Blockwise Joint Conformal Prediction for Simultaneous Multi-Component Condition Monitoring of Hydraulic Systems

Bùi Thanh Lâm[1*]

[1*]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, No. 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam.

Corresponding author(s). E-mail(s): lamt78789@gmail.com;

Abstract

Automated condition-monitoring systems often assess several components at every operating cycle. A separate 95% prediction set for each component does not, however, provide 95% reliability for the complete system state or for a sequence of cycles. This study formulates the operating-condition block as the unit of conformal calibration. For every calibration block, the proposed score takes the maximum nonconformity over four hydraulic components and ten cycles. A finite-sample order statistic then yields Cartesian component-wise prediction sets with simultaneous block-level coverage under exchangeable blocks. The method was evaluated on 1,440 stable cycles from the UCI hydraulic test-rig data, arranged as 144 complete factorial condition blocks. Thirty block-disjoint 72/43/29 training/calibration/test partitions were run with logistic-regression and random-forest backbones. At nominal 95% coverage, conventional marginal sets covered the complete four-component state in only 0.817 and 0.818 of cycles and covered whole blocks in 0.499 and 0.546, respectively. Blockwise calibration increased mean cycle coverage to 0.994 and 0.985 and mean block coverage to 0.971 and 0.946. The price was strongly backbone-dependent: mean Cartesian set sizes were 103.9 and 13.7 out of 144 possible system states. A post-hoc unstable-cycle stress test reduced random-forest block coverage to 0.830, demonstrating that the guarantee should not be extrapolated across regime shifts. The results show that reliability claims in multi-component diagnosis must match the operational decision unit, while set efficiency and exchangeability require explicit auditing.

Keywords: conformal prediction, fault diagnosis, condition monitoring, hydraulic systems, uncertainty quantification, industrial automation

1 Introduction

Condition monitoring converts multi-sensor measurements into maintenance decisions before a component failure becomes a production or safety event. Machine-learning methods now support this task across rotating machinery, process systems, and predictive-maintenance applications [21, 35, 41]. Much of the literature is organised around point-prediction accuracy. That objective is necessary but incomplete for an automated diagnostic layer: an operator also needs to know when a reported state is unreliable, which alternatives remain plausible, and what level of risk a downstream decision is accepting. Recent fault-diagnosis work has consequently emphasised probability calibration, model uncertainty, and distribution shift [22, 33, 37].

Conformal prediction provides model-agnostic set-valued outputs with a finite-sample marginal coverage guarantee under exchangeability [1, 13, 36]. Its use in industrial monitoring is growing. Published applications address online alarm-flood classification [25], false-alarm control for

principal-component and autoencoder detectors [10], uncertainty-aware process diagnosis [15], and class-conditional open-set diagnosis [16]. These studies establish that conformal tools can make industrial diagnostic uncertainty explicit. They also leave a practical question: what event should the advertised coverage refer to?

An industrial asset can have several simultaneously degraded components. Moreover, maintenance decisions may be triggered only after a state persists over several cycles. If four component-wise sets each have 95% marginal coverage, the probability that all four contain their true labels can be much lower; under independence it is $0.95^4 = 0.8145$. Requiring correctness over ten cycles compounds the mismatch. Calibrating individual sensor rows or cycles therefore cannot support a whole-system, whole-window reliability statement. Dependence-aware conformal research likewise shows that validity hinges on how the exchangeability structure is defined for calibration [3, 8].

This paper studies three questions: how much reliability is lost when component-wise coverage is interpreted as a system-level claim; whether calibration at the operating-condition-block level restores the declared whole-window coverage; and how much diagnostic ambiguity that protection costs. Its contributions are:

1. A precise distinction among component-wise, simultaneous per-cycle, and simultaneous per-block diagnostic coverage.
2. An application-specific block-level split-conformal construction that converts any collection of component classifiers into interpretable component-wise prediction sets while controlling omission of any true label over an entire decision window.
3. A leakage-resistant benchmark on a complete factorial hydraulic condition-monitoring experiment. All ten cycles from a component-state combination remain in one of proper training, calibration, or testing.
4. A 30-split, two-backbone assessment of coverage, set efficiency, empty sets, calibration- and window-size sensitivity, and the consequences of evaluating not-yet-stable cycles outside the calibration regime.

The aim is not to introduce another high-accuracy point classifier. It is to show that the operational unit attached to an uncertainty guarantee changes both the measured reliability and the usefulness of the resulting diagnostic sets.

2 Related work

2.1 Hydraulic condition monitoring

Hydraulic systems couple pumps, valves, accumulators, cooling circuits, and feedback control. Their multi-rate pressure, flow, temperature, vibration, and power measurements provide complementary views in industrial sensing [32], while nonlinear couplings and hidden faults make hydraulic diagnosis difficult [9]. The experimental data introduced by Helwig et al. [17, 18] have become a common benchmark because four component conditions are varied simultaneously across a full factorial design. Early work used multivariate statistics and automated feature selection [18, 31]. Later studies explored real-time recurrent models [20], multi-rate sensor fusion [19, 23], early classification [2], hybrid predictive-maintenance pipelines [6], and functional data analysis for multi-fault classification [39]. Recent deep architectures continue to improve feature extraction and sensor fusion [34, 40].

These works mainly evaluate a component label or a combined point prediction. High point accuracy does not specify the probability that every component state required by a maintenance decision is covered. The present study retains simple, auditable backbones and directs methodological attention to that system-level risk.

