GFIS: A Physics-Guided Hybrid AI Digital Twin for Methane Yield Forecasting, Soft Sensing, and Stability-Aware Optimization in Anaerobic Digestion

Original Research Article - Springer-style Manuscript Draft

Akash S. Chatake1*, Dipti Jadhav2, Navnath K. Mane3

1. Department of Mathematics / Artificial Intelligence and Machine Learning, Shri Siddheshwar Women's Polytechnic, Solapur, Maharashtra, India; M.Tech Artificial Intelligence and Machine Learning, BITS Pilani WILP, India.

2. Department of Computer Science and Engineering / Information Technology, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Navi Mumbai, Maharashtra, India.

3. Department of Computer Engineering, Shri Siddheshwar Women's Polytechnic, Solapur, Maharashtra, India.

*Corresponding author: Akash S. Chatake; Email: 2024ab05226@wilp.bits-pilani.ac.in

Author note: Final author order, affiliation wording, and co-author consent must be confirmed before journal submission.

Manuscript status: This is a publication-ready structural draft. Numerical values in the Results section are intentionally marked TBD until GFIS experiments are executed. Do not submit to a journal before replacing placeholders with verified results.

Abstract

Purpose: Anaerobic digestion is a nonlinear and multistage biochemical process in which methane yield depends on operational conditions, feedstock composition, microbial state, and process stability. Mechanistic models such as ADM1 provide interpretability but are difficult to calibrate and deploy as lightweight real-time decision-support systems. Pure machine-learning models can forecast methane production but may produce physically implausible outputs under unseen operating regimes. This study proposes the Green Fuel Intelligence System (GFIS), a physics-guided hybrid AI digital twin for methane-yield forecasting, soft sensing, and stability-aware scenario simulation in anaerobic digestion.

Methods: GFIS combines an XGBoost tabular learner, an LSTM sequence learner, ensemble fusion, a physics-guided methane feasibility constraint based on volatile solids and methane-potential bounds, and a soft-sensor module for VFA/ALK estimation and stability classification. The framework is designed for time-aware evaluation using plant-scale operational/SCADA data and synthetic physics-bounded scenario trajectories. The proposed software stack includes deterministic preprocessing, leakage-safe temporal splits, feature engineering, SHAP explainability, constrained what-if simulation, and a lightweight FastAPI/Streamlit/SQLite deployment layer.

Results: The manuscript defines a complete experimental protocol comparing mean, linear, random forest, XGBoost, LSTM, hybrid ensemble, and physics-guided GFIS variants using MAE, RMSE, R2, physics violation rate, VFA/ALK regression error, macro-F1, calibration error, robustness tests, and scenario-response evaluation. Final numerical results are to be inserted after execution of the reproducible pipeline.

Conclusion: GFIS is positioned as a reproducible physics-guided AI framework that links methane forecasting, biochemical inference, physical consistency checking, and digital-twin simulation. The contribution is not a single model but a validated integration strategy for reliable AI-assisted decision support in sustainable biogas systems.

Keywords: anaerobic digestion; methane yield prediction; physics-informed machine learning; digital twin; LSTM; XGBoost; soft sensor; VFA/ALK ratio; SHAP; biogas optimization

1. Introduction

Anaerobic digestion (AD) is a biological waste-to-energy process that converts organic substrates into methane-rich biogas under oxygen-free conditions. It is relevant to renewable energy generation, organic waste valorization, circular bioeconomy, wastewater treatment, and greenhouse-gas mitigation. Although the technology is mature, AD remains difficult to optimize because the process is governed by coupled biochemical reactions, microbial population dynamics, feedstock heterogeneity, pH buffering, volatile fatty acid accumulation, and temperature-sensitive kinetics.

Methane yield is a critical performance indicator because it directly reflects energy recovery. However, methane production is not controlled by one variable alone. It depends on interacting factors such as temperature, pH, organic loading rate (OLR), hydraulic retention time (HRT), total solids (TS), volatile solids (VS), carbon-to-nitrogen ratio (C/N), alkalinity, VFA concentration, gas composition, and feedstock variability. These dependencies are nonlinear and time delayed. As a result, conventional statistical models often fail to capture process dynamics, while high-fidelity mechanistic models may be too complex for lightweight operational decision support.

