Toward Reliable Intrusion Monitoring for O-RAN Based Networks with Conformal False Alarm Control

Anonymous Authors[1]

Abstract

Autoencoder-based intrusion detectors provide a practical option for Open Radio Access Network (O-RAN) environments because they can be trained from benign traffic and deployed as monitoring components, such as eXtended Applications. However, alarm thresholds of the intrusion detectors are typically determined by heuristic rules, which makes them difficult to integrate into confidence-gated closed-loop management. In this type of management, uncontrolled false alarms may trigger disruptive mitigation actions. In this paper, we propose a covariance-aware conformal detector for unsupervised intrusion detection with a false alarm rate (FAR) constraint. The method uses a Mahalanobis distance in the reconstruction error space of the autoencoder as the nonconformity score and calibrates the threshold by split conformal prediction. Under exchangeability between the benign calibration samples and future benign test samples, the proposed detector controls the FAR within a user-specified budget. For 5th Generation traffic, the proposed score increases the detection rate compared to the conformal baseline at the same target FAR. The quadratic score also decomposes exactly into feature-wise diagnostic terms, providing low-cost per-alarm triage without post-hoc explanation models. The online detector operates with microsecond-scale latency, which enables deployment to Near-Real-Time RAN Intelligent Controller.

1. Introduction

Open Radio Access Network (O-RAN) architectures are essential to 5th Generation (5G) and emerging 6th Generation (6G) deployments. These architectures support intelligent monitoring applications, known as eXtended Applications (xApps), on the Near-Real-Time RAN Intelligent Controller (Near-RT RIC) (Hung et al., 2025; Amachaghi et al., 2024). Autoencoder-based intrusion detectors are particularly well-suited to this monitoring point, because they can be trained exclusively on benign traffic and identify zero-day attacks through reconstruction error without labeled data (Chinnasamy et al., 2025; Kye et al., 2022). Moreover, the simple feedforward structure of autoencoders aligns well with the strict latency requirements of Near-RT RIC.

In 5G/6G environments, intrusion detectors must operate under reliability constraints, yet their thresholds are often selected without formal control over the false alarm rate (FAR). Existing works typically report empirical metrics such as accuracy, precision, recall, or F1-score, without providing a principled threshold-selection rule for a user-specified FAR budget (Amachaghi et al., 2024; Wen et al., 2024b). A similar issue arises in rule-based RAN security systems, where false positives are identified as a fundamental challenge (Wen et al., 2024a). Uncontrolled FAR can have direct consequences in automated security operations, where alerts may trigger Security Orchestration, Automation, and Response (SOAR) workflows for user-equipment disconnection (Gao et al., 2024) or slice-level mitigation actions (Lekidis, 2024). The 3rd Generation Partnership Project (3GPP) TS 23.288 specifies the analytics outputs of the Network Data Analytics Function (NWDAF), which may include confidence information or probability assertions (3GPP, 2024a). 3GPP TS 28.104 places management data analytics within open-loop or closed-loop management systems (3GPP, 2024b). These specifications indicate that the reliability of analytics is relevant to automated network management. In this context, target FAR control provides an operationally meaningful reliability signal for integrating AI-based network intrusion detection systems (NIDS) with confidence-aware management decisions.

Conventional heuristic thresholds, such as fixed percentiles (Borghesi et al., 2019; Ruff et al., 2021) or k-sigma rules (Montgomery, 2020), provide no finite-sample FAR guarantee. In addition, scalar nonconformity scores such as Mean Absolute Error (MAE) treat each feature residual independently (Zong et al., 2018) and can miss joint devi-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ations that characterize certain attack classes (Choi et al., 2025). To address these limitations, we combine split conformal prediction (CP) (Angelopoulos & Bates, 2023) with covariance-aware residual scoring. Split CP calibrates a threshold from held-out benign samples and provides a finite-sample marginal FAR guarantee under exchangeability. As the nonconformity score, we use a Mahalanobis distance in the reconstruction-error space of the autoencoder, with the benign residual covariance estimated by Ledoit-Wolf shrinkage (Ledoit & Wolf, 2004). Since the Mahalanobis score is a scalar nonconformity score, it can be calibrated by split CP without additional overhead. Although the Mahalanobis score has been widely used for anomaly and out-of-distribution detection (Lee et al., 2018), and conformal calibration has been applied to autoencoder-based detection (Cai et al., 2022), existing approaches do not directly address O-RAN-oriented intrusion monitoring with residual-space covariance-aware scoring, target-FAR calibration, and low-cost per-alarm diagnostics.

