# Forecasting Biogas Production Using Transformer-Based Deep Learning

Author: Lukas Bormann

2025

## Abstract

This project investigates the use of transformer-based deep learning models for time-series forecasting of biogas production in industrial anaerobic digestion (AD) systems. A custom transformer architecture was implemented from scratch in PyTorch and trained on multivariate plant data with a 2-hour sampling interval. The model forecasts up to 60 time steps (5 days) into the future using sequence-to-sequence attention mechanisms. A wide hyperparameter search combined with iterative evaluation of $R^2$ and MAE over increasing horizons enabled model refinement under real-world conditions. Compared to classical models such as Random Forests and Artificial Neural Networks, which focus on one-step-ahead predictions, the transformer demonstrates robust performance over extended windows. The approach is evaluated with step-wise performance metrics and stress-tested under varying input lengths and architectural depths. Results suggest that non-autoregressive multi-step prediction is feasible and informative for operational planning in AD plants, representing a promising direction for long-range forecasting in energy systems.

**Keywords:**
*Transformer, Time-series forecasting, Biogas production, Anaerobic digestion, Machine learning, Deep learning, Attention mechanism, PyTorch, $R^2$, MAE, Hyperparameter optimization, Optuna, Renewable energy, Data-driven modeling*

# Contents

# 1 Introduction

Biogas has become a strategic pillar in the global transition toward renewable energy, with production volumes reaching 38.1 billion cubic meters in 2020—equivalent to 1.46 exajoules of energy—driven by over 18,000 operational plants across Europe and large-scale deployments in Asia. [1] In Germany alone, over 9,700 plants produce 87 TWh of energy annually, and the global biogas market is projected to grow from USD 133.6 billion in 2024 to USD 191.2 billion by 2032. [2] As anaerobic digestion (AD) facilities shift toward advanced waste-based feedstocks and high-purity biomethane injection into natural gas grids, accurate forecasting of gas output becomes increasingly critical for operational reliability, financial planning, and grid integration. Although the biochemical processes underlying AD have been understood since the early 20th century, their manifestation in industrial contexts involves a multitude of interacting and nonlinear variables—ranging from microbial dynamics to feedstock variability—which render biogas yield prediction a complex and unresolved challenge. [3, 4]

Recent efforts have focused on applying data-driven methods to model these complex dependencies. Supervised machine learning (ML) approaches have shown considerable success when applied to historical plant data with dense temporal resolution. In particular, ensemble methods such as Random Forest (RF) and gradient-boosted trees have repeatedly outperformed linear models and traditional neural networks when forecasting biogas output for short-term horizons, typically one timestep (2 to 24 hours) into the future. Yildirim and Ozkaya, for instance, achieved an $R^2$ of 0.924 using RF for single-step predictions on a large-scale facility using 1-hour resolution data [5]. Li et al. evaluated multi-feature models on daily biogas data and reported $R^2$ values between 0.48 and 0.74 using RF and ANN models for 1-day-ahead forecasts, with accuracy sharply degrading beyond that horizon [6]. Thus, although current models perform well for immediate or next-day predictions, their efficacy deteriorates when extended to multi-step ahead forecasting, a capability crucial for mid-term operational planning.

In contrast to traditional architectures, transformer models are designed to capture long-range temporal dependencies without sequential recurrence. Originally developed for natural language processing, transformers use self-attention mechanisms to globally weigh input features across time, which has enabled state-of-the-art performance in sequence modeling tasks across domains. In time-series applications, this allows forecasting models to jointly consider entire sequences of past observations and generate multiple future timesteps in parallel, without relying on autoregressive rollouts.

