Anomaly Detection on the ATLAS TDAQ Network: Hand-Built Models versus Modern Attention Models

M. Phiri[1] S. H. Connell[1] M. E. Pozo Astigarraga[2]

mitchell.phiri@cern.ch

Proposal defence, 15 October 2026

[1]Faculty of Engineering and the Built Environment
University of Johannesburg

[2]ATLAS Trigger and Data Acquisition
CERN

Problem

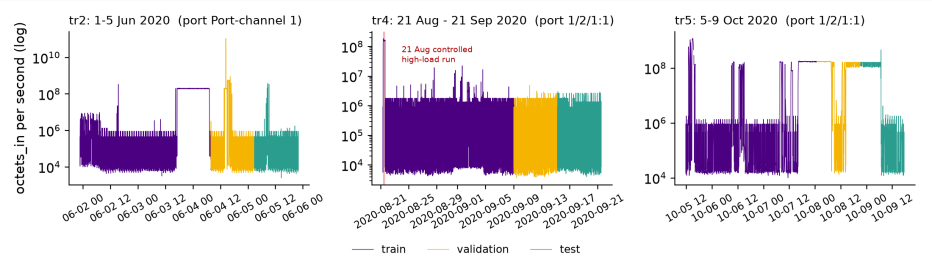
- ATLAS data-acquisition network at CERN: about 150 read-out nodes, 2000+ farm servers, 10 000+ monitored interfaces
- Telemetry: SNMP interface counters per port; no anomaly labels
- Deployed alerting is threshold-based and raises many false alarms
- **Question:** which detectors give the best balance of detections, false alarms and cost, and do modern attention models beat the hand-built (pre-attention) model families?

Source: ATLAS Collaboration, Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System, CERN-LHCC-2017-020 (2017).

Literature and gap

- Methods: thresholds, PCA, isolation forest; spectral (DMD, HAVOK); HTM; AE, VAE, VQ-VAE, LSTM; Transformers; pretrained time-series models
- Public benchmark practice: NAB-style windowed scoring
- Gap: no like-for-like comparison of pre- and post-attention models on unlabelled network telemetry with a fixed protocol and cost reported

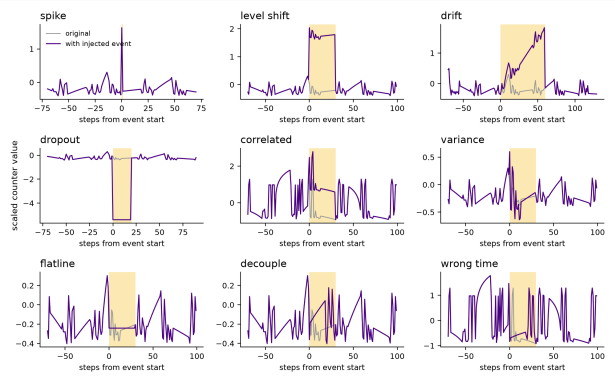
The data: three capture periods, one switch per period



Period	Dates (2020)	Native rate	Ports used
tr2	1–5 Jun	30 s	4
tr4	21 Aug–21 Sep	10 s	5
tr5	5–9 Oct	10 s	4

- Features: octets and packets, in and out, per port; all periods resampled to 60 s
- Chronological split 60/20/20, scaling from training data only
- Long high-load stretches appear in different segments (e.g. the 21 Aug controlled run)

No labels: synthetic anomalies with known positions



- **Why not real events?** Natural events cannot rank detectors: on tr1 every detector flags the same regime change; two real faults on tr2: one missed by all 14 detectors, the other caught by 4
- Injection gives the same events, at known places, for every model; it also gives false alarms and tuning
- Five amplitude types (first five panels) and four contextual types that stay in the normal range (last four); magnitude 1–3 training standard deviations
- Limit: events are designed, not observed

Method: 31 detectors, one protocol

- Models in five families: baselines (static threshold, EWMA as deployed, PCA, isolation forest) and spectral (FFT, DMD, HAVOK); hand-built neural (AE, VAE, VQ-VAE, hierarchical VQ-VAE, LSTM-AE) and HTM; attention (original recipe, my Time2Vec Transformer, eight modern variants); pretrained zero-shot (Chronos $\times 3$, TimesFM, Prophet, Moirai, MOMENT)
- Same data, split, scaling, injected events and threshold rule for every model: alarm above the 99.5% quantile of validation scores
- Metrics: event recall (an alarm during the event or in the 10 steps after it), false alarms per 1000 steps, NAB-style score; cost: parameters, fit time, scoring time
- Same tuning budget (12 trials) per neural model; three training seeds, three injection seeds

Results: recall on injected amplitude events (magnitude 2)

