Reading Machines from the Meter

Machine condition, state and cycles from standard-rate electrical telemetry, inscribed on a CPU

Sekos AI Research · September 2026 · Preprint

`CPU only · no GPU` `closed-form readers · no neural training` `8 data sources` `reports every 1–10 s` `preregistered · SHA-256 sealed, adaptive`

Abstract

Most industrial machines already sit behind a meter or a clip-on current sensor that reports averaged power every 1 to 10 seconds. Commercial machine-monitoring products turn that signal into run, idle and off states, cycle counts, utilization and downtime. We ask what else such telemetry carries and how to read it without collecting large labelled datasets or training networks. We treat each operating period of a machine as a *phrase of work*, discover the machine's states and phrases from resolution rules read off the instrument itself, and read conditions with classifiers fitted in closed form, which we call *inscription*: class means and a shrinkage covariance are estimated from labelled phrases in one pass and new phrases are read by whitened distance. This is shrinkage linear discriminant analysis with group-balanced means. It is still learned from data, but no neural network is trained, no iterative optimisation is run for these readers, and no GPU is used; every result in this paper was computed on a laptop CPU, and fitting a reader takes milliseconds.

Experiments were preregistered in files sealed by SHA-256 before they ran (without an external timestamp, and with some evaluation folds reused across successive preregistrations, so this is transparently documented adaptive development rather than confirmatory testing); analyses added after a result was seen are labelled post hoc; failed gates are reported; and internal audits of the manuscript and one external review were incorporated. Across eight data sources (two simulators, the PLAID appliance corpus, the UCI hydraulic test rig, a metro air compressor, a milling machine, household appliances from a repair centre, and a centrifugal pump test bed):

- **Operations metrics.** Against valve ground truth on a metro compressor (1.5 million reports over 7 months), a label-free threshold baseline, applied after our preregistered state readers fell short, identifies loaded running versus idle at 0.9955 accuracy (our own state-discovery reader reaches 0.962; k-means 0.994) with 0.031 false state changes per hour, counts load cycles within 1.4 % per day, estimates daily utilization within 0.18 percentage points (median) and finds 99.6 % of stops longer than five minutes. Stops shorter than a minute are not recoverable from 10 s reports by any reader we tried.
- **Condition.** From the motor's electrical power alone, linear discriminant reads identify valve condition on the hydraulic rig at 0.979 to 0.999 under time-blocked, forward-only and held-out-combination evaluation (9.7 points above nested-tuned gradient-boosted trees, time-blocked); retrospectively, with each failure held out, they recognise the failure phrases of all four operator-reported air-leak failures of the compressor (95 % lower bound on recall 0.40 for four events) with at most 0.9 % false positives in the preceding week,

where tree ensembles at their default threshold flag none. This is recognition of completed phrases, not a validated early warning; and they detect 58 % of fridge malfunctions at 0.994 specificity (tree ensembles 38–42 %).

- **Against the state of the art.** On the same held-out splits, the strongest open time-series classifiers (MiniRocket, MultiRocket-Hydra, QUANT, HIVE-COTE 2) match the inscribed reads on valve lag (0.991–1.000; preregistered comparison p = 0.75; equivalent within ±2 points for three of them and ±2.5 for QUANT, margins chosen post hoc) at roughly 100 to 70 000 times the CPU time of the classification step, catch at most half of the compressor's failure phrases and fewer fridge malfunctions, score below our reads on appliance identification in unseen homes (0.70–0.78 against 0.89), and read the accumulator better than any read of our own phrase features (on pump leakage they beat our whitened read, while shrinkage LDA on our features is level with them). Their advantage is their feature map. Inscribing our read on those maps, still in one closed-form pass, beat the maps' own heads on the rig (accumulator 0.571 against 0.513, preregistered, p = 0.019; up to 0.746 and 0.862 on accumulator and pump leakage, the best time-blocked results of any method here) and read accumulator pre-charge above a circular-shift null (p = 0.0099), which no read of our own features did. It did not help on tool wear.
- **Limits.** Averaged telemetry reads a condition when its change in average power or cycle shape exceeds the machine's normal run-to-run variation. Load-changing faults do (air leaks, valve lag, impeller damage, cavitation, tool wear). Friction faults (bearings, misalignment, unbalance) raise power by about one percent, inside a healthy machine's 1.3 to 2.2 % session-to-session spread, and were not reliably detectable.
- **Evaluation.** On cyclic machine data, random resampling inflates condition accuracy by up to 59 points relative to time-blocked evaluation.

---

1. Introduction

A clip-on current transformer or a revenue meter is the cheapest sensor a factory can add. Energy monitors in this class report averaged quantities (active and reactive power, current, power factor) every one to ten seconds. They keep no waveform. The machine-monitoring products built on them report a common set of operations metrics: whether a machine is running, idle or off; how many cycles it completed and how long each took; its utilization; its downtime events; and, increasingly, alerts about its condition. Published performance claims in this category are rare. Where one exists it concerns run-versus-idle detection (accuracy above 98 % with "a false positive rate below 0.5 % per hour" after threshold calibration); condition alerts are described qualitatively, and downtime reasons are typically entered by operators.

Two questions follow. First, how much of a machine's operational state is present in such telemetry at all, and how accurately can the standard operations metrics be recovered? Second, can the knowledge of the people who run a machine (who label what it was doing after the fact, in their own words, on a fraction of the time) be used without training models that need GPUs and large labelled sets?

We organise the answer around the **phrase of work**: the operating period that carries a machine's state (a load cycle, a program run, a compressor on-period, a turn-on sequence).

Four observations drive the methods:

1. Every resolution choice a state-discovery pipeline makes (how close two readings must be to be the same state) can be read off the instrument's own temporal residual, so no threshold needs tuning.
2. A condition that changes a machine's dynamics is a change of poles; a class of phrases shares its poles, and the order of the dynamics that is observable at a given reporting rate and noise can be estimated before any classifier is fitted.
3. A classifier can be fitted in closed form rather than trained iteratively: class means and a shrinkage metric are written into a Frame in one pass (we call this *inscription*), and a new phrase is read by its whitened distance to each class.
4. A state decoder should report how its certainty grows with look-back horizon; the horizon at which its belief collapses depends on how time is phrased.

Contributions

- A label-free phrase-discovery pipeline for standard-rate telemetry whose parameters come from the instrument's noise law, dwell law and averaging law, validated on a known universe and tested on real machines (§2.2, §4.5).
- Inscribed whitened class reads (closed form), class Frames for dynamics, and a tensor-calculus lift that converts a CPU-trained tree ensemble into an explicit lookup tensor with one ridge solve (§2.3–2.5).
- Observability diagnostics (Hankel rank against a simulation-calibrated noise edge, Cramér–Rao) that indicate whether a condition is likely to be readable at a reporting rate (§2.6, §4.6).
- A horizon-collapsing state decoder (§2.7, §4.4).
- A preregistered evaluation across eight data sources with random, time-blocked, forward-only, held-out-combination and held-out-unit protocols, the operations metrics used by the monitoring industry, and every failed gate (§3–5).
- A preregistered comparison with the open state of the art in time-series classification on the same held-out splits, and a feature-map lift that holds a comparator's fitted feature map in a Frame and reads it by inscription, computed exactly in the dual (§2.9, §4.9, §4.10).

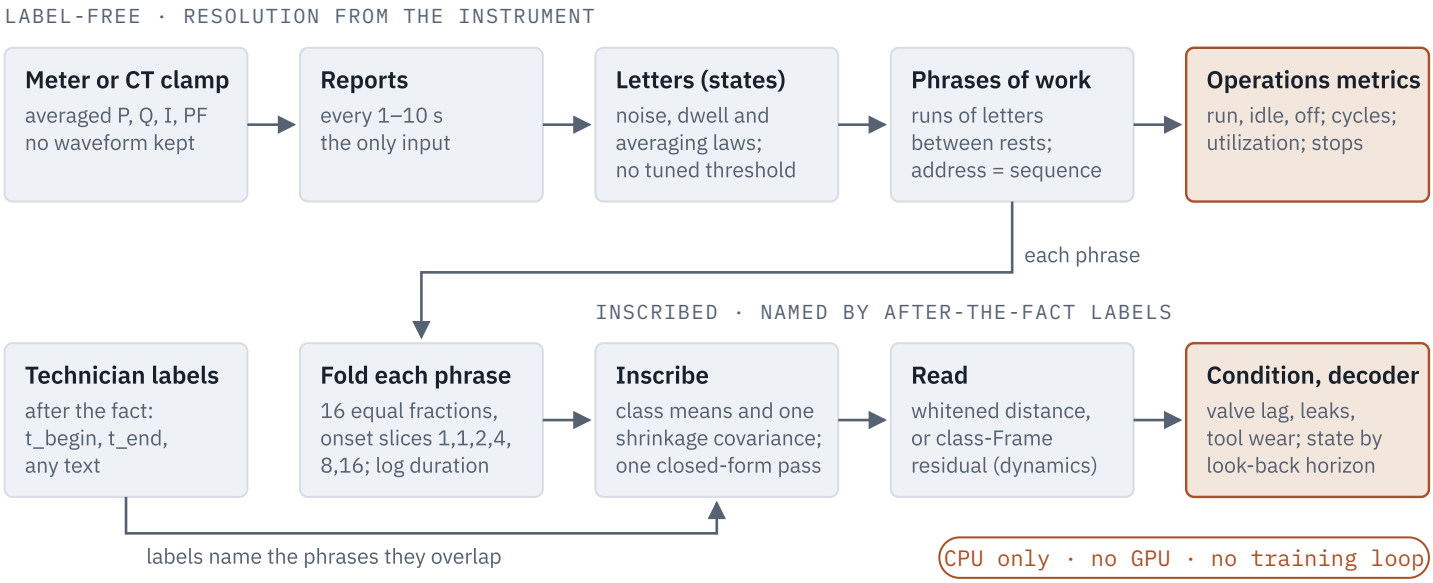

Figure 1. How a machine is read from its meter. The top row needs no labels: states (letters) and phrases of work are found with resolution rules taken from the instrument's own noise, and give the standard operations metrics. The bottom row uses labels a technician writes after the fact: each phrase is folded to a fixed-length vector, the classes are inscribed in one closed-form pass, and new phrases are read by whitened distance. No step uses a GPU.

Source: §2; code in machine_phrase_v1/pipeline.py, hydraulic_phrase_v1/hyd.py

2. Methods

2.1 Phrases of work

A *report* is one averaged reading from the meter. A *phrase* is the operating period that carries a state: a load cycle of the hydraulic rig (60 s), a compressor on-period, a milling cycle, a washing-machine or fridge cycle, a 15 s pump snapshot. A phrase is folded into a fixed-length vector per channel: means over 16 equal fractions of the phrase (duration-invariant shape), means over onset slices of 1, 1, 2, 4, 8 and 16 reports (the transient, on a logarithmic clock), and the log duration.

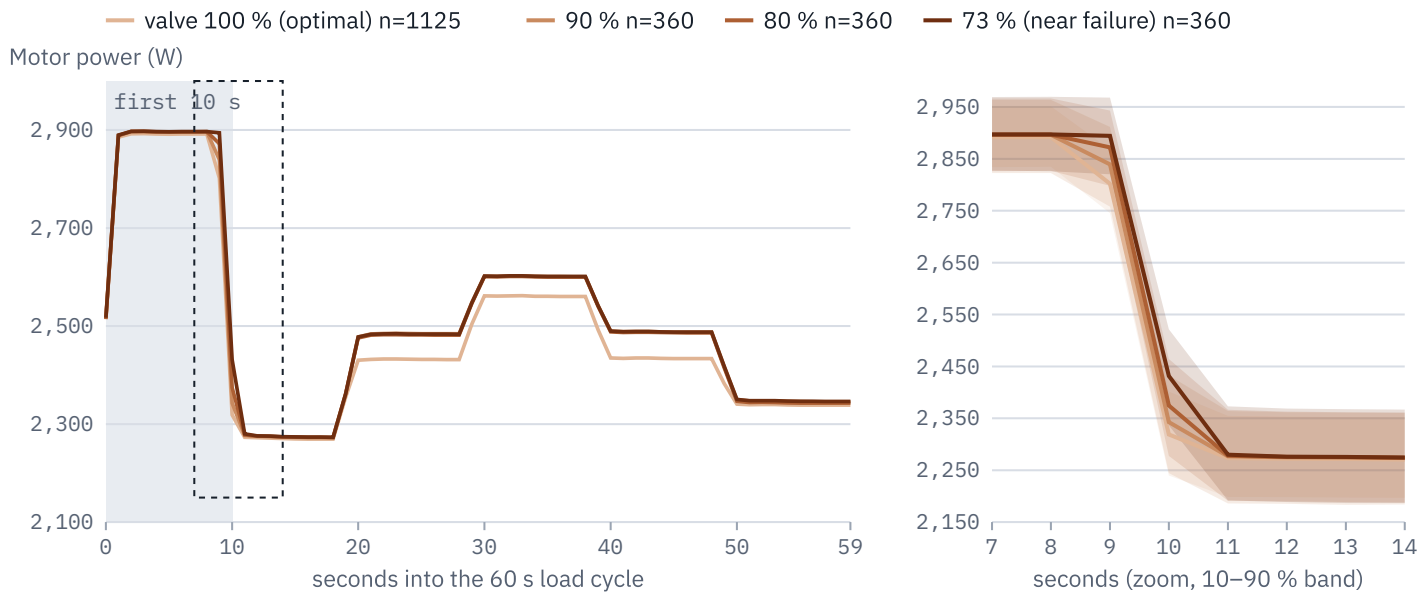

Figure 2. One phrase of work: the hydraulic rig's 60 s load cycle as its motor's electrical power, averaged to one report per second, mean over all cycles at each valve condition. The four conditions are the same machine doing the same work; they differ in how late the power steps down after the high-load phase (right, seconds 7–14, with the 10–90 % band). That is why the first 10 s of a cycle carry the valve state (Figure 12).