Despite their success in energy, finance, and health domains, transformers have not yet been applied to the task of biogas production forecasting. Prior works rely heavily on sequential input-output mappings, where each prediction is conditioned on previ-

ous outputs, compounding error over time. Just a few studies have explored whether transformer-based models can improve multi-step prediction robustness by leveraging full sequence-to-sequence modeling capabilities in this context, showing promising results. [7]

This project investigates how effectively a transformer-based model can forecast biogas production over extended horizons in a single forward pass. Specifically, it examines how model capacity, prediction depth, and interpretability are interrelated, and what trade-offs exist between forecast length, model size, and predictive accuracy. The hyperparameter space is treated as a dynamic network of interdependent design decisions, each affecting convergence and generalization. This work serves as an initial baseline for future transformer-based forecasting in anaerobic digestion, already achieving both improved predictive accuracy and longer forecast horizons compared to classical approaches.

# 2 Methods

The Transformer model architecture implemented in this work was developed from scratch in Python using PyTorch, adhering to the encoder–decoder structure proposed by Vaswani et al. (2017). All components, including multi-head self-attention, position-wise feedforward layers, and sinusoidal positional encodings, were implemented without reliance on high-level forecasting libraries. The framework allows for complete configurability across architectural depth, dimensionality, sequence span, and output strategy.

The input data consists of a multivariate time series logged at 2-hour intervals, already aligned and cleaned. Temporal feature engineering (e.g., cyclic encoding of hour or day) was applied to enrich the representation while preserving generality. The input features included operational variables such as substrate loading, temperature, and gas production rates, selected through correlation analysis and kernel principal component analysis (KPCA). For further details on dataset composition and preprocessing, see the associated database paper. [5, 6]

The model receives as input a fixed-length sequence, an output forecast length, and a learning rate—all treated as configurable parameters. The architecture supports full-sequence decoding in parallel (non-autoregressive), which significantly reduces inference time and avoids error accumulation across successive steps, particularly when used in combination with modern high-throughput hardware and optimized deep learning libraries.

A broad hyperparameter search was conducted using the Optuna framework during early experimentation. Architectural and training parameters were sampled across wide ranges, followed by convergence toward empirically robust configurations. The input sequence length `sequence_length` $\in [20, 260]$ was kept variable throughout, as no consistent performance trend was observed. The output length `output_length` $\in \{20, 30, 60\}$ was ultimately fixed to 60, after early trials indicated no degradation in predictive accuracy. The model dimensionality `d_model` $\in \{64, 128, 256\}$ and the feedforward size `dim_feedforward` $\in \{256, 512, 1024\}$ were treated as internal buffer capacities, providing architectural flexibility across trials. The number of attention heads `nhead` $\in \{2, 4, 8\}$ was also kept open to adjust the attention mechanism's parallelism. Encoder depth `num_encoder_layers` $\in \{2, 3, 4\}$ was preserved as a free parameter under the hypothesis that longer-range prediction may benefit from greater representational depth. Dropout regularisation was applied with `dropout` $\in [0.1, 0.4]$ to mitigate overfitting risks, given the use of a single dataset.

The training objective was the mean squared error (MSE), with maximisation of the coefficient of determination ($R^2$) as a secondary goal. Models were trained for 15 epochs using mini-batches with `batch_size` $\in \{32, 64\}$, and an 80/20 train–validation split. The

evaluation script computed a comprehensive set of metrics—mean absolute error (MAE), $R^2$, and step-wise degradation—across the full output forecast horizon.

To support modular experimentation, the entire training pipeline was built with reusable components for dataset preprocessing, model checkpointing, performance tracking, and visual diagnostics. This design enables fast iteration and transfer to similar forecasting tasks in other industrial contexts.

All code, trained models, and evaluation results are available in the accompanying GitHub repository `[https://github.com/adg88lu/Master]`. The repository includes complete training configurations, hyperparameter sets, data scalers, saved model checkpoints, loss curves, and additional diagnostic tools such as scatter plots and evaluation helpers. The following section presents the final results derived from the evaluated model configurations.

