GFIS Dissertation Research Report

Project title: Green Fuel Intelligence System (GFIS): A Physics-Guided AI Digital Twin for Methane Yield Prediction and Stability Monitoring in Anaerobic Digestion
Student: Akash Shivadas Chatake
BITS ID: 2024AB05226
Program: M.Tech. Artificial Intelligence and Machine Learning
Institution: BITS Pilani WILP
Supervisor: Dr. Dipti Jadhav
Report status: Working research report for dissertation development

Abstract

Anaerobic digestion is a renewable-energy process that converts organic waste into methane-rich biogas, but its operation is difficult to optimize because methane production depends on nonlinear, delayed, and interacting biochemical variables. The Green Fuel Intelligence System (GFIS) is developed as a physics-guided AI digital twin for methane-yield prediction, stability monitoring, scenario simulation, and operating-condition optimization. The current implementation combines synthetic anaerobic-digestion data generation, feature engineering, XGBoost tabular prediction, PyTorch LSTM time-series modeling, physics-bound feasibility checking, VFA/ALK soft sensing, FastAPI services, Streamlit interaction, and HTML-based digital-twin demonstrations. The system achieved an XGBoost R2 score of 0.7556 with RMSE 6.7933 and MAE 5.5517 on the synthetic evaluation set, while reporting zero physics-bound violations. Controlled 48-hour stress experiments further demonstrate nominal operation, organic overload warning, pH shock response, low-temperature disturbance, high-solids feed behavior, and corrective recovery operation. The result is a research-grade prototype that moves beyond black-box methane prediction toward physically consistent and operator-facing decision support for sustainable biogas systems.

1. Research Motivation

Anaerobic digestion (AD) is a practical route for organic-waste treatment and renewable methane generation. However, methane yield is sensitive to process variables such as temperature, pH, organic loading rate (OLR), hydraulic retention time (HRT), total solids (TS), volatile solids (VS), C/N ratio, moisture, and prior methane production. These variables interact in nonlinear ways and often affect methane production after a delay. This creates a strong case for a digital twin that can learn from data, respect physical limits, and support operator decisions.

Mechanistic models such as ADM1 are scientifically strong but complex to calibrate and operate in real time. Pure machine-learning models can be efficient but may produce infeasible predictions outside the training region. GFIS addresses this gap by combining data-driven learning with a physics-guided methane feasibility layer and a soft-sensor stability module.

2. Research Objectives

The dissertation work is structured around the following objectives:

1. Build an anaerobic-digestion data pipeline with synthetic and extensible real-data support.
2. Engineer process, lag, rolling, interaction, and methane-history features.
3. Train tabular and time-series methane-yield prediction models.
4. Add a physics-guided upper-bound layer based on volatile solids.
5. Develop a VFA/ALK soft sensor and stability classifier.
6. Evaluate models using R2, RMSE, MAE, and physics violation count.
7. Implement what-if simulation, 48-hour plant traces, and constrained optimization.
8. Expose the system through FastAPI, Streamlit, and browser-based digital-twin interfaces.
9. Produce dissertation-ready evidence, reports, architecture notes, and demonstration artifacts.

3. Literature Foundation

ADM1 is used as the mechanistic reference point for understanding AD stages such as hydrolysis, acidogenesis, acetogenesis, and methanogenesis. GFIS does not solve the full ADM1 differential-algebraic system; instead, it uses ADM1-informed process reasoning to define variables, synthetic data behavior, scenario design, and temporal modeling requirements.

VFA/ALK ratio is used as a practical early-warning stability indicator. Literature commonly treats low values as stable, a region around 0.3 to 0.4 as disturbance-prone, and higher values as serious acidification risk. GFIS therefore classifies process status as:

Stability class	VFA/ALK range	Interpretation
Stable	< 0.30	Healthy operating region
Warning	0.30 to < 0.45	Disturbance or overload risk
Critical	$>= 0.45$	Acidification or process failure risk

Machine-learning literature in anaerobic digestion supports the use of tree ensembles, boosting models, neural networks, LSTM/RNN models, explainability, and optimization. GFIS uses XGBoost for nonlinear tabular relationships and LSTM for temporal response modeling.

