fire

MDPI

Article

Simulation-Based Evaluation of a Hierarchical Decision Rule for Standby Fire-Pump Condition Assessment

Zengyi Yu [1], Yanpeng Wang [1], Jiahao Wu [1], Huaifeng Wang [1], Xinyu Han [1] and YiNuo Huo [2,*]

[1] CNOOC Safety Technology Services Co., Ltd., Tianjin 300456, China; yuzy3@cnooc.com.cn (Z.Y.); wangyp40@cnooc.com.cn (Y.W.); wujh36@cnooc.com.cn (J.W.); wanghf40@cnooc.com.cn (H.W.); hanxy10@cnooc.com.cn (X.H.)
[2] Hefei Institute of Public Safety, Tsinghua University, Hefei 230601, China
* Correspondence: huoyinuo@tsinghua-hf.edu.cn

Abstract

Standby fire-pump monitoring must distinguish supported abnormalities from insufficient measurement evidence. This simulation study evaluates a hierarchical decision rule that combines diagnostic probabilities with measurement-quality checks and compatible normal references. The rule returns normal, abnormal, or indeterminate states and records the domains supporting an abnormal decision. On 1800 fresh synthetic windows, it accepted 1365 with one observed binary error, compared with 1021 accepted windows and no observed errors under strict single-domain corroboration. Hydraulic evidence supported acceptance in 259 of 360 hydraulic-labeled windows (71.94%); the broader acceptance count included cases supported only by other domains. The main constraint was deferral of 240 of 360 normal windows (66.67%). At a hypothetical 5% abnormal prevalence, the expected deferral was 640.1 windows per 1000, representing unresolved windows rather than measured inspection time. The results show how relaxing single-domain corroboration changes acceptance and the reported evidence within the specified generator. Evaluation on synchronized pump measurements and recorded inspection outcomes is needed to determine whether the resulting verification requests are useful in practice.

Keywords: fire risk management; standby fire pump; condition assessment; multimodal monitoring; inspection support; selective classification

1. Introduction

A standby fire pump must contribute to the required water supply when a fire occurs, although deterioration can develop while the equipment is rarely operated. Inspection and testing therefore support the availability of this protective function [1]. Condition monitoring can help identify equipment or measurements requiring attention [2], but an equipment-condition assessment is only one part of that task. Adequate fire-water delivery also depends on the source, distribution network and required demand [3]. This study examines a decision rule for interpreting pump-monitoring evidence within that wider inspection process.

Vibration, hydraulic and electrical measurements observe different aspects of the pump and drive [4–7]. Their responses can be coupled: hydraulic changes can produce torque-related electrical signatures [8], while pressure and noise measurements have been used to investigate blockage and cavitation [9,10]. Simultaneous departures in several domains may therefore warrant a joint check. Conversely, normal vibration does not establish hydraulic delivery when pressure or flow is unavailable. An assessment should

Check for updates

Academic Editor: Zihe Gao

Received: 20 September 2026
Revised: 23 October 2026
Accepted: 30 October 2026
Published: 11 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

preserve these distinctions between observed departures, their possible causes and gaps in the measurements.

Fire-pump performance testing compares hydraulic and power curves with references across operating points [11]. Learned diagnostic methods, including CNN–GRU models for firefighting pumps [12], classify patterns in measured signals. Periodic encoders, multisensor fusion, and missing-modality methods provide alternative signal representations [13–19]. An inspection decision additionally requires a link between the predicted condition and the available measurements. Selective prediction provides a reject option [20], but confidence alone does not establish whether a relevant domain was observed, whether its measurements are usable or whether they depart from a compatible normal reference. The assessment problem considered here is how to combine these requirements while retaining the domains that support an accepted decision.

We propose a hierarchical rule that reports normal (N), abnormal (A), or indeterminate (I) states and separately records the domains supporting abnormality. It combines two changes to strict single-domain corroboration: probabilities are summed across abnormal classes for the condition decision, and all eligible alarm domains are retained instead of requiring a single-domain explanation. Normal acceptance requires clear evidence in all three domains; abnormal acceptance requires sufficient aggregate abnormal-class confidence and at least one eligible alarm. We evaluate how these choices change accepted errors, supported hydraulic evidence, and deferred windows at the original operating points and matched acceptance budgets. The reported domains identify proposed verification tasks for subsequent inspection.

Historical measurements from physical pumps provide a separate comparison of normal-reference use in an earlier assessment system. The proposed hierarchy is evaluated on controlled synthetic sequences. The primary rules and analysis plan were frozen, meaning fixed before the fresh evaluation windows and predictions were examined. Engineering-only assessment isolates the contribution of the diagnostic gate, while matched acceptance budgets control the number of accepted windows. Subsequent controls, threshold sensitivity, and class-mixture analyses are identified as post hoc. The fresh windows test behavior within the same generator and configuration family; the fidelity and usefulness of synthetic data for new measurements require independent evidence [21–23]. Table 1 summarizes the position of the assessment relative to neighboring approaches.

Table 1. Position of the assessment relative to neighboring approaches.

Approach	Primary Output	Additional Question Addressed in This Study
Performance-curve testing [11]	Pressure, flow, power and efficiency relative to reference curves	Which short-window departures should prompt a duty-point or measurement check?
Learned diagnosis [12–17]	A predicted fault class and its probability	Does the observed domain support accepting an abnormal or normal conclusion?
Selective prediction [20]	Acceptance with an error–coverage tradeoff	Can multiple supported alarm domains be retained without forcing a single-domain explanation?
Hierarchical assessment	N/A/I state, supported domain set and verification task	What condition evidence is retained at a given acceptance or review workload?

2. Materials and Methods

2.1. Assessment Scope and Inspection Tasks

The assessment unit is a monitoring window associated with an operating state, acquisition configuration, and reference version. The measured domains are vibration (V),

pressure and flow (H), and electrical trends (E). Each output contains a condition state, eligible domains, an alarm-domain set, evidence values, and a reason code. N means that every assessed domain is eligible and clear, and the diagnostic probability supports normality. A means that at least one eligible domain supports an alarm and the aggregate diagnostic probability supports abnormality. It means that neither acceptance condition is met. These states describe the available evidence for the window; they do not by themselves determine whether the pump can meet its required duty.

Figure 1 places the assessment within the fire-water inspection process. An H alarm calls for checking pressure and flow measurements, valve state, and delivery at the required duty. V and E alarms identify corresponding mechanical and electrical checks. The record also preserves unavailable or unsuitable measurements and disagreement between the diagnostic and reference evidence, allowing reacquisition, reference review or technical review to be requested. These are proposed follow-up tasks; pump-control and return-to-service decisions remain subject to the responsible operator's procedures.

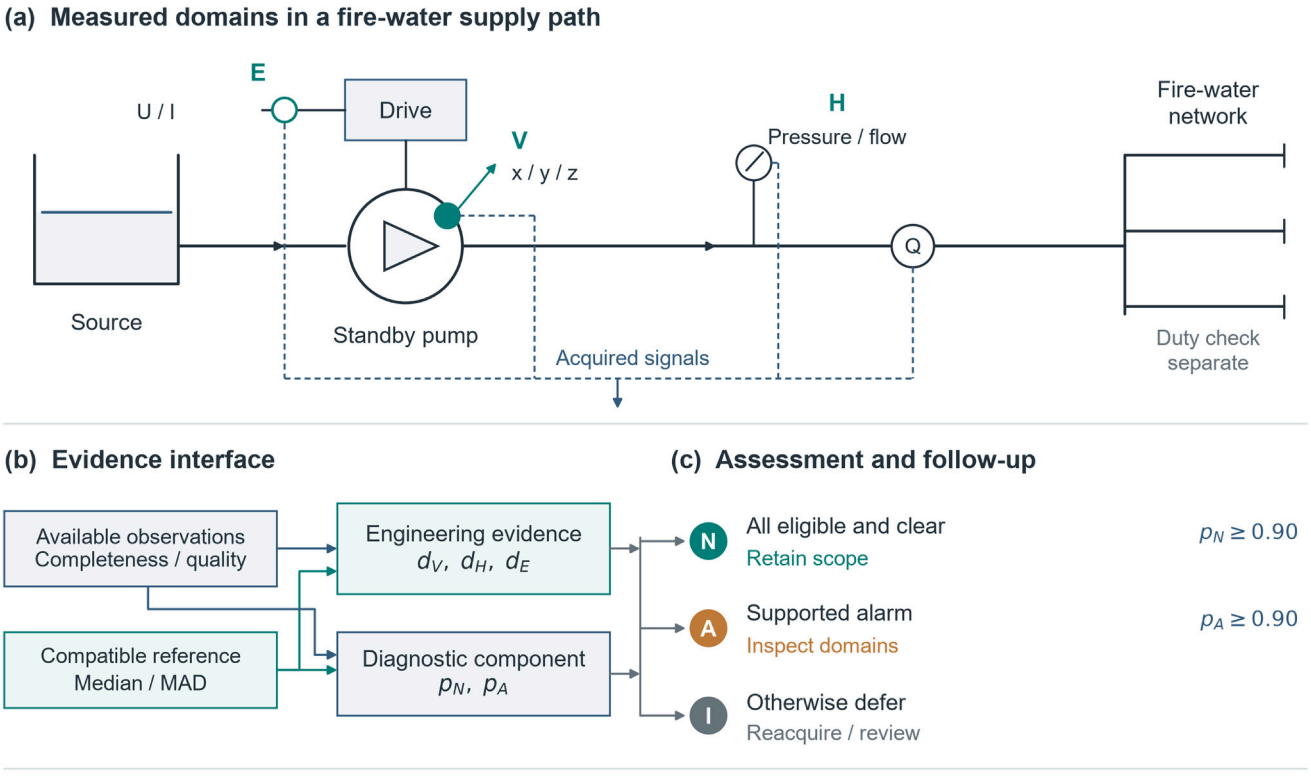

Figure 1. Measured domains and assessment interface for a conceptual fire-water supply path. (**a**) Vibration (V), hydraulic (H), and electrical (E) measurements support pump-condition assessment; required hydraulic duty is checked separately. Dashed lines denote signal transfer, not piping. (**b**) Available observations and a compatible normal reference supply engineering departures and diagnostic probabilities. Blue arrows carry available-observation inputs; teal arrows carry compatible-normal-reference inputs. (**c**) A normal conclusion requires all three domains (V, H, and E) to be eligible and clear with sufficient normal-class confidence; an abnormal conclusion requires at least one eligible alarm with sufficient aggregate abnormal-class confidence. Other records are deferred. Follow-up tasks retain the supported domains and unresolved evidence rather than issuing pump-control commands.

