

Anomaly Detection and Forecasting for the KFA71 Extraction Kicker

M. Algelly

F. Velotti, K. Papastegios, P. Ellison, A. Kalousis, V. Kain

5th ICFA Beam Dynamics Workshop on Machine Learning for Particle Accelerators

10/04/2025



Content

**1. KFA71
System Overview**
Key components,
challenges.

2. Data & Labeling
Description and
techniques

3. Approach & Models
Waveform analysis,
VAE/CVAE.

4. Continual Learning
Technique and practical
case.

**5. Conclusion &
Outlook**
Insights and future
directions.

System Overview – The KFA71 Extraction Kicker

Purpose: Fast-pulsed magnet system to extract particle beams from the Proton Synchrotron (PS).

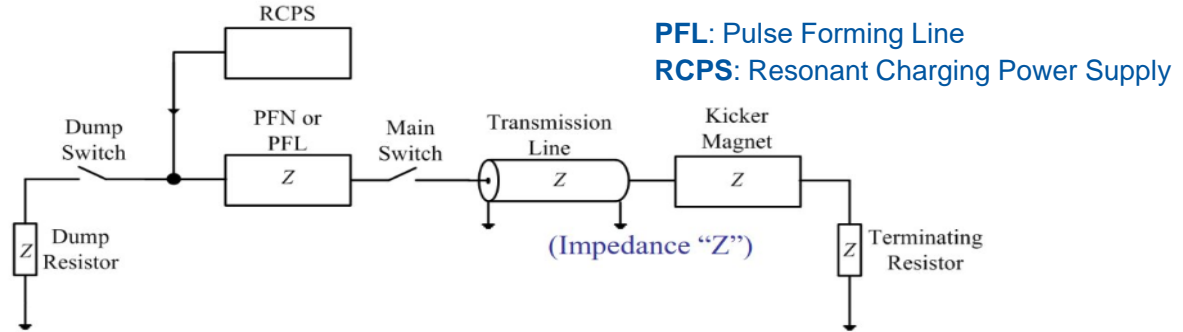
Components: 12 generator modules operating simultaneously in vacuum tanks.

Output:

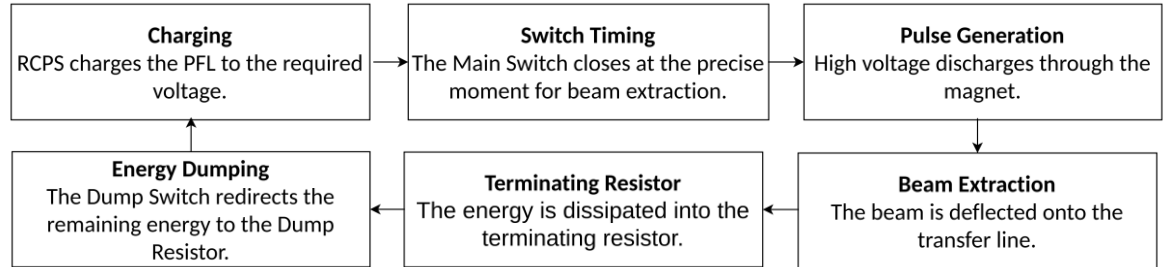
High-voltage pulses (~80 kV peak).

Key Info:

- System installed in the 1970s.
- Complex sub-components: HV switches (thyratrons), cables, transmission lines, ferrite magnets.
- Aging system = higher anomaly rates.
- Reactive maintenance, not proactive or predictive.



Schema: *Simplified schematic of a kicker module*[1]



Data Description

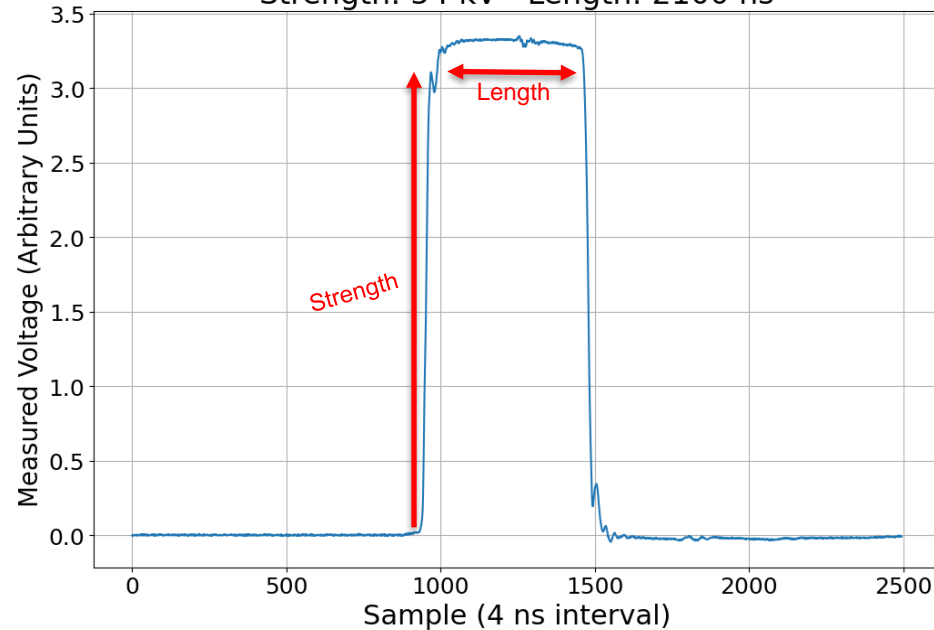
Waveform Characteristics:

- **Sampling Rate:** 1 sample every 4 ns for 10 μ s
→ 2500 sample points
- **Signal Content:** Short rise and fall times, short plateau region.
- **12 generators** → 12 waveforms per cycle
- **Pulse Settings:** Includes desired pulse strength and length

Waveforms have been stored in NXCALS since the end of September 2024

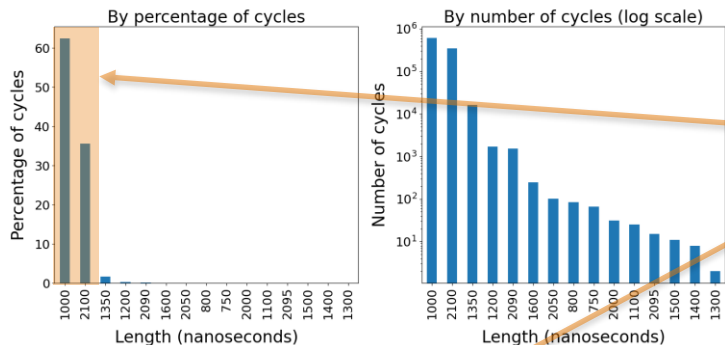
→ Current analysis and training focus on October & November 2024 data.

