University of St. Gallen

School of Computer Science

to obtain the title of

Master of Science in Computer Science

Master Thesis on

Self-Supervised Transfer Learning for Acoustic Anomaly Detection in Pumped-Storage Hydropower Plants

---

Submitted by:

Stefan Krummenacher

19-608-991

Approved on the application by:

Prof. Dr. Bruno Rodrigues

and

Dr. Bernhard Bermeitinger

Date of Submission:

22 September 2026

Acknowledgements

I thank my supervisor, Prof. Dr. Bruno Rodrigues, and Jinyuan Zhang for their constant and close guidance throughout this thesis. Their availability at every stage, their careful readings, and their insistence on justified choices over convenient ones shaped this work from the first framing of the question to the final chapter. I also thank my co-supervisor, Dr. Bernhard Bermeitinger, for his support of this work.

I further thank Jürgen Sutterlüti and Benoit Jouan of Gantner Instruments for their support with the measurement chain, and Bernhard Küng of illwerke vkw AG, who made the system deployment and the data access possible on the plant side.

This thesis was carried out within a project supported by Land Vorarlberg, illwerke vkw AG, and Vorarlberg industry, together with the Vorarlberg Chamber of Commerce (WKV) and the Federation of Austrian Industries (IV).

St. Gallen, 22 September 2026 Stefan Krummenacher

Abstract

Pumped-storage hydropower plants balance grids with a high share of intermittent renewables, and their condition monitoring rests on fixed alarm thresholds of unknown false-alarm rate. On an acoustic and vibration campaign at one pump-turbine unit, this thesis asks whether self-supervised pre-training on public sound corpora improves anomaly detection when adapted to scarce plant data. The monitor works in two steps. It identifies the operating state from audio and vibration, label-free by clustering or through a model bank fitted once at commissioning. It then scores each window against that state's normal reference under a per-state conformal threshold, with a finite-sample, distribution-free false-alarm guarantee. The answer is differentiated. A frozen BEATs encoder needs no plant training data and detects 174 of the 180 controlled hammer strikes at a median first-alarm latency below one second, while industrial pre-training recovers the operating-state structure. Adapting either encoder to the plant does not pay, and a distilled 0.78 MB student carries the transient head onto a plant server without a graphics processor. Deployability is decided by the calibration cadence, not the representation. A threshold carried across a change of recording setup loses false-alarm control, reaching a realised rate of 1.00, while daily recalibration is not deployable. Three label-free sentinels decide when that one calibration has expired: the state bank's own rejection rate fires on every foreign setup, and a watchdog on the monitor's realised alarm rate catches within-era threshold decay the others cannot see. Together they hold the realised rate between 0.027 and 0.075 on every held-out day at unchanged strike sensitivity. The evidence base is one machine and one labelled campaign, so what transfers is the procedure and its budget, not the constants fitted here.

Table of Contents

List of Abbreviations

AE	Autoencoder
ARI	Adjusted Rand Index
AUC	Area Under the ROC Curve
BEATs	Bidirectional Encoder representation from Audio Transformers
CWRU	Case Western Reserve University (bearing corpus)
DCASE	Detection and Classification of Acoustic Scenes and Events
FAR	False-Alarm Rate
FPR	False-Positive Rate
HMM	Hidden Markov Model
LOF	Local Outlier Factor
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
MIMII	Malfunctioning Industrial Machine Investigation and Inspection
OC-SVM	One-Class Support Vector Machine
PSHP	Pumped-Storage Hydropower Plant
RMS	Root Mean Square
SCADA	Supervisory Control and Data Acquisition
SSL	Self-Supervised Learning
TF-C	Time-Frequency Consistency
UTC	Coordinated Universal Time

List of Figures

List of Tables

Chapter 1

Introduction

This chapter states the problem that motivates acoustic monitoring of pumped-storage plants, the single research question the thesis pursues, and the six contributions that answer it.

1.1 Motivation

A failing machine often sounds different before it trips an alarm threshold. The machines of interest are pumped-storage hydropower plants (PSHPs), which support grid balancing in electricity systems with a high share of intermittent renewables by pumping water uphill and releasing it on demand [40]. These are large rotating machines, so an unplanned outage is costly twice over: once as the repair, again as lost balancing capacity.

Avoiding such outages depends on noticing trouble early, a task that remains challenging under current practices. Condition monitoring often rests on fixed alarm thresholds, from supervisory control and data acquisition (SCADA) process tags to the vibration limits codified for hydropower machine sets [25]. A rule of that kind fires only once a quantity has drifted far from its setpoint, which can sometimes be too late to act. Listening offers an earlier signal: microphones are low-cost, non-contact, retrofittable to a running unit, and sensitive to early mechanical and hydraulic signatures [6, 28].

The catch is the data. Faults are rare and seldom cleanly labelled, and healthy operation dominates every record, so a learning model sees almost no examples of what it must detect. Each plant also has its own acoustic character: models trained on generic industrial audio transfer poorly to hydropower conditions [28], and portability across plants is reported only with plant-specific modification [45]. Scarcity is the first obstacle.

The second appears only once a detector has to keep running. A threshold fitted on healthy data is valid for the sensing chain it was fitted on, and that chain does not stay fixed: sensors are replaced, cables re-routed, gains adjusted. Scores then move for reasons that have nothing to do with the machine, and the threshold cannot tell the two

causes apart. How often a monitor must recalibrate, and how it notices that its own notion of normal has expired, is therefore as much part of the problem as detection itself.

1.2 Research question and goals

If the obstacle is too few in-domain examples, one way around it is to bring knowledge from elsewhere. This thesis addresses the scarcity problem through transfer learning, guided by a single research question: *Does self-supervised pre-training on public industrial sound datasets improve acoustic anomaly detection when fine-tuned on scarce PSHP data?* An encoder pre-trained on large external audio corpora carries an inductive bias about how sound is structured, and that bias may stand in for the in-domain examples a freshly instrumented plant cannot supply. The study builds on the classical PSHP acoustic baselines of Khamaisi et al. [28], who benchmark from-scratch k-means, one-class support vector machine (OC-SVM), and long short-term memory (LSTM) autoencoder models on Rodundwerk II recordings; the question is what transferred self-supervised representations gain over those.

1.3 Contributions

Answering the research question requires more than a single comparison. This thesis makes six contributions, each stated below as delivered in the subsequent chapters. Chapter 7 checks the list against what was measured.

Operating-state identification. A pumped-storage unit emits distinct sounds per operating state. This work contributes a dual-arm mechanism to reliably recover the machine state from audio and vibration. State transitions are explicitly tracked to ensure transient anomalies during state changes can still be detected.

State-conditioned detection with false-alarm guarantees. Machine state selects the normal reference and decision threshold. This work contributes a distribution-free calibration approach that transforms a chosen false-alarm budget into a strict decision boundary for each state, ensuring the detector's true alarm rate is operationally bound.

Adaptive calibration cadence. A threshold remains valid only for the specific sensing configuration it was fitted on. This thesis introduces label-free drift sentinels that automatically announce when recalibration is required, addressing signal drift without relying on fixed schedules.

A comprehensive transfer-learning ablation. The transfer pipeline is systematically ablated across all eight representations, three adaptation strategies, and both sensing modalities, with every representation carried through state identification, the controlled-event benchmark, and the calibration replay. This exposes the distinct strengths of general-audio versus industrial pre-training for transient versus sustained anomaly detection.

A controlled-event benchmark. The evaluation relies on a newly introduced protocol of 180 controlled acoustic strikes. The evaluation methodology ensures that detection metrics are strictly paired with false-alarm bounds evaluated on out-of-sample normal operations.

A deployable monitoring artifact. The theoretical design is realised as a tested, end-to-end monitoring artifact capable of continuous processing.

Together these contributions close a specific gap: self-supervised transfer, state-aware scoring, and acoustic PSHP monitoring under scarce data have not been combined before, nor has the question of how long such a calibration stays valid once the installation around it changes.

Three boundaries define the scope. The contribution is airborne-acoustic *anomaly detection* through self-supervised transfer, used together with vibration; airborne audio is the deployment *setting* and a constraint, noise-sensitive in this machine room [28] and placement-sensitive, and it is not claimed superior to vibration or other modalities. Source localisation, including beamforming and time-difference-of-arrival, lies outside this thesis. The third is the evidence base: all recordings come from one machine at one site, and the labelled evidence is a single campaign of controlled acoustic events covering two of the four operating states. Portability is therefore argued for and built in rather than demonstrated across plants, so what transfers is the procedure and the false-alarm budget it is given, not the numeric constants fitted here.

1.4 Thesis structure

Chapter 2 states the problem and the sensing setting, positions the work against related research and the commercial market, and closes with a primer on the techniques the design uses. Chapter 3 presents the design, from the measurement setup and preprocessing through the two arms of operating-state detection to the state-conditioned detector, its calibration cadence, and the evaluation and deployment design. Chapter 4 describes

the implemented system and the few mechanisms inside it that decide whether it works. Chapter 5 opens with the campaign, the instrumentation eras it was recorded in, and what the delivered data can support. It then reports the three experiments and the evidence blocks around them: the adaptation ablation, the falsified alternatives, the scarcity measurement and the computational cost. Chapter 6 reads those results as one argument, weighs what the campaign delivered against what the design assumed, defends the results against the suspicion of overfitting, states what would transfer to another plant, and collects the limitations. Chapter 7 concludes and names the open work.

Chapter 2

Fundamentals

This chapter establishes the foundations the method builds on, and it moves from the problem to the tools. Section 2.1 formulates the problem and the sensing setting. Section 2.2 positions the work against related research and states the gap. Section 2.3 reviews what the market already offers and the limits of what it discloses. Section 2.4 then introduces the techniques the design of Chapter 3 builds on, and it comes last because it is the immediate run-up to that chapter.

2.1 Theoretical background

Acoustic anomaly detection for condition monitoring is framed as an unsupervised, normal-only learning task: a model is trained exclusively on recordings of healthy machine operation and must flag deviations from that learned normality at test time [38, 41]. The framing follows from operational reality. Abnormal events are rare, heterogeneous, and expensive to label, so complete supervised fault catalogues rarely exist.

Running such a detector carries two opposing costs, and their asymmetry fixes the operating point. False alarms are the frequent cost. Each one sends an operator to the machine, wastes effort, and, repeated often enough, erodes trust in the system. A missed fault is the rare but far larger cost. It can progress to an unplanned outage, which on a grid-balancing asset is expensive twice over, as the repair and as lost balancing capacity. Detecting a developing fault early is therefore the payoff that justifies monitoring at all, but only if false alarms stay at a level operators tolerate. Behaviour at low false-positive rates matters more than average ranking quality across the whole threshold range, which is why the partial area under the receiver operating characteristic curve (partial AUC) over a restricted false-positive range $[0, p]$ has become the common objective for this task

[38]. Formally, this is detection as a hypothesis test that maximises the detection rate subject to a bounded false-positive rate, following the Neyman–Pearson lemma [29].

PSHP units complicate this picture, because they run in several distinct operating states rather than one steady state. The same machine alternates between pumping and generating, passes through start-up and shut-down, and behaves differently under different loads [40, 45]. Background-noise characteristics and spectral energy shift markedly between stable-load and transient operation [28, 62]. "Normal" is therefore itself multimodal in the statistical sense, and one notion of normality has to cover all of it.

The faults of primary interest in hydro and pump-turbine units are cavitation, bearing wear, and rotor imbalance, all of which manifest in the vibration and acoustic domains [35]. Cavitation arises from force fluctuations in the flow, bearing degradation from mechanical surface imperfections, and imbalance from an unequal distribution of rotor mass about the rotating centreline. The same fault families recur in multi-sensor studies of pumped-storage plants and related rotating machinery [45, 48].

These families do not share one frequency range, which makes the sensing modality a question of complementarity rather than superiority. Cavitation is the difficult case. It is a high-frequency phenomenon, and the accelerometers deployed in this study reach only a few kilohertz (Section 3.2), below the band in which cavitation becomes diagnostic. The established route confirms how demanding that band is: Escaler et al. [15] detect cavitation with piezoelectric accelerometers (with natural frequencies in the tens of kilohertz) or with acoustic-emission sensors resonant near 200 kHz, combining both with envelope demodulation. However, their reported diagnostic bands lie as low as 5–15 kHz. Because the microphones used in this study are rated to roughly 20 kHz, their bandwidth overlaps this diagnostic range. This overlap motivates a central hypothesis of the present work: airborne acoustic sensing may capture high-frequency faults like cavitation without requiring the specialised, contact-mounted sensors conventionally used. This is posed as an open question rather than an established result, given that airborne capture reaches only the lower part of the cavitation band and remains sensitive to background noise and microphone placement. Airborne sensing is attractive more generally because it is low-cost, non-contact, and easily retrofitted to existing assets [1, 6, 28, 36].

Two constraints then bound what is achievable. The first is data scarcity. Healthy operation dominates the record, and labelled anomalies are rare and plant-specific, leaving little material from which to learn a fault model. The second follows from the first. Because PSHP signatures differ substantially from generic industrial machinery,

models trained on generic industrial audio generalise poorly to plant acoustics [28]. The classical PSHP baseline responds by benchmarking from-scratch models on plant-specific Rodundwerk II recordings, and Section 2.2 sets out which models those are and what the baseline leaves open. Whether transfer learning from large pre-trained audio representations can lift both limitations at once is the question this thesis takes up.

2.2 Related work and research gap

Two lines of research bear on this thesis, and every work below is read the same way: what it does, what it leaves open, and how this thesis differs. The first monitors hydropower machinery with classical or supervised models. The second transfers self-supervised audio models, but validates only on generic-machine benchmarks.

The direct predecessor is the classical PSHP baseline this thesis extends, a study on the same plant [28]: k-means, OC-SVMs, and an LSTM autoencoder, each trained from scratch on plant-specific recordings, with no pre-trained representation and no conditioning of the scoring on the operating state. This thesis keeps that plant and that task and changes exactly those two things. Closest to the idea of state-dependent normality is the neural mapping of Zhao et al. [65]. Back-propagation networks map overall vibration levels across the extended operating range of a pump-turbine, so that operating zones accelerating wear can be identified. The evidence is vibration from the plant's installed monitoring, which fixes the sensing points and the band, and the output marks unfavourable zones rather than stating how often a healthy machine will be flagged. This thesis maps normality in the same operating-point sense, but from airborne sound and against an explicit false-alarm target.

The rest of the hydro literature confirms that operating state matters and stops there. Statistical multi-sensor monitoring partitions operation into operating states and fits state-specific thresholds on hand-defined features [45]. Load-aware acoustic monitoring folds operating-state information into purpose-built lightweight models [62, 63]. Both are trained on the target plant, so neither relieves the data scarcity a single plant imposes.

Nearest to the calibration side is Toshkova et al. [52]. Unsupervised learning identifies the operating conditions, and warning and alarm thresholds are then set per condition by fitting generalised extreme-value models to each condition's baseline, on vibration and process data from industrial power plants. Per-condition thresholds are exactly the structure adopted here; what differs is the guarantee. An extreme-value fit models the tail of a baseline, whereas the thresholds in this thesis are read off held-out normal

scores and carry a finite-sample, distribution-free coverage statement, the split-conformal construction introduced in Section 2.4. Modality and scale differ as well: while Toshkova et al. [52] rely on vibration and process channels across multiple plants, the present work focuses on airborne sound from a single machine.

The second line asks a different question: how far a pre-trained audio model can be reused with little task-specific training. Frozen embeddings from BEATs, a transformer encoder pre-trained on the large public general-audio corpus AudioSet, paired with a nearest-neighbour back-end reach training-free detection that rivals fine-tuned systems [12, 44]. Low-rank adaptation of the same backbone, which trains a small added update per layer instead of every weight, has been reported to match or exceed full fine-tuning [66]; Section 2.4 sets out the machinery behind both. The prevailing discriminative paradigm shapes embeddings into compact sub-clusters before applying distribution-based scoring [61]. All three transfer strategies (frozen features, low-rank adaptation, and sub-cluster embeddings) are evaluated on Detection and Classification of Acoustic Scenes and Events (DCASE) benchmarks and generic-machine corpora, where the recordings come from benchmark rigs and ample normal training material exists for each machine type. This thesis transplants them into the opposite situation: one plant, several operating states, and the pump-to-turbine transition sitting inside the data a detector must call normal.

Two recent studies lie at the intersection of these fields, but leave the core gap open. One applies a ResNet with self-attention to hydro-turbine acoustics, but remains a *supervised* fault classifier rather than a normal-only detector [60]. The other fine-tunes a pre-trained audio spectrogram transformer on *acoustic-emission* signals from a hydraulic test rig, again in a supervised setting and from supervised rather than self-supervised pre-training [51]. Each holds one ingredient, hydro acoustics in the first case and pre-trained audio transfer in the second, and neither couples both with state-aware, normal-only detection. A forward-citation sweep over the anchor works of both lines found no study combining self-supervised audio transfer with operating-state-aware scoring for hydropower acoustics under scarce local data. The claim rests on an absence of counter-evidence, not on proven impossibility. That intersection is the position of this thesis, and Chapter 3 builds the method that occupies it.

2.3 Commercial solutions

A fielded product could in principle serve as a baseline, so the market deserves a look. Hydropower-specific condition monitoring already covers acoustics. The closest match is Voith OnCare.Acoustic, which pairs an acoustic sensor with an edge gateway and cloud analytics and sorts machine sound into warnings and alarms, with a reported deployment at the Budarhals plant in Iceland [56]. A related pilot at the same site targets cavitation [57]. ANDRITZ covers process signals, performance trending, and cavitation monitoring with Metris DiOMera [3] and vibration condition monitoring with the separate Metris Vibe [4], and a smaller offering fuses acoustic, vibration, and efficiency data into a single runner-health indicator [39]. Beyond hydropower, acoustic, multimodal, sensor-agnostic, and contact-vibration products target rotating equipment in general [2, 7, 8, 9, 37, 46]. They are relevant here only because their fault scope — bearings, shafts, pumps, and motors — recurs in hydropower units.

Vibration-based monitoring differs from all of these in one respect: its methods are public. For hydropower and pump-storage machine sets, ISO 20816-5 defines how shaft and bearing vibration is measured and which levels indicate acceptable operation [25]. The wider ISO/TC 108 family specifies condition-monitoring programmes, vibration signature analysis, and diagnostic interpretation [22, 23, 24]. These documents disclose their procedures in full. No standard yet covers deep-learning acoustic anomaly detection, so data-driven acoustic methods sit outside the established consensus.

The decisive observation is what none of the products states. They describe an output, a warning, an alarm, or a health score, but not the features, the model class, or the decision logic. Siemens Senseye names technique categories without giving architecture or evaluation [46]. 3DSignals alone names classical signal-processing steps beside its proprietary classifiers, still without training data or metrics [8]. Reported benefits come from vendor material and single case reports, rarely with detection rates, false-alarm rates, or independent validation.

The peer-reviewed literature confirms the pattern: documented predictive maintenance in hydropower is rare [10], acoustic detection is studied mostly at laboratory scale [42], and fielded maintenance models tend to be opaque [53]. Open work closes only part of the gap: it documents hydropower condition monitoring in full [16] and supplies open acoustic benchmarks [38, 41], none of them specific to pumped-storage hydropower. No fielded product can therefore serve as a reproducible baseline. Chapter 3 answers

this disclosure gap together with the research gap of Section 2.2, with a method whose representation, calibration, and evaluation are stated in full.

2.4 Technical background

Pre-training, self-supervision, and transfer. Three ideas underpin the approach. *Pre-training* trains a model on a large dataset before the target task, so it learns general structure a small target dataset cannot provide. *Self-supervised* learning does this without human labels, building the learning signal from the data itself, for example by predicting masked parts of an audio clip. *Transfer learning* then reuses the pre-trained model on a new, smaller task. Together they make a large public corpus usable for a data-scarce target such as a single plant.

Two pre-trained representations. The candidates differ in the public data they learn from. BEATs is a transformer pre-trained self-supervised on AudioSet, a large public general-audio corpus [12]. Time-Frequency Consistency (TF-C) is a self-supervised objective for time series: it learns embeddings by encouraging the time-domain and the frequency-domain view of the same signal to agree, so the representation captures structure consistent across both [64]. TF-C can instead be pre-trained on public industrial sound such as MIMII, a dataset of normal and malfunctioning machine sound used both as an anomalous-sound-detection benchmark and as pre-training material [41]. BEATs reads audio only, so a separate branch handles the vibration signal, with handcrafted features or with TF-C, which as a time-series method applies to vibration as well. General audio is included deliberately, because the largest public audio corpora come from it.

Adapting a pre-trained model. How much of a pre-trained backbone to retrain is a spectrum, from a frozen encoder through a small trainable low-rank update per layer, low-rank adaptation (LoRA), to full fine-tuning of every weight [19]. Figure 2.1 draws it; Section 3.4.3 states which options a data-scarce plant can afford.

Scoring, calibration, and deployment. Three further pieces are named here and developed in Chapter 3. Because the training data contain only normal examples, the score measures how far a new sample lies from what was learned as normal: the distance to the nearest normal embeddings [44], the Mahalanobis distance to the normal distribution [18], or the reconstruction error of a model that rebuilds its input [13]. Section 3.4.4 names the family this design uses.

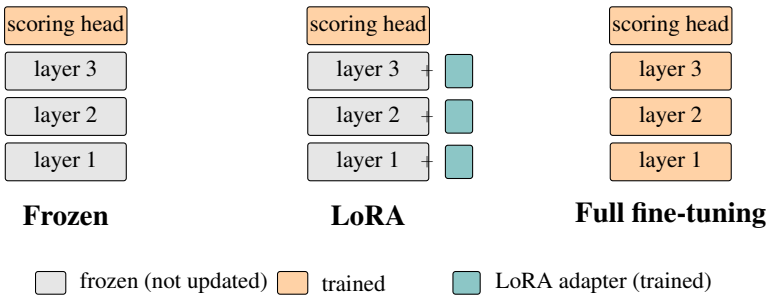

Figure 2.1: The adaptation spectrum. A frozen encoder trains only a small scoring head; full fine-tuning updates every weight and is data-hungry. LoRA sits between them, adding a small trainable low-rank adapter to each layer while the pre-trained weights stay frozen. For a pre-trained backbone with scarce target data, frozen and LoRA are the practical options.

A score is not yet a decision. Conformal prediction sets the threshold from held-out normal data and, under an exchangeability assumption, attaches a distribution-free guarantee on the false-alarm rate [5, 30]; Section 3.4.5 states what it costs and why exchangeability decides how the calibration set is chosen. A strong representation may finally come from a model too heavy for a plant server, so knowledge distillation trains a small student to reproduce a larger teacher's behaviour [17], which Section 3.6 weighs against post-training quantization.

Chapter 3

Design

A pumped-storage unit produces almost no faults and many distinct normal sounds, and a fresh installation offers only a short record of healthy operation before a monitor must run. Under that constraint the chapter builds the design that answers the research question of Chapter 1: self-supervised transfer, combined with state-aware embedding scoring and per-state calibration, on a pump-turbine where target-normal data are scarce.

The baseline is classical and multimodal: a rich set of physics-informed features, one-class detectors scored individually and as a majority vote, and per-state false-alarm control. Self-supervised transfer is the layer this thesis adds on top, not a replacement for it.

One concrete pipeline is the reference: a frozen general-audio encoder (BEATs) embeds each audio window, the embedding is scored by its distance to a per-state normal reference, and a per-state conformal threshold decides. Prior work shows that several choices inside that pipeline change the outcome, so the study varies each as an axis instead of fixing it by assumption.

The chapter follows the runtime flow: the reference pipeline and its design axes (Section 3.1), the measurement setup and preprocessing (Section 3.2), Step 1 in its two arms (Section 3.3), Step 2 with its calibration cadence and drift sentinel (Section 3.4), how the detector is judged (Section 3.5), and how it becomes small enough to run at the plant on scarce healthy data (Section 3.6).

3.1 System overview

A recording becomes an anomaly decision through a standard condition-monitoring pipeline (Figure 3.1) [54]. This thesis contributes to specific stages of this pipeline rather than its structure.