The present V/H/E interface assumes an electrically driven pump with three-phase voltage and current measurements. Extending it to a diesel-driven pump would require a

different definition of driver evidence. A structurally absent electrical domain should not be treated as an intermittent sensor-acquisition failure.

2.2. Compatible References and Observable Quality

Engineering assessment uses 11 summaries calculated from finite observed samples in the original channel units: uncentered RMS for each of the three vibration channels, mean pressure and flow, and RMS over each of the six voltage/current trend channels. The electrical samples are low-frequency RMS trends rather than raw 50 Hz waveforms. Missing positions do not enter these summaries. For each simulated configuration, 32 separate normal windows define the median center and median absolute deviation (MAD) in Equations (1) and (2). These robust summaries limit the influence of extreme reference values [24]; the scale floor prevents division by a very small dispersion.

$$\mu_{p,j} = \text{median}_k v_{p,k,j},\ D_{p,j} = \text{median}_k \left| v_{p,k,j} - \mu_{p,j} \right| \tag{1}$$

$$s_{p,j} = \max\left(1.4826 D_{p,j},\ 0.01\left|\mu_{p,j}\right|,\ 10^{-5}\right),\ B_p^{\text{raw}} = \left[\boldsymbol{\mu}_p; \mathbf{s}_p\right] \tag{2}$$

The 11 centers and 11 scales form a 22-dimensional reference vector. Engineering departures use this vector in original feature units. The diagnostic network instead receives a separately standardized reference and standardized sequences with missing positions filled by zero; its internal summaries therefore differ from the observed-only engineering features (Section S3). Training data alone determine standardization. Equation (3) defines each domain departure as the largest absolute standardized feature departure. Values below 4 are clear, values from 4 to below 6 are warning-level, and values of 6 or above are alarms. These dimensionless boundaries define one research operating point, not absolute equipment limits or specified tail probabilities.

$$z_j = \frac{v_j - \mu_{p,j}}{s_{p,j}},\ d_k = \max\left\{ |z_j| \ :\ j \in J_k \right\} \tag{3}$$

Here p identifies the compatible configuration; j is a feature; J_k is the feature set of domain k ∈ {V, H, E}.

A domain is eligible only if it passes both acquisition and reference checks. It must be available, every value marked observed must be finite, and every channel must contain at least three observations, at least 0.80 completeness, and a peak-to-peak range greater than 10^{-12}. The observable quality in Equation (7) must be at least 0.50. Extracted features and reference centers must be finite, and reference scales must be finite and positive. A calculable feature alone does not make a domain eligible. Compatibility is supplied by the matching simulated configuration and its normal-reference set. The evaluation does not test how an unmatched operating configuration would be recognized or assigned a new reference.

The completeness floor requires at least four-fifths of each channel's scheduled samples. The quality floor additionally screens the combined completeness and stability score; it is not a probability that a measurement is reliable. Section 3.5.1 varies the quality floor and other decision thresholds, while the completeness floor remains fixed at 0.80.

Let $O_{k,c}$ be the finite observed positions in channel c of domain k, and $A_{k,c}$ the adjacent pairs for which both samples are observed. With T_k positions and n_k channels, completeness and normalized roughness are defined as follows. The mean and population standard deviation in Equation (5) use only $O_{k,c}$.

$$C_{k,c} = \frac{|O_{k,c}|}{T_k},\ C_k = \frac{1}{n_k}\sum\nolimits_{c=1}^{n_k} C_{k,c} \tag{4}$$

$$R_{k,c} = \frac{\underset{(t,t+1)\in A_{k,c}}{\text{median}} |x_{k,c,t+1} - x_{k,c,t}|}{\max\{\sigma_{k,c},\ 0.01|\mu_{k,c}|,\ 10^{-6}\}} \tag{5}$$

$$S_{k,c} = \frac{1}{1+R_{k,c}},\ S_k = \frac{1}{n_k}\sum\nolimits_{c=1}^{n_k} S_{k,c} \tag{6}$$

A channel with fewer than three observations or no valid adjacent pair receives zero stability. Equation (7) multiplies the binary availability indicator by the geometric mean of mean completeness and mean stability. Only adjacent observed pairs on the original time grid contribute to roughness; differences are not taken across missing positions. The score uses neither injected SNR nor fault labels.

$$q_k = m_k\sqrt{C_k\, S_k} \tag{7}$$

Let V_k denote the joint acquisition and reference-validity conditions above, excluding the separate quality cutoff. Equation (8) combines availability, validity, and the observable-quality requirement to define eligibility.

$$e_k = m_k\, 1\{V_k\}\, 1\{q_k \geq 0.50\} \tag{8}$$

2.3. Binary Condition and Set-Valued Localization

The diagnostic component returns five probabilities for normal, mechanical, bearing/impact, hydraulic, and electrical classes. The two vibration-related classes map to V, while hydraulic and electrical map to H and E. For the condition decision, the four abnormal probabilities are summed to obtain p_A and p_N is the normal probability. Aggregation allows uncertainty between abnormal subtypes without treating it as equal uncertainty about whether the window is abnormal.

$$p_A = \sum\nolimits_{c\,\in\,F} p_c = 1 - p_N \tag{9}$$

F denotes the four abnormal classes.

Let K = {V, H, E}. The alarm set A(x) contains eligible domains with a reference departure of at least 6. The condition decision is then defined by Equation (10). Invalid diagnostic probabilities yield I before either acceptance condition is tested.

At the primary threshold of 0.90, the two acceptance conditions are mutually exclusive: normal acceptance requires every domain to be clear, whereas abnormal acceptance requires an eligible alarm. The frozen code selects the five-class argmax before using its normal or aggregate abnormal probability, which is equivalent to Equation (10) at 0.90. Section S7 documents the uniform-probability tie at 0.80; no decision disagreement occurred on the saved windows at any reported setting. The confidence gate does not guarantee 90% decision accuracy.

$$A(x) = \{k \in K : e_k = 1,\ d_k \geq 6\}$$
$$\delta(x) = \begin{cases} N,\ p_N \geq 0.90\ \textit{and}\ \forall k:\ e_k = 1,\ d_k < 4 \\ A,\ p_A \geq 0.90\ \textit{and}\ A(x) \neq \emptyset \\ \text{I, otherwise} \end{cases} \tag{10}$$

Table 2 separates condition acceptance from domain reporting. An abnormal window may be accepted when one domain supports an alarm even if another is unavailable; the unavailable domain remains unresolved in the record. A normal conclusion requires all three domains to be eligible and clear, because an unobserved domain cannot supply evidence of normality. Warning-only evidence remains indeterminate, while a warning in one domain does not veto an adequately supported alarm elsewhere. Unlike supervised

multi-label learning [25], the alarm set records threshold-supported measurement domains rather than jointly diagnosed fault causes. Acceptance is therefore evaluated with both retained errors and deferral [26].

Table 2. Hierarchical acceptance rule at the fixed operating point.

Condition	Assessment	Reported Scope and Follow-Up
$p_N \geq 0.90$; all V/H/E eligible and clear	Normal	Retain the assessed scope and reference version.
$p_A \geq 0.90$; at least one eligible alarm	Abnormal	Retain all alarm domains and their verification tasks.
Otherwise	Indeterminate	Preserve the reason and request a targeted check.

Each record retains its acquisition identifier, reference version, eligibility checks, departures, probabilities, state and domain set. Changes in the measurement–condition relationship can undermine a fixed diagnostic model [27]. Reference updates require reviewed normal windows at a compatible operating state and explicit versioning; no automatic drift correction or reference update is simulated.

For hydraulic interpretation, the record also retains pressure and flow in their original units, their signed feature departures from Equation (3), and the percentage difference from a positive reference median. Availability, reference validity, and engineering eligibility remain separate fields. Direction refines the requested check but does not alter Equation (10): a small decrease can remain within the clear range, while a calculable departure from inadequate data need not support acceptance. Missing or invalid descriptors remain undefined.

2.4. Diagnostic Probability Interface

The diagnostic component is a baseline-conditioned, quality-aware periodic convolutional network (PBQ-PCN). Three separate encoders map vibration, hydraulic and electrical sequences of unequal lengths to 48-dimensional representations. The encoders combine dilated convolutions [28] and depthwise separable convolutions [29] eature-wise linear modulation [30] conditions the representations on normal references, and four-head self-attention [31] combines the modality tokens. Availability masks identify missing modalities, a distinct problem from low-quality measurements [32]; observable quality is retained as a separate input before five-class probabilities are supplied to the acceptance rule. The implementation registers 218,235 parameters. Figure S1 and Section S3 in the Supplementary Material give the complete architecture, equations and training settings.