Our main contributions are summarized as follows:

- **Covariance-aware conformal scoring in reconstruction-error space**: We use a Mahalanobis nonconformity score in the autoencoder reconstruction-error space, capturing inter-feature residual dependencies. Ledoit-Wolf shrinkage stabilizes covariance estimation before precision-matrix inversion without dataset-specific ridge tuning.
- **Target-FAR calibration for O-RAN intrusion monitoring.** We replace heuristic threshold selection with split conformal calibration, which maps the FAR budget to a deployable decision threshold.
- **Closed-form per-alarm diagnostic scores**: The score decomposes exactly into feature-wise terms, giving each alarm a diagnostic signature without post-hoc explanation models.
- **Validation on 5G traffic against Near-RT RIC timing constraints.** We validate the method on a 5G dataset, collected over a 5G wireless testbed. The method improves the detection performance over baselines at the same target FAR budget. The online pipeline is conducted within the 10 ms, supporting its feasibility as an xApp-level monitoring component.

2. Covariance-Aware Conformal Detection Framework

We consider an autoencoder-based NIDS, in which a security analytics component processes per-flow feature vectors extracted from 5G/6G network traffic and reports calibrated alarms to downstream management functions. Let $\mathbf{x} \in \mathbb{R}^d$ denote such a feature vector with d features. An autoencoder $f : \mathbb{R}^d \to \mathbb{R}^d$ is trained on a benign traffic set $\mathcal{D}$. For any sample $\mathbf{x}$, the reconstruction error vector is defined as

$$\mathbf{e}(\mathbf{x}) = \mathbf{x} - f(\mathbf{x}). \tag{1}$$

The proposed framework operates in this reconstruction-error space. A separate held-out benign set, denoted by $\mathcal{H} = \{\mathbf{x}_i\}_{i=1}^{n}$, which is disjoint from D, is reserved for estimation and calibration after training. $\mathcal{H}$ is partitioned into two disjoint subsets, which are estimation set $\mathcal{H}_{\text{est}}$ of size n_{est} and calibration set $\mathcal{H}_{\text{cal}}$ of size $n_{\text{cal}} = n - n_{\text{est}}$.

2.1. Covariance-Aware Nonconformity Score

From $\mathcal{H}_{\text{est}}$, we compute the mean $\boldsymbol{\mu}_e$ and sample covariance $\boldsymbol{\Sigma}_e$ of the reconstruction errors as

$$\boldsymbol{\mu}_e = \frac{1}{n_{\text{est}}} \sum_{i=1}^{n_{\text{est}}} \mathbf{e}_i, \tag{2}$$

$$\boldsymbol{\Sigma}_e = \frac{1}{n_{\text{est}}} \sum_{i=1}^{n_{\text{est}}} (\mathbf{e}_i - \boldsymbol{\mu}_e)(\mathbf{e}_i - \boldsymbol{\mu}_e)^\top, \tag{3}$$

where $\mathbf{e}_i = \mathbf{e}(\mathbf{x}_i)$ for $\mathbf{x}_i \in \mathcal{H}_{\text{est}}$. To stabilize matrix inversion under finite-sample estimation, we apply Ledoit-Wolf shrinkage (Ledoit & Wolf, 2004) as

$$\tilde{\boldsymbol{\Sigma}}_e = (1 - \lambda^*)\boldsymbol{\Sigma}_e + \lambda^* \frac{\operatorname{tr}(\boldsymbol{\Sigma}_e)}{d} \mathbf{I}, \tag{4}$$

where $\lambda^* \in [0, 1]$ is the Ledoit-Wolf shrinkage intensity, $\operatorname{tr}(\cdot)$ denotes the trace operator, and $\mathbf{I}$ is the $d \times d$ identity matrix. The shrinkage intensity is estimated analytically to minimize the expected Frobenius-norm error of the covariance estimator without cross-validation. The regularized precision matrix $\tilde{\boldsymbol{\Sigma}}_e^{-1}$ is obtained by inverting Eq. (4). Unlike a fixed ridge term, Ledoit-Wolf shrinkage determines the shrinkage intensity from the residual samples themselves, adapting the covariance estimate to the available sample size without requiring dataset-specific tuning.

Let $\mathbf{v} = \mathbf{e}(\mathbf{x}) - \boldsymbol{\mu}_e$ for any sample $\mathbf{x}$. We define the nonconformity score as

$$s(\mathbf{x}) = \mathbf{v}^\top \tilde{\boldsymbol{\Sigma}}_e^{-1} \mathbf{v} \tag{5}$$

where $s(\mathbf{x}) \geq 0$ always. Note that $s(\mathbf{x})$ is the squared Mahalanobis distance in reconstruction-error space relative to the benign residual statistics. Directions with high benign-error variance are given less weight, while directions that are rarely seen in benign traffic are given more weight by the precision matrix. This amplifies anomalous joint deviations even when no single feature error is individually large. When $\tilde{\boldsymbol{\Sigma}}_e^{-1} \propto \mathbf{I}$, all residual dimensions are weighted equally, and the score reduces to a scaled squared Euclidean norm of the reconstruction-error vector.