# 3 Results and Discussion

The evaluation of forecasting performance was conducted iteratively, with each training run contributing both to model refinement and to insights into architectural dynamics. Rather than treating model performance as a static output, the experimental setup followed an adaptive progression—where evaluation results directly informed the next configuration. Consequently, the results and discussion are interwoven, as findings from one trial guided subsequent modifications in model design and training scope.

Initially, a wide-range hyperparameter search was performed using the Optuna framework. This search covered a broad space of model configurations—including variation in input sequence length, model width (`d_model`), encoder depth (`num_encoder_layers`), attention heads (`nhead`), and learning rate. The objective was to assess which type of Transformer architecture would emerge as optimal under minimal constraint. Early trials favored relatively shallow models with moderate embedding sizes (e.g. 128 or 256), which motivated a more focused examination of those regions.

As additional evaluation metrics were introduced—specifically the coefficient of determination ($R^2$) and mean absolute error (MAE) across forecast horizons (e.g. after 20, 40, and 60 timesteps)—the analysis became more granular. Model performance was no longer assessed solely via final validation loss, but through its robustness across the full multi-step output window, offering a more comprehensive view of temporal generalization.

Over the course of the experiment, the forecasting horizon was progressively extended. Initial runs targeted an output length of 30 timesteps (2.5 days), while later trials tested 60-step forecasts (5 days), as earlier models maintained acceptable performance even with extended outputs. This provided initial evidence for the Transformer's capacity to model long-range dependencies without significant degradation in accuracy.

To probe model stability and boundary behavior, additional stress tests were conducted. These involved exaggerated configurations: models with extremely long input sequences (up to 260 timesteps), minimal or excessive encoder depth, and varied feedforward dimensions. These setups were intended to expose breaking points in learning dynamics and to verify the model's ability to remain stable under extreme design conditions.

After approximately ten Optuna trials, the hyperparameter search began to exhibit non-deterministic or locally biased behavior, without clear convergence toward any specific region of the parameter space. As a result, the most stable and promising configurations were manually fixed for further evaluation. Subsequent training runs focused on controlled refinements around these baselines to reduce noise introduced by overly random parameter sampling.

The following table presents all evaluated model configurations alongside their key hyperparameters and resulting performance metrics, including average $R^2$, stepwise $R^2$ degradation, and corresponding MAE.

Table 3.1: Hyperparameters and performance metrics for Transformer model comparison

| **Parameter / Metric** | **Try 1** | **Try 2** | **Try 3** | **Try 4** | **Try 5** | **Try 6** |
|---|---|---|---|---|---|---|
| *sequence_length* | 20 | 140 | 260 | 100 | 220 | 120 |
| *output_length* | 10 | 30 | 60 | 60 | 60 | 60 |
| *batch_size* | 32 | 64 | 64 | 32 | 32 | 64 |
| *d_model* | 256 | 128 | 64 | 128 | 128 | 512 |
| *nhead* | 4 | 2 | 4 | 2 | 4 | 4 |
| *num_encoder_layers* | 4 | 2 | 3 | 4 | 2 | 2 |
| *dim_feedforward* | 256 | 512 | 512 | 512 | 256 | 256 |
| *dropout* | 0.2 | 0.2 | 0.2 | 0.2 | 0.4 | 0.2 |
| *learning_rate* | 1E-05 | 1E-05 | 1E-05 | 1E-05 | 1E-05 | 1E-05 |
| *num_epochs* | 15 | 15 | 15 | 15 | 15 | 15 |
| *total_model_params* | 2.983.169 | 1.072.001 | 580.609 | 1.468.545 | 940.417 | 5.627.905 |
| *folder_processing_time* | 816.87 | 495.72 | 1.477.77 | 863.76 | 934.95 | 1500.03 |
| *average* $R^2$ | / | / | 0.11 | 0.66 | 0.49 | 0.50 |
| $R^2$ *@ 20 timesteps* | / | / | 0.11 | 0.66 | 0.49 | 0.55 |
| $R^2$ *@ 40 timesteps* | / | / | 0.16 | 0.67 | 0.50 | 0.49 |
| $R^2$ *@ 60 timesteps* | / | / | 0.05 | 0.65 | 0.47 | 0.46 |
| *average MAE* | / | / | 3.14 | 2.60 | 2.98 | 2.37 |
| *MAE @ 20 timesteps* | / | 2.25 | 3.89 | 2.61 | 2.43 | 1.45 |
| *MAE @ 40 timesteps* | / | 0.09 | 1.78 | 3.25 | 1.79 | 2.64 |
| *MAE @ 60 timesteps* | / | / | 3.76 | 1.94 | 4.72 | 3.01 |

