

# **Green Fuel Intelligence System (GFIS): A Physics-Guided Hybrid AI Digital Twin for Methane Yield Prediction, Soft Sensing and Stability-Aware Simulation in Anaerobic Digestion**

AIMLCZG628T: Dissertation

by

**Akash Shivadas Chatake**  
**2024AB05226**

Dissertation work carried out at  
**Shri Siddheshwar Women's Polytechnic, Solapur, Maharashtra**

Submitted in partial fulfilment of **M.Tech. Artificial Intelligence and Machine Learning** degree programme

Under the Supervision of

**Dr. Dipti Jadhav**

**Ramrao Adik Institute of Technology, D Y Patil Deemed to be University,  
Nerul, Navi Mumbai, Maharashtra**



**Birla Institute of Technology & Science, Pilani, Vidya  
Vihar, Pilani, Rajasthan – 333031**

**June 2026**

## Abstract

This mid-semester progress report presents the status of **Green Fuel Intelligence System (GFIS)**, a physics-guided hybrid artificial intelligence framework for methane-yield prediction, VFA/ALK soft sensing, stability classification and digital-twin-ready simulation in anaerobic digestion. The project was accepted at the abstract stage with an Excellent faculty evaluation. The evaluator specifically encouraged the work to move beyond black-box machine learning toward Physics-Informed Machine Learning (PIML), soft sensing of difficult-to-measure chemical states, and a simulation dashboard that can later mature into a true digital twin when real-time plant data ingestion becomes available.

GFIS addresses a biological engineering problem rather than a simple web prediction task. Anaerobic digestion is affected by temperature, pH, organic loading rate, hydraulic retention time, total solids, volatile solids, C/N ratio, moisture, alkalinity and reactor history. A purely data-driven model can learn useful nonlinear relationships, but it can also produce physically impossible methane-yield estimates when it extrapolates beyond the training region. GFIS therefore combines machine learning with process constraints: it predicts methane yield, audits the prediction against a volatile-solids-based methane feasibility bound, estimates VFA/ALK as a soft-sensor state, and maps the process into Stable, Warning or Critical operational states.

The current prototype includes a structured data pipeline, XGBoost tabular model, PyTorch LSTM branch, validation-weighted ensemble, physics-guided feasibility audit, VFA/ALK soft sensor, FastAPI service layer, Streamlit dashboard, browser control-room interface and 48-hour controlled scenario simulator. The current metrics are preliminary and are based on a synthetic/literature-inspired anaerobic digestion dataset prepared for controlled mid-semester validation. The tabular model gives  $R^2 = 0.7556$ ,  $RMSE = 6.7933$  and  $MAE = 5.5517$ . The physics-guided ensemble records zero methane-bound violations during evaluation. The current system is deliberately described as a simulation dashboard and digital-twin-ready research prototype, not as a fully synchronized industrial digital twin, because continuous real-time IoT/SCADA ingestion from a physical biogas plant remains part of the next phase.



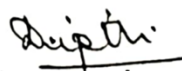
**Signature of the Student:**

---

**Name:** Akash Shivadas Chatake

**Date:** 19/06/2026

**Place:** Solapur



**Signature of the Supervisor:**

---

**Name:** Dr. Dipti Jadhav

**Date:** 19/06/2026

**Place:** Navi Mumbai

## Table Of Contents

|                                                       |    |
|-------------------------------------------------------|----|
| Abstract                                              | 2  |
| List Of Figures                                       | 4  |
| List Of Tables                                        | 4  |
| 1. Introduction                                       | 5  |
| 1.1 Broad Area Of Work                                | 5  |
| 1.2 Background                                        | 5  |
| 1.3 Scope Of Work                                     | 5  |
| 2. Literature Survey                                  | 6  |
| 3. Proposed System And Scientific Baseline            | 6  |
| 3.1 Custom Physics-Informed Loss Function             | 7  |
| 3.2 Interactive Simulation Dashboard And Soft Sensing | 7  |
| 4. Proposed Gfis Methodology                          | 9  |
| 4.1 Data And Feature Pipeline                         | 9  |
| 4.2 Hybrid Ai Model Architecture                      | 9  |
| 4.3 Physics-Guided Loss Function                      | 10 |
| 4.4 Methane Feasibility Constraint                    | 10 |
| 4.5 Vfa/Alk Soft-Sensor Formulation                   | 11 |
| 4.6 Stability Classification Rule                     | 11 |
| 5. Implementation Details                             | 12 |
| 5.1 Software Architecture                             | 12 |
| 5.2 Tools And Platforms                               | 12 |
| 5.3 Demonstration Endpoints                           | 13 |
| 6. Results And Discussion                             | 14 |
| 7. Digital-Twin-Ready Simulation Experiments          | 15 |
| 8. Data Strategy And Open Dataset Plan                | 16 |
| 9. Assumptions, Constraints And Risk Controls         | 17 |
| 10. Future After Mid Semester                         | 17 |
| 11. Conclusion And Future Scope Of Work               | 18 |
| 12. References                                        | 18 |

### **List of Figures**

Figure 1: GFIS Industrial Simulation Dashboard and Control Room UI

Figure 2. GFIS physics-guided hybrid AI research pipeline.

Figure 3. GFIS stability decision path.