2.2. False Alarm Control via Conformal Prediction

Let $s_{(m)}$ be the m-th smallest calibration score for the calibration samples $\{\mathbf{x}_i\}_{i=1}^{n_{\text{cal}}} \subset \mathcal{H}_{\text{cal}}$, i.e.,

$$s_{(1)} \leq \cdots \leq s_{(n_{\text{cal}})} \tag{6}$$

which are the ordered nonconformity scores. Given a target FAR budget $\beta \in (0, 1)$, the split conformal threshold is given by

$$\tau(\beta) = s_{(m)}, \tag{7}$$

where

$$m = \lceil (1 - \beta)(n_{\text{cal}} + 1) \rceil . \tag{8}$$

A sample $\mathbf{x}$ is flagged as an intrusion if and only if $s(\mathbf{x}) > \tau(\beta)$.

By the finite-sample guarantee of split CP (Vovk et al., 2005; Angelopoulos & Bates, 2023), if the calibration samples and a benign test sample are exchangeable, then

$$\Pr\Big(s(\mathbf{x}) > \tau(\beta)\Big) \leq \beta, \tag{9}$$

where Pr denotes the probability. In 5G/6G environments, this assumption may be weakened by temporal drift, including diurnal traffic variation, service changes, mobility patterns, or software updates. In such cases, recalibration with recent benign traffic can help recover the conformal guarantee, provided that the updated calibration samples are exchangeable with future benign test samples.

2.3. Per-Alarm Score Decomposition

When $s(\mathbf{x}) > \tau(\beta)$, the anomaly score is decomposed into per-feature contributions to identify which features are driving the alarm. Let $\mathbf{u} = \tilde{\mathbf{\Sigma}}_e^{-1}\mathbf{v}$ be the precision-weighted residual vector where v_j and u_j denote the j-th components of $\mathbf{v}$ and $\mathbf{u}$, respectively. We define the score-decomposition term for feature j as

$$\phi_j(\mathbf{x}) = v_j u_j = v_j \left[\tilde{\mathbf{\Sigma}}_e^{-1}\mathbf{v}\right]_j. \tag{10}$$

The decomposition is exact because the feature-wise terms sum to the original quadratic score as follows

$$\sum_{j=1}^{d} \phi_j(\mathbf{x}) = \mathbf{v}^\top \mathbf{u} = s(\mathbf{x}). \tag{11}$$

It provides an exact feature-wise decomposition of the anomaly score without post-hoc approximation. Since the precision matrix mixes correlated residual features, individual terms $\phi_j(\mathbf{x})$ may be negative. Therefore, we interpret them as score-decomposition terms for triage rather than causal feature attributions. Note that this Online scoring requires $O(d^2)$ operations per sample, dominated by the precision matrix–vector product, while $\boldsymbol{\mu}_e$ and $\tilde{\mathbf{\Sigma}}_e^{-1}$ are computed once offline before deployment.

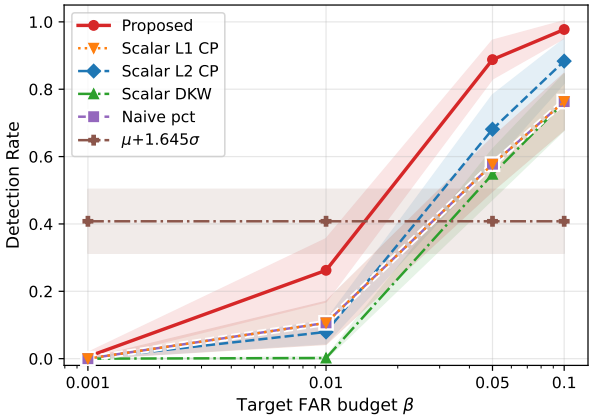

Figure 1. Detection rate across target FAR budget β (Shaded band: ± 1 standard deviation over 10 seeds).

3. Experiments

We evaluate the proposed framework by addressing deployment-relevant perspectives. First, we examine whether covariance-aware residual scoring improves the detection rate over baselines under the same autoencoder backbone and also measure whether the online detector satisfies Near-RT RIC timing constraints. Second, we test whether split conformal calibration controls FAR on held-out benign 5G traffic across target budgets. Third, we examine the per-alarm diagnostic signatures obtained from the quadratic score decomposition.