Source: hydraulic_phrase_v1/data/eps1_1hz.npy, profile.npy (figures/extract.py)

2.2 States and phrases without labels: the instrument's own resolution

Per channel, the **noise law** $\sigma(v)^2 = a^2v^2 + b^2$ (proportional noise plus a quantisation floor) is fitted from binned medians of squared third differences of consecutive reports, which cancel locally quadratic dynamics while white noise survives with a known gain ($\text{var}\,\Delta^3\varepsilon = 20\sigma^2$). The variance-stabilising coordinate $g(v) = \text{asinh}(a{\cdot}v/b)/a$ makes noise one unit at every level, so a 1 W quantum at 40 W and at 2000 W are treated alike. **Letters** (states) are density components of g: core bins hold at least 20 reports within $\rho = 2/\sqrt{\pi}$ (the expected difference of two unit-noise reports); rare states are kept. The **dwell law** keeps a component only if at least half of its reports persist for two or more reports, which removes transition samples. The **averaging law** re-reads a one-report run that lies on the mixing line of its neighbouring letters (within 3σ) as a transition. Runs of a letter are chunks; chunk sequences between rest are phrases; a phrase's address is its letter sequence. After-the-fact labels (t_begin, t_end, any string) name the discovered phrase they overlap; a discovered type is *self-named* when 95 % of its labelled instances agree, and part of a label budget can be spent querying the least certain phrases.

2.3 Inscription: closed-form fitting

A **whitened class read** stores, for each class, the mean of its phrases with each group of recordings (a run, a unit, a date) weighted once, and one Ledoit–Wolf shrinkage covariance of the within-group and between-group deviations. A new phrase is assigned to the class with the smallest Mahalanobis distance. This is shrinkage linear discriminant analysis with group-balanced means; we report plain shrinkage LDA beside it throughout. Both are single closed-form passes over the data: nothing is iterated and no gradient is taken. We call this *inscription*: the class geometry is written into a Frame once and read thereafter. Three

auxiliary components used in some experiments are small iterative CPU solves and are not inscription: the pole fits of §2.4 and §2.6 (a bounded one-dimensional search and nonlinear least squares over one to three poles), logistic-regression ablations, and the gradient-boosted tree ensemble that serves as a baseline and as the teacher of the lift (§2.5).

2.4 Class Frames for dynamics

When a condition changes a machine's dynamics, a phrase is $x_k = c_0 + \Sigma\, c_m\, z_m^k$. A class Frame is the column space of $[1, z_c^k]$ with the class pole z_c fitted jointly over the class's labelled phrases (shared pole, per-phrase level and amplitude). A phrase is read by the smallest noise-whitened residual energy over class Frames.

2.5 The tensor-calculus lift

To bring a trained tree ensemble into the same form, it is queried on a *siphon set* (training phrases, same-class interpolations across groups, cross-class probes) and one ridge regression is fitted from the phrase's per-axis bin indicators, on the ensemble's own split lattice, to its output margins. The result is an explicit lookup tensor, and its R^2 against the ensemble measures how much of the ensemble's calculus is additive per axis.

2.6 Observability diagnostics

By Kronecker's theorem, the Hankel matrix of a phrase has rank equal to the order of the minimal linear system that generates it; the steady level is the pole $z = 1$, so order = rank − 1. A mode is observable when its singular value clears the noise edge $1.5\sigma(\sqrt{L} + \sqrt{C})$. That edge is the scale of the largest singular value of an $L \times C$ matrix of independent noise; the noise entries of a Hankel matrix are not independent (each report recurs along an anti-diagonal), and the spectral norm of random Hankel matrices grows faster, by a logarithmic factor in the square case (Meckes 2007 [25]). The edge is therefore a heuristic, checked against simulations with known order in §4.6, not a bound. It bears on what our features retain; it does not show that no method could read a condition from the measurements. Poles are initialised by shift-invariance ESPRIT and refined by variable projection, and their precision is compared with the Cramér–Rao bound.

2.7 Horizon-collapsing decoder

At decision time the decoder reads a horizon h of the stream: within a phrase, a Gaussian class read restricted to the observed reports (either the first h seconds of the phrase or the last h seconds before the decision; Figure 12); across phrases, the sum of per-phrase class scores; and a forward filter with a switching prior estimated from the training schedule, which needs no horizon.

2.8 Compute

Every experiment ran on the CPU of one Apple M5 Max laptop; its integrated GPU was not used, and no code path uses a GPU. The tree-ensemble baselines (XGBoost, random forest) also ran on the CPU. Table 1 and Figure 3 give measured costs.

Table 1. CPU cost (single process). Source: `field_datasets_v1/result_ops.json`.

Operation	Data	Time
Fit state letters	215 k compressor reports (February)	16 ms
Read states	1.52 M reports (7 months)	57 ms (0.04 µs per report)
Whitened class read, fit + predict one fold	hydraulic: 1 995 phrases × 60 values	6 ms
same	milling: 953 phrases × 265 values	17 ms
same	pumps: 411 phrases × 120 values	87 ms
XGBoost baseline, same folds (incl. process start)	as above	2.4 s / 1.4 s / 5.9 s

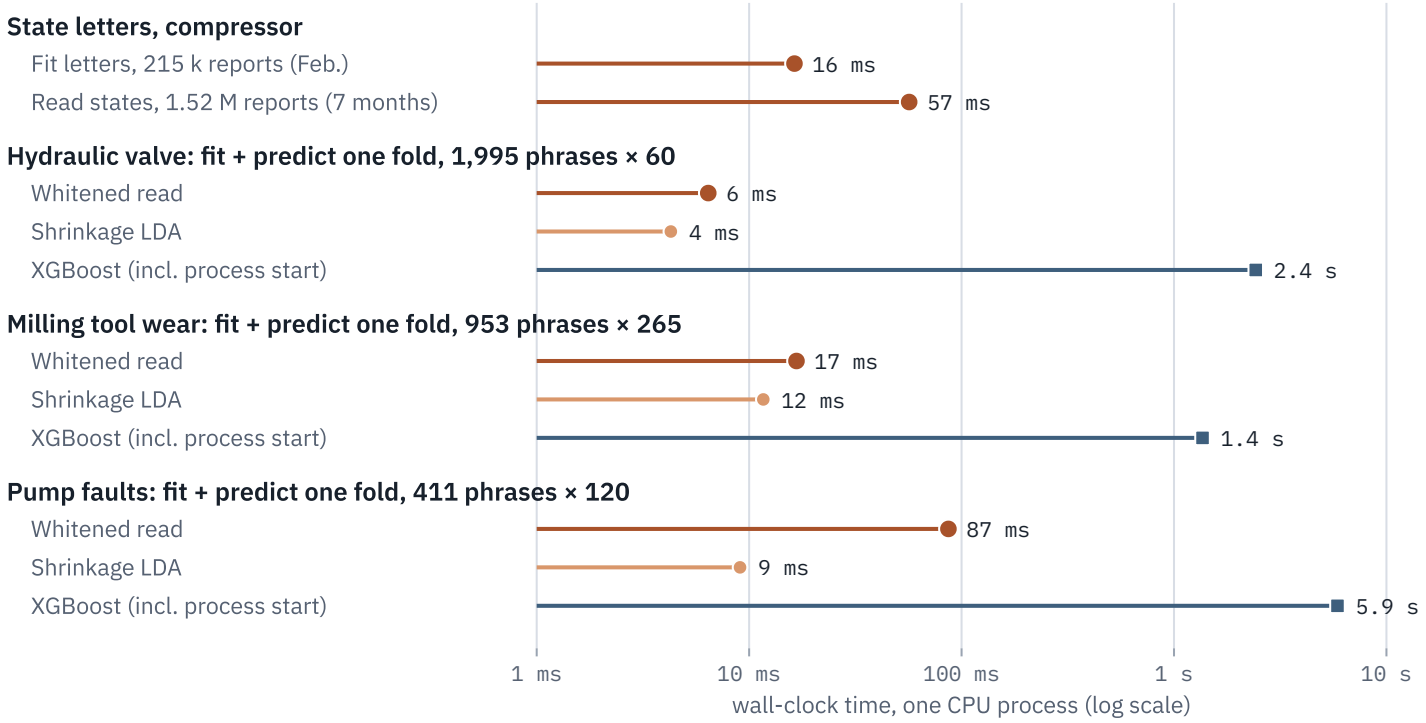

Figure 3. Measured CPU cost. Reading seven months of compressor current into states takes 57 ms (0.04 µs per report). An inscribed condition reader is fitted and applied in milliseconds; the XGBoost baseline on the same fold takes seconds (it runs in a separate process, whose start-up is included).

Source: field_datasets_v1/result_ops.json · Apple M5 Max CPU, GPU unused

2.9 Inscription on a borrowed feature map

The strongest open time-series classifiers are not deep networks trained by gradient descent. MiniRocket and MultiRocket-Hydra apply a fixed bank of random convolution kernels whose biases are set from quantiles of the training series (no labels), pool each output, and fit a closed-form ridge classifier on the pooled features; QUANT takes quantiles over dyadic intervals of the series and its derivatives and fits extra-trees. Their feature maps can therefore be held by a Frame as they are. The *feature-map lift* keeps a comparator's fitted map, and the scaler its own head sees, and replaces the head with an inscribed read: the whitened read of §2.3 unchanged (group-balanced class means and

one Ledoit–Wolf metric), or the same with plain class means. With 10^4–10^5 features and 10^3 training phrases the metric cannot be formed, so the Ledoit–Wolf shrinkage and every Mahalanobis distance are computed exactly in the dual: with deviation rows R (n × d), $\Sigma = (1 - s)R^TR/n + s\cdot\mu\cdot I$ and, by the Woodbury identity, $\Sigma^{-1} = (1/a)[I - R^T(a/b\cdot I + RR^T)^{-1}R]$ with $a = s\cdot\mu$ and $b = (1 - s)/n$; the shrinkage s itself depends on R only through the Gram matrix RR^T. The dual read agrees with the primal one (sklearn's LedoitWolf) to $1.4\cdot10^{-15}$ relative on a test problem. The read is still one closed-form pass; the map it reads is borrowed, and we say so wherever it is used.

3. Data and protocol

Table 2. Data sources.

Source	Machine and signal	Labels (how obtained)	Size
Simulated machine	6 states, 4 phrase types, 5 s reports of P and Q with 1.5 % noise and 1 W quantisation	ground truth; arbitrary 6-character technician labels	3 seeds × 10 sessions × 40 phrases (+ 3 fresh seeds)
Simulated dynamics	phrases of known order (0, 1, 3) and time constants	ground truth	3 seeds × 10 sessions × 30 phrases (+ 3 fresh seeds)
PLAID 2018 submetered [1, 2]	appliance V and I, 30 kHz, folded to turn-on phrases	appliance type	1 876 records, 16 types, 56 homes
UCI hydraulic rig [5]	motor electrical power (EPS1), averaged to 1 Hz	experimenter-set valve, pump leakage, accumulator, cooler	2 205 cycles of 60 s
MetroPT-3 [6]	compressor motor current, ≈ 10 s reports; digital valve signals	operator failure reports; valve signals as independent state truth	1.52 M reports, Feb–Aug 2020
CNC milling tool life [7]	spindle and axis current, 500 Hz, averaged to 1–10 s	cycles to tool failure	968 cycles, 14 tools
SMART-PDM [8]	washers and fridges, native 1 Hz P, Q, S, PF, I	repair-centre technician labels	96 washer cycles, 1 111 fridge cycles
Twente pumps [9]	3-phase V and I, 20 kHz, reduced to 1 s reports	installed mechanical faults	893 snapshots, 2 pumps, 65 conditions

Protocols. Each experiment was preregistered: the hypotheses, methods, metrics and gates were written and sealed by SHA-256 before any label-dependent computation. Analyses added after a result was seen are marked post hoc in the text: the threshold reader and chunk rule of Table 3, the reviewer-requested analyses on the hydraulic rig (tuned baselines, forward-only splits, ablations, circular-shift null, sensitivity grid), the duration control on the compressor, and the excess-power detector of §4.3, whose design followed a label-informed look at the pump data. A change made after a result was seen was sealed as a new file naming the failure it addressed; failed runs were kept. Held-out units are the strictest the data allow: time blocks with 10-cycle purges and forward-only splits (hydraulic rig), held-out combinations of the other faults, leave-one-home-out (PLAID), leave-one-tool-out (milling), leave-one-failure-out (compressor), leave-one-date-out and held-out repair episodes (appliances), and unseen operating speed or unseen condition (pumps). Random k-fold is reported only for comparison with prior literature.

Confidence intervals are group bootstraps (2 000 resamples); paired comparisons on the hydraulic rig use a block-level sign-flip permutation test with Holm correction over the comparison family, with one comparison named as primary in advance. Per-cycle tests such as McNemar's were not used because consecutive cycles are autocorrelated.