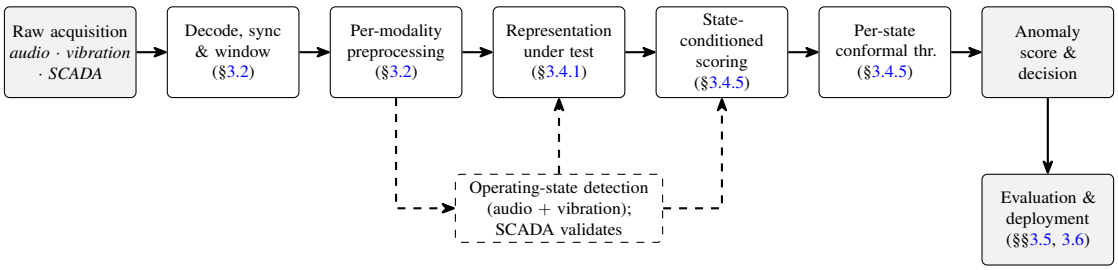

Figure 3.1: End-to-end data pipeline. The monitor works in two steps: a first step (dashed) detects the operating state from audio and vibration, and a second step scores each window against the normal reference of that state before a per-state conformal threshold decides.

Raw acquisition collects audio, vibration, and SCADA channels; logs are aligned to a shared timebase, windowed, and preprocessed. Step 1 infers each window's operating-state label from sound and vibration (Section 3.3). The representation stage (Section 3.4.1) extracts embeddings or features, state-conditioned scoring (Section 3.4.5) compares each window against the normal reference of its detected state, and a per-state conformal threshold yields the decision.

Table 3.1 names the design axes varied in this study. Varying one axis while the rest stay fixed isolates the effect of representation transfer, multimodality, or conditioning under scarcity. How the system is judged is not an axis but a commitment, settled in Section 3.5.

3.2 Data and preprocessing

The pipeline is grounded in a multi-channel acoustic and vibration campaign on a pump-turbine unit at the Rodundwerk II plant. Nine microphones (50 kHz) and two tri-axial accelerometers deliver 15 measurement channels (Figure 3.2). The acquisition system records 21 channels because it is configured for four accelerometer input groups, two of which hold no sensor. Both accelerometers sit on the turbine level, at the 0° and 90° positions of the turbine ring, as recorded in the operator's installation log of 25 June 2026. The two vibration streams keep the acquisition system's generator and turbine labels, which name recording groups rather than levels. The hammer strikes of Section 5.2.2 confirm the placement: a strike at the turbine-ring 0° or 90° position registers roughly six to eight times stronger at the accelerometer at that position than at the other, while strikes at the generator-ring positions stay at the noise floor of both. Process and SCADA

Table 3.1: Design axes of the comparison. Five axes, each varied around the reference option set in bold; deployment carries no bold reference because both variants ship (Section 3.6).

Axis	Options	What it compares	Section
Representation	handcrafted, from-scratch, TF-C, **BEATs**	pre-trained vs not; general vs industrial pre-training data	§3.4.1
Modality	**audio**, vibration, fusion	whether vibration adds over audio; fusion vs single modality	§3.4.2
Adaptation	**frozen**, LoRA, full fine-tuning	how much to adapt under scarce data	§3.4.3
Scoring	one-class, reconstruction, **embedding distance**	which back-end on the same representation	§3.4.4
Operating state	state-agnostic, **per-state**, per-state sub-cluster	whether state conditioning lowers false alarms	§3.4.5
Deployment	distillation, quantization	a compact on-premise model	§3.6

signals (shaft speed, active power, pressure) log the operating point and only validate operating-state detection; they never score anomalies at run time.

The two modalities cover complementary bands: accelerometers are usable only to roughly 3.7 kHz, missing high-frequency cavitation signatures, while the 50 kHz microphones respond up to 20 kHz. Preprocessing is tailored to each representation's requirements.

Audio (pre-trained backbones). The AudioSet-pretrained transformer (BEATs) and the 2D variants of TF-C ingest audio transformed into 128-dimensional log-Mel spectrograms at a 16 kHz sampling rate [6, 12, 28]. This specific downsampling is required because these backbones were originally pre-trained on human-centric and environmental datasets that standardise on 16 kHz, forcing the plant's audio to match their expected input space. The 1D variant of TF-C operates directly on the raw time-domain signal and its frequency-domain equivalent.

Handcrafted baseline. Because the 16 kHz downsampling for the SSL models discards all acoustic content above 8 kHz, the original 50 kHz full-band audio signals are retained for the baseline. This allows the handcrafted extraction to capture high-frequency phenomena (such as early-stage cavitation or mechanical scraping) that the pre-trained transformers are completely blind to [15]. The spectral estimates are split by design:

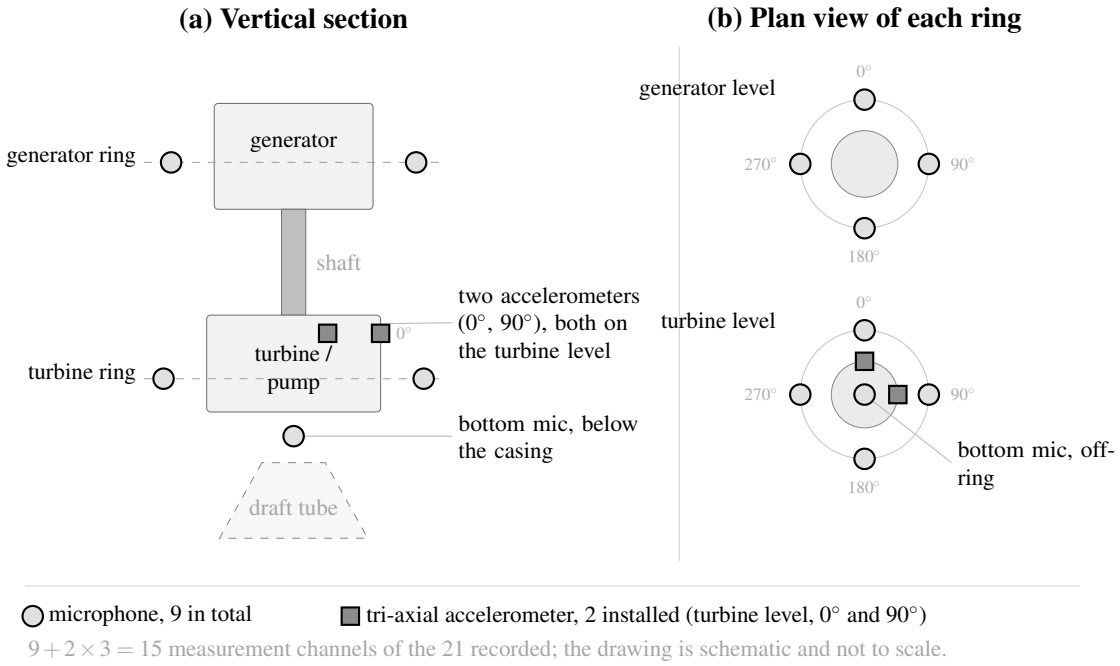

Figure 3.2: As-built sensor geometry on the machine set. The two installed accelerometers sit on the turbine level at the 0° and 90° microphone positions; the two further accelerometer input groups of the acquisition system carry no sensor. The recording configurations changed during the campaign, dividing it into three instrumentation eras (Section 5.1.1).

broadband features trade frequency resolution for lower variance via an averaged Welch estimate, whereas narrowband machine-frequency features demand a high-resolution, single-segment estimate to guarantee at least one genuine bin per physical band.

Vibration. Vibration features are extracted via spectral harmonics and Hilbert-envelope demodulation [1]. Raw time-domain vibration is typically dominated by the machine's fundamental rotation speed; demodulation strips this away to expose the subtle, high-frequency structural resonances excited by bearing defects or mechanical impacts.

Process data. SCADA channels are recorded at lower, often irregular cadences (e.g., once per second or on-change). They are resampled to the shared timebase and normalised to align perfectly with the short acoustic windows, ensuring the operational state validation exactly matches the audio segment under test.

Scoring input. Finally, the classical baseline scorers (such as OC-SVM, Isolation Forest, and kNN) do not ingest raw data. Anomaly scoring algorithms suffer from the curse of dimensionality on raw waveforms (a one-second window carries 50 000 samples per

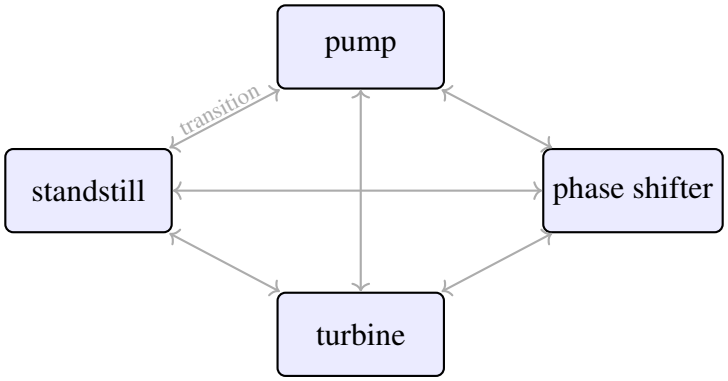

Figure 3.3: The operating-state model. Nodes denote steady operating states, while edges represent transitions (kept as a separate segment class rather than discarded). Within a state, load variability appears as sub-clusters. Every steady state carries its own normal reference and conformal threshold.

channel at the plant's 50 kHz rate); instead, they operate purely on the highly compressed, descriptive embeddings or feature vectors produced by the extraction pipelines, making distance calculations tractable and robust.

3.3 Step 1: Operating-state detection

A pumped-storage unit's acoustic signature shifts with its operating state (pump, turbine, standstill, phase shifter) and load [28, 62]. A state-agnostic detector pooling every operating state into one normal reference is exposed to the condition shift the anomalous-sound literature documents [38], which is why multi-sensor plant monitoring partitions operation by operating state [45]; pooled scoring can then mistake a load-driven change for an anomaly. The state must therefore be recovered before scoring. Deriving this state purely from sound and vibration, rather than relying on live SCADA feeds, ensures the monitor can run as an independent, non-intrusive retrofit system.

The monitor infers the state from audio and vibration; Figure 3.3 sketches the state model it works with. Two arms implement this step, differing in their reliance on commissioning labels:

- **Clustering arm**: Unsupervised. k-means clusters the fused features, smoothed by a sticky hidden Markov model and a minimum dwell-time filter. A five-second dwell floor is derived directly from the plant's statistics: genuine state transitions take between 20 and 138 seconds, whereas transient turbine load-step blips typically flicker for six to nine seconds, allowing a short filter to absorb the noise without

missing a real transition. The clusters are named via majority vote over the fit day's operational labels; labels touch only this naming step and the evaluation, never the fitting, and without them clusters keep numeric identifiers.

- **Model bank arm**: Fits one small normality model per operating state on windows carrying operational labels. At run time, a window goes to the state whose model fits best. Windows that fit no model well are flagged as `no_mode_fits` (Section 3.4.6).

The bank arm requires one day of operational labels for commissioning but agrees better with the operational reference; the clustering arm trades that agreement for independence from labels. Both arms are mechanisms, not representations: each consumes the feature streams of Section 3.2, so any representation of Section 3.4.1 can drive either arm, and the evaluation runs both arms under every representation (Section 5.2.1). The output is one operating-state label per analysis window.

3.4 Step 2: State-conditioned anomaly detection

Step 2 takes the operating-state label from Step 1 and runs a state-conditioned anomaly detector built from three components: representation extraction, adaptation strategy, and scoring back-end. The state label routes the window to the correct normal reference and threshold (Section 3.4.5). Table 3.2 details the pre-training and plant-side fitting for each representation.

3.4.1 Representation

The representation dictates how the signal is encoded. Four poles are evaluated, forming a gradient from physical priors to large-scale transfer learning:

1. **Handcrafted features**: Time-domain statistics, spectral harmonics, and Hilbert envelopes encode prior physics knowledge [1, 31]. They are interpretable but presuppose known fault taxonomies.

2. **From-scratch deep models**: An autoencoder trained purely on target-domain normal data [28]. Its success is bound by the availability of target data.

3. **Industrial SSL (TF-C)**: Pre-trained on public machine sounds (MIMII), transferring industrial inductive biases to the pump-turbine [41, 64].

4. **General-audio SSL (BEATs)**: Pre-trained on AudioSet [12], this backbone offers powerful embeddings even when kept frozen [44].

Table 3.2: Training and data view of the representations. Self-supervised pre-training happens offline on public data; on the plant, only the detectors, adapters, or a fusion head are fitted.

Branch / repr.	Pre-trained on	Target-side (PSHP) step	Input format	What is trained
Handcrafted (both)	none	one-class detector fit	RMS, kurtosis, harmonics	detector only
From-scratch deep (audio)	none	trained on PSHP normal	16 kHz log-Mel / spectra	full model on PSHP
TF-C (audio SSL)	MIMII; scratch control: none	adapted: frozen / LoRA / full	time- and frequency-view	adapter or full
BEATs (general SSL)	AudioSet (general)	adapted: frozen / LoRA	16 kHz log-Mel, 128 bands	backbone frozen, or LoRA
Vibration (TF-C SSL)	CWRU, Paderborn	adapted: frozen / LoRA / full	time- and frequency-view	adapter or full
Fusion head	frozen branch outputs	trained on PSHP	concatenated embeddings	fusion head only

This contrast tests a central tension in transfer learning: whether a model pre-trained specifically on industrial machines generalises better to a novel turbine than a massive foundation model exposed to a completely unstructured, universal soundscape.

The vibration branch uses handcrafted features or a vibration-native TF-C model (pre-trained on bearing datasets [47]) to keep the core audio SSL question separate. The no-pre-training counterpart exists on both sides: for audio as the from-scratch pole above and additionally as a control that trains the TF-C architecture from scratch on plant audio alone (Section 5.2.5); for vibration, the handcrafted features are the no-pre-training option, and no from-scratch deep vibration model is trained.

3.4.2 Multimodal detection

Airborne sound and structure-borne vibration provide distinct, overlapping views of mechanical events. Multimodality is built into the pipeline to test whether audio-only anomaly detection (the primary focus of the SSL backbone) is sufficient, or if vibration adds indispensable diagnostic power [58]. Audio-only, vibration-only, and fusion detectors compete equally. Fusion is evaluated at the feature level (concatenating descriptors before scoring) and via a cross-attention head. Feature-level fusion strictly requires per-modality standardisation because acoustic and accelerometer levels have incomparable

scales; an unscaled concatenation would allow the modality with the larger numerical range to dominate every distance calculation.

3.4.3 Adaptation

For the pre-trained poles (BEATs, TF-C), the adaptation strategy dictates how much of the network is updated on the scarce plant data. Three approaches are compared:

- **Frozen encoder**: The backbone is locked and only a lightweight scoring back-end is fitted. Under severe data scarcity, this often matches or exceeds fine-tuned systems [44], making it the default.
- **Low-Rank Adaptation (LoRA)**: Offers a middle ground by injecting and training small rank-decomposition matrices while keeping the backbone frozen, updating only a fraction of the parameters [19, 66].
- **Full fine-tuning**: Updates the entire network. This bounds the upper capacity but is highly prone to overfitting on scarce normal data, serving as a reference point rather than a deployable candidate.

3.4.4 Anomaly scoring

The scorer is responsible for translating a window's encoded representation into a single scalar anomaly score. Conceptually, this works by building a reference profile of "normal" behaviour during the calibration phase; at run time, the scorer calculates how far a new window deviates from that known normal reference. Three algorithmic families are considered:

- **Classical one-class detectors**: OC-SVM [11] and Isolation Forest [32], which attempt to draw explicit decision boundaries around the normal data.
- **Reconstruction error**: Deep autoencoders that learn to reconstruct normal data and flag inputs they fail to reconstruct well [13].
- **Embedding distance estimators**: Nearest neighbour (kNN) and Mahalanobis distance, which rely purely on geometric distances in the representation space [18].

Embedding distance (kNN with cosine metric or Mahalanobis) is chosen as the reference back-end for the SSL models. The reason is structural: SSL models are

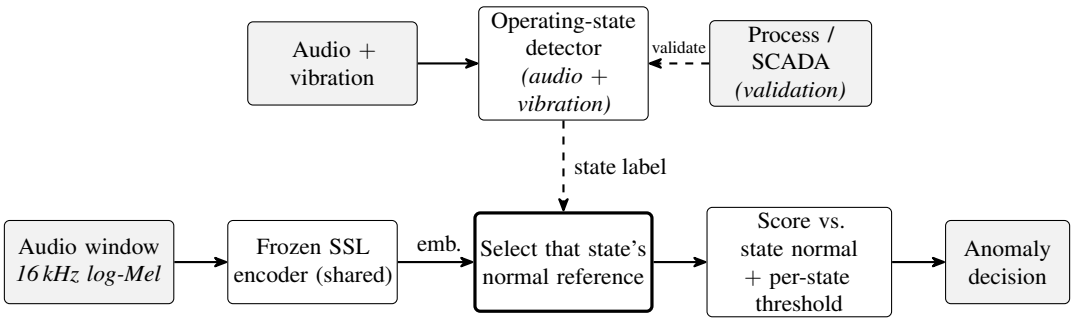

Figure 3.4: Two-step gated architecture. Shown here for the SSL reference pipeline: one shared encoder serves every state. *Step 1* detects the state. *Step 2* embeds the audio; the state selects the normal reference, and a conformal threshold decides. For the classical baseline, the embedding step is replaced by handcrafted feature extraction and the normal reference by a majority-voting ensemble, though the state-routing principle remains identical.

pre-trained to map semantically similar sounds to tightly clustered coordinates in a high-dimensional metric space. Because the space itself is already well-structured, simple geometric distance estimators are highly effective and avoid the over-fitting risks of complex boundary-drawing algorithms like OC-SVM. Keeping this distance back-end strictly disjoint from the encoder is also what permits the powerful representations (like BEATs) to remain completely frozen without requiring downstream fine-tuning.

Conversely, the classical handcrafted baseline employs a majority-voting ensemble of OC-SVM, Isolation Forest, and an LSTM autoencoder to suppress individual false alarms across the handcrafted features.

3.4.5 State conditioning and thresholds

A state-agnostic threshold exposes the monitor to false alarms driven by load changes [62]. Instead, the detected state (including load sub-clusters) routes the window to a state-specific normal reference and threshold (Figure 3.4). State-agnostic scoring is evaluated alongside it strictly as a baseline counterfactual, to measure exactly how much false-alarm control is lost when this routing is removed.

Thresholds are set using inductive split-conformal prediction [5, 30]. Rather than assuming a Gaussian distribution, conformal calibration provides a finite-sample, distribution-free guarantee: with n calibration windows, the realised false-alarm rate is pinned to within $1/(n+1)$ of the nominal target α. In practice, this allows the plant operator to simply define a tolerated false-alarm budget (e.g., 5%), and the threshold automatically

shapes itself to the data to guarantee exactly that rate without requiring any manual tuning. Windows matching no known state confidently are routed to a conservative fallback. Transitions are treated as their own segment class if labelled references exist, or flagged as `near_transition` to optionally suppress spurious alarms.

3.4.6 Calibration cadence and the drift sentinel

A threshold is only valid if the sensing chain remains stable. The instrumentation eras (Section 5.1.1) break this assumption: a threshold carried across an era boundary no longer controls its false-alarm rate, the failure E3 measures (Section 5.2.4). Rather than recalibrating daily (which defeats the label-free premise), the system recalibrates discretely, triggered by a *drift sentinel*.

The sentinel system monitors the plant's stability without labels, prompting an era-level recalibration when three checks fail:

- **Mode-fit sentinel (s1)**: Monitors the `no_mode_fits` rate from the model bank against a bootstrap quantile of the commissioning day's rate. When the machine setup shifts, the bank's members fail to fit incoming windows, the rejection rate spikes, and the sentinel fires.
- **Level sentinel (s2)**: Watches the microphone level against a robust margin and cross-checks it against vibration. This attributes a one-modality volume step directly to the sensing chain (e.g., gain changes) rather than to the machine.
- **Alarm-rate sentinel (s3)**: Watches the monitor's own realised alarm rate under the frozen thresholds against a fixed factor of twice the nominal budget. Because score distributions can drift even inside a recognised operating state (invisible to the other two sentinels), the label-free alarm rate reads this failure directly.

Attempts to adapt continuously (such as normalising features individually per recording, continually shifting thresholds over time, or fine-tuning the neural network on the live data stream) were falsified during the evaluation; they conflate genuine machine degradation with sensor drift and inadvertently "normalise away" the anomaly signal. The vibration branch, unaffected by microphone gain changes, requires no recalibration and provides era-robust redundancy.

3.5 Evaluation design

The evaluation protocol defines how detector performance is measured against real-world constraints. Table 3.3 positions this design against three reference practices, highlighting how it explicitly tests state conditioning, finite-sample guarantees, and era-robustness while requiring no runtime labels. The evaluation groups validation windows by contiguous recording segments to prevent data leakage between overlapping windows [27, 43].

Table 3.3: Where this design sits against three reference practices. The fixed-threshold baseline matches the transient-detection rule running in the parallel source-localisation project and is measured against this design in Section 5.2.2 (hand-set MAD threshold).

Property	Fixed threshold	Global one-class	Ranking metrics only	This work
Threshold conditioned on operating state	no	no	no operating point	yes (§3.4.5)
Finite-sample, distribution-free guarantee	no	no	no	yes (§3.4.5)
Calibration cadence for setup changes	not addressed	not addressed	not addressed	once per era (§3.4.6)
No labels needed at run time	yes	yes	n/a	yes (§3.3)
States reported in operator vocabulary	no	no	no	yes (§3.3)
Separate transient/sustained alarm paths	transient only	no	no	yes (§3.5)

The split is three-way: data is separated to fit per-state normal references, establish the conformal calibration quantile, and evaluate the scored windows. Cross-era results are strictly isolated to measure drift, not primary performance.

Because genuine, cleanly labelled faults are rare in production hydropower plants, the evaluation relies on four complementary pillars (Table 3.4). The verdict therefore never depends on a single source of evidence.

The controlled-event campaign provides the labelled benchmark. It is scored per strike: detection within a tolerance window of each registered strike, first-alarm latency, and a false-alarm rate measured strictly outside the padded strike intervals. To prevent deflation, these intervals are strictly excluded from calibration.

Table 3.4: The four evaluation pillars. Ordered by the label assumption each requires.

	Pillar	What it measures	Source
1	Per-state false-alarm rate on normal	FAR held under each state's own normal	[14, 29]
2	Degradation / novelty trend	drift away from learned normal, without labels	[11, 59]
3	Controlled acoustic events	controlled or simulated labelled cases; gap to real faults declared	[21, 28]
4	Real fault case studies	direct confirmation where labels exist	(plant data)

The decision layer translates the raw scores (calculated every second) into actual operational alarms. It uses two parallel mechanisms to solve the trade-off between catching short events and avoiding noise:

- **Sustained alarms (Persistence)**: This path targets slowly developing degradation (e.g., a failing bearing). Because it is meant for long-lasting faults, it demands that the anomaly threshold is breached for several seconds *in a row* before sounding an alarm. This "persistence" requirement suppresses random statistical noise [26], allowing the system to use a relatively sensitive threshold (e.g., the nominal budget $\alpha = 5\%$) without flooding the operator with false alarms.

- **Transient alarms (Impulse path)**: Short, impulsive events (like the acoustic hammer strikes) only last a fraction of a second; the persistence filter above would completely ignore them. To catch them, this parallel path triggers an alarm immediately on a single anomalous window. However, because it lacks the protection of a duration filter, it enforces a much stricter statistical threshold (e.g., $\alpha = 0.1\%$) so it only fires on extreme outliers.

3.6 Deployment and data-scarcity design

A production monitor must run on low-resource, on-premise edge servers without dedicated GPUs, and it must be calibratable from limited commissioning data.

3.6.1 Deployment and compactness

Compactness is a hard requirement. The heavy audio backbones (e.g., BEATs) dominate the inference cost compared to lighter encoders (e.g., TF-C). To achieve an edge-feasible

footprint (a constraint that both the acoustic edge-detection and the industrial IoT literature treat as binding [34, 50]), the design compares two compression strategies:

- **Knowledge distillation**: Training a lightweight, fast student model to replicate the heavy teacher's anomaly scoring behaviour [17].
- **Quantization**: Post-training 8-bit integer (INT8) quantization of the model weights to reduce memory bandwidth and latency.

Detection performance is then explicitly weighed against size, memory footprint, and latency.

3.6.2 Data-scarcity analysis

Given the sheer dominance of normal operation recordings in PSHPs, the primary challenge is determining the absolute minimum healthy data required to establish a reliable detector. The core hypothesis is that the inductive bias of a pre-trained encoder should substitute for missing target examples [6, 33].