3.1. Experimental Setup

We evaluate the proposed method on the 5G Wireless Network Intrusion Detection Dataset (5G-NIDD) (Samarakoon et al., 2022). 5G-NIDD contains real network traffic collected over a 5G wireless testbed infrastructure, providing a deployment-relevant setting for modern mobile networks. The dataset contains 431,650 flows, including 227,571 benign and 204,079 attack flows. It spans five attack types, such as ICMPFlood, SYNScan, TCPConnectScan, UDPScan, and UDPFlood. Benign samples are split into a training set and a holdout set with a 60/40 ratio. The holdout set is partitioned into $\mathcal{H}_{\text{est}}$ ($n_{\text{est}} = 2{,}000$) for covariance estimation, $\mathcal{H}_{\text{cal}}$ ($n_{\text{cal}} = 69{,}000$) for threshold calibration, and a benign test set ($n_{\text{test}} = 20{,}000$) used for empirical FAR evaluation. All attack samples are excluded from training, covariance estimation, and calibration, and are used only for detection-rate evaluation. The autoencoder has a symmetric $69 \to 48 \to 32 \to 48 \to 69$ architecture with ReLU activations and dropout after the first encoder and decoder layers. It is trained on benign data with a batch size of 256. Experiments are repeated over 10 random seeds.

We compare against five baselines. Scalar L1 CP applies split CP (Angelopoulos & Bates, 2023) to the scalar MAE

Table 1. Comparisons of Empirical FAR (%).

β	Proposed	Scalar L1 CP	Scalar L2 CP	Scalar DKW	Naive pct	μ+1.645σ
0.001	0.107 ± 0.029	0.108 ± 0.038	0.105 ± 0.029	0.000 ± 0.000	0.110 ± 0.039	3.076 ± 0.488
0.010	1.012 ± 0.080	1.010 ± 0.069	1.011 ± 0.067	0.502 ± 0.073	1.012 ± 0.070	3.076 ± 0.488
0.050	4.995 ± 0.246	5.011 ± 0.280	4.965 ± 0.273	4.511 ± 0.278	5.013 ± 0.280	3.076 ± 0.488
0.100	9.965 ± 0.276	9.990 ± 0.270	10.002 ± 0.277	9.454 ± 0.289	9.991 ± 0.270	3.076 ± 0.488

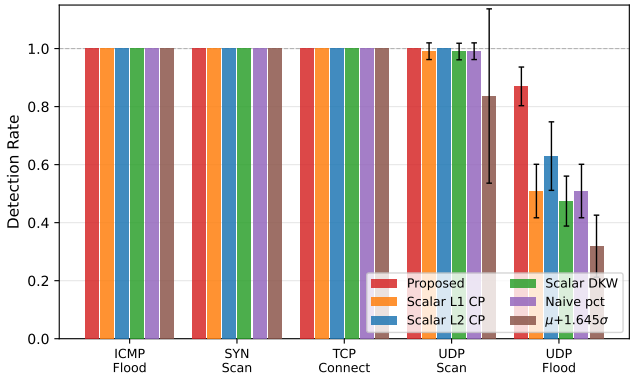

Figure 2. Per-attack detection rate at β=0.05. Error bars show $\pm$1 standard deviation

score. Scalar L2 CP applies split CP to the squared Euclidean norm of the residual vector, corresponding to the limit case of the proposed score when $\hat{\mathbf{\Sigma}}_e^{-1} \propto \mathbf{I}$. This baseline isolates the contribution of covariance structure beyond scalar residual aggregation. Scalar DKW applies a Dvoretzky-Kiefer-Wolfowitz (DKW)-based conservative threshold (Umsonst et al., 2022) to the MAE score with confidence parameter $\alpha = 0.05$. Naive percentile thresholds the MAE score at the empirical $(1-\beta)$ quantile without the conformal rank correction. $\mu + 1.645\sigma$ thresholds the MAE score at $\mu + 1.645\sigma$, corresponding to the 5% upper tail under a Gaussian assumption.

3.2. Intrusion Detection Performance and Latency

Figure 1 shows the detection rate as a function of the target FAR budget β. The proposed scoring consistently improves over scalar reconstruction-error scoring across the evaluated range. At $\beta = 0.05$, it achieves detection rate $88.8\% \pm 5.7\%$, compared with $57.7\% \pm 7.9\%$ for Scalar L1 CP and $68.4\% \pm 10.4\%$ for Scalar L2 CP, corresponding to gains of 31.1 and 20.4 pp, respectively. As shown in Figure 2, the performance gain is mainly driven by UDPFlood, which comprises 86% of attack samples. On UDPFlood, the proposed method improves the detection rate from 50.9% to 87.0%. The improvement remains visible under uniform weighting across attack types, where the macro-average detection rate increases from 90.0% to 97.4%. These results suggest that scalar reconstruction-error scores can miss useful structure in the residual vector. By accounting for feature-wise residual variance and cross-feature residual correlations, the Mahalanobis score provides a stronger nonconformity measure under the same conformal calibration procedure.

