International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 4 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

WORKLOAD SHIFTS AND THRESHOLD CALIBRATION IN 5G OPEN RADIO ACCESS NETWORK TELEMETRY ANOMALY DETECTION: A RUN-LEVEL EVALUATION

Raymond GBEGBE

Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire

**Corresponding author: Raymond GBEGBE | E-mail: raymond.gbegbe@inphb.ci*

ABSTRACT

Legitimate workload variations can alter 5G Open Radio Access Network telemetry without indicating faults, causing false alarms in detectors calibrated under a single operating regime. This study evaluates workload-induced normality shift on the SpotLight dataset using 73 physical runs and run-grouped nested validation (ten outer splits and four inner folds). Four KPI-selection strategies—AllPlatform, discriminability-based (Disc), stability-based (Stable), and drift-aware (DAKS)—are assessed with PCA, Isolation Forest, One-Class SVM, and robust Mahalanobis detectors. For PCA and a source-calibrated threshold, the normal 5-UE false-positive rate is 92.40% with AllPlatform, compared with 1.94% for Disc, 1.52% for Stable, and 4.04% for DAKS. DAKS attains AUPRC 0.181, AUROC 0.521, and MCC −0.007, without consistent superiority over the baselines. Recalibrating the DAKS–PCA threshold on normal 5-UE development data lowers the false-positive rate to 1.39% but reduces anomaly recall from 33.45% to 28.43%. These findings distinguish operating-point adaptation from improved anomaly discrimination and highlight the importance of run-level leakage control. Generalization beyond SpotLight is unverified.

Keywords: 5G Open RAN; network telemetry; workload shift; anomaly detection; nested validation; threshold calibration.

1. INTRODUCTION

The disaggregation and programmability of Open RAN make performance-counter monitoring central to fault management, quality assurance, and intelligent network operations [1], [2]. Telemetry-based anomaly detection is attractive because numerical indicators can be collected from software components and network functions without requiring application-layer payload inspection. Yet the distributions of these indicators depend on ordinary changes in active users, traffic mix, resource scheduling, and platform utilization.

An anomaly score calibrated on observations obtained under a particular legitimate workload may consequently respond to a new, equally legitimate workload as though it represented a fault. This mechanism is a normality shift: a change in the normal operating distribution, distinct from a change in the underlying failure process [3], [4]. An operationally useful detector must therefore control false alerts while retaining the ability to rank genuine anomalies above normal observations.

The operational significance of distribution shift is also documented in longitudinal cellular-network measurements, where model degradation and KPI drift can persist despite periodic retraining [5]. Complementary analyses distinguish global changes from feature-specific shifts [6, 7], while recent work studies 5G network-cluster anomaly detection under distribution shifts [8]. These findings motivate a workload-conditioned evaluation rather than treating all changes in network telemetry as evidence of faults.

Feature selection is one possible mitigation. Ranking KPI by discriminability may retain variables that separate known anomalous development runs from normal ones, whereas ranking by workload stability may suppress benign load-sensitive signals. Drift-Aware KPI Selection (DAKS) combines the two objectives. Its previously reported benefits, however, were sensitive to how physical runs were assigned to training and evaluation. The central question is not whether a particular selector can achieve a low FPR in one partition, but whether it confers a reproducible advantage under strict evaluation boundaries.

This study therefore evaluates the effect of workload shift, telemetry selection, and threshold calibration under a run-grouped nested design. Its contributions are: (i) a controlled comparison of source-calibrated false-alarm rates across normal loads; (ii) an assessment of Disc, Stable, and DAKS with four detector families; (iii) an ablation separating score ranking from the operating threshold; and (iv) a run-level assessment of selection stability and the sensitivity of conclusions to repeated partitions. The empirical outcome is deliberately reported even where it contradicts the anticipated benefit of DAKS.

1.1. Background and Study Positioning

O-RAN anomaly detection has been investigated with supervised and federated models, reconstruction-based approaches, and network telemetry collected in experimental deployments [9], [10]. The SpotLight benchmark is especially useful because its traces reflect both regular operating conditions and labeled anomaly sessions [11]. Related work on concept drift and normality-shift adaptation emphasizes that model reliability can deteriorate when operational context changes without a fault [3], [4], [12], [13].

