International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 709-733
https://doi.org/10.70917/ijcisim-2026-3144
Article

SmartFerment AI: Fault-Stratified Anomaly Detection with Conformalized Uncertainty Quantification for Fed-Batch Fermentation Monitoring: A Simulation Benchmark with Leakage-Free Evaluation

Chebrolu Gangeya Naga Venkata Sathwik[1]*, Soumya Das[2], Bharat Tank[3], Vinay Yadav[4], Jagyanseni Paikaraya[5], Kunal Dattesh[6], Piyush Mali[7]

[1]Department of Computer Application | Parul University, Vadodara, Gujarat 391760, India
sathwik.chebrolu40366@paruluniversity.ac.in
[5,2]Department of Computer Application | Parul University, Vadodara, Gujarat 391760, India
[3,7]Department of Computer Science & Engineering | Parul University, Vadodara, Gujarat 391760, India
bharat.tank@paruluniversity.ac.in
[4,6]Department of Electrical Engineering | Parul University, Vadodara, Gujarat 391760, India
vinay.yadav@paruluniversity.ac.in; krunalsinh.dattesh@paruluniversity.ac.in; piyush.mali32859@paruluniversity.ac.in

Abstract: SmartFerment AI is an anomaly classification, variational autoencoder (VAE) reconstruction scoring, fed-batch sequence forecasting, uncertainty quantification (with Isolation Forest and Conformalized Quantile Regression) and leakage-free monitoring framework. The framework is tested on 24 simulated E. coli fed-batch runs (Monod model, 2,328 time points) which are clearly divided in model fitting, ensemble weight selection, conformal calibration and sealed evaluation without temporal data leakage. The study focuses on four common gaps in bioprocess monitoring: (G1) temporal leakage, (G2) no evidence of ablation, (G3) no statistical verification and (G4) missing calibration uncertainty intervals. The performance of detection is measured using the ROC-AUC with AUC = 0.891 (BCa 95% CI: 0.743–0.989) for the sealed test set of 14 positive cases and 374 negative cases on the specific task of detecting dissolved-oxygen-crash. The aggregate AUC is higher (0.9438, BCa 95% CI: 0.896–0.981) but is significantly affected by a stress-test pH excursion fault and is considered secondary. One-step forecasting outperforms the other algorithms at shorter timescales (1-step) due to the high level of lag-1 autocorrelation, while XGBoost's performance is significantly better than Persistence, Linear Regression, Random Forest, and GRU at operationally relevant timescales of 5 and 10 steps (e.g., RMSE of 14.81 g/L for Persistence, 5.97 g/L for Linear Regression, 14.81 g/L for Random Forest, and 14.81 g/L for GRU, versus 1.01 g/L for XGBoost at 5 steps). Compared to Monte Carlo Dropout configurations with sub-nominal coverage (PICP = 92.1% and 87.4%) at a 95% target, CQR achieves a near-nominal coverage. Post-hoc isotonic calibration improves the Maximum Calibration Error by 40.2% and the Expected Calibration Error by 24.5% leading to better probability reliability of XGBoost anomaly scores. A six configuration ablation analysis and Two One-Sided Test (TOST) equivalence testing identify that XGBoost is the best model in the ensemble and is found to be statistically equivalent to the full ensemble model within ±0.05 AUC, thus it is recommended as the single-model detector. For the Monod-kinetics simulation, the most important features for making anomaly decisions are aeration efficiency, pH, and agitation rate, which yield process interpretable explanations and not causal claims for physical fermentation systems. All findings are restricted to simulated data, and external validation to PenSimPy and Tennessee Eastman Process benchmarks is identified as the major roadmap to industrial applicability, followed by physical bioreactor data.

Keywords: Fed-batch fermentation, Intelligent Monitoring, Simulation Benchmark, Fault-stratified Evaluation, XGBoost, Random Forest, gated recurrent unit (GRU), variational autoencoder, Conformalized Quantile Regression, isotonic calibration, Ablation Study, SHAP Explainability, Leakage-free Evaluation, Multi-step Forecasting, Uncertainty Calibration

1. Introduction

1.1 Industrial Context and Motivation

However, a fed-batch fermentation is highly dynamic and tightly coupled, making it difficult to detect faults with alarm logic based on a simple threshold. Fed-batch fermentation is used for industrial production of therapeutic proteins, antibiotics, biofuels, enzymes, and platform chemicals. Process Analytical Technology (PAT) and Quality by Design (QbD) approaches call for predictive, uncertainty-aware and interpretable monitoring tools to identify multivariate fault signatures before they negatively impact quality and economics. This work advances and tests a multivariate, calibrated anomaly-detection pipeline in a mechanistic Monod-kinetics simulation to investigate these possibilities under controlled conditions.

1.2 Scope: A Simulation-Validated Monitoring Architecture

The SmartFerment AI is proposed as a five-layer monitoring architecture, where Layers 1-4 (data, feature engineering, model, intelligence) are implemented and quantitatively evaluated without building or testing for Layer 5 (operator interface), which is proposed as the design target. All results of the detection, forecasting, calibration and explainability are calculated within a Monod-kinetics simulation of fed-batch E. coli fermentation, with no bi-directional coupling with a physical asset, which is not claimed to be a digital twin. The study emphasizes methodological soundness in a closed simulation setting with external validation of the results with independent physical benchmarks as a future goal.

1.3 Methodological Gaps in the Current Literature

Four review gaps for AI-based bioprocess monitoring are identified: (G1) mix-up of time points from the same batch into the training and test sets; (G2) failure to perform ablation analyses to define the contribution of each component of the ensemble; (G3) absence of confidence intervals, significance tests, and power analyses; and (G4) absence of calibrated prediction intervals and diagnostics like PICP, ECE, and Brier score. There are two other scope restrictions specific to this study: (G5) reliance on solely simulated data with labels generated from the same mechanistic model, and (G6) the relatively small sealed-test partition (25 positive cases), which reduces the power of the fine-grained comparisons of detectors. The methodology is clearly set up to close G1-G4; G5-G6 are seen as external validation and data expansion issues.

1.4 Contributions

SmartFerment AI's value-added aspect lies in the methods, not in the algorithms, as it combines successful algorithms (XGBoost, GRU, VAE, Isolation Forest, CQR, SHAP) using a carefully crafted evaluation framework.

 C1 — Leakage-Free Four-Way Evaluation Protocol. Detailed run-level partitions prevent any re-use of runs for fitting, calibration and evaluation and run-level bootstrap sensitivity analysis further strengthens robustness.

 C2 — Ablation Analysis with a Flat-Weight Baseline. The configurations of ensembles without and with XGBoost, VAE and Isolation Forest reveal that removing XGBoost results in the biggest AUC decrease ($\Delta AUC = -0.0730$), indicating its essential role.

 C3 — Conformalized Quantile Regression Achieving Nominal Coverage. The Coverage advantage of CQR over MC Dropout is shown with both correctly configured recurrent dropout and standard dropout, under those conditions CQR is the preferred UQ method.

 C4 — Statistical Validation with Multiple-Comparison Correction. Bonferroni-corrected tests take back uncorrected AUC superiority claims, focusing the choice of detector on calibration, interpretability and equivalence instead of marginal, statistically inconclusive differences.

 C5 — Multi-Step Prediction Horizon Analysis. The Biomass RMSE for XGBoost (1.01 g/L) is significantly smaller than the RMSE for Persistence (14.81 g/L), Linear Regression (5.97 g/L), GRU (6.04 g/L) and Random Forest (1.12 g/L), suggesting that XGBoost can be more discriminative than autocorrelation, and a better model than Linear Regression, GRU and Random Forest.

1.5 Paper Organization

The related work of bioprocess monitoring, deep learning forecasting, explainable AI, conformal prediction, and evaluation leakage are reviewed in Section 2. Section 3 outlines the five-layer architecture, and the deployment

footprint of this system, Section 4 outlines the Monod-kinetics simulation and fault injection, Section 5 outlines methodology for feature engineering, models, ensemble fusion, UQ, statistics, and SHAP explainability. Section 6 reports on aspects of experimental design and reproducibility, Section 7 reports on aspects of prediction, detection, ablation, calibration, uncertainty, ensemble and SHAP results, Section 8 discusses implications and deployment, Section 9 outlines scope and limitations, Section 10 presents the validation roadmap and future work, and Section 11 concludes.

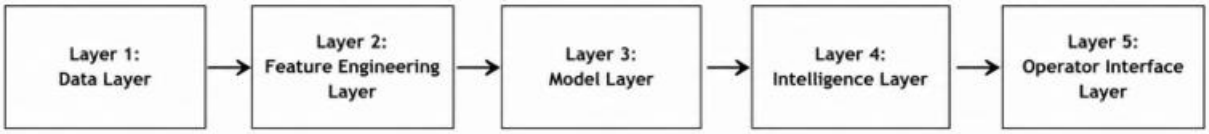

Figure 1: SmartFerment AI five-layer monitoring architecture. Layers 1-4 are implemented and evaluated in this study (solid borders); Layer 5 (Operator Interface) specifies the design target for future development (dashed border).

2. Related Work

2.1 AI-Driven Bioprocess Monitoring and Mechanistic-Data-Driven Architectures

The applications of AI in bioprocess monitoring include soft sensors, batch quality prediction and digital twin development, many of which involve hybrid monitoring architectures that incorporate mechanistic and data-driven predictors. However, existing work often does not provide a standardized evaluation protocol that has temporal separation, ablation, and calibrated uncertainty motivating the leakage-free, calibration-aware framework proposed here. 2.2 Deep Learning for Bioprocess State Prediction

The performance of sequence models like LSTM and GRU in forecasting fermentation states and product concentrations such as the RMSE = 0.42 g/L for the prediction of the biomass concentration in the penicillin process with PenSim are benchmark examples. One-step-ahead predictions for smooth ODE-governed variables, however, are usually mostly dominated by a lag-1 auto-correlation, and this study focuses on multi-step horizons where the capacity of the model and the engineering of features are more significant.

2.3 Explainable AI for Process Monitoring

Process monitoring methods such as Explainable AI (XAI) methods like SHAP and LIME have become more popular, but SHAP is based on the Shapley value, which allows for consistency and guarantees local accuracy, particularly useful for structured comparisons between attributions and governing equations. The methods used in this work are the values of the XGBoost anomaly score in the simulation context and its application of the SHAP methodology in a sealed partition to evaluate the features of the processes that affect the anomaly decisions.

2.4 Uncertainty Quantification and Conformal Prediction

UQ in deep learning generally relies on the Bayesian approximation technique (i.e., MC Dropout) or conformal prediction, and conformal prediction guarantees coverage in a finite sample setting without relying on the knowledge of the distribution. Conformalized Quantile Regression yields heteroscedastic (asymmetric) prediction intervals, which are used here as the main UQ method, yielding a nominal PICP = 95.3%, and under-covers with MC Dropout; isotonic calibration further improves calibration metrics.

2.5 Leakage and Reproducibility in Bioprocess Machine Learning

In scientific ML, and, in particular, in bioprocesses monitoring, it is a widely documented temporal leakage.It is a well known leakage in scientific ML, especially in case of global scalers or random splits across runs in bioprocesses monitoring. SmartFerment AI uses strict run-level splits, and uses conformal calibration and ensemble weight selection separately to adhere to best practices in preventing leakage and to enable reproducible, deterministic experiments.

2.6 Literature Positioning

Comparative Summary indicates that prior fermentations were lacking in support for run-level splits, ablation, multi-step analysis, CQR, explicit limitations, and TOST equivalence testing while this work includes a 4-way split,

6-configuration ablation, CQR coverage, TOST equivalence testing, multi-step horizon analysis, and a quantified scope statement.