2.2 Set-valued and conformal classification

Set-valued classifiers trade ambiguity for controlled error [30]. Split conformal classification separates proper training from calibration, evaluates a nonconformity score on held-out labels, and selects a finite-sample quantile [1, 29]. This wrapper leaves the underlying classifier unchanged. Risk-controlling prediction sets generalize the same principle to task-specific losses, including multi-label decisions [4]; conformal multi-label algorithms have also been studied directly. In particular, multilabel confidence sets can represent label interactions rather than calibrating each label independently [7, 24, 38].

Most conformal guarantees are marginal over a new exchangeable observation. Time dependence, drift, and grouping require care. Block and resampling strategies have been developed for dependent data [8], and hierarchical conformal methods address observations nested within groups [11], while weighted and adaptive approaches address some forms of nonexchangeability [3, 14]. In fault diagnosis, recent methods quantify uncertainty or reject unfamiliar samples [16, 22, 26]. Our construction uses a simpler setting: condition blocks are randomly partitioned from a balanced factorial experiment, and the maximum loss inside each block is conformalized. The resulting guarantee is a direct split-conformal rank argument at the block level, not a claim of robustness to arbitrary temporal drift.

Max-score aggregation for joint events is itself a general conformal device, not the novelty claimed here. Closest in spirit, JANET constructs rectangular joint regions for multi-step time-series regression and can control the familywise or K-familywise error across forecast horizons [12]. The present problem differs in its output and sampling unit: it requires Cartesian sets for several multiclass component states, evaluated across repeated diagnostic cycles, while whole factorial condition blocks—not time-index permutations—form the exchangeable units. Table 1 makes these boundaries explicit. The contribution is the operational formulation, auditable block-disjoint implementation, and industrial evidence for this classification setting.

3 Problem formulation

Let a condition block B_j contain L operating cycles. Each cycle has a feature vector $X_{j\ell} \in \mathbb{R}^d$ and a vector of K component states $Y_{j\ell} = (Y_{j\ell 1}, \ldots, Y_{j\ell K})$. In this study, $K = 4$ and $L = 10$. A fitted component classifier supplies probabilities $\widehat{p}_k(c \mid x)$ for class $c \in \mathcal{Y}_k$. We use the standard probability-complement nonconformity score

$$s_k(x, c) = 1 - \widehat{p}_k(c \mid x). \tag{1}$$

A set-valued diagnosis returns $C_k(x) \subseteq \mathcal{Y}_k$ for every component. Three coverage events should not be conflated:

$$E_{j\ell k}^{\text{component}} = \{Y_{j\ell k} \in C_k(X_{j\ell})\}, \tag{2}$$

$$E_{j\ell}^{\text{cycle}} = \bigcap_{k=1}^{K} E_{j\ell k}^{\text{component}}, \tag{3}$$

$$E_{j}^{\text{block}} = \bigcap_{\ell=1}^{L} \bigcap_{k=1}^{K} E_{j\ell k}^{\text{component}}. \tag{4}$$

Component-wise conformal calibration targets the first event. Automated release of a complete system state requires the second; a persistence-based maintenance rule requires the third.

4 Materials and methods

4.1 Hydraulic test-rig data

The UCI *Condition monitoring of hydraulic systems* data set [17] was measured on a hydraulic rig with connected primary working and secondary cooling-filtration circuits. Each 60-s cycle contains 17 sensor streams: six pressures and motor power at 100 Hz, two volume flows at 10 Hz, and four temperatures, vibration, cooling efficiency, cooling power, and an efficiency factor at 1 Hz. The labels describe cooler efficiency (3, 20, or 100%), valve switching behaviour (73, 80, 90, or 100%), internal pump leakage (0, 1, or 2), and accumulator pressure (90, 100, 115, or 130 bar). The data are distributed under CC BY 4.0 with DOI 10.24432/C5CW21.

The source contains 2,205 cycles and an author-provided stability flag. We excluded 756 cycles for which a static condition might not yet have been reached. This decision was made before model evaluation. The stable subset contains all

$$3 \times 4 \times 3 \times 4 = 144$$

component-state combinations. Every combination has ten stable cycles except a repeated nominal condition with 19. To create equal-size blocks, the first ten stable cycles of each combination were retained, yielding 1,440 observations in 144 blocks. The excluded unstable cycles were preserved only for the post-hoc regime-shift stress test.

4.2 Cycle features and classifiers

Seventeen summaries were computed independently for every sensor and cycle: mean, standard deviation, minimum, 5th, 25th, 50th, 75th and

Table 1 Positioning relative to representative joint and grouped conformal methods. The comparison distinguishes the protected event from the geometry of the reported prediction set.

Method	Primary task	Calibration structure	Protected event	Output form
Cauchois et al. [7]	Multiclass and multilabel classification	Exchangeable observations	True class or label vector is included	Class or structured label set
Dunn et al. [11]	Prediction with nested data	Two-layer hierarchical samples	Coverage under a hierarchical sampling model	Prediction set for a future response
JANET [12]	Multi-step time-series regression	Independent series or time permutations	FWER/K-FWER across forecast horizons	Rectangular joint prediction region
This study	Multi-component condition classification	Exchangeable operating-condition blocks	Every true component state in every cycle is included	Cartesian product of component-wise class sets

95th percentiles, maximum, root mean square, median absolute deviation, range, mean absolute first difference, first-difference standard deviation, normalized linear slope, second-half minus first-half mean, and lag-one autocorrelation. This produced $17 \times 17 = 289$ features. Extraction is cycle-local and uses neither labels nor information from other partitions.