O-RAN monitoring research includes early attack detection through RAN intelligence [14], explainable reconstruction-based traffic anomaly detection [15], and experimentally grounded platforms for control and monitoring [16]. Work on KPI-driven handover monitoring [17] and IDS xApps [18] further illustrates the range of detection targets. These studies are valuable context, but their task-specific evaluation settings are not interchangeable with the run-grouped benign-workload comparison considered here.

A related line of research addresses distribution drift in O-RAN and operational networks through drift-handling frameworks, generative modeling, and anomaly-aware adaptation [12, 19, 20]. Dynamic concept adaptation for anomaly streams has also been studied outside O-RAN [21]. Together with the cellular-network evidence from LEAF [5], these works motivate distinguishing three operations that are often conflated: detecting a distribution shift, selecting features insensitive to benign changes, and recalibrating an alarm threshold.

A separate systems consideration is the cost of requesting and transporting telemetry. Explainability-guided KPI reduction has been studied as a means of lowering O-RAN performance-indicator queries [22]. In the present study, the KPI budget is an algorithmic input-size constraint: the experiments do not measure E2 transport overhead, and no such saving is claimed.

Two distinct forms of adaptation must be distinguished. Feature selection changes the information provided to a detector; threshold adjustment changes the decision boundary applied to a fixed score. A monotonic change of threshold cannot improve AUROC or AUPRC when the score ordering is unchanged, although it can alter FPR, precision, recall, and MCC at a selected operating point. Separating these effects is essential when evaluating a workload-adaptive monitoring strategy.

2. MATERIALS AND METHODS

2.1. SpotLight Dataset and Analysis Unit

We use the SpotLight platform telemetry archive [11, 23] and the anomalous-session annotations documented in the validation report. The corpus comprises 73 unique physical executions. Each execution contains a temporally ordered sequence of telemetry

observations; multiple rows within the same execution are correlated and therefore are not treated as independent experimental units. The primary experiments use platform KPIs rather than pooling platform and radio variables indiscriminately. The platform-KPI monitoring architecture is summarized in Fig. 1.

The 16 anomalous runs include 12 MIXED sessions and four single-family sessions (MAC, NETWORK, PDCP, RADIO). The 7-UE condition comprises only two physical runs and is therefore treated as an exploratory stress condition, not a confirmatory generalization test.

2.2. Nested Validation and Leakage Control

We evaluate ten outer partitions (seeds 1, 7, 13, 23, 31, 42, 57, 71, 89, and 99). Within each outer development set, four run-grouped folds tune the KPI budget K and penalty coefficient λ. The searched values are $K \in \{25, 50, 75, 100\}$ and $\lambda \in \{0.5, 1, 3, 10\}$. The internal selection criterion maximizes average AUPRC subject to mean normal 5-UE FPR ≤ 5%; ties are resolved using recall at the FPR constraint, followed by the smaller feature budget. The outer-test runs are withheld from preprocessing estimation, KPI ranking, model fitting, and tuning. The run-level nested-validation workflow is summarized in Fig. 2.

The leakage audit reports PASS for all ten outer splits: no physical run has conflicting external roles within a split, and no inner-validation execution overlaps an outer-test set. The ten outer partitions nevertheless reuse some physical runs across seeds. Aggregate split means and their dispersion therefore describe allocation sensitivity; they do not constitute ten independent replications.

2.3. Descriptive Workload-Shift Characterization

A separate descriptive examination of normal SpotLight runs compares 35 physical executions at 1 UE with 20 at 5 UE (Figure 3). The example processing-activity counter (l1app_main_ebbupool_act_2_mean) exhibits a workload-dependent change in run-level medians. Across the 212 platform KPIs meeting the descriptive missingness criterion, 49 have an absolute run-level Cliff's delta of at least 0.474. The example KPI and the cohort-wide summary are shown for post hoc characterization only: they were not used to select KPIs, tune hyperparameters, or calibrate any detector. All model-development transformations remain restricted to the corresponding nested-training data.

2.4. KPI Selection and Detectors

2.4.1. Preprocessing and Selection Strategies

Source-normal 1-UE training runs determine median-based imputation, robust scaling by the interquartile range, and clipping of normalized extremes. These transformations are fitted within development data, never on the outer test. The compared selectors are AllPlatform (all eligible platform indicators), Disc (univariate anomaly discriminability), Stable (low sensitivity to normal workload changes), and DAKS (joint discriminability and workload-sensitivity ranking).