This hypothesis is tested via a *data-scarcity curve*. The anomaly detectors are evaluated across varying fractions of target-normal training data (from 5%, eight clips, up to 100%). This curve answers two key questions:

- **The pre-training gap**: The vertical gap between the SSL models and the from-scratch models measures the exact value of the pre-trained weights.
- **The saturation point**: The horizontal plateau identifies the "enough" threshold (how many hours of data a commissioning engineer actually needs to collect per state).

The curve's left extreme models the zero-shot or first-shot setting [38], testing the expectation that frozen pre-trained embeddings stay competitive even at minimal data fractions and degrade gracefully [44].

Chapter 4 describes how this design was built, and which mechanisms inside the implementation decide whether it works.

Chapter 4

Implementation

The design of Chapter 3 was implemented from scratch as one self-contained repository: the reader for the plant's binary logger format, the window grid, the featurizers, the state layer, the anomaly layer, the runtime artifact and the evaluation harnesses are all new code. This cohesive architecture ensures deployment feasibility: a monitor intended for real-world calibration and continuous operation must function as a single software artifact rather than a disconnected collection of analysis scripts.

To ensure correctness, the implementation is validated through strict static type checking and a comprehensive suite of over 1,600 automated tests. This rigorous validation is critical because many pipeline errors in anomaly detection fail silently, yielding mathematically incorrect anomaly scores rather than explicit crashes.

Figure 4.1 shows how the components connect, and Table A.1 in Appendix A resolves each fundamental concept of Chapter 3 to a named module. This chapter therefore stays on the six mechanisms that decide whether the system works at all: the derived time base, the label-free state rejection, the conformal threshold with its data floor, the leakage guard, the event-free calibration rule and the drift sentinel. Two are quoted as the code that runs, because there the whole argument is a few lines long; the other four are named by function and file.

The chapter follows the data path: ingestion and the time base (Section 4.1), the representations (Section 4.2), the operating-state layer (Section 4.3), the anomaly layer (Section 4.4), and the runtime and evaluation tooling (Section 4.5).

4.1 Data ingestion and the time base

The plant logger writes a proprietary binary container comprising metadata, a channel-descriptor block, and continuous frames consisting of an unsigned 64-bit nanosecond

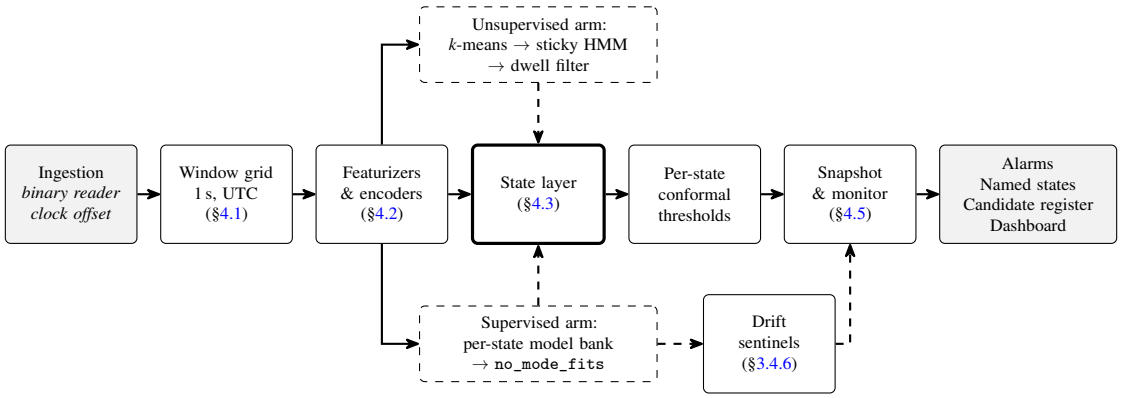

Figure 4.1: One dashed edge, from the model bank to the sentinels, is what makes recalibration self-triggering rather than scheduled. Solid arrows carry every window; dashed arrows carry control signals that need no labels at run time. The state layer has two interchangeable arms, and the supervised arm additionally exports its whole-bank rejection rate, which is the quantity the first drift sentinel watches.

timestamp followed by one 32-bit float per channel. The ingestion reader parses these blocks and strictly validates token lengths to prevent silent channel dropping or corruption.

Recording discovery parses directory structures without reading file payloads to efficiently index several hundred gigabytes of data. Continuous measurement sessions are automatically segmented wherever the gap between consecutive file bursts exceeds fifteen minutes, indicating that acquisition was physically stopped and restarted. A quality guard rejects any session where more than 5 % of the windows are invalid, which successfully removes short, unusable restart fragments from the evaluation set. One session exceeds that guard for a different reason: the 30 June turbine morning carries 8 % invalid windows from two channel gaps that other streams bridge, and it is admitted with a documented override and its invalid windows masked (Section 5.1.1), because the guard exists to reject restart artefacts, not masked channel gaps.

The acquisition hardware's clock required an algorithmic correction, as its internal timestamps were misconfigured to count local-time seconds since the year 2000 while declaring them as Unix nanoseconds. To handle this robustly, `src/rowii/io/dataset.py` computes a robust median time shift across all files in a run, rounds it to the nearest whole hour, and applies the correction dynamically. This approach guarantees that mislabelled outlier files are ignored and that future datasets with correctly configured clocks are not incorrectly shifted. The second time-base repair concerns the controlled-event protocol rather than the sensors: the strike log was recorded in local time against sensor timestamps

in UTC, and the two-hour offset is corrected once at ingestion of the ground-truth tables; Section 5.1.3 states what missing this offset would cost.

The correction is applied at the ingestion point, which guarantees that all downstream components operate in true UTC. The pipeline then constructs a one-second UTC window grid across the intersection of all active streams. Because whole sessions exceed system memory, the data is loaded and featurized in chunks. Each window is tagged with a validity flag and its originating burst-segment identifier. This identifier is strictly preserved through the pipeline to ensure that future train-test splits remain disjoint and leakage-safe.

4.2 Features and representations

The implemented system processes two distinct feature pipelines: a physics-informed handcrafted baseline and a set of learned neural representations.

Handcrafted featurizers. The handcrafted branch extracts deterministic features from both audio and vibration, keeping every channel strictly distinct to preserve position-specific signatures.

- **Audio**: The pipeline extracts 135 columns across nine microphone channels (four on the generator ring, five on the turbine ring). Features include a log level, spectral centroid, 95-percent rolloff, and energy per octave and machine-frequency band. These are fed by two distinct spectral estimates: an averaged Welch estimate for broadband stability, and a high-resolution, single-segment estimate for precise narrowband machine frequencies.

- **Vibration**: Of the twelve accelerometer inputs, eight pass the liveness rule; the pipeline extracts 96 columns from them, computing a log level, kurtosis, and band energies up to 4 kHz per input. Six of those inputs belong to the two installed sensors; the other two are sensor-less inputs at the electronic noise floor, so 24 of the 96 columns carry no machine information. Inputs with zero variance across a batch are dropped rather than zero-filled, which keeps constant inputs out of the distance computations. On 8 July a ninth input becomes intermittent and so passes the rule, which widens the vector to 108 columns; the model bank of Section 5.2.1 refuses the widened set as outside its feature contract, whereas the deployed monitor described below drops the novel columns.

- **Fusion**: The two modalities are concatenated into a single 231-dimensional vector (135 audio and 96 vibration columns, the 24 sensor-less columns included). To

prevent the modality with the larger numerical range from dominating the distance metric, strict per-modality standardisation is applied prior to concatenation.

Learned representations. Three neural backbones provide the alternative, data-driven embeddings:

- **General-audio encoder (BEATs)**: A transformer backbone with 90.4 million parameters (361.5 MB on disk). The implementation provides two deployment paths for this model: a strictly frozen encoder (no plant-specific training) and a LoRA-adapted version that injects trainable low-rank matrices for domain fine-tuning.
- **Distilled student**: A compressed student network with 192,643 parameters (0.78 MB), trained to mimic the frozen BEATs teacher's embeddings for edge deployment.
- **Industrial encoder (TF-C)**: A contrastive model pre-trained on machine sounds, comprising 479,560 parameters (1.9 MB). A vibration-native twin of the same architecture, pre-trained on public bearing corpora, encodes the accelerometer streams. It embeds the channel mean of each accelerometer stream, sensor-less inputs included; the consequence for the 8 July session is described in Section 5.2.4.

Because these models expect single-channel inputs, each microphone stream is mono-mixed prior to encoding. While ring identity (turbine vs. generator) survives, the precise spatial arrangement within the ring is discarded (see Section 3.4.1). The standardise-and-concatenate rule of the handcrafted fusion also builds a second, embedding-level fusion space from the frozen general-audio embeddings and the handcrafted vibration features.

Finally, to guarantee reproducibility, all extracted features and embeddings are cached on disk under a robust hash covering the entire generation configuration, ensuring that a stale cache cannot be mistakenly used for an altered pipeline.

4.3 Operating-state layer

This section details the implementation of the operating-state layer, which is responsible for continuously decoding the machine's current state from the raw acoustic and vibration feature streams.

Unsupervised state tracking. The unsupervised arm operates in three sequential stages. First, a k-means clustering step groups the data using a fixed seed for reproducibility.

Next, a sticky hidden Markov model (HMM) refines these assignments. The HMM's transition matrix is manually initialized to strongly favour remaining in the current state, and the estimation parameters are constrained so that training cannot overwrite this prior. Finally, a duration filter cleans the timeline by removing any state segments shorter than a required five-second minimum dwell time.

Supervised model bank. The supervised arm is a per-state model bank. At calibration, one small model per operating state is fitted on process-labelled windows using one of three interchangeable families: a diagonal Gaussian, a cosine nearest-neighbour model, or a two-component Gaussian mixture. At monitoring time, the model bank operates entirely without labels. For each incoming feature window, the system evaluates the distance under every state model and assigns the window to the state yielding the minimum distance.

The bank is commissioned per representation: the evaluation covers all eight representations under a pre-declared family rule, Gaussian members for handcrafted feature spaces and nearest-neighbour members for embedding spaces (Section 5.2.1).

More importantly, the system implements a strict rejection loop. A window is flagged as `no_mode_fits` only if its distance exceeds the individual conformal threshold of every single member of the bank, not simply when its closest match happens to be poor. This conjunction provides a label-free drift signal indicating that the machine has entered an operating state completely unseen during the commissioning phase (see Section 4.5). A model whose threshold is mathematically infinite (because its calibration set was too small) can never contribute a rejection. This failure direction is intentional: the signal under-fires rather than over-fires when calibration data is thin. The bank logs any such incomplete models so that a missing state is visible rather than silent.

Both arms output named states. The snapshot persists a map from cluster or model identifier to the process vocabulary, built at calibration by majority vote over the labelled fit windows. At monitoring time, the output speaks the operator's vocabulary with no process data present. Clusters without a clear majority keep their numeric identifier rather than borrow a name they do not deserve. Consequently, the clustering arm names only those states it cleanly isolates, whereas the bank names all four natively.

Ground truth labelling. A single rule-based classifier produces every process label that this chapter and the next depend on. It supplies the labelled fit windows that name both arms' states and fit the bank, and it serves as the ground truth that Section 5.1.2 scores against. It reads the plant's own process channels, runs offline only, and is never

called by the deployed monitor. Table A.2 in Appendix A lists every threshold of its four ordered rules, each anchored to a measured plant quantity rather than to a convention. A window is marked as `unknown` when its own speed or power reading is missing from the SCADA export, and separately whenever its validity flag (Section 4.1) is false, because a state no detector could see is not a state that can be evaluated.

An independent re-implementation of the collaborating team's own SCADA rules, applied unchanged to the same recordings, agrees with these labels on 97.3 % of windows across the shared measurement days. Every disagreement falls within about a minute of a state change, which locates the difference in how the two rules handle a transition rather than in the underlying data.

4.4 Anomaly layer

The anomaly scoring mechanism operates strictly on a per-state basis. For every detected operating state, the implementation maintains a dedicated reference matrix containing the valid fit-side windows. When a new feature window arrives, the system computes its anomaly score by measuring its distance to this specific reference matrix, utilizing either a cosine nearest-neighbour model or a diagonal Mahalanobis distance with shrinkage. The remaining scorer families of the evaluation (one-class classifiers, density estimators, and the reconstruction models) run in the offline evaluation harness only; the runtime snapshot admits just these two, whose fitted state is purely numerical (Section 4.5).

The conformal thresholds require computing an order statistic on the state's held-out calibration scores [5, 30]. With n calibration scores and a target false-alarm rate α, the threshold corresponds to the value at index $\lceil (n+1)(1-\alpha) \rceil$ of the sorted scores. In software, this calculation requires a small floating-point tolerance ($\varepsilon = 10^{-9}$) subtracted before the ceiling function to prevent double-precision rounding errors from incorrectly pushing exact integer boundaries to the next rank.

Crucially, the theoretical conformal guarantee requires a strict data floor: the order statistic only exists when $n \geq 1/\alpha - 1$. If a state has too few calibration windows to support the requested α, the implementation enforces a safety branch that sets the threshold to infinity and flags the state as low-confidence. This explicitly prevents a data-scarce state from issuing false alarms under a statistical promise the calibration set cannot actually keep.

Data splitting strategy. To build the models and thresholds, the implementation partitions the data into three strictly disjoint sets: a fit set (to construct the reference matrices),

a conformal calibration set (to compute the thresholds), and a scoring set (for evaluation). Crucially, this partition is performed at the boundary of continuous burst segments rather than at the individual window level. Because consecutive one-second windows from the same burst are highly correlated near-duplicates, splitting within a segment would inadvertently leak identical acoustic signatures into both the training and evaluation sets.

Leakage prevention. Data leakage in anomaly detection often fails silently, producing falsely optimistic results rather than explicit errors. For example, if a nearest-neighbour model evaluates a window that was also present in its reference set, it will return a distance near zero. Similarly, if a calibration window accidentally leaks into the scoring set, the apparent detection rate artificially inflates without raising the false-alarm rate. To definitively prevent this, the implementation enforces a programmatic disjointness check (`_assert_three_way_disjoint`). This safeguard actively halts execution before any window is scored or a snapshot is built if the three sets share even a single overlapping window.

State-agnostic comparison. As an experimental baseline, the system also implements a state-agnostic scoring arm. Instead of building per-state models, a single pooled reference matrix and threshold are fitted indiscriminately across all operating states. This allows the evaluation to quantify exactly how much accuracy is gained by the state-aware architecture.

4.5 Runtime and evaluation tooling

Deploying the anomaly detector to a real-world power plant requires packaging the calibrated models into a standalone artifact and providing robust tooling for execution and evaluation.

Snapshot persistence. Once the monitor is calibrated, its entire state must be saved so it can be deployed to the power plant. This saved state is called a *monitor snapshot*. Technically, it consists of a compressed array file and a JSON sidecar document, which together bundle all necessary data: the standardisation statistics, the HMM transition matrices, the per-state reference matrices, the conformal thresholds, and the state-name map. Crucially, no Python object pickling is used, and the loading mechanism strictly refuses pickled payloads. Because the snapshot is designed to be copied and loaded months later at an active facility, arbitrary code execution via pickle poses an unacceptable security risk. This security constraint dictates which anomaly scorers are eligible for

runtime deployment: only those whose fitted state consists purely of numerical arrays are allowed. A numerical round-trip check verifies that a reloaded snapshot reproduces labels, scores, and thresholds exactly.

Monitor execution and threshold regimes. A dedicated monitor command applies a persisted snapshot to a new recording. The snapshot's column names serve as a strict feature contract: missing columns abort the run with an explicit error, while novel columns present in the new recording but absent from the snapshot are dropped with a warning. This ensures that a model is never influenced by a sensor it was not calibrated on. The command then decodes the operating states and scores each window under one of three threshold regimes: *frozen* (applies the stored thresholds unchanged), *recalibrate* (refits each state's threshold on the calibration split of the new recording), and *rolling* (derives thresholds dynamically from a trailing window of calibration scores). Every scored row logs its active regime to ensure full traceability.

Runtime options and event handling. Two runtime options support the operating-state layer and the evaluation campaigns. First, because state transitions often trigger spurious anomaly spikes, transition alarms can optionally be suppressed. Second, an *event-free calibration* rule (`_apply_calibration_exclusion`) ensures that labelled faults do not corrupt the threshold calculation. Rather than discarding labelled event windows that happen to fall into the calibration split, this rule explicitly moves them into the scoring split. Leaving an anomaly inside the calibration set would artificially raise the detection threshold, rendering the anomaly structurally undetectable and quietly degrading the true-positive rate.

Drift sentinels. Three drift sentinels automatically determine when a model requires recalibration. The first sentinel monitors the whole-bank rejection rate (as described in Section 4.3). Its trigger band is established during commissioning via a segment-level bootstrap (`_bootstrap_rate_pct`), resampling whole bursts rather than individual windows to account for temporal correlation. The second sentinel detects robust level steps on the raw microphone streams and cross-checks them against a vibration channel; if the acoustic level steps while vibration remains flat, the shift is attributed to instrumentation drift rather than physical machine behaviour. The third sentinel tracks the raw alarm rate of the frozen model against a fixed materiality factor, providing a simple volume-based fallback.

Evaluation tooling. Four supporting tools complete the implementation. An *event harness* scores alarm tables against labelled fault intervals to compute per-event detection rates, first-alarm latency, and out-of-event false-alarm rates. A *replay driver* orchestrates the monitor and harness sequentially over the recorded days, simulating a day-granular retrospective deployment. An *annotation kit* aids manual review by rendering high-resolution spectrograms directly from the raw samples, preserving the full twenty-kilohertz acoustic band. Finally, an interactive *dashboard* visualizes the state timeline, anomaly scores, discrete alarms, and raw audio snippets on a single page, making the entire pipeline reviewable without relying solely on aggregate tables.

With the data ingestion, anomaly scoring, and runtime pipelines fully established, Chapter 5 now measures the real-world performance of these architectural decisions against the experimental baselines.

Chapter 5

Evaluation

This chapter reports the experimental evaluation of the monitor. The design decisions behind it, the leakage-safe split scheme, the per-state threshold calibration, the evidence strategy, and the deployment and compactness requirements, are set out in the Design chapter (Sections 3.5 and 3.6) and are not restated here. Section 5.1 describes the evidence base before anything is measured on it, the campaign and its three instrumentation eras, the setup of the three experiments, and the metrics the plant data allow; Section 5.2 reports those three experiments in the same order and then the supporting evidence blocks, the adaptation ablation, the falsified alternatives, the scarcity measurement and the computational cost.

5.1 Experimental setup

The evaluation uses real-world recordings from Rodundwerk II [55], characterised by strong machine-room background noise and state-switching variability [28]. The six-day dataset covers turbine, pump, phase-shifter, and standstill operations. Crucially, the final day contains controlled acoustic anomalies.

The evaluation follows three sequential experiments. E1 evaluates whether the operating state can be identified from the sensor streams. E2 measures anomaly detection performance conditioned on that state. E3 assesses calibration robustness across changes to the physical sensor installation.

5.1.1 The campaign, its eras, and what can be verified

The recordings span three distinct *instrumentation eras* due to configuration changes in the acquisition hardware. Era A covers 25 and 27 June; Era B covers 29 June to 1 July; Era C covers the controlled-event day on 8 July.

These boundaries are confirmed by physical signatures: between Era A and B, broadband acoustic levels stepped up by ~ 2 dB while vibration levels remained flat, indicating

a microphone gain change. Between Era B and C, one sensor-less accelerometer input that had been constant through Eras A and B became intermittent: on 8 July it toggles between its rail value and zero and does not respond to the hammer strikes, so the change is a wiring change rather than a working sensor. It still marks the boundary, and it matters for the vibration encoder (Section 5.2.4). Because an autonomous monitor cannot anticipate when plant operators adjust sensors, crossing these eras provides a realistic test case for the drift sentinels designed in Section 3.4.6.

Consequently, accuracy claims for E1 and E2 are strictly bounded within a single era (a fitted detector belongs to the instrumentation it was trained on). Cross-era performance is measured separately in E3 to evaluate the recalibration strategy. The commissioning baseline is formed from the 1 July sessions (Era B), meaning these recordings are strictly separated into calibration and scoring splits. All other days act entirely as held-out test sets.

Figure 5.1 summarises the campaign: per day, it visualises the active operating states, the machine load, and the role of every recording session in the evaluation pipeline. Notable data realities include excluded sessions (e.g., missing SCADA exports on 27 June) and partial-load turbine runs (30 June).

The 30 June day comprises a turbine morning and a pump midday with a full-day process export. Its era membership is established by the same evidence the boundaries rest on: the day's microphone and vibration level chains sit inside the era-B envelope of its calendar neighbours, clearly separated from the era-C step. Its turbine session carries two single-chunk recording gaps on the two turbine-labelled streams, absent from the delivery itself and bridged by the other channels; the affected windows are masked.

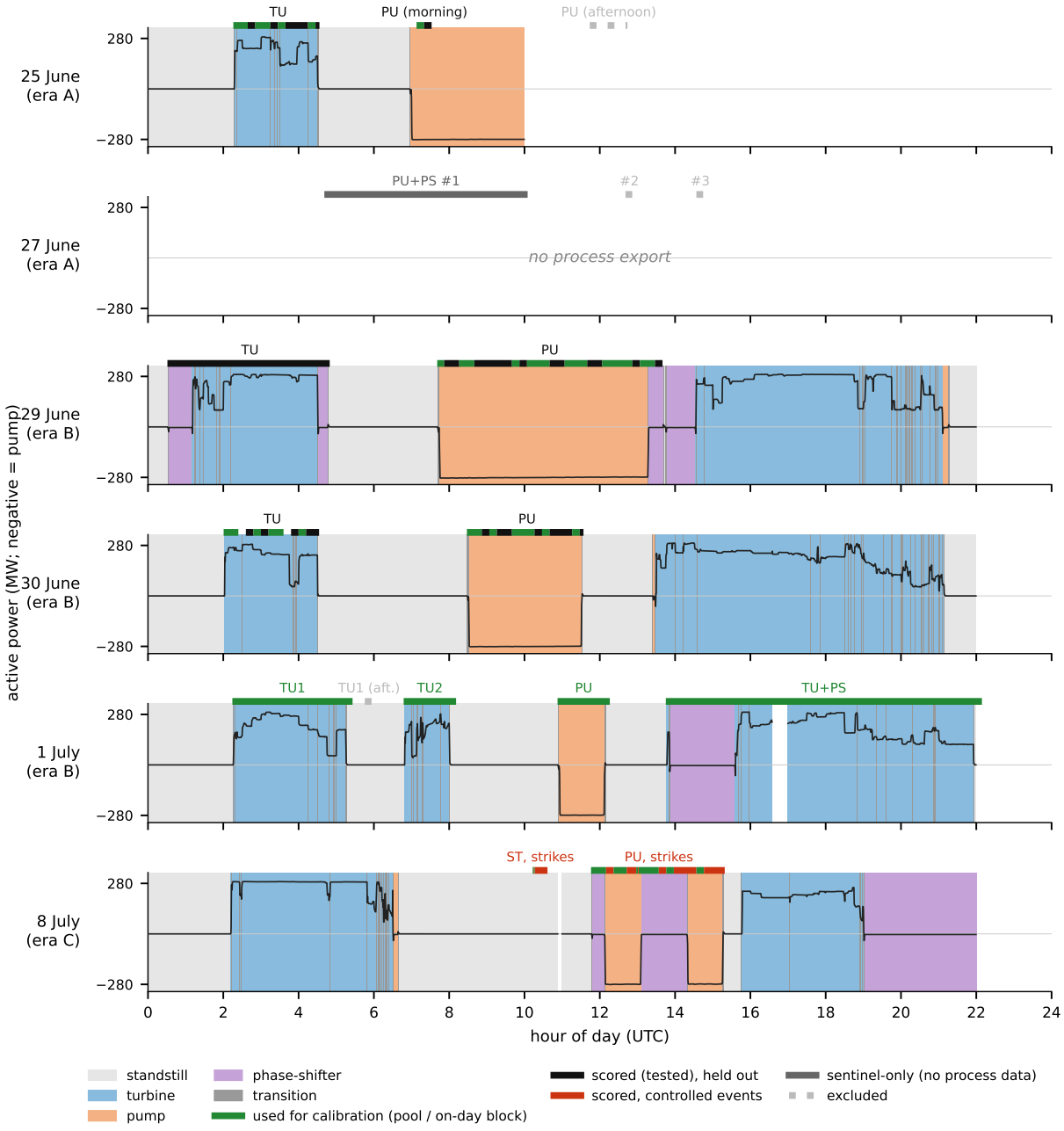

Figure 5.1: The campaign at a glance: states, loads, and what every window was used for. Six measurement days across three instrumentation eras: SCADA operating states as colour bands, active power as the line (negative is pump operation), and session usage as green (calibration), black (scored held out), red (controlled events) and grey (excluded or sentinel-only); narrow gaps are ramp windows masked by the ground-truth rule.