4. System Architecture

GFIS is organized as a layered digital twin:

Layer	Implementation
Data layer	Synthetic AD dataset with methane yield, VFA/ALK, and stability labels
Preprocessing layer	Missing-value repair, scaling, lag features, rolling features, LSTM windows
AI layer	XGBoost tabular model, PyTorch LSTM model, ensemble prediction
Physics layer	Volatile-solids-based methane upper bound and violation count
Soft-sensor layer	VFA/ALK prediction and Stable/Warning/Critical classification
Simulation layer	OLR sweep, 48-hour plant trace, stress experiments, optimization
Interface layer	FastAPI, Streamlit, animated HTML control room, industrial simulation page
Persistence layer	SQLite logging for experiments, predictions, and simulations
Deployment layer	Ubuntu VM notes, Docker scaffold, PostgreSQL-ready direction, ROS2/Gazebo blueprint

5. Dataset and Feature Engineering

The current prototype uses a synthetic anaerobic-digestion dataset designed around literature-informed process behavior. It contains process variables such as temperature, pH, OLR, HRT, TS, VS, C/N ratio, ambient temperature, moisture, methane yield, VFA/ALK ratio, and stability label.

Feature engineering includes:

- Lag features for process variables.
- Rolling mean features.
- Methane-yield history features.
- OLR/HRT ratio.
- VS/TS ratio.
- Temperature-pH interaction.
- Scaled tabular matrices.
- LSTM-ready time windows.

The synthetic-data approach allows controlled experimentation and complete software validation. The final dissertation should state this limitation clearly and position GFIS as a deployable prototype that can be recalibrated with plant or public laboratory data.

6. Model Development

The model stack includes:

- **XGBoost regressor:** Learns nonlinear tabular relationships among process variables.
- **PyTorch LSTM:** Learns sequential methane-yield patterns from time windows.
- **Validation-weighted ensemble:** Combines LSTM and tabular predictions using validation performance.
- **Physics-guided ensemble:** Applies the volatile-solids feasibility bound.
- **Soft sensor:** Estimates VFA/ALK ratio and stability condition.

The current validation selected a zero LSTM ensemble weight because the tabular model performed strongest on the present synthetic dataset. The LSTM branch remains research-relevant because a real plant telemetry dataset will likely contain stronger temporal dependencies than the synthetic tabular benchmark.

7. Evaluation Results

The current trained build used 2400 samples and 24 LSTM epochs.

Model	R2	RMSE	MAE
LSTM	0.2942	11.5712	9.0685
XGBRegressor	0.7556	6.7933	5.5517
Validation-weighted ensemble	0.7554	6.8114	5.5699
Physics-guided ensemble	0.7554	6.8114	5.5699

Physics consistency:

- Physics violation count: 0
- Physics-bound logic: predictions above feasible methane yield are clipped and counted
- Current result: no infeasible predictions in the evaluated test set

8. Explainability

Feature-importance analysis shows that methane history and volatile solids are major contributors to methane prediction.

Top features:

Rank	Feature	Importance
1	methane_yield_lag1	0.3551
2	methane_yield_roll3	0.1198
3	VS	0.0729
4	pH	0.0694
5	temperature_pH_interaction	0.0474

This is technically meaningful: methane history captures process inertia, VS represents biodegradable organic matter, pH affects microbial stability, and temperature-pH interaction reflects coupled biological sensitivity.

9. Research Scenario Experiments

GFIS now includes controlled 48-hour research stress scenarios. These experiments move the work beyond one-point prediction and demonstrate digital-twin behavior under different plant conditions.