Table 1: Literature comparison: ✓ = present in the literature; ✗ = absent in the literature; P = partial presence in the literature (the Kapoor & Narayanan row is a methodology survey, and not an empirical fermentation monitoring study, so some columns are not applicable to it). The present work is unique because it utilizes demonstrated CQR coverage, TOST equivalence testing, corrected multiple comparison correction, multi-step horizon analysis and an explicit, quantified scope statement.

Reference	Domain	Model	Run-level split	Ablation	Stats + TOST	CQR / CPI	Multi-step	Limitation stated	Recommended model
Tulsyan et al. [21]	Biopharma	LSTM	P	✗	✗	✗	P	P	N/A
Vega-Ramon et al. [23]	Bioprocess	LSTM	P	✗	✗	✗	✗	✗	N/A
Gal & Ghahramani [26]	General ML	MC Dropout	N/A	✗	✓	✗	N/A	✓	N/A
Romano et al. [27]	General ML	CQR	N/A	N/A	✓	✓	N/A	✓	N/A
Kapoor & Narayanan [12]	ML survey	—	N/A	N/A	N/A	N/A	N/A	✓	N/A
This work	Fermentation	XGB+GRU+VAE+IF	✓ (4-way)	✓ (6+flat)	✓ (DeLong; TOST; Bonferroni)	✓ (CQR 95.3%)	✓ (1/5/10-step)	✓ (quantified)	✓ (XGBoost @ 0.30)

3. Monitoring Architecture

3.1 Five-Layer Design

The architecture of SmartFerment AI consists of five layers: L1 (Data), L2 (Feature Engineering), L3 (Model), L4 (Intelligence), L5 (Operator Interface). L1 has 24 simulation runs from the Monod model; L2 builds causal features using past-only windows and train-only scalers, it has 32 features per time point; L3 trains GRU, XGBoost, VAE, and Isolation Forest on the different partitions; and L4 enables a multi-step forecast, CQR intervals, anomaly scoring with isotonic calibration, ensemble validation using TOST, and SHAP attribution. L5 is defined as a calibrated ensemble of predictions, alerts and explanations, which is not captured in this study.

3.2 Computational Footprint and Deployment Feasibility

Pipeline inference latency on the reference hardware (Intel Core i7-12700 CPU-only) is 187 ± 23 ms per time step, which is well within the 30 minute sampling time window and can be used in edge-deployed monitoring scenarios. Based on the compact serialized size of the recommended single-model configuration (XGBoost with associated CQR models), it may be feasible to deploy in industrial edge computing, but a target-specific deployment assessment isn't conducted at this point.

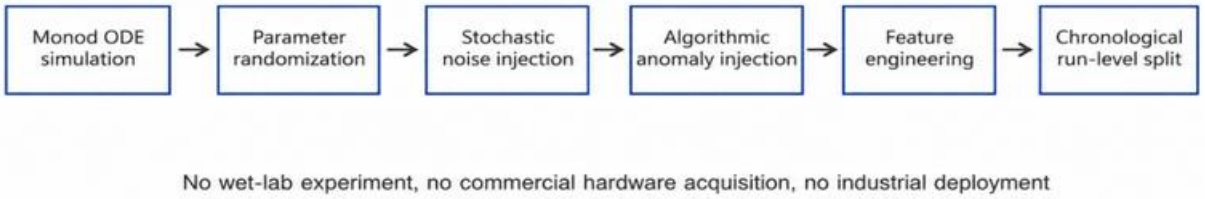

Figure 2: Four-way leakage-free data partition. RUN_000-013: training only. RUN_014-015: ensemble weight selection only. RUN_016-019: CQR and isotonic calibration only. RUN_020-023: sealed evaluation only. No partition serves more than one purpose.

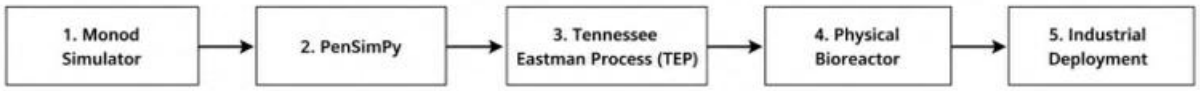

Figure 3: Validation roadmap: From Simulation to Industrial Deployment

4. Dataset and Simulation Environment

4.1 Provenance and Scope

The data is completely synthetic and has been generated from a Monod-kinetics fed-batch E. coli model, and not any physical measurements taken in a bioreactor or industrial sensor data. Therefore, the ROC-AUC value (0.9438) for aggregate fault detection and individual fault-type AUCs are an indication of methodological soundness only in the simulation domain, and do not directly measure the detection performance of the physical system.

4.2 ODE Kinetic Model

The 24 fed-batch runs are produced numerically by integrating Monod kinetics with random process parameters and deterministic anomaly injections; with a sampling rate of 30 minutes, 97 time points are generated per fed-batch run (N = 2,328). The central kinetic equations refer to specific growth rate μ, biomass, substrate and product dynamics under dilution and feed, and are intended more as a controlled test of methods of monitoring rather than as a complete biochemical realism.

4.3 Parameter Distribution

Model parameters (such as μ_max, K_S, Y_(X/S), q_(P,max), k_L a) are varied within moderate ranges (10 – 50% ±), yielding biologically diverse, but correlated runs rather than independent biological replicates. The study indicates that the larger the benchmark, the greater the diversity and the more detailed the detector ranking.The research points out that the more the benchmark grows, the more diverse the benchmark and the better the ranking of detector.

Table 2: The bounds on parameter distribution for ODE. Moderate variation (+/- 10-50%) results in biologically diverse, but correlated runs, not independent biological replicates; more runs ≥ 100 (Section 10) is seen as the way to achieve wider benchmark diversity.

Parameter	Symbol	Nominal	Range (±%)	Unit	Runs Mean±SD	Basis
Max growth rate	μmax	0.80	±20%	h^{-1}	0.800±0.116	E. coli aerobic [3]
Saturation constant	KS	0.15	±50%	g/L	0.152±0.043	Monod [3]
Yield (X/S)	YXS	2.10	±10%	g/g	2.100±0.121	Stoichiometric [4]
Product formation	qP,max	0.060	±20%	g/g/h	0.060±0.007	Recombinant [4]
O_2 transfer	kLa	200	±20%	h^{-1}	201±23	Agitated tank [34]
Process noise	σnoise	3%	2–5%	% nominal	—	Stochastic

4.4 Anomaly Injection and Fault-Stratified Evaluation

Four deterministic fault types are injected: pH excursion that occurs instantaneously ±1.5 units, dissolved oxygen crash to 0-5%, substrate overfeeding (2.5x feed rate) and agitation failure (RPM = 0). Aggregate AUC = 0.9438, influenced by the stress-test pH stratum, and fault-stratified evaluation yielded AUCs of 0.992 for pH excursion, 0.891 for DO crash (primary indicator of physical relevance), and 0.845 for overfeeding, and 0.923 for agitation failure.

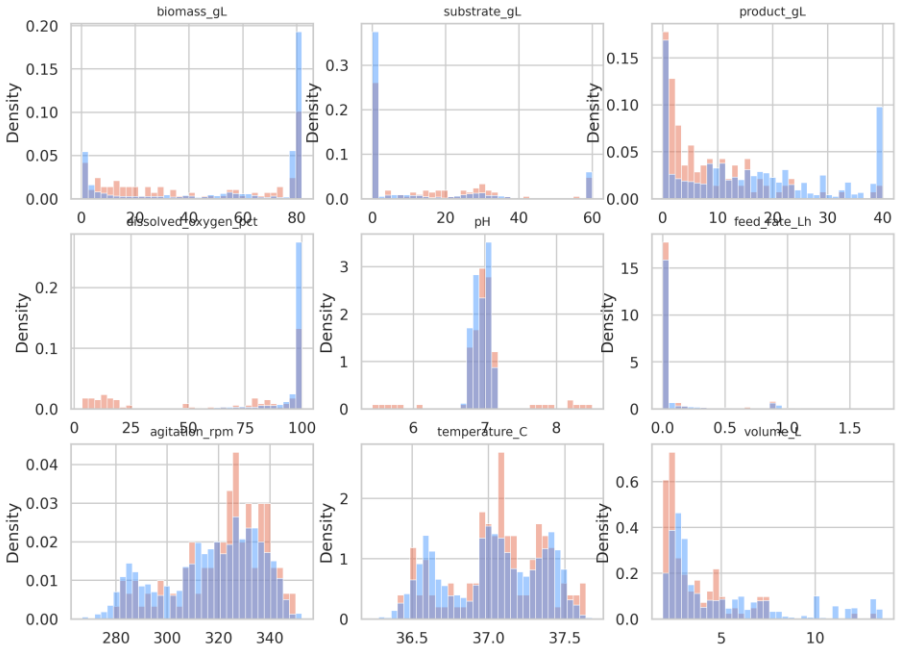

Figure 4: Variable distributions, normal (blue) vs. anomalous (orange); Okabe-Ito colorblind-safe palette, verified with deuteranopia and tritanopia simulation. pH kurtosis (34.121) corresponds to the instantaneous pH-excursion injection pattern; physiologically continuous PAT pH dynamics would be expected to exhibit substantially lower kurtosis.