The delivered data also have gaps, and three of them are permanent. The measurement day of 27 June carries no operational data at all, so no state ground truth exists for it and

it can serve as false-alarm control but never as labelled evidence. The pump session on the afternoon of 25 June falls outside the operational export as well and leaves the replay set for the same reason. A further session survives only as a short acquisition-restart fragment whose share of invalid windows exceeds the ingestion guard, so it is dropped by rule rather than by judgement. Each exclusion follows a rule stated before any result was seen, and together they narrow what can be claimed.

Two defects in the time base had to be repaired first, a shifted acquisition epoch and a strike protocol logged in local time against sensor timestamps in UTC; Section 4.1 states both repairs and the checks behind them, and Section 5.1.3 states what the second costs if it is missed.

One narrowness no processing can widen. All recordings come from one machine at one site, so every number in this chapter is a number about this unit.

5.1.2 E1: operating-state identification

E1 compares the two arms of the operating-state layer of Sections 3.3 and 4.3: the unsupervised clustering arm and the commissioning-time model bank with its explicit rejection output. Both arms run under every representation of Chapter 3: E1 scores them on all eight, the bank's member family following a pre-declared rule (Gaussian for handcrafted feature spaces, nearest-neighbour for embedding spaces).

The comparison quantifies the value of the supervised commissioning step. While the unsupervised arm isolates states purely by acoustic structure, the supervised bank can explicitly name the states and detect out-of-distribution windows.

Both arms are scored against SCADA-derived state labels on held-out days within their respective eras. The primary metric is the adjusted Rand index (ARI) [20], computed after mapping each cluster to its majority true state. This prevents penalising an unsupervised model that correctly subdivides a single state (e.g., turbine operation at different loads) into distinct, valid clusters. Accuracy and macro-averaged F_1 [49] are reported alongside.

5.1.3 E2: anomaly detection

The evaluation for anomaly detection relies on a controlled acoustic campaign executed on 8 July 2026. Because faults cannot be provoked during active service, repeatable hammer strikes on the machine and nearby landmarks were conducted. The protocol

comprised two sessions (standstill and pumping at −279 MW) with 90 strikes each, yielding 180 protocol strikes in total.

These 180 strikes form the individual test cases. An impulse-detection pipeline processed the raw microphone signals to assign exact second-resolution timestamps to the 180 slots (`scripts/strike_register`). This register resolved 172 strikes by measurement: a located trace in the signal, from the author's blind listening marks, the impulse detector, and spectral templating verifying one another. For the remaining eight slots no locatable trace exists, and their timestamps are instead inferred in a listening pass guided by the protocol's strike rhythm: six distant building landmarks masked by pump noise, and two standstill guide-vane slots too faint to measure.

To prevent data leakage, a strict event-free calibration rule is enforced: labelled event windows (the registered strikes padded by 5 s) are categorically excluded from the calibration split (Section 3.5). A threshold fitted on event windows would inadvertently incorporate the anomaly it is meant to detect.

For all other campaign days without controlled events, the evaluation measures false-alarm control rather than detection sensitivity. The conformal calibration guarantees a bound on the nominal false-alarm budget, which is verified empirically by checking if the realised false-alarm rate across event-free days stays within the chosen threshold.

5.1.4 E3: robustness across instrumentation eras

E3 evaluates robustness by replaying the recorded days chronologically and comparing three calibration regimes:

1. **Always-frozen:** Fits thresholds once at commissioning and never updates them. This is the deployment ideal, but its vulnerability to hardware drift motivates E3.

2. **Always-recalibrate:** Refits thresholds on every monitored day. This establishes an upper bound on false-alarm control but is practically non-deployable, as it would silently absorb slowly developing faults into the new normal reference.

3. **Once-per-era (Sentinel-driven):** Fits thresholds at commissioning and recalibrates only when a label-free drift sentinel fires. This is the proposed architectural solution.

To isolate the effect of calibration cadence from representation choice, the replay commits to a single primary arm, handcrafted fusion with nearest-neighbour scoring; the commitment was fixed before the final measurement days arrived, so no representation

could be chosen in hindsight. Section 5.2.4 additionally repeats the full replay on every representation. The three label-free sentinels of Section 3.4.6 govern the *once-per-era* regime:

- **Mode-fit sentinel (s1):** Monitors the fraction of windows rejected by the model bank, compared against a segment-block bootstrapped threshold from the commissioning pool.
- **Level sentinel (s2):** Watches for robust median-absolute-deviation (MAD) step changes in audio levels, cross-checked against vibration channels to rule out physical machine changes.
- **Alarm-rate sentinel (s3):** A watchdog on the frozen arm's raw alarm rate, triggering if it exceeds a materiality factor (twice the nominal budget). This provides a volume-based fallback for failure modes that evade the first two sentinels.

Regime results are reported for every held-out day. Within-era days carry the primary numbers, because they measure the regime under the instrumentation it was calibrated for, while cross-era days are read as drift evidence: a frozen threshold that alarms on most windows of a foreign era is not a detector, and its nominal detection rate there is meaningless. The replay is a retrospective, day-granular simulation over recorded files rather than an online system (Section 4.5), and every threshold here, the sentinel threshold as much as the per-state conformal ones, is an empirical fit to this machine and this instrumentation that another plant would have to measure again on its own commissioning data.

5.1.5 Metrics

The metrics follow from what the plant data contain. No confirmed in-service fault appears in any recording, so there is no anomalous population to rank against a normal one, and three metrics are primary. The realised false-alarm rate against the nominal budget, per state and per day, measures validity, the property the per-state conformal calibration claims [5, 30] and the only detection-related one measurable on unlabelled normal days. Per-strike detection on the labelled benchmark, together with first-alarm latency, measures sensitivity under the tolerance and the event-free calibration rule of Section 5.1.3. The state metrics of Section 5.1.2 come third, since the anomaly layer is conditioned on the state layer and inherits its errors. Two further quantities characterise

the operation: alarm episodes per day measure the review workload (Section 5.2.3), and the calibration windows a state needs before its threshold stabilises measure how much healthy data a newly instrumented state requires.

Accuracy is among the state metrics and is not the honest one. On the 25 June turbine session 7 750 of 8 275 evaluated windows carry the turbine state, so a detector that labels every window turbine and detects nothing at all already scores 0.937. Macro-averaged F_1 weights each state equally and exposes what accuracy hides, whether the short standstill and transition stretches are recovered at all. Ranking metrics are kept only where a ranking population exists: clip-level area under the receiver operating characteristic curve (AUC) for the public-data scarcity study of Section 5.2.6, and nothing on the plant recordings, where the partial variant and the operating-point F_1 of Section 3.5 have no labelled population to be computed over.

5.2 Results

The research question asked: *Does self-supervised pre-training on public industrial sound datasets improve acoustic anomaly detection when fine-tuned on scarce PSHP data?* The results answer it in three experiments and the evidence blocks that support them.

They follow the order in which the experiments were introduced: E1 the state layer, within each recording session and then across held-out rotations, both arms under every representation; E2 the controlled-event benchmark of 8 July with the two controls that fix how its numbers may be read, the hand-set-threshold comparison, the scorer-family sweep, the reaction time, and a short part that prices the alarm policy; E3 the three calibration regimes, the sentinels that drive the third of them, and the replication of the cadence result across all representations. A fourth part collects the adaptation ablation and the alternatives the campaign falsified, and the scarcity measurement and the computational cost follow. Every number is a within-era number unless the text marks it otherwise; cross-era numbers are read as drift evidence in the sense of Section 5.1.1, and interpretation belongs to Chapter 6.

5.2.1 E1: operating-state identification

Table 5.1 scores the clustering arm strictly within each recording session (no transported model, no labels at any point) across all eight representations, while Table 5.2 scores the commissioning-time model bank against the clustering arm on held-out rotations, simulating deployment conditions.

Table 5.1: Operating-state identification within each session: state-level adjusted Rand index (ARI) of the clustering arm. Compared: all eight representations, each fitted and scored within the same session against process-derived state labels. ° single ground-truth class, every entry 1.000 by construction; windows are the fusion-arm count (counts per representation differ slightly by channel validity); best value per session in bold.

			Handcrafted (baseline)			Pre-trained (SSL) and student				
Session	**Era**	**Windows**	**Audio**	**Vib.**	**Fusion**	**BEATs**	**TF-C aud.**	**TF-C vib.**	**Fus. BEATs**	**Student**
25 Jun, turbine	A	8 275	0.684	0.153	**0.687**	0.000	0.000	0.000	0.000	0.676
25 Jun, pump, morning°	A	1 439	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
29 Jun, turbine	B	15 564	0.878	**0.916**	0.894	0.875	0.826	0.859	0.893	0.835
29 Jun, pump	B	21 510	0.954	**0.961**	**0.961**	0.959	0.954	0.955	0.958	0.956
30 Jun, turbine, morning	B	7 733	0.706	0.712	**0.717**	0.628	0.679	0.223	0.654	0.654
30 Jun, pump, midday	B	11 204	0.884	0.909	0.889	0.913	0.908	0.914	**0.917**	0.912
1 Jul, turbine, morning	B	11 039	0.714	0.734	**0.749**	0.000	0.000	0.000	0.000	0.000
1 Jul, turbine, second	B	4 561	**0.466**	0.406	0.446	0.000	0.000	0.000	0.436	0.000
1 Jul, turbine and phase shifter	B	27 821	0.929	**0.941**	0.930	0.922	0.907	0.920	0.928	0.919
1 Jul, pump	B	4 725	0.838	0.858	0.843	0.859	0.862	**0.868**	0.858	0.776
8 Jul, pump, strikes	C	12 808	0.956	**0.968**	0.957	0.962	0.966	0.940	0.964	0.965
8 Jul, standstill, strikes°	C	1 415	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5.1 reads along the sessions, not along the representations. On every session that contains more than one operating state, all eight representations identify the states almost equally well: the ARI stays between 0.83 and 0.97, and the choice of representation barely matters. The picture changes on the three sessions that hold only turbine operation. There the learned encoders fail completely, at an ARI of 0.000, while the handcrafted features still find the short standstill and transition stretches (0.45 to 0.75).

The reason is what each representation preserves. Running against standing is mostly a difference in loudness, and the pre-training of BEATs and TF-C is built to ignore loudness; two loud states at the same load differ in spectral content, which the encoders do preserve. The distilled student confirms the mechanism: re-trained on the plant's own audio, it reaches 0.676 exactly where its teacher reads 0.000, because learning from plant recordings brings the loudness structure back at the price of being plant- and era-specific rather than a general-audio model.

Macro-averaged F_1 follows the same pattern (0.32 to 0.82). Accuracy does not: it stays between 0.94 and 0.99 throughout and hides every failure, as Section 5.1.5 predicted. The two sessions marked ° contain only one state, so their 1.000 entries carry no information.

Table 5.2: Model bank against the transported clustering arm on held-out rotations, all eight representations. Compared: per cell, bank ARI / clustering-arm ARI on identical held-out windows, both arms fitted on the same rotation pool; bank families follow the pre-declared rule (Gaussian for handcrafted features, nearest-neighbour for embeddings). ° degenerate single-class session; n/a[c] = the era-C vibration channel set (108 against 96 columns, because the intermittent sensor-less input of 8 July passes the liveness rule) is refused by the bank's feature contract, whereas the monitor of Section 4.5 drops novel columns; best bank value per row in bold.

Held-out session	Era	HC audio	HC vib.	HC fusion	BEATs	TF-C aud.	TF-C vib.	Fus. BEATs	Student
29 Jun, turbine	B	0.91/0.65	0.95/0.69	**0.99**/0.70	**0.99**/0.41	0.97/0.43	0.95/0.57	**0.99**/0.33	0.98/0.33
29 Jun, pump	B	0.96/0.98	**0.97**/0.98	0.71/0.96	0.92/0.96	0.16/0.96	0.81/0.59	-0.01/0.90	0.59/0.91
30 Jun, turbine, morning	B	**0.99**/0.24	0.35/0.20	0.20/0.23	**0.99**/0.10	**0.99**/0.18	0.93/0.18	0.86/0.06	**0.99**/0.04
30 Jun, pump, midday	B	0.83/0.96	0.87/0.90	0.05/0.82	0.94/0.80	0.33/0.82	**0.97**/0.95	0.07/0.63	0.93/0.81
1 Jul, turbine and phase shifter	B	0.46/0.45	0.31/0.14	0.01/0.26	0.55/0.56	**0.86**/0.10	0.06/-0.05	0.57/0.33	0.52/0.50
1 Jul, pump	B	0.85/0.73	0.86/0.81	0.86/0.20	**0.87**/0.66	0.41/0.72	0.61/0.74	0.23/0.49	0.68/0.32
Mean, era-B rotations: bank		0.835	0.720	0.470	**0.877**	0.620	0.722	0.452	0.783
Mean, era-B rotations: cluster		0.668	0.620	0.528	0.582	0.534	0.494	0.456	0.486
25 Jun, turbine	A	**0.99**/0.09	0.85/0.06	0.16/0.05	0.96/0.04	0.88/0.05	0.91/0.04	**0.99**/0.03	**0.99**/0.04
25 Jun, pump, morning°	A	1.00/0.00	1.00/1.00	0.00/0.00	1.00/1.00	1.00/1.00	1.00/1.00	0.00/0.00	1.00/1.00
8 Jul, pump, strikes	C	0.96/0.98	n/a[c]	n/a[c]	0.71/0.95	0.42/0.95	**0.98**/0.98	n/a[c]	0.19/0.95

The bank has no within-session counterpart to Table 5.1: fitted and scored on one session's own labels it would answer no deployment question, so its setting is the held-out rotation, the bank-only numbers are the first entry of every cell below, and Appendix Table A.6 carries them for all three member families. Table 5.2 answers the deployment question: fitted on one day, how well does each arm work on another? The bank wins on six of the eight representations in the era-B mean, and by the largest margin exactly where the deployed system runs, on the frozen BEATs embeddings (0.877 against 0.582). One day of labels at commissioning therefore buys a measurably better state layer, and the deployed pairing (BEATs embeddings for audio, handcrafted features for vibration) is a measured optimum rather than a curated choice. The comparison also shows where the bank loses: on both fusion feature spaces the clustering arm is ahead, so a bank is only as good as the match between its representation and its member family, and that family rule belongs to commissioning. The two 30 June rotations, held out like every other day, sit at 0.99 and 0.94 for the BEATs bank.

Two rows need care. The 25 June pump rotation contains a single state, so its 1.000 entries prove nothing. The honest cross-era row is 25 June turbine: the bank holds 0.85 to 0.99 while the transported clustering arm falls to 0.03–0.09, because clustering must

split the one turbine state into four groups and the ARI charges every split, whereas the bank's members are fixed, named states. On the era-C strike session the ordering flips (bank 0.71, clustering 0.95). The bank loses those points because it misassigns 956 phase-shifter windows and rejects 19.5 % of the session as fitting no member; that rejection is deliberate: it is the very signal the first drift sentinel of E3 watches. The same row shows that the refusal is a property of the channel handling, not of the modality: the handcrafted vibration space is refused because the intermittent input adds twelve columns, while the TF-C vibration bank, which embeds the channel mean of each stream, holds 0.98 across the boundary.

The bank is therefore the primary operating-state layer. It names all four states in operator vocabulary, it reports every window it cannot place, and its better assignment carries through to fewer false alarms: on the 29 June turbine rotation, the same scorer under the same budget realises a pooled false-alarm rate of 0.011 under the bank's label-free assignment against 0.075 under the clustering arm's.

5.2.2 E2: anomaly detection

Table 5.3 reports the labelled benchmark on the register basis, every one of the 90 protocol strikes per session scored as its own test case. Every cell is zero-shot on the strike day: all references were fitted on era-B sessions and only the thresholds were recalibrated, on that day's own strike-free windows and under the event-free calibration rule of Section 5.1.3. The split mirrors both the test design and the deployment: keeping the references on era B is what makes the test zero-shot, since refitting them on the strike day would turn the transfer question into a within-day one; and recalibrating only the thresholds is exactly what the once-per-era cadence does at an era boundary, because thresholds can be refreshed label-free from a block of normal windows while the references would need a new commissioning block.

Standstill. Resolving the protocol strike by strike separates the representations far more sharply than a minute-wide reading does. The frozen BEATs encoder finds all 90 strikes from $\alpha = 0.05$ upward and 87 at the strictest budget. Its 0.78 MB distilled student stays within four strikes of that, at a fraction of the realised rate (0.012 against 0.086), and the audio TF-C encoder reaches 85 of 90. The handcrafted baselines fall far behind: fusion resolves 17 strikes, audio none at all. The audio arm still alarms on 6.2 % of the session's windows, enough to cover a minute-wide interval by coincidence, yet never within one second of a strike. The vibration side splits by feature space. The handcrafted

Table 5.3: Per-strike detection on the 180-strike register, 8 July 2026. Compared: all eight representations against the three handcrafted baselines, zero-shot on the strike day (era-B references, event-free recalibrated thresholds); a strike counts as detected if an alarm falls within 1 s of its registered instant, and the rate is the realised share of alarming windows outside every padded event interval. Best per session block and budget in bold; † at standstill the handcrafted-vibration state layer leaves no scored window, so no rate exists and nothing could alarm; ‡ on 8 July the channel mean of the turbine accelerometer stream that feeds this encoder is dominated by an intermittent sensor-less input (Section 5.1.1), so this row rests mainly on the generator-stream embedding.

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Representation	**Strikes**	**Rate**	**Strikes**	**Rate**	**Strikes**	**Rate**
Standstill session, 90 strikes						
Handcrafted audio (baseline)	0/90	0.003	0/90	0.062	0/90	0.114
Handcrafted vibration (baseline)†	0/90	—	0/90	—	0/90	—
Handcrafted fusion (baseline)	12/90	0.022	17/90	0.077	20/90	0.157
BEATs frozen (audio SSL)	**87/90**	0.003	**90/90**	0.086	**90/90**	0.132
TF-C (audio SSL)	79/90	0.000	85/90	0.043	85/90	0.114
TF-C (vibration SSL)‡	18/90	0.003	26/90	0.046	31/90	0.120
Fusion on BEATs (SSL)	0/90	0.000	18/90	0.058	20/90	0.092
Distilled student	85/90	0.003	86/90	0.012	88/90	0.031
Pump session at −279 MW, 90 strikes						
Handcrafted audio (baseline)	35/90	0.017	47/90	0.063	54/90	0.111
Handcrafted vibration (baseline)†	0/90	—	2/90	—	10/90	—
Handcrafted fusion (baseline)	2/90	0.011	3/90	0.046	4/90	0.093
BEATs frozen (audio SSL)	**83/90**	0.008	**84/90**	0.049	**84/90**	0.104
TF-C (audio SSL)	32/90	0.008	49/90	0.076	62/90	0.134
TF-C (vibration SSL)‡	0/90	0.016	5/90	0.060	14/90	0.120
Fusion on BEATs (SSL)	35/90	0.024	54/90	0.085	77/90	0.161
Distilled student	75/90	0.016	79/90	0.068	83/90	0.114

vibration branch leaves no scored window at standstill; the vibration SSL encoder scores the session and resolves 26 of 90 strikes. On this day the channel mean of its turbine stream is dominated by an intermittent sensor-less input, so the resolved strikes rest on the generator-stream embedding, that is, on the accelerometer at the turbine-ring 0° position: the pre-trained embedding picks up structure-borne traces of the impacts that the handcrafted harmonics-and-envelope features, built for rotating machinery, do not represent.

Pumping. Pump operation reorders the field. The frozen encoder keeps 84 of 90 strikes at the nominal budget; the audio TF-C encoder drops to 49, handcrafted fusion to 3, and machine noise also masks the vibration SSL encoder (5 of 90). Handcrafted audio rises to 47 strikes, but mostly at the positions closest to the microphones: it reacts to raw loudness, not to the structural signature. The BEATs-based fusion quantifies the dilution mechanism directly: the same embeddings that resolve 84 strikes alone resolve 54 once fused, because concatenating a modality that cannot sense an airborne impact dilutes the distance the impact induces. Fusing for robustness and detecting short transients are competing objectives, which sharpens the two-path alarm design of Section 3.5.

Unresolved strikes. The six strikes missed by the frozen encoder during pumping are exactly the six landmark slots whose timestamps are inferred rather than measured: the ones neither the annotator nor a validated targeted search could locate in the signal under pump noise (Section 5.1.3). On the 84 slots a human could resolve at all, the encoder detects 84 of 84. What remains is therefore a limit of physical sensor coverage, not a detector failure.

The from-scratch pole. The from-scratch autoencoder has no entry in Table 5.3 for a structural reason: the runtime snapshot refuses pickled model payloads (Section 4.5), so no monitor path exists for a scorer whose fitted state is a network. Evaluated offline under the same pooled era-B fit, it holds the rate once recalibrated on the strike day (0.063 at standstill, 0.066 during pumping) and breaks completely under frozen thresholds at the era-C standstill session, alarming on every scored window (1.000).

Two controls fix how Table 5.3 may be read; both are results, not caveats.

The calibration rule. Event-free calibration is a precondition of a controlled-event benchmark, not a refinement of one. With the labelled event windows left on the calibration side, the detected share drops from 1.00 to 0.38 at standstill and from 0.92 to

0.77 during pumping. The reason is mechanical: a threshold fitted on event windows incorporates the very anomaly it is meant to detect.

The frozen control. The second control asks what the benchmark would look like with no recalibration at all, the era-B thresholds carried unchanged onto the era-C day. Handcrafted fusion then alarms on 62.3 % of the pump session's windows outside the padded intervals and on 95.9 % of the standstill session's, and handcrafted audio on 68.5 and 100 %. A detector that flags six to ten of every ten windows finds every strike trivially and detects nothing, which is why the benchmark recalibrates thresholds in the first place, and why every detection number in this chapter carries its realised rate.

The closing comparison is the hand-set threshold, the first reference practice of Table 3.3: a median-plus-k scaled-median-absolute-deviation (MAD) rule on the high microphone band, fitted once on the commissioning pool, never recalibrated, and run over the same two strike sessions on the same register basis (Table 5.4).

Table 5.4: Hand-set MAD threshold (practice baseline) on the two strike sessions. Compared: the fixed-threshold reference practice of Table 3.3 against the calibrated system of Table 5.3, on the same register basis and 1 s tolerance. The threshold is median $+ k$ scaled median absolute deviations of the high-band microphone energy, computed once on the commissioning pool; best detection per column in bold.

	Standstill		Pump	
Multiplier	**Strikes**	**Rate**	**Strikes**	**Rate**
$k = 5$	0/90	0.000	0/90	0.004
$k = 10$	0/90	0.000	0/90	0.000
$k = 20$	0/90	0.000	0/90	0.000
$k = 30$	0/90	0.000	0/90	0.000
$k = 40$	0/90	0.000	0/90	0.000
$k = 2.409$, tuned to 1 %	**36/90**	0.000	**36/90**	0.009

No conventional multiplier between 5 and 40 resolves a single strike in either session. Those constants are chosen for raw impulse resolution; this monitor scores the mean band energy of a one-second window, so the excursion the constant was designed for is averaged away before the comparison starts. Retuned to flag 1 % of the commissioning pool ($k = 2.409$), the same rule reaches 36 of 90 strikes per session, well ahead of handcrafted fusion and less than half of the frozen encoder. What the retuned rule cannot do is keep its own error rate, which is the property the conformal design exists for. Its flag rate holds where it was tuned (1.04 and 1.08 % on the two in-sample 1 July sessions)

and drifts everywhere else, down to a factor of nine below its target; per operating state it spans 0.0 % on several steady states to 26.4 % on transition windows. One threshold serves every state, and it holds its intended rate on almost none of them.

Validity on the unlabelled days. Per-state conditioning delivers what it promises wherever the day's calibration population is stable: every calibratable state realises a false-alarm rate between 0.000 and 0.105 at $\alpha = 0.05$. The measured exception is 30 June, whose main state drifts within the day itself and reaches 0.340 (Table 5.5; Section 5.2.4 carries the detailed analysis). The pooled comparison arm only looks similar: its aggregate stays near the budget (0.035 to 0.228), but its individual states range from 0.000 to 0.311: over- and under-alarming states cancel, and the average hides a violation in every component. A second comparison repeats the sub-cluster finding of the Design chapter: calibrating on the detected sub-clusters realises 0.069 where the coarse process labels realise 0.188 on the same windows.