Figure 4. Model comparison on the current anaerobic digestion dataset.

Figure 5. 48-hour digital-twin stress scenario summary.

### **List of Tables**

Table 1. Evaluator feedback and GFIS technical response.

Table 2. GFIS model comparison metrics.

Table 3. Stability threshold interpretation.

Table 4. Controlled 48-hour scenario summary.

Table 5. Data strategy for final dissertation phase.

# **1. Introduction**

## **1.1 Broad Area of Work**

The broad area of this dissertation is artificial intelligence for renewable-energy process intelligence, with a specific focus on anaerobic digestion and methane-rich biogas production. The work combines machine learning, process monitoring, soft sensing, physics-guided prediction and simulation-based decision support. It is positioned in the Machine Learning domain, but it is not restricted to conventional regression because anaerobic digestion is a biological and biochemical process with hard feasibility boundaries.

The project is also aligned with the wider objective of applying AI to sustainability and waste-to-energy systems. Small and medium biogas plants often do not have complete laboratory instrumentation or advanced SCADA infrastructure. They may measure temperature, pH, feed quantity and basic solids parameters, but delayed or missing chemical measurements make instability difficult to detect early. GFIS is designed as a practical bridge between measurable process data and decision support for methane yield and reactor health.

## **1.2 Background**

Anaerobic digestion converts organic substrates into biogas through hydrolysis, acidogenesis, acetogenesis and methanogenesis. Methane production depends on substrate composition, temperature regime, pH, organic loading rate (OLR), hydraulic retention time (HRT), total solids (TS), volatile solids (VS), C/N ratio, moisture, alkalinity and past reactor state. A machine-learning model may learn correlations from these variables, but it does not automatically understand that methane output cannot exceed substrate-driven biochemical feasibility limits.

This creates a research gap. Prediction accuracy alone is not sufficient for a process where unsafe recommendations may disturb reactor stability. A strong system should also answer whether a prediction is physically plausible, whether acid accumulation risk is increasing, and whether the operating condition should be treated as normal, warning or critical. GFIS addresses this gap by combining methane-yield modelling with a physics feasibility audit and VFA/ALK soft sensing.

## **1.3 Scope of Work**

The current mid-semester scope is a working research prototype. It includes dataset preparation, feature engineering, methane-yield prediction, preliminary LSTM time-series modelling, physics-bound checking, VFA/ALK soft sensing, stability classification, controlled scenario simulation and local demonstration through API/dashboard interfaces. The current validation uses synthetic/literature-inspired anaerobic digestion data for controlled methodological testing.

The scope does not yet include full industrial deployment, calibrated plant-scale validation or continuous IoT/SCADA synchronization. These are planned as future work. For this reason, this report uses the term “digital-twin-ready simulation dashboard” rather than claiming a complete industrial digital twin.

## 2. Literature Survey

Mechanistic modelling of anaerobic digestion is commonly associated with the Anaerobic Digestion Model No. 1 (ADM1), which remains an important scientific reference for biochemical states and kinetic interactions. ADM1-style modelling is valuable because it represents digestion mechanisms explicitly. However, it has a high parameter burden, requires careful calibration, and can be difficult to deploy directly for lightweight real-time monitoring in small plants. GFIS therefore uses mechanistic digestion science as a constraint source and modelling inspiration, while implementing a practical machine-learning surrogate for mid-semester demonstration.

Machine learning has been widely explored for methane-yield prediction because anaerobic digestion is nonlinear and affected by multiple interacting variables. Tree-based models are often strong for tabular process variables because they capture nonlinear feature interactions without requiring heavy feature transformations. Recurrent neural networks such as LSTM become more relevant when sequential histories are available, because digestion response is delayed by retention time and biological adaptation. The current GFIS results follow this pattern: XGBoost performs best on the present tabular dataset, while the LSTM branch is retained for future continuous time-series data.

Physics-informed machine learning provides a middle path between pure mechanistic modelling and unconstrained black-box prediction. In PIML, domain priors can be introduced through losses, constraints, architectures or post-prediction audits. For GFIS, the first physics-informed step is a volatile-solids-based methane upper-bound audit and a penalty for impossible predictions. This is appropriate for mid-semester because it turns process knowledge into an explicit, inspectable component of the learning workflow.

Soft sensors are important in biological process control. Variables such as VFA/ALK ratio may be expensive, delayed or unavailable in small biogas plants, whereas temperature, pH, feed loading and solids measurements are easier to collect. A soft sensor estimates a difficult-to-measure state using easier online variables. GFIS applies this principle to estimate VFA/ALK and convert it into Stable, Warning or Critical process states.

Digital-twin terminology must be used carefully. A true industrial digital twin requires continuous synchronization between a physical asset and its digital model. The present GFIS system is a digital-twin-ready simulation dashboard. It provides the software architecture, model layer and scenario engine needed for a future IoT-enabled twin, but it is not yet a full industrial twin because real-time physical plant synchronization is not implemented.

## 3. PROPOSED SYSTEM AND SCIENTIFIC BASELINE

The GFIS architecture operates as a digital-twin-ready simulation dashboard. Rather than relying on simple web-based regression, the core system is powered by a Physics-Guided Neural Network (PGNN) that is mathematically constrained by domain-specific biological constraints.