Open state-of-the-art comparators. Two independent literature searches found no published model with public code evaluated under held-out-unit protocols on the hydraulic rig, the compressor, the milling data, the appliance data or the pump test bed; published figures on the hydraulic rig (typically 99–100 %) come from random splits. On PLAID the best leave-one-house-out result with public code is WRG-NILM [18] (weighted recurrence graphs of V and I, and a CNN). We therefore ran the strongest open models ourselves on our folds, inputs and metrics, in one further preregistration sealed before any of them ran: MiniRocket [19], MultiRocket-Hydra [20] and QUANT [21] (aeon 1.6.0, defaults), HIVE-COTE 2 [22], the time-series foundation model Mantis-8M [23] (frozen embeddings and a random forest), WRG-NILM (the authors' code, unmodified), hidden-Markov and k-means state readers for the operations metrics, and Isolation Forest and ECOD [24] for the compressor failures. Each comparator sees the raw report series of a phrase (variable-length phrases resampled to a fixed length, with log duration as an added constant channel), not our fold vector. The primary comparison, named in advance, is the whitened read against MultiRocket-Hydra on the valve, time-blocked. The feature-map lift of §2.9 was sealed as an addendum after the first comparator results were seen, with its own primary comparison (accumulator, time-blocked, whitened read on the MultiRocket-Hydra map against MultiRocket-Hydra), and is labelled post hoc-motivated.

---

4. Results

4.1 Operations metrics against independent ground truth

The compressor logs a digital outlet-valve signal that is active exactly when it runs under load, which is independent of the motor current. Table 3 and Figures 4 and 5 score readers of the motor current alone against it.

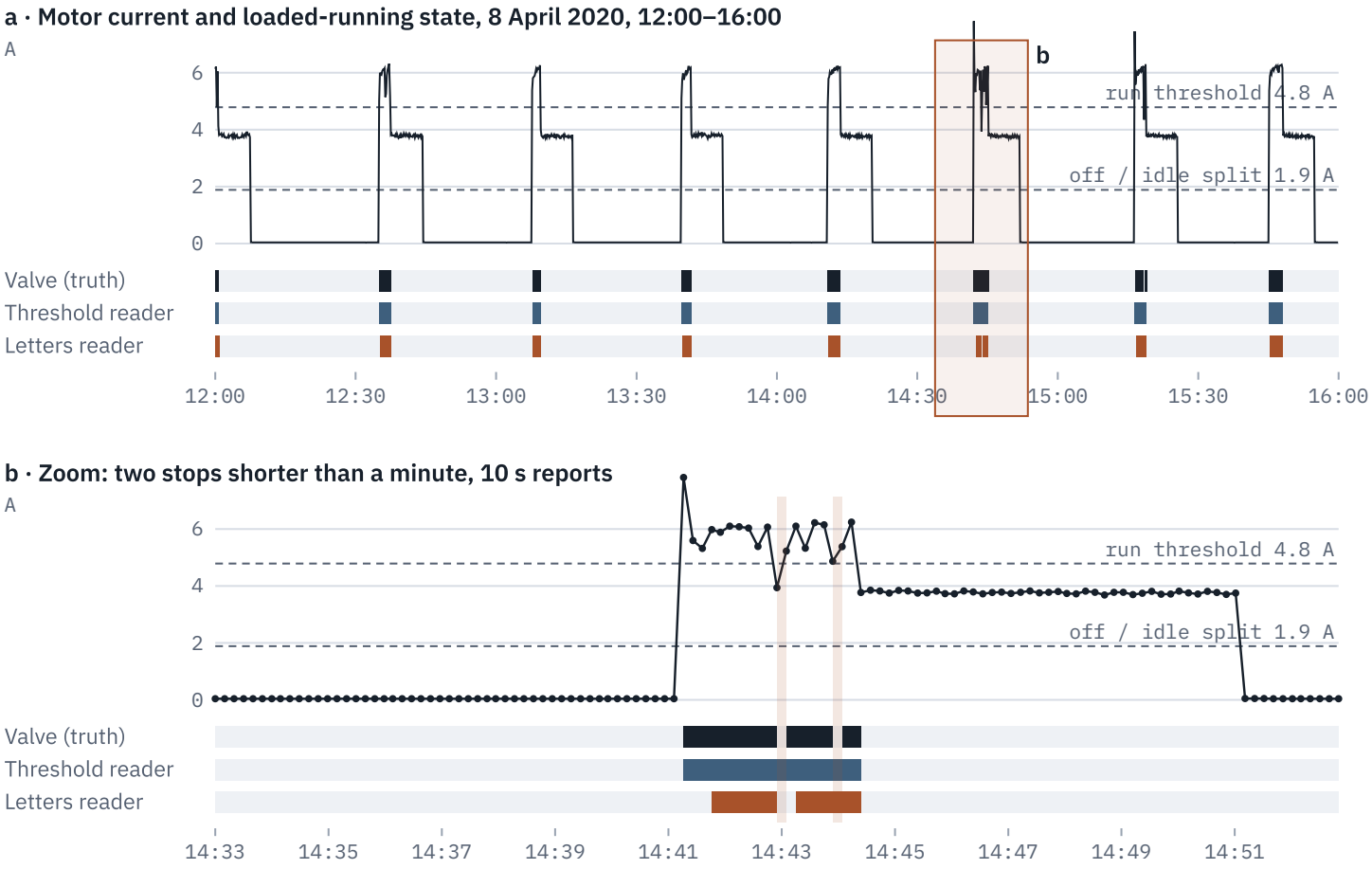

Figure 4. Reading loaded-running state from motor current alone, scored against the compressor's outlet-valve signal. (a) Four ordinary hours: both readers follow the valve closely. (b) A 20-minute zoom (box in a): the valve closes twice for 10–20 s; the threshold reader merges the three runs into one and the letters reader splits them once. Stops this short last one or two reports and were not recoverable at 10 s reporting by any reader (Table 3).

Source: MetroPT-3 CSV; readers from field_datasets_v1/ops.py, states_v2.py (figures/extract.py)

Table 3. Operations metrics, MetroPT-3 compressor (1.50 M scored reports, 5 116 h).
Sources: `result_states*.json`, `result_ops.json`.

Metric	Threshold reader, label-free, debounced	Inscribed letters (§2.2) + chunk rule	Published claims in the product category
Running vs idle accuracy	**0.9955**	0.962	above 0.98
False state changes per hour	**0.031**	1.75	not published as a count
Idle reports read as running	**0.08 %**	0.12 %	"below 0.5 % per hour"
Daily cycle-count error (median)	1.4 %	**1.3 %**	not published
Cycle duration, correlation with truth	**0.9998**	0.946	not published
Daily utilization error, median / 90th pct. (points)	**0.18 / 0.69**	0.96 / 6.9	not published
Whole-period utilization (truth 15.63 %)	**15.57 %**	12.03 %	not published
Stops ≥ 5 min: recall / precision	**0.996 / 0.966**	0.996 / 0.749	not published
Micro-stops ≤ 1 min: recall	0.03–0.09	0.10 (before the chunk rule)	claimed, no figure
Read latency	0.04 µs per report (letters); both instantaneous at report rate		dashboard latency under 5 s (claimed)

Notes. The threshold reader places two thresholds on the current histogram without labels (a three-level Otsu split) and ignores one-report blips; it was chosen after our preregistered readers failed (below) and is a generic method. The inscribed letter reader was preregistered in two rounds: the first failed (0.392 accuracy) because its resolution was set by the very quiet off state; the preregistered correction (resolution from the noise law, §2.2) reached 0.953, and 0.962 with the pipeline's own chunk rule applied post hoc. Its remaining gap is a real slow wander of the loaded state (5.6 to 6.2 A) that a noise-derived resolution splits into separate states. The published claims in the right-hand column concern other machines and are shown only as a reference scale. Two standard unsupervised state readers, fitted on February like ours (preregistered with the comparators, §3), bracket these results: three-cluster k-means on current reaches 0.9938 accuracy with 0.36 false state changes per hour and a 3.5 % median daily cycle-count error, close to the threshold reader; a three-state Gaussian hidden Markov model reaches 0.773 with 122 false changes per hour: its highest-mean state holds only 0.49 of the loaded reports.

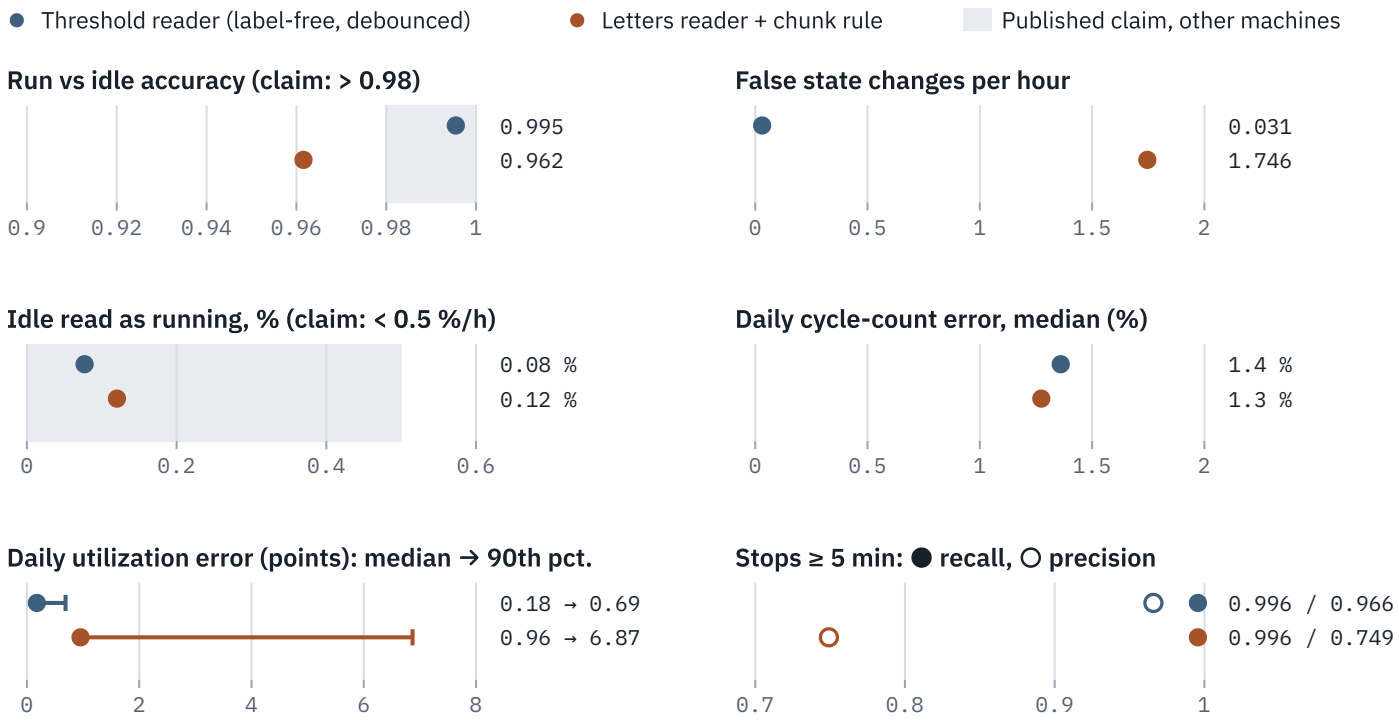

Figure 5. The standard operations metrics of machine monitoring, recovered from 10 s motor-current reports and scored against the compressor's valve signal over 1.52 M reports. The shaded regions mark the only published figures in the product category, which concern other machines; they are a reference scale, not a head-to-head result. The threshold reader was applied after the preregistered readers failed (§4.1).

Source: `field_datasets_v1/result_states_v2.json`, `result_states_v2_posthoc_chunks.json`, `result_ops.json`

4.2 Condition from power

Table 4. Condition from standard-rate electrical telemetry, strictest available protocol (Figure 6). Majority = always predicting the most frequent training class. Sources: `hydraulic_phrase_v1/result_ext.json`, `result_tuned.json`, `field_datasets_v1/result_*.json`.

