Project Update #3 (February 23, 2026)

CS 6678 - Advanced Machine Learning
Instructor: Dr. Leslie Kerby

George Lake

Current Project Focus

Research Question
Can machine learning distinguish between benign sensor failures and malicious cyber attacks in Industrial Control Systems (ICS)? Existing anomaly detection research treats this as binary classification (normal vs. abnormal), but operational response to failures versus attacks differs fundamentally; misclassification leads to either alarm fatigue or undetected persistence.

Changes since last update
The pipeline has progressed from preprocessing infrastructure (Update #2) to a five-notebook pipeline through binary classification.
The following Jupyter notebooks were created:

- WaDi Notebook 1 - Data Collection
- WaDi Notebook 2 - Fault Injection
- WaDi Notebook 3 - Curate Validate
- WaDi Notebook 4 - Feature Engineering
- WaDi Notebook 5 - Binary Classification

Work Since Last Update

Completed Fault Injection Pipeline (WaDi Notebook 2)
Implemented a five-type synthetic fault injection framework covering drift, bias, precision degradation, stuck-at, and intermittent dropout. The resulting dataset contains ≈ 1.4M labeled rows (80.75% normal, 12.50% attack, 6.75% synthetic fault). Reference Failure Injection Design Documing - WaDi

Figure 1 illustrates each fault type applied to a synthetic sensor signal over a 10-minute window. Each panel overlays the faulted signal (colored) against normal behavior (light blue). Drift accumulates gradually, bias introduces a sudden offset, precision degradation adds noise, stuck-at freezes the sensor, and intermittent dropout simulates signal loss.

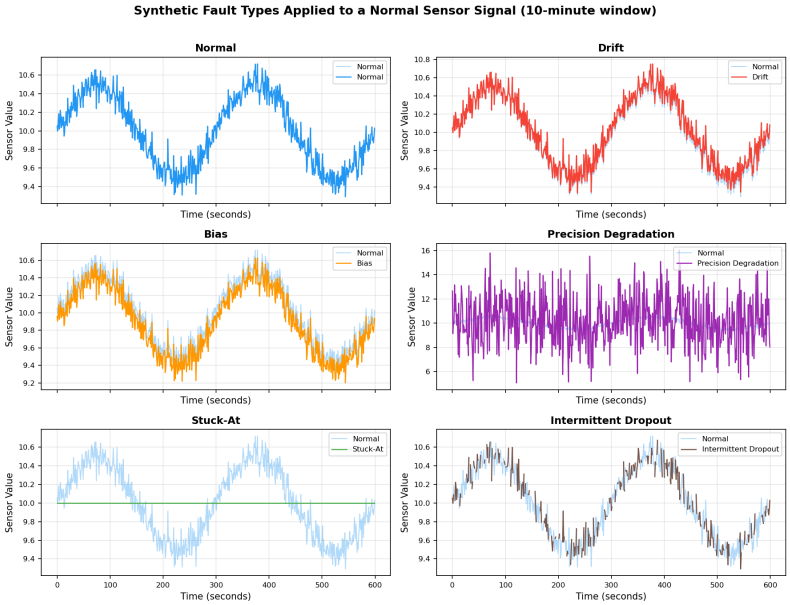

Figure 1: Five synthetic fault types applied to a normal sensor signal.

Completed Curation, Validation, and Feature Engineering (WaDi Notebooks 3-4)
Build a validation suite covering structural integrity, label integrity, sensor physical plausibility bounds, and temporal leakage audit. Feature engineering produced 570 features (95 raw + 475 rolling statistics at 60-second windows) after dropping three zero-variance sensors. Normalization was fit exclusively on the training split to close the leakage risk identified during pipeline design.

Executed Binary Classification (WaDi Notebook 5)
Trained a Logistic Regression baseline and a Random Forest primary model on the binary task (normal vs. anomaly). Key results:

- Random Forest OOB score: 0.9996 - excellent fit on training distribution
- High recall operating point (threshold=0.2): attack recall=1.0, but ≈ 220k false positives on ≈ 223k normal holdout rows
- No threshold produced acceptable precision and recall simultaneously on the test split.
- Logistic Regression achieved normal recall of 0.18. Linear boundary insufficient for this feature space

Figure 2 shows confusion matrices at threshold=0.2 (high recall) and threshold=0.6 (high precision). At threshold=0.2, the model achieves great attack recall but correctly identifies only 4,173 of 223,251 normal holdout rows. Raising the threshold to 0.6 improves normal identification to 26,877 rows, but at the cost of missing 61,069 attacks. No threshold resolves the underlying problem, The model is not learning a meaningful boundary between normal and anomalous behavior. It is treating the Oct 6-9 normal holdout period as anomalous because that period is distributionally distinct from the Sep 25-Oct 6 training period.