Experimentation began with Try 1, a large model (2.98 million parameters) using only 20 input and 10 output timesteps. However, it yielded no usable evaluation metrics—highlighting that minimal lookback windows limit the transformer's ability to generalize. In Try 2, we extended the historical context to 140 timesteps and predicted 30 steps ahead, with MSE values exceeding 10 and no valid $R^2$, suggesting that merely increasing sequence length is insufficient without proper model balance.

Try 3 utilized the longest lookback of 260 timesteps with a 60-step output and finally yielded valid metrics: an average $R^2 = 0.11$ and MAE = 3.14, indicating some learning but still weak performance, as shown in Table 3. In contrast, Try 4 emerged as the best configuration, achieving an average $R^2 = 0.66$ and MAE = 2.60 with just 100 input timesteps—demonstrating that more historical data is not necessarily better, and that architectural balance is critical. From this point on, we fixed the prediction horizon to 60 timesteps (5 days) to standardize comparisons.

Try 5 and Try 6 revisited longer lookback windows (220 and 120 timesteps, respectively)

and included increased model complexity (e.g., 4 attention heads in both, and 5.6 million parameters in Try 6). However, performance deteriorated: Try 5 achieved $R^2 = 0.49$, MAE = 2.98 and Try 6 reached $R^2 = 0.50$, MAE = 2.37. These outcomes suggest that larger models with more attention heads may overfit or become inefficient when not aligned with proper training dynamics or input regularization. Also, bigger datasets may be needed with cross-validation for more robust training. Signs of overfitting can be observed in the training and validation loss curves.

Compared to existing work in biogas yield prediction, our study significantly extends both the temporal scope and the methodological depth. Previous studies, such as those by Yildirim and Ozkaya [5] and Li et al. [6], focus exclusively on single-step-ahead prediction, forecasting either one hour or one day into the future based on a static window of past inputs. In contrast, the transformer-based architecture developed here was trained to generate multi-step parallel forecasts spanning up to five days (60 timesteps) ahead, using input sequences as long as 260 timesteps (approximately 21 days). This not only demands greater temporal abstraction but also tests the model's ability to retain accuracy over long forecast horizons.

Furthermore, while prior studies report $R^2$ values of up to 0.92, these are measured for immediate next-step predictions—where statistical learning models such as Random Forests typically perform well. By contrast, this work addresses the more challenging task of maintaining predictive fidelity across multiple future steps, and evaluates degradation explicitly through step-wise $R^2$ and MAE. The Transformer's ability to model the entire output horizon jointly—rather than relying on autoregressive chaining—enables direct optimization of long-range forecasts, better matching the practical demands of multi-day planning and early anomaly detection in industrial AD settings.

This approach thus directly responds to the core research question: it evaluates how transformer models behave when configured for non-sequential, multi-output forecasting, and demonstrates their potential advantages—such as long-range modeling and flexibility in sequence design—as well as their current limitations, including training complexity and sensitivity to data structure, when compared with classical forecasting alternatives.