The scorer family. Table 5.5 holds the split, the conditioning and the conformal harness fixed and varies the scorer alone. No family dominates: the ordering changes from cell to cell, each of the five lands below the budget in some cells and above it in others, and on the 30 June session every family misses the budget in both conditioning arms and under both representations, the adapted encoder included. The nearest-neighbour scorer this chapter uses by default is not the empirical best performer; it is the worst of the five in seven of the sixteen cells. A cross-day check sharpens the point: carried from one day's calibration onto another, the Mahalanobis arm holds the nominal budget on ten of twelve ordered day pairs (0.003 to 0.134), the nearest-neighbour arm on one (0.017 to 0.857). The contribution claimed here is therefore the conditioning and calibration layer around a swappable scorer, not one scoring rule.

Two arms sit at the edges of the sweep. The majority-voting ensemble of the classical baseline lands inside the picture rather than above it (0.056, 0.092, 0.027 and 0.250 on the four handcrafted-fusion cells). The reconstruction pole scores log-Mel input and is the strongest arm of the whole sweep on the richest session (0.019 on 1 July) while aggregating to 0.253 on 25 June: a model fitted from scratch on a single session needs a data volume that only the richest session supplies, which is exactly the constraint the transferred representations exist to relax.

Reaction time. Figure 5.2 reports the per-slot first-alarm latency on the register. The two pre-trained audio encoders answer well under a second on the one-second grid:

Table 5.5: Scorer families on sixteen within-day cells. Compared: five scorer families on the same split, conditioning and conformal harness per cell (one session, one representation, one conditioning arm); realised false-alarm rate at $\alpha = 0.05$, aggregated over a cell's scored states. Lowest rate per cell in bold; † one state has no usable nearest-neighbour reference, so that column scores fewer windows.

Session	Conditioning	k-NN	Mahalanobis	OC-SVM	Isolation forest	LOF
Handcrafted fusion						
25 June	per state	0.065	0.065	0.092	**0.063**	0.093
25 June	pooled	0.073	0.058	0.047	0.058	**0.017**
29 June	per state	0.066	0.112	0.069	0.115	**0.064**
29 June	pooled	**0.035**	0.087	0.091	0.106	0.089
30 June	per state	0.340	0.260	**0.252**	0.257	0.289
30 June	pooled	**0.228**	0.459	0.322	0.456	0.260
1 July	per state	0.069	0.043	**0.028**	0.033	0.033
1 July	pooled	0.143	0.040	**0.027**	0.043	0.027
BEATs with low-rank adapters						
25 June	per state	0.069†	0.005	0.006	**0.003**	0.018
25 June	pooled	0.066	0.071	0.079	0.068	**0.026**
29 June	per state	0.136†	0.088	0.081	**0.072**	0.119
29 June	pooled	0.108	0.101	0.099	**0.089**	0.109
30 June	per state	0.359	0.324	**0.132**	0.244	0.353
30 June	pooled	0.344	0.436	0.341	0.480	**0.305**
1 July	per state	**0.028**	0.057	0.082	0.055	0.059
1 July	pooled	0.030	**0.020**	0.031	0.024	0.054

medians of 0.522 and 0.566 s for the frozen encoder, 0.503 and 0.585 s for its distilled student. The handcrafted arms sit between one and three seconds and reach far fewer strikes. Re-scored on a 0.25-second hop, with windows and calibration untouched, the same frozen encoder answers at 0.114 and 0.109 s and detects the same strikes. The reaction time achievable on this data is therefore set by the scoring grid, not by the detector, and the finer hop also buys the ability to say which strike an alarm belongs to.

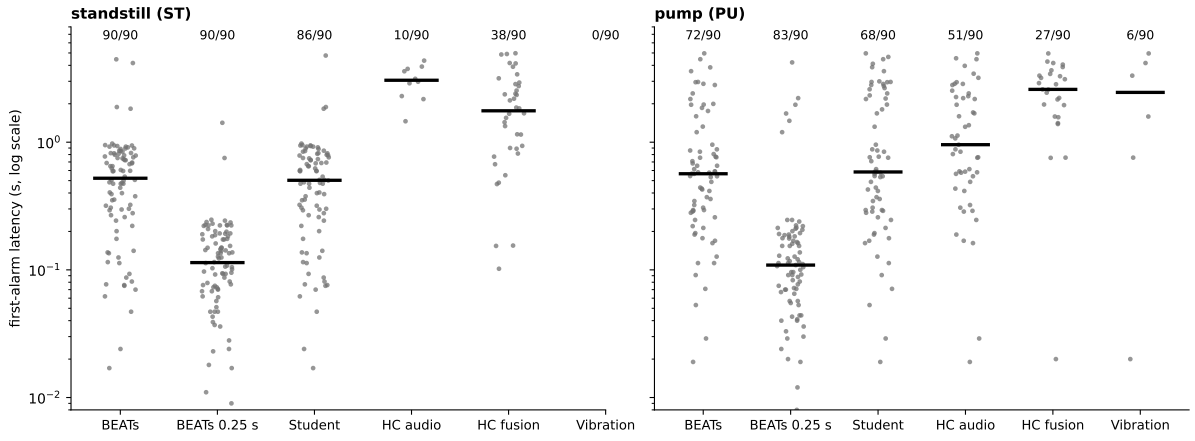

Figure 5.2: First-alarm latency on the strike register. Compared: per-slot first-alarm latency of six representations at $\alpha = 0.05$ for standstill (ST) and pump (PU), log scale; each point is one strike's gap to its first alarm, the bar the median, the count above each column the strikes with a first alarm inside a five-second horizon, and BEATs 0.25 s the fine-hop re-scoring of the frozen encoder.

The candidate register. Across the thirteen monitored sessions the three register criteria of Appendix A.3 yield 159 review candidates (69 impulse, 63 transient, 27 sustained). One entry stands out on measured grounds alone: a broadband transient during turbine operation on 29 June at 02:00:23 UTC falls inside the load-swing window in which the collaborating team independently reported click impulses, and the band-energy search locates the same event as the dominant impulse of its minute on both microphone rings. None of this is ground truth; the register exists for expert review, and its assessment vocabulary records an honest no-finding as a complete answer.

5.2.3 From window p-values to operator alarms

A per-window decision at a nominal budget of $\alpha = 0.05$ is not an operator alarm. By construction it flags about 5 % of all windows, which on a full recording day is hundreds to thousands of one-second flags: the designed budget behaving exactly as specified, and

Table 5.6: Alarm episodes under five alarm policies. Compared: episode counts per session under rising policy strictness, on the once-calibrated fusion monitor's scored windows; an episode is a run of consecutive flagged windows. The two 1 July sessions are in-sample, and the two 8 July sessions contain the controlled strikes.

Session	Scored windows	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.01$, ≥ 3 s	$\alpha = 0.01$, ≥ 5 s	$\alpha = 0.001$
25 June, turbine	3 967	82	32	1	0	0
25 June, pump (morning)	719	15	2	0	0	0
29 June, turbine	15 498	274	96	15	5	7
29 June, pump	9 916	273	48	0	0	1
30 June, turbine	3 394	134	26	0	0	0
30 June, pump	5 346	197	40	2	0	4
1 July, turbine and phase shifter	29 299	323	45	6	1	3
1 July, pump	2 157	179	32	0	0	7
8 July, pump, strikes	6 554	241	40	5	3	3
8 July, standstill, strikes	1 193	106	54	0	0	0

unusable as a work list. Table 5.6 prices the two policy knobs of Section 3.5, the budget and the persistence requirement, by counting the episodes each session emits under five policies of rising strictness.

Table 5.6 carries two lessons pointing in opposite directions. First, review workload is plannable through persistence filtering. At a strict budget ($\alpha = 0.01$) with a three-second persistence requirement, seven of the eight normal sessions emit six or fewer alarm episodes, down from hundreds at the unconstrained nominal budget. The monitor thus behaves as an operational tool only when the policy layer explicitly demands temporal continuity.

Second, the exact same persistence rule destroys transient anomaly detection. The three-second rule leaves zero episodes on the standstill strike session, despite the distilled student successfully resolving 85 of the 90 strikes at the window level. Because a hammer strike is a sub-second transient, persistence filtering inherently discriminates against short mechanical impacts while effectively suppressing isolated statistical flukes.

The consequence is a mandatory two-path alarm architecture. A sustained-anomaly path pairs a strict budget with persistence filtering to produce a manageable daily worklist. A parallel transient path uses the same scores and strict budget but omits the persistence requirement to catch impacts. Neither path can substitute for the other.

5.2.4 E3: robustness across instrumentation eras

The replay steps through the recorded sessions in chronological order (every session except the standstill strike session, which the pinned replay set leaves to the per-strike evaluation of E2), and Table 5.7 reports the three regimes on identical windows together with the sentinel rate that decides the third.

Table 5.7: Calibration regimes on the replayed sessions, primary arm. Compared: the realised false-alarm rate under the three regimes of Section 5.1.4 on identical windows per session, handcrafted fusion at $\alpha = 0.05$; the s1 column is the share of windows fitting no model-bank member against the fixed threshold 0.0805, and a decision marked (s3) was triggered by the alarm-rate sentinel alone. Lowest realised rate per session in bold; * in-sample (commissioning pool); 27 June carries no process data, so only s1 and s2 are evaluated; 8 July rates are computed outside every padded event interval.

Session	Era	Frozen	Recal.	Once	s1	Decision
25 June, turbine	A	0.056	**0.027**	**0.027**	0.250	recalibrate
25 June, pump, morning	A	1.000	**0.031**	**0.031**	0.184	recalibrate
27 June, pump and phase shifter	A	n/a	n/a	n/a	0.327	recalibrate
29 June, turbine	B	0.075	**0.044**	0.075	0.077	frozen
29 June, pump	B	0.301	**0.032**	**0.032**	0.669	recalibrate
30 June, turbine	B	0.376	**0.056**	**0.056**	0.013	recalibrate (s3)
30 June, pump	B	0.671	**0.044**	**0.044**	0.107	recalibrate
1 July, turbine and phase shifter*	B	**0.019**	0.200	**0.019**	0.021	frozen
1 July, pump*	B	0.192	**0.096**	**0.096**	0.025	recalibrate (s3)
8 July, pump, strikes	C	0.623	**0.046**	**0.046**	0.195	recalibrate

*In-sample: the session is part of the commissioning pool the thresholds were fitted on.

Always-frozen vs recalibration. The always-frozen regime severely degrades over time and across instrumentation eras, reaching false-alarm rates up to 1.000 (every window alarming) when applied to era A pump data. In contrast, the continuous recalibration regime naturally holds the false-alarm rate near the nominal budget of $\alpha = 0.05$. However, as established in the design, continuous recalibration is not deployable because it would inadvertently absorb and conceal slowly developing faults into the new normal reference.

Once-per-era with sentinels. The once-per-era regime bridges this gap. Driven by the three label-free sentinels, it triggers recalibrations when the sensing configuration changes: on the 8 July strike session (era C) the recalibration drops the false-alarm rate from 0.623 frozen to 0.046, a factor of 13.5, without manual intervention or scheduled refits, and every held-out session of the column sits at or below 0.075.

Reading the decision column. Eight of the ten decisions read recalibrate, three of them inside era A, which is more than one calibration per era; this is a property of the replay rather than of the rule. Each session is scored independently against the original commissioning calibration, and a triggered recalibration is never carried into the next session, so the sentinels meet every era-A session with the same era-B thresholds and fire on each in turn. The per-era trigger counts are therefore upper bounds; operated as designed, the first session of a foreign era is recalibrated once and later sessions score against that fresh calibration.

The alarm-rate sentinel earns its place on 30 June. On that turbine session the first two sentinels report a day the system should trust: the bank recognises 98.7 % of the windows, the lowest rejection rate of the whole replay, and the level sits far inside its margin. Yet the frozen thresholds alarm on 37.6 % of the scored windows. One recognised operating state carries the entire failure, flagging 44 % of its 6 646 windows at process conditions indistinguishable from the pool days (power, head, guide-vane opening, flow and both pressure taps agree), while the spectral profile tilts consistently on both microphone arrays. State membership and threshold validity are therefore different quantities, and only the monitor's own realised alarm rate sees this failure: the sentinel fires at twice the nominal budget and alone turns 0.376 into 0.056. Two notes belong next to it. The alarm-rate rule was added after an early replay pass exposed exactly this failure mode, with its factor fixed before the replay was rerun under it; that is the same epistemic status as the first two sentinels, which were designed knowing that instrumentation eras exist. And a day of sustained genuine anomalies would raise the same rate; the trigger log records which sentinel fired, so a deployment can distinguish threshold breakage from the machine itself before acting.

The breakage is representation-dependent. Under the same pooled protocol on the same 30 June session, the committed fusion arm's 0.376 stands against 0.101 for handcrafted audio, 0.091 for handcrafted vibration, 0.100 for the frozen BEATs encoder, 0.065 for the BEATs-based fusion, and 0.048 for the audio TF-C embedding. The breakage concentrates in the committed handcrafted-fusion space. Two caveats keep that axis honest: the fusion commitment predates the day's recording, and no representation transports universally: each alternative fails on other sessions of the same rotation data in turn. Under the symmetric pool rotation the other era-B days improve substantially (29 June pump from 0.301 to 0.033, 1 July pump to 0.002), so pool diversity pays where pool volume does not.

Table 5.8 settles the cadence question across the whole representation axis by repeating the entire replay, sentinels included, on every representation.

Table 5.8: The calibration-cadence result replicated on every representation. Compared: the three regimes of Section 5.1.4 rerun per representation over the same nine replayed sessions with the same label-free sentinels; columns count sessions whose always-frozen rate misses the $\alpha = 0.05$ budget, the worst frozen rate, the era-C strike-session pair, and the worst once-per-era rate. Handcrafted fusion is the pre-committed primary arm (Table 5.7); the full per-session matrix is Table A.7 in Appendix A.

	Frozen misses	Worst	Era-C strike session		Worst
Representation	**budget on**	**frozen**	**frozen**	**once**	**once**
Handcrafted audio (baseline)	8/9	0.927	0.685	0.017	0.104
Handcrafted vibration (baseline)	4/9	0.112	**0.041**	0.049	0.112
Handcrafted fusion (baseline)	8/9	1.000	0.623	0.046	0.096
BEATs frozen (audio SSL)	7/9	0.686	0.216	0.049	0.100
TF-C (audio SSL)	5/9	0.384	0.121	0.008	**0.031**
TF-C (vibration SSL)	8/9	0.907	0.907	0.016	0.049
Fusion on BEATs (SSL)	8/9	0.896	0.506	0.024	**0.024**
Distilled student	6/9	0.510	0.061	0.016	0.090

Always-frozen misses the budget on four to eight of the nine replayed sessions under every representation, so the cadence is not an artefact of the committed arm; and the once-per-era rule holds every representation at or below 0.112 across the whole replay. The two excursions above 0.10 are stated rather than hidden: handcrafted vibration on 29 June turbine (0.112) and handcrafted audio on the same session (0.104), both without a sentinel firing. The table also separates era robustness from modality: handcrafted vibration is the calmest branch of the replay (four misses, and the only representation that holds the era-C session frozen, at 0.041), while the vibration SSL encoder breaks completely at the same boundary (0.907 frozen). The break has a known cause: the encoder embeds the channel mean of each stream, and on 8 July the mean of the turbine stream is dominated by the intermittent sensor-less input, whereas the monitor's feature contract (Section 4.5) drops that input's novel columns from the handcrafted branch. Surviving the boundary is here a property of the channel handling, not of the accelerometers.

Across days. Frozen calibration fails between days as well as between eras. Carried from each day's own calibration onto each other day, eleven of the twelve ordered pairs miss the nominal budget. The transfer is directional: 30 June onto 1 July reads 1.7 %, the single passing cell, while the reverse direction reads 85.7 %, the worst of the grid. And

the era tag alone does not predict the failure, since the worst same-era cell is worse than the mildest cross-era one (Figure A.1, Appendix A).

Rolling recalibration. Recalibrating thresholds from a trailing window of the running session, with references left frozen, realises 0.051 to 0.055 on 29 June turbine and 0.047 and 0.083 on the two 30 June sessions: most of the frozen breakage vanishes with no sentinel involved. The structural price is unchanged, and it is why the cadence question is settled in Section 3.4.6 rather than here. A trailing window must be assumed normal permanently, so a slowly developing fault raises its own threshold and disappears; and rolling refreshes neither the references nor the bank, so an instrumentation change still needs the era-level recalibration.

Vibration stability. Unlike the microphones, the accelerometers see no gain change at the era boundaries. On the 8 July strike session the never-refitted handcrafted vibration branch holds 0.041 frozen, on the very session where the acoustic branch reads 0.623: the quantitative form of the deliberate redundancy of Section 3.4.6: one branch survives the boundary that breaks the other.

5.2.5 Adaptation and falsified alternatives

Table 5.9 evaluates the adaptation of the BEATs encoder, fitted solely on the 1 July session and scored across the held-out days.

Table 5.9: Adaptation of the BEATs encoder. Compared: frozen, low-rank-adapted and fully fine-tuned variants of the same encoder, all plant-side fitting on 1 July only; realised false-alarm rate at $\alpha = 0.05$, per-state conformal calibration with nearest-neighbour scoring. Best rate per session in bold.

Encoder	Fitted on plant audio	1 July (adapt. day)	29 June	30 June	25 June
Frozen	nothing	0.058	**0.130**	0.362	**0.007**
Low-rank adaptation	rank-8 adapters (query, value)	0.028	0.136	**0.359**	0.069
Full fine-tuning	all 90.4 million parameters	**0.025**	0.364	0.362	0.051

Both adaptation methods halve the false-alarm rate on their own fit day (from 0.058 to 0.025–0.028). Off that day the picture reverses. Full fine-tuning collapses to 0.364 on 29 June: moving all 90.4 million parameters to fit a few hours of data distorts the embedding geometry and breaks threshold transportability. Low-rank adaptation avoids that collapse, holding 0.136 against the frozen encoder's 0.130 on 29 June. Yet on the

cross-era 25 June session the completely frozen encoder leads both adapted variants by an order of magnitude (0.007 against 0.069 and 0.051). The 30 June column adds a third point: all three variants read 0.359 to 0.362 on the session whose calibration population drifts within the day (Section 5.2.4), so adaptation does not repair a drifting calibration either; that failure lives in the threshold population, not in the encoder. Adaptation is therefore strictly a same-condition tool, and under data scarcity the deployment choice is the completely frozen, zero-shot encoder.

Table 5.10 documents the alternatives that were tested and did not survive, each with the number that killed it and the structural reason; the cross-attention row is the one exception, a coverage limit rather than a falsification.

These falsifications rule out simple normalisation and adaptation shortcuts for drift. The once-per-era calibration cadence (Section 5.2.4) emerged as the architectural survivor specifically because it sidesteps these structural dead ends.

Table 5.10: Alternatives tested and falsified. Compared: each candidate against the configuration it would replace, on the sessions named in its row; realised false-alarm rates at $\alpha = 0.05$ unless another quantity is named, and an arrow reads as the value before and after the change under test.

Alternative	Measured	Why it does not carry
Session normalisation	29 June turbine 0.075 → 0.300; 1 July pump 0.097 → 1.000	The normalisation window spans several operating states, so the state mix dominates the statistic meant to remove level offsets.
Level-only recalibration	29 June turbine, audio 0.481 → 0.483; vibration 0.112 → 0.800	Cosine scoring is already level-invariant, so on audio there is no lever to pull; on vibration, the label-free offset absorbs the state mix instead of the session gain.
Continued pre-training on plant audio	1 July state ARI 0.907 unchanged; 29 June frozen 0.305 → 0.478; from-scratch control 0.425	Added training reshapes the score distribution rather than transferring knowledge. A control trained from scratch on the same audio outperforms the continued checkpoint.
Multi-day adaptation pooling	25 June: frozen 0.007, pooled adapters 0.072, pooled student 0.069	More adaptation data across days cannot repair a calibration, because the days differ in what the threshold has to fit, not in how much of it there is.
Full fine-tuning	0.364 off-day (Table 5.9)	All 90.4 million parameters overfit to one session.
Level sentinel	Never fires; largest excursion 0.223 against a margin of 0.921	The level step between eras is real but smaller than the margin derived from the commissioning blocks' scatter.
Larger vibration pre-training corpus	State ARI 0.920 → 0.926 on 1 July and 0.859 → 0.860 on 29 June	Quadrupling healthy bearing data yields no gain; the limitation is the data *type*, not size.
Cross-attention fusion head	1 July per state 0.025, 0.054, 0.033 and 0.052 against 0.079, 0.047, 0.043 and 0.105 for feature-level fusion	A coverage limit: only the richest session (1 July) has enough data to calibrate the head across all states. Simpler feature-level fusion is required for sparser days.
Mixture clustering in place of k-means	25 June fusion state ARI 0.704 against 0.687, but 0 of 122 standstill windows recovered	The higher index results from a cleaner two-way split of dominant states, while entirely missing minority states like standstill.

5.2.6 Data-scarcity analysis

The data-scarcity question, how little healthy data a detector needs, is evaluated on a public industrial-sound proxy corpus [41], as the plant recordings lack the labelled abnormal clips necessary to compute detection performance curves. The protocol varies the normal fitting budget over five points (5 % to 100 % of the available pool) and measures clip-level AUC.

The resulting curve (Figure 5.3) separates the representations structurally at the extreme low-data end. With only eight normal fitting clips, the frozen BEATs encoder reaches 0.911 AUC, outperforming all other representations. Multiplying the data budget by twenty only yields a marginal gain to 0.953. In contrast, the audio TF-C encoder starts significantly lower (0.773) but exhibits the steepest learning curve, passing 0.908 at 42 clips and only overtaking the frozen encoder at the maximum budget with 0.958. The handcrafted and log-Mel baselines flatten early and never cross 0.771. For an industrial deployment with no recorded history, the absolute minimum-data performance dictates architecture selection, making the zero-shot frozen encoder the optimal choice.

A second constraint comes from the conformal calibration mechanics: separating well and being calibrated are two different budgets. At $\alpha = 0.05$, detection on the proxy corpus is zero for every representation up to 42 fitting clips, not because the representations fail but because those cells hold too few calibration clips for the split-conformal order statistic to exist at all (at least 19 calibration points are needed). On this corpus the calibration budget therefore binds before the separation budget. The plant-side counterpart asks how many normal windows a state needs before its threshold stabilises: 20 seconds to 5 minutes of effective recording suffice for the dominant operating states, while short minority states often never accumulate the 20-window minimum. Commissioning must therefore be budgeted by accumulated recording time per operating state, not by raw data volume.

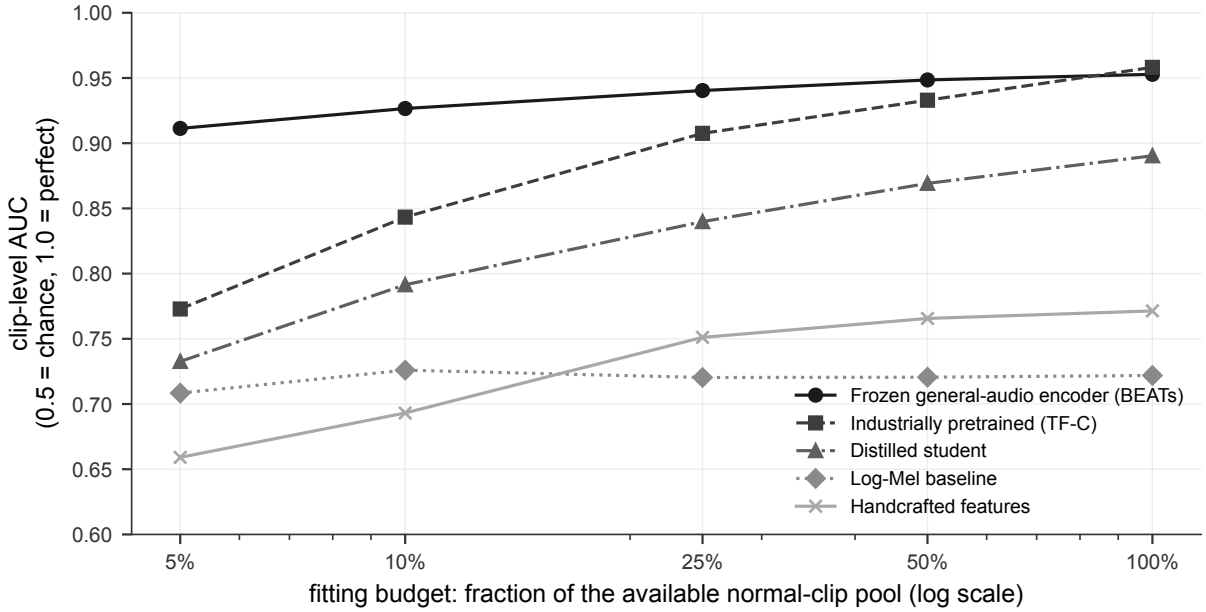

Figure 5.3: The data-scarcity curve. Compared: clip-level AUC of five representations against the normal fitting budget (log scale, markers at 8, 16, 42, 84 and 168 clips) on the public industrial-sound proxy corpus. This is public-proxy evidence measured on a different machine type, not evidence about Rodundwerk II.