Two deliberately different backbones were used:

1. class-balanced multinomial logistic regression after feature standardization; and
2. a class-balanced random forest with 300 trees, square-root feature subsampling, and minimum leaf size two [5].

One classifier was fitted for each component. Probability scores from different learning algorithms can have different sharpness and calibration [27]; using two backbones tests whether the reliability result is model-agnostic while exposing efficiency differences. Implementations used scikit-learn [28].

4.3 Block-disjoint partitions

For each of 30 repetitions, seed $20260726 + r$ uniformly permuted the 144 condition blocks. The first 72 blocks formed proper training, the next 43 formed conformal calibration, and the remaining 29 formed testing. Thus no component-state combination or one of its adjacent stable cycles appeared in more than one partition. The partition was not optimized using labels, accuracy, or coverage. Conditional on the proper-training blocks, calibration and test blocks are a uniform random partition of the remaining factorial conditions.

4.4 Compared conformal strategies

For n calibration scores $a_1, \ldots, a_n$, define the finite-sample conformal quantile

$$Q_{1-\alpha}(a_{1:n}) = a_{(\lceil (n+1)(1-\alpha) \rceil)}, \tag{5}$$

where $a_{(r)}$ is the rth order statistic and the quantile is infinity if $r > n$.

Four strategies were compared.

Marginal.

For each component k, Eq. (5) is applied to its calibration cycle scores at level α. This targets component-wise, not joint, coverage.

Bonferroni.

Each component is calibrated at α/K. A union-bound argument can protect a simultaneous cycle event when the relevant calibration units are exchangeable, but it does not control an entire L-cycle block.

Joint-cycle.

The score for calibration cycle (j, ℓ) is

$$R_{j\ell}^{\text{cycle}} = \max_{1 \le k \le K} s_k(X_{j\ell}, Y_{j\ell k}).$$

A common threshold is calibrated over pooled cycles. This directly aligns the score with simultaneous component coverage, but treats cycles as calibration units despite within-block dependence.

Joint-block.

The primary method defines one score per calibration block,

$$R_j^{\text{block}} = \max_{\substack{1\le \ell \le L \\ 1 \le k \le K}} s_k(X_{j\ell}, Y_{j\ell k}), \tag{6}$$

and sets

$$q_\alpha = Q_{1-\alpha}\left(R_1^{\text{block}}, \ldots, R_m^{\text{block}}\right). \tag{7}$$

The component sets remain simple and interpretable:

$$C_k(x) = \{c \in \mathcal{Y}_k : s_k(x, c) \le q_\alpha\}. \tag{8}$$

Proposition 1 (Simultaneous block coverage) *Condition on the proper-training data and fitted classifiers. If the m calibration blocks and one future block are exchangeable, and ties are handled by the non-randomized higher quantile in Eq. (5), then the sets in Eq. (8) satisfy*

$$\Pr\left[\bigcap_{\ell=1}^{L}\bigcap_{k=1}^{K}\{Y_{m+1,\ell k} \in C_k(X_{m+1,\ell})\}\right] \ge 1-\alpha.$$

Proof Apply the same score function in Eq. (6) to all calibration blocks and the future block. Exchangeability makes the rank of the future block score among the $m+1$ scores uniform up to conservative handling of ties. The order statistic in Eq. (7) therefore exceeds the future block score with probability at least $1-\alpha$. That inequality is equivalent to every true component label in every cycle satisfying $s_k \le q_\alpha$, which is exactly the stated event. □

The proposition is a split-conformal rank result at the chosen block unit. It does not assume independent cycles inside a block. It does require the calibration and future blocks to be exchangeable after fixing the trained model.

Excluding the backbone's probability evaluation, calibration requires $O(mLK)$ true-label score comparisons, stores m block maxima, and sorts those m values once. At inference, set construction requires one comparison for every candidate component class. No conformal refitting is needed when a new cycle arrives; the wrapper adds only these class-wise comparisons to the four classifier evaluations.

Training-only normalization ablation.

As a predeclared score-design check, each component/class score was divided by its 90th-percentile true-label score in the proper-training data, lower-bounded at 10^{-6}. The fixed scale depends on neither calibration nor test blocks, so the block-rank argument is unchanged. Supplementary Table S6 reports this ablation; it was not adopted as the primary method because its efficiency effect was not consistent across backbones.

4.5 Evaluation and stress test

Nominal coverage levels were 90% and 95%. We recorded component-wise coverage, simultaneous per-cycle coverage, simultaneous whole-block coverage, the product $\prod_k |C_k(x)|$ (the number of complete system states consistent with the Cartesian output), singleton-system rate, and empty-system rate. Point classifiers were evaluated with accuracy, macro-F1, and balanced accuracy. Results are means and standard deviations over the 30 fixed partitions; 2.5th–97.5th percentiles across partitions are supplied in the machine-readable results. Because the partitions overlap in observations, repetitions quantify split sensitivity and are not treated as 30 independent physical experiments.

For a post-hoc stress test, models and thresholds calibrated on stable blocks were applied to source cycles flagged as not yet stable. Only unstable cycles whose component-state combination belonged to that repetition's test blocks were included. This is deliberately outside the primary exchangeable-block claim and tests how clearly coverage failure exposes a regime mismatch.