### 3.1 Custom Physics-Informed Loss Function

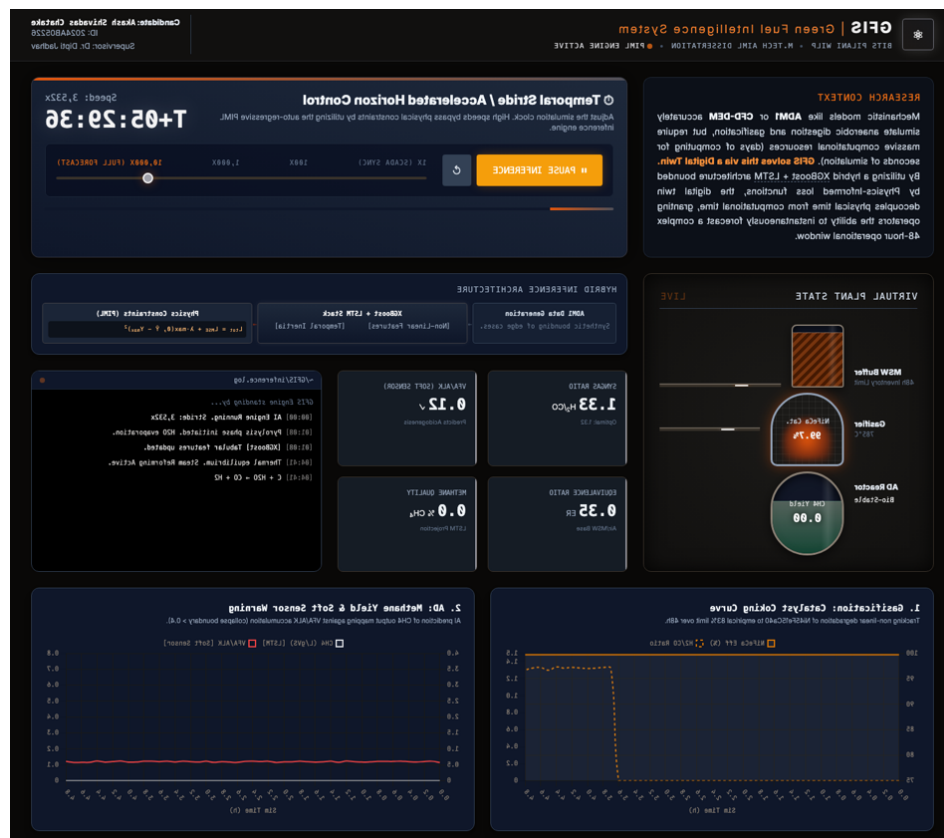
The central scientific claim of the GFIS framework is its departure from standard Mean Squared Error (MSE) optimization. The model employs a composite loss function that actively penalizes the network for predicting biophysical impossibilities. The global optimization objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{data} + \lambda_1 \mathcal{L}_{physics\_bounds} + \lambda_2 \mathcal{L}_{VFA\_constraint}$$

Where  $\mathcal{L}_{data}$  represents the standard empirical supervised error mapping inputs to target methane yields. The regularization terms  $\mathcal{L}_{physics\_bounds}$  and  $\mathcal{L}_{VFA\_constraint}$  establish explicit biophysical penalty boundaries. These constraints actively penalize the network for predicting states such as negative volatile fatty acid concentrations, thermodynamically impossible methanogenesis rates, or violations of volatile-solids-based methane feasibility thresholds.

### 3.2 Interactive Simulation Dashboard and Soft Sensing

At the current stage of the dissertation, GFIS has been locally implemented as an interactive, what-if simulation control room utilizing a FastAPI backend and a sophisticated web frontend. This dashboard allows operators to synthetically manipulate input variables to conduct stability-aware what-if analysis. Concurrently, the integrated VFA/ALK soft sensor evaluates the operational telemetry to classify the biological stability of the digester into distinct actionable states: Stable, Warning, or Critical (sour condition).