5.2.7 Computational cost

Table 5.11 details the end-to-end computational cost per one-second window, including feature extraction and inference on a standard CPU without graphics acceleration.

Table 5.11: Inference cost per one-second window. Compared: parameters, on-disk size and end-to-end latency (from the raw 50 kHz signal, including feature extraction, on a CPU without graphics acceleration) across the deployable representations; lowest cost per column in bold.

			Latency per window (ms)	
Representation	**Parameters**	**Size (MB)**	**Batch 1**	**Batch 256**
Frozen BEATs encoder	90 354 032	361.5	30.15	33.31
The same encoder, 8-bit quantized	4 893 968	124.26	29.81	29.96
Industrially pre-trained, TF-C	479 560	1.94	1.45	1.56
Distilled student	192 643	0.779	0.78	0.49
Log-Mel baseline	0	**0.0**	**0.48**	**0.23**
Handcrafted features	0	**0.0**	2.99	2.87

The measurements support two primary deployment conclusions. First, knowledge distillation and post-training quantization are not interchangeable. The distilled student

compresses the frozen encoder's size by a factor of 464 (down to 0.779 MB) and its latency by a factor of 39 (down to 0.78 ms). Conversely, 8-bit quantization reduces storage footprint but leaves CPU latency virtually unchanged at 29.81 ms.

Second, the entire architecture is heavily over-provisioned for real-time edge execution. Even the most expensive representation (the frozen 90.4M parameter encoder) answers in 30.15 ms per one-second window. Therefore, the system easily supports real-time monitoring on a central processor without a dedicated graphics accelerator. Notably, a zero-parameter representation is not strictly the cheapest to compute: the heavy handcrafted feature extraction costs 2.99 ms, nearly four times the latency of running the distilled neural student.

Chapter 6

Discussion

The preceding chapters report five measurements: a state layer, a detector conditioned on that layer, a controlled-event benchmark, a calibration-cadence experiment, and an alarm policy. Each is modest alone; together they describe one system, and this thesis's argument lives in how they constrain each other rather than in any single table.

Section 6.1 reads the five strands as one argument. Section 6.2 weighs what the campaign delivered against what the design assumed, and answers from the same inventory why the reported numbers are not an artefact of tuning. Section 6.3 states what a second plant must supply; Section 6.4 collects the limitations.

6.1 Reading the evidence together

An anomaly score means nothing except relative to a stated normal; that normal is stable only if the operating state is known; and the whole is deployable only if the system notices when its normal has expired. Each strand supplies one link and fails without the others, so the evidence reads as one chain, not five findings.

6.1.1 State identification works two ways, and the trade is stated

Two arms recover the operating state, and the useful fact is not which wins but that they fail differently. On held-out days the commissioning-time model bank leads; the unsupervised clustering arm ties or wins two of the six era-B rotations and takes the strike-day zero-shot comparison outright (Table 5.2). The bank carries a fixed set of fitted states, so under drift it degrades honestly: a poor fit surfaces as a rejection, not as a confident wrong label. The clustering arm re-partitions whatever it is given, so it wins on a clean day and returns a partition unrelated to the operating states on an unfamiliar one.

The trade between them is supervision, not accuracy. One day of process labels at commissioning buys the bank names for all four states and the explicit rejection output; the clustering arm needs no labels and names only what it isolates. Neither arm reads

process data at run time, so the choice reduces to what the first day of a new installation can supply, a question a plant operator can answer.

One further E1 observation generalises: the representation that wins the state task is not the one that wins the detection task. Handcrafted features and the audio TF-C encoder recover the four-state day; the BEATs encoder collapses to an adjusted Rand index near 0.000 where the states differ mainly in level, because a general-audio objective has little reason to preserve stationary loudness. Identifying the state and detecting the transient are two jobs, and this thesis uses one representation for each.

The honest weakness: the layer is good at the states that dominate a session and weak at the short ones, which accuracy hides and macro-averaged F_1 shows. The same shortage of recorded time also starves those states' calibration floors, so one cause underlies both, and more commissioning data repairs both.

6.1.2 Conditioning is what makes the false-alarm control real

The false-alarm guarantee is real only per state. Per-state conformal calibration holds every calibratable state near its budget wherever the day's calibration population is stable (the 30 June exception is the subject of Section 5.2.4); a single pooled threshold over the same windows reaches 0.311 in one state (Section 5.2.2). The pooled arm fails structurally: one state absorbs almost all flags while the rest sit far below budget, so the average looks fine while every component is violated. An average over states is not a guarantee for any state.

The sharper result: calibrating on the detected sub-clusters holds the budget better than on the official process labels, 0.069 against 0.188. The reason is exchangeability. A turbine at low load and one near full load are different acoustic regimes; one label across both makes the calibration set a mixture, which is exactly where the conformal guarantee stops holding. The better label was the wrong partition.

The scorer, by contrast, does very little work. No family dominates the sixteen cells of Table 5.5, the default nearest-neighbour scorer is the worst of the five in seven of them, and within a single day classical one-class scorers reach comparable false-alarm control. The contribution is therefore the conditioning and calibration layer around a swappable scorer (Section 3.4.4), not a scoring rule.

The hand-set threshold makes the same point from practice (Table 5.4). Averaged over one-second windows, its conventional multiplier resolves nothing; retuned, it detects but cannot keep its own error rate off the day it was tuned on, while the calibrated system

reaches its events at rates stated in advance. One number is a promise, the other a guess, and only the promise survives a day nobody tuned it on.

6.1.3 The controlled-event benchmark measures detection at a budget fixed in advance

The strike day is the campaign's only labelled evidence, and it is read under the strictest arrangement the data allow: references from a different instrumentation era, no refit on the day, event windows excluded from calibration. Under those conditions the standstill session resolves in full (90 of 90 at the nominal budget, 87 at the strictest) and the pump session reaches 84 of 90; the missing six are the landmark slots no method and no human resolves.

Two qualifications give those numbers their meaning. Pump is harder because the machine raises the broadband floor, so the same impact is a smaller excursion against a louder background: a property of the machine, not of the method. And without the day's threshold recalibration the frozen configuration flags most windows and would find every strike trivially (Section 5.2.2), which is why no detection rate in this thesis appears without its false-alarm rate.

The vibration branch contributes redundancy, not detection: it resolves almost no strikes, but holds its false-alarm rate at 0.041 across the very era boundary that breaks the acoustic branch, because the microphone-level step never reaches the accelerometers. Two sensing chains that fail on different days are worth more than accuracy in one. The full-coverage rerun adds the finer split: the vibration SSL encoder hears what the handcrafted vibration features cannot score, resolving 26 of the 90 standstill strikes, yet it breaks at the era boundary the handcrafted branch survives, because its channel mean admits the intermittent sensor-less input that the monitor's feature contract drops from the handcrafted branch (Table 5.8); what survives an era is the channel handling, not the modality. The audio TF-C encoder closes the pattern from the other side: industrial pre-training buys steady-state structure, general-audio pre-training buys transient sensitivity, and neither buys both.

6.1.4 A budget becomes a work list, along two paths

A per-window budget is a statistical object; an operator needs a list. Table 5.6 prices the conversion: under a strict budget with a persistence requirement, seven of the eight normal sessions emit six or fewer episodes. That is a work list. The hundreds of window

flags at the nominal budget never were one, because at $\alpha = 0.05$ the monitor flags 5 % of windows by instruction.

The same table breaks the rule it establishes: the persistence requirement leaves zero episodes on the standstill strike session although the strikes are detected at window level. A hammer strike lasts well under three seconds, so persistence deletes this class of anomaly entirely. The consequence is a two-path design, matching the two physical classes of event: a developing fault persists, a mechanical impact does not. Alarm design trades false-alarm suppression against detection delay [26]; the two classes sit on opposite sides of that trade, so no single policy serves both, and the thesis stops looking for one.

6.1.5 Drift becomes the deployment recipe rather than a nuisance

Carried across an era boundary, a frozen threshold loses false-alarm control completely, up to a realised rate of 1.00. Refitting every day repairs it and is not deployable: it demands a fresh block of normal data daily and would absorb a slowly developing fault into the new normal. The cadence the design commits to sits between: calibrate once, and let the system say when that calibration has expired. Under it, every held-out day stays at or below 0.075 and the strike detection is unchanged.

The drift signal was never designed as one. The bank rejects windows fitting no member so that the state layer can be honest about operating points it has never seen, and that rejection rate turned out to be the drift detector: thresholded once at 0.0805 from the commissioning bootstrap, it rises on every foreign-era day. Inside the era it stays quiet on every held-out turbine day and fires on both held-out pump days, the state the pool covers with a single session: two conservative extra triggers, and the right kind of error, since an unnecessary recalibration costs one commissioning block while a missed boundary costs false-alarm control for as long as it goes unnoticed. The level sentinel never fires; its margin of 0.921 sits far above the largest session excursion of 0.223 (Section 5.2.4), and reporting that silence keeps the cadence story from looking cleaner than it is.

The 30 June turbine session added the failure neither sentinel sees: thresholds broke inside a recognised operating state, invisible to state membership and unreflected in the level. The response is the third label-free trigger, on the monitor's own realised alarm rate (Section 5.2.4). The lesson generalises better than the constant: a conformal monitor carries its own drift signal in the gap between nominal and realised alarm rate, and a cadence that watches only its inputs is blind to that gap. That the alarm-richest normal

session is also the one whose sentinel came closest to firing (Section 5.2.3) is independent evidence that both instruments measure drift rather than noise.

6.1.6 The research question, answered in the form the evidence supports

The thesis asks whether self-supervised pre-training on public industrial sound datasets improves acoustic anomaly detection when fine-tuned on scarce plant data (Chapter 1). It contains three claims, about pre-training, about the pre-training corpus, and about fine-tuning; the measurements answer them differently, and that differentiation is the finding.

Pre-training wins where target data are absent altogether. The frozen BEATs encoder carries no plant-trained parameters at all and is the strict-budget event head: at standstill it detects all 90 strikes from the nominal budget upward and 87 of 90 at the strictest budget, at a realised rate of 0.003 there, on a day from an instrumentation era its calibration never saw. On the public proxy it reaches a clip-level AUC of 0.911 from eight normal clips (Section 5.2.6). The pre-trained representation separates before the plant has supplied anything.

Pre-training does not win the structure task; the pre-training corpus decides which task a representation can do. The BEATs encoder is unusable for state identification, the audio TF-C encoder recovers it, and on the pump strike session the ordering reverses. The question as posed therefore splits: industrial-sound pre-training helps the task the industrial corpus resembles and does not transfer to the other.

Adaptation answers in the negative, with a structural reason (Table 5.9). Full fine-tuning wins on the day it was adapted to and collapses to 0.364 on 29 June, the held-out day of its own instrumentation era, which is what memorising a few hours of one machine looks like. Low-rank adaptation avoids that collapse and stays within 0.006 of the frozen encoder on the same day. On the cross-era day of 25 June neither adapted arm holds up: frozen realises 0.007 against 0.069 and 0.051, an order of magnitude apart. Adaptation pays parameters for very little, and for less the further the day sits from the one it was fitted on. The best adaptation under scarcity is none, a usable engineering result rather than an absence of one.

At the boundary of the claim, transfer buys no within-day advantage in false-alarm control over classical one-class scorers on handcrafted features. It buys three other things: it removes the plant training set, supplies the strict-budget transient head, and survives compression. The distilled student holds 0.78 MB against 361.5 MB for its teacher, runs

roughly forty times faster, matches the teacher on the state task, and resolves 86 of the 90 standstill strikes and 79 of 90 pump strikes at $\alpha = 0.05$. The useful part of a 90.4-million-parameter model therefore reaches a plant server without a graphics processor, the form in which the answer to the research question can be deployed.

6.1.7 The configuration the evidence selects

Read as one decision per design axis of Table 3.1, the measurements select a single deployable pipeline; Table 6.1 states each choice with the number that decides it. This is the constructive form of the answer to the research question: self-supervised transfer earns exactly two places in the deployed monitor, the frozen general-audio encoder as the detection head and its distilled student as the edge variant, while every other axis is settled by conditioning, calibration and cadence rather than by learning.

Table 6.1: The configuration the evidence selects, one decision per design axis. Compared: the axes of Table 3.1 and the operating decisions around them, each resolved by the measurement named in the last column.

Decision	Selected option	Decisive evidence
Operating-state layer	model bank, commissioned on one labelled day	held-out mean 0.877 against the clustering arm's 0.582; names and rejection output (Table 5.2)
Detection representation	BEATs frozen (audio SSL)	90/90 and 84/90 strikes zero-shot at controlled rates (Table 5.3)
Modality	audio primary; vibration as an independent redundant branch, not fused	fusion dilutes strikes (84 to 54); vibration holds 0.041 frozen at the era boundary
Adaptation	none: encoder stays frozen	frozen 0.007 against 0.069/0.051 adapted, cross-era (Table 5.9)
Scoring	commodity choice; nearest-neighbour or Mahalanobis in the runtime	no family dominates sixteen cells; Mahalanobis transports best (10/12 day pairs)
Threshold conditioning	per-state split-conformal on detected sub-clusters	0.069 against 0.188 on process labels; pooled hides 0.311 in one state
Calibration cadence	once per era, guarded by three label-free sentinels	held-out sessions at or below 0.075; frozen breaks up to 1.000 (Table 5.7)
Alarm policy	two paths: strict budget with persistence for sustained faults, without persistence for transients	persistence yields 0–15 episodes per day and deletes every strike (Table 5.6)
Edge deployment	distilled student beside the full encoder	86/90 and 79/90 strikes at 0.78 MB and roughly forty times lower latency (Table 5.11)

6.2 Validity: what the campaign delivered and why the results are not overfitting

The design of Chapter 3 was written before the recordings existed and assumed more than the campaign produced: labelled operation across a series of days, at least one real in-service fault to anchor the fourth evaluation pillar (Table 3.4), one homogeneous instrumentation, continuous process data alongside the sensor streams, and a data-scarcity curve measured on plant recordings.

What arrived was six measurement days in three instrumentation eras, with process data permanently missing for one day and one further session, one session lost to an acquisition-restart fragment, and no fault (Section 5.1.1). Exactly one day carries controlled acoustic events, in two of the four operating states: 180 strikes, 172 of them resolved to measured timestamps by the register pipeline, with the ground truth self-annotated from the written protocol and verified against the audio (Section 5.1.3). That is the whole of the labelled evidence. Stating it plainly matters more than working around it; the design did work around it, and three inversions carried the thesis.

The instrumentation eras were the largest threat to any accuracy claim and became the strongest result. A monitor calibrated on one setup and run on another is what a deployed system faces after a sensor swap, a re-routed cable, or a gain change; a homogeneous campaign would have produced a cleaner table and a weaker thesis, leaving the cadence and the sentinels nothing to be tested against. The single strike day became a benchmark rather than an anecdote by being kept out of every fitting path: the one thing a single data point supports is an out-of-sample test of a system calibrated elsewhere. And the unlabelled days became the false-alarm control and the candidate register, which is not a consolation: validity under a stated budget is exactly what conformal calibration guarantees and needs no anomaly labels, so the unlabelled days measure what the method promises while the labelled day measures sensitivity.

One initial expectation proved wrong in an instructive way: a working monitor does not raise few alarms, it raises exactly as many as its budget specifies. The repair was the explicit policy layer (Section 5.2.3) rather than a better scorer, and the misreading is worth recording because an operator handed the raw flag rate would make it too.

What could not be shown remains unshown. There is no evidence here about detecting a real developing fault: none occurred, and the controlled events are impacts rather than degradations. The scarcity curve had to move to a public corpus, since such a curve needs abnormal clips to rank against; its plant-side counterpart instead asks how many normal

windows a state needs before its threshold is valid, and answers 19 to 319 for the states that stabilise within the measured grid.

Those same days are also the objection. The design choices and the constants of the pipeline were iterated on the very days the results are reported on, which gives the author researcher degrees of freedom over exactly that evidence. Four lines of defence answer it.

The first is the split scheme of Section 3.5: blocked by recording segment, three-way, and disjoint by construction, so no window that helped set a threshold can also test it. The scheme is enforced, not assumed. Windows consumed by a threshold fit are never alarmed, since that would break exchangeability in the anti-conservative direction, and disjointness is re-asserted in code before any window is scored (Section 4.4).

The second is that the headline numbers are held out by day, not by window. The state comparison is a mean over six held-out rotations rather than one favourable split, and it reports the two rotations where the weaker arm wins. The strike benchmark is held out in the strongest sense the campaign allows: foreign era, no refit of any kind, event-free calibration on top.

The third is architectural. The audio branch has no plant-trained parameters, so the only objects fitted on plant data are per-state reference sets and one quantile per state. There is very little there to overfit, and the ablation confirms it from the other direction, the variant that does fit plant data heavily, full fine-tuning, collapsing off-day. The design's default is the option with the fewest degrees of freedom, decided before that collapse was measured.

The fourth is what the thesis reports rather than hides. Table 5.10 is a table of failures, each with the number that killed it and the mechanism behind that number. A pipeline tuned until it looked good would not carry it, because the table is made of the settings that did not survive.

The residual risk is real and cannot be argued away from inside the data. Three constants carry that exposure: the minimum dwell time, the nominal budget, and the sentinel threshold. Each has a defence that is not a score. The dwell floor of five seconds comes from measured transition durations on this machine, real ramps of 20 to 138 seconds against load-step blips of six to nine seconds: the plant's own statistics rather than a sweep for a result (Section 4.3). The budget is the engineer's choice and is not fitted at all. The sentinel threshold comes from a bootstrap over the commissioning pool alone, never saw a foreign era, and so owes nothing to the days it is evaluated on.

The standing defence is procedural. Measurement days arriving after the design is frozen are declared held out before anyone looks at them and replayed through the once-calibrated pipeline with no refitting, under an intake checklist fixing what is verified first: the process timebase against the audio timestamps, the state of the accelerometer inputs, the acquisition header that decides the era, and process coverage. Recording days that arrive in the future are the material this defence meets, with the three-sentinel set now fixed in advance, and the commitment is to report the number replay produces rather than reconstruct a claim afterwards.

6.3 What transfers to another plant

Four things transfer unchanged. The first is the procedure: record a commissioning block, run the calibration, obtain thresholds meeting a chosen budget; the engineer picks the tolerated false-alarm rate and the procedure returns the threshold that delivers it, with a finite-sample and distribution-free guarantee, so nothing is hand-tuned at the new plant. The second is the architecture: one shared frozen encoder with a small per-state reference and threshold at the scoring head (Figure 3.4), keeping the per-state cost at a reference set rather than a replicated backbone. The third is the sentinel construction: fit the model bank, measure its rejection rate on the commissioning pool, take a bootstrap quantile, watch that rate. The fourth is the two-path alarm policy, which follows from the physics of the two event classes rather than from this plant. Nothing in that recipe refers to this machine.

What does not transfer is every number. The per-state thresholds, the reference sets, the sentinel quantile of 0.0805 and the dwell floor are empirical fits to this unit and this instrumentation, and a new plant re-measures all of them on its own commissioning data. A method whose constants transferred without re-measurement would claim something about machines it has never heard.

The data a new plant has to supply are minutes, not a training set. The conformal floor of Section 3.4.5 fixes that lower bound rather than leaving it to judgement: 19 to 319 windows for the states that stabilise within the measured grid, about twenty seconds to about five minutes of effective recording per state, while a minority of state cells needs the full pool or never stabilises. On the public proxy the frozen encoder reaches a clip-level AUC of 0.911 from eight normal clips, evidence about a different machine type in a different room and labelled as such throughout (Section 5.2.6). The two bound the scarcity question from both sides, and neither number is large.

A new deployment then runs as a short loop. Record one commissioning block covering the states the unit runs, with process data for it; fit the model bank on those labelled windows, name the states, take the bootstrap sentinel quantile; fit per-state references and conformal thresholds at the chosen budget on the detected sub-clusters, freeze the vibration references, and ship one snapshot file to the plant server. From then on the monitor runs label-free and refits when the sentinel fires. Only the first step needs a human with process access; where even that is unavailable, the unsupervised arm runs the same loop with weaker agreement, naming only the states it isolates.

6.4 Limitations

The limitations are collected here, so that a reader deciding what to believe need not assemble them from six chapters.

One machine, one site. The evidence base is a single unit at one plant. Portability is built into this design and argued for, not demonstrated across plants. Nothing here measures what happens at a second unit.

A thin labelled inventory. The 180 controlled strikes, 90 per session over two sessions, cover standstill and pump operation only, so the labelled benchmark says nothing about two of the four states. All events are mechanical impacts, a different class of signal from the developing degradations condition monitoring exists to catch. No in-service fault occurred, so the fourth evaluation pillar is realised as a planned qualitative study rather than a confirmed case.

Under-sampled transitions. Ramp windows carry their own label class, but the recordings contain too few ramps to calibrate a transition reference well. In the default runtime they are scored against the neighbouring steady operating state and flagged rather than suppressed, while impulse-like anomalies inside a ramp still reach the transient path. A sustained anomaly developing during a ramp is the case this evaluation cannot yet quantify.

Single-annotator ground truth. The protocol was logged at minute granularity and no per-strike timestamps existed, so the second-resolution reference behind the per-strike results was annotated by the author, then corrected for duplicates introduced by the overlapping snippet padding. Six pump-session landmark strikes carry no blind marks: neither the annotator nor a validated targeted search could locate them under pump noise. The procedure is documented and versioned, and remains a single-annotator reference.

Unexplained era boundaries. The plant operator supplied no setup-change log, so the header change, the microphone level step and the accelerometer-input state change are reported as observed facts rather than classified as deliberate reconfiguration or unintended drift. The cadence result inherits that limit: it shows the mechanism works when a boundary occurs, and says nothing about how often boundaries occur in a production installation.

Discarded array geometry. The learned encoders take a single channel, so each microphone stream is mono-mixed before encoding and whatever the geometry carries within a ring is lost there (Section 4.2). The handcrafted branch keeps the channels. Removing this limitation means replacing the encoder rather than tuning it, so the comparison between the two representation families is also one between a single mixed channel and the full array.

A register without ground truth. The candidate register holds 159 candidates from eleven sessions, and every entry awaits the expert listening pass. Its one highlighted candidate is corroborated only by a window-level coincidence with the collaborating team's independent observation and by the register's own band-energy criterion, not by any listening judgement. Nothing in the reported quantitative results depends on a register entry, which is the point, but the register therefore contributes no confirmed detection.

A re-implemented practice baseline. The fixed-threshold comparison is an independent re-implementation of a method type inside this thesis's own one-second window harness. It is not a measurement of the collaborating project's detector, which operates at raw impulse resolution, and is not offered as one. The comparison is fair to the method type within this harness and says nothing about impulse-resolution implementations of it.

Six measurement days. The campaign spans six measurement days within seven weeks, and every number in this thesis rests on them. The replay therefore samples few days and only the era boundaries they happen to contain; how the monitor behaves over months of operation, across seasons, or around maintenance interventions is not measured.

Retrospective replay and operator knobs. The replay is a day-granular retrospective simulation over recorded files rather than an online service. The alarm policy contains operator choices, the budget per state and the persistence rule, that this thesis prices but does not set. Both are stated at that scope everywhere they are reported.

With the evidence weighed and its limits stated, Chapter 7 closes the thesis: the answer to the research question, the contribution check, and the work that follows.

Chapter 7

Final considerations

This chapter closes the thesis: it answers the research question in the form the measurements support, checks the announced contributions against what was delivered, states the boundary of the claim, and names the work that follows.