For KPI j, the discriminability score is defined by the direction-invariant univariate development AUROC in (1). The workload-sensitivity score in (2) combines robust normal-workload location and scale differences. DAKS ranks KPI according to the penalized utility in (3) and retains the K largest scores. In (2), m and I denote the median and interquartile range; superscripts (s) and (w) identify source- and alternative-workload normal data, and ε>0 stabilizes denominators. The handling of constant indicators must follow the experimental scripts.

$$A_j = 2\mathrm{abs}\left(\mathrm{AUROC}_j - \frac{1}{2}\right) \quad (1)$$

$$W_j = \frac{\mathrm{abs}\left(m_j^{(s)} - m_j^{(w)}\right)}{I_j^{(s)} + \varepsilon} + \mathrm{abs}\left(\ln\left(\frac{I_j^{(w)} + \varepsilon}{I_j^{(s)} + \varepsilon}\right)\right) \quad (2)$$

$$S_j^{\mathrm{DAKS}} = \frac{A_j}{1 + \lambda W_j} \quad (3)$$

2.4.2. Detector Families and Calibration

Selected KPI subsets are evaluated with four anomaly-scoring models: PCA reconstruction error, Isolation Forest, One-Class SVM, and a robust Mahalanobis distance using a median

location and Ledoit–Wolf covariance estimator. PCA uses the normalized reconstruction error in (4), where z_t denotes the scaled KPI vector and the hatted vector is its PCA reconstruction. Detector hyperparameters are primarily held fixed to isolate the feature-selection effect.

$$r_t = \frac{1}{K}\|\mathbf{z}_t - \hat{\mathbf{z}}_t\|_2^2 \quad (4)$$

Two calibration settings are distinguished. Source calibration uses held-out normal 1-UE runs; adaptive calibration uses a separate normal 5-UE development subset. For either fixed scoring model, the decision threshold follows the calibration quantile in (5), using only the appropriate normal calibration runs. The selected KPI subset, detector parameters, and score ordering remain unchanged.

$$\tau = Q_{1-\alpha}\left(\{r_t : t \in \mathcal{D}_{\text{cal}}^N\}\right), \quad \alpha = 0.01 \quad (5)$$

3. RESULTS AND DISCUSSION

3.1. Workload-Induced False Alarms and PCA Ranking

Table 2 reports the results of the nested experiment. Source calibration combined with all eligible platform KPIs yields 92.40% FPR under normal 5-UE traffic. The very high false-alarm rate demonstrates a severe operating-point mismatch under benign workload shift. All three KPI-selection strategies sharply reduce this failure mode, but DAKS does not lead the comparison: Stable records the lowest FPR (1.52%), while Disc records the highest AUPRC (0.200).

The discrepancy between the 5-UE FPR and anomaly-ranking metrics should not be ignored. Discriminability measured by AUPRC is not preserved by every selection mechanism that reduces false alarms. In particular, a low FPR does not necessarily indicate that anomalous sessions can be reliably ranked above legitimate traffic. DAKS–PCA has AUROC 0.521 and MCC −0.007, providing insufficient evidence for operationally reliable discrimination.

3.2. Transfer across Anomaly Detectors

The trend observed with PCA does not become uniformly favorable when the detector family changes. DAKS achieves a marginal AUPRC advantage over Stable only for robust Mahalanobis (0.164 versus 0.162, rounded), with negative MCC and AUROC 0.508. For the other three detector families, its AUPRC remains below Stable. None of the tested DAKS configurations attains AUROC 0.60 or positive MCC. Thus, the use of multiple detector families strengthens the negative finding rather than supporting universal transfer.

3.3. Threshold Adaptation and Operating-Point Trade-Off

Adaptive calibration lowers DAKS–PCA 5-UE FPR from 4.04% to 1.39%, but recall falls from 33.45% to 28.43%. An unchanged anomaly score ordering means that AUPRC and AUROC do not change, independently of the chosen threshold. Hence the improvement is an operating-point correction, not increased anomaly separability. This distinction is central for network operations: suppressing spurious alarms may reduce operational burden, but a deployment decision must also quantify the detection loss for genuine incidents.