The matched models use cross-entropy training and validation-based early stopping. PBQ-observed uses the measured quality input; PBQ-constant sets this input to one for available domains; Concat-observed retains observable quality but bypasses cross-modal attention. Their input information and training settings are matched, as specified in Section S1. PBQ-observed is primary because its checkpoints and normalization were fixed before the fresh rule evaluation. The controls in Sections 3.4 and 3.5.4 test whether the more complex component is needed. Reconstruction and consistency losses belong only to the historical benchmark in Section S6.

2.5. Evaluation Measures

Section S5 defines the classification and calibration measures. Calibration concerns agreement between probabilities and outcome frequencies [33]; it is distinct from the measurement sufficiency required for acceptance. Historical engineering indicators remain in Section S4 and are not additional inputs to the current 11-summary assessment.

Assessment maps the normal source class to N and all four abnormal classes to A; confusion between abnormal subtypes is not a binary condition error. Coverage is the fraction of all windows assigned N or A. Selective error divides false-normal and false-abnormal outputs by the accepted count. A false-normal output accepts a truly abnormal window as N; a false-abnormal output accepts a truly normal window as A. Indeterminate outputs remain in the coverage denominator but are excluded from the selective-error denominator and are not counted as corrected.

Deferral is also reported separately for the two true condition classes: abnormal-window deferral is the number of truly abnormal windows assigned I divided by all truly abnormal windows; normal-window deferral uses the corresponding normal-class counts. A false-normal output can provide false reassurance, whereas a deferred abnormal window retains unresolved evidence. Deferral avoids accepting that window as normal but does not establish that the underlying condition has been verified or resolved. A false-abnormal output or deferral of a normal window can lead to additional checks, including checks of measurement quality. These outcomes provide decision-related risk proxies; they are not failure-on-demand probabilities or direct estimates of fire risk.

$$\text{Coverage} = \frac{n_{acc}}{n}, \text{ selective error} = \frac{n_{err}}{n_{acc}} \tag{11}$$

In Equation (11), n is the total number of windows, n_{acc} is the number accepted and n_{err} is the number of accepted binary errors.

Localization is evaluated separately from the binary condition. Domain inclusion asks whether the injected source domain appears in the alarm set of a truly abnormal window accepted as A. Set size reports how narrowly the evidence is localized. Supported-H coverage uses all hydraulic-labeled windows as its denominator and counts only those accepted as A with H in their alarm set. It therefore differs from both overall hydraulic-class acceptance and localization among accepted cases. Deferred-window counts provide a review-workload proxy; review duration, cost, and outcomes were not measured.

For localization, let I_+ comprise truly abnormal windows accepted as abnormal, g_i their injected source domain, and A_i the reported alarm set. Inclusion and exact-singleton rates have the same denominator, $|I_+|$:

$$R_{inc} = \frac{\sum_{i \in I_+} 1\{g_i \in A_i\}}{|I_+|}, \; R_{exact} = \frac{\sum_{i \in I_+} 1\{A_i = \{g_i\}\}}{|I_+|} \tag{12}$$

Mean set size is the average number of reported domains over the same windows. These measures evaluate agreement with the injected source domain, not independent confirmation of every additional affected component. Empty denominators are reported as undefined.

2.6. Historical Records and Simulated Sequences

The historical archive contains 184 measurement records from seven physical pumps: 154 normal and 30 abnormal. A fixed 91-record subset compares correct, incorrect, and uncertain outcomes with and without normal-reference information in the earlier assessment system. These historical measurements, the synthetic windows used to evaluate the proposed hierarchy, and the separate single-channel laboratory study in Section S2 are distinct sources of evidence. The separate laboratory study uses the dataset described by Bruinsma et al. [34] and its NLN-EMP version 4 release [35]; its surrogate generation uses iterative amplitude and Fourier projections [36]. The materials available for this revision comprise aggregate historical outcomes and engineering reports, but not a record-level dataset linking synchronized vibration, hydraulic and electrical inputs, acquisition masks

and quality information, and configuration-compatible normal references. The new hierarchy could therefore not be reconstructed on the historical archive from these materials. A record-by-record inventory would be needed to determine the availability of each required input; no count of incomplete original records is inferred from the materials supplied for revision.

The five-class mechanism-based generator uses seven recurring configuration profiles. Each nominal 6 s window contains 3 × 256 vibration points, 2 × 64 hydraulic points and 6 × 64 electrical trend values. Fault scenarios introduce harmonic or impulse patterns, hydraulic changes, and electrical imbalance; missing points, missing domains, noise, and drift alter acquisition. Vibration is represented on a compressed frequency scale with a nominal sampling rate of 42.67 Hz. The scenario definitions and split-specific degradation settings are given in Tables S2 and S3.

The mechanisms are motivated by reported response types. Experiments on a centrifugal fire pump associate shaft-seat looseness with rotational harmonics and impacts [4]; bearing-fault models describe impulses exciting damped structural responses [5]. Pressure fluctuations, cavitation-related vibration and noise, and hydraulic torque/current oscillations motivate the hydraulic and cross-domain terms [8–10]. The generator encodes these signatures through prescribed perturbations rather than solving coupled pump–fluid–motor dynamics. Its amplitudes, frequency compression, and coupling coefficients are benchmark choices, not parameters fitted to the historical records. Table S2 and the accompanying text specify the dominant and secondary terms. In particular, a hydraulic-labeled window can trigger a vibration or electrical alarm without meeting the H eligibility and alarm requirements, which is why supported-H coverage is reported separately.

Figure 2 separates the historical comparison from synthetic development and evaluation. Table 3 summarizes the evidence sources and their distinct evaluation roles. Each of the three paired model runs uses 600 training and 200 validation windows, with 32 additional normal-reference windows per configuration excluded from those counts. The trained checkpoints and normalizers are retained. Fresh evaluation uses 600 new windows per run with seeds 292026, 292027, and 292028 generated after the primary rules and analysis plan were frozen. The pooled 1800-window holdout contains 360 normal windows (20%) and 1440 abnormal windows (80%), with 360 in each abnormal class. Training, validation, and evaluation use separate random realizations within the same seven-profile configuration family, rather than a split that holds out previously unseen pump configurations.

Table 3. Evidence sources and their distinct evaluation roles.

Evidence Set	Records and Separation	Role in the Paper
Historical monitoring comparison	91 records drawn from the 184-record archive	Earlier system with versus without normal-reference information.
Model development	600 training + 200 validation per run; 32 normal-reference windows per configuration	Training, normalization and checkpoint selection; three paired runs.
Earlier computational tests	Separate 3 × 600-window holdouts	Fusion controls and identification of the restrictive single-domain rule.
Fresh hierarchy evaluation	3 × 600 new windows; seeds 292026–292028	Frozen decision rules, matched coverage and domain-set analysis.

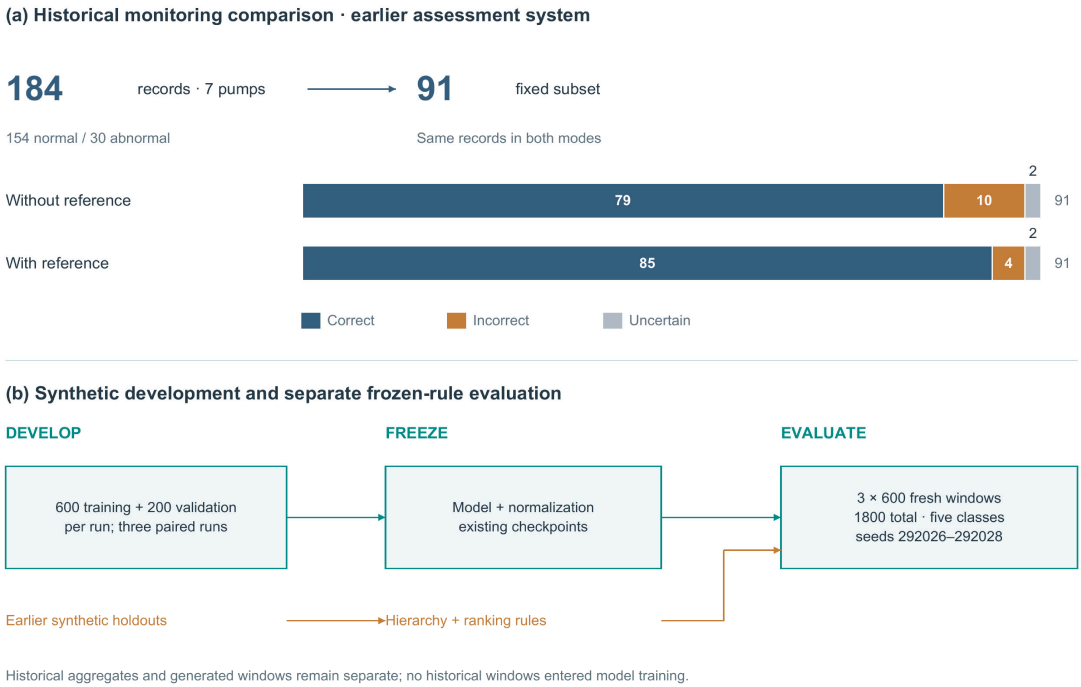

Figure 2. Historical measurement evidence and the frozen simulation evaluation. (**a**) The earlier assessment used a fixed 91-record subset of 184 measurement records from seven physical pumps. Normal-reference use changed the correct/incorrect/uncertain counts from 79/10/2 to 85/4/2 on the same records. (**b**) Computational development used 600 training and 200 validation windows per paired run. Existing model checkpoints and normalization were retained, and the hierarchy and ranking rules were fixed before assessment on three fresh 600-window synthetic holdouts. Earlier and fresh holdouts are not pooled. No raw historical windows entered computational training.