7.1 Conclusion

The thesis asked one question: *Does self-supervised pre-training on public industrial sound datasets improve acoustic anomaly detection when fine-tuned on scarce PSHP data?* The measurements answer it in three parts, and the split is itself the finding.

Pre-training helps, and helps most where the plant supplies least. The frozen BEATs encoder carries no plant-trained parameters. It detects every controlled event of the standstill session from the nominal budget upward and 87 of 90 at the campaign's strictest, at a realised false-alarm rate of 0.003 there, on a day from an instrumentation era its calibration never saw, with median first-alarm latency below one second. On the public proxy corpus it reaches a clip-level AUC of 0.911 from eight normal clips. Where data are scarce and transients must be caught at a strict budget, transfer is what makes a detector work without a plant training set.

Pre-training does not transfer everything; the pre-training corpus decides what it does transfer. The BEATs encoder is unusable for identifying operating states that differ mainly in level, while the audio TF-C encoder recovers that structure and is in turn the weakest representation on the pump strike session. Identifying the state and detecting the transient are two jobs, and this thesis uses one representation for each.

Fine-tuning, the third element of the question, answers in the negative. Full fine-tuning wins on the day it was adapted to and falls to 0.364 on a day it was not, while low-rank adaptation degrades gracefully, staying close to the frozen encoder. Under scarcity the best adaptation is none.

One result cuts across all three parts. Within a single day, classical one-class scorers on handcrafted features reach comparable false-alarm control, making the scoring method a commodity choice. What deploys is the layer around it: the operating state that fixes what normal means, the per-state conformal calibration turning a chosen budget into a threshold, and the cadence that says when that threshold has expired. Chapter 6 argues that reading and collects the limitations in full.

Six contributions were announced in Chapter 1; each is checked here against what was delivered.

Operating-state identification. Delivered with both arms scored on sessions neither was fitted on. The commissioning-time model bank leads at 0.877 in the within-era mean and 0.96 on the cross-era rotation, and the trade between the arms is stated as one in supervision rather than accuracy. The lead is an average, not a rule: on the era-C strike session the clustering arm reaches 0.948 against the bank's 0.710.

State-conditioned detection with false-alarm guarantees. Delivered as per-state split-conformal calibration on detected sub-clusters. A single pooled threshold over the same windows realises 0.311 in its worst state.

Adaptive calibration cadence. The headline result. A frozen threshold loses control at an era boundary and daily recalibration is not deployable; the label-free rejection sentinel fires on every session recorded outside the calibration era, with two conservative extra triggers inside it. The 30 June turbine session showed that thresholds can also decay inside an era without that sentinel noticing, and the third trigger on the monitor's own realised alarm rate, added in response to that finding, catches exactly that day. One calibration per era plus its sentinels holds every held-out session at or below 0.075.

A comprehensive transfer-learning ablation. Delivered as task-dependent winners rather than a single one, across all eight representations, and a 0.78 MB distilled student carrying the strict-budget transient head on a plant server without a graphics processor.

A controlled-event benchmark. Delivered as a per-strike register resolving 172 of the 180 protocol strikes to measured timestamps (the remaining eight inferred from the strike rhythm), event-free calibration, and 90 of 90 standstill and 84 of 90 pumping strikes detected at budgets fixed in advance.

A deployable monitoring artifact. Delivered as named states, a two-path alarm policy, a versioned candidate register with its review tooling, and a dashboard; a strict budget with a persistence requirement turns a day of window flags into between zero and fifteen episodes per normal session.

What the thesis claims is narrow, and is stated at that width: a procedure that runs on one machine at one site, calibrated once per instrumentation era from a commissioning block and monitored for its own expiry. What would transfer to a second plant is that procedure and the false-alarm budget it is given, not the numeric constants fitted here. It does not claim a result across plants, nor the detection of a developing in-service fault: none occurred, and the labelled evidence is one campaign day of mechanical impacts in two of the four operating states. The drift mechanism is shown at the boundaries the recordings happen to contain, and the rate at which such boundaries occur in a production installation is not measured here.

7.2 Future work

The work that follows is given in the order in which it would repay the effort: two items strengthen the evidence, two close known method gaps, one follows directly from the 30 June finding, and four extend the representation and the system around it.

1. **An expert listening pass over the candidate register**, since only an expert audition can turn its entries into confirmed detections rather than false-alarm rates.

2. **A pre-registered replay of further recording days** as they become available, since days that informed no design decision are the strongest available defence against the suspicion of tuning.

3. **A transition-conditioned reference** fitted on ramps of their own, covering the one case this evaluation cannot quantify: an anomaly developing while the machine changes state.

4. **A state-naming path that reads no process data**, removing the single supervised step from the deployment loop.

5. **Load-aware commissioning and conditioning.** The exploratory checks of Section 5.2.4 bound the shape of the solution: finer conditioning starves its own calibration floor, pool diversity transports where pool volume does not, and the

drift driver is acoustic and unlogged, which is why the alarm-rate watchdog is the catch-all. The consequence is a commissioning recording that deliberately drives each state's load envelope until every intended sub-state clears the conformal floor, an excess-alarm report over state and load band when the watchdog fires, and recalibration of only the flagged states as a refinement. A controlled-event campaign inherits the same rule: microphones must be placed so that every structure the protocol intends to excite is acoustically covered.

6. **An encoder that keeps the microphone array**, since the learned branch mono-mixes each ring and discards whatever the differences between its microphones carry.

7. **Deeper fusion on more data**, since the cross-attention head calibrated tighter than feature-level fusion only where a state carried enough windows to fit it, and one session is too thin a base for that advantage.

8. **Integration with source localisation**, since an operator acting on an alarm also needs to know where the sound came from.

9. **A wider set of encoders** (among them M2D-X, a masked-modelling audio encoder line), since the representation axis is where the measured differences are largest.

7.3 Data and code availability

The implementation described in Chapter 4 is maintained as one repository at https://github.com/StefanKrum/rowii-monitor, covering the ingestion, state, anomaly and evaluation code, the runtime tools and the dashboard, together with a demonstration replay of four recorded sessions that runs locally in a browser. Access to the repository is available upon request, since the data from illwerke vkw AG is confidential. Its automated test suite runs without plant data; tests needing real recordings carry a separate marker and are skipped when no recordings are present, so the pipeline can be installed and exercised without access to the campaign. The recordings themselves are proprietary plant data of the operator, are not redistributed with this thesis, and are available on reasonable request through the research partners, subject to the operator's approval. The public-proxy scarcity study of Section 5.2.6 uses an openly available industrial-sound corpus and is therefore reproducible independently of any plant access.

Appendix A
Reference tables

This appendix holds three reference tables that the main text points to but does not need to display: the map from design concept to implementing module, the thresholds of the SCADA ground-truth rule, and the criteria that build the candidate register.

A.1 Design concept to module map

Table A.1 resolves each load-bearing concept of Chapter 3 to the module of the implementation that carries it. It is the lookup table for a reader who wants to find a mechanism described in Chapter 4 in the source tree.

A.2 The SCADA ground-truth rule

Table A.2 lists every threshold of the rule-based process classifier of Section 4.3, grouped by the four stages in the order they are applied. The rule supplies the labelled fit windows of both state-detection arms and the ground truth that Section 5.1.2 scores them against, so its thresholds are stated in full rather than summarised.

A.3 Candidate-register criteria

The candidate register is the shortlist of windows handed to a human reviewer on the days that carry no controlled acoustic events. Three criteria admit a window to it, listed in Table A.3. Two of them read the p-values a monitor of Section 4.5 already produces, on two different representations. The third reads the raw microphone signal instead, and exists because a shortlist assembled from embeddings alone can structurally miss short mechanical impulses.

Four rules turn the three criterion outputs into one list. Overlapping candidates are resolved in favour of the more extreme of the pair, where extremity is the lowest p-value and an impulse pair is scored through the more conservative of its two rings' z-scores. An impulse candidate within ± 2 s of a surviving transient candidate is dropped before the merge, because the two paths are different signal-processing views of one physical event rather than two independent detections, so a window that satisfies an embedding criterion and the impulse criterion is listed once. At most 25 candidates are kept per

Table A.1: Design concept to module map. The load-bearing concepts of Chapter 3 each resolve to a named module, which is what allows the runtime artifact to be assembled from library code alone. Library paths are relative to `src/rowii/`, script paths to the repository root.

Design concept	Module	Role
Ingestion and time base (§3.2)		
Logger format	`io/gantner.py`	Validating reader for the plant's binary container
Discovery, time base	`io/dataset.py`	Name-only session index; derives the acquisition-clock offset per run
Window grid	`signals/windows.py`	Shared one-second UTC grid over the intersection of the streams
Run preparation	`pipeline.py`	Discover, grid, chunked featurize, validity mask, feature cache
Representations (§3.4.1)		
Handcrafted features	`signals/features.py`	Audio and vibration featurizers, machine-frequency bands, per-stream z-scoring, fusion
Frozen audio encoder	`signals/beats.py`	Embeddings from the pre-trained transformer, no plant-specific training
Compact encoders	`adapt/student.py`, `tfc/wrapper.py`	Distilled student and industrially pre-trained contrastive encoder
Operating states, Step 1 (§3.3)		
Unsupervised arm	`state/cluster.py`, `state/smooth.py`, `state/segments.py`	k-means, sticky hidden Markov smoothing, minimum-dwell filter
Supervised arm	`state/modebank.py`	Per-model bank; label-free assignment and whole-bank rejection
State naming	`eval/metrics.py`	Majority-vote map from cluster or model identifier to the process vocabulary
Process labels	`scada/labels.py`	Rule-based state ground truth, used for fitting and validation only
State-conditioned anomaly layer, Step 2 (§3.4.5)		
Splits and references	`anomaly/references.py`	Segment-blocked splits and per-state normal reference matrices
Scorers	`anomaly/scorers.py`	Cosine nearest-neighbour and shrinkage Mahalanobis distance
Conformal thresholds	`anomaly/conformal.py`	Split-conformal threshold, p-values, low-confidence flag
Sweeps	`anomaly/sweep.py`	Per-state and pooled arms, leakage guard, false-alarm tables
Candidate register	`scripts/candidate_kit.py`	Criterion-based shortlist, review assets, versioned assessments
Calibration cadence (§3.4.6)		
Drift sentinels	`anomaly/sentinels.py`	Label-free day-level triggers on rejection rate and level step
Replay driver	`scripts/run_once_calibrated.py`	Chronological replay with the once-per-era recalibration decision and the alarm-rate watchdog
Runtime		
Deployable artifact	`runtime/snapshot.py`	Pickle-free snapshot; save, load and numeric round-trip
Monitor	`scripts/monitor.py`	Applies one snapshot to a new recording under three threshold regimes
Evaluation and communication (§3.5)		
Event harness	`eval/events.py`, `scripts/eval_events.py`	Per-event detection, first-alarm latency, false-alarm rate outside events
Annotation kit	`scripts/annotation_kit.py`	Clips, spectrograms and templates for second-resolution annotation
Analysis suite	`scripts/analyze_days.py`	Cross-day figures computed from the evaluation caches
Dashboard assets	`scripts/make_demo_assets.py`	Self-contained replay page with state timeline and audio snippets

Table A.2: The SCADA ground-truth rule. Every threshold is anchored to a measured plant quantity rather than to a convention: the speed gate to the 378.8 rpm nominal, the phase-shifter power window to the roughly 3.5 MW measured idling power, and the transition buffer to the 20 s shortest real transition. An independent re-implementation of the collaborating team's own rules agrees with the labels this rule produces on 97.3 % of windows. Implemented as `rowii.scada.labels.gt_labels` and grouped by its four stages in the order applied: base classification, phase-shifter promotion, ramp override, and transition buffer.

Test	Value	Rationale
Speed gate	378.8 rpm nominal; standstill $< 5\,\%$, steady operating state $\geq 95\,\%$	Excludes run-up/run-down windows that already draw power from being read as steady.
Steady-operation power floor	2.0 MW	With the speed gate, separates standstill from loaded turbine or pump.
Turbine/pump split	flow dominance; power sign as fallback	Pump power is occasionally logged positive at this plant.
Phase-shifter candidate power	$\leq$ 5.0 MW at nominal speed	Wider than the 2.0 MW floor; measured idling power sits near 3.5 MW.
Phase-shifter dwell	$\geq$ 600 s (10 min), contiguous	Excludes the brief zero-power crossing an ordinary state change makes.
Phase-shifter valve gate	inlet valve reading ≤ 10 (closed)	Confirms draft-tube dewatering, not just unloaded spinning.
Ramp override	$\lvert dP/dt \rvert > 1.0$ MW/s	Flags a mid-ramp window as transition even if its power looks steady.
Transition buffer	$\pm$10 s around every operating-state boundary	Shorter than the shortest real transition (20 s); a ramp window belongs to no steady operating state's normal.

session, ranked by extremity. The controlled acoustic events of the campaign day are excluded, since the register exists for the days without them.

Two properties keep the register honest about its own search. Every raw single-stream peak is written to the search trace whether or not it found a partner on the other ring, so a reader can see not only what the impulse criterion selected but everything it looked at. And the register is explicitly not ground truth: it is a handover artifact for manual review, and its assessment vocabulary records an honest no-finding as a complete answer. What the register held is reported in Section 5.2.2; the expert listening pass over it remains open.

Table A.3: The three register criteria read two different kinds of evidence, which is why one shortlist needs all three. The two embedding criteria consume monitor p-values on two representations; the impulse criterion consumes the raw signal and requires agreement between the two microphone rings, so a single-channel sensor artefact cannot enter the register. Implemented in `scripts/candidate_kit.py`, with peak detection in `rowii.anomaly.impulse.detect_impulses`.

Criterion	Evidence	Rule
Sustained	p-values of the fusion monitor	Episode of at least three consecutive scored seconds at $p < 0.01$. Episodes are grouped by strict window-index adjacency, so a skipped or invalid grid window never bridges two of them.
Transient	p-values of the pre-trained-encoder monitor	Single scored window at $p < 0.001$, deduplicated within ± 5 s of a more extreme neighbour.
Impulse	raw 5–20 kHz band energy on both microphone streams	Band energy on 10 ms frames at 5 ms hop, detrended by a one-second rolling median and standardised as a median-absolute-deviation z-score. Accepted at $z \geq 6$ only when both microphone rings carry a coincident peak within 0.15 s.

A.4 Session inventory

Table A.4 lists every recording session of the campaign with its instrumentation era, its evaluated window count and its role in the evaluation splits, complementing the timeline view of Figure 5.1 with exact numbers.

Table A.4: Session inventory. Evaluated windows are the fusion-arm counts of Table 5.1 (per-representation counts differ slightly by channel validity); one further acquisition-restart fragment is dropped by the ingestion guard and carries no session row.

Session	Era	Eval. windows	Role
25 Jun, turbine	A	8 275	held out
25 Jun, pump, morning	A	1 439	held out (single ground-truth class)
25 Jun, pump, afternoon	A	—	excluded: outside the operational export
27 Jun, pump and phase shifter	A	—	sentinel-only: no process export
29 Jun, turbine	B	15 564	held out
29 Jun, pump	B	21 510	held out
30 Jun, turbine, morning	B	7 733	held out (channel gaps masked)
30 Jun, pump, midday	B	11 204	held out
1 Jul, turbine, morning	B	11 039	commissioning pool
1 Jul, turbine, second	B	4 561	commissioning pool
1 Jul, turbine and phase shifter	B	27 821	commissioning pool
1 Jul, pump	B	4 725	commissioning pool
8 Jul, pump, strikes	C	12 808	controlled events
8 Jul, standstill, strikes	C	1 415	controlled events (single ground-truth class)

A.5 Canonical representation names

Every table, figure and text passage of this thesis uses exactly one name per representation; Table A.5 maps these names to the repository's variant keys for reproducibility.

Table A.5: Canonical representation names and repository keys. The baseline tag marks the handcrafted comparison arms; the hand-set MAD threshold (practice baseline) of Table 5.4 is a decision rule rather than a representation and carries no variant key.

Canonical name	Repository key	Pre-training
Handcrafted audio (baseline)	`audio`	none
Handcrafted vibration (baseline)	`vibration`	none
Handcrafted fusion (baseline)	`fusion`	none
BEATs frozen (audio SSL)	`audio-beats`	AudioSet
TF-C (audio SSL)	`audio-tfc`	MIMII
TF-C (vibration SSL)	`vibration-tfc`	CWRU, Paderborn
Fusion on BEATs (SSL)	`fusion-beats`	AudioSet (audio side)
Distilled student	`audio-student`	distilled from BEATs on era-B plant audio
From-scratch autoencoder	`logmel` + LSTM autoencoder	none

A.6 Full-coverage matrices

Tables A.6 and A.7 carry the complete grids behind the summary views of Chapter 5: every model-bank family on every rotation, and every representation's replay under the three calibration regimes.

Table A.6: Model-bank ARI for all three member families. Compared: bank state-ARI per held-out rotation, representation and member family; the family rule of Table 5.2 picks Gaussian members for the first three columns and nearest-neighbour members for the remaining five. ° degenerate single-class session; n/a = era-C vibration channel set (108 against 96 columns, the intermittent sensor-less input of 8 July included) refused by the bank's feature contract.

Held-out session	HC audio	HC vib.	HC fusion	BEATs	TF-C aud.	TF-C vib.	Fus. BEATs	Student
Gaussian members								
29 Jun, turbine	0.91	0.95	0.99	0.75	0.63	0.55	0.71	0.98
29 Jun, pump	0.96	0.97	0.71	0.31	0.36	-0.04	0.81	0.34
30 Jun, turbine, morning	0.99	0.35	0.20	0.64	0.21	0.18	0.15	0.99
30 Jun, pump, midday	0.83	0.87	0.05	0.07	0.46	0.02	0.20	0.17
1 Jul, turbine and phase shifter	0.46	0.31	0.01	0.69	0.81	0.02	-0.06	0.99
1 Jul, pump	0.85	0.86	0.86	0.56	0.58	0.82	0.86	0.21
25 Jun, turbine	0.99	0.85	0.16	0.54	0.04	0.03	0.02	0.99
25 Jun, pump, morning°	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
8 Jul, pump, strikes	0.96	n/a	n/a	0.72	0.87	-0.00	n/a	0.71
Nearest-neighbour members								
29 Jun, turbine	0.98	0.99	0.90	0.99	0.97	0.95	0.99	0.98
29 Jun, pump	0.66	0.09	0.03	0.92	0.16	0.81	-0.01	0.59
30 Jun, turbine, morning	0.85	0.98	0.63	0.99	0.99	0.93	0.86	0.99
30 Jun, pump, midday	0.78	0.99	0.37	0.94	0.33	0.97	0.07	0.93
1 Jul, turbine and phase shifter	0.21	0.38	0.46	0.55	0.86	0.06	0.57	0.52
1 Jul, pump	0.85	0.67	0.16	0.87	0.41	0.61	0.23	0.68
25 Jun, turbine	0.97	0.98	0.83	0.96	0.88	0.91	0.99	0.99
25 Jun, pump, morning°	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00
8 Jul, pump, strikes	0.53	n/a	n/a	0.71	0.42	0.98	n/a	0.19
Gaussian-mixture members								
29 Jun, turbine	0.82	0.99	0.94	0.94	0.72	0.68	0.99	0.97
29 Jun, pump	0.98	0.82	0.24	0.59	0.52	0.38	0.27	0.95
30 Jun, turbine, morning	0.84	0.99	0.54	0.64	0.23	0.11	0.60	0.83
30 Jun, pump, midday	0.96	0.84	0.05	0.89	0.49	0.81	0.05	0.92
1 Jul, turbine and phase shifter	0.96	0.16	0.10	0.82	0.68	0.29	0.70	0.99
1 Jul, pump	-0.01	-0.01	0.00	0.54	0.24	0.28	-0.00	0.16
25 Jun, turbine	0.94	0.89	0.44	0.60	0.12	0.81	0.67	0.87
25 Jun, pump, morning°	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
8 Jul, pump, strikes	0.82	n/a	n/a	0.94	0.73	0.19	n/a	0.82

Table A.7: Calibration regimes for every representation: always-frozen / once-per-era realised false-alarm rate per replayed session. Compared: the same nine-session replay as Table 5.7, repeated per representation at $\alpha = 0.05$; the summary view is Table 5.8. ° the 25 June pump session is in the replay set but carries a single ground-truth class for the state layer.

Session	HC audio	HC vib.	HC fusion	BEATs	TF-C aud.	TF-C vib.	Fus. BEATs	Student
25 Jun, turbine	0.25/0.00	0.07/0.03	0.06/0.03	0.48/0.02	0.38/0.01	0.08/0.05	0.35/0.01	0.51/0.01
25 Jun, pump, morning°	0.39/0.06	0.01/0.03	1.00/0.03	0.10/0.01	0.12/0.01	0.32/0.00	0.90/0.01	0.00/0.00
29 Jun, turbine	0.51/0.10	0.11/0.11	0.07/0.07	0.23/0.07	0.24/0.03	0.24/0.03	0.16/0.01	0.33/0.02
29 Jun, pump	0.92/0.00	0.01/0.04	0.30/0.03	0.69/0.04	0.23/0.01	0.24/0.02	0.21/0.01	0.38/0.01
30 Jun, turbine, morning	0.10/0.00	0.09/0.09	0.38/0.06	0.10/0.10	0.05/0.01	0.09/0.00	0.07/0.01	0.09/0.02
30 Jun, pump, midday	0.93/0.01	0.02/0.06	0.67/0.04	0.12/0.04	0.02/0.02	0.06/0.01	0.15/0.01	0.02/0.02
1 Jul, turbine and phase shifter	0.07/0.06	0.06/0.06	0.02/0.02	0.05/0.05	0.03/0.02	0.07/0.05	0.02/0.02	0.09/0.09
1 Jul, pump	0.00/0.00	0.00/0.00	0.19/0.10	0.03/0.03	0.00/0.00	0.00/0.00	0.08/0.00	0.00/0.00
8 Jul, pump, strikes	0.68/0.02	0.04/0.05	0.62/0.05	0.22/0.05	0.12/0.01	0.91/0.02	0.51/0.02	0.06/0.02

A.7 Supplementary replay figures

Three figures supplement the replay results of Section 5.2.4. Figure A.1 carries the day-by-day transfer grid behind the across-days finding, and Figures A.2 and A.3 plot the three regimes and the first sentinel across the replay.

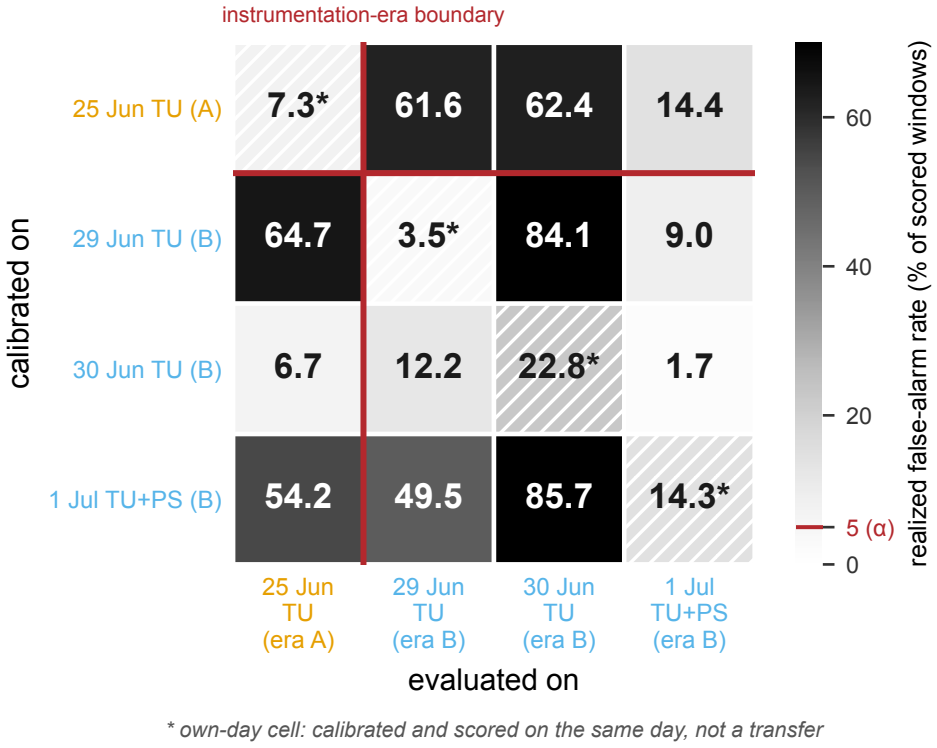

Figure A.1: Day-by-day transfer of a single-day calibration. Compared: each day's own calibration carried onto each other day (handcrafted fusion, nearest-neighbour scoring, pooled reference, $\alpha = 0.05$); only the twelve off-diagonal cells are transfer measurements, and the starred diagonal is each day's own within-day split.