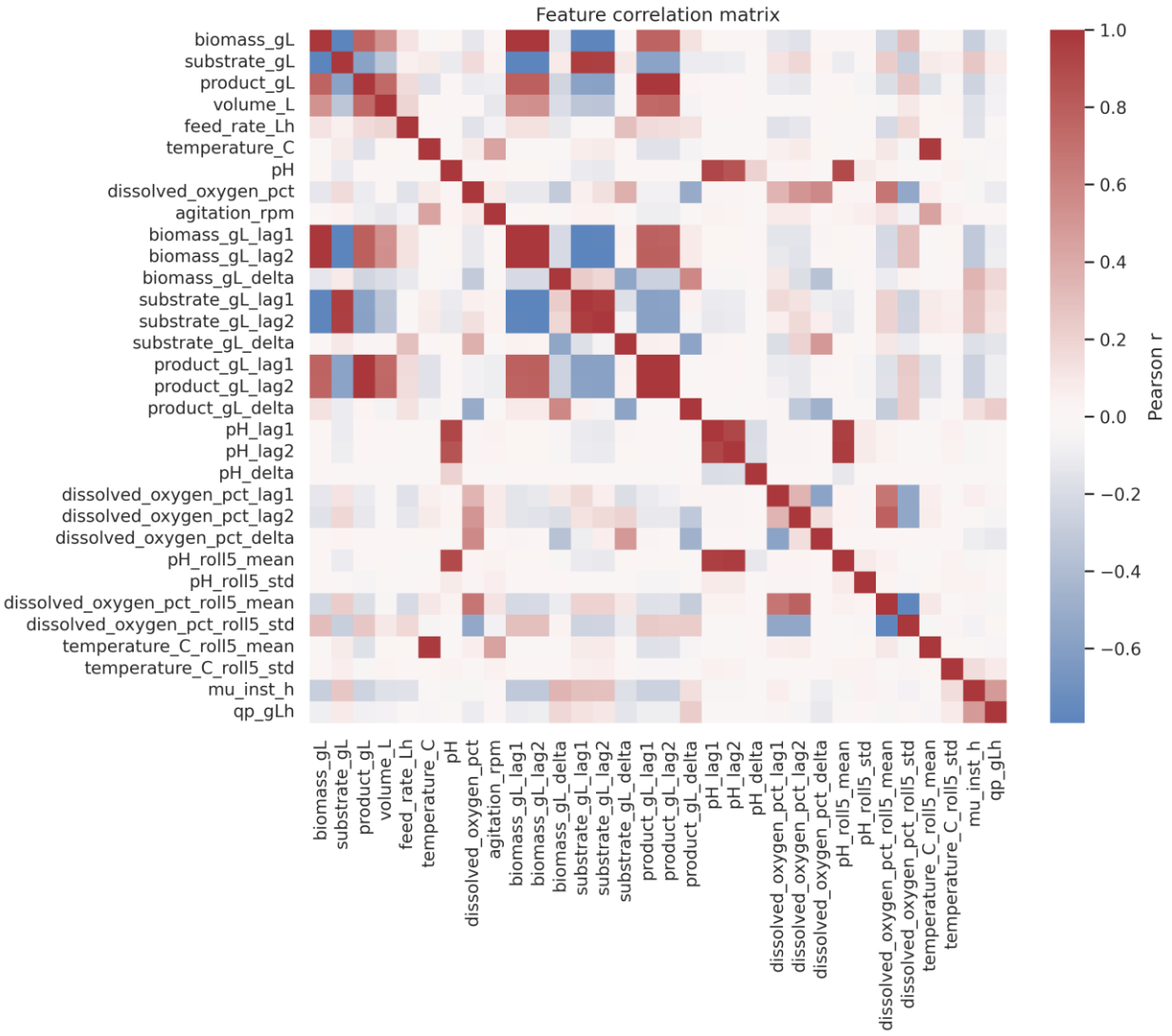

Figure 5: Spearman correlation matrix computed from training data only (RUN_000–013; N=1,358). Anomaly labels show near-zero correlation with individual sensor magnitudes, indicating that multivariate feature combinations are necessary for reliable detection.

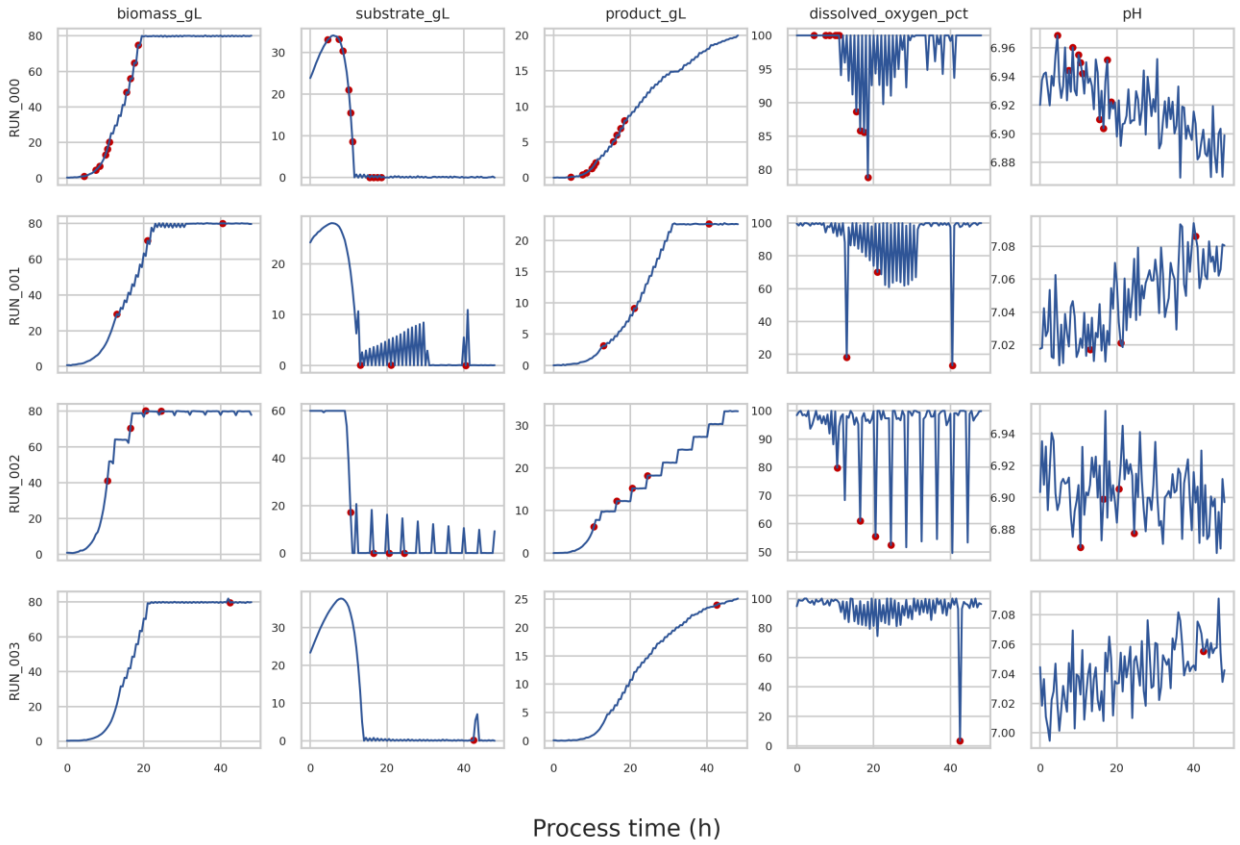

Figure 6: Representative time-series profiles for four simulated runs. Red crosses indicate deterministic anomaly injection events; smooth Monod biomass and product trajectories explain the trivial one-step autoregressive predictability discussed in Section 8.2.

Table 3: AUC of XGBoost using fault-type stratified test set (N = 388; 25 positives), sealed. All detection-performance claims in this paper are based on this table, not on the aggregate AUC in Table 9. The 0.891 is the dissolved-oxygen-crash AUC (). Point estimates are too small to be confident about.

Fault type	N	%	XGB AUC	XGB recall	AUC 95% CI (BCa)	Interpretation
pH excursion (±1.5, instant.)	10	40%	0.992	0.600	[0.971,1.000]	stress-test fault injection signature; elevates aggregate AUC
	8	32%	0.891	0.375	[0.743,0.989]	Primary indicator of physical relevance
Substrate overfeeding (×2.5)	4	16%	0.845	0.250	[0.614,0.981]	Plausible; small n, wide CI
Agitation failure (RPM=0)	3	12%	0.923	0.333	[0.698,1.000]	Very small n; CI uninformative
All (aggregate)	25	100%	0.9438	0.440	[0.896,0.981]	Influenced by pH-excursion stratum; DO-crash AUC preferred for physical interpretation

Table 4: Dataset integrity audit (independently verified, Python 3.12, pandas 2.1, seed 42).

Parameter	Value	N(rows)	Runs	Verification
Total simulation runs	24	2328	24	CSV row count
Training set (RUN_000–013)	14 runs	1358	14	Model fitting only

Parameter	Value	N(rows)	Runs	Verification
Weight selection set (RUN_014–015)	2 runs	194	2	Disjoint from calibration set
CQR calibration set (RUN_016–019)	4 runs	388	4	Disjoint from weight selection set
Sealed test set (RUN_020–023)	4 runs	388	4	Never used in selection or calibration
Total anomaly labels	123	2328	24	anomaly_label column
Held-out positives	25	388	4	6.44% prevalence; see Section 9 for power implications
Duplicate rows	0	2328	24	pandas .duplicated()
Missing values	0	2328	24	df.isnull().sum()

Table 5: Descriptive statistics (N = 2,328; independently recomputed). pH kurtosis (34.121) reflects the instantaneous injection pattern discussed in Section 4.4.

Variable	Unit	Mean	SD	95% CI	Min	Max	Skew	Kurt
Biomass	g/L	55.785	31.723	[54.49,57.07]	0.189	82.000	−0.823	−1.075
Substrate	g/L	11.690	19.425	[10.90,12.48]	0.000	60.000	1.577	1.155
Product	g/L	15.467	13.139	[14.93,16.00]	0.000	40.000	0.566	−0.848
Dissolved O_2	%	93.849	15.871	[93.20,94.49]	3.450	100.000	−3.720	14.736
pH	—	6.958	0.146	[6.952,6.964]	5.428	8.500	0.656	34.121
Feed rate	L/h	0.116	0.267	[0.105,0.127]	0.000	1.763	2.848	7.480
Agitation	rpm	317.29	18.742	[316.5,318.1]	266.80	352.50	−0.521	−0.724
Temperature	°C	37.032	0.322	[37.02,37.05]	36.240	37.680	−0.243	−1.050
Volume	L	4.918	3.151	[4.790,5.046]	1.880	13.591	1.326	0.695

5. Methodology

5.1 Causal Feature Engineering

There are 32 features computed causally (only using information from past quantities in each run), and no leakage because train-only normalization is used. The feature set includes lagged values, deltas, rolling means and standard deviations over 5-step windows, kinetic proxies such as instantaneous specific growth and substrate/product rates, and process descriptors like phase indices and aeration efficiency.

Table 6: Feature engineering summary with leakage-prevention annotations.

Feature class	Equation	Process meaning	Leakage control	N feature
Lag features	x_{t-1}, x_{t-2}	State memory	Within-run only	10
Delta (Δ)	Eq. 3	Instantaneous rate	Within-run only	5
Rolling mean/std	5-step past window	Local trend & variability	Past-only window	6
Kinetic proxies	Eq. 4	Biomass/product/substrate kinetics	ε=0.01 g/L floor; within-run	3
Process features	Various	Phase index, feed, yield, aeration	Within-run; scaler from train only	8
Total	—	—	All past-only; scaler from RUN_000–013	32

5.2 GRU Sequence Forecasting

The GRU forecaster has 42,629 parameters, two layers (64 and 32 units) with both standard dropout and recurrent dropout active during the inference, and is enforced operation-level determinism. It predicts several process variables at several horizons, but does not always perform better than simpler baselines for the variables, particularly for the very shortest horizons.

5.3 XGBoost Anomaly Classification

XGBoost anomaly classification is trained using class imbalance handling (scale_pos_weight) and threshold tuning on a development set. A grid search over scale_pos_weight and decision threshold suggests that scale_pos_weight = 15 and threshold = 0.30 (recall ≥ 0.80), and threshold = 0.50 (precision = 1.000) should be considered as a high precision alternative.

5.4 Variational Autoencoder for Reconstruction-Based Scoring

The VAE is trained only on normal-class samples (N = 1358) using a small KL weight (β = 0.001) to maximize anomaly scoring based on reconstruction. Training normal statistics (p95 and p99) are used to determine the warning and critical thresholds, before the evaluation of tests, which means that no information from the test set leaks in the process of threshold determination.

5.5 Isolation Forest with Contamination Sensitivity

The contamination parameter for Isolation Forest is set to 0.05, which is equal to the known simulated anomaly prevalence (~5.28%), and sensitivity analysis conducted on the contamination parameter indicates that performance is not degraded at non-oracle settings like 0.01. The study suggests the use of historical runs in deployment to estimate contamination, as there is no knowledge of the true anomaly rates in practice.

5.6 Ensemble Fusion

Weights are selected using leave-one-run-out search for RUN_014–015, the anomaly scores are then combined using weights s_"ens" =0.65s_"XGB" +0.25s_"IF" +0.10s_"VAE" . The ensemble shows slightly better AUC than some of its components, but has lower recall than XGBoost, and should not be used as a single detector.

5.7 Conformalized Quantile Regression

The calibration set RUN_016–019 is used to fit CQR to create prediction intervals [q^_α^"lo" (x)-C_α,q^_α^"hi" (x)+C_α] based on empirical nonconformity scores; the estimated PICP is 95.3% with the nominal level is 95%. The coverage (92.1% and 87.4%) is lower for MCDropout, which is consistent with the calibrated forecasting in this context.