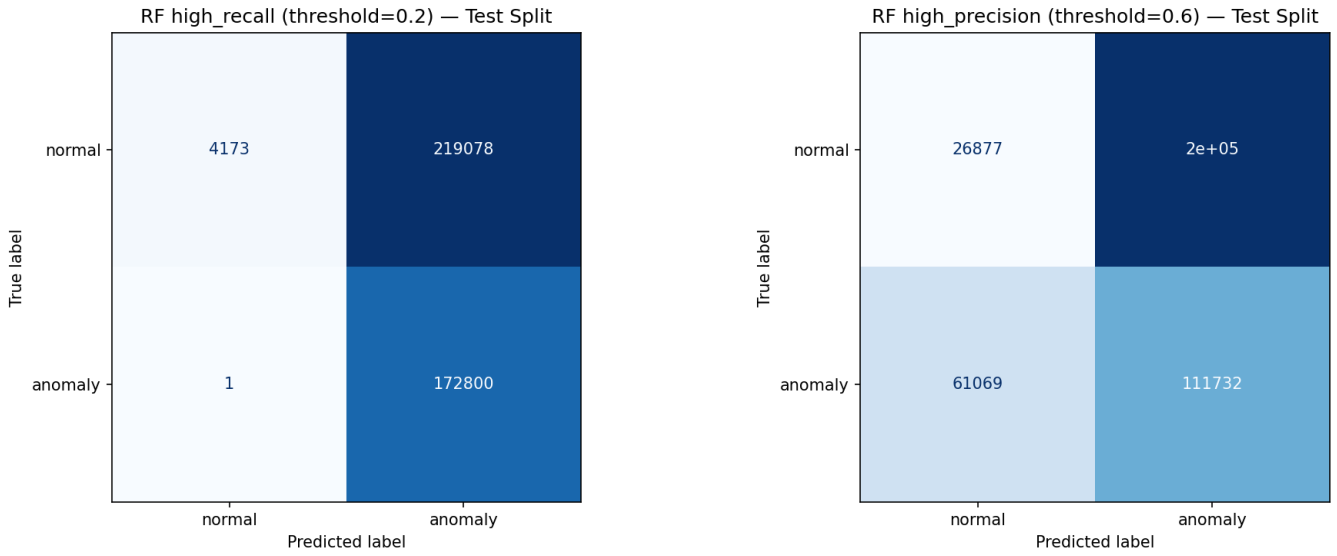

Figure 2: RF confusion matrices at threshold=0.2 (left) and threshold=0.6 (right).

What Was Learned

Binary Classification is Defeated by Temporal Distribution Shift
The test set normal rows (Oct 6-9, the pre-attack period) are distributionally distinct from the training rows (Sep 25-Oct 6). The model learned the earlier "normal" behavior correctly. The pre-attack period already looks anomalous by that standard. This is a documented WaDi dataset property identified in prior research (Turrin - Sections 6.4, 6.5). They find that only 18 sensors maintained consistent statistical behavior between training and test splits. I believe that this is not a model failure, but a dataset property that makes supervised binary classification on WaDi inherently limited regardless of model quality.

Three-Class Problem
Because binary detection is limited by this distribution shift, the three-class question (given an anomaly signal, is it an attack or sensor fault?) is more operationally useful. It does not require normal rows in the test set. The test set is pure, real attack data. The model must distinguish those attacks from the synthetic fault patterns it learned during training. This will be the primary framing of my project.

Published Results Use Incompatible Evaluation Protocols
Prior WaDi papers reporting strong F1 scores (0.75-0.98) use unsupervised, one-class evaluation: train on normal data only, evaluate on the attack period, and apply point-adjust scoring that credits the model for detecting an entire attack window when it fires at any point within it. These results are not directly comparable to a supervised three-class evaluation.