Example Waveform from the First Generator
Strength: 54 kV - Length: 2100 ns

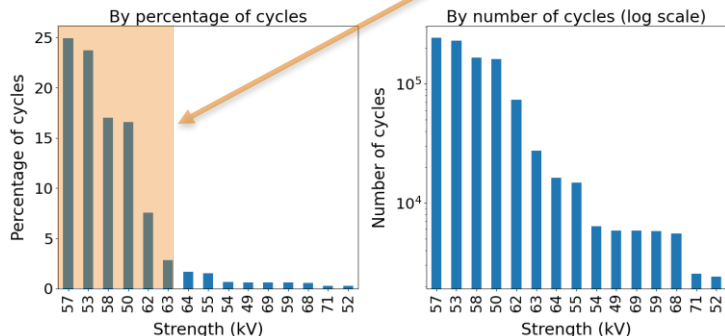


Settings Distribution – Strength and Length

Top 15 Length Settings for October 2024



Top 15 Strength Settings for October 2024



Key Challenges and Implications:

Imbalanced Data:

- ~98% of waveforms share 2 length values.
- ~80% of waveform share 6 voltage values.
- Risk: **Overfitting** to dominant settings.

Consequences:

- Rare configurations **misclassified** as anomalies.
- Reduces detection accuracy and increased biases.

Recommendations

- **Dataset Balancing:** Sampling, augmentation, reweighting.
- **Performance Monitoring:** Focus on rare settings.
- **Leverage Diversity:** Use rare configurations to improve robustness.

Labeling Process: Challenges and Key Steps

Context:

- How to label a subset of waveforms from millions of records ?

Two Approach:

• Comparing IPOC Data:

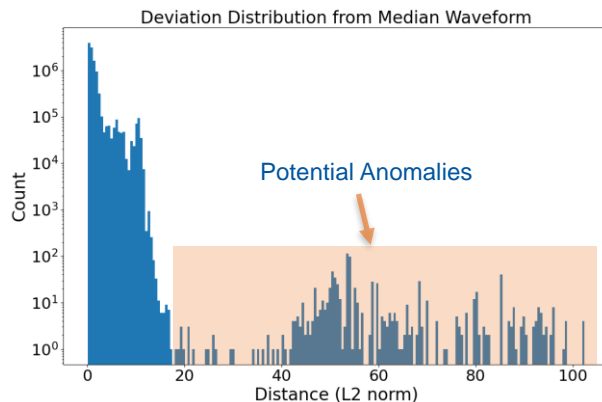
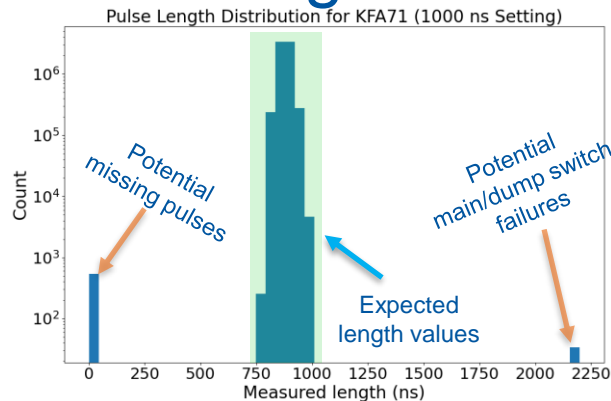
- Measured pulse properties against expected settings.
- Detect issues like *missing pulses* or *faulty shot*.

• Median Waveform Computation:

- Group waveforms by *strength* and *length*.
- Compute median waveforms.
- Compute deviations (e.g., L2 norm).

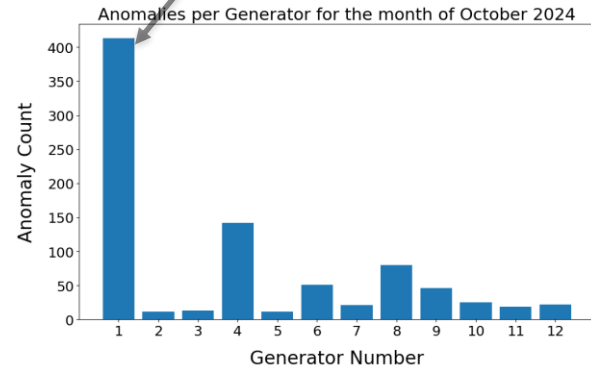
Outcome:

- Preliminary set of anomalies for evaluation.
- Manual verification feasible due to reduced candidate anomalies.



In the following sections, we focus specifically on the first generator of the KFA71

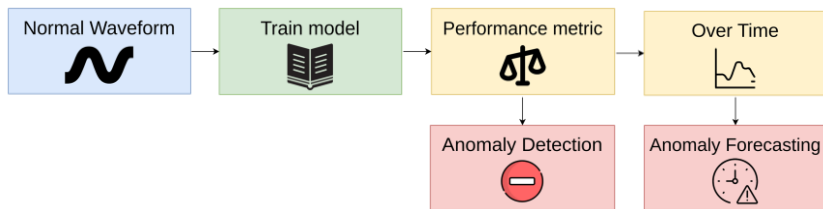
→ more than **400k** cycle where this generator should have pulsed.



Proposed Approach: Anomaly Detection & Prediction

General Idea:

1. Train a model on nominal waveforms.
2. Detect deviations using performance metrics.
3. Monitor trends to identify drifts or early anomalies.



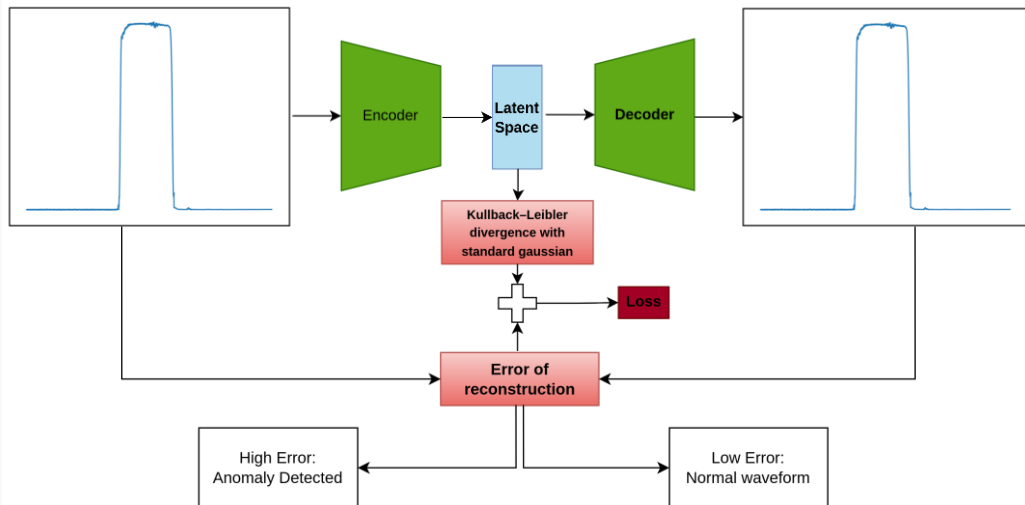
Mathematical Formulation[1]:

$$\mathbb{E}[L] = (1 - \epsilon) \cdot \mathbb{E}[l_w(X_{good})] + \epsilon \cdot \mathbb{E}[l_w(X_{anom})]$$

- $l_w(X)$: Loss for waveform X with model's weights w .
- ϵ : Fraction of anomalies.