Machine and condition	Protocol	Whitened read	Shrinkage LDA	XGBoost	Random forest	Majority
Hydraulic rig: valve lag (4 levels)	11 time blocks	0.991	0.998	0.888 (tuned 0.894)	0.814 (tuned 0.886)	0.510
same	forward only (train past, test future)	0.979	0.999	0.808	0.734	0.495
same	unseen combination of the other 3 faults	0.996	0.999	0.943	–	0.510
Hydraulic rig: pump leakage (3)	11 time blocks	0.726	0.833	0.732	0.772	0.554
same	forward only	0.583	0.794	0.667	0.673	0.540
Hydraulic rig: accumulator (4)	11 time blocks	0.330	0.453	0.446	0.390	0.185
same	forward only	0.265	0.289	0.264	0.246	0.209
Compressor: air-leak failures	each of 4 failures held out	recall 1.00 on all 4	recall 1.00 on all 4	0.00	0.00	0.00
Milling: last 20 % of tool life (macro-F1)	each of 14 tools held out	0.748	0.742	0.747	0.770	0.444
Fridges: malfunction (recall at specificity)	each date held out	0.58 at 0.994	0.50 at 0.995	0.42 at 0.996	0.38 at 0.994	0
Washers: fault vs working (balanced accuracy)	each date held out	0.738	0.738	0.739	0.817	0.5
Pumps: 26 mechanical faults	unseen operating speed	0.021	0.021	0.023	0.024	0.156

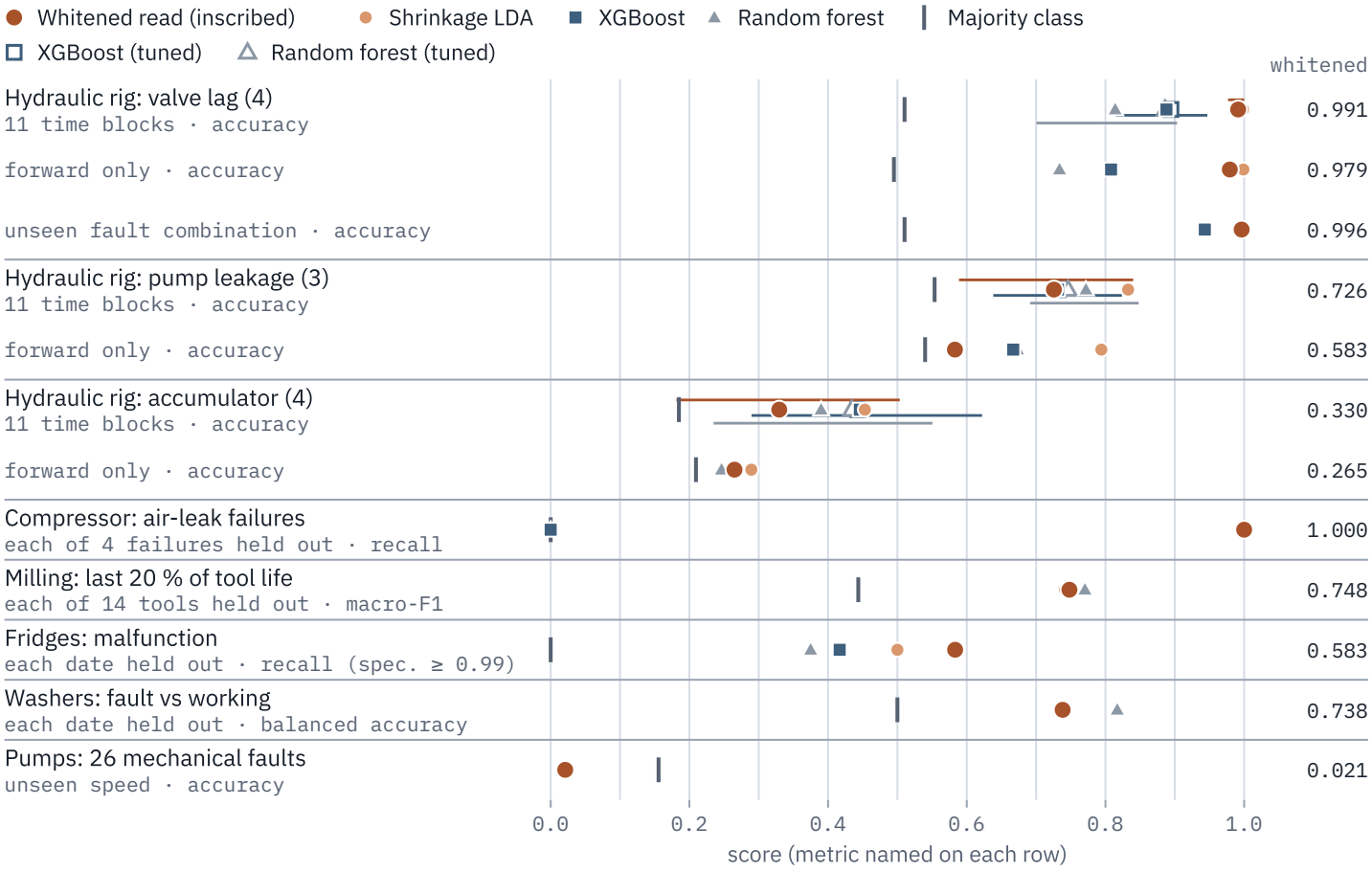

Figure 6. Every condition result under the strictest protocol each dataset allows, on one scale. Linear discriminant reads (copper) lead on valve lag and compressor failures; tree ensembles lead on washers and are level on milling; nobody reads pump mechanical faults at an unseen speed, where every method falls below always predicting the most common class (grey tick). Horizontal bars are 95 % group-bootstrap intervals, drawn where they were computed (hydraulic rig, 11 time blocks).

Source: hydraulic_phrase_v1/result_ext.json, result_tuned.json, result_stats.json; field_datasets_v1/result_metro.json, result_haas_raw.json, result_smartpdm.json, result_twente.json

Valve condition. The primary comparison (named in advance) is the whitened read against XGBoost on the valve, time-blocked: +0.103 mean per-block difference, t-interval [0.028, 0.179], block sign-flip p = 0.004. Because this comparison was named primary in advance, the unadjusted p is its preregistered test; within the 21-comparison family of the addendum, Holm correction gives p = 0.078, so it does not carry a family-wise guarantee at 0.05. The advantage is shared by plain shrinkage LDA and logistic regression; it is a property of linear discriminant reads of the 1 Hz power profile, not of the group balancing, which did not help on the valve and scored lower on pump leakage (0.726 against 0.846 without it; p = 0.039 uncorrected, 0.66 after Holm correction). Across nine settings of block count and purge width (Figure 7) the whitened read stays between 0.975 and 0.996 while XGBoost ranges from 0.663 to 0.933. With the label sequence circularly shifted against the signal, the read falls to 0.346 on average (p = 0.0099 over 100 shifts), so it reads the signal rather than the schedule (Figure 8).

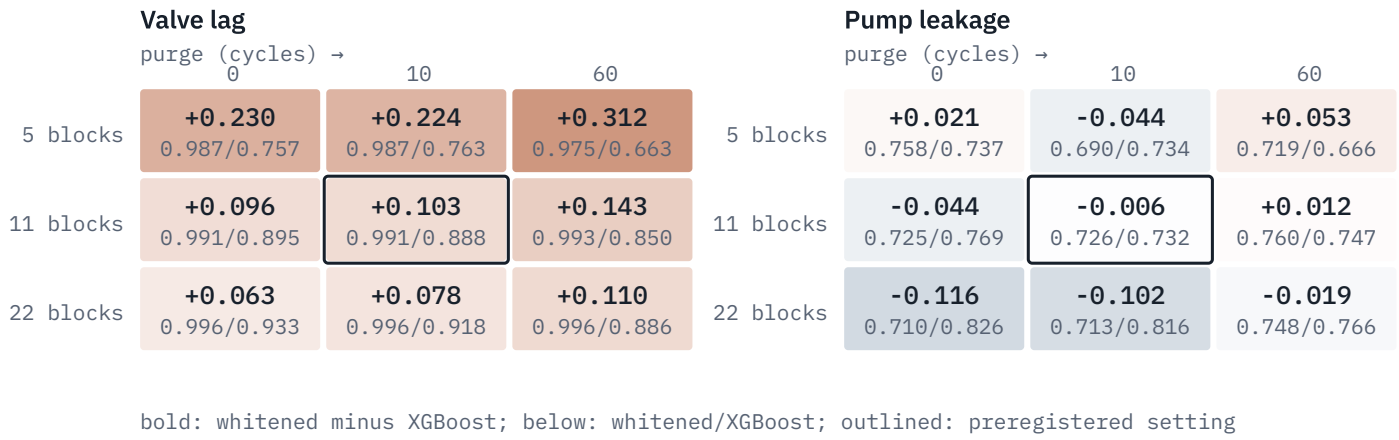

Figure 7. The time-blocked comparison under nine choices of block count and purge width. On valve lag the whitened read stays between 0.975 and 0.996 and leads everywhere; on pump leakage the sign changes with the setting. Copper cells favour the whitened read, slate cells XGBoost. XGBoost's own seed variation on the valve is small (sd 0.0039 over 5 seeds).

Source: hydraulic_phrase_v1/result_sens.json

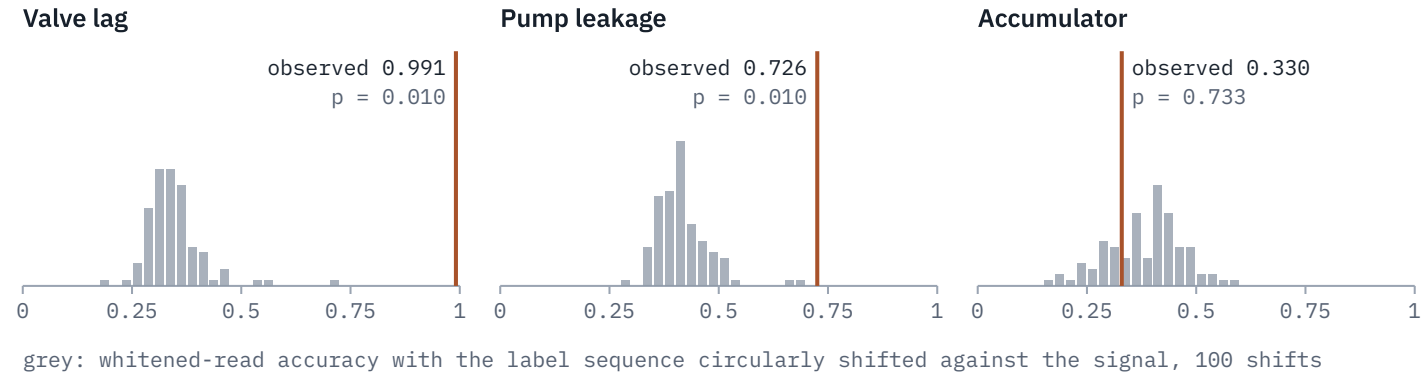

Figure 8. Does the read use the signal or the schedule? Shifting the label sequence in time keeps its runs and persistence but breaks its link to the signal. Valve and pump reads sit far above all 100 shifted runs; the accumulator read does not, so its time-blocked score is not evidence of reading the accumulator.

Source: hydraulic_phrase_v1/result_null.json; draws recomputed with the same seed (figures/extract.py)

Compressor failures. Each failure window contains one to three very long on-periods (577 to 10 937 reports; the compressor runs continuously during an air leak). The inscribed reads flag all seven failure phrases with at most 0.9 % false positives in the week before; the tree ensembles, trained on one to three positive phrases per fold, predict the majority class at their default threshold (they were not compared at a matched false-alarm budget). Four failures are few: under an idealised model of four independent events, 4 of 4 has a 95 % lower bound on recall of 0.40. The phrases are read once they are complete, and some last more than a day; this is retrospective recognition, and a chronological replay reporting detection delay and false alarms per healthy hour is needed before it can be called early warning. A rule that flags any on-period longer than the longest normal one catches none, because a normal on-period of 18 518 reports exists (Figure 9); the read uses the current profile within the phrase.

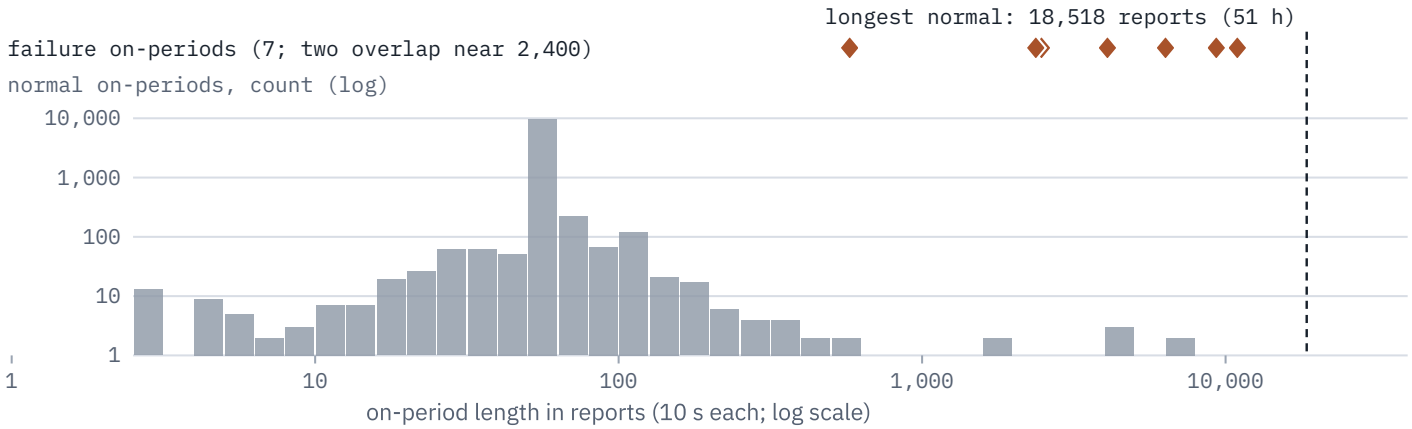

Figure 9. Why a duration rule fails. The compressor runs continuously during an air leak, so failure on-periods are long, but the normal record already contains an on-period of 18,518 reports, longer than any failure phrase. The inscribed reads use the current profile inside each phrase and flag all seven failure phrases (Figure 6); a rule on length flags none. Median normal on-period: 55 reports; 10,504 normal on-periods in all.

Source: MetroPT-3 CSV, phrases from field_datasets_v1/metro.py (figures/extract.py); result_metro_duration_control.json

Where trees are better. On washers the random forest leads (0.817 balanced accuracy against 0.738), and on milling tool wear all methods are level. Remaining-tool-life regression reaches $R^2 = 0.48$ with a tool held out (XGBoost on the 1 s phrase), below the 93.9 % "prediction accuracy" reported by the dataset's authors [7], which used vibration as well as current features and does not state whether tools were held out; the two numbers are not directly comparable.