3.4. Physical-Run Uncertainty and Selector Stability

A paired comparison at the physical-run level reports that DAKS increases 5-UE FPR relative to Stable by 2.28 percentage points (95% bootstrap interval 0.65–4.46) and relative to Disc by 1.92 points (0.96–3.26). The paired per-session AUPRC difference relative to Stable is +0.0088 with a 95% interval from −0.0042 to 0.0217; this interval includes zero. This per-session average is not equivalent to the aggregate AUPRC in Table 2 because individual runs contain different numbers of rows and anomaly prevalences.

Stable yields substantially higher feature-set agreement across outer partitions than DAKS (mean Jaccard 0.697 versus 0.376). The chosen DAKS budgets range from 25 to 100 and λ spans the full tested grid from 0.5 to 10. The resulting hyperparameter variability is inconsistent with a claim that this dataset identifies one robust, broadly transferable operating configuration.

3.5. Interpretation and Operational Implications

3.5.1. Workload Shift Is Not an Anomaly

The primary observation is that a source-calibrated detector can conflate legitimate changes in O-RAN platform utilization with faults. The very high AllPlatform 5-UE FPR illustrates the danger of deploying a fixed operating point outside its calibration regime. This failure is not specific to a particular alarm threshold or model architecture; it arises from a mismatch between the monitored normality distribution and the one used to determine the score scale and operating point. The descriptive run-level shifts shown in Fig. 3 establish that legitimate workload regimes exhibit distinct platform-telemetry distributions; this observation does not, by itself, identify which KPIs are informative for anomaly discrimination.

3.5.2. Why DAKS Does Not Provide Consistent Benefit

DAKS attempts to strike a compromise between anomaly-relevant features and features invariant to benign workload changes. That compromise is not automatically superior to either objective considered separately. Disc retains stronger anomaly ranking under PCA, whereas Stable better controls load-induced false alarms and produces more repeatable feature sets. A composite utility score can therefore underperform when the penalty coefficient is difficult to identify from small development samples, when workloads have heterogeneous responses across KPI, or when the available anomaly families fail to represent future faults. These are plausible interpretations rather than mechanisms directly demonstrated by the present experiment.

3.5.3. Evaluation Design and Generalization

The nested design reduces a critical threat: test information must not determine feature rankings, the budget K, the stability weight λ, or the threshold. Nonetheless, run-grouped partitions drawn from one experimental environment are not a substitute for deployment-level generalization. Some runs recur across outer seeds; uncertainty intervals should be based on physical-run-level units and interpreted as conditional on this corpus. Moreover, only two 7-UE Ping runs are available. The experiment is therefore best read as a controlled case study of evaluation sensitivity and workload shift, not as proof of general O-RAN performance.

3.5.4. Implications for 5G Network Management

From an operational perspective, the clearest actionable result is that calibration strategy and KPI selection should be assessed separately. A load-conditioned alarm threshold can stabilize false-positive behavior when independent normal calibration data from the target regime are available, but it cannot recover separability absent from the anomaly score. In a 5G network-management pipeline, workload context may inform model selection or operating-point calibration; the reliability of such a design must be assessed jointly with missed-alarm costs, normal-regime coverage, and the latency of adaptation. The current study does not validate a deployed xApp or demonstrate closed-loop performance.

3.6. Limitations and Next Validation

This analysis is restricted to one SpotLight environment and a limited number of independent physical runs. The anomaly corpus contains 16 runs, and the 7-UE stress scenario contains only two. The nested validation therefore supplies stronger within-corpus evidence but not external validation across sites, hardware platforms, or traffic generators. Model hyperparameters are largely fixed to isolate KPI selection. The conclusions apply to this prespecified comparison and cannot rule out all other model configurations.

A confirmatory extension should use a genuinely independent 5G/O-RAN telemetry corpus or a separately collected testbed campaign with documented workload labels. It should (i) define the physical-run unit before model fitting; (ii) preregister primary metrics and acceptable operating constraints; (iii) compare fixed source, unsupervised recalibration, and load-conditioned thresholds without access to test labels; (iv) report

uncertainty based on independent runs; and (v) include a sufficiently large set of high-load normal runs and distinct anomaly families. No such external evidence is claimed in this manuscript.