- USAD: Unsupervised Anomaly Detection on Multivariate Time Series
- Graph Neural Network-Based Anomaly Detection in Multivariate Time Series
- MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks

Figure 3 shows the top 20 features by Random Forest importance. Rolling window statistics (60-second min, max, mean) on analytic instrument transmitters (AIT sensors) dominate over raw sensor readings, indicating the model detects sustained deviations over time rather than instantaneous spikes. AIT sensors measure water quality indicators such as chlorine, pH, and conductivity, the process variables likely to be manipulated in a water distribution attack.

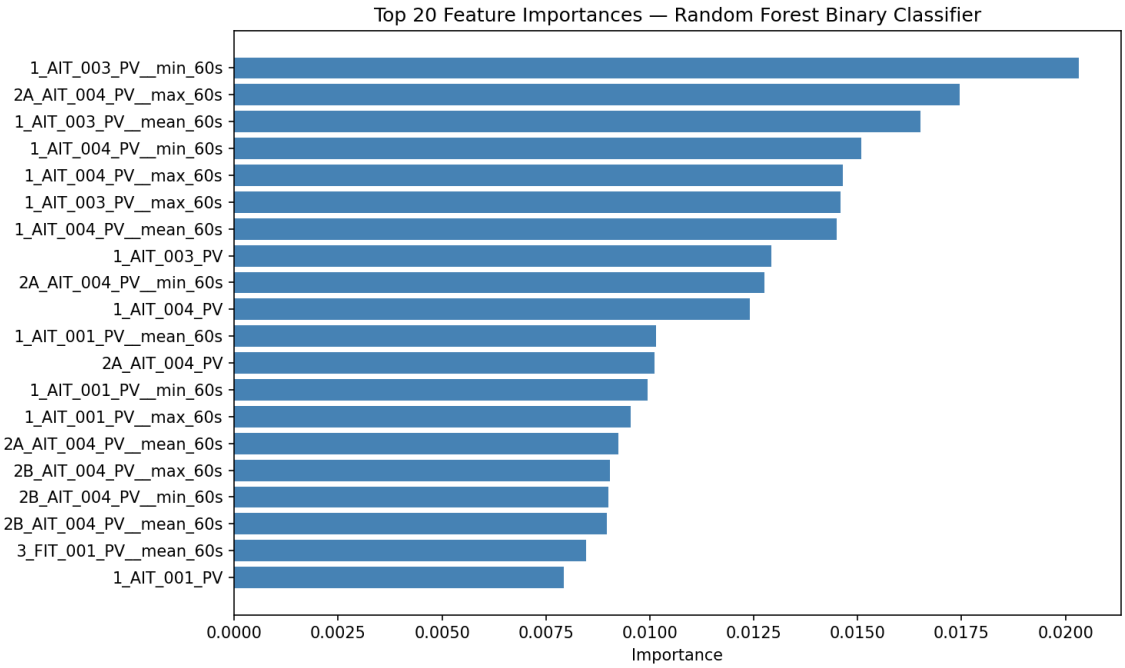

Figure 3: Top 20 Random Forest feature importances from binary classification

Current Challenges or Risks

No Ground Truth for Fault Realism Validation
Synthetic faults are parameter-justified by literature but cannot be validated against real sensor failure events, no labeled real-failure data exists for WaDi. The risk is that synthetic patterns are either too separable from attacks (trivially easy) or insufficiently realistic (undermining the research). Notebook 6 results will reveal which applies.

Evaluation Metric Selection for the Three-Class Problem
The primary metric is attack recall, how many of the ≈174k real attack rows are correctly classified as attacks rather than faults. Macro-averaged F1 may be misleading given class imbalance and the synthetic nature of the fault class, so metric framing requires careful consideration before reporting results.

Next Steps

- **Execute Notebook 6 - Three-Class Classification**. Train on normal + synthetic faults, evaluate on real attack rows only. Key question: does the model correctly classify attacks as attacks rather than faults?
- **Analyze three-class results**. Focus on attack recall and confusion patterns — which fault types are most attack-like in the feature space?
- **Resolve metric framing**. Per-class precision/recall/F1 with emphasis on attack-class metrics; macro and weighted averages as supplementary.
- **Begin paper outline**. Draft methodology section while pipeline decisions are fresh.