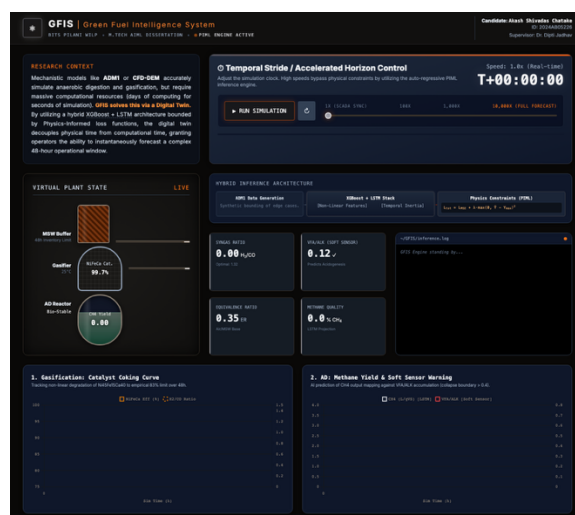
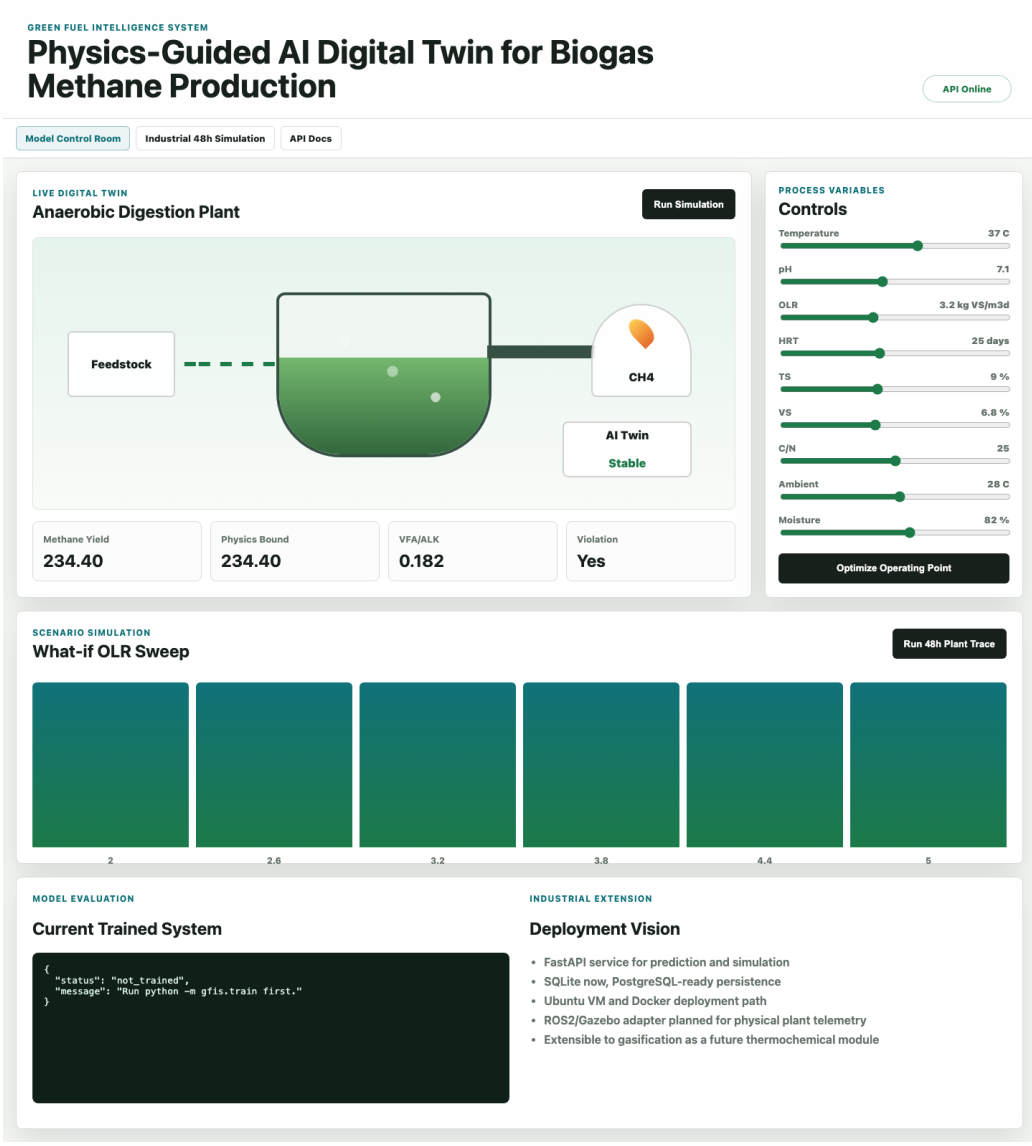
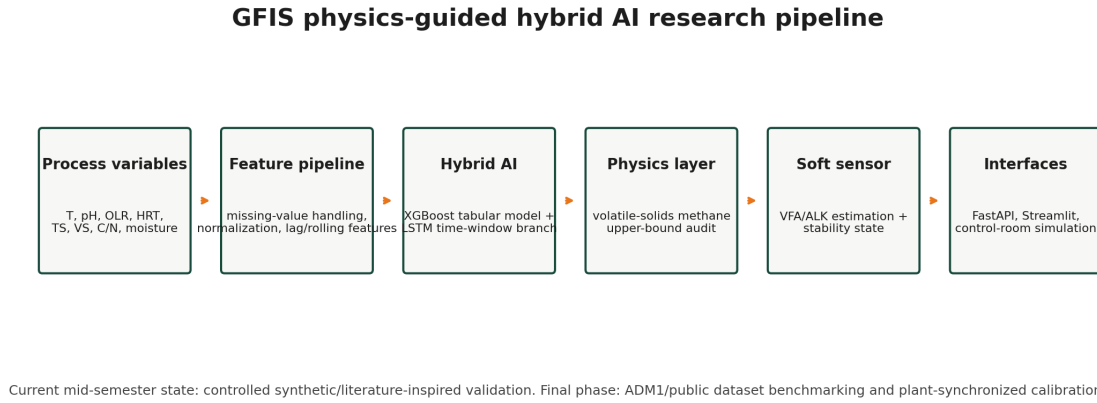


Figure 1: GFIS Industrial Simulation Dashboard and Control Room UI



## 4. Proposed GFIS Methodology

Figure 1 shows the current research pipeline from process measurements to AI prediction, physics checking, soft sensing and operator-facing interfaces.