4.6 Post-hoc robustness analyses

Two additional analyses were specified after the primary results had been inspected and are therefore labelled post hoc. First, the protected window was varied over $L \in \{1, 5, 10\}$. For each primary 72/43/29 split, the classifier continued to use all

ten cycles in every proper-training block, while calibration and testing used the first L source-order cycles per block. This holds the fitted model and block allocation fixed while changing the decision window and its conformal threshold.

Second, calibration size was varied over $m \in \{29, 43, 57\}$ with 58 fixed proper-training blocks, 29 fixed test blocks, and nested prefixes of a 57-block calibration pool. Blocks outside a smaller prefix were unused. Thus, within a repetition, the fitted classifier and test set were identical across the three calibration sizes. Both analyses used the same 30 seeds and the nominal 95% joint-block method. They assess sensitivity of this benchmark result; they are not independent confirmation experiments.

4.7 Computational assistance and accountability

OpenAI Codex assisted the preparation workflow with literature discovery, code generation, analysis scripting, and manuscript drafting. It did not make autonomous authorship or accountability decisions. Before submission, the named human authors must independently inspect the data, code, results, citations, and prose; correct any errors; approve the final disclosure; and accept full responsibility for the submitted work.

5 Results

5.1 Point-diagnostic performance

Table 2 confirms that the backbones were informative but not uniformly easy across components. Both classified cooler condition perfectly. Pump leakage was also well separated, with mean macro-F1 of 0.973 for logistic regression and 0.979 for the random forest. Accumulator diagnosis favoured the random forest (0.878 versus 0.755 macro-F1), whereas valve diagnosis remained the most difficult random-forest target (0.696 macro-F1). These differences are useful for the set-valued analysis: a reliability wrapper must accommodate heterogeneous task difficulty rather than hiding it in one aggregate accuracy.

5.2 Coverage depends on the declared unit

Table 3 and Fig. 1 show the 95% results. Marginal component sets yielded mean simultaneous cycle coverage of only 0.817 (logistic) and 0.818 (random forest), close to the 0.95^4 independence calculation despite substantial dependence among tasks. Whole-block coverage fell further to 0.499 and 0.546. Thus reporting only four component-wise coverage values would materially overstate the reliability of the complete diagnosis.

Bonferroni and joint-cycle calibration improved simultaneous cycle coverage. Their block coverage nevertheless remained between 0.775 and 0.884. The primary joint-block method raised mean block coverage to 0.971 for logistic regression and 0.946 for the random forest. Across splits, the corresponding 2.5th–97.5th percentile ranges were 0.859–1.000 and 0.809–1.000, illustrating finite test-set variability with only 29 held-out blocks. At the 90% target, joint-block mean block coverage was 0.914 and 0.915 for the two backbones.

5.3 Coverage–efficiency trade-off

The stronger event was not free. Logistic joint-block sets contained a mean of 103.9 complete states and had a median split-level mean of 144, the entire state space. The same method with random-forest scores averaged 13.7 states and had a median split-level mean of 6.0. The large standard deviation (25.8) reflects a few random-forest partitions that required very conservative thresholds. Therefore the conformal construction is model-agnostic in validity, but not in usefulness. A high-coverage set can be too broad to guide maintenance.

The training-only normalization ablation reinforced that conclusion. At 95% joint-block coverage, it reduced the logistic mean set size from 103.9 to 33.9 while maintaining 0.966 mean block coverage, but increased the random-forest mean size from 13.7 to 47.5 with 0.953 mean block coverage. A normalization that helps one score geometry can harm another; it is not a universal efficiency correction.

Marginal methods were sharper but sometimes empty: at 95%, their mean empty-system rates were 0.081 for logistic regression and 0.098 for the

Table 2 Point-prediction performance over 30 block-disjoint repetitions. Entries are mean ± standard deviation.

Model	Target	Balanced accuracy	Macro-F1	Accuracy
Logistic regression	Cooler	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	Valve	0.766 ± 0.052	0.745 ± 0.058	0.755 ± 0.059
	Pump	0.974 ± 0.026	0.973 ± 0.029	0.973 ± 0.029
	Accumulator	0.767 ± 0.061	0.755 ± 0.063	0.762 ± 0.062
Random forest	Cooler	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	Valve	0.727 ± 0.074	0.696 ± 0.086	0.703 ± 0.085
	Pump	0.981 ± 0.019	0.979 ± 0.023	0.979 ± 0.022
	Accumulator	0.885 ± 0.067	0.878 ± 0.069	0.884 ± 0.065

Table 3 Set-valued performance at nominal 95% coverage over 30 block-disjoint repetitions. Set size is the Cartesian number of complete four-component states; its maximum is $3 \times 4 \times 3 \times 4 = 144$. Entries are mean ± standard deviation.

Model	Method	Cycle coverage	Block coverage	Mean set size	Singleton rate
Logistic regression	Marginal	0.817 ± 0.061	0.499 ± 0.117	3.75 ± 1.39	0.088 ± 0.071
	Bonferroni	0.946 ± 0.029	0.775 ± 0.097	11.50 ± 7.20	0.007 ± 0.009
	Joint-cycle	0.938 ± 0.038	0.776 ± 0.109	16.51 ± 17.57	0.015 ± 0.020
	Joint-block	0.994 ± 0.010	0.971 ± 0.050	103.87 ± 48.06	0.001 ± 0.002
Random forest	Marginal	0.818 ± 0.082	0.546 ± 0.146	2.48 ± 1.29	0.134 ± 0.084
	Bonferroni	0.936 ± 0.054	0.793 ± 0.126	4.16 ± 1.96	0.045 ± 0.047
	Joint-cycle	0.950 ± 0.046	0.884 ± 0.074	5.80 ± 6.16	0.065 ± 0.059
	Joint-block	0.985 ± 0.025	0.946 ± 0.058	13.65 ± 25.78	0.016 ± 0.021

random forest. Bonferroni empty rates were 0.017 and 0.021. Neither joint-cycle nor joint-block produced empty system sets in these experiments. Empty sets are not a mathematical error, but an automated workflow must map them to an explicit escalation or out-of-model response.