The Anaerobic Digestion Model No. 1 (ADM1) is the dominant mechanistic reference model for AD. ADM1 formalizes hydrolysis, acidogenesis, acetogenesis, methanogenesis, gas-liquid equilibrium, pH, inhibition effects, and multiple substrate and microbial states. It provides biochemical interpretability and supports mechanistic simulation. Nevertheless, ADM1 requires careful parameterization and calibration, and its dynamic differential-algebraic structure makes direct real-time deployment difficult in many practical settings.

Machine-learning (ML) approaches offer an alternative by learning predictive mappings from plant data. Tree-based models such as XGBoost are strong for engineered tabular features, while sequence models such as LSTM networks are suitable for time-dependent process histories. Yet purely data-driven approaches may violate domain constraints. For example, a black-box model can predict negative methane or methane above feasible biochemical limits. Such errors are not merely numerical inaccuracies; they reduce trust and limit adoption in process-engineering contexts.

Physics-informed machine learning addresses this weakness by embedding scientific knowledge into the learning or inference process. In AD, the most defensible physical constraint is that predicted methane cannot exceed feasible methane potential implied by substrate availability and volatile solids. Similarly, VFA/ALK ratio can be used as a stability indicator, because VFA accumulation and insufficient alkalinity buffering signal process imbalance. A digital twin adds a system-level layer: rather than only predicting a point estimate, it provides a virtual representation for scenario simulation, stability warning, and operating-region exploration.

This paper proposes GFIS - the Green Fuel Intelligence System - as a physics-guided hybrid AI digital twin for anaerobic digestion. GFIS integrates tabular prediction, temporal learning, physics-constrained methane forecasting, soft sensing of stability indicators, explainability, scenario simulation, and safe operating optimization.

1.1 Contributions

The main contributions of this work are: (1) a hybrid methane-yield forecasting architecture combining XGBoost, LSTM, and ensemble fusion; (2) a physics-guided feasibility layer that constrains methane predictions using volatile-solids-based methane-potential bounds; (3) a soft-sensor module for estimating VFA/ALK ratio and classifying AD stability into Stable, Warning, and Critical states; (4) a digital-twin simulation workflow for what-if analysis of OLR, temperature, pH, HRT, moisture, and feedstock-related changes; (5) a complete evaluation design covering forecasting accuracy, physical consistency, soft-sensor performance, calibration, robustness, explainability, and constrained optimization; and (6) a lightweight implementation plan using Python, XGBoost, PyTorch, SHAP, FastAPI, Streamlit, and SQLite.

2. Related Work

ADM1 remains the canonical reference for mechanistic AD modelling. It represents the biochemical conversion chain from complex organic matter to methane and carbon dioxide through hydrolysis, acidogenesis, acetogenesis, and methanogenesis. Its strength is interpretability and process-level detail. However, its complexity, calibration burden, and sensitivity to uncertain kinetic parameters make it difficult to use as a lightweight decision-support layer in ordinary software prototypes.

ML models have been increasingly used to forecast biogas production from operational data, laboratory measurements, and SCADA histories. Tabular methods such as random forest, gradient boosting, and XGBoost can capture nonlinear feature interactions, while neural networks can learn complex mappings when sufficient data are available. Sequence models are especially relevant because anaerobic digestion exhibits lag effects: changes in feedstock or loading may influence methane output only after a delay.

Physics-informed ML integrates domain knowledge, constraints, conservation laws, or governing relations into learning systems. In GFIS, the physics is expressed through biochemical feasibility rather than a full mechanistic solver. The model output is checked against methane-potential limits derived from volatile solids and substrate assumptions. This reduces physically unrealistic predictions and improves model credibility.

Soft sensors estimate variables that are difficult, delayed, or expensive to measure directly. In anaerobic digestion, VFA and alkalinity are important stability indicators, but they are not always available as real-time measurements. A soft sensor can infer VFA/ALK ratio from accessible process variables and model outputs, enabling early warnings and operational decision support.