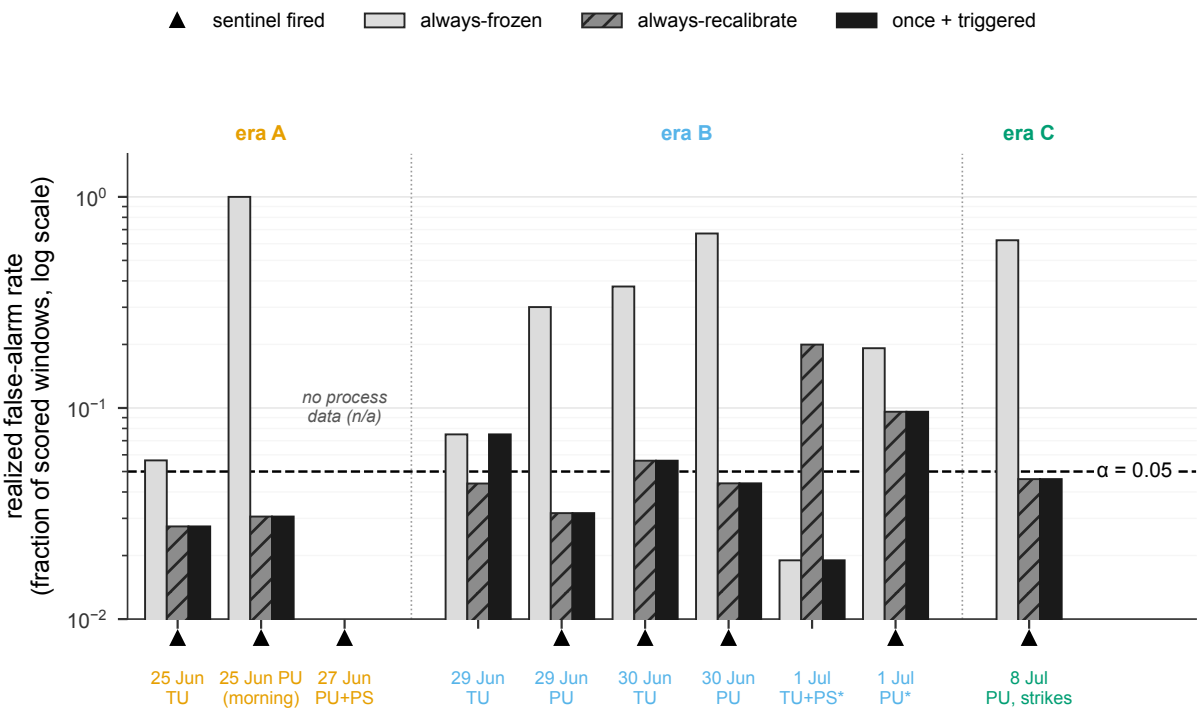

Figure A.2: False-alarm rate under the three calibration regimes across the replay. Same data as Table 5.7, one bar group per replayed session in chronological order (log scale); the dashed line is the nominal budget, and a triangle marks a session on which a label-free sentinel fired.

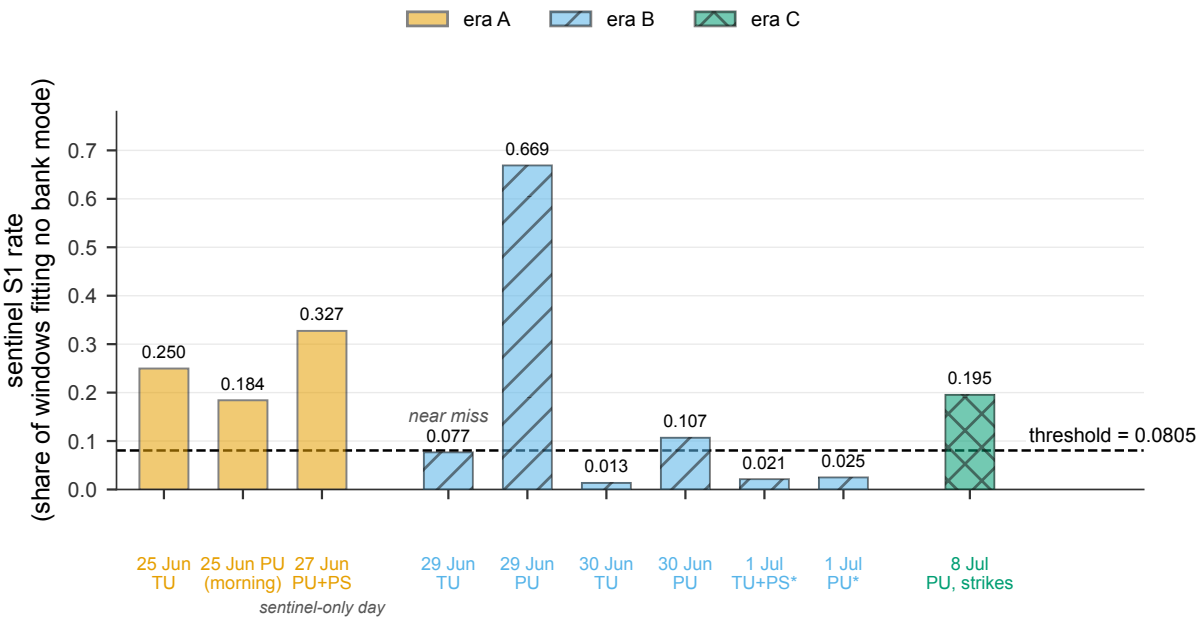

Figure A.3: Sentinel s1, the model-bank rejection rate per replayed session. The share of each session's windows that fit no bank member, against the fixed threshold of 0.0805 from the commissioning-pool bootstrap (dashed line); a bar above the line fires the sentinel.

Bibliography

[1] Khalid M. Almutairi and Jyoti K. Sinha. "A Comprehensive 3-Steps Methodology for Vibration-Based Fault Detection, Diagnosis and Localization in Rotating Machines". In: *Journal of Dynamics, Monitoring and Diagnostics* 3.1 (2024), pp. 49–58. DOI: 10.37965/jdmd.2024.484.

[2] Analog Devices. *Making Sense of Sounds: How AI Can Boost Your Machine Uptime*. 2024.

[3] ANDRITZ. *Metris DiOMera*. 2024.

[4] ANDRITZ. *Metris Vibe Condition Monitoring*. 2024.

[5] Anastasios N. Angelopoulos and Stephen Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. 2022. arXiv: 2107.07511 [cs.LG].

[6] Ali Akbar Taghizadeh Anvar and Hossein Mohammadi. "A Novel Application of Deep Transfer Learning with Audio Pre-Trained Models in Pump Audio Fault Detection". In: *Computers in Industry* 147 (2023), p. 103872. ISSN: 0166-3615. DOI: 10.1016/j.compind.2023.103872.

[7] Augury. *Machine Health*. 2024.

[8] Automation.com. *What Does a Good Machine Sound Like? A Look at 3DSignals*. 2017.

[9] Baker Hughes. *System 1 Condition Monitoring Software (Bently Nevada)*. 2024.

[10] Alessandro Betti, Emanuele Crisostomi, Gianluca Paolinelli, Antonio Piazzi, Fabrizio Ruffini, and Mauro Tucci. "Condition Monitoring and Predictive Maintenance Methodologies for Hydropower Plants Equipment". In: *Renewable Energy* 171 (2021), pp. 246–253. ISSN: 0960-1481. DOI: 10.1016/j.renene.2021.02.102.

[11] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Computing Surveys* 41.3 (2009), pp. 1–58. DOI: 10.1145/1541880.1541882.

[12] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. *BEATs: Audio Pre-Training with Acoustic Tokenizers*. 2022. arXiv: 2212.09058 [eess.AS].

[13] Gabriel Coelho, Luís Miguel Matos, Pedro José Pereira, André Ferreira, André Pilastri, and Paulo Cortez. "Deep Autoencoders for Acoustic Anomaly Detection: Experiments with Working Machine and in-Vehicle Audio". In: *Neural Computing and Applications* 34.22 (2022), pp. 19485–19499. ISSN: 1433-3058. DOI: 10.1007/s00521-022-07375-2.

[14] E. Di Fiore, A. Ferraro, A. Galli, V. Moscato, and G. Sperlì. "An Anomalous Sound Detection Methodology for Predictive Maintenance". In: *Expert Systems with Applications* 209 (2022), p. 118324. DOI: 10.1016/j.eswa.2022.118324.

[15] Xavier Escaler, Eduard Egusquiza, Mohamed Farhat, François Avellan, and Miguel Coussirat. "Detection of Cavitation in Hydraulic Turbines". In: *Mechanical Systems and Signal Processing* 20.4 (2006), pp. 983–1007. ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2004.08.006.

[16] Fausto Pedro García Márquez, Alfredo Peinado Gonzalo, and Mayorkinos Papaelias. "A Comprehensive Review of Condition Monitoring Systems for Hydropower Stations: Technologies, Applications, and Future Trends". In: *Electric Power Systems Research* 251 (2026), p. 112339. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2025.112339.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. Mar. 2015. arXiv: 1503.02531.

[18] H. Hojjati and N. Armanfard. "Self-Supervised Acoustic Anomaly Detection Via Contrastive Learning". In: *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 3253–3257. DOI: 10.1109/ICASSP43922.2022.9746207.

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. June 2021. arXiv: 2106.09685.

[20] Lawrence Hubert and Phipps Arabie. "Comparing Partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075.

[21] Rony Ibrahim, Ryad Zemouri, Antoine Tahan, Bachir Kedjar, Arezki Merkhouf, and Kamal Al-Haddad. "Early Detection of Rotor Faults in Large Hydrogenerators Using Vibration Measurements, Variational Autoencoder, and Euclidean Distance". In: *IEEE Transactions on Industry Applications* 61.6 (2025), pp. 9023–9032. ISSN: 1939-9367. DOI: 10.1109/TIA.2025.3571883.

[22] International Organization for Standardization. *ISO 13373-1:2002 Condition monitoring and diagnostics of machines – Vibration condition monitoring – Part 1: General procedures*. Standard. 2002. URL: https://www.iso.org/standard/21831.html.

[23] International Organization for Standardization. *ISO 13379-1:2012 Condition monitoring and diagnostics of machines – Data interpretation and diagnostics techniques – Part 1: General guidelines*. Standard. 2012. URL: https://www.iso.org/standard/39836.html.

[24] International Organization for Standardization. *ISO 17359:2018 Condition monitoring and diagnostics of machines – General guidelines*. Standard. 2018. URL: https://www.iso.org/standard/71194.html.

[25] International Organization for Standardization. *ISO 20816-5:2018 Mechanical vibration – Measurement and evaluation of machine vibration – Part 5: Machine sets in hydraulic power generating and pump-storage plants*. Standard. 2018. URL: https://www.iso.org/standard/67919.html.

[26] Iman Izadi, Sirish L. Shah, David S. Shook, Sandeep R. Kondaveeti, and Tongwen Chen. "A Framework for Optimal Design of Alarm Systems". In: *IFAC Proceedings Volumes* 42.8 (2009), pp. 651–656. ISSN: 1474-6670. DOI: 10.3182/20090630-4-ES-2003.00108.

[27] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. "Leakage in Data Mining: Formulation, Detection, and Avoidance". In: *ACM Trans. Knowl. Discov. Data* 6.4 (2012). ISSN: 1556-4681. DOI: 10.1145/2382577.2382579.

[28] Karim Khamaisi, Nicolas Keller, Stefan Krummenacher, Valentin Huber, Bernhard Fässler, and Bruno Rodrigues. *From Noise to Knowledge: A Comparative Study of Acoustic Anomaly Detection Models in Pumped-storage Hydropower Plants*. Sept. 2025. arXiv: 2509.22881 [cs.LG].

[29] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Yuta Kawachi, and Noboru Harada. "Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2019), pp. 212–224. ISSN: 2329-9290. DOI: 10.1109/TASLP.2018.2877258.

[30] Rikard Laxhammar and Göran Falkman. "Inductive Conformal Anomaly Detection for Sequential Detection of Anomalous Sub-Trajectories". In: *Annals of Mathematics and Artificial Intelligence* 74.1-2 (2015), pp. 67–94. DOI: 10.1007/s10472-013-9381-7.

[31] Chuan Li, René-Vinicio Sánchez, Grover Zurita, Mariela Cerrada, and Diego Cabrera. "Fault Diagnosis for Rotating Machinery Using Vibration Measurement Deep Statistical Feature Learning". In: *Sensors* 16.6 (2016). ISSN: 1424-8220. DOI: 10.3390/s16060895.

[32] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[33] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn W. Schuller. "Audio Self-Supervised Learning: A Survey". In: *Patterns* 3.12 (2022), p. 100616. ISSN: 2666-3899. DOI: 10.1016/j.patter.2022.100616.

[34] F. Lo Scudo, E. Ritacco, L. Caroprese, and G. Manco. "Audio-Based Anomaly Detection on Edge Devices via Self-Supervision and Spectral Analysis". In: *Journal of Intelligent Information Systems* 61.3 (2023), pp. 765–793. DOI: 10.1007/s10844-023-00792-2.

[35] Rati Kanta Mohanta, Thanga Raj Chelliah, Srikanth Allamsetty, Aparna Akula, and Ripul Ghosh. "Sources of Vibration and Their Treatment in Hydro Power Stations-A Review". In: *Engineering Science and Technology, an International Journal* 20.2 (2017), pp. 637–648. ISSN: 2215-0986. DOI: 10.1016/j.jestch.2016.11.004.

[36] Robert Müller, Fabian Ritz, Steffen Illium, and Claudia Linnhoff-Popien. "Acoustic Anomaly Detection for Machine Sounds Based on Image Transfer Learning". In: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (Volume 2)*. 2021, pp. 49–56. DOI: 10.5220/0010185800490056.

[37] Neuron Soundware. *Acoustic Machine Diagnostics AI*. 2024.

[38] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. *Description and Discussion on DCASE 2024 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*. June 2024. DOI: 10.48550/arXiv.2406.07250. arXiv: 2406.07250 [eess]. (Visited on 01/29/2026).

[39] Oxmaint. *Hydro Turbine Maintenance: Cavitation Monitoring, Repair and Performance Tracking*. 2024.

[40] Juan I Pérez-Díaz, Manuel Chazarra, Javier García-González, Giovanna Cavazzini, and Anna Stoppato. "Trends and Challenges in the Operation of Pumped-Storage Hydropower Plants". In: *Renewable and Sustainable Energy Reviews* 44 (2015), pp. 767–784. DOI: 10.1016/j.rser.2015.01.029.

[41] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. *MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection*. Sept. 2019. DOI: 10.5281/ZENODO.3384388. (Visited on 09/30/2024).

[42] Ahmad Qurthobi, Rytis Maskeliūnas, and Robertas Damaševičius. "Detection of Mechanical Failures in Industrial Machines Using Overlapping Acoustic Anomalies: A Systematic Literature Review". In: *Sensors* 22.10 (2022), p. 3888. DOI: 10.3390/s22103888.

[43] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure". In: *Ecography* 40.8 (2017), pp. 913–929. DOI: 10.1111/ecog.02881.

[44] Phurich Saengthong, Tomoya Nishida, Kota Dohi, Natsuo Yamashita, and Yohei Kawaguchi. *Retaining Mixture Representations for Domain Generalized Anomalous Sound Detection*. Oct. 2025. DOI: 10.48550/arXiv.2510.25182. arXiv: 2510.25182 [eess]. (Visited on 01/29/2026).

[45] Sojin Shin, Cheolgyu Hyun, Seongpil Cho, and Phill-Seung Lee. "Multi-Sensor Data-Based Anomaly Detection and Diagnosis of a Pumped Storage Hydropower Plant". In: *Structural Engineering and Mechanics* 88.6 (2023), pp. 569–581. DOI: 10.12989/sem.2023.88.6.569.

[46] Siemens. *Senseye Predictive Maintenance Product Data Sheet v1.6*. 2025.

[47] Wade A. Smith and Robert B. Randall. "Rolling Element Bearing Diagnostics Using the Case Western Reserve University Data: A Benchmark Study". In: *Mechanical Systems and Signal Processing* 64-65 (2015), pp. 100–131. ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2015.04.021.

[48] Giancarlo Sperlì and Andrea Vignali. "Anomaly Detection in Cyber-Physical Systems: A Case Study on Pump Health Monitoring". In: *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. 2024, pp. 361–364. DOI: 10.1109/ICASSPW62465.2024.10627724.

[49] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. "Classification of Imbalanced Data: A Review". In: *International journal of pattern recognition and artificial intelligence* 23.04 (2009), pp. 687–719.

[50] Yuting Sun, Tong Chen, Quoc Viet Hung Nguyen, and Hongzhi Yin. *TinyAD: Memory-efficient Anomaly Detection for Time Series Data in Industrial IoT*. Mar. 2023. DOI: 10.48550/arXiv.2303.03611. arXiv: 2303.03611 [cs]. (Visited on 01/29/2026).

[51] Lisa Maria Svendsen, Vignesh V. Shanbhag, and Rune Schlanbusch. "Hydraulic Seal Wear Classification by Fine-Tuning a Transformer-Based Audio Model Using Acoustic Emission". In: *Sensors* 26.9 (2026), p. 2856. ISSN: 1424-8220. DOI: 10.3390/s26092856.

[52] Daniela G. Toshkova, Matthew F. Asher, Paul Hutchinson, and Nicholas A. J. Lieven. "Automatic Alarm Setup Using Extreme Value Theory". In: *Mechanical Systems and Signal Processing* 139 (2020), p. 106417. ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2019.106417.

[53] Aysegul Ucar, Mehmet Karakose, and Necim Kırımça. "Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends". In: *Applied Sciences* 14.2 (2024), p. 898. DOI: 10.3390/app14020898.

[54] Govind Vashishtha, Sumika Chauhan, Mert Sehri, Radoslaw Zimroz, Patrick Dumond, Rajesh Kumar, and Munish Kumar Gupta. "A Roadmap to Fault Diagnosis of Industrial Machines via Machine Learning: A Brief Review". In: *Measurement* 242 (2025), p. 116216. ISSN: 0263-2241. DOI: 10.1016/j.measurement.2024.116216.

[55] illwerke vkw. *Rodundwerk II*. 2025. (Visited on 02/07/2025).

[56] Voith. *OnCare.Acoustic*. 2024.

[57] Voith. *Pilot project on innovative cavitation monitoring in hydropower plants launched in Iceland*. 2020.

[58] Haibo Wan, Xiwen Gu, Shixi Yang, and Yanni Fu. "A Sound and Vibration Fusion Method for Fault Diagnosis of Rolling Bearings under Speed-Varying Conditions". In: *Sensors (Basel, Switzerland)* 23.6 (2023), p. 3130. ISSN: 1424-8220. DOI: 10.3390/s23063130.

[59] Han Wang, Dongdong Wang, Haoxiang Liu, and Gang Tang. "A Predictive Sliding Local Outlier Correction Method with Adaptive State Change Rate Determining for Bearing Remaining Useful Life Estimation". In: *Reliability Engineering and System Safety* 225 (2022), p. 108601. ISSN: 0951-8320. DOI: 10.1016/j.ress.2022.108601.

[60] Wei Wang, Jiaxiang Xu, Xin Li, Kang Tong, Kailun Shi, Xin Mao, Junxue Wang, Yunfeng Zhang, and Yong Liao. "ResNet + Self-Attention-Based Acoustic Fingerprint Fault Diagnosis Algorithm for Hydroelectric Turbine Generators". In: *Processes* 13.8 (2025), p. 2577. ISSN: 2227-9717. DOI: 10.3390/pr13082577.

[61] Kevin Wilkinghoff. "Sub-Cluster AdaCos: Learning Representations for Anomalous Sound Detection". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9534290.

[62] Xiong Xu, Jiazeng Deng, Haijun Lin, Zhongwen Li, and He Wen. "Lightweight Anomalous Detection of Hydro Turbine Operation Sound Using Fusion Network Enhanced by Load Information". In: *IEEE Transactions on Instrumentation and Measurement* 74 (2025), pp. 1–13. ISSN: 1557-9662. DOI: 10.1109/TIM.2025.3533632.

[63] Xiong Xu, He Wen, HaiJun Lin, ZiXin Li, and Cong Huang. "Online Detection Method for Variable Load Conditions and Anomalous Sound of Hydro Turbines Using Correlation Analysis and PCA-adaptive-K-means". In: *Measurement* 224 (2024), p. 113846. ISSN: 0263-2241. DOI: 10.1016/j.measurement.2023.113846.

[64] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. "Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3988–4003.

[65] W. Zhao, M. Egusquiza, M. Valero, D. Valentin, A. Presas, and E. Egusquiza. "On the Use of Artificial Neural Networks for Condition Monitoring of Pump-Turbines with Extended Operation". In: *Measurement* 163 (2020), p. 107952. ISSN: 1873-412X. DOI: 10.1016/j.measurement.2020.107952.

[66] Xinhu Zheng, Anbai Jiang, Bing Han, Yanmin Qian, Pingyi Fan, Jia Liu, and Wei-Qiang Zhang. “Improving Anomalous Sound Detection via Low-Rank Adaptation Fine-Tuning of Pre-Trained Audio Models”. In: *2024 IEEE Spoken Language Technology Workshop (SLT)*. 2024, pp. 969–974. DOI: 10.1109/slt61566.2024.10832335.

Declaration of Authorship

"I hereby declare,

- that I have written this thesis independently,
- that I have written the thesis using only the aids specified in the index;
- that all parts of the thesis produced with the help of aids have been declared;
- that I have handled both input and output responsibly when using AI. I confirm that I have therefore only read in public data or data released with consent and that I have checked, declared and comprehensibly referenced all results and/or other forms of AI assistance in the required form and that I am aware that I am responsible if incorrect content, violations of data protection law, copyright law or scientific misconduct (e.g. plagiarism) have also occurred unintentionally;
- that I have mentioned all sources used and cited them correctly according to established academic citation rules;
- that I have acquired all immaterial rights to any materials I may have used, such as images or graphics, or that these materials were created by me;
- that the topic, the thesis or parts of it have not already been the object of any work or examination of another course, unless this has been expressly agreed with the faculty member in advance and is stated as such in the thesis;
- that I am aware of the legal provisions regarding the publication and dissemination of parts or the entire thesis and that I comply with them accordingly;
- that I am aware that my thesis can be electronically checked for plagiarism and for third-party authorship of human or technical origin and that I hereby grant the University of St.Gallen the copyright according to the Examination Regulations as far as it is necessary for the administrative actions;
- that I am aware that the University will prosecute a violation of this Declaration of Authorship and that disciplinary as well as criminal consequences may result, which may lead to expulsion from the University or to the withdrawal of my title."

By submitting this thesis, I confirm through my conclusive action that I am submitting the Declaration of Authorship, that I have read and understood it, and that it is true.

St. Gallen, 22 September 2026 Signature:

Declaration of Auxiliary Aids

The creation of this thesis involved third-party software, including AI-based tools. I declare all uses of third-party software that contributed non-trivially to the thesis's content and are non-standard in such research. Further, by the University of St. Gallen's decree II.B.1.20 section 5.3, I explicitly declare the following uses:

- Anthropic Claude (used through the Claude Code environment, in the model versions current between 2025 and September 2026) has been used, under my direction, for implementation support in the accompanying software repository (drafting, refactoring and testing of code), for data-analysis tooling, for structuring and organizing content, as well as revising thesis text. All generated code and text were reviewed by me, all quantitative statements were verified against the underlying measurement artifacts, and the scientific decisions, claims and conclusions are my own.
- Google Deepmind Antigravity (powered by Gemini, in the model versions current between 2025 and September 2026) has been used, under my direction, with a specific focus on structural editing and text revision. It assisted in condensing the manuscript, ensuring adherence to scientific writing principles, validating the consistency of the thesis text against the repository results, and orchestrating the final layout of figures and tables. All revisions and structural changes were reviewed and approved by me.
- Google NotebookLM (Gemini Notebook), in its versions current between 2025 and August 2026, has been used for source-grounded literature questions.
- Semantic Scholar has been used for literature search and citation-metadata lookup; Zotero has been used for reference management.