*Figure 2. GFIS physics-guided hybrid AI research pipeline.*

### 4.1 Data and Feature Pipeline

The input vector for GFIS is constructed from measurable or derivable anaerobic digestion variables:

$$x_i = [T_i, pH_i, OLR_i, HRT_i, TS_i, VS_i, C/N_i, moisture_i, ambient_i, lagged\ reactor\ states_i]$$

The current implementation prepares raw and processed synthetic/literature-inspired datasets, handles missing values, normalizes features, creates lag features and rolling means, and produces time windows for the LSTM branch. Once plant data is available, the split strategy must preserve time order to avoid leakage. Synthetic and real samples should not be mixed inside the same final test set because that would inflate apparent generalization.

### 4.2 Hybrid AI Model Architecture

GFIS uses a hybrid model structure. The XGBoost branch learns nonlinear tabular relationships among operating variables. The LSTM branch is retained for sequential plant data, where retention-time effects and historical reactor states can influence methane yield. The ensemble combines model outputs, and the physics-guided layer audits whether predictions remain feasible.

The current strongest model is XGBoost because the mid-semester dataset is primarily tabular. This does not invalidate the LSTM branch. It indicates that LSTM should be re-evaluated when continuous time-series data from a plant, ADM1 benchmark or public dynamic dataset becomes available.

### 4.3 Physics-Guided Loss Function

The central scientific upgrade is to define the GFIS objective explicitly. The full multi-head objective can be written as:

$$L_{GFIS} = L_{CH4} + \lambda_{phys} L_{feas} + \frac{\lambda_{soft} L_{VFA}}{ALK} + \lambda_{stab} L_{stab} + \lambda_{reg} \|\theta\|_2^2$$

The methane prediction loss is:

$$L_{CH4} = \left(1 - N\right) \sum_i (y_i - y_{hat_i})^2$$

The feasibility penalty is:

$$L_{feas} = (1 / N) \sum_i [ \max(0, y_{hat_i} - Y_{max_i})^2 + \max(0, -y_{hat_i})^2 ]$$

The VFA/ALK soft – sensor loss is:

$$L_{VFA/ALK} = \left(1 - N\right) \sum_i (r_i - r_{hat_i})^2$$

The stability classification loss is:

$$L_{stab} = \left(1 - N\right) \sum_i CE(s_i, s_{hat_i})$$

Here,  $y_i$  is measured methane yield,  $y_{hat_i}$  is predicted methane yield,  $Y_{max_i}$  is the methane feasibility bound,  $r_i$  is the observed VFA/ALK ratio when available,  $r_{hat_i}$  is the estimated soft-sensor ratio, and  $s_i$  is the stability class. If current training uses separate model heads, the methane model can use the reduced objective  $L_{CH4} + \lambda_{phys} L_{feas} + \lambda_{reg} \|\theta\|_2^2$ , while soft-sensor and stability objectives are trained separately.

### 4.4 Methane Feasibility Constraint

The current physics constraint is a volatile-solids-based upper bound:

$$Y_{max_i} = VS_i \times BMP_{feedstock}(i) \times \eta_T(T_i)$$

Here,  $VS_i$  is volatile solids,  $BMP_{feedstock}(i)$  is the biochemical methane potential of the feedstock class, and  $\eta_T(T_i)$  is a temperature-efficiency factor in the range 0 to 1. This is not claimed as a complete biochemical closure. It is a feasibility audit that prevents methane prediction from exceeding a substrate-driven practical bound.

For final dissertation calibration, the same equation can be extended as:

$$Y_{max\_i} = VS\_i \times BMP\_feedstock(i) \times \eta_{T(T\_i)} \times \eta_{pH(pH\_i)} \times \eta_{HRT(HRT\_i)}$$

This extension is appropriate once real plant or benchmark simulation data is available.

#### 4.5 VFA/ALK Soft-Sensor Formulation

The general soft-sensor form is:

$$r\_hat_i = predicted\_VFA/ALK_i = g\_phi(x_i)$$

For interpretability, GFIS can also use the following engineering surrogate:

$$\begin{aligned} r\_hat\_i = & \exp(\beta_0 + \beta_1 OLR\_i + \beta_2 (VS\_i / TS\_i) \\ & + \beta_3 HRT_i^{-1} + \beta_4 |T\_i - T_{opt}| \\ & + \beta_5 positive\_part(pH_{opt} - pH\_i)) \end{aligned}$$

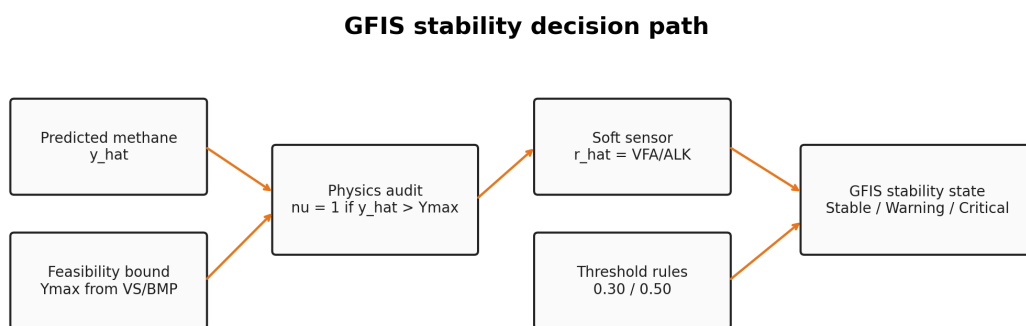
where  $positive\_part(z) = \max(0, z)$ . This equation encodes expected instability drivers: overload, higher volatile fraction, shorter retention time, temperature deviation and acid stress. If online alkalinity becomes available, the upgraded form can subtract a buffering term:

$$r\_hat\_i = \exp(previous\_terms - \beta_6 Alk\_i)$$

This keeps the soft sensor scientifically honest: it is not a universal law of anaerobic digestion, but an interpretable surrogate that can be calibrated as measured VFA/ALK data becomes available.