Digital twins provide virtual representations of physical systems for monitoring, simulation, and decision support. GFIS is not intended to be a full mechanistic plant clone. Instead, it is an operational surrogate twin that combines data-driven forecasting, physics consistency checks, soft sensing, and what-if simulation for AD process intelligence.

3. Materials and Methods

GFIS follows a modular surrogate-twin architecture. Raw plant data, laboratory data, SCADA histories, and synthetic physics-bounded scenario data are processed into tabular features and sequence windows. XGBoost learns nonlinear structured relationships from engineered features. LSTM learns temporal dependencies from process histories. Their outputs are fused in an ensemble layer. The physics layer checks methane feasibility. A soft-sensor branch estimates VFA/ALK ratio and stability state. The digital twin layer performs what-if simulation and constrained optimization.

Figure 1. Overall GFIS architecture

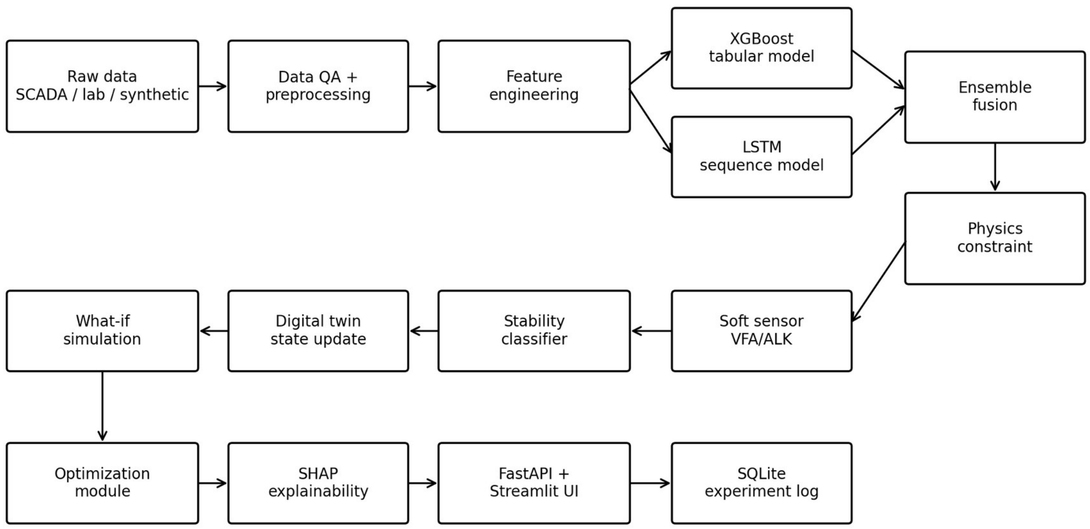

Fig. 1 Overall GFIS architecture integrating data processing, hybrid learning, physics constraint, soft sensing, digital twin simulation, explainability, and deployment.

3.1 Data sources and strategy

GFIS is designed to use a hybrid data backbone: real or public AD time-series data as the empirical anchor; synthetic scenario trajectories to extend coverage across operating regimes; literature-supported methane-potential and stability thresholds for physical constraints; and optional facility-level context data for external interpretation. Since no single public dataset is expected to cover all possible feedstocks and instability regimes, synthetic trajectories are used to cover rare or unsafe scenarios in a controlled manner.

3.2 Target variables

The primary target variable is methane yield or methane production rate. Secondary targets are VFA/ALK ratio and stability class. Let the feature vector at time t be $x_t = [temperature_t, pH_t, OLR_t, HRT_t, TS_t, VS_t, C/N_t, moisture_t, alkalinity_t, gas\_flow_t, ...]$. The forecasting task is $y_t = f(x_t, x_{t-1}, ..., x_{t-k})$, where y_t is methane yield or methane flow at the prediction horizon.

3.3 Preprocessing and feature engineering

The preprocessing pipeline includes time alignment and resampling, duplicate removal, range validation, missing-value imputation with missingness indicators, leakage-safe chronological splitting, scaling for neural models, outlier tagging, and construction of train, validation, and test split manifests. A row-wise random split is not appropriate because it can leak future information in time-series forecasting. Therefore, blocked chronological or rolling-origin evaluation is used.