2.7. Frozen Rules and Comparators

For an abnormal candidate, the strict joint rule requires a confident five-class label, an alarm in the corresponding domain, and no alarm in another eligible domain. Its low hydraulic acceptance on the earlier computational holdouts motivated the hierarchy. Those development holdouts are kept separate from the fresh evaluation, for which the rules, comparison variants, and ranking definitions were fixed in advance.

Engineering-only assessment accepts A for any eligible alarm and N only when all three domains are eligible and clear. Five-class confidence-only assessment accepts the argmax class, mapped to N/A, when its probability reaches 0.90. Strict joint retains this confidence requirement: a normal candidate additionally requires every domain to be eligible and clear, whereas an abnormal candidate requires an alarm in its mapped domain and none in another eligible domain. Hierarchical assessment follows Equation (10). The direct-classifier reference maps the five-class argmax to N/A without confidence or engineering gates.

Two changes distinguish the hierarchy from strict joint assessment: summing abnormal probabilities and replacing single-domain corroboration with a supported alarm set. A crossed comparison varies these factors separately; an additional bridge removes only the veto on other alarm domains. The bridge retains candidate-domain corroboration, so its result is not conflated with the broader change in the hierarchical rule.

2.8. Matched Acceptance and Reproducibility

Matched acceptance compares methods at the same accepted count within each run. Native ranking orders each method's naturally accepted windows by its own decision

margin: engineering departure, diagnostic confidence, or the smaller normalized margin for a joint rule. Hash ranking orders record identifiers deterministically and provides a comparison independent of these margins; the same hash also breaks native-ranking ties. Neither ranking uses true class labels to select windows. Section S1 specifies the scores and ordering procedure.

Fixed target coverages of 10%, 20%, …, 90% are applied to each 600-window run. A target is unreachable if a method cannot supply that many naturally accepted windows; an indeterminate output is never converted to acceptance to fill a budget. The common-capacity comparison uses the smallest natural accepted count among the four primary rules within each run, under both native and hash ranking. The additional controls and component comparisons retain these original per-run budgets. Equal counts can therefore contain different records and different class mixtures.

Pooled counts describe the 1800 controlled windows; means and sample standard deviations describe three paired generation-and-training runs. They are not uncertainty intervals for a population of field pumps. Record identifiers, predictions, eligibility reasons, and protocol hashes are retained for reconciliation. Boundary tests check implementation consistency separately from empirical performance.

2.9. Computational Checks and Reporting

The primary evaluation comprises the frozen rule comparison, matched-acceptance rankings, and domain-set analysis. Additional controls, threshold sensitivity, illustrative cases, class-mixture scenarios and fixed-rule comparisons of diagnostic components were performed afterward using the saved holdout. These analyses are grouped in Section 3.5 and specified in Section S9. Class-mixture scenarios reweight normal and abnormal outcomes while retaining the equal mixture of four fault classes, following a prior-probability-shift assumption [37]. Conditional uncertainty for natural decisions uses 5000 paired bootstrap replicates [38] stratified by model run, class, configuration, and severity, with models and references held fixed. Ranked subsets retain observed counts without bootstrap intervals.

OpenAI Codex assisted with English editing, manuscript and figure preparation, and computational analysis. During revision, Codex desktop version 26.1002.7124.0 (OpenAI; used on 16 October 2026) assisted with static source inspection and the preparation of read-only consistency-check scripts. Using Python 3.12.14 and NumPy 2.3.5, these scripts checked the public archive hashes, recomputed engineering evidence and eligibility from 1800 saved synthetic windows, reconstructed the original rule and component decisions, and reconciled saved threshold, class-mixture, and laboratory prediction summaries. The archived 5000 paired bootstrap count arrays were reproduced using the original strata and seed. These checks reproduced saved calculations without retraining models, running model inference, generating new observations, or selecting thresholds. Section S10 describes the checked files and the limits of this verification.

The original release separately records eight boundary tests preceding protocol freezing, reconciliation of saved predictions and rankings, and reproduction of PBQ-observed outputs before alternate-component inference. These historical records were inspected; the boundary-test suite and checkpoint inference were not rerun during revision. The checks concern computational consistency and reporting, not physical fidelity or field performance.

3. Results

3.1. Historical Reference Comparison

On the same 91 historical measurement records, normal-reference use increased correct outcomes from 79 to 85 and reduced incorrect outcomes from 10 to 4; uncertain outcomes remained at two (Table 4). Accuracy over all 91 records increased from 86.81% to 93.41%, while accuracy over the 89 conclusive outcomes increased from 88.76% to 95.51%. Coverage remained 97.80%. The improvement in this earlier system therefore reflects changed conclusive decisions rather than additional deferral.

Table 4. Earlier system outcomes on the fixed historical measurement comparison.

Mode	Correct/91	Incorrect/91	Uncertain/91	Coverage/%
Without normal reference	79	10	2	97.80
With normal reference	85	4	2	97.80

3.2. Accepted Errors and Deferred Windows

The hierarchy accepted 1365/1800 windows (75.83%) with one false-abnormal decision and no observed false-normal decisions (Table 5). Compared with strict joint assessment, it retained 344 additional decisions for one additional error, reducing deferrals from 779 to 435. Engineering-only and confidence-only accepted more records but retained 13 and 34 errors, respectively. The hierarchy's selective binary error was 0.07%.

At the common accepted total of 1021, native ranking retained 1, 5, 0, and 0 errors for engineering-only, confidence-only, strict joint, and hierarchical assessment, respectively; hash ranking retained 10, 24, 0, and 0 (Figure 3). On the common 10–50% coverage grid under native ranking, engineering-only and both joint rules had no observed errors (Table S4). The larger engineering-only and confidence-only error counts under hash ranking show that the selected subsets depend materially on the ordering rule. These comparisons describe the retained records, whose class allocations are examined in Section 3.3.

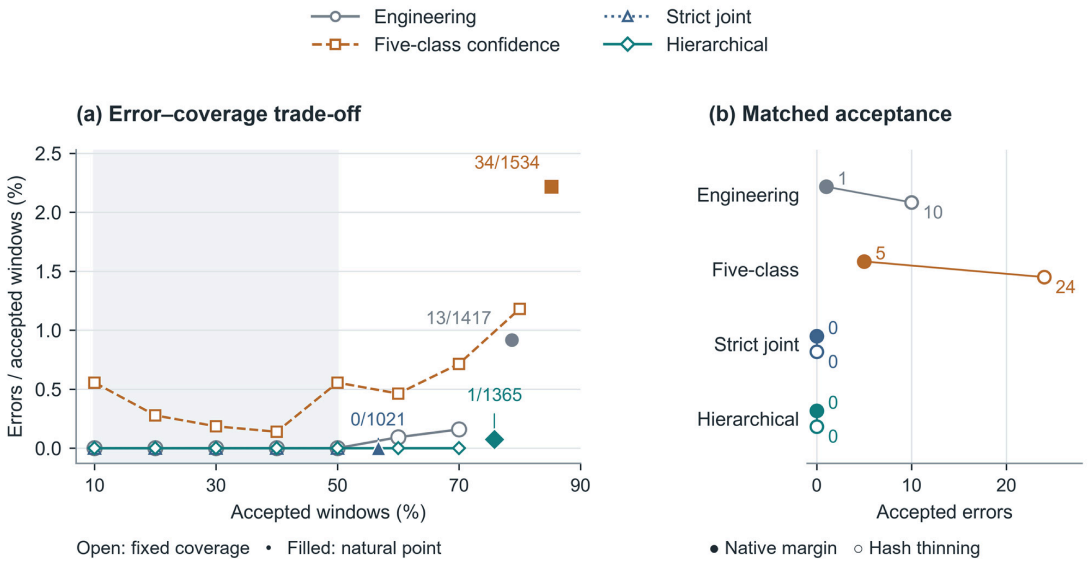

Figure 3. Acceptance and binary errors on the fixed synthetic holdout (1800 windows). (**a**) Open markers show attainable fixed-coverage targets; filled markers show natural operating points, labeled errors/accepted windows. The shaded range is attainable by all four rules. (**b**) Error counts at the common budget of 1021 accepted windows (345, 342, and 334 across runs), using native-margin ranking or hash-only thinning. Matched counts do not imply identical records. Zero is an observed error count, not a guaranteed rate.

Table 5. Fixed operating points on the fresh synthetic holdout.

Method	Accepted/1800	Errors/Accepted	FN/1440	FP/360	Deferred
Direct classifier	1800 (100.00%)	132/1800 (7.33%)	56	76	0
Confidence only	1534 (85.22%)	34/1534 (2.22%)	14	20	266
Engineering only	1417 (78.72%)	13/1417 (0.92%)	1	12	383
Strict joint	1021 (56.72%)	0/1021 (0.00%)	0	0	779
Hierarchical	1365 (75.83%)	1/1365 (0.07%)	0	1	435

N = 1800, including 360 normal and 1440 abnormal windows. FN is abnormal accepted as normal; FP is normal accepted as abnormal. Indeterminate windows are deferred, not counted as correct. Errors/accepted is the binary selective error denominator.

3.3. Supported Hydraulic Evidence and Domain Sets

The primary hydraulic measure is supported-H coverage: 259 of 360 hydraulic-labeled windows (71.94%) were accepted as abnormal with H in the reported alarm set (Figure 4). The hierarchy accepted 323 hydraulic-labeled windows in total (89.72%), but 64 were supported only by other domains. Strict joint assessment accepted 23/360. The broader acceptance count therefore measures recognition of abnormality in the hydraulic-labeled class, not confirmation of a hydraulic departure in every accepted case.