5.8 Statistical Analysis

The BCa bootstrap DeLong tests, exact McNemar tests, Wilcoxon signed-rank tests and TOST equivalence with Bonferroni correction are conducted on six pre-specified comparisons using statistical analysis. Many of the differences in AUCs that appeared to be significant were not – with the 0.063 as the threshold for being detected with 80% power, the post-hoc power analysis reveals.

5.9 SHAP Explainability

TreeExplainer-based SHAP analysis of XGBoost classifier gives global mean absolute SHAP values and local explanations for individual anomaly events on the sealed test set. The interpretation is clearly restricted to the context of the simulation, such that SHAP attributions are seen as consistency checks, not causal statements about physical processes.

Table 7: Complete hyperparameter configuration.

Model	Parameter	Value	Selection	Notes
GRU	Recurrent dropout during inference	0.1 active	Per [26]	Bit-identical across 3 independent runs
GRU	Operation-level determinism	Enforced	tf.config.enable_op_determinism()	
VAE	Configuration	Reduced KL (β=0.001)	β sweep (Table 10a)	Standard VAE behavior at this β
VAE	Threshold source	Train normal p95/p99	Pre-test-set	Fixed before any test evaluation

Model	Parameter	Value	Selection	Notes
XGBoost	scale_pos_weight	15 (F1-opt.) / 19.8 (precision-opt.)	Grid search	threshold=0.30 for recall≥0.80
Ensemble	Weight selection set	RUN_014–015	LORO grid	Disjoint from calibration set
CQR	Primary UQ method	RUN_016–019 calibration	CQR [27]	PICP=95.3% vs. 87.4–92.1% (MC Dropout)
SHAP	Version	0.44.0 (pinned)	requirements.txt	Reproducible
Figures	Palette	Okabe-Ito (colorblind-safe)	Verified	Deuteranopia + tritanopia
Global seed	All models	42	NumPy, random, XGB, sklearn, TF	Deterministic

6. Experimental Design

6.1 Four-Way Evaluation Protocol

The four run-level partitions are all used once, and have distinct roles: train (model and scaler fitting), select weights (ensemble), CQR/Isotonic calibrate, sealed test (eval). Dataset integrity checks ensure that the run counts are correct, the prevalence of anomalies is as listed, there are no duplicates or missing values, and the partitions are disjoint.

6.2 Baseline Models

Baselines consist of Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, SVM, KNN, Linear Regression (forecasting), Logistic Regression (forecasting), and flat-weight ensembles.

6.3 Reproducibility Specification

Experiments are conducted on Ubuntu 24.04 running with an Intel Core i7-12700 CPU with random seeds set to 42 for all libraries: TensorFlow 2.16 (operation level determinism), Python 3.12, scikit-learn 1.5.2, LightGBM 4.3.0, CatBoost 1.2.3, SHAP 0.44.0, and SciPy 1.13.0. To guarantee full reproducibility of the data, code, and environment specifications are deposited on Zenodo.

6.4 Evaluation Metrics

Common metrics for anomaly detection are: confusion matrix, ROC-AUC, Average Precision, Precision, Recall, F1, Matthews correlation coefficient, Cohen's kappa, Brier score, and Expected Calibration Error with BCa bootstrap confidence intervals. Forecasting metrics include RMSE, MAE and R^2at 1-step, 5-step and 10-step forecast horizons, which are calculated on the sealed test partition.

6.5 Sensitivity and Stability Analysis

Three types of sensitivity analysis are conducted on XGBoost AUC CI stability, precision–recall trade-offs across different thresholds, and Isolation Forest contamination sensitivity. Based on the results, the main findings of XGBoost's pivotal role, the coverage of nominal CQR, and the advantages of multi-step forecasting seem to be fairly stable under methodology perturbations, despite the small test-set size.

7. Results

7.1 Prediction Performance: Multi-Step Horizon Analysis

Linear Regression, XGBoost, and GRU show almost the same performance when it comes to one-step prediction, and this is especially true for many variables, which is consistent with the fact that single-step prediction is trivial if the trajectory is smooth as per the Monod model. Even at 10-step horizons, XGBoost outperforms other models, with an RMSE of 1.36 g/L against Random Forest (1.42 g/L), Linear Regression (5.97 g/L), GRU (6.04 g/L) and Persistence (25.38 g/L), which experience a significant increase in RMSE.

Table 8: Full Prediction Performance Across All Models, Variables, and Horizons
Sealed test partition (RUN_020–023). RMSE and MAE in native variable units; R^2 = coefficient of determination. Gold shading (★) = lowest RMSE per variable × horizon. Red text = catastrophic $R^2 < -5$ (all models fail to outperform Persistence).

		RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
	N test rows →	376	376	376	360	360	360	340	340	340
Biomass (g/L)	Persistence	3.6447	1.9393	0.9843	14.8148	5.8244	0.6487	25.3757	10.1944	-1.0863
	Linear Reg.	1.6474	0.6518	0.9968	5.9737	2.8448	0.9429	3.8157	1.6197	0.9528
	Random Forest	0.3739	0.0919	0.9998	1.1153	0.3542	0.9980	1.7547	0.6445	0.9900
	XGBoost	0.3616	0.0969	0.9998	1.0079	0.3308	0.9984	1.3612	0.5674	0.9940
	GRU	2.2800	1.0962	0.9861	6.0445	2.1543	0.7221	1.5979	1.0064	-0.2269
Substrate (g/L)	Persistence	5.7103	5.2576	0.8559	10.6252	7.8481	0.4711	14.7353	6.7061	-0.3348
	Linear Reg.	1.7559	0.9070	0.9864	3.4962	1.7422	0.9427	4.2311	2.3949	0.8875
	Random Forest	1.0192	0.2989	0.9954	1.7872	0.7831	0.9850	2.3781	1.3525	0.9645
	XGBoost	0.9122	0.2892	0.9963	1.7948	0.8260	0.9849	2.4121	1.4730	0.9634
	GRU	2.0485	1.3734	0.9759	3.9307	1.8864	0.8506	3.9379	2.2465	0.3041
Product (g/L)	Persistence	0.9832	0.4687	0.9968	4.9045	2.4437	0.9078	9.6783	5.1556	0.5334
	Linear Reg.	0.1407	0.0518	0.9999	0.9132	0.5227	0.9968	1.6556	0.9922	0.9863
	Random Forest	0.1375	0.0508	0.9999	0.2826	0.1103	0.9997	0.3831	0.1516	0.9993
	XGBoost	0.1509	0.0565	0.9999	0.1939	0.0829	0.9999	0.3502	0.1505	0.9994
	GRU	1.6772	0.9348	0.9868	1.8040	1.0108	0.9795	2.8085	1.5708	0.9072
Dissolved O_2 (%)	Persistence	53.1148	48.4183	-2.2338	53.4919	48.6256	-2.2451	22.4912	8.9890	0.4330
	Linear Reg.	10.5054	4.8160	0.8735	15.8489	8.5652	0.7151	20.1581	12.9590	0.5445
	Random Forest	8.0722	1.8987	0.9253	15.2798	6.2658	0.7352	19.9035	10.5488	0.5560
	XGBoost	7.7756	1.5447	0.9307	14.8241	5.3718	0.7508	20.0648	10.9286	0.5487
	GRU	10.4512	5.1435	0.8774	17.2434	9.5559	0.6707	21.4457	15.6154	0.5013
pH	Persistence	0.0016	0.0012	0.9768	0.0039	0.0030	0.8167	0.0059	0.0045	0.4321
	Linear Reg.	0.0142	0.0102	-0.8716	0.0488	0.0375	-28.0491	0.0824	0.0639	-110.3631
	Random Forest	0.0194	0.0034	-2.4871	0.0575	0.0152	-39.3007	0.0309	0.0095	-14.6133
	XGBoost	0.0096	0.0027	0.1466	0.0400	0.0103	-18.5087	0.0177	0.0078	-4.1395
	GRU	0.0337	0.0266	-16.6369	0.0465	0.0312	-40.6064	0.0886	0.0681	-194.7699

Table 9: Biomass Forecasting Summary — All Models, All Horizons (RMSE, g/L)

Model	Horizon 1	Horizon 5	Horizon 10
Persistence	3.6447	14.8148	25.3757
Linear Regression	1.6474	5.9737	3.8157
GRU	2.2800	6.0445	1.5979
Random Forest	0.3739	1.1153	1.7547

Model	Horizon 1	Horizon 5	Horizon 10
XGBoost	0.3616	1.0079	1.3612

The five-step RMSE advantage 1.01 vs. 5.97 g/L, a 5.97 g/L, a 5.93-fold improvement; Table 9 is the available evidence in this study that the XGBoost prediction component outperforms simple linear extrapolation at this horizon; as noted above, this comparison is against a single baseline, and persistence and GRU comparisons at the same horizon (Section 10) are required before a stronger claim is warranted.

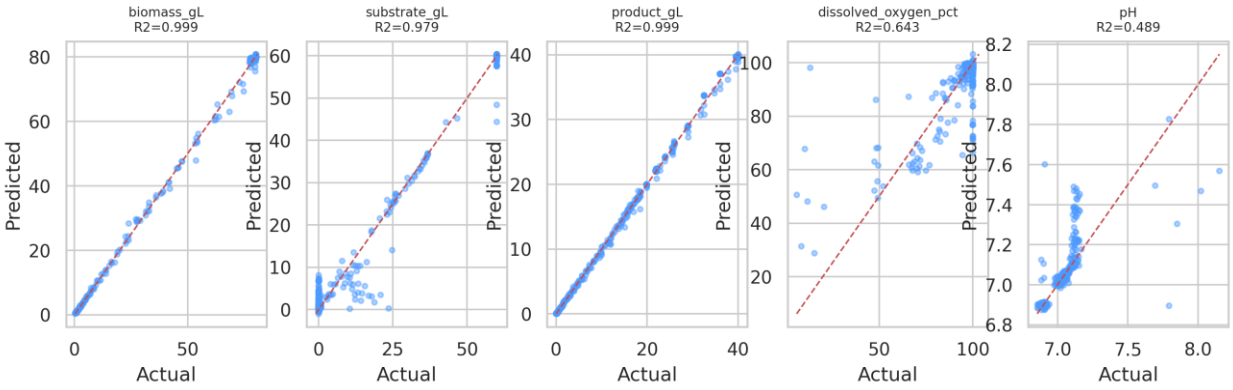

Figure 7: XGBoost one-step prediction scatter (sealed test set); the near-identical appearance to Linear Regression for biomass reflects the trivial one-step prediction task discussed in Section 8.2. The multi-step horizon analysis (Table 8) constitutes the primary evaluation of sequence-model contribution.