Feature group	Examples	Role
Operational variables	temperature, pH, OLR, HRT	process state
Feedstock variables	TS, VS, C/N, moisture	substrate quality
Lagged features	previous methane, previous pH, previous OLR	temporal memory
Rolling features	rolling mean, rolling slope, rolling	regime trend

	volatility	
Interactions	OLR x temperature, VS x HRT, pH x alkalinity	nonlinear coupling
Stability proxies	VFA trend, alkalinity trend, pH volatility	instability warning

Table 1 Candidate feature groups for methane-yield forecasting and stability soft sensing.

Figure 2. Reproducible GFIS data and experiment pipeline

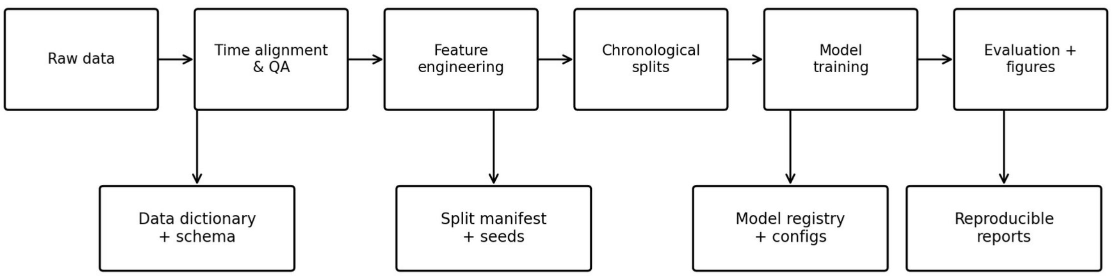

Fig. 2 Reproducible GFIS data and experiment pipeline.

3.4 XGBoost tabular model

XGBoost is used as the main tabular learner. It is trained on engineered features and optimized using validation-based hyperparameter tuning. Candidate hyperparameters include number of trees, maximum depth, learning rate, subsampling ratio, column sampling ratio, and regularization terms.

3.5 LSTM sequence model

The LSTM branch receives sliding windows of multivariate process histories. For a window length k, the input tensor has shape [batch_size, k, number_of_features]. The LSTM output is passed through fully connected layers to predict methane yield at the chosen horizon. Dropout and early stopping are used to reduce overfitting.

3.6 Ensemble fusion

The first ensemble strategy is weighted late fusion: y_hat_t = w1 * y_hat_t^XGB + w2 * y_hat_t^LSTM, where w1 + w2 = 1. The weights are selected using validation-set performance. A stacked meta-learner may be added as an advanced variant.

3.7 Physics-guided methane feasibility constraint

The physics layer encodes the principle that methane prediction cannot exceed feasible methane production implied by volatile solids and methane potential. If elemental composition C_aH_bO_cN_d is available, a Buswell-type theoretical methane potential can be used: TBMP_CH4 = 22.4 * (a/2 + b/8 - c/4 - 3d/8) / (12a + b + 16c + 14d). In operational settings, elemental composition is often unavailable. A practical upper-bound formulation is y_t^max = eta_t * VS_t * BMP_t^prior, where VS_t is volatile-solids availability, BMP_t^prior is the feedstock methane-potential prior, and eta_t is a conversion-efficiency factor. The physics penalty is L_total = L_pred + lambda * E[max(0, y_hat_t - y_t^max)^2 + max(0, -y_hat_t)^2]. This penalizes predictions above the upper bound and negative methane predictions.

3.8 Soft sensor and stability classification

The soft-sensor module estimates VFA/ALK ratio using accessible operational features and model outputs. Stability classes are defined as Stable when VFA/ALK < 0.4, Warning when 0.4 <= VFA/ALK <= 0.8, and Critical when VFA/ALK > 0.8. This two-stage design provides both regression output and decision-level stability interpretation.