Figure 4. Acceptance and alarm-domain scope on 1800 synthetic windows. (**a**) Natural acceptance by source class (360 windows per class); labels give accepted counts, including incorrect N or A decisions. All five strips use the same percentage scale. (**b**) Hydraulic-source windows partitioned into accepted A with H, accepted without H, and deferred; strict joint assessment has zero accepted-without-H windows. (**c**) Counts of alarm sets above a membership matrix for the 1245 true-abnormal windows accepted as A by the hierarchy; the false-A window is excluded. Dark dots indicate included vibration (V), hydraulic (H), and electrical (E) domains. The injected source is included in 1178 sets; 915 are exact source-domain singletons and the mean set size is 1519/1245 = 1.22. Acceptance and reported alarm sets are not exact causal fault localization.

Among 1245 truly abnormal windows accepted as A, the alarm set included the injected source domain in 1178 cases (94.62%); 915 (73.49%) were exact singleton matches, and the mean set size was 1.22. In the remaining 67 accepted abnormal windows, the alarm set did not include the injected source domain. Binary acceptance therefore did not always identify the source domain, and additional alarm domains indicate measured departures rather than independently confirmed faults.

At the native-ranked matched budget of 1021 accepted windows, the hierarchy retained 236 hydraulic-labeled and 52 true-normal windows, versus 23 and 119 for the strict rule. Equal totals therefore concealed a shift toward accepting abnormal windows rather than an improvement in every class. Outcomes of deferred windows remain unresolved; no technician correction is assumed.

3.4. Rule Decomposition and Model Controls

Relaxing single-domain corroboration accounted for most of the increase in acceptance (Table 6). Binary aggregation with the strict domain condition changed the accepted count from 1021 to 1030. Removing only the veto on other alarm domains retained 1278 windows, including 250 hydraulic-labeled cases. Retaining maximum-class confidence but accepting any eligible alarm increased these counts to 1340 and 311, with no observed binary errors. Adding binary aggregation to the alarm-set rule yielded the hierarchy's 1365 accepted windows, 323 hydraulic-labeled cases, and one error. Thus, the two rule changes had different effects: the domain rule drove most of the acceptance increase, while binary aggregation admitted additional decisions with an observed error cost.

Table 6. Rule variants on the same fresh holdout.

Rule Variant	Accepted/1800	Binary Errors	Hydraulic Accepted/360
Strict joint	1021	0	23
Binary confidence + strict domain	1030	0	25
Remove other-domain veto only	1278	0	250
Maximum-class confidence + alarm set	1340	0	311
Binary confidence + alarm set	1365	1	323

On the separate model-control holdouts, macro-F1 was 89.59 ± 2.95% for PBQ-observed, 90.04 ± 2.25% for PBQ-constant and 89.71 ± 2.43% for Concat-observed (Table S1 in Section S1). Concat-observed also had the lowest mean Brier score and calibration error. The added attention and quality-weighting paths therefore showed no consistent advantage across these diagnostic measures.

3.5. Post Hoc Analyses

The following analyses reuse the saved holdout after the primary evaluation. They examine interpretation and sensitivity without changing the primary thresholds, model selection, or acceptance budgets. Complete methods are provided in Section S9, and the full sensitivity grid is retained in Section S7.

3.5.1. Class Conditional Decisions and Threshold Sensitivity

Binary-confidence filtering retained 1603 records with 51 errors; adding a complete-acquisition gate reduced these to 1269 and 24 (Table 7). At matched acceptance of 1021, their native-ranked error counts were two and six, versus zero for the hierarchy; hash-only counts were 33, 18, and zero. Ranking affected both error and class composition.

Table 7. Class-conditional decisions and fixed-budget errors on the fresh synthetic holdout.

Rule	True Normal N/A/I	True Abnormal N/A/I	Normal Accepted %	Abnormal Accepted %	Matched Errors Native/Hash
Engineering only	141/12/207	1/1263/176	42.50	87.78	1/10
Five-class confidence	223/20/117	14/1277/149	67.50	89.65	5/24
Strict joint	119/0/241	0/902/538	33.06	62.64	0/0
Hierarchical	119/1/240	0/1245/195	33.33	86.46	0/0
Binary confidence	223/37/100	14/1329/97	72.22	93.26	2/33
Binary + acquisition gate	202/16/142	8/1043/389	60.56	72.99	6/18

N, A, and I denote normal, abnormal, and indeterminate outputs. True-class denominators are 360 normal and 1440 abnormal windows. Accepted means N or A, including errors; I is deferred. The last column retains the original matched total of 1021 (345, 342, and 334 by run). Its errors refer to different ranked subsets, not the natural class tables. All added class-conditional analyses are post hoc.

The hierarchy accepted 33.33% of normal and 86.46% of abnormal windows. Of the 360 normal windows, 119 were correctly accepted as N, one was incorrectly accepted as A, and 240 (66.67%) were deferred. Of the 1440 abnormal windows, 1245 were accepted as A, none were accepted as N, and 195 (13.54%) were deferred. The strict rule retained the same 119 correct normal decisions but only 902 abnormal decisions, so its lower overall acceptance mainly reflected abnormal-case deferral. Native matched subsets also differed in composition, retaining 52 true-normal windows under the hierarchy and 119 under the strict rule (Table S8).

Across runs, the hierarchy accepted 460, 451, and 454 of 600 windows, with zero, zero, and one observed error, respectively; binary-confidence filtering retained 16, 12, and 23 errors (Table S5; Figure S2a,b). Of the 435 hierarchical deferrals, the implementation assigned 196 to insufficient binary confidence, 118 to incomplete clear evidence, 120 to lack of a supported alarm, and one to absence of a usable domain (Table S7). These are exclusive decision-reason categories, not mutually exclusive physical causes. Section S7 separately reports overlapping acquisition defects.

The alarm boundary had the largest effect among the tested decision-threshold changes (Figure S2c,d; Table S6). With the other primary settings fixed, alarm boundaries of 5, 6, and 7 accepted 1412, 1365, and 1316 windows, respectively, with 11, 1, and 0 observed errors and 388, 435, and 484 deferred windows. All errors at these settings were false-abnormal decisions. An alarm boundary of 8 further reduced acceptance to 1255 with no observed errors. At alarm boundary 6, varying confidence from 0.80 to 0.975 changed acceptance from 1372 to 1320 and the error count from two to one. These post hoc comparisons retain the original operating point; Section S7 reports the full tested grid.

Separately, raising the clear boundary from 4 to 5 introduced one false-normal decision. At the tested quality floors of 0.40, 0.50 and 0.60, the hierarchy accepted 1365 windows with one observed error and deferred 435 windows. A floor of 0.70 deferred two additional windows without changing the error count (Table S6). This tests the quality cutoff in the acceptance gate, whereas the model controls test how quality enters the diagnostic network. The two comparisons address different uses of quality information.

3.5.2. Illustrative Inspection Tasks

The saved cases link signed departures to inspection tasks (Figure 5). Pressure and flow were 1.77% and 3.18% below reference in the normal case, but all domain departures remained below 4. In the hydraulic case, the respective differences were −6.54% and −8.93%; vibration and hydraulic departures of 28.87 and 8.93 produced {V, H}, prompting a joint mechanical and hydraulic check.