Scenario	Mean methane yield	Max VFA/ALK	Final stability	Warning/Critical hours	Physics violation hours
Nominal	199.342	0.197	Stable	0	0
Overload	200.932	0.335	Warning	48	0
pH shock	152.935	0.284	Stable	0	0
Low temp	183.397	0.227	Stable	0	0
Low HRT	194.579	0.278	Stable	0	0
High TS/VS	213.273	0.251	Stable	0	0
Recovery	196.719	0.152	Stable	0	0

Interpretation:

- Organic overload pushes the soft sensor into Warning state for the full 48-hour horizon.
- pH shock and low-temperature disturbance reduce methane yield while remaining physically feasible.
- High-solids feed increases methane potential while still respecting the physics layer.
- Corrective recovery demonstrates operator action through reduced OLR, higher HRT, and near-neutral pH.

10. Working Software Prototype

The implementation is not only a notebook. It includes deployable software surfaces:

- FastAPI backend with prediction, simulation, training, evaluation, optimization, soft-sensor, and stress-test endpoints.
- Streamlit dashboard for model interaction.
- HTML model control room connected to the API.
- Industrial 48-hour simulation page for viva explanation.
- SQLite database logging.
- Docker and Ubuntu VM deployment notes.

Current local demo URLs:

- API root: `http://127.0.0.1:8010`
- API docs: `http://127.0.0.1:8010/docs`
- Model control room: `http://127.0.0.1:8520/index.html`
- Industrial 48-hour simulation: `http://127.0.0.1:8520/industrial_simulation.html`
- Streamlit dashboard: `http://127.0.0.1:8510`

11. Dissertation Contribution

GFIS contributes a practical research prototype with the following strengths:

1. Combines machine learning with physical feasibility checking.
2. Uses both tabular and temporal methane-prediction models.
3. Adds a soft sensor for VFA/ALK-based stability monitoring.
4. Provides what-if simulation and constrained optimization.
5. Generates repeatable research evidence through stress experiments.
6. Exposes the model through API, dashboard, and browser-based interfaces.
7. Provides a deployment path for future plant data integration.

12. Limitations and Future Work

Current limitations:

- The system is validated on synthetic/literature-inspired data, not live plant data.
- The LSTM branch needs richer real time-series telemetry to show its full value.
- SHAP plots are prepared as a pipeline goal; the current build uses model-native feature importance fallback.
- Gasification is implemented as a demonstration/extension concept, not as the core validated AD model.

Future work:

- Integrate a public or plant-specific anaerobic-digestion dataset.
- Add uncertainty estimation and prediction intervals.

- Add Random Forest, SVR, Linear Regression, and neural-network baselines.
- Add richer acidification dynamics for Critical-state stress testing.
- Export dashboard screenshots for the final dissertation.
- Build a PostgreSQL deployment path and connect to IoT/SCADA telemetry.

13. Conclusion

GFIS demonstrates a credible dissertation-scale AI system for renewable-energy process intelligence. The work combines data generation, ML modeling, physics guidance, soft sensing, simulation, optimization, deployment planning, and live interfaces. The current results show that the XGBoost component is the strongest predictor on the synthetic dataset, while the physics layer enforces feasibility and the soft sensor adds operational meaning. The stress-test experiments and industrial simulation strengthen the project as a 16-week dissertation by showing not only code execution, but a research workflow that can be defended, extended, and connected to real anaerobic-digestion plant data.

References

1. Batstone, D. J., Keller, J., Angelidaki, I., Kalyuzhnyi, S. V., Pavlostathis, S. G., Rozzi, A., Sanders, W. T. M., Siegrist, H., and Vavilin, V. A. (2002). The IWA Anaerobic Digestion Model No. 1 (ADM1). Water Science and Technology. DOI: `10.2166/wst.2002.0292`.
2. Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD.
3. Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.
4. Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS.
5. Recent review literature on machine-learning solutions in anaerobic digestion, including tree-based models, boosting, SVM, neural networks, LSTM/RNN methods, explainability, and optimization.
6. Literature on VFA/alkalinity thresholds for anaerobic digestion stability and Buswell-style theoretical methane potential estimation.