4.3 What averaged telemetry can and cannot see

In our data, a condition was read when its change in average power or cycle shape exceeded the machine's normal run-to-run variation. On the pump test bed, friction-type faults (bearings, loose or soft feet, stator faults, misalignment, unbalance) raise active power consistently, by +0.1 to +1.6 % at 75 and 100 % speed, while the three healthy recordings of the same pump differ from each other by 1.3 to 2.2 % (Figure 10). Load-changing faults move power far more: impeller damage −2 to −8 %, cavitation −2 to −41 %. A detector that compares each snapshot with the machine's own healthy baseline flagged friction faults in 44 % of snapshots at a 25 % false-alarm rate on held-out healthy sessions (thresholds between healthy sessions ranged from 0.07 to 2.25 % of power); its two-sided form flagged load-changing faults in 94 % of snapshots, at a 56 % false-alarm rate. Both designs followed a label-informed look at this data. A related limit appeared on the hydraulic rig with our own phrase features: accumulator pre-charge is read at 0.25–0.29 when training on the past and testing the future (majority 0.21), and the whitened read's time-blocked score does not exceed its circular-shift null ($p = 0.73$). This is a limit of the feature vector, not of the telemetry: on a borrowed convolution or quantile feature map the same inscribed read clears that null (§4.10).

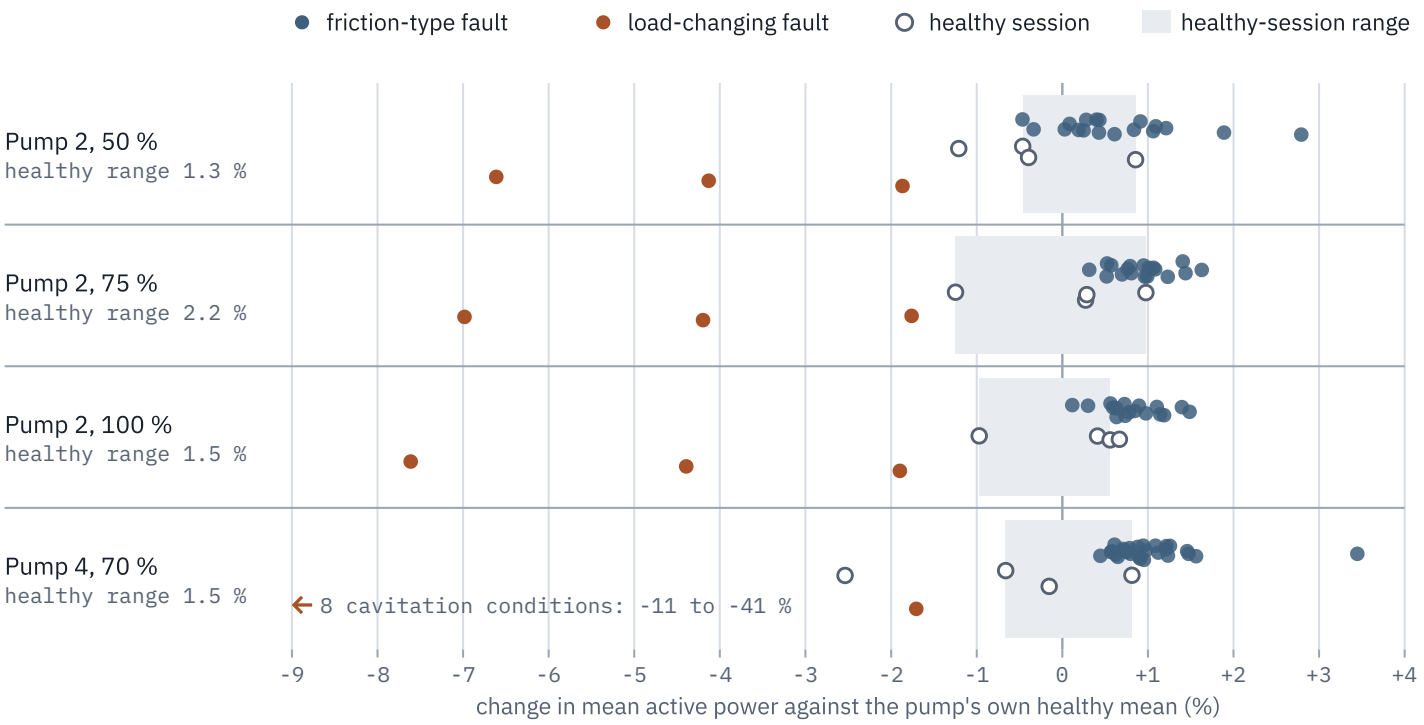

Figure 10. What averaged power can see on the pump test bed. Each dot is one installed condition's mean active power against the same pump's healthy mean. Friction-type faults (bearings, loose or soft feet, stator shorts, misalignment, unbalance) mostly land inside or just beyond the range spanned by the three healthy recordings; impeller damage and cavitation move power far more. "New motor" is omitted, as it is not a fault.

Source: field_datasets_v1/data/twente_1s.npz, channel 4 = active power (figures/extract.py)

Table 5. Reporting interval, hydraulic rig, time-blocked (whitened read / XGBoost). Source: `result_rates*.json`.

Report every	Valve	Pump leakage	Accumulator
0.01 s (100 Hz)	0.992 / 0.939	0.732 / 0.741	0.586 / 0.434
0.1 s	0.992 / 0.947	0.727 / 0.718	0.582 / 0.447
1 s	0.991 / 0.888	0.726 / 0.732	0.330 / 0.446
2 s	0.975 / 0.839	0.665 / 0.774	0.277 / 0.397
5 s	0.730 / 0.713	0.544 / 0.776	0.257 / 0.313
10 s	0.417 / 0.614	0.496 / 0.737	0.249 / 0.230

Valve lag is a transient of a few seconds (Figure 11): readable at 1 to 2 s reporting, largely lost at 5 to 10 s. Pump leakage lives in levels and survives averaging for tree methods. On the milling machine, 10 s reporting loses little (macro-F1 0.708 to 0.747 against 0.742 to 0.770 at 1 s).

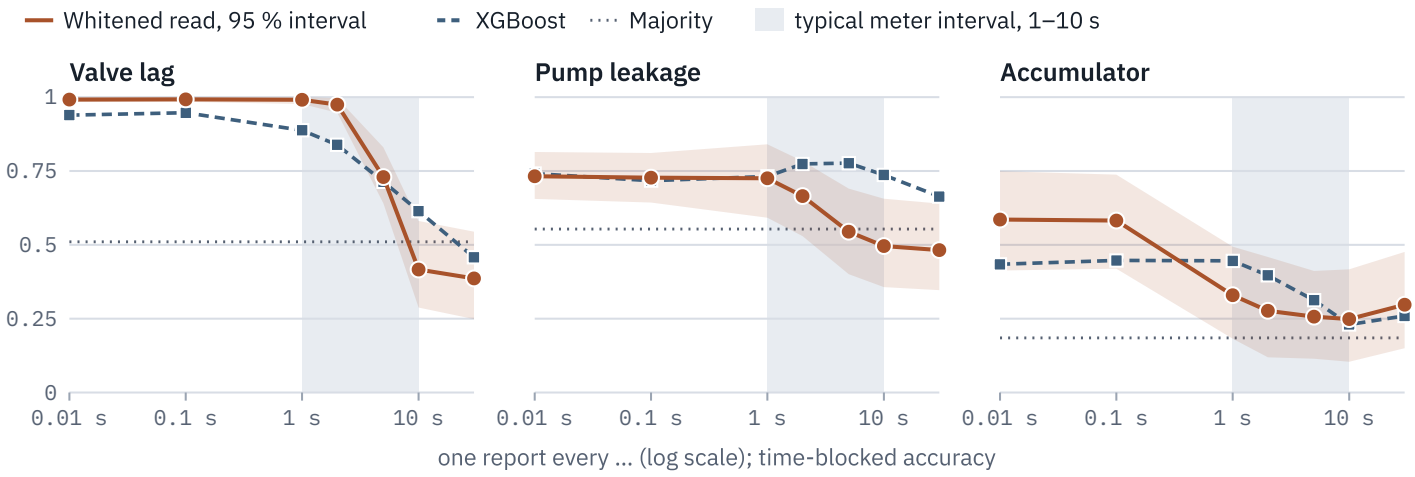

Figure 11. How long can the meter average before the condition disappears? Valve lag is a few-second transient: readable at 1–2 s reports, largely gone at 5–10 s. Pump leakage lives in power levels and survives averaging for XGBoost, not for the whitened read. The accumulator is weak at every interval.

Source: `hydraulic_phrase_v1/result_rates*.json, result_ext.json`

4.4 The decoder collapses with the phrase

Table 6. Valve state from the first h seconds of a cycle versus the last h seconds before the decision (hydraulic rig, time-blocked). Source: `result_ext.json`, `result_decoder.json`.

Horizon h	1 s	5 s	10 s	20 s	30 s	60 s
First h seconds of the phrase	0.203	0.241	**0.984**	0.991	0.991	0.991
Last h seconds before the decision	0.215	0.218	0.216	0.270	0.378	0.991

The valve state lives in the opening high-load step of the cycle. Anchored at the phrase onset, the read collapses within 10 s; anchored at the decision time, it needs the whole 60 s cycle. Summing evidence across cycles lowers valve accuracy (0.991 at one cycle, 0.661 at ten) because the valve changes every 15 cycles on average; a forward filter with a switching prior keeps 0.990 without choosing a horizon (0.983 with cycle order shuffled inside the test block, so it does not rely on schedule persistence). Pump leakage needs the whole cycle from either anchor, and the accumulator never collapses. Both preregistered decoder gates (strict monotonicity within a cycle; accumulation beyond one cycle helps) failed as written.

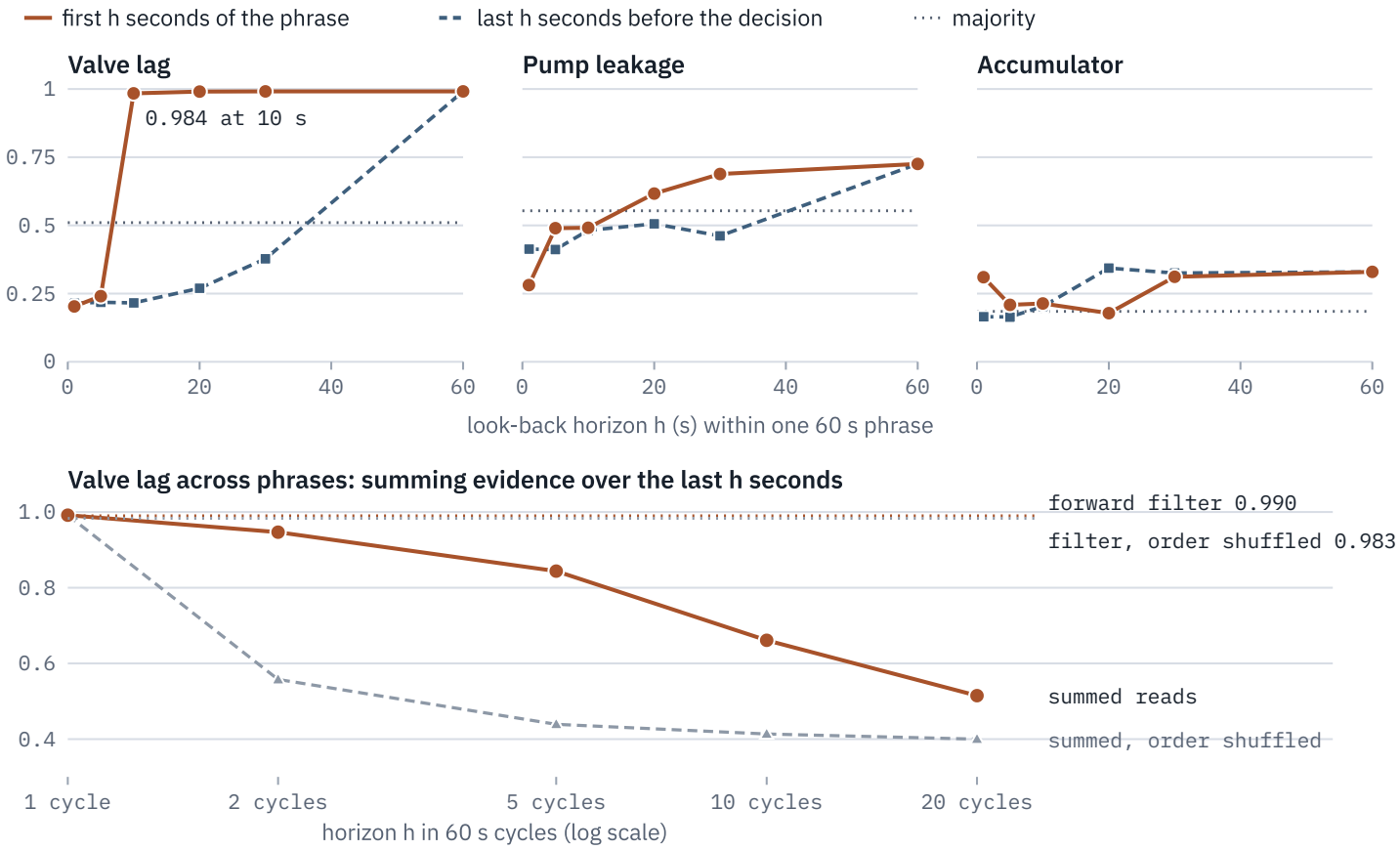

Figure 12. The decoder collapses with the phrase. Top: anchored at the start of a cycle, the valve read is decided within 10 s, because the valve state lives in the opening high-load step (Figure 2); anchored at the decision time, it needs the whole cycle. Bottom: summing evidence over more cycles hurts, because the valve changes every 15 cycles on average; a forward filter with a switching prior keeps the one-cycle accuracy without choosing a horizon, and keeps it with cycle order shuffled inside each test block.