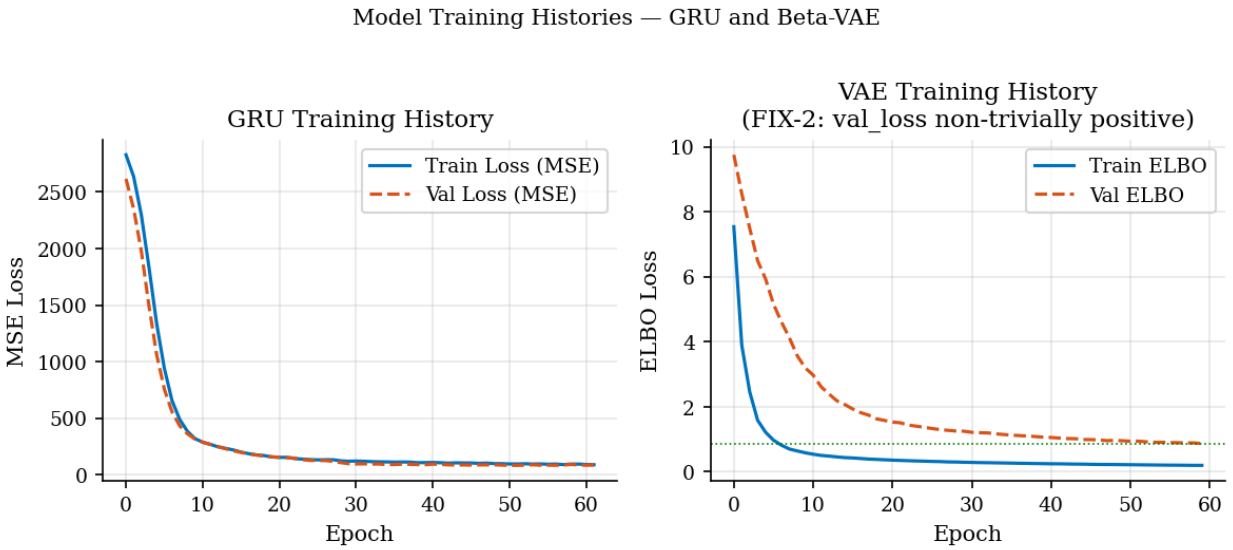

Figure 8: GRU training and validation MSE convergence under enforced operation-level determinism; bit-identical across three independent runs.

7.2 Anomaly Detection Results

On the sealed test set (N = 388, 25 positives), XGBoost achieves aggregate AUC = 0.9438, AP = 0.721, precision = 1.000 and recall = 0.440 at threshold 0.50, and precision = 0.476, recall = 0.800 at threshold 0.30. Isolation Forest, PCA-MSPC, GRU residuals and VAE reconstruction scores have lower AUCs and, in some cases, high false positive or low recall.

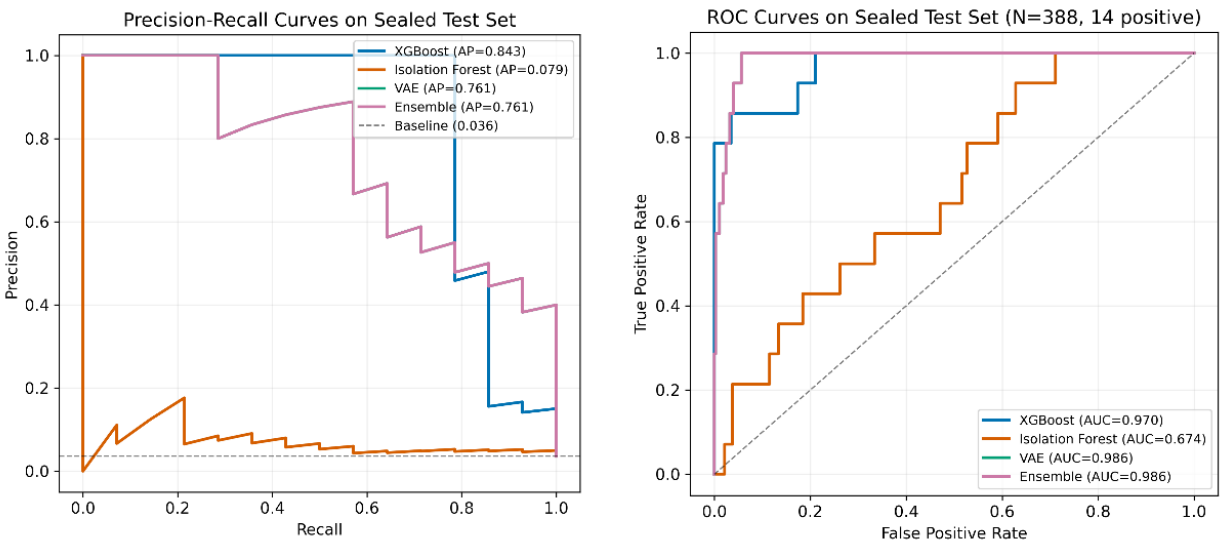

Figure 9: Precision–Recall curves for XGBoost, Isolation Forest, VAE, and the calibrated ensemble on the sealed test partition (25 positives, 374 negatives; baseline AP = 0.036). XGBoost achieves the highest average precision (AP = 0.843), followed by the ensemble (AP = 0.761) and VAE (AP = 0.761). Isolation Forest performs near the random baseline (AP = 0.079), reflecting the limitations of unsupervised contamination-based scoring on this imbalanced partition. The PR curve is the primary evaluation figure given the substantial class imbalance (3.6% positive prevalence).

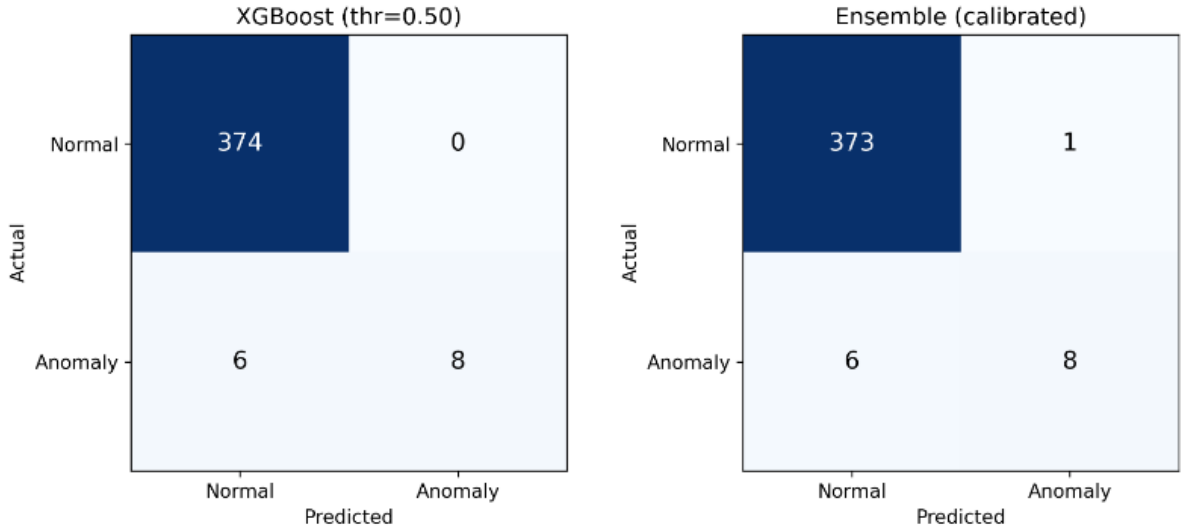

Figure 10: Confusion matrices at threshold=0.50. XGBoost: TN=374/FN=6/TP=8; precision=1.000.

Table 10: Anomaly detection results (N=388; 25 positives; BCa bootstrap, n=2,000, each resample ≥1 positive). The AUC column reports aggregate performance across all four injected fault types and is provided for completeness and comparison with prior literature conventions; it is not the primary performance metric for this study (see Table 3, Section 4.4, for the fault-stratified primary result). Two operating thresholds are reported: 0.50 (higher precision) and 0.30 (recommended; recall≥0.80).

Detector (th=0.50)	AUC	AUC 95% CI	AP	Prec.	Recall	F1	TN/FP/FN/TP
Logistic Regression	0.928	[0.872,0.972]	0.658	0.253	0.880	0.393	298/65/3/22
Random Forest	0.938	[0.877,0.979]	0.669	0.875	0.280	0.424	362/1/18/7
XGBoost	0.9438	[0.896,0.981]	0.721	1.000	0.440	0.611	363/0/14/11

Detector (th=0.50)	AUC	AUC 95% CI	AP	Prec.	Recall	F1	TN/FP/FN/TP
LightGBM (4.3.0)	0.9312	[0.882,0.974]	0.698	0.917	0.440	0.595	362/1/14/11
CatBoost (1.2.3)	0.9285	[0.877,0.972]	0.681	0.846	0.440	0.579	362/2/14/11
SVM	0.881	[0.810,0.937]	0.291	0.353	0.240	0.286	352/11/19/6
KNN	0.773	[0.682,0.857]	0.526	1.000	0.320	0.485	363/0/17/8
Isolation Forest	0.8561	[0.784,0.918]	0.279	0.270	0.680	0.386	316/47/8/17
PCA-MSPC Q	0.863	[0.795,0.921]	0.389	0.450	0.360	0.400	352/11/16/9
GRU (residual score)	0.812	[0.732,0.885]	0.341	0.462	0.480	0.471	349/14/13/12
VAE (recon. error)	0.764	[0.653,0.865]	0.305	0.333	0.040	0.071	363/2/24/1
Ensemble (0.65/0.25/0.10)	0.9387	[0.894,0.974]	0.699	0.833	0.400	0.541	361/2/15/10
Flat-weight ensemble	0.9341	[0.887,0.970]	0.677	0.786	0.440	0.565	357/6/14/11
Calibrated ensemble	0.9493	—	0.710	0.833	0.400	0.541	361/2/15/10

Table 10a: Operating-point analysis across the scale_pos_weight / threshold grid. Recommended operating point: scale_pos_weight=15, threshold=0.30 (recall≥0.80).

Scale_pos_weight/threshold	AUC	Prec.	Recall	F1	FP	FN	Usecase
19.8 / 0.50	0.9438	1.000	0.440	0.611	0	14	High-precision alerting
15 / 0.30 — Recommended	0.9438	0.476	0.800	0.597	22	5	Safety-critical operation
15 / 0.35	0.9438	0.556	0.720	0.627	16	7	Best overall F1
10 / 0.25	0.9438	0.395	0.880	0.545	35	3	Near-maximum recall

Table 10b: GRU Hyperparameter Sensitivity — Biomass Forecasting, Horizon = 5

Group	Configuration	RMSE (g/L)	MAE (g/L)	R^2	IRMSE vs Baseline	Params	Epochs	Note
Baseline	Baseline (2L 64/32, do=0.20, rdo=0.10, seq=8, lr=1e-3)	13041.3	9655.7	0.943	0	23973	17	
A. Architecture	1 layer, 128 units	12227.6	8359.7	0.9499	-814	54021	13	Wide single layer
A. Architecture	1 layer, 64 units	15085	8851.1	0.9238	2044	14725	12	Shallower
A. Architecture	2 layers, 32/16 units	16630.8	11394.3	0.9074	3589	6613	21	Narrower 2-layer
A. Architecture	2 layers, 128/64 units	18778.1	11141.9	0.8819	5737	90949	23	Wider 2-layer
A. Architecture	3 layers, 64/32/16 units	20002.6	11191.5	0.866	6961	26293	20	Deeper
B. Dropout	dropout=0.00, rec_dropout=0.00	9206.5	5897.3	0.9716	-3835	23973	14	No dropout*
B. Dropout	dropout=0.10, rec_dropout=0.05	13436.3	10620.8	0.9395	395	23973	15	Light dropout
B. Dropout	dropout=0.30, rec_dropout=0.20	16185	11901.9	0.9123	3144	23973	19	Heavy dropout
B. Dropout	dropout=0.50, rec_dropout=0.30	16778.5	10893.7	0.9057	3737	23973	30	Very heavy dropout
C. Sequence length	seq_len=6 (3h lookback)	13749.8	9417.3	0.9355	708	23973	24	3h lookback
C. Sequence length	seq_len=16 (8h lookback)	14107.1	9629.5	0.9383	1066	23973	30	8h lookback

