions with machine learning approach,” *Environ. Res. Commun.*, vol. 7, no. 6, p. 065016, 2025, doi: 10.1088/2515-7620/ade03b.
- [11] K. Yetilmezsoy, K. Karakaya, M. Bahramian, S. A. Abdul-Wahab, and B. İ. Goncaloğlu, “Black-, gray-, and white-box modeling of biogas production rate from a real-scale anaerobic sludge digestion system in a biological and advanced biological treatment plant,” *Neural Comput. Appl.*, vol. 33, no. 17, pp. 11043–11066, 2021, doi: 10.1007/s00521-020-05562-7.
- [12] K. Yetilmezsoy, F. I. Turkdogan, I. Temizel, and A. Gunay, “Development of ann-based models to predict biogas and methane productions in anaerobic treatment of molasses wastewater,” *Int. J. Green Energy*, vol. 10, no. 9, pp. 885–907, 2013, doi: 10.1080/15435075.2012.727116.
- [13] S. Swami, S. Suthar, R. Singh, A. K. Thakur, L. R. Gupta, and V. S. Sikarwar, “Integration of anaerobic digestion with artificial intelligence to optimise biogas plant operation,” *Environ. Dev. Sustain.*, vol. 27, no. 5, pp. 9773–9803, 2025, doi: 10.1007/s10668-023-04326-2.