Source: hydraulic_phrase_v1/result_ext.json (decoder_windows), result_decoder.json

4.5 Self-classification with arbitrary after-the-fact labels

On the simulated machine (5 s reports, 1.5 % noise, 1 W quantisation), the label-free pipeline finds all six states, exact phrase boundaries and all four phrase types, and names them from arbitrary strings: 360 of 360 test phrases per rung are labelled correctly (Figure 13) and self-classified with labels that are exact, noisy, jittered by ±2 reports, or given for only 20 % of phrases plus 10 % queried at the smallest margin, with nested batch labels set aside. The same holds on three seeds never used during development. When fouling is a slow drift labelled by a technician's threshold, labels are needed (self-classified share 0.39) and the gate is missed by one or two phrases per seed; on the fresh seeds a new error appeared (short three-report phrases misread in two seeds).

On the real hydraulic rig the same pipeline finds four states and the rig's load program ("1 s low, 9 s high, 50 s low" in 55 % of cycles) but not its conditions (valve 0.51 against a 0.51 majority rate): the rig's faults do not change which states it passes through; they reshape the same phrase. Self-classification worked when a condition changed the states a machine visits and did not when a condition only reshaped the phrase; two settings do not establish this as a general law.

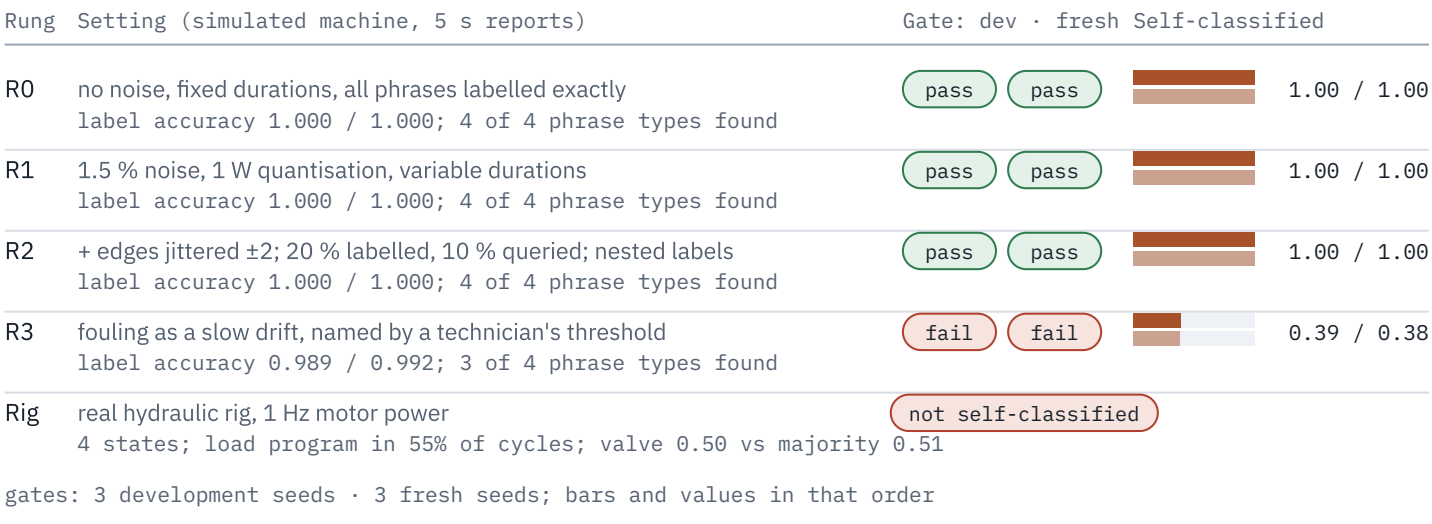

Figure 13. Self-classification with arbitrary after-the-fact labels. On the simulated machine every phrase type is found and named from technician strings at R0–R2, on the three development seeds and on three seeds never used during development. At R3 fouling no longer changes which states the machine visits, so labels have to do the work and the gate is missed. On the real rig the conditions reshape the same phrase rather than changing it, and are not self-classified.

Source: machine_phrase_v1/result_R*.json, result_fresh_R*.json; hydraulic_phrase_v1/result_selfclass.json

4.6 Dynamics and bounds (simulated)

With a heating element whose fouling lengthens its time constant at equal steady power, reading fouling from a single phrase improves in steps (Figure 14): steady-level control 0.48, velocity, acceleration and jerk features 0.60, per-phrase pole threshold 0.67, refined poles 0.90, whole-phrase whitened poles 0.993, and **class Frames 1.000**, all with 10 % of phrases labelled (class Frames also 1.000 at 5 %, and again on fresh seeds). Per-phrase time constants reach 5.6 % median error against a Cramér–Rao bound of 5.0 %. Only one of a motor's three transient modes clears the noise edge in a single phrase; pooling phrases with shared poles recovers two. For a slowly drifting fault, reading each phrase through the session's condensate (a polynomial in phrase order whose degree is chosen by BIC) raises accuracy from 0.837 to 0.970, against a 0.978 ceiling with the true threshold.

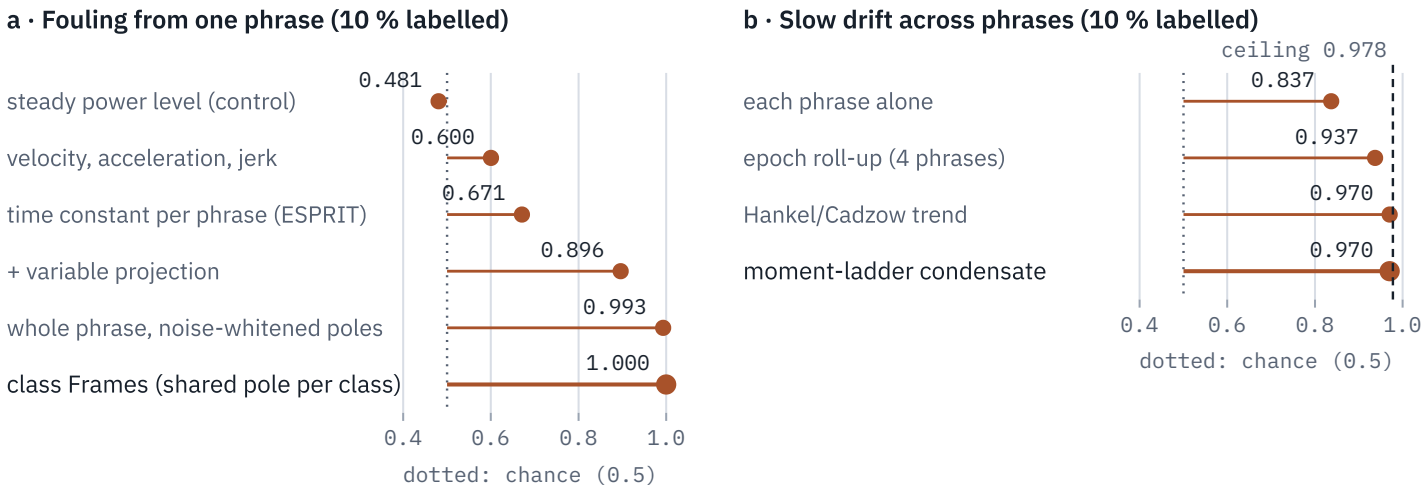

Figure 14. Simulated dynamics, where the truth is known. (a) A heating element whose fouling lengthens its time constant at equal steady power: power level alone is near chance, derivative features help little, and reading the phrase as poles climbs to class Frames at 1.000 (also 1.000 on fresh seeds). (b) When fouling is a slow drift, reading each phrase through the session's trend recovers most of the gap to the true-threshold ceiling.

Source: machine_dynamics_v1/result_R1.json, result_R2.json, result_fresh_R1.json; failed_fix3/, failed_fix4/ result_R1.json (earlier sealed steps of the same ladder)

4.7 Appliance identification and the tensor lift (PLAID)

Folding each appliance record into its turn-on phrase (Figure 15) raised XGBoost from 0.779 to 0.886 accuracy on held-out homes (+0.107, CI [0.074, 0.144]). The lift of that ensemble into a lookup tensor reached 0.891 (+0.005 [−0.012, 0.027] against the ensemble, non-inferior at a 0.02 margin); on held-out records its R^2 against the ensemble's margins has a median over homes of 0.92 (order 1) and 0.95 (with pair tables), a record-weighted mean of 0.66 and 0.82, and a plain mean of −0.65 and 0.13: the fit is good on most homes and poor on a few small ones. The lift result came from the eighth preregistration on the same folds, after earlier versions failed.

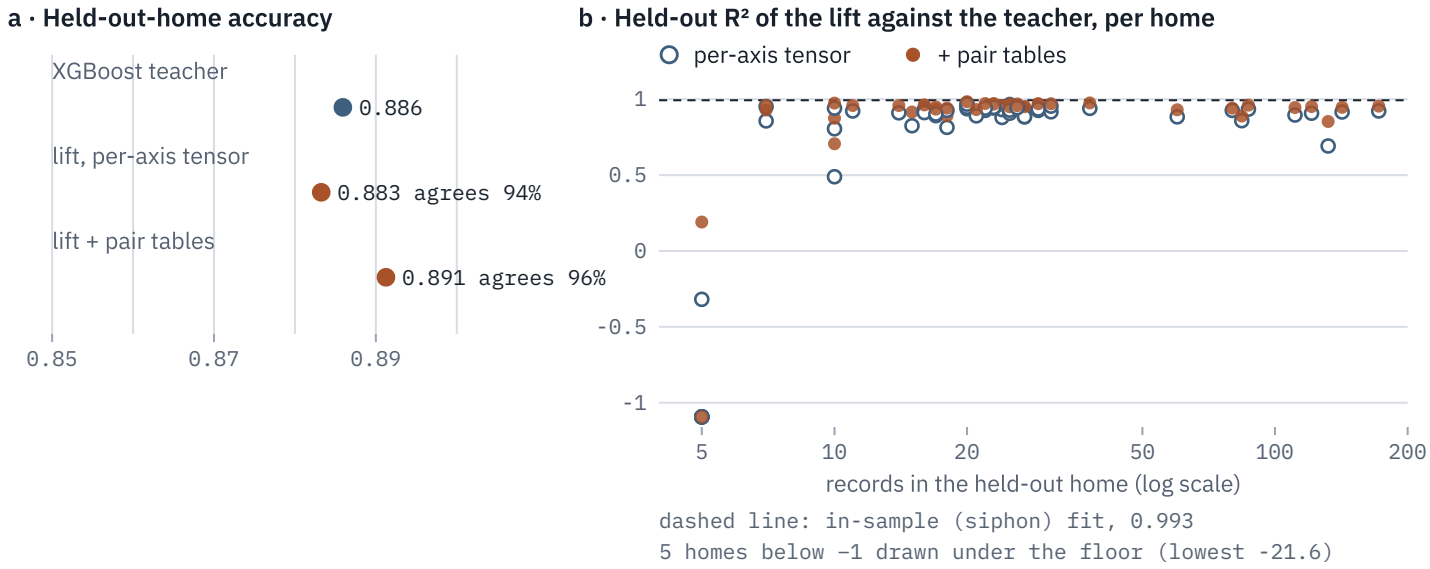

Figure 15. Lifting a trained tree ensemble into an explicit lookup tensor (PLAID, 56 homes, each held out in turn). (a) The lift matches its teacher's accuracy. (b) How much of the teacher's calculus the tensor reproduces on unseen homes: most homes sit near the in-sample fit, and a few small homes fall far below, which is why the plain mean of held-out R^2 is low while its median is high.

Source: nilm_plaid_v1/result_v8_R3.json, result_heldout_r2.json, run_heldout_r2.log (figures/extract.py)

4.8 Evaluation protocol changes the conclusion

On the hydraulic rig (Figure 16), XGBoost scores 0.971 on pump leakage and 0.963 on the accumulator under random five-fold resampling, and 0.732 and 0.446 under time-blocked evaluation; across methods the gap reaches 59 points. Random folds reward recognising neighbouring cycles of the same condition run. Holding out a combination of the other faults inflates slow factors in the same way. Published figures on the hydraulic rig (typically 99–100 %) come from random splits; on PLAID, random splits give 95–99 % while holding houses out gives far less (in one study, up to 98.5 % with houses shared between training and test, and at most about 80 % on unseen houses [17]).

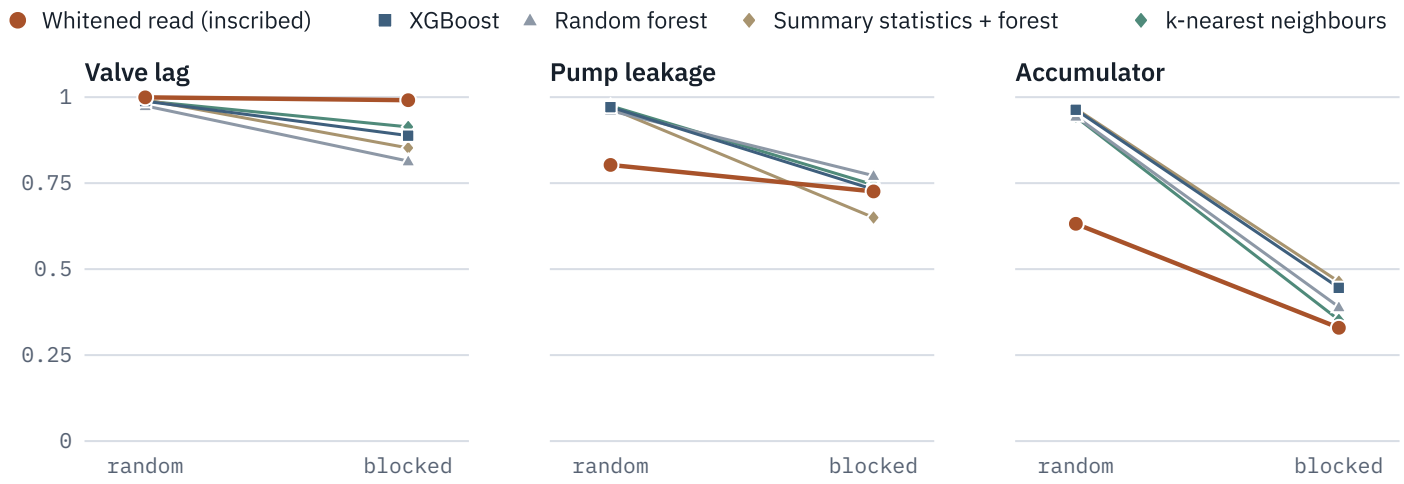

Figure 16. The protocol changes the conclusion. Random five-fold resampling puts neighbouring cycles of the same condition run on both sides of the split, so methods that recognise neighbours look nearly perfect on pump leakage and the accumulator. Holding out blocks of time removes that; the largest drop is 59 points (k-nearest neighbours, accumulator). The valve read, which depends on the signal, barely moves.