Model: VAE Components

- **Encoder**: Maps data to a compressed probabilistic representation. Output two vector: μ & σ^2 .
- **Latent Space**: Probabilistic representation: $z \sim \mathcal{N}(\mu, \sigma^2)$.
- **Decoder**: Reconstructs data: Decoder(z) = \hat{x}
- **Loss**: Mean Squar Error + Kullback-Leibler divergence with $\mathcal{N}(0, I)$



ML Models – Conditional Variational Autoencoders (CVAE)^[1]

Contextual Data:

Use contextual information:

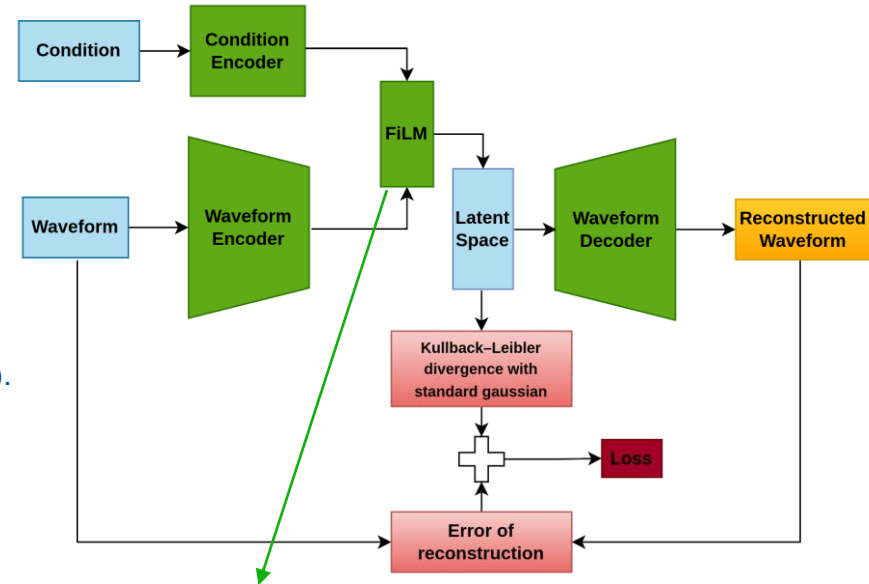
Settings (e.g., strength, length)

Specificities:

- **Condition Encoder:** Encodes condition into its own latent space.
- **FiLM-style Fusion**[2]: Modulates waveform latent using condition (γ , β).
- **Latent Awareness:** Conditioning modulates the latent space using context, which could help encode condition-related structure

Advantages:

- **More general model:** More robust reconstruction, especially for rare pulses or unseen pulses.
- **Improved Robustness:** Higher anomaly score for subtle anomalies.



Feature-wise Linear Modulation

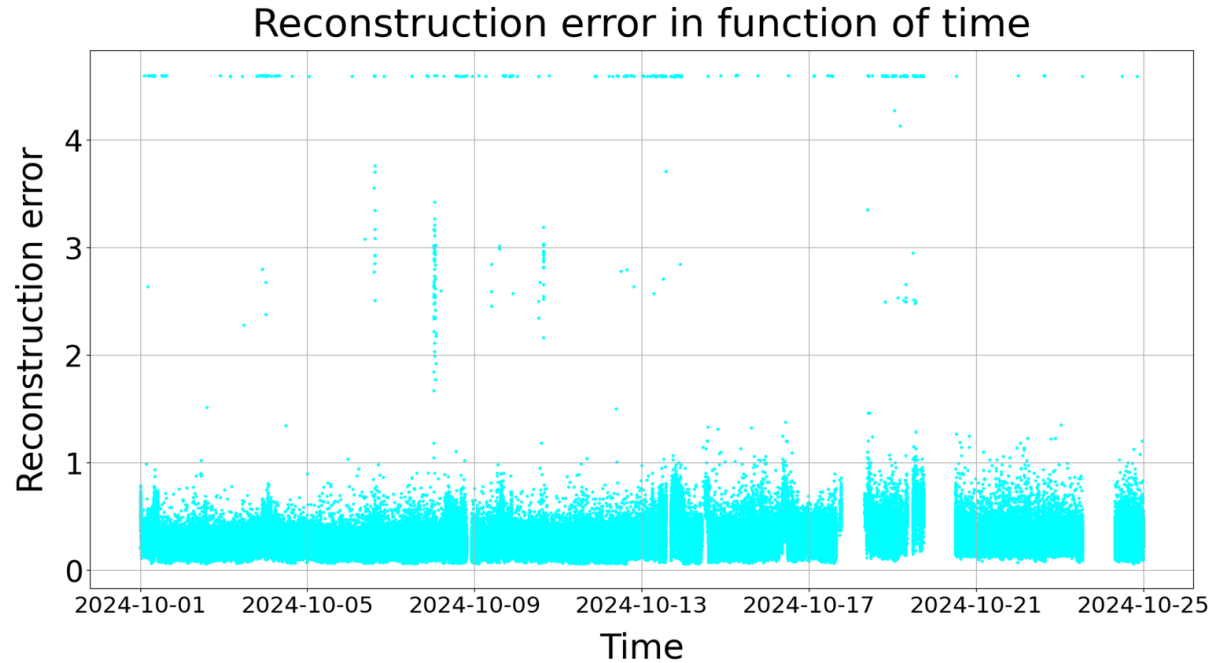
$$\begin{aligned}\mu &= \gamma_{\mu}(\mu_{cond}) \cdot \mu_{wave} + \beta_{\mu}(\mu_{cond}) \\ \sigma^2 &= \gamma_{\sigma}(\sigma_{cond}^2) \cdot \sigma_{wave}^2 + \beta_{\sigma}(\sigma_{cond}^2)\end{aligned}$$

[1] Pol et al., *Anomaly Detection With CVAE*, arXiv:2010.05531 (2020)

[2] E. Perez et al., *FiLM: Visual Reasoning with a General Conditioning Layer*, arXiv:1709.07871 [cs.CV], 2017. doi:10.48550/arXiv.1709.07871