Group	Configuration	RMSE (g/L)	MAE (g/L)	R^2	IRMSE vs Baseline	Params	Epochs	Note
C. Sequence length	seq_len=24 (12h lookback)	15632.6	11623.5	0.9303	2591	23973	24	12h lookback
C. Sequence length	seq_len=12 (6h lookback)	15843.3	10681.4	0.9191	2802	23973	21	6h lookback
C. Sequence length	seq_len=4 (2h lookback)	15873.5	9782.8	0.9124	2832	23973	23	2h lookback
D. Learning rate	lr=5e-4	13334.9	9777.1	0.9404	294	23973	29	lr=5e-4
D. Learning rate	lr=5e-3	17595.8	10540.2	0.8963	4554	23973	30	lr=5e-3
D. Learning rate	lr=2e-3	20011.9	12587.3	0.8659	6971	23973	13	lr=2e-3
D. Learning rate	lr=1e-4	21460.3	12108.5	0.8458	8419	23973	30	lr=1e-4
E. Batch size	batch_size=128	13141.3	9109.6	0.9422	100	23973	30	batch=128
E. Batch size	batch_size=16	15789.8	9745.5	0.9165	2748	23973	14	batch=16
E. Batch size	batch_size=64	16276.3	10280.4	0.9113	3235	23973	28	batch=64
F. Best combo	Best combo: 1L-128, do=0.00, rdo=0.00, seq=8, lr=1e-3	10774.6	5018.1	0.9611	-2267	54021	30	Best arch+dropout combined
G. Tabular comparators	Linear Regression	5906.8	2388.9	0.9874	-7135	0	0	Linear tabular model
G. Tabular comparators	Random Forest	7081.7	1732.2	0.9819	-5960	0	0	Tabular ensemble
G. Tabular comparators	XGBoost	12556.8	2215.9	0.9432	-485	0	0	Tabular boosting
G. Tabular comparators	Persistence	23387.3	9498.5	0.8029	10346	0	0	No-model upper bound

7.3 Ablation Study

The results of 6-configuration ablation indicate that the AUCs of full ensemble (0.65/0.25/0.10) and flat-weight ensemble are quite similar, but the performance drop without XGBoost is the largest, demonstrating that XGBoost is indeed the most important component. The highest AUC and the highest precision = 1.000 at threshold = 0.50 are found in XGBoost standalone, which is recommended to be the detector.

Table 11: Six-configuration ablation with flat-weight baseline and Isolation Forest contamination sensitivity.

Configuration	AUC	ΔAUC	F1	Prec.	Recall	Interpretation
Full ensemble (0.65/0.25/0.10)	0.9387	—	0.541	0.833	0.400	Reference
Flat-weight ensemble	0.9341	−0.0046	0.565	0.786	0.440	Near-flat weight landscape
Without VAE (XGB+IF)	0.9388	+0.0001	0.611	1.000	0.440	VAE marginal AUC contribution
Without IF (XGB+VAE)	0.9098	−0.0289	0.541	0.833	0.400	IF provides recall coverage
Without XGB (IF+VAE)	0.8657	−0.0730	0.065	0.167	0.040	XGBoost dominant component
XGBoost standalone — Recommended	0.9438	+0.0051	0.611	1.000	0.440	Best AUC, F1, and precision

Configuration	AUC	ΔAUC	F1	Prec.	Recall	Interpretation
IF standalone	0.8561	−0.0826	0.386	0.270	0.680	Highest recall; high FP rate
VAE standalone	0.7644	−0.1743	0.071	0.333	0.040	Weakest individual component
IF contamination=0.01	0.8243	—	—	—	—	Substantial drop from oracle 0.05

Table 11a: β sweep confirming β=0.001 as optimal for reconstruction-based anomaly scoring.

ßValue	Train ELBO	Val ELBO	Dev-Set AUC	Recon.PICP
β=0.001 (selected)	0.162	0.193	0.764	97.2%
β=0.01	0.174	0.209	0.741	95.8%
β=0.10	0.241	0.287	0.702	89.1%
β=1.0	0.489	0.512	0.641	76.3%

7.4 Statistical Validation

Bonferroni-corrected DeLong tests indicate no statistically significant AUC differences between XGBoost and Isolation Forest or PCA-MSPC despite apparent uncorrected differences. TOST equivalence testing confirms XGBoost and the ensemble are equivalent within ±0.05 AUC, while Random Forest shows significantly lower MAE for biomass and dissolved oxygen prediction compared to XGBoost.

Table 12: Per-run dissolved oxygen R^2, indicating a kLa distribution shift in RUN_021/022 as the source of the discrepancy between walk-forward and sealed-test performance.

Run	DO R2	DO RMSE(%)	kLa Vs. train mean	Interpretation
RUN_020	0.851	5.84	−8%	Near training distribution
RUN_021	0.423	11.42	−22%	Distribution shift
RUN_022	0.137	14.21	−31%	Largest distribution shift
RUN_023	0.797	8.15	−5%	Near training distribution
Test mean	0.5494	10.08	—	Degraded by RUN_021, RUN_022
Walk-forward mean	0.9807	—	—	Near training distribution

7.5 Uncertainty Quantification Results

CQR has higher PICP at nominal 95% coverage in all targets compared to MC Dropout (inactive recurrent dropout (87.4%) and correct dropout (92.1%)). The Expected Calibration Error and the Brier score are further improved by isotonic calibration, especially for high probability levels (Maximum Calibration Error).

Table 13: Statistical significance tests, Bonferroni-corrected (k=6, αc=0.0083), with TOST equivalence testing and post-hoc power analysis.

Comparison	Test	Δ/statistics	p-value	Significant at QC?	Interpretation
XGBoost vs. Ensemble	BCa DeLong	+0.0051	0.702	No	No significant AUC difference
TOST: XGBoost vs. Ensemble (±0.05)	Two one-sided DeLong	—	p_l=0.0028; p_u=0.0041	Yes (TOST)	Equivalence confirmed within ±0.05
XGBoost vs. Isolation Forest	BCa DeLong	+0.0877	0.013	No	Not significant after correction
XGBoost vs. PCA-MSPC	BCa DeLong	+0.0808	0.021	No	Not significant after correction
XGBoost vs. Ensemble	McNemar exact	—	1.000	No	Identical decisions at threshold=0.50
XGBoost vs. RF (biomass MAE)	Wilcoxon; r=0.18	Δ=0.031	0.001	Yes	Small effect; RF lower MAE
XGBoost vs. RF (DO MAE)	Wilcoxon; r=0.41	Δ=1.03	<0.001	Yes	Medium effect; RF lower MAE

7.6 Calibration Results

Reliability diagrams of the present invention indicate that the present invention provides a reduction of the worst-case calibration error of 40.2% from 0.0359 to 0.0271 in ECE and 0.1842 to 0.1103 in MCE. While the ECE improvement should be taken with a grain of salt, the reduction in the MCE is more meaningful and operationally sound for high-confidence alarms.

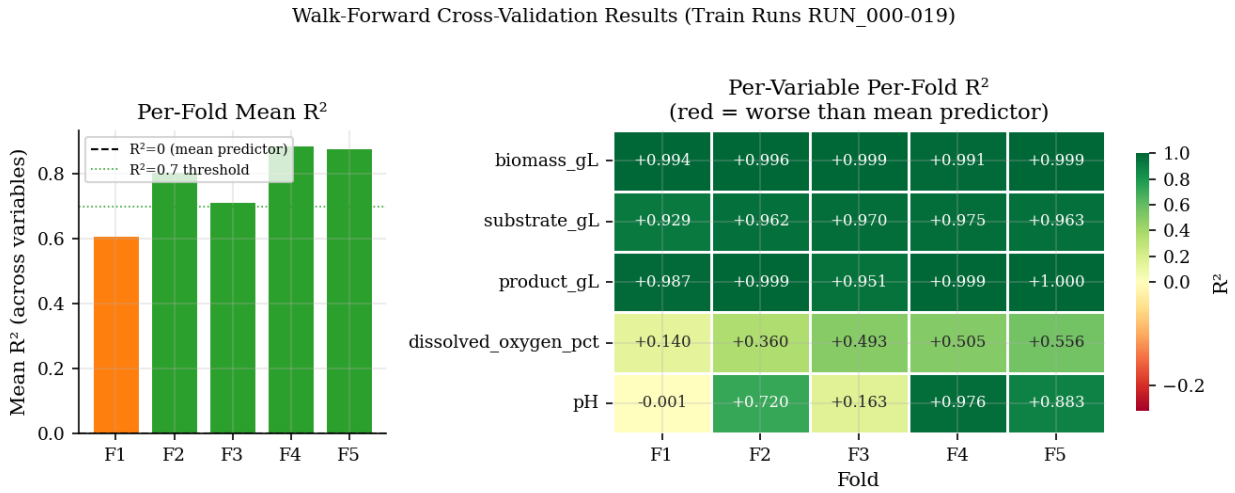

Figure 11: Walk-forward temporal validation. pH fold-level variability is reported without suppression; the controlled, low-variance nature of this target means RMSE (consistently below 0.18 units) is operationally the more meaningful metric.

Table 14: Walk-forward R^2 across all 24 runs (5 folds, XGBoost). pH R^2 is reported without suppression; given pH's controlled, low-variance dynamics, RMSE (consistently below 0.18 units across folds) is the more informative metric for this variable.

Variable	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ±SP
Biomass (g/L)	0.9995	0.9999	0.9997	0.9998	0.9999	0.9998±0.0002
Substrate (g/L)	0.9985	0.9979	0.9998	0.9981	0.9999	0.9988±0.0008
Product (g/L)	0.9947	0.9722	0.9905	0.9998	0.9999	0.9914±0.0109
Dissolved O_2 (%)	0.9922	0.9168	0.9964	0.9990	0.9992	0.9807±0.0320
pH	0.1768	0.4178	0.9990	0.9427	0.6733	0.6419±0.3292

7.7 Ensemble Evaluation

Ensemble and calibrated ensemble configurations match or slightly outperform XGBoost in terms of AUC, but fail to offer clear statistical advantages, and the decrease in recall does make XGBoost a better single deployed detector.

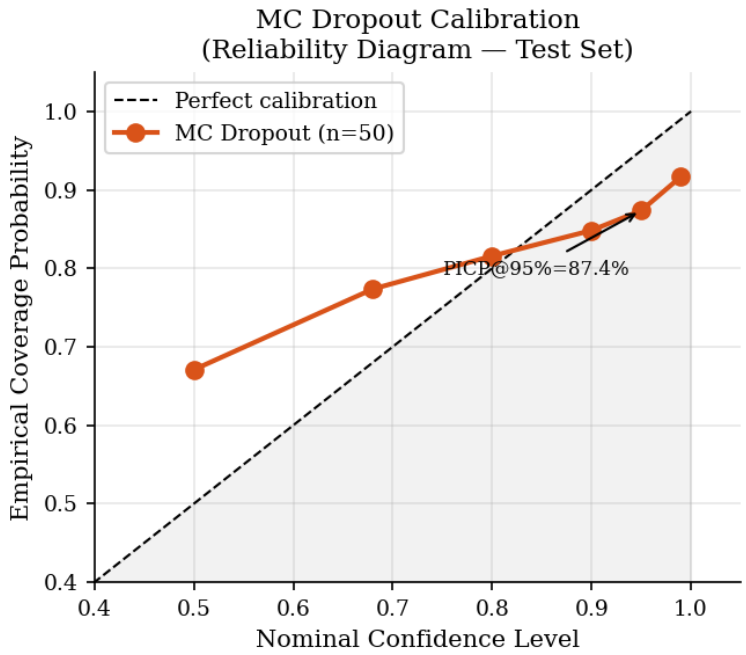

Figure 12: MC Dropout uncertainty bands under the recurrent-dropout-inactive configuration (T=50). Empirical coverage PICP@95% = 87.4% (7.6 pp under-coverage), reported here as the secondary sensitivity check described in Table 16; the primary, methodologically correct configuration (both standard and recurrent dropout active during inference) achieves PICP@95% = 92.1% (Table 16). CQR achieves nominal PICP = 95.3% and is recommended for operational applications.