While this work demonstrates that transformer-based models can handle long-range forecasting with reasonable accuracy, it also reveals the intricate relationship between training time, model complexity, and prediction stability. The increase in parameter count and sequence length offers new potential but also introduces architectural fragility and interpretability challenges. These observations highlight the importance of systematic follow-up studies that fix individual parameters and isolate their impact—ideally through large-scale training experiments and scatter plot analyses of loss surfaces and predictive fidelity. The next logical step is to test not only simultaneous forecasting but also autoregressive multi-step chaining, helping clarify where the balance lies between stepwise accuracy and long-horizon generalization.

# 4 Outlook

Unlike previous studies that focus solely on single-step forecasts, this work explores how far ahead a Transformer-based model can predict without collapsing in accuracy. While one-step-ahead models may achieve high $R^2$ scores—as reported in classical machine learning literature—they offer limited foresight for real-world operational planning. By contrast, the model presented here predicts entire output sequences of up to 60 timesteps (approximately five days) in parallel, enabling a richer view into system dynamics and future trajectories.

This design, however, also introduces new challenges. As the forecasting horizon extends, the risk of overfitting and representational drift increases, particularly for deeper or wider models. Future work should explore hybrid strategies that combine short-term precision with long-range structure—for example, by mixing direct sequence forecasting with autoregressive rollouts. Another promising direction is to investigate how predictive accuracy degrades as a function of output depth and architecture size, which would require large-scale parametric sweeps and model stability analyses.

In addition, studies could assess whether transformer-based architectures can be trained to issue multiple rolling predictions in sequence, potentially adapting their horizon based on recent performance. Such an approach could provide an adaptive forecasting mechanism tailored to operational reliability requirements. Overall, this work forms a solid foundation for future studies into long-horizon time-series modeling using attention-based architectures in biogas forecasting and beyond.

# Bibliography


[1] *Global Bioenergy Statistics - Worldbioenergy*. URL: https://www.worldbioenergy.org/global-bioenergy-statistics/ (visited on 06/15/2025).

[2] *Biogas Market Size, Share & Growth | Industry Analysis Report [2032]*. URL: https://www.fortunebusinessinsights.com/industry-reports/biogas-market-100910 (visited on 06/15/2025).

[3] Rimika Kapoor et al. "Evaluation of Biogas Upgrading Technologies and Future Perspectives: A Review". In: *Environmental Science and Pollution Research International* 26.12 (Apr. 2019), pp. 11631–11661. ISSN: 1614-7499. DOI: 10.1007/s11356-019-04767-1. PMID: 30877529.

[4] Ahinara Francisco López et al. "From Biogas to Biomethane: An In-Depth Review of Upgrading Technologies That Enhance Sustainability and Reduce Greenhouse Gas Emissions". In: *Applied Sciences* 14.6 (6 Jan. 2024), p. 2342. ISSN: 2076-3417. DOI: 10.3390/app14062342. URL: https://www.mdpi.com/2076-3417/14/6/2342 (visited on 06/15/2025).

[5] Oznur Yildirim and Bestami Ozkaya. "Prediction of Biogas Production of Industrial Scale Anaerobic Digestion Plant by Machine Learning Algorithms". In: *Chemosphere* 335 (Sept. 1, 2023), p. 138976. ISSN: 0045-6535. DOI: 10.1016/j.chemosphere.2023.138976. URL: https://www.sciencedirect.com/science/article/pii/S0045653523012432 (visited on 06/14/2025).

[6] Chao Li et al. "Exploring Available Input Variables for Machine Learning Models to Predict Biogas Production in Industrial-Scale Biogas Plants Treating Food Waste". In: *Journal of Cleaner Production* 380 (Dec. 20, 2022), p. 135074. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2022.135074. URL: https://www.sciencedirect.com/science/article/pii/S0959652622046480 (visited on 06/14/2025).

[7] Yongming Han et al. "Time Series Prediction of Anaerobic Digestion Yield and Carbon Emissions from Food Waste Based on iTransformer Model". In: *Chemical Engineering Journal* 513 (2025). URL: https://urn.kb.se/resolve?urn=urn:nbn:se:du-50555 (visited on 06/15/2025).