Model	tr4	tr5	tr2
Baselines			
Static threshold	0.72	0.60	0.54
EWMA (as deployed)	0.89	0.63	0.15
PCA	0.72	0.61	0.09
Isolation forest	0.17	0.25	0.12
Spectral			
FFT signature	0.45	0.17	0.01
DMD	0.75	0.45	0.09
HAVOK	0.45	0.29	0.02
Hand-built neural			
AE	0.98	0.86	0.22
VAE	0.98	0.86	0.10
VQ-VAE	0.74	0.42	0.16
VQ-VAE (EMA)	0.84	0.69	0.23
Hierarchical VQ-VAE	0.88	0.77	0.25
LSTM-AE	0.95	0.37	0.07
HTM	0.73†	0.37†	–

Model	tr4	tr5	tr2
Attention, earlier			
Transformer AE (original recipe)	0.97	0.85	0.69
Time2Vec Transformer (Oct 2020)	0.78	0.41	0.14
Attention, modern			
RoPE+RMSNorm+SwiGLU	0.97	0.96	0.63
+ gated attention	0.98	0.95	0.87
+ gated + top-k	0.99	0.95	0.87
+ mixture of experts	**0.99**	0.96	0.89
+ sparse latent	0.99	0.95	0.92
all four combined	0.99	**0.97**	**0.93**
causal forecaster	0.83	0.67	0.10
causal + multi-token	0.87	0.73	0.08
Pretrained, zero-shot			
Chronos-Bolt tiny	0.61	0.94	0.77
Chronos-Bolt small	0.63†	–	–
Chronos-Bolt base	0.67†	–	–
TimesFM	0.55†	–	–
Prophet	0.73†	–	–
Moirai 2	0.44†	0.83†	–
MOMENT	0.80†	0.36†	–

Mean over ports and seeds; five amplitude event types; threshold from validation. Bold = best in column. † fewer ports than the rest (zero-shot, HTM, MOMENT: 2–3). – not run. These metrics cover amplitude (spatial) anomalies only.

Can the ranking be trusted? Sensitivity of the evaluation design

Design change	Rank correlation
Threshold quantile 0.99 / 0.999	0.98 / 0.97
Detection window 5 / 20 steps	1.00 / 1.00
Event length $\times 0.5$ / $\times 2$	1.00 / 0.98
Threshold from train+val	0.90
Magnitude 1 / 3	0.92 / 0.90

- Spearman correlation with the base ranking of nine models
- The best model stays first in every variation
- Absolute recall moves most with event magnitude, so only the ordering is robust

Cost next to accuracy (tr4)

Model	Parameters	Fit (s)	Score (ms/1k)	Recall	FA/1000
Static threshold	–	<0.1	0.1	0.72	2.0
PCA	–	<0.1	1.7	0.72	1.8
AE	21 k	4	3.4	0.98	3.9
LSTM-AE	40 k	14	8.9	0.95	4.4
Transformer AE (original)	26 k	32	12.4	0.97	1.2
Modern Transformer, plain	83 k	19	6.7	0.97	0.8
Modern Transformer, all four	42 k	96	13.6	0.99	2.0
Time2Vec Transformer (Oct 2020)	260 k	497	184.4	0.78	2.5
Chronos-Bolt tiny (zero-shot)	8.7 M	none	297.8	0.61	5.1
MOMENT (zero-shot)†	35.3 M	none	190.0	0.80	6.5

Parameters = learned weights. Fit = seconds to train on one port (about 27 000 one-minute steps, one RTX 2080 Ti, indicative). Score = milliseconds per 1000 steps. Recall and false alarms (FA) as before; † 3 ports.

- The modern Transformer with all four ingredients is the most accurate (0.99) with 42 k parameters; the plain modern variant trains fastest among the attention models and has the fewest false alarms in the table
- My earlier Time2Vec Transformer has 3–6 times more parameters than these two, trains 5–26 times slower and scores lower (0.78)
- Zero-shot models need no training but are large and slow to score

Gaps and limits

- **Gap:** anomalies that stay in the normal range are mostly missed. The top amplitude model (modern Transformer, all four) scores: wrong-time 0.05, flat-line 0.27, variance 0.51, decoupled counters 0.58; the best single model on wrong-time reaches 0.38
- A small dedicated flat-line monitor lifts flat-line recall from 0.27 to 0.92 (+0.5 false alarms per 1000); the other monitors are weak and need better designs
- Models trained on Aug–Sep 2020 keep their recall on Nov 2019 data (0.98), but raise 4–7 times more alarms on clean data
- **tr2:** on this period several models alarm too rarely (AE recall 0.22) although they still rank events well (AUC 0.91–0.95). The alarm threshold is set on a validation day that looks different from the rest of the data. Results depend on whether the training part includes the high-load stretch (modern Transformer: recall 0.10 without, 0.95 with); the cause is still being tested, so tr2 is reported separately
- **Limits:** all events are synthetic; data are LS2 plus controlled load runs, no real production data taking; three seeds
- Next: better monitors, real data-taking runs from CERN, significance tests

Summary

- A fair, reproducible comparison of 31 detectors on the ATLAS TDAQ network, with cost reported
- Early findings: modern attention models lead on amplitude events, the ranking is robust to protocol choices, and contextual anomalies need dedicated monitors
- Questions