3.9 Digital twin simulation and optimization

The digital twin receives current state variables and allows what-if perturbations. Example simulation questions include: what happens if OLR increases by 10%, what happens if temperature drops below target regime, what happens if pH shifts toward inhibition, and what is the safest operating region for high methane yield. The optimization objective is maximize_u [CH4_hat(u) - alpha * Risk(u) - beta * Violation(u)], where u is a candidate operating configuration, Risk(u) is predicted instability risk, and Violation(u) is the physics-consistency penalty.

3.10 Explainability and implementation

SHAP is used to explain the XGBoost component and selected ensemble outputs. Global explanations identify dominant drivers of methane yield and instability. Local explanations are used for critical cases. The proposed implementation stack is Python, Pandas, NumPy, Scikit-learn, XGBoost, PyTorch, SHAP, FastAPI, Streamlit, SQLite, and Docker.

4. Experimental Protocol

The evaluation protocol is designed to test prediction accuracy, physical consistency, stability inference, robustness, and decision-support utility. The forecasting models include mean predictor, linear regression, ridge/lasso, random forest, XGBoost, LSTM, weighted XGBoost-LSTM ensemble, and physics-guided GFIS ensemble. The soft-sensor study evaluates VFA/ALK regression and stability classification. Rolling-origin or blocked chronological splits are used to avoid temporal leakage.

Study	Comparison	Purpose
Model ablation	XGBoost vs LSTM vs ensemble	quantify hybrid benefit
Physics ablation	unconstrained vs clipped vs penalty	quantify physical consistency gain
Feature ablation	raw vs engineered features	quantify feature engineering value
Soft sensor ablation	direct classifier vs VFA/ALK regression + classifier	quantify soft-sensor benefit
Robustness ablation	full data vs reduced/noisy/missing data	quantify robustness

Table 2 Planned ablation studies for GFIS validation.

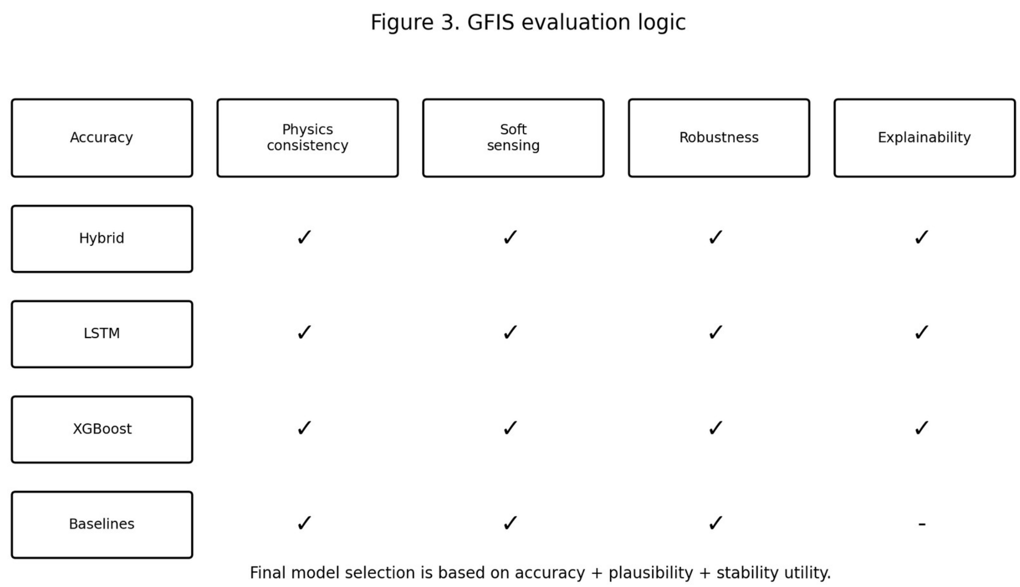

Fig. 3 Evaluation logic used to select the final GFIS model.

For methane regression, the primary metrics are MAE, RMSE, R2, and optionally sMAPE where denominator stability is acceptable. For physical consistency, the metrics are violation rate, mean excess over bound, negative prediction rate, and clipped-mass fraction. For VFA/ALK regression, MAE, RMSE, and R2 are used. For stability classification, precision, recall, macro-F1, weighted-F1, confusion matrix, AUROC where meaningful, and expected calibration error are used. Bootstrap confidence intervals and paired statistical tests are planned for headline comparisons.