#### 4.6 Stability Classification Rule

Figure 2 summarizes the GFIS stability decision path. The rule combines predicted methane yield, physics feasibility and soft-sensed VFA/ALK risk.



The rule is intentionally auditable: prediction quality, physics feasibility, and chemical-stability risk are reported together.

Figure 3. GFIS stability decision path.

The physics audit is:

$$nu_i = 1 \text{ if } y_{hat_i} > Y_{max_i}, \text{ otherwise } 0$$

The stability class is defined as:

*Stable* :  $r_{hat_i} < 0.30$  and  $nu_i = 0$

*Warning* :  $0.30 \leq r_{hat_i} < 0.50$  or  $nu_i = 1$

*Critical* :  $r_{hat_i} \geq 0.50$  or repeated feasibility violations persist

These thresholds are GFIS control thresholds for the current prototype, not universal constants. They must be recalibrated when measured plant data is available.

| Predicted VFA/ALK ratio    | GFIS state | Interpretation                        | Suggested action                                 |
|----------------------------|------------|---------------------------------------|--------------------------------------------------|
| $r_{hat} < 0.30$           | Stable     | Buffering dominates acid accumulation | Continue normal operation                        |
| $0.30 \leq r_{hat} < 0.50$ | Warning    | Early overload or acidification risk  | Check OLR, pH, temperature and feed consistency  |
| $r_{hat} \geq 0.50$        | Critical   | High instability risk                 | Reduce loading, inspect buffering and hydraulics |

## 5. Implementation Details

### 5.1 Software Architecture

The current GFIS codebase is organized into backend, dashboard, data, models, reports, deployment and documentation modules. The backend exposes prediction, simulation, plant-run, optimization, soft-sensor and evaluation endpoints. The dashboard and browser interfaces provide operator-facing views. SQLite logging is used at prototype stage, while the architecture remains PostgreSQL-ready.

The backend layer is implemented in Python. It loads trained model artifacts, performs feature preparation, executes prediction, applies the physics feasibility audit, estimates VFA/ALK state, and returns a structured response. The service layer is separated from the UI so that the same logic can be used by Streamlit, browser-based control room, API clients or future plant ingestion services.

### 5.2 Tools and Platforms

The implementation uses Python, Pandas, NumPy, Scikit-learn, XGBoost, PyTorch, FastAPI, Streamlit and browser-based HTML/CSS/JavaScript. Docker and AWS App Runner notes are prepared for deployment. ROS2/Gazebo extension notes are retained for future simulation integration if a more detailed plant simulation is required.

The immediate focus is not to introduce unnecessary infrastructure, but to keep the research pipeline reproducible. The current system can be demonstrated locally, and the deployment path is prepared for migration to a cloud-hosted setup.

### 5.3 Demonstration Endpoints

The GFIS prototype has been deployed on AWS using CloudFront for the web portal and AWS App Runner for the Streamlit research dashboard and FastAPI backend. The deployed system can be accessed through the following links:

- GFIS Portal:  
<https://d2c2mw7lwtswdx.cloudfront.net/>
- GFIS Level 2 Research Workbench:  
<https://d2c2mw7lwtswdx.cloudfront.net/level-2/>
- AI Model Control Room:  
<https://d2c2mw7lwtswdx.cloudfront.net/level-2/apps/model-control-room/>
- Industrial Digital Twin Simulation:  
[https://d2c2mw7lwtswdx.cloudfront.net/level-2/apps/model-control-room/industrial\\_simulation.html](https://d2c2mw7lwtswdx.cloudfront.net/level-2/apps/model-control-room/industrial_simulation.html)
- Streamlit Research Dashboard:  
<https://ea4js82me5.us-east-1.awsapprunner.com/>
- GFIS API Documentation:  
<https://y3kgjnwetp.us-east-1.awsapprunner.com/docs>
- GFIS API Health Endpoint:  
<https://y3kgjnwetp.us-east-1.awsapprunner.com/health>

These endpoints demonstrate that the project is a working software system, not only a written proposal. For submission, screenshots and evidence reports should be attached separately or inserted into the final dissertation report after supervisor review.