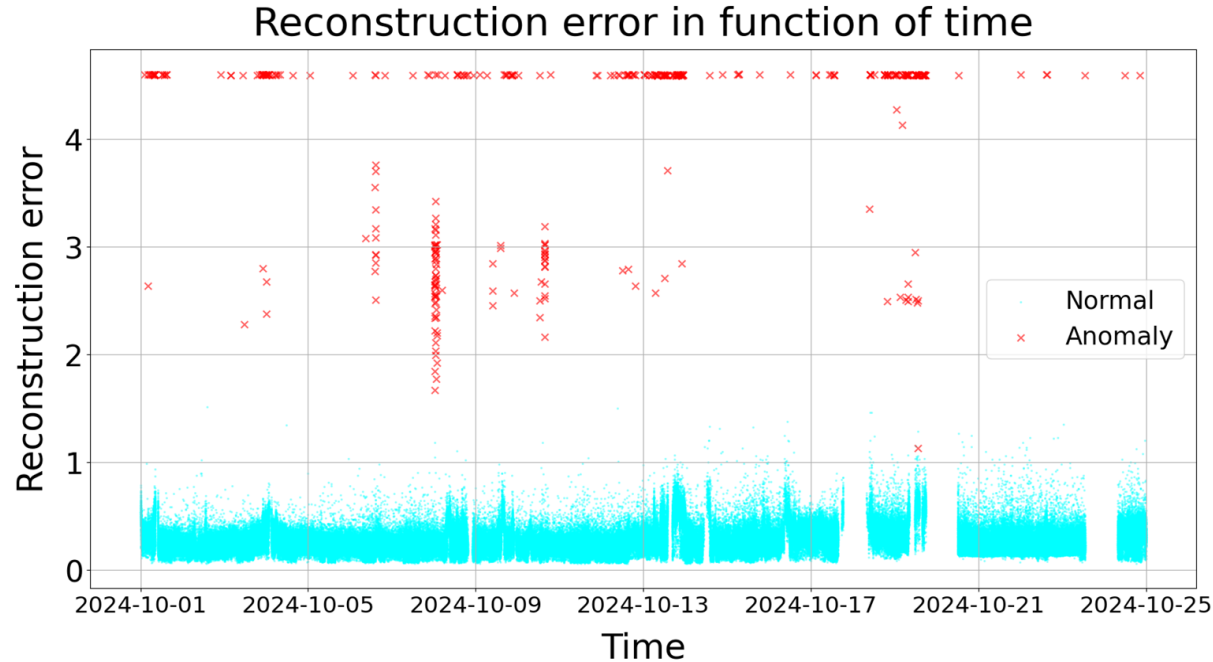
Results – Reconstruction errors

- **High Reconstruction Errors:** Observed in some data.



Results – Reconstruction errors

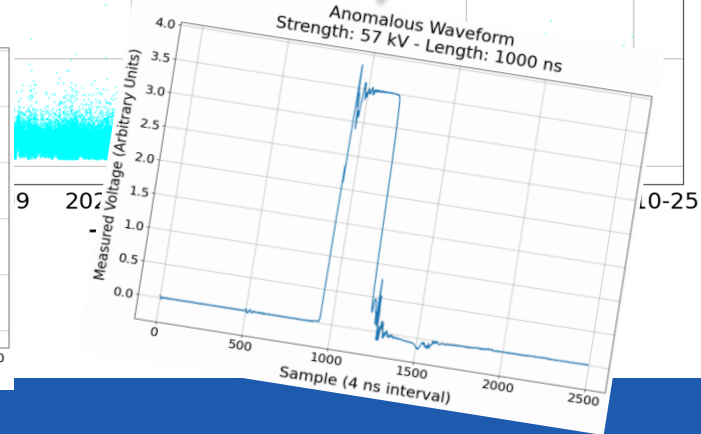
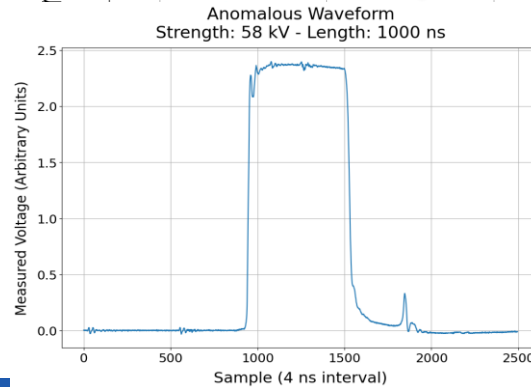
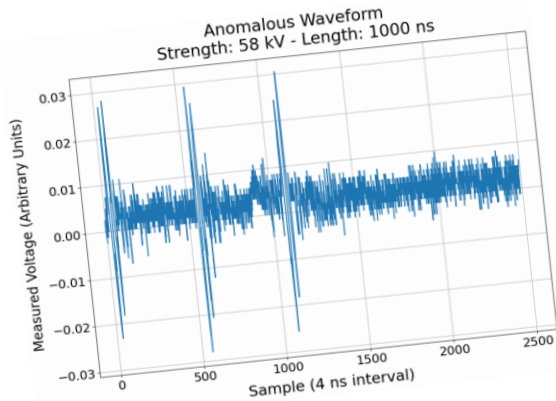
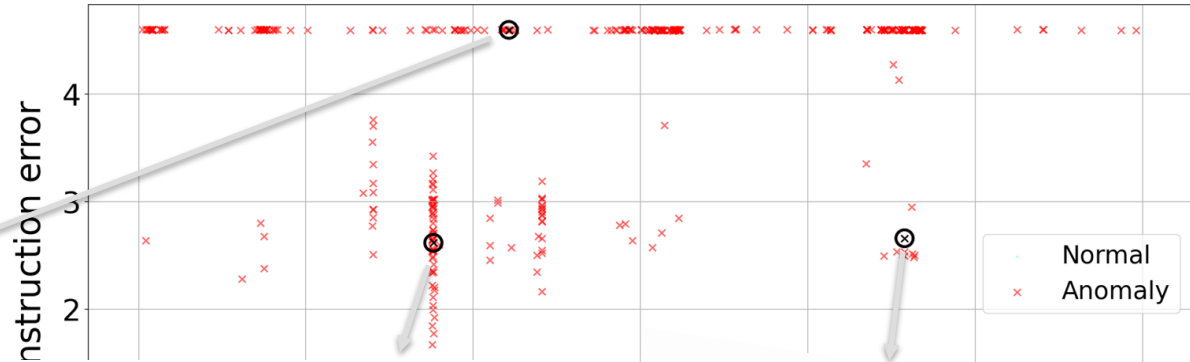
- **High Reconstruction Errors:** Observed in some data.
- **Known Anomalies:** Associated with high errors.