4. CONCLUSIONS

This run-level study investigated workload-induced normality shift in 5G O-RAN anomaly detection using SpotLight telemetry. Under nested validation, selecting a compact KPI subset substantially reduced the false-positive rate seen when all eligible platform indicators were used under a legitimate 1-UE-to-5-UE workload transition. However, DAKS did not outperform discriminability-only or stability-only selection consistently and failed to provide reliable anomaly ranking across four detector families. Adaptive thresholds reduced false alarms but also reduced anomaly recall, leaving ranking metrics unchanged. The evidence therefore supports careful separation of feature selection, operating-point calibration, and physical-run-level evaluation, rather than a claim of general DAKS superiority. Independent-site or independent-dataset validation is required before these findings can support broader deployment guidance.

DATA AND CODE AVAILABILITY

The public SpotLight dataset is available at https://github.com/netsys-edinburgh/SpotLight. The nested-validation report identifies the scripts daks_nested_validation_experiment.py and analyze_daks_nested_results.py and the files outer_test_results.csv, per_physical_run_results.csv, selected_configuration_by_split.csv, and leakage_audit.csv. Public release and exact environment details for the author-generated result package must be verified before submission.

REFERENCES

[1] Polese, M., Bonati, L., D'Oro, S., Basagni, S. and Melodia, T. (2023). Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges. IEEE Communications Surveys & Tutorials, 25(2), 1376–1411. doi: 10.1109/COMST.2023.3239220.

[2] Abdalla, A. S., Upadhyaya, P., Shah, V. K. and Marojevic, V. (2022). Toward Next Generation Open Radio Access Networks: What O-RAN Can and Cannot Do! IEEE Network, 36(6), 206–213. doi: 10.1109/MNET.108.2100659.

[3] Han, D., et al. (2023). Anomaly Detection in the Open World: Normality Shift Detection, Explanation, and Adaptation. Proceedings of the 30th Network and Distributed System Security Symposium (NDSS). doi: 10.14722/ndss.2023.24830.

[4] Hinder, F., Vaquet, V. and Hammer, B. (2024). Feature-based analyses of concept drift. Neurocomputing, 600, 127968. doi: 10.1016/j.neucom.2024.127968.

[5] Liu, S., Bronzino, F., Schmitt, P., Bhagoji, A. N., Feamster, N., Garcia Crespo, H., Coyle, T. and Ward, B. (2023). LEAF: Navigating Concept Drift in Cellular Networks. Proceedings of the ACM on Networking, 1(CoNEXT2), Article 7, 1–24. doi: 10.1145/3609422.

[6] Kulinski, S., Bagchi, S. and Inouye, D. I. (2020). Feature Shift Detection: Localizing Which Features Have Shifted via Conditional Distribution Tests. Advances in Neural Information Processing Systems, 33, 19523–19533.

[7] Tziouvaras, A., Bertalanic, B., Floros, G., Kolomvatsos, K., Sarigiannidis, P. and Fortuna, C. (2026). A Representation Learning Approach to Feature Drift Detection in Wireless Networks. IEEE Transactions on Emerging Topics in Computing, 14(1), 80–93. doi: 10.1109/TETC.2025.3644604.

[8] Liu, X., Luo, C. and Shao, R. (2026). 5G Mobile Networked Clusters Anomaly Detection Under Distribution Shifts: A Causal Perspective. IEEE Transactions on Cognitive Communications and Networking, 12, 3600–3617. doi: 10.1109/TCCN.2025.3632390.

[9] Alves, P. V. A., et al. (2023). Machine Learning Applied to Anomaly Detection on 5G O-RAN Architecture. Procedia Computer Science, 222, 81–93. doi: 10.1016/j.procs.2023.08.146.

[10] Attanayaka, D., Porambage, P., Liyanage, M. and Ylianttila, M. (2023). Peer-to-Peer Federated Learning Based Anomaly Detection for Open Radio Access Networks. Proceedings of IEEE ICC, 5464–5470. doi: 10.1109/ICC45041.2023.10278993.

[11] Sun, C., Pawar, U., Khoja, M., Foukas, X., Marina, M. K. and Radunovic, B. (2024). SpotLight: Accurate, Explainable and Efficient Anomaly Detection for Open RAN. Proceedings of the 30th ACM International Conference on Mobile Computing and Networking (MobiCom 2024), 923–937. doi: 10.1145/3636534.3649380.

[12] Gudepu, V., Chintapalli, V. R., Castoldi, P., Valcarenghi, L., Tamma, B. R. and Kondepu, K. (2024b). The Drift Handling Framework for Open Radio Access Networks: An Experimental Evaluation. Computer Networks, 243, 110290. doi: 10.1016/j.comnet.2024.110290.