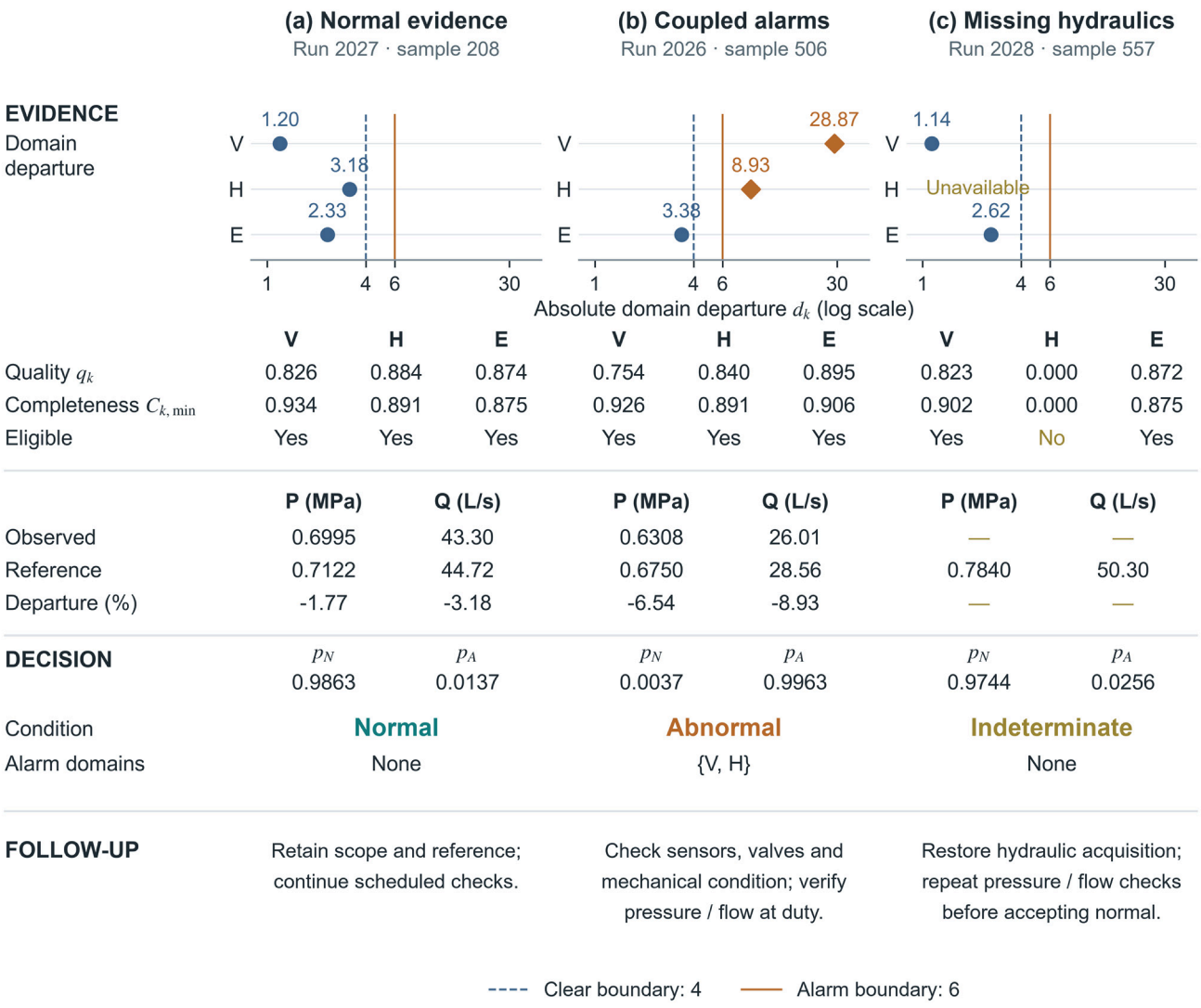

Figure 5. Aligned evidence, decisions and proposed checks for the same three synthetic windows. V, H and E denote vibration, hydraulic and electrical evidence. Top plots show absolute standardized domain departures on shared logarithmic axes, with the unchanged clear and alarm boundaries. Pressure P and flow Q are shown alongside their compatible normal-reference medians; signed departures are percentages of those medians, not required-duty margins. In (**c**), the reference remains valid but hydraulic measurements are unavailable; dashes mean undefined, not zero. Original N/A/I decisions and alarm sets are unchanged. Follow-up items are proposed verification actions, not observed repairs or measured operational benefits. Blue circles indicate observed domain departures below the clear boundary, and orange diamonds indicate departures at or above the alarm boundary. Ochre labels and dashes identify unavailable hydraulic evidence. Condition labels are teal for Normal, orange for Abnormal and ochre for Indeterminate.

The third case had a normal probability of 0.974 but no observed pressure or flow. Its reference remained valid; hydraulic reacquisition was required. The plotted percentages are reference differences, not required-duty shortfalls. In the two observed cases, reference scales happened to be approximately 1% of their medians, making standardized and percentage departures numerically similar but not interchangeable.

3.5.3. Class Mixture and Conditional Uncertainty

At a hypothetical 5% abnormal prevalence, the hierarchy would accept 359.9 and defer 640.1 windows per 1000, compared with 345.3 and 654.7 for the strict rule (Figure 6). It would retain 2.64 false-abnormal decisions per 1000 total windows. The acceptance dif-

ference falls from 19.11 percentage points in the 80%-abnormal test design to 1.45 points at 5%, with differences calculated before rounding the displayed counts. This reduction follows from the class-conditional results: the hierarchy and strict rule accept similar proportions of normal windows, while their largest difference concerns abnormal windows. Reweighting preserves those conditional outcomes and changes only the assumed class mixture.

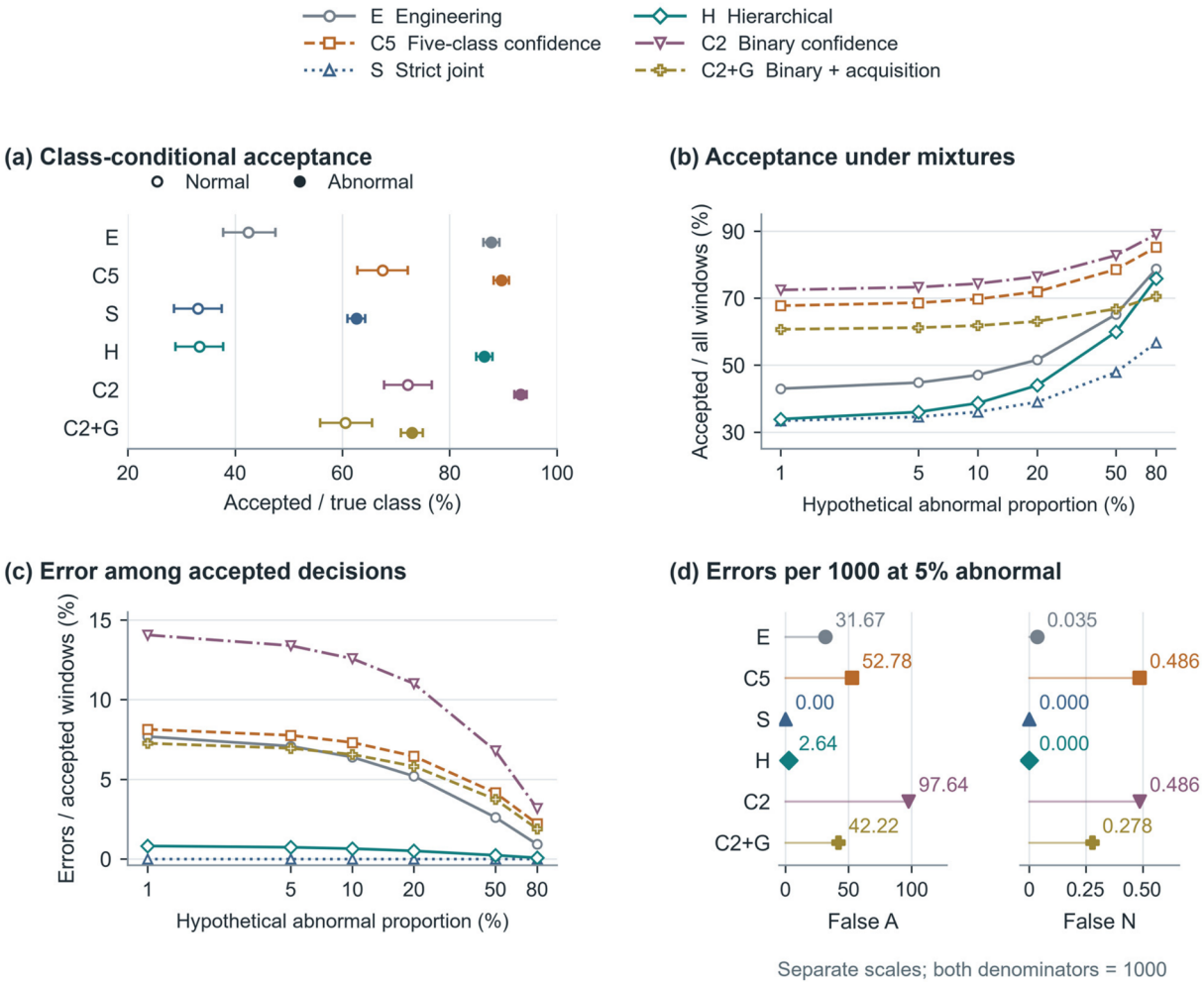

Figure 6. Class-conditional acceptance and hypothetical population mixtures for six unchanged rules. (**a**) Acceptance within true Normal and Abnormal classes with pointwise conditional 95% stratified-bootstrap percentile intervals (5000 paired replicates); these intervals condition on the saved models and synthetic design. (**b**,**c**) Overall acceptance and errors among accepted decisions after reweighting to 1%, 5%, 10%, 20%, 50%, and 80% abnormal proportions. Both proportion axes are logarithmic; within-class profiles and the equal four-fault mixture remain fixed. (**d**) False-A and false-N decisions per 1000 total windows at the hypothetical 5% proportion, shown on separate horizontal scales. A false A is a true-Normal window accepted as Abnormal; a false N is a true-Abnormal window accepted as Normal. These are scenario weights, not measured field prevalence, and observed zeros do not establish zero future risk.

Across the six scenarios and six rules (Figure 6), normal-case deferral increases the expected number of deferred windows as normal windows dominate. Low selective error therefore does not, by itself, imply a small number of review requests; converting windows into actual inspection tasks requires a separate operational evaluation.

The hierarchy's conditional 95% bootstrap intervals were 28.89–37.78% for normal acceptance and 84.93–87.99% for abnormal acceptance. Its acceptance difference from

engineering-only was −2.89 percentage points (interval −3.61 to −2.17), alongside fewer errors (Table S9 in Section S8). For the separate event of a truly abnormal window being accepted as N, zero events among all 1440 abnormal windows gave a one-sided 95% upper bound of 0.208% under the independent synthetic-draw assumption. That denominator includes deferred abnormal windows; the bound does not estimate the probability that an accepted N decision is wrong or a field missed-fault rate.

3.5.4. Fixed Rule Comparison Across Diagnostic Components

The archived PBQ-observed replay reproduced the saved probabilities, decisions, domain sets, and rankings before the alternate components were evaluated. On the same 1800 windows, the unchanged hierarchy accepted 1369 with PBQ-constant and 1362 with Concat-observed, compared with 1365 for PBQ-observed (Table 8). Each retained one false-abnormal decision and no observed false-normal decisions. All three had no observed errors at the original matched budget of 1021 under either ranking.

Table 8. The unchanged hierarchy applied to three existing diagnostic components.

Diagnostic Component	Accepted /1800	FN/FP	Hydraulic Accepted/360	Supported H /360	Matched Errors Native/Hash
PBQ-observed	1365	0/1	323	259	0/0
PBQ-constant	1369	0/1	324	260	0/0
Concat-observed	1362	0/1	324	261	0/0

All values use the same 1800 synthetic windows and original per-run budgets. Supported H counts hydraulic-labeled windows accepted as abnormal with H in their alarm set; it is not the count of all H alarms. These post hoc inferences use frozen checkpoints and normalizers, with no retraining.