Results – Reconstruction errors

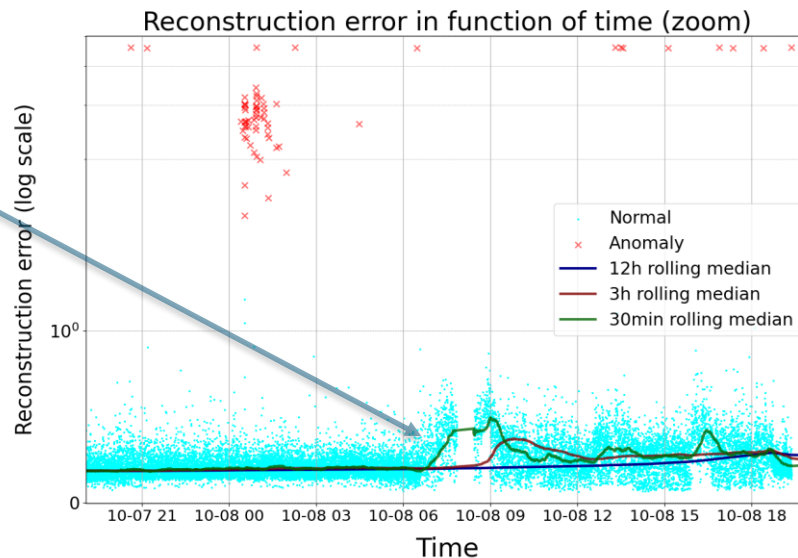
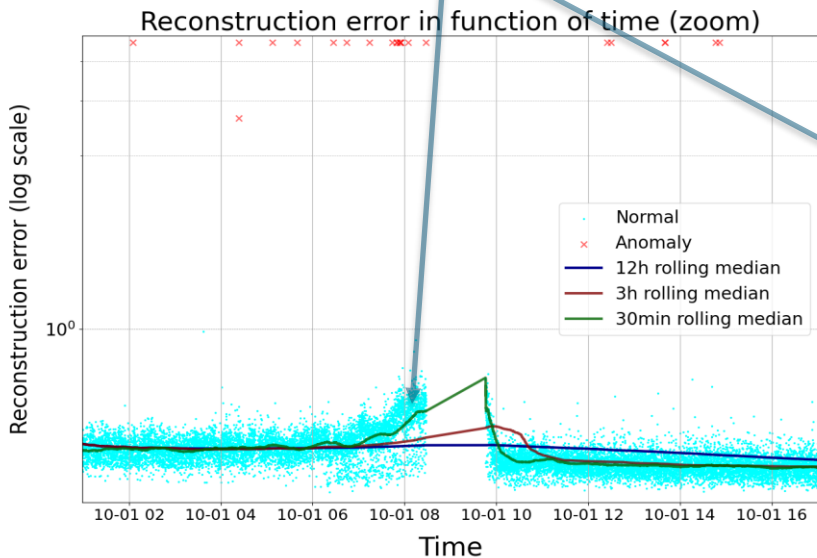
- **High Reconstruction Errors:** Observed in some data.
- **Known Anomalies:** Associated with high errors.

Reconstruction error in function of time



Preliminary Results – Anomaly Forecasting

Observation: Gradual rise in reconstruction error before failures.

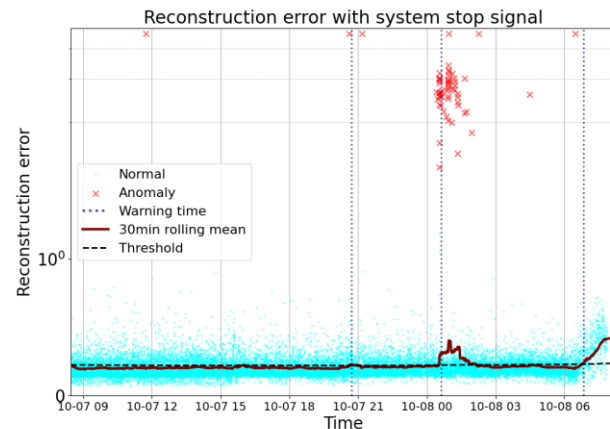
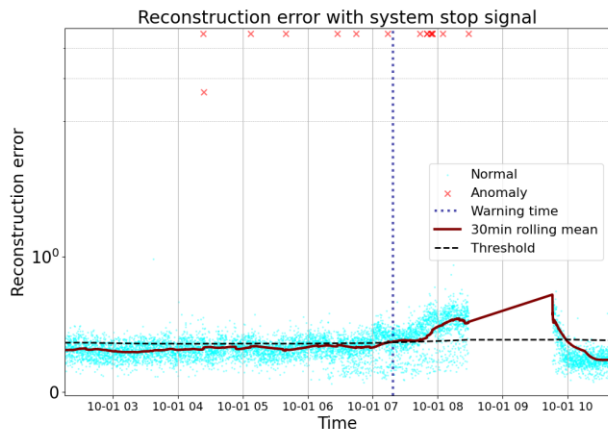


Next Step
Threshold-based methods for *Anomaly Forecasting*

Preliminary Results – Anomaly Forecasting

Threshold-based methods:

- **Risk:** 30 min rolling mean of reconstruction error
 - *Mean gives more weight to anomalies compared to median for risk calculation.*
- **Threshold:** 12h Rolling median + Rolling median absolute value (MAD).
 - *Median and MAD is chosen for a more stable threshold.*
- **Condition:** If *Risk > Threshold* for 5 min continuously → **Warning**.
- **Visualization:** Vertical dash lines show the time warnings occur.



Continual Learning

What is it?

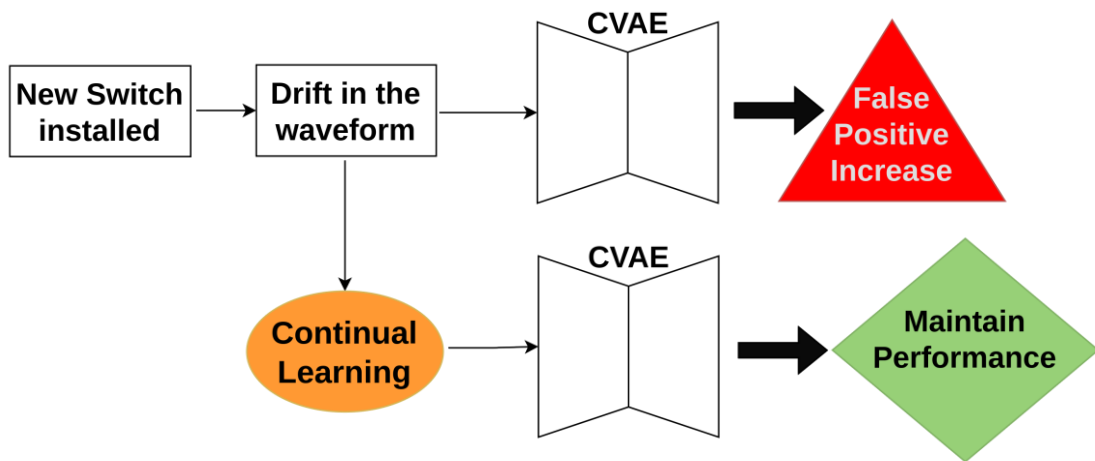
Methods for models to adapt to new data while retaining past knowledge.