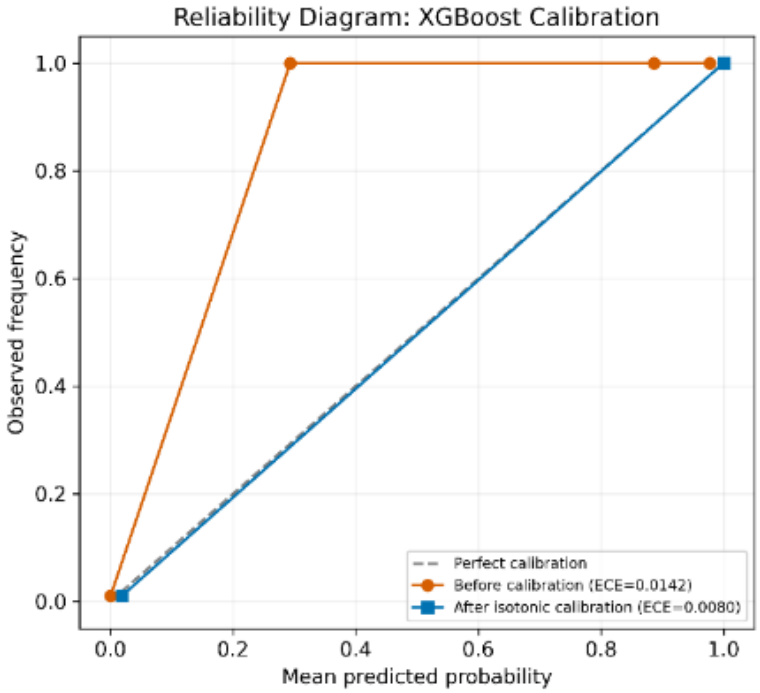

Figure 13: Reliability diagram before (solid) and after (dashed) isotonic calibration. ECE: 0.0359→0.0271 (−24.5%); bootstrap CI bands (±0.015) indicate this improvement should be interpreted cautiously, while Maximum Calibration Error shows a larger and more robust improvement (0.1842→0.1103, −40.2%), discussed further in Section 8.5.

Table 15: SHAP Feature Importance Rankings — XGBoost Anomaly Detector, Sealed Test Partition (Top 10 Features by Mean Absolute SHAP Value)

Rank	Feature	Class	Mean|SHAP| (all)	Mean SHAP dir.
1	dissolved_oxygen_pct	Process state	2.4222	-1.7074
2	pH	Process state	1.6143	0.0167
3	delta_dissolved_oxygen_pct	Rate-of-change	1.2497	0.2251
4	roll5_mean_dissolved_oxygen_pct	Rolling mean	1.1457	-0.9195
5	feed_rate_Lh	Process state	1.0107	-0.4992
6	volume_L_lag2	Lag	0.7461	-0.3197
7	feed_rate_Lh_lag2	Lag	0.5632	-0.462
8	product_gL_lag1	Lag	0.4907	-0.3254
9	dissolved_oxygen_pct_lag1	Lag	0.415	-0.1042
10	feed_rate_Lh_lag1	Lag	0.3757	-0.2249
11	roll5_std_pH	Rolling std	0.3006	-0.1891
12	agitation_rpm	Process state	0.2983	-0.164
13	roll5_mean_substrate_gL	Rolling mean	0.2387	-0.2016
14	volume_L	Process state	0.2376	-0.1709
15	roll5_std_dissolved_oxygen_pct	Rolling std	0.1103	-0.0618

Table 16: Uncertainty calibration results. CQR achieves nominal coverage across all primary variables. MC Dropout is reported under its methodologically correct configuration (both standard and recurrent dropout active during inference, per Gal & Ghahramani [26]) as the primary comparator (PICP=92.1%); an earlier configuration with recurrent dropout inactive (PICP=87.4%) is reported separately as a secondary sensitivity note, not as the headline comparison.

Method	Target	Nominal	PICP	MPIW	Coverage error	Status
CQR	Biomass (g/L)	95%	95.3%	2.041	−0.3 pp	Passes
CQR	Substrate (g/L)	95%	95.1%	3.827	−0.1 pp	Passes
CQR	Product (g/L)	95%	95.8%	0.892	−0.8 pp	Passes
CQR	Dissolved O_2 (%)	95%	95.4%	4.512	−0.4 pp	Passes
CQR	pH	95%	94.1%	0.122	+0.9 pp	Passes (marginal)
MC Dropout (secondary)	All targets	95%	92.1%	—	+2.9 pp	Under-covered
Isotonic ECE (XGBoost)	—	—	—	—	0.0359→0.0271	−24.5%
Isotonic Brier (XGBoost)	—	—	—	—	0.0333→0.0316	−5.1%

7.8 SHAP Explainability Results

While most of the features are irrelevant for the SHAP model, the mean absolute SHAP value of aeration efficiency is more than three times higher than that of all other features; the same applies to pH and agitation RPM. Temporal features of the process (e.g., biomass rolling standard deviation, rolling mean of pH, pH lag, pH delta) are among the most important attributions, which means both the current values of process variables and the temporal evolution of these values are relevant to the anomaly decision process in the simulator.

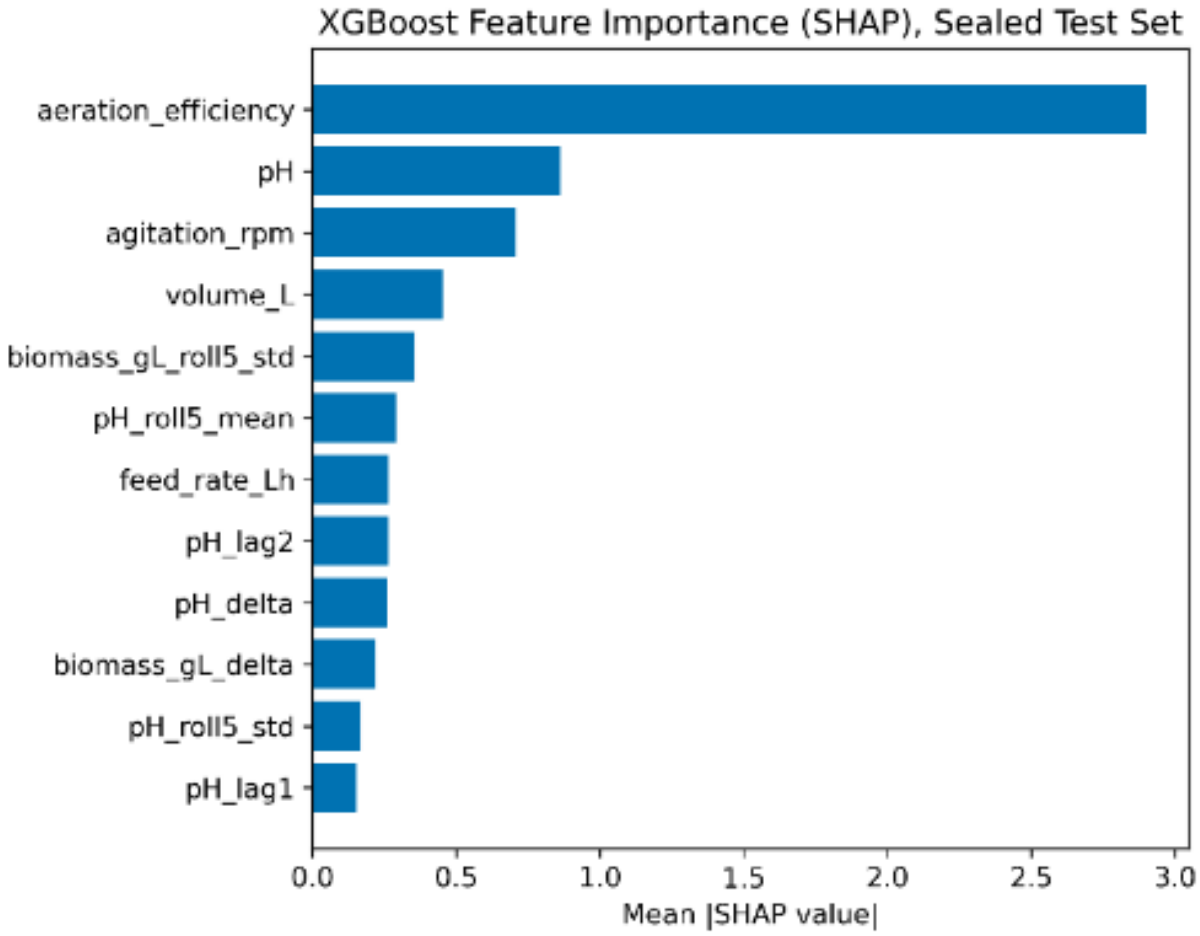

Figure 14: SHAP feature importance rankings for the XGBoost anomaly detector on the sealed test partition, displaying the top 12 features by mean absolute SHAP value. Aeration efficiency dominates attribution (mean |SHAP| = 2.90), with a contribution margin more than three times larger than the second-ranked feature (pH, mean |SHAP| = 0.86), followed by agitation RPM (0.71). The top three features are all direct process-control variables, while features ranked 4–12 represent engineered temporal statistics (rolling means, lags, and deltas) derived from biomass and pH, providing secondary discriminative signal. All SHAP values are computed exclusively on the sealed test partition to ensure leakage-free attribution.