5.4 Sensitivity to window and calibration size

The $L = 10$ robustness run reproduced every primary joint-block metric at the repetition level, providing a direct implementation check. As the protected prefix increased from one to five and ten cycles, logistic mean Cartesian set size increased from 26.68 to 90.36 and 103.87; the corresponding random-forest values were 10.88, 13.18, and 13.65. Mean block coverage remained near the 95% target: 0.938, 0.967, and 0.971 for logistic regression and 0.960, 0.959, and 0.946 for random forest. Thus a longer protected window was obtained by recalibration and wider sets, not by extrapolating a one-cycle claim (Supplementary Table S7).

Changing calibration size did not produce a monotone efficiency gain (Supplementary Table S8). With the 58-block training fit held fixed, random-forest mean set size was 29.60, 19.48, and 25.08 for $m = 29, 43, 57$; logistic sets remained broad at 130.12, 121.15, and 131.02. All six mean block coverages were between 0.960 and 0.989. At 95%, the finite-sample ranks are the largest calibration score for $m = 29$, second largest for $m = 43$, and third largest for $m = 57$. Additional blocks refine the attainable rank but also introduce new extreme scores, so sharpness need not improve monotonically in one finite data set.

5.5 Unstable-cycle stress test

Figure 2 and Table 4 compare the stable primary test with cycles that the data providers marked as potentially not yet at steady state. Logistic sets remained broad and retained a mean whole-block coverage of 0.942. The sharper random-forest sets were more sensitive: group-macro cycle coverage fell from 0.985 to 0.931, and whole-block coverage

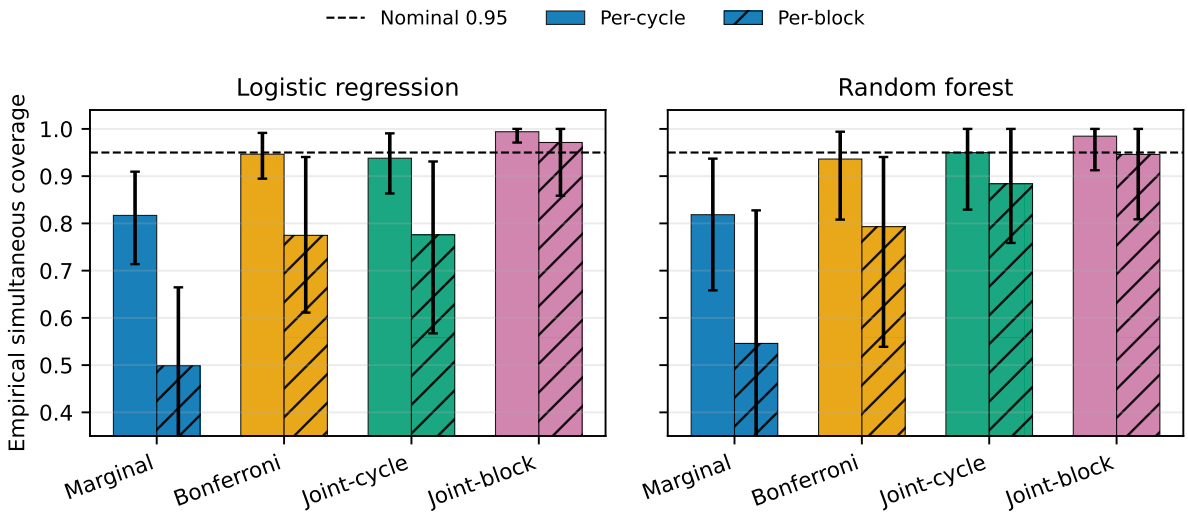

Fig. 1 Mean simultaneous coverage at nominal 95% across 30 block-disjoint partitions. Error bars show the 2.5th–97.5th percentiles over partitions. Solid bars evaluate a component vector for one cycle; hatched bars require all four labels to be covered in all ten cycles of a test block.

fell from 0.946 to 0.830. The micro cycle value was lower still (0.894) because several long transition segments received greater weight. These results do not invalidate Proposition 1; they demonstrate that unstable test cycles are not exchangeable with the stable calibration blocks.

6 Discussion

6.1 The risk statement must match the decision

The main result is conceptual but operationally consequential. A statement such as "each component has 95% coverage" is not interchangeable with "the complete system state has 95% coverage," and neither statement protects ten consecutive cycles. The empirical marginal cycle coverage near 0.82 makes the distinction concrete. Related joint-label reliability issues arise in multi-label classification and structured decisions [4, 24]; industrial monitoring adds temporal grouping and maintenance consequences.

The maximum score in Eq. (6) is intentionally strict. One omitted state at one cycle counts as block failure. This matches applications where a diagnostic report is released only after a persistence window, or where missing a plausible component state anywhere in the review window is unacceptable. A different operational loss could replace the maximum: for example, allowing one transient miss, weighting severe component states more heavily, or controlling the expected number of omitted labels. Risk-controlling prediction-set methods provide a broader framework for such alternatives [4].