[13] Xu, L., Han, Z., Zhao, D., Li, X., Yu, F. and Chen, C. (2024). Addressing Concept Drift in IoT Anomaly Detection: Drift Detection, Interpretation, and Adaptation. IEEE Transactions on Sustainable Computing, 9(6), 913–924. doi: 10.1109/TSUSC.2024.3386667.

[14] Xavier, B. M., Dzaferagic, M., Collins, D., Comarela, G., Martinello, M. and Ruffini, M. (2023). Machine Learning-Based Early Attack Detection Using Open RAN Intelligent Controller. Proceedings of IEEE ICC, pp. 1856–1861. doi: 10.1109/ICC45041.2023.10279349.

[15] Basaran, O. T. and Dressler, F. (2025). XAInomaly: Explainable and interpretable Deep Contractive Autoencoder for O-RAN traffic anomaly detection. Computer Networks, 261, 111145. doi: 10.1016/j.comnet.2025.111145.

[16] Parada, R., Vilajosana, X., Alfayoumi, S., Serra, J., Font-Bach, O. and Dini, P. (2025). An open testbed for O-RAN experimentation with AI-enabled control and monitoring. Internet of Things, 33, 101729. doi: 10.1016/j.iot.2025.101729.

[17] Azkaei, B., Joshi, K. C. and Exarchakos, G. (2025). Machine Learning-Driven Anomaly Detection for 5G O-RAN Performance Metrics. Proceedings of IEEE INFOCOM Workshops, pp. 1–6. doi: 10.1109/INFOCOMWKSHPS65812.2025.11152997.

[18] Yu, H. and Kim, H. (2026). Performance Analysis of an AI-Based IDS xApp for Cyberattack Anomaly Detection in O-RAN Near-RT RIC. Computer Modeling in Engineering & Sciences, 147(2), Article 52. doi: 10.32604/cmes.2026.082095.

[19] Gudepu, V., Chirumamilla, B., Chintapalli, V. R., Castoldi, P., Valcarenghi, L., Tamma, B. R., Kataria, D. and Kondepu, K. (2024a). GAN-Based Drift and Anomaly Detection for Open Radio Access Networks. Proceedings of IEEE HPSR, pp. 124–129. doi: 10.1109/HPSR62440.2024.10635911.

[20] Gudepu, V., Chirumamilla, B., Chintapalli, V. R., Castoldi, P., Valcarenghi, L., Tamma, B. R. and Kondepu, K. (2025). GEN-DRIFT: Generative AI-driven drift handling for beyond 5G networks. Computer Networks, 263, 111237. doi: 10.1016/j.comnet.2025.111237.

[21] Zhu, J., Cai, S., Deng, F., Ooi, B. C. and Zhang, W. (2023). METER: A Dynamic Concept Adaptation Framework for Online Anomaly Detection. Proceedings of the VLDB Endowment, 17(4), 794–807. doi: 10.14778/3636218.3636233.

[22] Tassie, C., Kim, B., Groen, J., Belgiovine, M. and Chowdhury, K. R. (2024). Leveraging Explainable AI for Reducing Queries of Performance Indicators in Open RAN. Proceedings of IEEE ICC, pp. 5413–5418. doi: 10.1109/ICC51166.2024.10622827.

[23] NetSys Edinburgh (2024). SpotLight: Open RAN Anomaly Detection Dataset. GitHub repository. Available at: https://github.com/netsys-edinburgh/SpotLight (accessed 15 November 2026).

TABLES AND FIGURES

Table 1. SpotLight run inventory.

Run category	Physical runs	Function in nested evaluation
Normal 1 UE	35	24 train / 5 calibration / 6 test in each outer split
Normal 5 UE	20	10 development / 10 test in each outer split
Normal 7 UE–Ping	2	Stress test only
Anomaly sessions	16	8 development / 8 test in each outer split

Table 2. Nested PCA evaluation.

Selector	Mean K	5-UE FPR source	5-UE FPR adapt.	7-UE FPR source	AUPRC	AUROC	Recall source	MCC
AllPlatform	176.1	92.40%	1.77%	41.08%	0.150	0.456	74.71%	−0.069
Disc	37.5	1.94%	1.45%	14.06%	0.200	0.533	21.89%	0.009
Stable	82.5	1.52%	1.14%	23.03%	0.182	0.511	29.89%	−0.026
DAKS	57.5	4.04%	1.39%	25.81%	0.181	0.521	33.45%	−0.007

Table 3. Detector transfer.