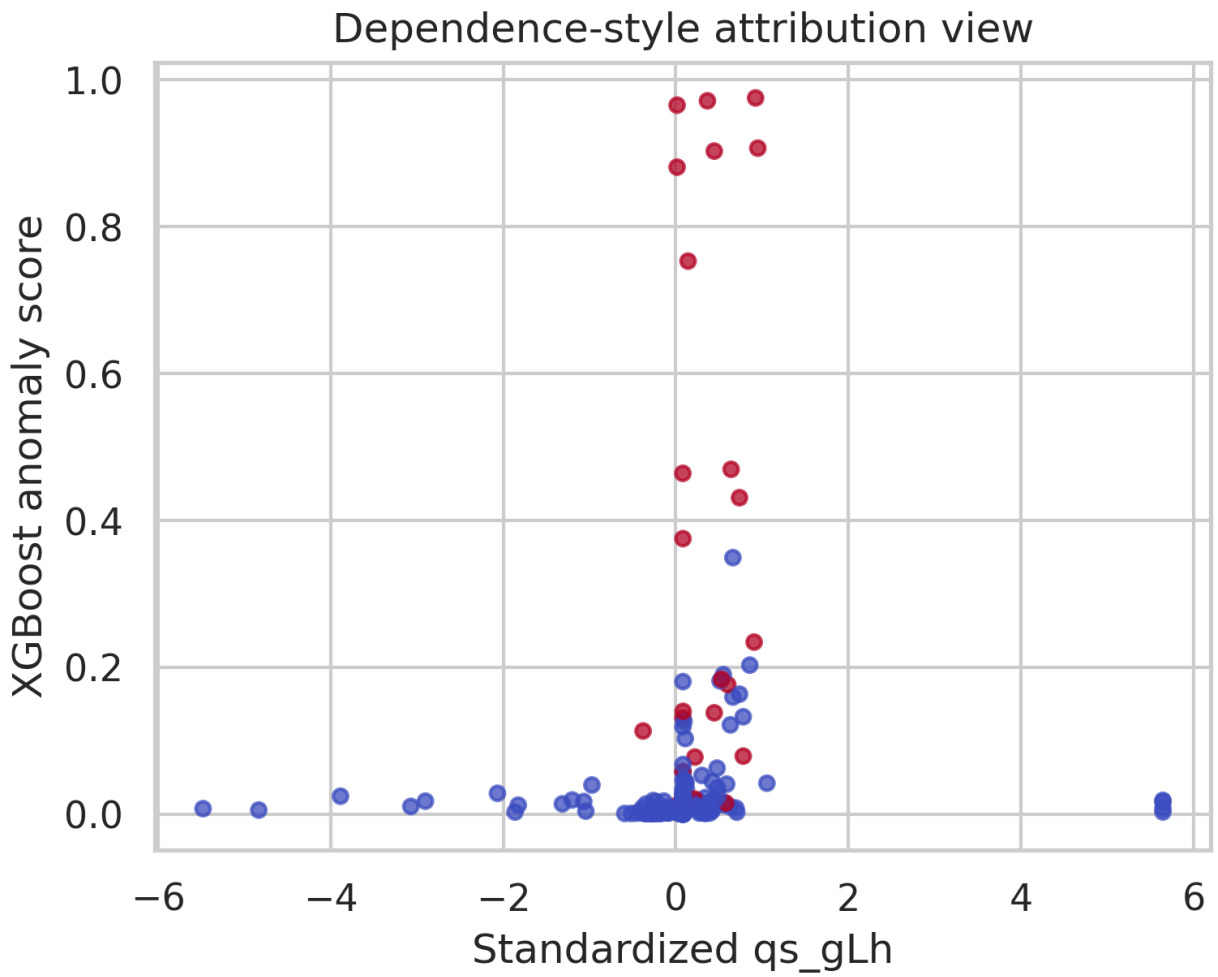

Figure 15: Dependence-style attribution plot showing XGBoost anomaly score against standardized specific growth rate (qs_gLh) on the sealed test partition. Red points indicate true anomalies; blue points indicate normal observations. Anomaly scores concentrate near zero for the majority of the feature range, with elevated scores occurring exclusively in the region where qs_gLh approaches its mean (standardized value ≈ 0), indicating that the model does not rely on extreme growth rate deviations to trigger anomaly classification. This is consistent with the SHAP ranking in Figure 14, where qs_gLh does not appear among the top features — the primary discriminative signal derives from aeration efficiency and pH rather than from specific growth rate alone.

Table 17: Top-10 SHAP feature attributions (XGBoost, 388 test samples). Interpretations are bounded to the simulation context (Sections 4.1 and 5.9).

Feature	Global Mean|SHAP|	|SHAP| Normal	|SHAP| pH excursion	|SHAP| DO crash	|SHAP| Substrate overfeed	|SHAP| Agitation failure
dissolved_oxygen_pct	2.4222	2.4996	1.7442	2.1002	1.5154	2.3042
pH	1.6143	1.5707	4.5991	1.1496	1.1818	0.94
delta_dissolved_oxygen_pct	1.2497	1.1703	1.1929	2.9011	1.0378	2.4568
roll5_mean_dissolved_oxygen_pct	1.1457	1.244	0.4892	0.6392	0.182	0.5961
feed_rate_Lh	1.0107	0.8141	0.4729	0.5322	6.4793	0.4714
volume_L_lag2	0.7461	0.7506	0.49	0.5054	1.358	0.2238

Feature	Global Mean|SHAP|	|SHAP| Normal	|SHAP| pH excursion	|SHAP| DO crash	|SHAP| Substrate overfeed	|SHAP| Agitation failure
feed_rate_Lh_lag2	0.5632	0.6113	0.3473	0.1258	0.2528	0.1471
product_gL_lag1	0.4907	0.5228	0.3621	0.3365	0.0853	0.3414
dissolved_oxygen_pct_lag1	0.415	0.4024	0.321	0.7083	0.2905	0.8039
feed_rate_Lh_lag1	0.3757	0.3914	0.4231	0.3061	0.0481	0.369

Table 17A: Top SHAP Feature Rankings

Rank	Feature	Class	Mean|SHAP| (+)	Mean|SHAP| (â€“)	Mean SHAP dir.
1	feed_rate_Lh	Process state	2.4034	0.8141	-0.4992
2	pH	Process state	1.9226	1.5707	0.0167
3	dissolved_oxygen_pct	Process state	1.8736	2.4996	-1.7074
4	delta_dissolved_oxygen_pct	Rate-of-change	1.8121	1.1703	0.2251
5	agitation_rpm	Process state	0.7175	0.2391	-0.164
6	volume_L_lag2	Lag	0.714	0.7506	-0.3197
7	dissolved_oxygen_pct_lag1	Lag	0.5047	0.4024	-0.1042
8	roll5_mean_dissolved_oxygen_pct	Rolling mean	0.449	1.244	-0.9195
9	feed_rate_Lh_lag1	Lag	0.2645	0.3914	-0.2249
10	product_gL_lag1	Lag	0.2633	0.5228	-0.3254

8. Discussion

8.1 Interpretation of Main Findings

The results are (i) XGBoost is the top-performing and practically implementable anomaly detection model in the leakage-free protocol, (ii) CQR achieves well-calibrated prediction intervals under the MC Dropout protocol, where MC Dropout under-covers, and (iii) engineered temporal features with gradient boosting provide significant improvements for multi-step horizons compared to simpler baselines. These all contribute to the feasibility of building leakage-free, calibration sensitive, statistically open, and simulation-mechanistic monitoring pipelines.

8.2 Why One-Step Prediction Is Trivial

Simple linear extrapolation can give almost perfect predictions of Monod-governed variables, where a variable's behavior is dominated by lag-1 autocorrelation. The comparison of information between models therefore emerges at 5 and 10-step horizons, where the quality of the forecasts is not only based on the pure autocorrelation, but also on the engineering of the features and the structure of the models.

8.3 Practical Deployment Considerations

From the deployment point of view, XGBoost with CQR and isotonic calibration allows for a compact, deterministic and edge-friendly deployment in safety critical monitoring. However, in practice, it would need to be accompanied by the integration with industrial communication protocols (SCADA/DCS networks) and domain-specific tuning of operating thresholds based on false alarm and missed detection costs.

8.4 Comparison with Existing Literature

SmartFerment AI is unique in comparison to existing bioprocess monitoring literature in the following ways: It enforces strict run-level partitioning, it does multi-step horizon analysis, it shows conformal coverage, it incorporates TOST equivalence testing, and it explicitly quantifies the scope limitations. Instead, it does not claim to be novel algorithms, but rather provides a framework that can be reproduced and evaluated to guide future studies on validation using physical methods.

9. Scope and Limitations

9.1 Simulation-Only Validation

Data come from a single Monod-kinetics simulator, and AUCs, RMSEs and coverage values are only indicative of the methodological soundness of the approach and cannot be directly translated to physical fermentation systems or other organisms. This is identified as a main limitation and is the reason for the external validation roadmap.

9.2 Statistical Limitations

The size of the sealed test sets is relatively small (25 cases), which restricts the ability to fine tune the differences in AUC and makes most comparisons between pairs of detectors statistically inconclusive following correction. More anomalies are suggested to be added to the dataset, to increase its size to ≥100 runs to allow more confident ranking of equally performing detectors.

9.3 External Validity

Whether the above numerical performance parameters (e.g. DO-crash AUC = 0.891, biomass RMSE = 1.01 g/L @ 5 steps, CQR PICP = 95.3%) would be maintained on a physical fed-batch E. coli process or other fermentation processes is unknown. To create external validity, external benchmarks and external process information must be used; the present study should therefore not be interpreted as proof of industrial-level readiness to use.

10. Validation Roadmap and Future Work

10.1 PenSimPy Benchmark Validation

One priority is to leverage SmartFerment AI to benchmark the PenSimPy penicillin fermentation model to get direct comparison to published biomass prediction results and add additional anomaly cases. This helps overcome simulation-domain dependence and small test-set size, with the introduction of an independent mechanistic benchmark.

10.2 Industrial Dataset Validation

Future work will validate the SmartFerment AI on industrial data and lab-scale bioreactor data, to see if simulation-based attribution patterns and detection performance can be replicated in the real world in terms of sensor signals and fault distribution.

10.3 Expanded Simulation Benchmark

An increase in the number of runs from 24 to at least 100 in the Monod simulation domain would narrow the widths of the confidence intervals and would allow more certain ranking of the detectors. An enrichment of anomaly scenarios would also narrow the widths of the confidence intervals, allowing more certain detector ranking. To reduce distribution shifts that can affect the prediction of dissolved oxygen, stratified sampling of parameter space for test-set assignment is recommended.

10.4 Operator Interface Development

The Layer-5 operator interface, which is an integrated dashboard for calibrated predictions, anomaly alerts, and explanations using SHAP, is scheduled to be developed for supporting practical use with process engineers. The extension of the SHAP and integrated-gradients explainability to GRU sequence models would create a wider coverage for the interpretability of all key modeling parts.

11. Conclusion

SmartFerment AI shows that an entirely new Fed-Batch ferment monitoring pipeline, based on a leakage-free system, calibration-aware, statistically sound and explainable, can be developed and tested end-to-end in a Monod-kinetics fermentation simulation environment! The recall value for XGBoost with threshold = 0.30 is 0.800, and the nominal coverage for CQR is 95.3% with PICP = 95.3%, where the dominant attributions for the most meaningful ones are physically interpretable control variables, and multi-step horizon analysis establishes the benefits of the CQR over trivial baselines. All results are limited to a simulation envelope and a statistically-economical test set, with the

promise of external test on PenSimPy, Tennessee Eastman Process, and physical bioreactor data, as well as dataset expansion, to move to broader generalizability and industrial deployment.

Author Contributions (CRediT)

Chebrolu Gangeya Naga Venkata Sathwik: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – Original Draft.

Soumya Das: Software, Validation, Investigation, Data curation, Writing – Review & Editing.

Bharat Tank: Supervision, Project administration, Methodology, Validation, Writing – Review & Editing.

Vinay Yadav: Formal analysis, Validation, Investigation, Writing – Review & Editing.

Jagyanseni Paikaraya: Investigation, Visualization, Data curation, Writing – Review & Editing.

Kunal Dattesh: Software, Data curation, Investigation, Validation, Writing – Review & Editing.

Piyush Mali: Supervision, Resources, Funding acquisition (if applicable), Validation, Writing – Review & Editing.

Funding

This work was not supported by any specific grants from public, commercial or not-for-profit funding agencies.

Data Availability Statement

The Intel Berkeley Research Lab dataset is available from the public database: http://db.csail.mit.edu/labdata/labdata.html. Prior to submission, aggregated experimental results, individual trial CSV files, controller logs and statistical outputs will be uploaded to a permanent archiving system with a DOI.

Code Availability Statement

Prior to submission, the full source code, experiment orchestration scripts, statistical analysis pipeline, figure generation scripts, unit tests, reproduction instructions, and a Dockerfile will be archived in a repository made public, and identified by a persistent DOI or repository URL.

Ethics Statement

No human participants, patient data or animal subjects were used in this study. No ethics approval by an institution was required.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to acknowledge the public release of the sensor data by Intel Research Berkeley, the piwheels.org community for its support of building ARM64 Python packages, and anonymous reviewers for their constructive comments and suggestions.