6.2 Validity does not imply usefulness

Logistic regression and random forest used the same partitions, score definition, and conformal algorithm. Their block coverage was similar, but their set sizes were radically different. This is consistent with conformal theory: validity is driven by exchangeable ranks, while efficiency depends on how well the score orders correct and incorrect labels [1, 29]. The logistic threshold often admitted nearly the full Cartesian state space. Its apparent robustness in the unstable stress test is therefore not necessarily desirable; a vacuous set is safe but uninformative.

For deployment, coverage and efficiency should be displayed together. Candidate backbones should be selected on a locked development protocol using set size or a maintenance-cost utility, not test coverage alone. Calibration size also matters. With 43 calibration blocks, the 95% block quantile is a high order statistic, so a single extreme block can enlarge every set. The post-hoc

Table 4 Post-hoc stability-regime stress test for the nominal 95% joint-block method. Entries are mean ± standard deviation across the same 30 partitions.

Model	Regime	Cycle coverage	Group-macro cycle coverage	Block coverage
Logistic regression	Stable	0.994 ± 0.010	0.994 ± 0.010	0.971 ± 0.050
	Unstable	0.939 ± 0.143	0.975 ± 0.078	0.942 ± 0.112
Random forest	Stable	0.985 ± 0.025	0.985 ± 0.025	0.946 ± 0.058
	Unstable	0.894 ± 0.164	0.931 ± 0.123	0.830 ± 0.241

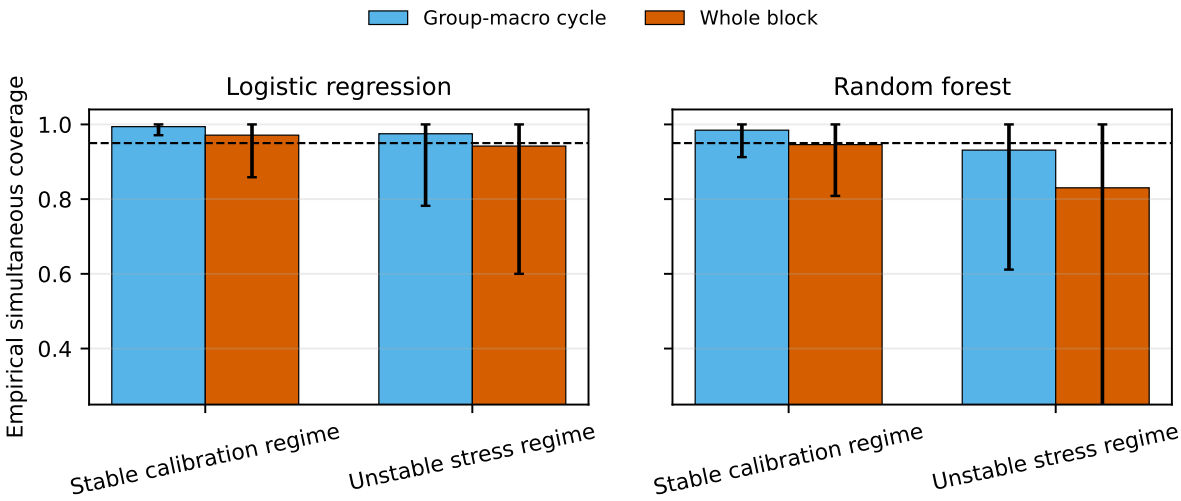

Fig. 2 Joint-block coverage at the nominal 95% level in the stable calibration regime and the not-yet-stable stress regime. Error bars show the 2.5th–97.5th percentiles across partitions. Broad logistic sets mask much of the shift; sharper random-forest sets expose a substantial loss of whole-block coverage.

calibration-size analysis showed this finite-sample stability issue rather than a monotone sharpness improvement. More independent operating-condition blocks improve rank resolution in principle, but they do not remove the need to predeclare the calibration design and audit extreme scores.

6.3 Exchangeability is an engineering assumption

Conformal prediction does not certify arbitrary future operation. The exchangeability assumption must be argued from data acquisition and deployment. Here it is plausible for a finite-population experiment because complete condition blocks are uniformly partitioned after the proper-training set is fixed. It is not plausible when stable calibration is applied to early transients without adjustment. The stress test makes this boundary visible, especially for the sharper random-forest model.

Practical systems should therefore monitor regime indicators, calibration age, sensor availability, and score drift. If operation changes, weighted, adaptive, or nonexchangeable conformal methods may be appropriate [3, 14, 26], but their assumptions and target risk must be validated for the specific asset. A fallback can abstain, request additional cycles, or route the case to a human diagnostician rather than silently reusing an invalid guarantee.

6.4 Limitations and threats to validity

First, the evidence comes from one laboratory hydraulic rig and a balanced factorial condition design. It does not establish performance on a production fleet, new fault mechanisms, different loads, or sensor ageing. Second, labels are constant within a block and the primary block length is fixed at ten; the post-hoc prefix analysis does

not validate arbitrary maintenance-window definitions. Third, the 289 handcrafted summaries favour transparency and reproducibility but do not exhaust spectral or learned representations. Fourth, separate component classifiers do not model label interactions directly. Fifth, only two backbones and the probability-complement score were studied. The very broad logistic sets show that score design is a material limitation.