Detector	Stable FPR	DAKS FPR	Stable AUPRC	DAKS AUPRC	DAKS AUROC	DAKS MCC
PCA	1.52%	4.04%	0.182	0.181	0.521	−0.007
Isolation Forest	0.86%	1.16%	0.169	0.165	0.501	−0.025
One-Class SVM	1.43%	5.84%	0.167	0.164	0.470	−0.030
Robust Mahalanobis	1.88%	4.18%	0.162	0.164	0.508	−0.021

Table 4. Threshold calibration.

Detector	FPR source	FPR adaptive	Recall source	Recall adaptive	AUPRC
PCA	4.04%	1.39%	33.45%	28.43%	0.181
Isolation Forest	1.16%	0.98%	14.80%	14.14%	0.165

One-Class SVM	5.84%	1.66%	38.46%	34.50%	0.164
Robust Mahalanobis	4.18%	1.40%	33.53%	28.79%	0.164

Table 5. KPI-selection stability.

Selector	Mean selected K	Mean Jaccard	Mean overlap
Disc	37.5	0.369	0.702
Stable	82.5	0.697	0.958
DAKS	57.5	0.376	0.720

Figures

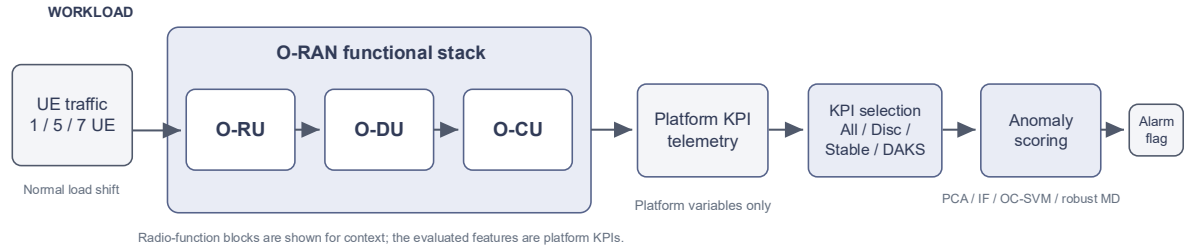

Fig. 1. O-RAN telemetry architecture.

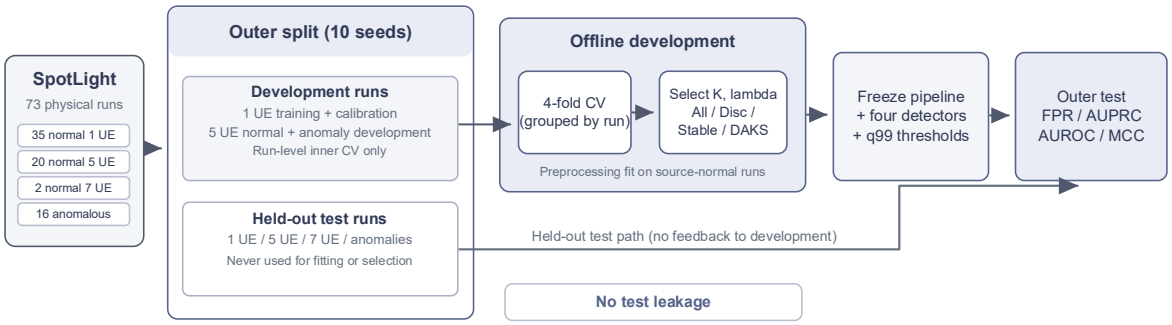

Fig. 2. Run-level nested validation.