Hydraulic acceptance ranged from 323 to 324, with H evidence retained in 259–261 cases. The similar aggregate behavior does not establish formal model equivalence or a consistent attention/quality advantage. PBQ-observed remains primary under the frozen protocol.

4. Discussion

4.1. Interpretation of the Simulated Results

The rule decomposition identifies how the hierarchy increases acceptance. Under the strict rule, an alarm in another eligible domain vetoes acceptance even when the candidate domain corroborates abnormality. Retaining concurrent alarms removes this restriction; allowing any eligible alarm further separates binary condition from a single predicted subtype. Binary aggregation alone changes acceptance much less. The principal finding is therefore a change in how measurement evidence is admitted and reported, rather than a demonstrated improvement in the diagnostic network.

Hydraulic-class acceptance and hydraulic support capture different consequences of that change. Of the 323 accepted hydraulic-labeled windows, 259 included H in the alarm set, and 64 were supported only by other domains. The latter count partly reflects the generator's imposed cross-domain responses. Even a set containing the injected source domain does not uniquely identify a physical cause: vibration RMS cannot distinguish every mechanical disturbance, and concurrent alarms may reflect coupled responses or measurement problems. Supported-H coverage, source-domain inclusion, and set size therefore need to be read together. Their physical interpretation depends on how well the imposed coupling represents measurements from operating pumps.

The diagnostic-component comparison likewise places the contribution at the decision-rule level. With the hierarchy unchanged, PBQ-observed, PBQ-constant, and Concat-observed accepted 1365, 1369, and 1362 windows, respectively, and each retained one false-abnormal decision. The additional quality and attention paths showed no consis-

tent advantage. PBQ-observed supplies the primary probabilities because its checkpoints and normalization were fixed before examination of the fresh holdout; retaining it avoids model reselection based on those outcomes. This study-design choice does not establish PBQ-observed as the preferred deployment architecture. Deployment selection would require comparison of robustness, computational requirements and maintenance needs, together with performance on independent measurements from physical fire pumps. Simpler components remain candidates for that comparison.

4.2. Proposed Integration with Inspection, Testing, and Maintenance

The proposed monitoring record would supplement an established inspection, testing and maintenance process. NFPA 20 addresses installation of stationary fire pumps, while NFPA 25 addresses inspection, testing and maintenance of water-based fire-protection systems [39,40]. For an N decision, the record would retain the condition evidence supported by the current measurements without replacing required inspections or tests or closing an unresolved equipment issue [11]. An A decision would prompt targeted verification of the supporting domains and corrective maintenance where findings warrant it: hydraulic delivery and valve configuration for H alarms, and mechanical or electrical checks for V or E alarms. An I decision would prompt reacquisition or technical review and remain unresolved until the evidence gap had been addressed. Each record would link the operating state, reference version, and decision evidence to the requested checks, verification findings, and subsequent actions. Reacquired measurements would undergo the same validity checks, and a revised normal reference would receive a new version identifier. Authorized personnel would determine impairment of the protective function, corrective measures, and return to service under the applicable procedures.

The connection to protective-function availability depends on what happens after a request. Periodic-inspection analyses consider fire-pump probability of failure on demand [1], while a Fire Protection Research Foundation study models unavailability as time during which a pump cannot operate on demand and specifies assumptions about detection and repair [41]. A benefit of the proposed rule would be that a defect is identified, confirmed, and corrected sooner. Evaluation would need detection, confirmation and restoration times, monitoring-related interruptions, pump redundancy and water-supply conditions. Return-to-service decisions would remain with the responsible operator. These outcomes were not measured here, so the window-level results cannot quantify an availability benefit.

Of the 1440 abnormal windows, 195 (13.54%) remained indeterminate. These windows were not falsely accepted as normal, but their underlying conditions remained unresolved; any safety benefit would depend on timely verification and action. Normal-window deferral also creates a substantial review demand: 240 of 360 normal windows (66.67%) remained indeterminate. At a hypothetical 5% abnormal prevalence, unchanged class-conditional outcomes would yield 640.1 deferred windows per 1000, compared with 654.7 for the strict rule, and 2.64 false-abnormal decisions per 1000. The acceptance difference would be only 1.45 percentage points. Several windows may concern one unresolved condition, and an acquisition problem may be resolved without equipment intervention. Feasibility therefore depends on how windows are grouped into review requests, which requests require attendance and how long resolution takes. The present counts cannot determine inspection hours or staffing needs.

The thresholds define one research operating point and have not been independently calibrated as fire-safety limits. The clear boundary of 4 and alarm boundary of 6 leave intermediate departures unresolved; confidence 0.90, quality 0.50 and completeness 0.80 impose additional acceptance requirements. The post hoc results in Section 3.5.1 illustrate the tradeoff: alarm boundaries of 5, 6, and 7 produced 1412, 1365, and 1316 accepted

windows with 11, 1, and 0 observed errors, respectively, leaving 388, 435, and 484 windows deferred. At these alarm settings, all errors were false-abnormal decisions. Raising the clear boundary to 5 introduced one false-normal decision. Quality floors of 0.40, 0.50 and 0.60 produced identical accepted and error counts; completeness was not varied. Selection for an operating installation would require independent data, an explicit balance among false reassurance, review requests and unresolved conditions, and evaluation on data separate from threshold calibration. The present values should not be used directly as safety limits for an operating fire-protection system.

4.3. Limitations and Future Validation

The main limitation is the construction of the evaluation distribution. Simulated faults introduce mean shifts, harmonic content, and imbalance, and the rule tests departures from references drawn from the same configuration family. The benchmark therefore favors abnormalities expressed through its selected features. Engineering-only assessment is essential for examining what the diagnostic gate adds under these conditions. Fresh random realizations do not test transfer to unseen pumps or operating regimes [42,43]. Unmatched operating states and drift outside the simulated degradation model remain untested; the nominal drift setting of 0.06 is not a validated physical tolerance. The compressed vibration representation at nominal 42.67 Hz also cannot reproduce the bandwidth of the historical recordings at approximately 13 kHz.

Sampling uncertainty is a separate limitation. The holdout contains 360 windows per class across three fixed model runs, and the bootstrap intervals condition on those models, references, and the generator. Comparisons involving zero versus one observed error provide limited evidence about rare events. The absence of observed false-normal decisions does not establish a field missed-fault rate. The separate laboratory-surrogate study in Section S2 failed to transfer across its retained condition files; because it uses a different generator and a single-channel task, it provides no validation of the present multimodal simulation.

The historical comparison concerns the earlier system. The materials supplied for revision do not contain the linked record-level inputs and compatible references required to reconstruct the new hierarchy on that archive, as detailed in Section 2.6. This limits the present evaluation without establishing how many original records lack any particular input. A subsequent study would require vibration, hydraulic, and electrical measurements linked to operating states, compatible normal references, and condition assessments established independently of the rule. The fixed rule could initially be evaluated alongside scheduled inspections and tests. Accepted errors and unresolved cases should be recorded separately, requested checks compared with measured findings, and resolution time linked to equipment disposition. This would test both transfer of the decision rule and the practical value of its verification requests.

5. Conclusions

This simulation study evaluated a hierarchical rule that separates acceptance of a normal or abnormal condition from the measurement domains supporting abnormality. On 1800 fresh synthetic windows, the rule accepted 1365 with one observed binary error, compared with 1021 accepted windows and no observed errors under strict single-domain corroboration. Hydraulic evidence supported 259 of 360 hydraulic-labeled windows (71.94%); the larger total of 323 accepted hydraulic-class windows included 64 supported only by other domains. Rule decomposition showed that relaxing single-domain corroboration accounted for most of the increase in acceptance.

The principal constraint was deferral of 240 of 360 normal windows (66.67%). At a hypothetical 5% abnormal prevalence, the expected 640.1 deferred windows per 1000 corresponded to an acceptance difference of only 1.45 percentage points from the strict rule. The practical cost of resolving those windows has not been measured. The study characterizes decision behavior within the specified simulation, with the historical comparison providing separate evidence about the earlier system. Synchronized measurements, independent condition assessments, and recorded inspection outcomes are needed to establish field performance and any contribution to fire-pump availability.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/fire9110503/s1, Sections S1–S10 provide generator assumptions and matched-model controls, the separate laboratory-surrogate study, the diagnostic architecture, historical indicators, metric definitions, training settings, post-hoc sensitivity and uncertainty results, the additional-analysis design, and the public-file coverage and revision checks. Figure S1: Implemented PBQ-PCN diagnostic component; Figure S2: Variation across the three frozen model runs and the complete post-hoc threshold grid; Table S1: Matched model controls on the earlier synthetic holdouts; Table S2: Mechanisms represented in the controlled five-class generator; Table S3: Acquisition degradation settings by split; Table S4: Accepted binary errors at the fixed coverage grid; Table S5: Natural operating points by model run; Table S6: One-factor threshold sensitivity of the hierarchy; Table S7: Mutually exclusive reasons for hierarchical deferral; Table S8: True class decisions at the original native ranked common budget; Table S9: Conditional paired differences at natural acceptance on the 1800 design windows.