Sixth, the deterministic first-ten rule equalised block sizes but discarded nine additional stable nominal cycles and may be sensitive to source ordering; the robustness check changed prefix length but not cycle-selection order. Seventh, the post-hoc unstable analysis violates the primary exchangeability design and is diagnostic rather than guaranteed. Eighth, repeated partitions reuse the same physical cycles; their standard deviations measure sensitivity to data allocation, not population sampling uncertainty. Finally, the method controls marginal coverage for a randomly drawn future block under the stated exchangeability condition. It does not provide conditional coverage for every specific fault combination or guarantee a small set.

7 Conclusion

Multi-component condition monitoring needs a reliability unit that matches the automated decision. On the hydraulic benchmark, conventional 95% component-wise conformal sets covered the complete state in only about 82% of cycles and approximately half of ten-cycle condition blocks. Calibrating the maximum error over all components and cycles moved empirical block coverage toward the 95% target. The accompanying increase in ambiguity ranged from moderate for random-forest scores to nearly vacuous for logistic scores.

The practical lesson is twofold. First, system-level coverage should be calibrated and reported explicitly rather than inferred from component metrics. Second, a valid prediction set must also be audited for sharpness and for the exchangeability of the intended deployment regime. Blockwise joint conformal prediction offers a simple, reproducible wrapper for that purpose, but it is a reliability contract tied to a declared data-generating regime, not a substitute for regime monitoring or engineering judgment.

Supplementary information. The accompanying reproducibility package contains the feature-extraction, partitioning, model-fitting, conformal-calibration, aggregation, and plotting code; exact split identifiers; raw repetition-level metrics; data hashes; and the complete study protocol. Raw UCI sensor files are not redistributed.

Acknowledgements. Apart from the generative-AI assistance disclosed below, the author has no acknowledgements to declare.

Statements and Declarations

Funding. The author received no specific funding for this work.

Competing interests. The author declares no competing interests.

Author contributions. Bùi Thanh Lâm: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, project administration, writing—original draft, and writing—review and editing.

Ethics approval. Not applicable. The study uses an openly licensed engineering test-rig data set and involves no human participants or animals.

Data availability. The source data are available from the UCI Machine Learning Repository under CC BY 4.0 at https://doi.org/10.24432/C5CW21. The exact downloaded archive SHA-256 and all preprocessing choices are recorded in the accompanying manifest.

Code availability. All analysis code, fixed random seeds, split identifiers, and result tables are included in the accompanying reproducibility archive. A public archival URL should be inserted after the authors deposit the final release.

Use of generative AI. OpenAI Codex was used to assist with literature discovery, code generation, analysis scripting, and manuscript drafting. It was not used as an author. The author accepts full responsibility for the accuracy, originality, and integrity of the final manuscript.

References

[1] Angelopoulos AN, Bates S (2023) Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591. https://doi.org/10.1561/2200000101

[2] Askari B, Carli R, Cavone G, et al (2022) Data-driven fault diagnosis in a complex hydraulic system based on early classification. IFAC-PapersOnLine 55(40):187–192. https://doi.org/10.1016/j.ifacol.2023.01.070

[3] Barber RF, Candès EJ, Ramdas A, et al (2023) Conformal prediction beyond exchangeability. The Annals of Statistics 51(2). https://doi.org/10.1214/23-aos2276

[4] Bates S, Angelopoulos A, Lei L, et al (2021) Distribution-free, risk-controlling prediction sets. Journal of the ACM 68(6):1–34. https://doi.org/10.1145/3478535

[5] Breiman L (2001) Random forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/a:1010933404324

[6] Buabeng A, Simons A, Frempong NK, et al (2021) A novel hybrid predictive maintenance model based on clustering, smote and multi-layer perceptron neural network optimised with grey wolf algorithm. SN Applied Sciences 3(5). https://doi.org/10.1007/s42452-021-04598-1

[7] Cauchois M, Gupta S, Duchi JC (2021) Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. Journal of Machine Learning Research 22(81):1–42. URL https://jmlr.org/papers/v22/20-753.html

[8] Chernozhukov V, Wüthrich K, Zhu Y (2018) Exact and robust conformal inference methods for predictive machine learning with dependent data. In: Proceedings of the 31st Conference On Learning Theory, pp 732–749, URL https://proceedings.mlr.press/v75/chernozhukov18a.html

[9] Dai J, Tang J, Huang S, et al (2019) Signal-based intelligent hydraulic fault diagnosis methods: Review and prospects. Chinese Journal of Mechanical Engineering 32(1). https://doi.org/10.1186/s10033-019-0388-9

[10] Diallo AR, Homri L, Dantan JY (2025) Reducing false alarms in fault detection: A comparative analysis between conformal prediction and classical methods applied to pca and autoencoders. Journal of Process Control 152:103495. https://doi.org/10.1016/j.jprocont.2025.103495

[11] Dunn R, Wasserman L, Ramdas A (2023) Distribution-free prediction sets for two-layer hierarchical models. Journal of the American Statistical Association 118(544):2491–2502. https://doi.org/10.1080/01621459.2022.2060112

[12] English E, Wong-Toi E, Fontana M, et al (2025) JANET: Joint adaptive prediction-region estimation for time-series. Machine Learning 114(8):177. https://doi.org/10.1007/s10994-025-06812-2

[13] Fontana M, Zeni G, Vantini S (2023) Conformal prediction: A unified review of theory and new challenges. Bernoulli 29(1). https://doi.org/10.3150/21-bej1447

[14] Gibbs I, Candès E (2021) Adaptive conformal inference under distribution shift. In: Advances in Neural Information Processing Systems, pp 1660–1672