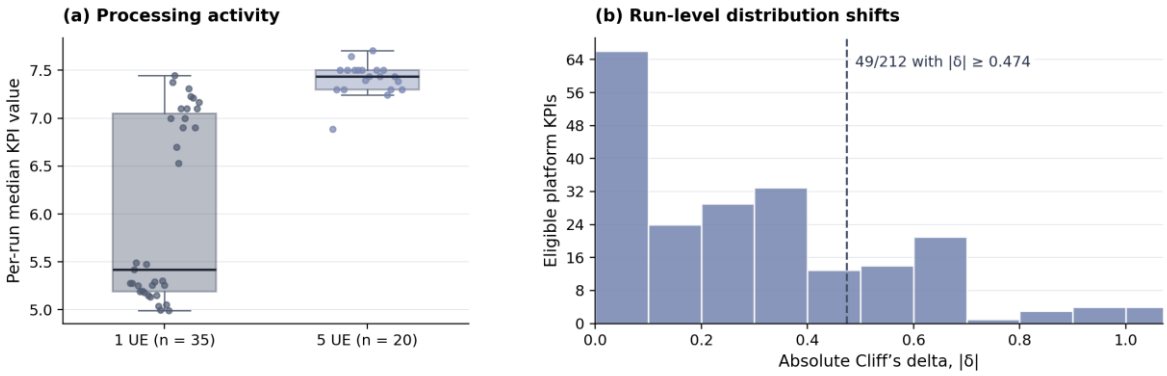

Fig. 3. Workload-induced KPI shifts.

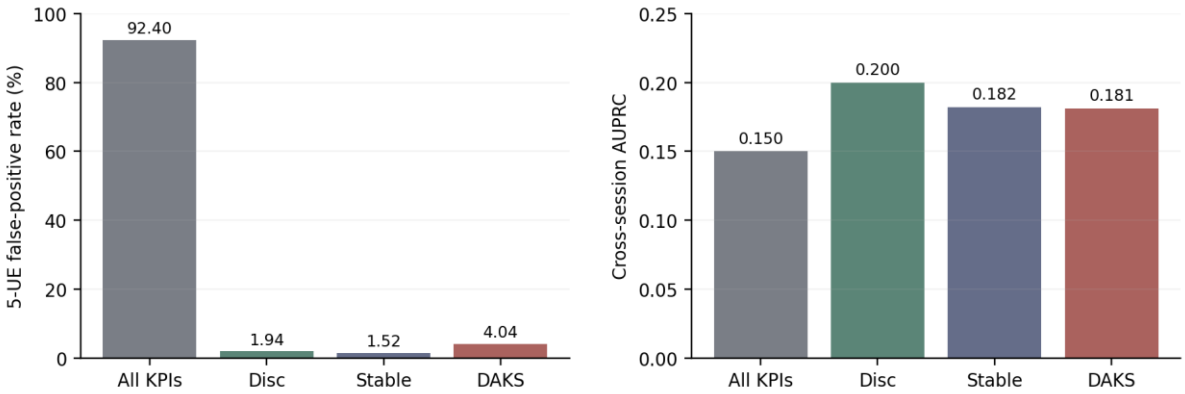

Fig. 4. PCA detector comparison.

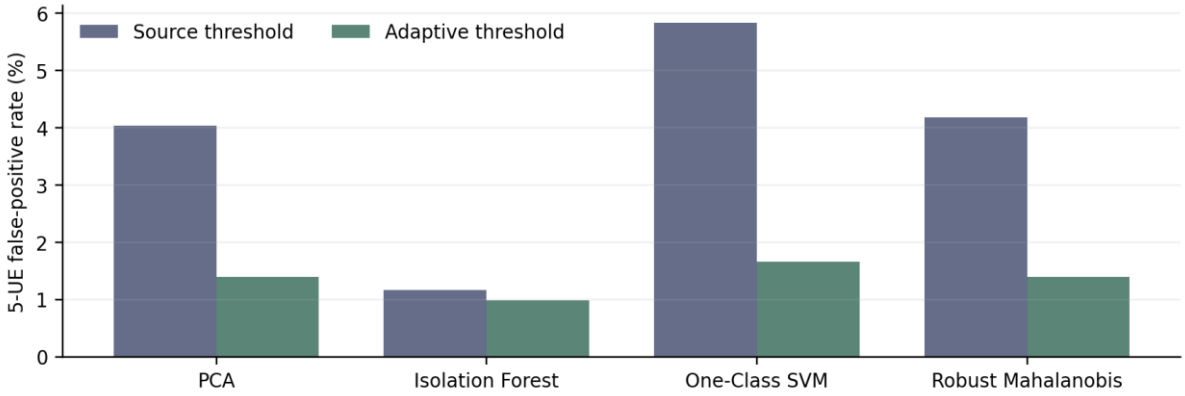

Fig. 5. Threshold calibration effects.