Source: hydraulic_phrase_v1/result_1hz_rand_*.json, result_1hz_block_*.json

4.9 Comparison with open state-of-the-art classifiers

Table 7 and Figure 17 put the strongest open time-series classifiers on the same held-out splits (§3). On valve lag they match the inscribed reads: the preregistered primary comparison, whitened read against MultiRocket-Hydra, time-blocked, is −0.006 per block (t-interval [−0.021, 0.008], sign-flip p = 0.75). A non-significant difference is not equivalence; with a margin of ±2 accuracy points chosen after the results were seen, two one-sided tests at 0.05 (90 % intervals) show equivalence for MultiRocket-Hydra [-0.018, 0.005], MiniRocket [-0.015, 0.016] and HIVE-COTE 2 [-0.018, 0.003], and for QUANT only at ±2.5 points [-0.020, 0.004] (`result_equivalence.json`), and QUANT, MultiRocket-Hydra and shrinkage LDA all reach 0.998–1.000. On pump leakage and the accumulator the convolution and quantile feature maps are clearly better than our fold vector: the whitened read is 0.087 below MiniRocket on pump leakage (per-block sign-flip p = 0.15, not significant over 11 blocks) and 0.17–0.18 below each of the three on the accumulator ($p \le 0.004$). Plain shrinkage LDA on our fold vector remains competitive on pump leakage (0.833 time-blocked, 0.794 forward-only, above every comparator's own head). The frozen foundation model is the weakest reader on valve lag and pump leakage. On the field data the open classifiers, at their default thresholds, flag at most half of the compressor's failure phrases (the two unsupervised detectors 0.50 mean recall, the supervised classifiers 0.00–0.31) where the inscribed reads catch all four; find fewer fridge malfunctions (0.17–0.50 recall at 0.994–0.995 specificity, against 0.58); tie on tool wear (0.706–0.750 macro-F1 against 0.748); lead on washers (QUANT 0.829 balanced accuracy, level with the random forest's 0.817 and above our 0.738); and, like every method, fall below the majority rate on pump faults at an unseen speed (0.023–0.146 against 0.156). On PLAID, with each home held out, every open comparator is below the lift of §4.7 (0.891): QUANT 0.781, WRG-NILM 0.757 (0.832 when the training epoch is chosen on the held-out home, as its published script does), MultiRocket-Hydra 0.739, Mantis 0.703 and MiniRocket 0.699; they read the steady-state cycles their literature uses, our reads the turn-on phrase

(§6). WRG-NILM's published 92.7 % was measured on the smaller 2014 release with 11 classes.

Table 7. Open state-of-the-art classifiers on the paper's held-out splits. Best score per row in bold where scores are comparable. Sources: `sota_compare_v1/result_*.json`; the paper's reads from Table 4. "Not run": HIVE-COTE 2 was stopped after the valve's time-blocked folds (§5), and is not run on PLAID or the compressor by design. On PLAID, MiniRocket, MultiRocket-Hydra and QUANT are the identical fits recorded as each map's own head in the lift run (Table 8). † WRG-NILM with the training epoch chosen on the held-out home, as its published script does (optimistic).

Machine and condition	Protocol	Whitened read	Shrinkage LDA	MiniRocket	MultiRocket-Hydra	QUANT	HIVE-COTE 2	Mantis-8M	Other
Hydraulic rig: valve lag (accuracy)	11 time blocks	0.991	0.998	0.991	0.998	**1.000**	0.999	0.843	–
same (accuracy)	forward only	0.979	**0.999**	0.985	**0.999**	**0.999**	not run	0.786	–
Hydraulic rig: pump leakage (accuracy)	11 time blocks	0.726	**0.833**	0.813	0.812	0.734	not run	0.595	–
same (accuracy)	forward only	0.583	**0.794**	0.707	0.754	0.704	not run	0.562	–
Hydraulic rig: accumulator (accuracy)	11 time blocks	0.330	0.453	0.510	**0.513**	0.504	not run	0.462	–
same (accuracy)	forward only	0.265	0.289	0.281	0.294	**0.363**	not run	0.334	–
PLAID: appliance type, 16 classes (accuracy)	each home held out	–	–	0.699	0.739	0.781	not run	0.703	our XGBoost lift 0.891; WRG-NILM 0.757 (0.832†)
Milling: last 20 % of tool life (macro-F1)	each tool held out	0.748	0.742	0.733	0.706	**0.750**	not run	0.739	random forest 0.770
Fridges: malfunction (recall; specificity)	each date held out	0.583 (0.994)	0.500 (0.995)	0.208 (0.995)	0.167 (0.995)	0.500 (0.995)	not run	0.417 (0.994)	random forest 0.375
Washers: fault vs working (balanced accuracy)	each date held out	0.738	0.738	0.816	0.791	0.829	not run	0.812	random forest 0.817
Pumps: 26 mechanical faults (accuracy)	unseen speed	0.021	0.021	0.110	0.091	0.023	not run	0.146	majority 0.156
Compressor: air-leak failures (recall; max FPR in the week before)	each failure held out	1.00 (0.009)	1.00 (0.009)	0.00 (0.000)	0.25 (0.000)	0.00 (0.000)	not run	0.31 (0.000)	IForest 0.50 (0.010); ECOD 0.50 (0.018)

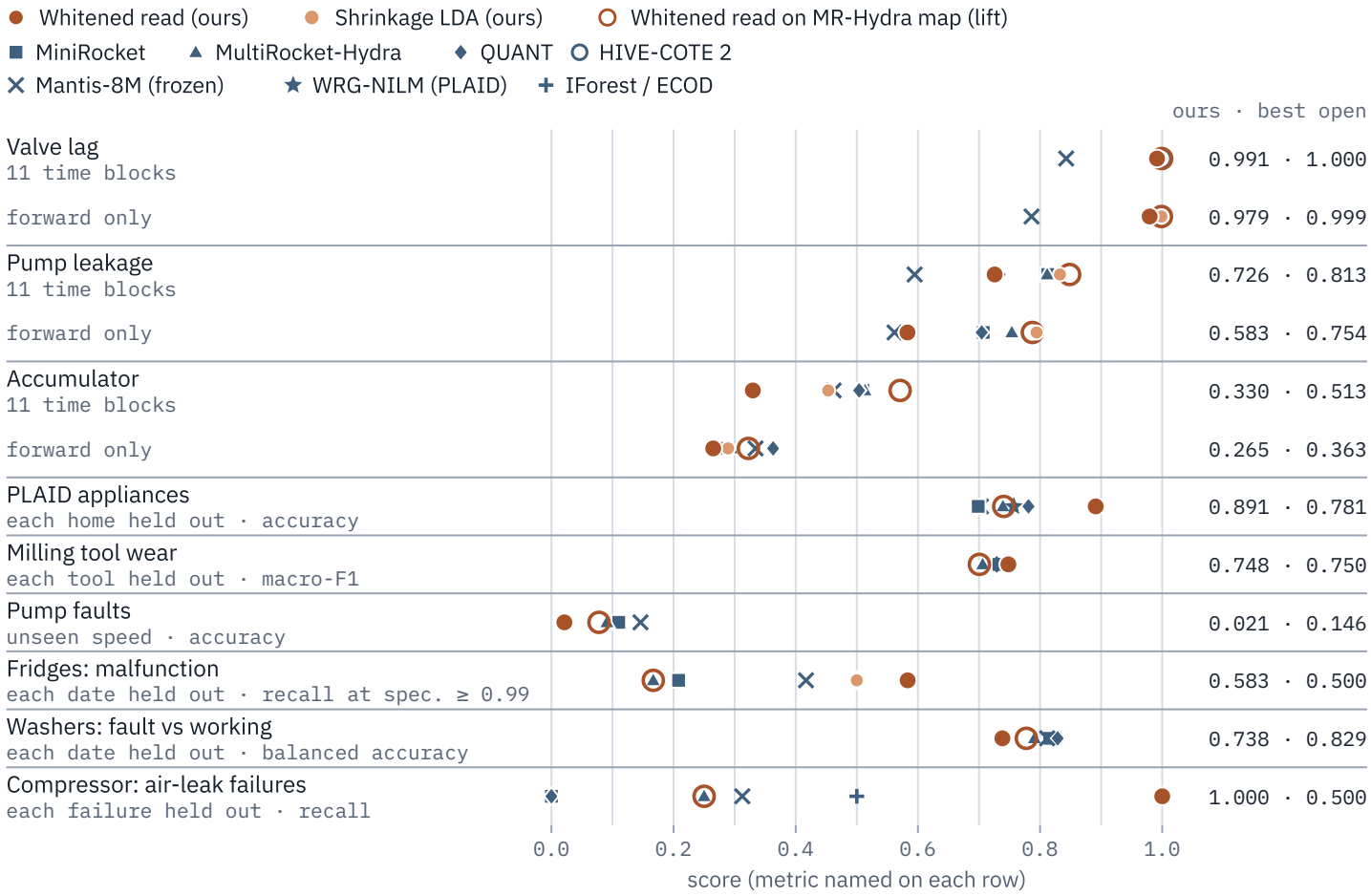

Figure 17. The same held-out splits, read by open state-of-the-art classifiers (slate) and by our inscribed reads (copper; light copper is plain shrinkage LDA). On valve lag they match (equivalent within 2–2.5 points, margin chosen post hoc). On pump leakage and the accumulator the convolution and quantile feature maps win; inscribing our read on the MultiRocket-Hydra feature map (open ring) shows how much of that advantage the Frame can take over, with no gradient step in the read. Right-hand column: our read (the whitened read; the XGBoost lift on PLAID), then the best open comparator, including the unsupervised detectors on the compressor.

Source: sota_compare_v1/result_*.json (preregistered, SHA-256 sealed); hydraulic_phrase_v1/result_ext.json; field_datasets_v1/result_*.json

Cost differs by orders of magnitude (Figure 3, Table 1). On one hydraulic fold the whitened read is fitted and applied in 6 ms; MiniRocket takes 1.9 s, MultiRocket-Hydra 7.0 s, QUANT 0.7 s and HIVE-COTE 2 405 s on the same CPU (means over folds, measured while other jobs shared the machine, so the comparators' times are upper bounds). HIVE-COTE 2 reached 0.999 on the valve (against the whitened read, per-block difference −0.007, sign-flip p = 0.25); at that cost it was stopped after the valve's time-blocked folds, a declared deviation (§5).

4.10 Inscription on a borrowed feature map

The comparators' advantage on pump leakage and the accumulator is their feature map, not their head. Table 8 keeps each comparator's fitted map and replaces its head with an inscribed read (§2.9). On the accumulator the preregistered test of the lift (whitened read on the MultiRocket-Hydra map against MultiRocket-Hydra, time-blocked) gives +0.058 per

block (t-interval [0.007, 0.108], sign-flip p = 0.019): 0.571 against 0.513. On pump leakage the same read gives 0.849 against 0.812 (+0.037, [0.002, 0.071], p = 0.030), and on the valve both are at ceiling (0.999 against 0.998). The largest scores come from the QUANT map read with plain class means: 0.746 on the accumulator (its own extra-trees head 0.504) and 0.862 on pump leakage (head 0.734), the highest time-blocked accuracies on the rig of any method in this study; they are the best of six lifted variants and carry no separate test. The lift does not always help: on the MiniRocket map the inscribed read is below MiniRocket's own ridge head on pump leakage (0.672 against 0.813). All four lifted reads clear the same circular-shift null that the paper's own accumulator read failed (PREREGISTRATION_ADD2; 100 shifts, p = 0.0099, the smallest attainable): on the accumulator 0.570 and 0.746 against null means of 0.387 and 0.371, on pump leakage 0.850 and 0.862 against 0.486 and 0.452. Accumulator pre-charge can therefore be read from 1 Hz motor power; the limit we had reported was that of our phrase features. (The null recomputes the read through the training Gram matrix; it reproduces the direct read to within one cycle, 0.570 against 0.571.) On the field data the lift is mixed. It gives the best washer result of any method, again the best of six variants (MiniRocket map, whitened read, 0.847 balanced accuracy against its head's 0.816); it is below the maps' own heads on tool wear and on fridges (where it also trades specificity); on the compressor no borrowed map catches more than 0.375 of the failure phrases on average, where our own phrase features catch all of them; and the pump faults stay at chance. On PLAID the lift ties the MultiRocket-Hydra head (0.740 against 0.739) and is below the MiniRocket and QUANT heads (0.656 against 0.699; 0.499 against 0.781).