Why does it matter:

- **Operational Shifts:** Upgrades, different beam types, new hardware components, varying high-voltage settings across years.

Key Requirements:

- **Adaptation:** Ongoing training with new waveforms while preserving knowledge of previously seen modes.
 - **Stability vs. Plasticity:** Avoid *catastrophic forgetting* while still learning new patterns and normal states.
- **Drift Detection:** Distinguish *natural drift* in data (e.g., new standard operation) from true anomalies.

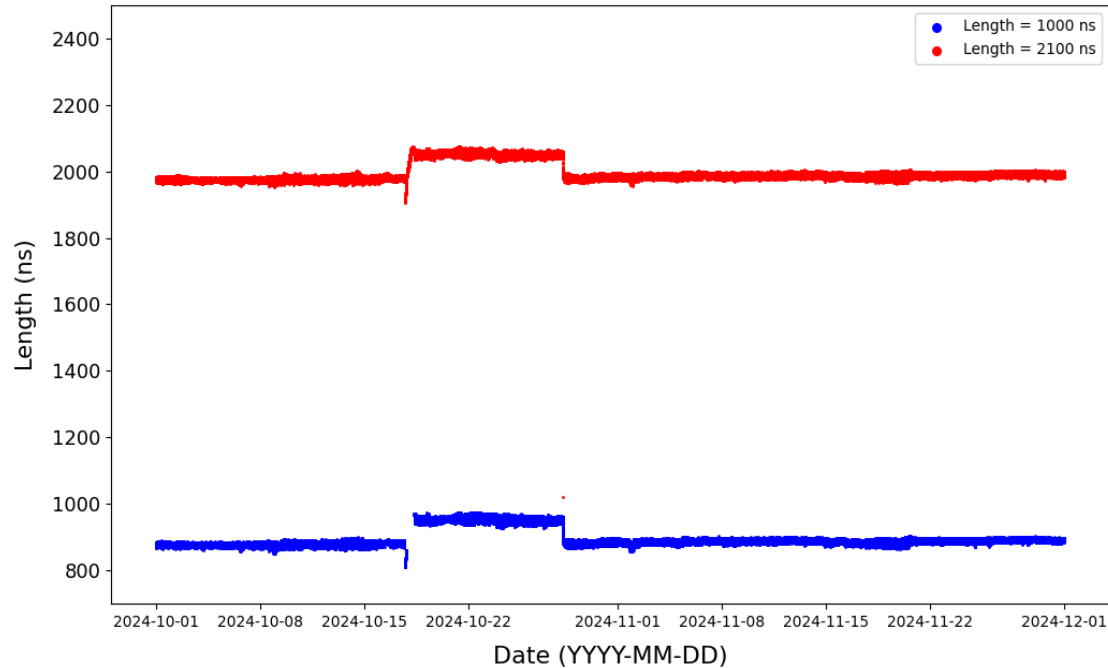


Motivation:

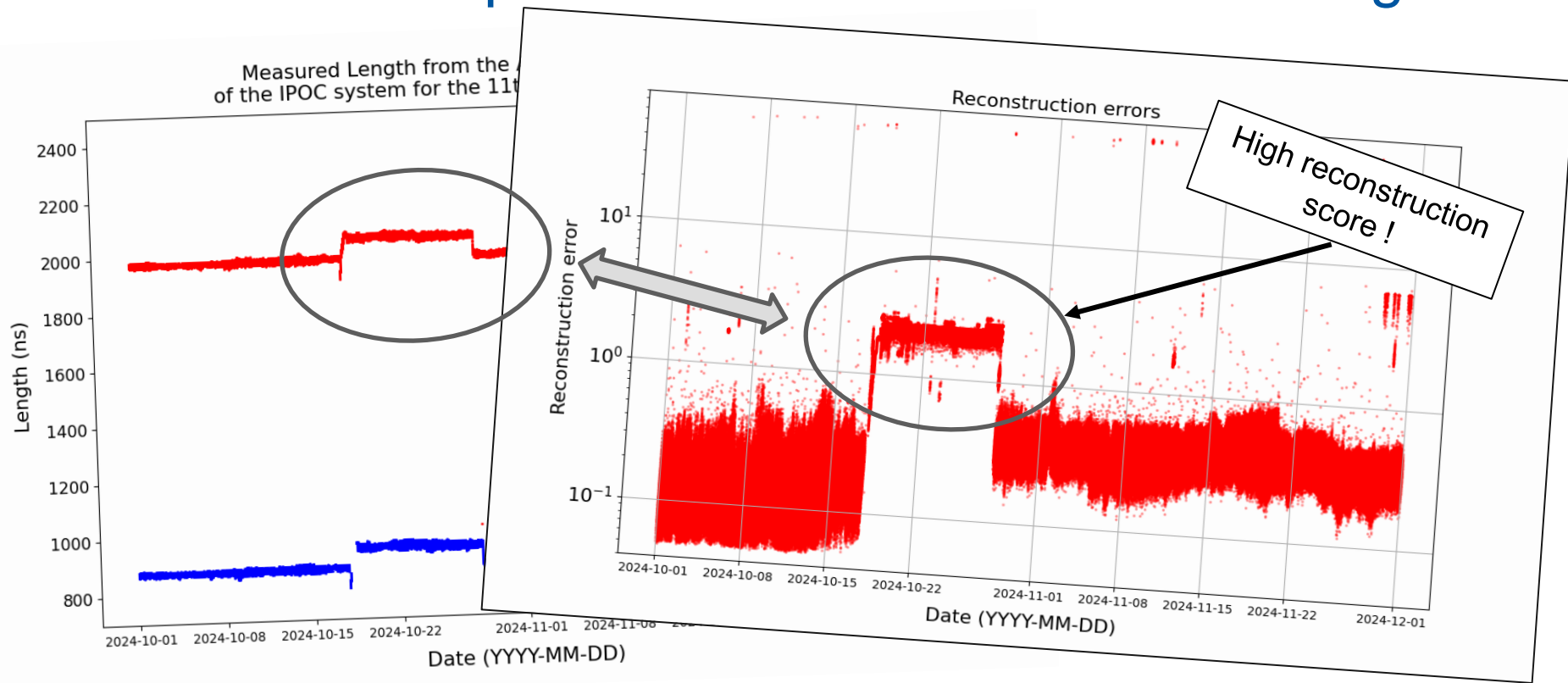
- Shift in operational condition can happen (very) often.
- Learning sequentially causes **catastrophic forgetting**.
- Retraining from scratch take **time** and resources.