[15] Heddoub A, Diallo AR, Homri L, et al (2025) Uncertainty-aware fault diagnosis with conformal prediction. IFAC-PapersOnLine 59(10):536–541. https://doi.org/10.1016/j.ifacol.2025.09.092

[16] Heddoub A, Homri L, Siadat A (2026) Class-conditional conformal prediction with discriminant-analysis backbones for reliable open-set fault diagnosis in safety-critical industrial systems. Journal of Process Control 161:103701. https://doi.org/10.1016/j.jprocont.2026.103701

[17] Helwig N, Pignanelli E (2015) Condition monitoring of hydraulic systems. https://doi.org/10.24432/C5CW21

[18] Helwig N, Pignanelli E, Schutze A (2015) Condition monitoring of a complex hydraulic system using multivariate statistics. In: 2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings. IEEE, pp 210–215, https://doi.org/10.1109/i2mtc.2015.7151267

[19] Huang K, Wu S, Li F, et al (2022) Fault diagnosis of hydraulic systems based on deep learning model with multirate data samples. IEEE Transactions on Neural Networks and Learning Systems 33(11):6789–6801. https://doi.org/10.1109/tnnls.2021.3083401

[20] Kim K, Jeong J (2020) Real-time monitoring for hydraulic states based on convolutional bidirectional lstm with attention mechanism. Sensors 20(24):7099. https://doi.org/10.3390/s20247099

[21] Lei Y, Yang B, Jiang X, et al (2020) Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing 138:106587. https://doi.org/10.1016/j.ymssp.2019.106587

[22] Lin YH, Li GH (2024) Uncertainty-aware fault diagnosis under calibration. IEEE Transactions on Systems, Man, and Cybernetics: Systems 54(10):6469–6481. https://doi.org/10.1109/tsmc.2024.3427345

[23] Ma X, Wang P, Zhang B, et al (2021) A multirate sensor information fusion strategy for multitask fault diagnosis based on convolutional neural network. Journal of Sensors 2021(1). https://doi.org/10.1155/2021/9952450

[24] Maltoudoglou L, Paisios A, Lenc L, et al (2022) Well-calibrated confidence measures for multi-label text classification with a large number of labels. Pattern Recognition 122:108271. https://doi.org/10.1016/j.patcog.2021.108271

[25] Manca G, Kunze FC, Fay A (2024) Addressing uncertainty in online alarm flood classification using conformal prediction. IEEE Access 12:165626–165652. https://doi.org/10.1109/access.2024.3492348

[26] Nanopoulos A, Buza K (2025) Conformal prediction for out-of-distribution time-series classification. Applied Intelligence 55(11). https://doi.org/10.1007/s10489-025-06708-7

[27] Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning - ICML '05. ACM Press, ICML '05, pp 625–632, https://doi.org/10.1145/1102351.1102430

[28] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12:2825–2830. URL https://jmlr.org/papers/v12/pedregosa11a.html

[29] Romano Y, Sesia M, Candès E (2020) Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems, pp 3581–3591

[30] Sadinle M, Lei J, Wasserman L (2018) Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association 114(525):223–234. https://doi.org/10.1080/01621459.2017.1395341

[31] Schneider T, Helwig N, Schütze A (2017) Automatic feature extraction and selection for classification of cyclical time series data. tm - Technisches Messen 84(3):198–206. https://doi.org/10.1515/teme-2016-0072

[32] Schütze A, Helwig N, Schneider T (2018) Sensors 4.0 – smart sensors and measurement technology enable industry 4.0. Journal of Sensors and Sensor Systems 7(1):359–371. https://doi.org/10.5194/jsss-7-359-2018

[33] Shi Y, Wei P, Feng K, et al (2025) A survey on machine learning approaches for uncertainty quantification of engineering systems. Machine Learning for Computational Science and Engineering 1(1). https://doi.org/10.1007/s44379-024-00011-x

[34] Sun J, Ding H, Li N, et al (2024) Intelligent fault diagnosis of hydraulic system based on multiscale one-dimensional convolutional neural networks with multiattention mechanism. Sensors 24(22):7267. https://doi.org/10.3390/s24227267

[35] Ucar A, Karakose M, Kırımça N (2024) Artificial intelligence for predictive maintenance applications: Key components, trustworthiness, and future trends. Applied Sciences 14(2):898. https://doi.org/10.3390/app14020898

[36] Vovk V, Gammerman A, Shafer G (2005) Algorithmic Learning in a Random World. Springer, https://doi.org/10.1007/b106715

[37] Xiao Y, Shao H, Liu B (2025) Evaluating calibration of deep fault diagnostic models under distribution shift. Computers in Industry 171:104334. https://doi.org/10.1016/j.compind.2025.104334

[38] Yang F, Gan X, Wang H, et al (2017) CP-RAkEL: Improving random k-labelsets with conformal prediction for multi-label classification. In: Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications, pp 266–279, URL https://proceedings.mlr.press/v60/yang17a.html

[39] Yildirim C, Franco-Pereira AM, Lillo RE (2025) Condition monitoring and multi-fault classification of hydraulic systems using multivariate functional data analysis. Heliyon 11(1):e41251. https://doi.org/10.1016/j.heliyon.2024.e41251

[40] Zhang D, Zheng K, Liu F, et al (2024) Fault diagnosis of hydraulic components based on multisensor information fusion using improved tso-cnn-bilstm. Sensors 24(8):2661. https://doi.org/10.3390/s24082661

[41] Zhao R, Yan R, Chen Z, et al (2019) Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing 115:213–237. https://doi.org/10.1016/j.ymssp.2018.05.050