Table 8. The feature-map lift: each comparator's own head against inscribed reads on the same fitted map. Best per row in bold. Source: `sota_compare_v1/result_lift_*.json` (PREREGISTRATION_ADD1, sealed after the first comparator results).

Machine and condition	Protocol	Feature map	Map's own head	Inscribed, group-balanced (L-white)	Inscribed, plain means (L-plain)	Paper's whitened read
Hydraulic rig: valve lag	11 time blocks	MiniRocket	0.991	**0.997**	0.994	0.991
		MR-Hydra	0.998	**0.999**	**0.999**	
		QUANT	**1.000**	0.992	0.995	
	forward only	MiniRocket	0.985	0.994	**0.998**	0.979
		MR-Hydra	**0.999**	**0.999**	**0.999**	
		QUANT	**0.999**	0.986	0.996	
Hydraulic rig: pump leakage	11 time blocks	MiniRocket	**0.813**	0.672	0.636	0.726
		MR-Hydra	0.812	0.849	**0.851**	
		QUANT	0.734	0.743	**0.862**	
	forward only	MiniRocket	**0.707**	0.673	0.624	0.583
		MR-Hydra	0.754	0.787	**0.788**	
		QUANT	0.704	**0.796**	0.771	
Hydraulic rig: accumulator	11 time blocks	MiniRocket	0.510	0.516	**0.520**	0.330
		MR-Hydra	0.513	0.571	**0.595**	
		QUANT	0.504	0.552	**0.746**	
	forward only	MiniRocket	0.281	0.307	**0.317**	0.265
		MR-Hydra	0.294	0.322	**0.329**	
		QUANT	0.363	0.411	**0.485**	
Milling: tool wear (macro-F1)	each tool held out	MiniRocket	**0.733**	0.705	0.705	
		MR-Hydra	**0.706**	0.701	0.704	
		QUANT	**0.750**	0.679	0.652	
Pumps: 26 faults (accuracy)	unseen speed	MiniRocket	**0.110**	0.105	0.097	
		MR-Hydra	**0.091**	0.078	0.079	
		QUANT	0.023	0.053	**0.062**	
PLAID (accuracy)	each home held out	MiniRocket	**0.699**	0.656	0.632	
		MR-Hydra	0.739	**0.740**	**0.740**	
		QUANT	**0.781**	0.499	0.484	
Fridges (malfunction recall)	each date held out	MiniRocket	0.208	**0.375**	**0.375**	

Machine and condition	Protocol	Feature map	Map's own head	Inscribed, group-balanced (L-white)	Inscribed, plain means (L-plain)	Paper's whitened read
		MR-Hydra	**0.167**	**0.167**	**0.167**	
		QUANT	**0.500**	0.125	0.125	
Washers (balanced accuracy)	each date held out	MiniRocket	0.816	**0.847**	0.841	
		MR-Hydra	**0.791**	0.778	**0.791**	
		QUANT	**0.829**	0.704	0.711	
Compressor failures (recall)	each failure held out	MiniRocket	0.00	0.25	0.00	
		MR-Hydra	0.25	0.25	0.25	
		QUANT	0.00	0.38	0.38	

The reading is therefore: inscription is a competitive head — on most of these maps and targets it matched or beat the ridge and extra-trees heads the maps were designed with, in one closed-form pass — while the quality of the feature map decides what can be read. Our fold vector is a good map for transients such as valve lag and a poor one for the accumulator; random convolutions and dyadic quantiles are better general-purpose maps, and a Frame can hold them.

5. Negative results

- Preregistered readers that failed: the first inscribed state reader on the compressor (0.392); the first letter pipeline on the simulator (nine sealed correction files, one superseded, before its first three rungs passed); the label-free CUSUM change alarm on the compressor (107 of 157 alarms outside any failure window, no better than shuffled days); both decoder gates on the hydraulic rig; lift non-inferiority on the valve (−0.009, CI too wide); several PLAID gates.
- Conditions that no method we tried read from averaged telemetry in our data: pump mechanical faults at an unseen speed (every method below the majority rate), friction faults against reinstallation drift, accumulator pre-charge with our own phrase features (it is readable on a borrowed feature map, §4.10), micro-stops shorter than a minute at 10 s reporting, unseen washer fault episodes.
- The group balancing we proposed did not help on the valve and scored lower on pump leakage (not significant after correction).
- On pump leakage and the accumulator our phrase features are a worse map than the open classifiers' random convolutions and quantiles (§4.9), and the feature-map lift did not improve on the comparators' heads on tool wear or on the pump faults (§4.10).
- HIVE-COTE 2 was stopped after the valve time-blocked folds (11 of 57 preregistered hydraulic folds; its field runs never started) because of its cost on a shared machine, about 400 s per fold; this is a deviation from the preregistration.

6. Limitations

- Two data sources are simulators we designed together with the methods; they establish mechanisms, not field performance.
- The results are adaptively developed evidence. Methods were revised after results on the same folds were seen (each revision sealed and kept), and the seals carry no external timestamp. A frozen pipeline evaluated once on new machines or episodes has not been run; it is the most important next step.
- Condition results are retrospective. Hydraulic cycles come with their 60 s boundaries, and compressor phrases are read when complete; neither measures segmentation errors in a live stream or detection delay. The compressor comparison is at each method's default threshold, not at a matched false-alarm budget.
- Attribution: the best operations numbers belong to a threshold baseline; the PLAID comparison mixes input window (turn-on against steady state) with method; the largest lifted numbers use borrowed feature maps. The paper's contribution is the evidence on what standard-rate telemetry carries and the closed-form readers on top of it, not a new category of statistical learning.
- Timing: our figures cover fitting and reading, not phrase folding; the comparators' include their feature transforms and, for XGBoost, process start-up, and were measured on a shared machine. A controlled end-to-end benchmark (warm processes, fixed threads, memory) has not been run.
- The operations metrics come from one compressor; published figures in the product category concern other machines.
- Several fault labels coincide with a single recording session or repair episode (hydraulic cooler, pump conditions, washer faults), which limits what the protocols can separate.
- The threshold reader of Table 3 and the chunk rule were applied after the preregistered readers failed.
- One set of gates on the simulated dynamics was re-set after results were seen, against independently computed information bounds; this is declared in the preregistration record.
- Many comparisons were made; only the valve comparison was primary. Block-bootstrap intervals rest on 11 blocks.
- Preregistration seals are SHA-256 files without an external timestamp.
- The lift of §2.5 needs a trained teacher; the inscribed reads do not.
- The open comparators ran with their published default settings, not tuned per dataset (the paper's own reads are not tuned either); tuning could raise them. Mantis was used frozen, not fine-tuned. Their wall-clock times were measured while other jobs shared the machine and are upper bounds.
- The feature-map lift was sealed after the first comparator results were seen; its largest numbers are the best of six variants (three maps × two reads) and only the named primary comparison carries a test. The feature maps it reads are the comparators' design, not ours.
- On PLAID the time-series comparators and WRG-NILM read the steady-state cycles their literature uses, while our reads use the turn-on phrase; the two inputs differ.

7. Reproducibility

Code, preregistrations (each sealed by SHA-256 before it ran), result files and failed runs are kept in Sekos's research repository, organised as six experiment sets (`machine_phrase_v1`, `machine_dynamics_v1`, `nilm_plaid_v1`, `hydraulic_phrase_v1`, `field_datasets_v1`, `sota_compare_v1`); they are available to reviewers on request. The comparators of §4.9 are public packages (aeon 1.6.0, mantis-tsfm with the public Mantis-8M weights, hmmlearn) and the authors' WRG-NILM code, used unmodified. Environment: Python 3.14.6, NumPy 2.4.6, SciPy 1.17.1, scikit-learn 1.8.0, XGBoost 3.4.1, PyTorch 2.12.1 (CPU), on an Apple M5 Max without a GPU. Public data: PLAID (figshare, CC BY 4.0), UCI hydraulic (CC BY 4.0), MetroPT-3 (UCI, CC BY 4.0), CNC milling tool life (figshare, CC BY 4.0), SMART-PDM (Zenodo, CC BY 4.0), Twente pumps (4TU, CC0). Every figure is drawn from these result files by two scripts (`figures/extract.py`, `figures/figures.py`); no plotted value is typed by hand.

References

1. J. Gao, S. Giri, E. C. Kara, M. Bergés. PLAID: a public dataset of high-resolution electrical appliance measurements for load identification research: demo abstract. Proc. ACM BuildSys '14, 198–199 (2014). doi:10.1145/2674061.2675032.
2. R. Medico, L. De Baets, J. Gao, S. Giri, E. Kara, T. Dhaene, C. Develder, M. Bergés, D. Deschrijver. A voltage and current measurement dataset for plug load appliance identification in households. Scientific Data 7, 49 (2020).
3. L. De Baets, J. Ruyssinck, C. Develder, T. Dhaene, D. Deschrijver. Appliance classification using VI trajectories and convolutional neural networks. Energy and Buildings 158, 32–36 (2018).
4. A. Faustine, L. Pereira, C. Klemenjak. Adaptive weighted recurrence graphs for appliance recognition in non-intrusive load monitoring. IEEE Transactions on Smart Grid 12(1), 398–406 (2021). doi:10.1109/TSG.2020.3010621.
5. N. Helwig, E. Pignanelli, A. Schütze. Condition monitoring of a complex hydraulic system using multivariate statistics. Proc. IEEE I2MTC 2015, 210–215. doi:10.1109/I2MTC.2015.7151267; dataset doi:10.24432/C5CW21.
6. N. Davari, B. Veloso, R. P. Ribeiro, P. M. Pereira, J. Gama. Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry. IEEE DSAA 2021; MetroPT-3 dataset, UCI Machine Learning Repository, dataset 791. (The later MetroPT dataset of B. Veloso et al., Scientific Data 9, 764 (2022), is a different, 1 Hz recording from 2022.)
7. G. Piecuch, T. Żabiński. A new open dataset from a milling process – data for classification and estimation of tool life. Scientific Data 12, 650 (2025); figshare doi:10.6084/m9.figshare.28589216.
8. T. Fonseca, P. Chaves, L. L. Ferreira, N. Gouveia, D. Costa, A. Oliveira, J. Landeck. Dataset for identifying maintenance needs of home appliances using artificial intelligence. Data in Brief 48, 109068 (2023). doi:10.1016/j.dib.2023.109068; Zenodo 7245198.
9. S. J. Bruinsma, R. D. Geertsma, R. Loendersloot, T. Tinga. Motor current and vibration monitoring dataset for various faults in an e-motor-driven centrifugal pump. Data in Brief 52, 109987 (2024); 4TU doi:10.4121/2b61183e-c14f-4131-829b-cc4822c369d0.

10. B. L. Ho, R. E. Kalman. Effective construction of linear state-variable models from input/output functions. Regelungstechnik 14(12), 545–548 (1966).
11. R. Roy, T. Kailath. ESPRIT—estimation of signal parameters via rotational invariance techniques. IEEE Trans. ASSP 37(7), 984–995 (1989).
12. G. H. Golub, V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM J. Numer. Anal. 10(2), 413–432 (1973).
13. O. Ledoit, M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88(2), 365–411 (2004).
14. T. Chen, C. Guestrin. XGBoost: a scalable tree boosting system. Proc. KDD '16, 785–794 (2016).
15. E. S. Page. Continuous inspection schemes. Biometrika 41(1/2), 100–115 (1954).
16. M. Ester, H.-P. Kriegel, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. KDD '96, 226–231 (1996).
17. F. Ghazali, A. Hacine-Gharbi, K. Rouabah, P. Ravier. Performance evaluation of the electrical appliances identification system using the PLAID database in independent mode of house. Proc. ICPRAM 2024.
18. A. Faustine, L. Pereira. Improved appliance classification in non-intrusive load monitoring using weighted recurrence graph and convolutional neural networks. Energies 13(13), 3374 (2020). doi:10.3390/en13133374; code github.com/sambaiga/WRG-NILM.
19. A. Dempster, D. F. Schmidt, G. I. Webb. MiniRocket: a very fast (almost) deterministic transform for time series classification. Proc. KDD '21, 248–257 (2021).
20. A. Dempster, D. F. Schmidt, G. I. Webb. Hydra: competing convolutional kernels for fast and accurate time series classification. Data Mining and Knowledge Discovery 37, 1779–1805 (2023); C. W. Tan, A. Dempster, C. Bergmeir, G. I. Webb. MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. DMKD 36, 1623–1646 (2022).
21. A. Dempster, D. F. Schmidt, G. I. Webb. QUANT: a minimalist interval method for time series classification. DMKD 38, 2377–2402 (2024).
22. M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, A. Bagnall. HIVE-COTE 2.0: a new meta ensemble for time series classification. Machine Learning 110, 3211–3243 (2021); M. Middlehurst, P. Schäfer, A. Bagnall. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. DMKD 38(4), 1958–2031 (2024).
23. V. Feofanov, S. Wen, M. Alonso, R. Ilbert, H. Guo, M. Tiomoko, L. Pan, J. Zhang, I. Redko. Mantis: lightweight calibrated foundation model for user-friendly time series classification. arXiv:2502.15637 (2025).
24. Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, G. H. Chen. ECOD: unsupervised outlier detection using empirical cumulative distribution functions. IEEE TKDE 35(12), 12181–12193 (2023); F. T. Liu, K. M. Ting, Z.-H. Zhou. Isolation forest. Proc. IEEE ICDM 2008, 413–422.
25. M. W. Meckes. On the spectral norm of a random Toeplitz matrix. Electronic Communications in Probability 12, 315–325 (2007).