Practical Example – Drift of the waveform's length

Measured Length from the AnalyserResults property
of the IPOC system for the 11th generator of the KFA71/79



Practical Example – Drift of the waveform's length



Elastic Weight Consolidation (EWC)[1]

Core idea: Prevent **important parameters** of the model from changing too much while still allowing the model to learn from new data.

How: Add a regularization term in the loss.

$$\mathcal{L}(\theta) = \underbrace{\mathcal{L}_{\text{new}}(\theta)}_{\text{Loss on new data}}$$

$$+ \frac{\lambda}{2} \sum_i \underbrace{F_i}_{\text{Fisher}} \cdot \underbrace{(\theta_i - \theta_i^*)^2}_{\text{Change from previous value}}$$

Fisher Information Matrix:

$$F(\theta) = \mathbb{E}_{x \sim p(x|\theta)} [\nabla_{\theta} \log p(x|\theta) \cdot \nabla_{\theta} \log p(x|\theta)^{\top}]$$

$$\Rightarrow F_{ij}(\theta) = \mathbb{E}_{x \sim p(x|\theta)} \left[\frac{\partial \log p(x|\theta)}{\partial \theta_i} \cdot \frac{\partial \log p(x|\theta)}{\partial \theta_j} \right]$$

$$\Rightarrow F_i(\theta) = \mathbb{E}_{x \sim p(x|\theta)} \left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta_i} \right)^2 \right]$$

$$\Rightarrow F_i(\theta) \approx \frac{1}{N} \sum_{n=1}^N \left(\frac{\partial \log p(x_n|\theta)}{\partial \theta_i} \right)^2$$

$$\Rightarrow F_i(\theta) \approx \frac{1}{N} \sum_{n=1}^N \left(\frac{\partial \mathcal{L}_{\text{MSE}}(x_n, \hat{x}_n)}{\partial \theta_i} \right)^2$$

Gaussian reconstruction assumption:

Training with MSE implies:

$$x \sim \mathcal{N}(\hat{x}, \sigma^2 I) \Rightarrow p(x | \theta) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|x - \hat{x}\|^2\right)$$

i.e., the data is assumed to follow a Gaussian distribution centered around the model output \hat{x} [2].

$$\text{So: } \log p(x | \theta) = -\frac{1}{2\sigma^2} \|x - \hat{x}\|^2 + \text{const}$$

$$\Rightarrow \frac{\partial \log p(x | \theta)}{\partial \theta_i} \propto -\frac{\partial \mathcal{L}_{\text{MSE}}}{\partial \theta_i}$$

$$\Rightarrow \left(\frac{\partial \log p(x | \theta)}{\partial \theta_i} \right)^2 \propto \left(\frac{\partial \mathcal{L}_{\text{MSE}}}{\partial \theta_i} \right)^2$$

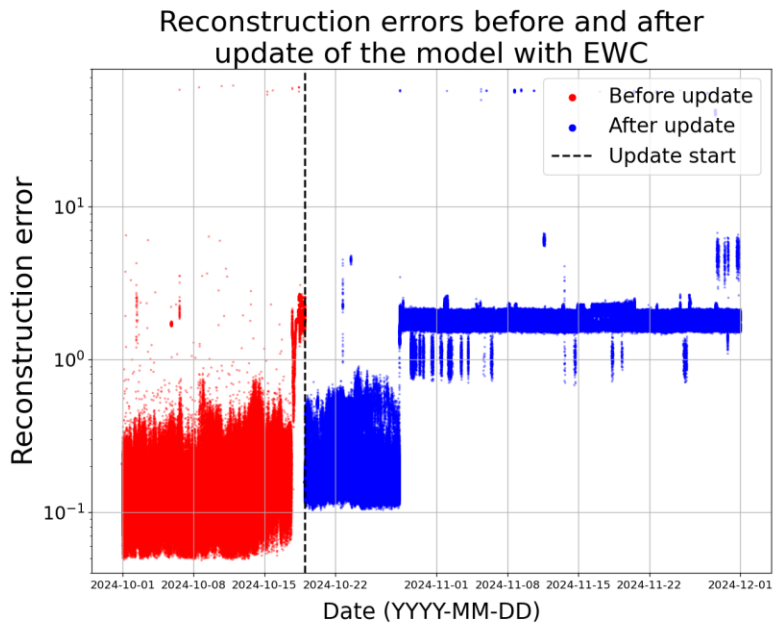


[1] Kirkpatrick et al. (PNAS, 2017), *Overcoming catastrophic forgetting in neural networks*.

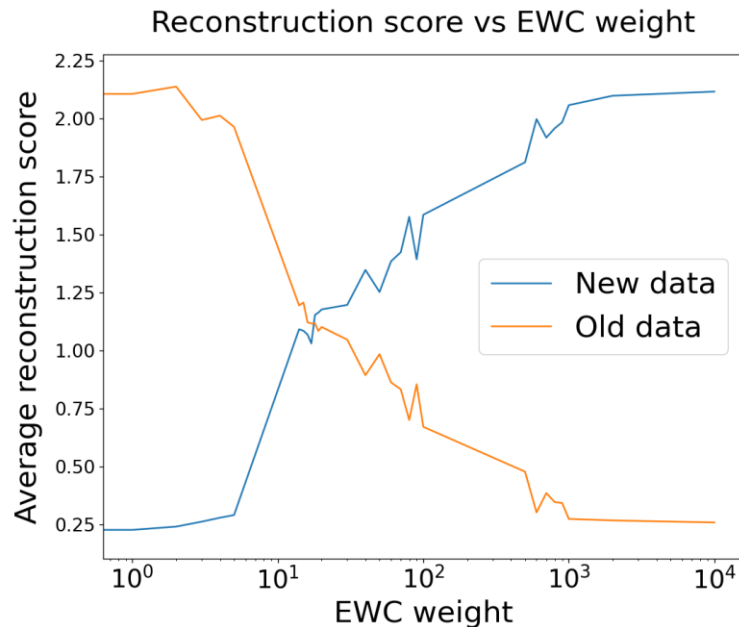
[2] Rybkin et al., *Simple and Effective VAE Training with Calibrated Decoders*, ICML 2021.