5. Results

This section is intentionally structured as a results reporting template. Actual numerical values must be inserted after experimental execution. No values are fabricated in this manuscript draft.

Model	**MAE**	**RMSE**	**R2**	**sMAPE**	**Notes**
Mean baseline	TBD	TBD	TBD	TBD	baseline
Linear regression	TBD	TBD	TBD	TBD	classical
Random forest	TBD	TBD	TBD	TBD	tree baseline
XGBoost	TBD	TBD	TBD	TBD	tabular learner
LSTM	TBD	TBD	TBD	TBD	sequence learner
XGBoost + LSTM ensemble	TBD	TBD	TBD	TBD	hybrid learner
Physics-guided GFIS	TBD	TBD	TBD	TBD	final model

Table 3 Methane forecasting performance reporting template.

Model	**Violation rate**	**Mean excess over bound**	**Negative prediction rate**	**Comment**

XGBoost	TBD	TBD	TBD	unconstrained
LSTM	TBD	TBD	TBD	unconstrained
Ensemble	TBD	TBD	TBD	unconstrained
Clipped ensemble	TBD	TBD	TBD	post-hoc correction
Physics-penalty GFIS	TBD	TBD	TBD	train-time constraint

Table 4 Physical consistency comparison reporting template.

Model	**MAE**	**RMSE**	**R2**	**Macro-F1**	**ECE**	**Comment**
XGBoost soft sensor	TBD	TBD	TBD	TBD	TBD	tabular soft sensing
Sequence soft sensor	TBD	TBD	TBD	TBD	TBD	temporal soft sensing
GFIS stability classifier	TBD	TBD	TBD	TBD	TBD	final classification layer

Table 5 Soft-sensor and stability classification reporting template.

6. Discussion

GFIS is designed to address the gap between mechanistic interpretability and data-driven deployability. ADM1 provides scientific structure but may be difficult to calibrate and deploy as a lightweight real-time tool. Pure ML models may be accurate but can behave unrealistically outside familiar data regimes. The proposed architecture therefore adopts a hybrid strategy: ML learns plant-specific nonlinear and temporal relationships, while physics defines feasible output boundaries and biochemical semantics.

The expected scientific value of the study lies in the comparison between unconstrained ML and physics-guided variants. If the physics-guided GFIS model maintains comparable or better forecasting accuracy while substantially reducing infeasible predictions, it supports the central hypothesis that physics-guided constraints improve practical reliability. Similarly, if the soft sensor estimates VFA/ALK ratio with useful accuracy and supports stability classification, GFIS becomes more than a methane prediction model: it becomes a process intelligence framework.

The digital twin component adds decision-support value. Operators and researchers rarely need only a point prediction; they need to know how the system may respond to changes in OLR, temperature, pH, HRT, or feedstock conditions. A surrogate digital twin enables fast, safe, and explainable what-if analysis. The proposed optimization module provides a practical route toward identifying high-methane, low-risk operating conditions.

Several limitations must be acknowledged. If training relies heavily on synthetic data, external validity depends on how well synthetic trajectories approximate real AD dynamics. If real plant data are incomplete or from a single site, generalization to other feedstocks and plant configurations may be limited. VFA/ALK thresholds are useful operational heuristics but may vary with plant design and substrate type.

7. Conclusion

This paper presents GFIS, a physics-guided hybrid AI digital twin framework for methane-yield forecasting, soft sensing, and stability-aware scenario simulation in anaerobic digestion. The proposed framework combines XGBoost, LSTM, ensemble learning, volatile-solids-based methane feasibility constraints, VFA/ALK soft sensing, SHAP explainability, and digital twin simulation. The contribution is a complete scientific-engineering pipeline that moves beyond simple prediction toward physically reliable, explainable, and decision-oriented AI for sustainable biogas systems.

Future work will include execution on plant-scale datasets, validation against external datasets, uncertainty quantification, integration with IoT sensor streams, reinforcement learning for control, and deployment in pilot-scale or industrial biogas environments.