Author Contributions: Conceptualization and writing—original draft preparation, Z.Y.; investigation and data curation, Y.W., H.W. and X.H.; software and validation, J.W.; supervision and writing—review and editing, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The synthetic windows and normal references, nine matched-model checkpoints, saved predictions and decision records, code and protocols for the primary hierarchy and its control, threshold, component and class-mixture analyses are available in Zenodo at https://doi.org/10.5281/zenodo.22858721 (version 1.0.0) (accessed on 29 October 2026). The deposit also contains source-selection manifests, derived statistics, and segment-level predictions for the separate NLN-EMP study. Its original laboratory waveforms are available from NLN-EMP version 4 at https://doi.org/10.4121/2b61183e-c14f-4131-829b-cc4822c369d0.v4 (accessed on 29 October 2026); generated surrogate waveforms, larger calibration pools, and serialized laboratory classifiers are not included in the Zenodo package. The historical monitoring records are not publicly available because of confidentiality restrictions; requests should be addressed to the corresponding author and are subject to authorization. Run-level outputs for the earlier full-objective benchmark summarized in Section S1 are also absent from the public package. Section S10 maps the available files to the analyses and states the scope of the revision checks.

Acknowledgments: OpenAI Codex assisted with English editing, manuscript and figure preparation, and computational analysis. Its use for the revision checks is described in Section 2.9. The authors are responsible for reviewing the assisted work and for the accuracy and integrity of the manuscript.

Conflicts of Interest: Zengyi Yu, Yanpeng Wang, Jiahao Wu, Huaifeng Wang, and Xinyu Han are employed by CNOOC Safety Technology Services Co., Ltd. The authors declare no other relationships or interests requiring disclosure in connection with this manuscript.

References

1. Sobral, J.; Ferreira, L.A. Availability of fire pumping systems under periodic inspection. *J. Build. Eng.* **2016**, *8*, 285–291. [CrossRef]

2. Jardine, A.K.S.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510. [CrossRef]
3. Kanta, L.; Brumbelow, K. Vulnerability, Risk, and Mitigation Assessment of Water Distribution Systems for Insufficient Fire Flows. *J. Water Resour. Plan. Manag.* **2013**, *139*, 593–603. [CrossRef]
4. Larin, A.N.; Larin, A.A.; Ushapivsky, I.L. Experimental investigations of vibrations of centrifugal fire pump with the defect (unfastened shaft in a bearing seat). *Saf. Fire Technol.* **2014**, *34*, 133–141. [CrossRef]
5. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics—A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520. [CrossRef]
6. Rapur, J.S.; Tiwari, R. On-line time domain vibration and current signals based multi-fault diagnosis of centrifugal pumps using support vector machines. *J. Nondestruct. Eval.* **2019**, *38*, 6. [CrossRef]
7. Nandi, S.; Toliyat, H.A.; Li, X. Condition monitoring and fault diagnosis of electrical motors—A review. *IEEE Trans. Energy Convers.* **2005**, *20*, 719–729. [CrossRef]
8. Han, Y.; Zou, J.; Gong, B.; Luo, Y.; Wang, L.; Presas, A.; Yuan, J.; Wang, C. The use of model-based voltage and current analysis for torque oscillation detection and improved condition monitoring of centrifugal pumps. *Mech. Syst. Signal Process.* **2025**, *222*, 111781. [CrossRef]
9. Tiwari, R.; Bordoloi, D.J.; Dewangan, A. Blockage and cavitation detection in centrifugal pumps from dynamic pressure signal using deep learning algorithm. *Measurement* **2021**, *173*, 108676. [CrossRef]
10. Čdina, M. Detection of cavitation phenomenon in a centrifugal pump using audible sound. *Mech. Syst. Signal Process.* **2003**, *17*, 1335–1347. [CrossRef]
11. Zhang, S.; Li, Y.; Chen, X.; Zhou, R.; Wu, Z.; Zarhmouti, T. A novel IoT-based performance testing method and system for fire pumps. *Water* **2024**, *16*, 792. [CrossRef]
12. Nguyen, D.-T.; Nguyen, T.-P.; Cho, M.-Y. Cloud-based AIoT intelligent infrastructure for firefighting pump fault diagnosis-based hybrid CNN-GRU deep learning technique. *J. Supercomput.* **2025**, *81*, 462. [CrossRef]
13. Liu, R.; Yang, B.; Zio, E.; Chen, X. Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mech. Syst. Signal Process.* **2018**, *108*, 33–47. [CrossRef]
14. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023; Available online: https://openreview.net/forum?id=ju_Uqw384Oq, https://github.com/thuml/Time-Series-Library/tree/4e938a1767106324dd753b2a44832bf870a0252e (accessed on 29 October 2026).
15. Shao, H.; Lin, J.; Zhang, L.; Galar, D.; Kumar, U. A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance. *Inf. Fusion* **2021**, *74*, 65–76. [CrossRef]
16. Cui, J.; Xie, P.; Wang, X.; Wang, J.; He, Q.; Jiang, G. M2FN: An end-to-end multi-task and multi-sensor fusion network for intelligent fault diagnosis. *Measurement* **2022**, *204*, 112085. [CrossRef]
17. Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J.T.; Peng, X. Provable dynamic fusion for low-quality multimodal data. *Proc. Mach. Learn. Res.* **2023**, *202*, 41753–41769.
18. Neverova, N.; Wolf, C.; Taylor, G.W.; Nebout, F. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1692–1706. [CrossRef]
19. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. SMIL: Multimodal learning with severely missing modality. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2302–2310. [CrossRef]
20. Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. *Proc. Mach. Learn. Res.* **2019**, *97*, 2151–2159.
21. Yoon, J.; Jarrett, D.; van der Schaar, M. Time-series Generative Adversarial Networks. In Proceedings of the Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32. Available online: https://papers.nips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html (accessed on 29 October 2026).
22. Kuo, N.I.-H.; Polizzotto, M.N.; Finfer, S.; Garcia, F.; Sönnerborg, A.; Zazzi, M.; Böhm, M.; Kaiser, R.; Jorm, L.; Barbieri, S. The Health Gym: Synthetic health-related datasets for the development of reinforcement learning algorithms. *Sci. Data* **2022**, *9*, 693. [CrossRef] [PubMed]
23. Ang, Y.; Huang, Q.; Bao, Y.; Tung, A.K.H.; Huang, Z. TSGBench: Time Series Generation Benchmark. *Proc. VLDB Endow.* **2023**, *17*, 305–318. [CrossRef]
24. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]
25. Zhang, M.-L.; Zhou, Z.-H. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837. [CrossRef]
26. Chow, C.K. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory* **1970**, *16*, 41–46. [CrossRef]
27. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **2014**, *46*, 44. [CrossRef] [PubMed]

28. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the Fourth International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2–4 May 2016; Available online: https://arxiv.org/abs/1511.07122 (accessed on 29 October 2026).
29. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
30. Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 3942–3951. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 29 October 2026).
32. Mckinzie, B.; Shankar, V.; Cheng, J.Y.; Yang, Y.; Shlens, J.; Toshev, A.T. Robustness in Multimodal Learning under Train-Test Modality Mismatch. *Proc. Mach. Learn. Res.* **2023**, *202*, 24291–24303.
33. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. *Proc. Mach. Learn. Res.* **2017**, *70*, 1321–1330.
34. Bruinsma, S.; Geertsma, R.D.; Loendersloot, R.; Tinga, T. Motor current and vibration monitoring dataset for various faults in an E-motor-driven centrifugal pump. *Data Brief.* **2024**, *52*, 109987. [CrossRef] [PubMed]
35. Bruinsma, S.; Geertsma, R.; Loendersloot, R.; Tinga, T. NLN-EMP—Motor Current and Vibration Monitoring Dataset for Various Faults in an E-Motor-Driven Centrifugal Pump, Version 4; 4TU. ResearchData. 2024. Available online: https://data.4tu.nl/datasets/2b61183e-c14f-4131-829b-cc4822c369d0/4 (accessed on 29 October 2026).
36. Schreiber, T.; Schmitz, A. Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.* **1996**, *77*, 635–638. [CrossRef] [PubMed]
37. Lipton, Z.; Wang, Y.-X.; Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. In Proceedings of the 35th International Conference on Machine Learning PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 3122–3130. Available online: https://proceedings.mlr.press/v80/lipton18a.html (accessed on 29 October 2026).
38. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
39. National Fire Protection Association. *NFPA 20: Standard for the Installation of Stationary Pumps for Fire Protection*, 2025 ed.; National Fire Protection Association: Quincy, MA, USA, 2025; Available online: https://www.nfpa.org/product/nfpa-20-standard/p0020code (accessed on 16 October 2026).
40. National Fire Protection Association. *NFPA 25: Standard for the Inspection, Testing, and Maintenance of Water-Based Fire Protection Systems*, 2026 ed.; National Fire Protection Association: Quincy, MA, USA, 2026; Available online: https://www.nfpa.org/product/nfpa-25-standard/p0025code (accessed on 16 October 2026).
41. Joglar, F.; Ontiveros, V.; Pennel, G. *Applying Reliability Based Decision Making to ITM Frequency*; Final Report, FPRF 2018-08; Fire Protection Research Foundation: Quincy, MA, USA, 2018; Available online: https://www.nfpa.org/-/media/Project/Storefront/Catalog/Files/Research/Research-Foundation/Reports/Detection-and-signaling/RF_ApplyingReliabilityBasedDecisionMakingToITMFrequency.pdf?hash=F7B2EBFD6BCDA4446CF3E8D3219BDF76&rev=472cd0155bf44242b35d3d8e004f3785 (accessed on 16 October 2026).
42. Li, X.; Zhang, W.; Ma, H.; Luo, Z.; Li, X. Domain generalization in rotating machinery fault diagnostics using deep neural networks. *Neurocomputing* **2020**, *403*, 409–420. [CrossRef]
43. Xiao, Y.; Shao, H.; Feng, M.; Han, T.; Wan, J.; Liu, B. Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer. *J. Manuf. Syst.* **2023**, *70*, 186–201. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.