Practical Example – Drift of the waveform's length

No EWC

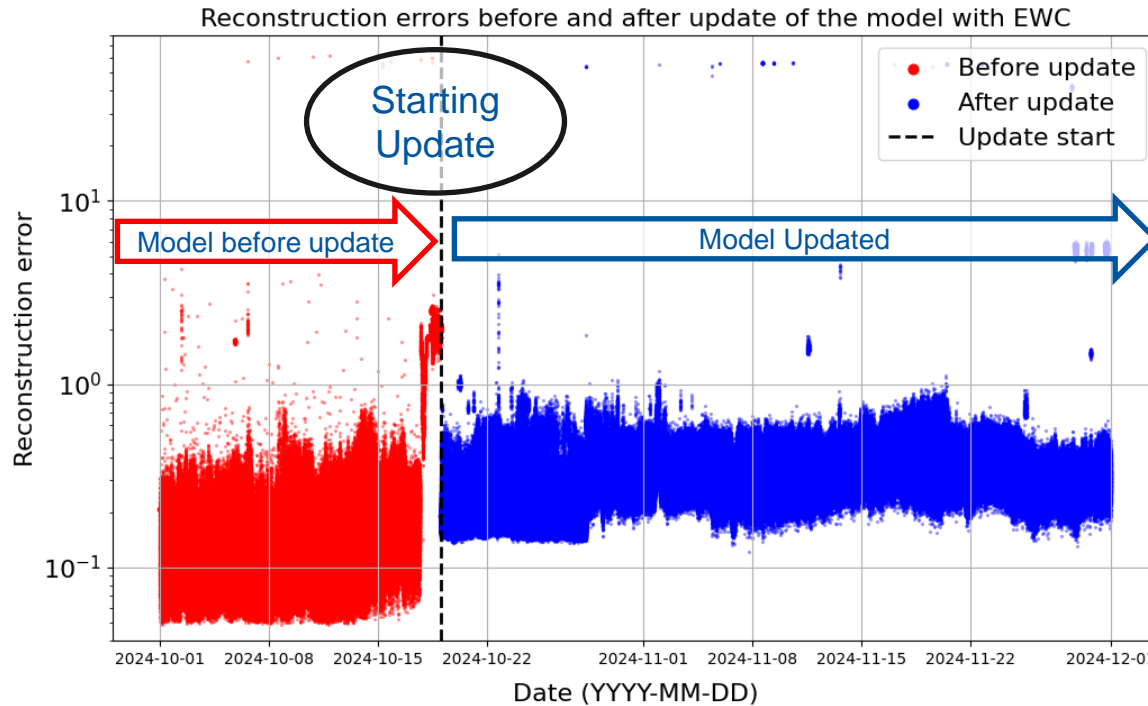


Only EWC with no Replay – Stability vs Plasticity



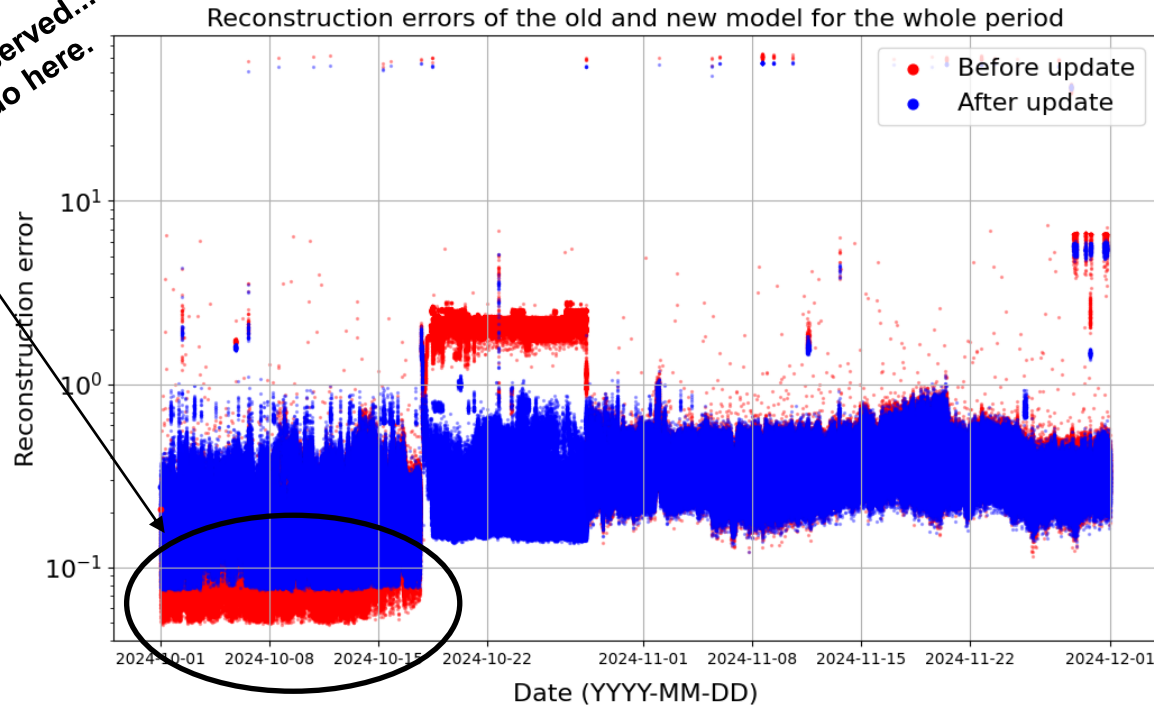
Practical Example – Drift of the waveform's length

EWC with Replay !



Practical Example – Drift of the waveform's length

Old data isn't fully preserved...
Some work left to do here.



Practical Example – Experimental Summary

Methods Compared:

- **No EWC**
 - ✗ **Problem:** *catastrophic forgetting*
→ Model quickly loses past knowledge.
- **EWC only:**
 - ✗ **Problem:** *stability vs plasticity trade-off*
→ Learns a blurry mix of old and new tasks.
This still needs deeper investigation.
- **EWC + Replay[1]:**
 - ✓ **Effective hybrid approach**
→ Retains past knowledge (EWC)
→ Adapts to new data (Replay)
→ Easy to apply in our setup (NXGALS)

Training Strategy & Timing

Phase	Data	Epochs	Time (~)
Initial training	~50,000 samples	100+	~40 min
Update (EWC + Replay)	~500 new samples + ~300 old samples	~50	~5 min

Next steps

- Reduce required sample count.
- Try **Replay** technique alone.
- Try **multiple sequential updates**.
- Explore other **CL techniques and model**.

Conclusions & Outlook

Conclusion

- **CVAE + FiLM: promising for diverse waveform reconstruction**
 - Captures waveform diversity, stays robust on rare settings.
- **Reconstruction-based anomaly score is useful for anomaly detection and forecasting**
 - Matches known anomalies, enables early failure forecasting.
- **Continual learning is a practical necessity in operation**
 - Handles drift & updates.
 - EWC + Replay = simple & effective with NXCALS.

Outlook

- **Model & Technique Refinement**
 - Tune CVAE, FiLM, EWC for better robustness and rare-case handling.
- **Forecasting Improvements**
 - Try ARIMA, LSTM
 - Add confidence/uncertainty estimation
- **Continual Learning Strategies**
 - Test long-term + multi-step updates
 - Try VCL and other CL methods
- **Operational Testing**
 - Deploy in operation (UCAP)
 - Run A/B tests on thresholds
- **Broader Application**
 - Apply to other systems using similar waveform data