## 6. Results and Discussion

The current model comparison is shown in Table 2 and Figure 3.

| 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 |

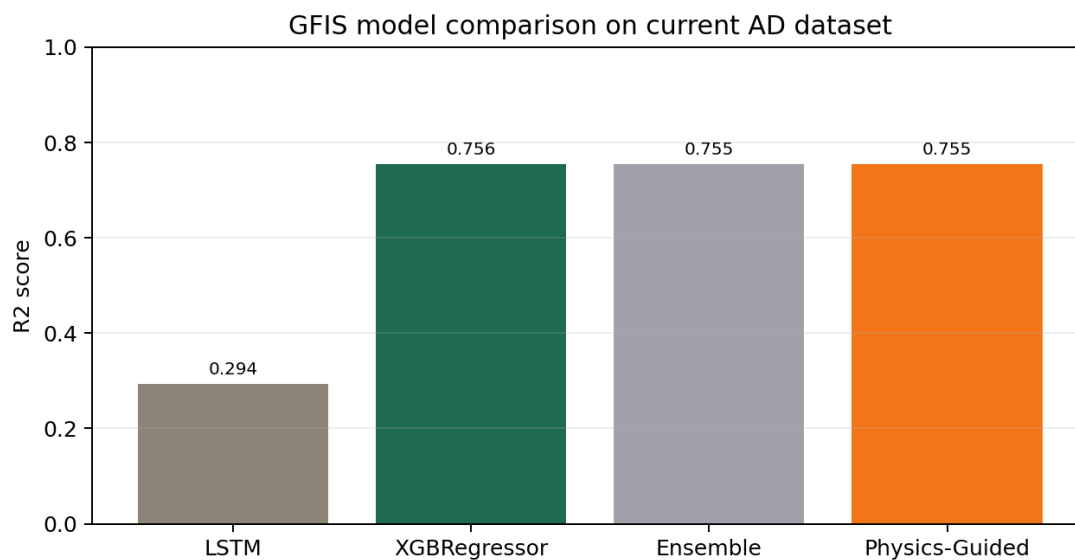


Figure 4. Model comparison on the current anaerobic digestion dataset.

The tabular model currently performs best. The physics-guided ensemble does not materially improve the numerical metric in this experiment because the tested predictions already remain within the volatile-solids feasibility bound. Its value is still important because it introduces process auditability and prevents silent acceptance of impossible methane-yield predictions.

The physics violation count during evaluation is zero. This is a useful early signal that the feasibility layer is working on the current dataset, but it should not be overclaimed as industrial validation. A stronger conclusion will require validation on public anaerobic digestion datasets, ADM1-generated benchmark cases and plant-measured data if available.

The main interpretation at mid semester is therefore:

- XGBoost is the strongest current predictor for tabular AD operating data.
- LSTM requires richer sequential data before its full value can be judged.
- Physics guidance is valuable as a safety and plausibility layer even when numeric metrics remain similar.
- Soft sensing provides the first bridge between prediction and operating-state diagnosis.

## 7. Digital-Twin-Ready Simulation Experiments

GFIS includes controlled 48-hour scenario experiments. These experiments test nominal mesophilic operation, organic overload, pH shock, low-temperature disturbance, reduced HRT washout risk, high-solids feedstock and corrective recovery operation. The purpose is not to claim that the simulator replaces a physical digester. The purpose is to verify that the modelling pipeline reacts sensibly to controlled changes in process variables.

| Scenario                      | Mean methane yield | Max VFA/ALK | Final stability | Warning/Critical hours |
|-------------------------------|--------------------|-------------|-----------------|------------------------|
| Nominal mesophilic operation  | 199.52             | 0.197       | Stable          | 0                      |
| Organic overload              | 201.09             | 0.337       | Warning         | 31                     |
| pH acidification shock        | 152.83             | 0.284       | Warning         | 48                     |
| Low-temperature disturbance   | 183.37             | 0.227       | Warning         | 48                     |
| Reduced HRT washout risk      | 194.78             | 0.277       | Warning         | 48                     |
| High-solids feedstock         | 213.41             | 0.250       | Warning         | 48                     |
| Corrective recovery operation | 196.95             | 0.152       | Stable          | 0                      |

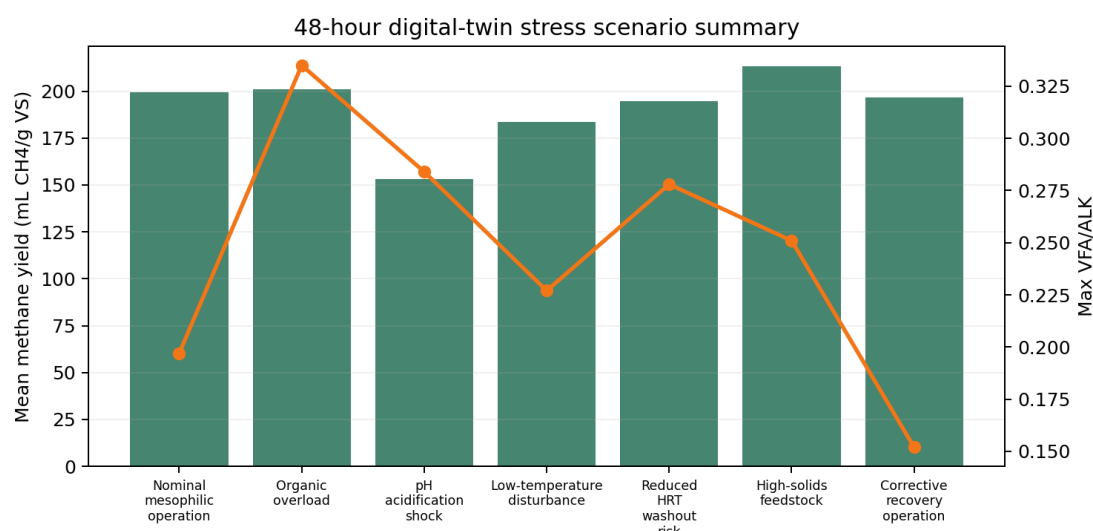


Figure 5. 48-hour digital-twin stress scenario summary.