List of Abbreviations

AD: Anaerobic Digestion; ADM1: Anaerobic Digestion Model No. 1; AI: Artificial Intelligence; ALK: Alkalinity; API: Application Programming Interface; BMP: Biochemical Methane Potential; COD: Chemical Oxygen Demand; ECE: Expected Calibration Error; GFIS: Green Fuel Intelligence System; HRT: Hydraulic Retention Time; LSTM: Long Short-Term Memory; MAE: Mean Absolute Error; ML: Machine Learning; OLR: Organic Loading Rate; RMSE: Root Mean Square Error; SCADA: Supervisory Control and Data Acquisition; SHAP: SHapley Additive exPlanations; TS: Total Solids; VFA: Volatile Fatty Acids; VS: Volatile Solids.

Statements and Declarations

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript draft.

Competing interests: The authors declare that they have no competing interests.

Ethics approval: Not applicable. This study does not involve human participants, animals, or clinical data.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The final study will use publicly available anaerobic digestion datasets and synthetic data generated through reproducible scripts. Dataset links, preprocessing scripts, split manifests, and generated synthetic data will be provided in the final reproducibility package.

Code availability: The final implementation will be made available as a structured Python codebase containing data processing, modelling, physics constraint, soft sensor, digital twin, optimization, explainability, API, and dashboard modules.

Author contributions: ASC conceptualized the GFIS framework, prepared the initial methodology, and will implement the experimental pipeline. DJ will provide academic supervision and technical review. NKM will support institutional review and evaluation. Final contributions must be verified before submission.

Acknowledgements: The author acknowledges BITS Pilani WILP for the academic dissertation framework and the supervisor for research guidance. Final acknowledgements will be revised after completion of experiments and manuscript review.

Use of AI tools: AI-assisted drafting and structuring tools were used to support language polishing, document organization, and planning. Scientific responsibility, verification, implementation, experiments, and final manuscript approval remain with the human authors.

References

1. Batstone DJ, Keller J, Angelidaki I, Kalyuzhnyi SV, Pavlostathis SG, Rozzi A, Sanders WTM, Siegrist H, Vavilin VA. The IWA Anaerobic Digestion Model No. 1 (ADM1). Water Science and Technology. 2002;45(10):65-73.

2. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics. 2019;378:686-707.

3. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nature Reviews Physics. 2021;3:422-440.

4. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735-1780.

5. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

6. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017.

7. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning. 2017.

8. Schroer LA, Just CL. Feature engineering and supervised machine learning to forecast biogas production during municipal anaerobic co-digestion. ACS ES&T Engineering. 2024.

9. Yildirim O, Ozkaya B. Prediction of biogas production of industrial scale anaerobic digestion plant by machine learning algorithms. Chemosphere. 2023;335:138976.

10. Wang et al. Tree-based automated machine learning to predict biogas production for anaerobic co-digestion. 2021.

11. Angelidaki I, Sanders W. Assessment of the anaerobic biodegradability of macropollutants. Reviews in Environmental Science and Bio/Technology. 2004.

12. Buswell AM, Mueller HF. Mechanism of methane fermentation. Industrial and Engineering Chemistry. 1952.

13. NASA. Digital twins and their role in simulation and decision support. NASA Science resources.

14. FastAPI Documentation. FastAPI framework features and OpenAPI support.

15. Streamlit Documentation. Streamlit application framework.

16. SQLite Documentation. SQLite serverless database engine.

Appendix A. Core Implementation Skeleton

The following pseudocode illustrates the core physics penalty that will be used in the GFIS LSTM or neural fusion branch:

```
import torch

def physics_penalty(y_pred, y_max):
```

```
    upper = torch.relu(y_pred - y_max) ** 2
    lower = torch.relu(-y_pred) ** 2
    return (upper + lower).mean()

def total_loss(y_pred, y_true, y_max, lam=0.2):
    pred_loss = torch.mean((y_pred - y_true) ** 2)
    phys_loss = physics_penalty(y_pred, y_max)
    return pred_loss + lam * phys_loss
```