The scenario results demonstrate the reason for including stability state along with methane yield. Organic overload may not immediately reduce average methane yield, but the VFA/ALK risk increases. Low-temperature and reduced-HRT cases show prolonged warning states even when methane prediction remains within a plausible range. Corrective recovery returns to stable operation with lower VFA/ALK risk. This behavior is consistent with the research direction suggested during abstract evaluation: GFIS should help interpret process stability, not only predict a gas-yield number.

## 8. Data Strategy and Open Dataset Plan

The current system uses synthetic/literature-inspired data for controlled validation. v05 defines a four-layer data strategy for the final dissertation phase.

| Data layer                                    | Purpose                                                        | Status                                        |
|-----------------------------------------------|----------------------------------------------------------------|-----------------------------------------------|
| Controlled prototype data                     | Reproducible development and debugging                         | Completed for mid-semester prototype          |
| ADM1 or reduced mechanistic benchmark         | Stress-test model behavior under known biochemical assumptions | Planned                                       |
| Public empirical anaerobic digestion datasets | External validation where accessible                           | Planned search and curation                   |
| Plant-synchronized data                       | Final movement toward true digital twin                        | Future work, dependent on sensor availability |

The target real-data schema should include timestamp, digester ID, feedstock class, temperature, pH, OLR, HRT, TS, VS, alkalinity, gas flow, methane fraction, VFA if available,



alarm states, actuator events and maintenance notes. If the final dataset includes both synthetic and real data, the final test set should be clearly separated and reported honestly. The dissertation should avoid presenting synthetic-only metrics as industrial validation.

## 9. Assumptions, Constraints and Risk Controls

The current prototype is transparent about its limitations. The dataset is synthetic/literature-inspired and suitable for controlled research demonstration, but final claims must not be overstated. The current version is not claimed as a full industrial digital twin because it is not yet connected to a physical biogas plant or continuous real-time IoT/SCADA stream.

The physics layer is an upper-bound feasibility audit, not a full ADM1 simulator. It checks whether methane prediction exceeds a substrate-driven practical bound, but it does not yet model all biochemical intermediates. The soft-sensor thresholds are initial GFIS control thresholds and require plant calibration. The LSTM result is preliminary because the available data is not yet a strong continuous plant time series.

Risk controls planned for the next phase include:

- Preserve time order in training and testing when time-series data is used.
- Report synthetic, benchmark and real-data results separately.
- Add SHAP or equivalent explainability for the tabular model.
- Calibrate VFA/ALK thresholds when measured VFA/ALK or alkalinity data becomes available.
- Keep “digital twin” wording conservative until a live plant data loop is implemented.

## 10. Future after Mid Semester

The planned work after mid semester is:

1. Improve dataset quality by adding public empirical data and benchmark simulation cases.
2. Extend the physics equations and cite assumptions clearly.
3. Improve the LSTM branch using realistic time windows and time-aware validation.
4. Calibrate the VFA/ALK soft sensor using measured or literature-supported values.
5. Add SHAP explainability plots and interpretation for the XGBoost branch.
6. Improve the dashboard with clearer plant-state telemetry, alerts and saved reports.
7. Prepare deployment so that the backend, dashboard and documentation portal can be demonstrated remotely.
8. Finalize the dissertation report with architecture diagrams, screenshots, model results and limitations.

## 11. Conclusion and Future Scope of Work

GFIS has progressed from an accepted abstract to a working research prototype. The system now demonstrates methane-yield prediction, physics-guided feasibility checking, VFA/ALK soft sensing, stability classification, controlled scenario simulation and local software interfaces. The report strengthens the scientific core by making the loss function, constraints, soft-sensor equations and data strategy explicit while preserving the graphs and evidence from the previous expanded work.

The future scope is to evolve GFIS from a simulation dashboard into a true IoT-enabled digital twin when live plant data becomes available. The final dissertation should prioritize scientific honesty: GFIS is already stronger than a simple prediction web application, but it should claim full industrial digital-twin status only after continuous physical-plant synchronization and real-data validation are implemented.

## 12. 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. Anaerobic Digestion Model No. 1 (ADM1). Water Science and Technology.
2. Appels, L., Baeyens, J., Degreve, J., and Dewil, R. Principles and potential of the anaerobic digestion of waste-activated sludge. Progress in Energy and Combustion Science.
3. Mao, C., Feng, Y., Wang, X., and Ren, G. Review on research achievements of biogas from anaerobic digestion. Renewable and Sustainable Energy Reviews.
4. Chen, Y., Cheng, J. J., and Creamer, K. S. Inhibition of anaerobic digestion process: A review. Bioresource Technology.
5. Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics.
6. Lund, H., and Mathiesen, B. V. Energy system analysis of 100 percent renewable energy systems. Energy.
7. GFIS project evidence reports, local implementation outputs and scenario experiment records prepared in the M.Tech. Semester 4 workspace.