Considering resource-constrained deployment platforms, all latency measurements were performed using CPU-only inference with an Intel(R) Core(TM) i5-14400F CPU at 2.50 GHz and 8 GB RAM. Without hardware acceleration, the online pipeline, including the autoencoder forward pass and Mahalanobis scoring, completes in 56 μs on average and 137 μs at the 99th percentile for $d = 69$, measured over 10,000 single-sample trials. This provides approximately 179$\times$ headroom at the average and 73$\times$ at the 99th percentile relative to the 10 ms Near-RT RIC control cycle (Hung et al., 2025).

3.3. False Alarm Rate Certification

Table 1 reports empirical FAR on an independent benign test set that is disjoint from calibration. The proposed method, Scalar L1 CP, and Scalar L2 CP track the target FAR budget closely across all evaluated β values. Small deviations around the target are expected because the conformal guarantee in Eq. (9) is marginal for a future benign test sample, whereas the table reports a finite-sample estimate over $n_{\text{test}} = 20{,}000$ benign samples. The Naive percentile baseline is numerically close to the CP methods because the conformal rank correction changes the nominal quantile level only by $O(1/n_{\text{cal}})$. The distinction is theoretical but important. Split CP selects the $\lceil (1-\beta)(n_{\text{cal}}+1) \rceil$-th order statistic and provides a finite-sample marginal FAR guarantee under exchangeability, whereas the naive empirical quantile does not. The $\mu + 1.645\sigma$ heuristic has an approximately fixed FAR because its threshold does not depend on the target budget. It exceeds tight budgets ($\beta = 0.001$ and 0.010) and underuses the looser budgets ($\beta = 0.05$ and 0.10). Scalar DKW remains systematically below the target FAR, confirming its conservative behavior. This conservativeness follows from the DKW correction $\epsilon_{n_{\text{cal}}} = \sqrt{\log(2/\alpha)/(2n_{\text{cal}})}$, which is approximately 0.0052 for $\alpha = 0.05$ and $n_{\text{cal}} \approx 69{,}000$. At $\beta = 0.001$, where $\epsilon_{n_{\text{cal}}} > \beta$, the DKW threshold is set above the calibration score range, so the detection rate becomes zero.

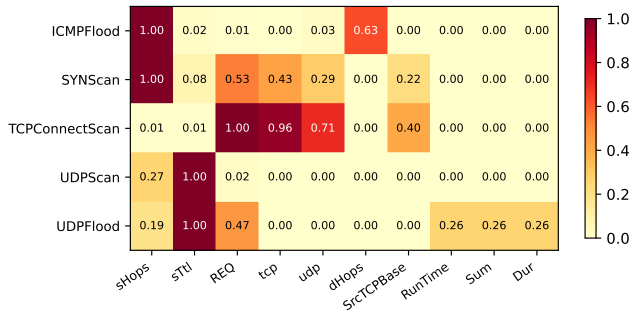

Figure 3. Per-attack score-decomposition terms $|\phi_j|$ for the top-10 features. Each row is normalized by its row maximum.

3.4. Per-Attack Score Decomposition

The completeness identity in Eq. (11) holds for any $\mathbf{v}$ and $\tilde{\boldsymbol{\Sigma}}_e^{-1}$, so each flagged sample carries an exact per-feature breakdown of its anomaly score at no additional cost. Figure 3 shows the normalized mean absolute score-decomposition terms $|\phi_j|$ per attack type for the top-10 features.

The dominant score-decomposition terms differ across attack classes. Hop-count fields (`sHops`, `dHops`) dominate ICMPFlood. For SYNScan, hop-count and protocol-indicator features (`sHops`, `tcp`, `REQ`) consistently rank among the largest terms. `tcp` and `REQ` are prominent for TCPConnectScan. While `sTtl` is the largest term for UDPScan across most seeds, UDPFlood does not show a single consistently dominant feature across seeds. Instead, multiple flow-level features receive comparable diagnostic scores. While `sTtl` is the largest term for UDPScan across most seeds, the

To assess whether the feature rankings are stable across training runs, we repeat the score-decomposition analysis over 10 independent random seeds. The mean pairwise Spearman rank correlation of the per-attack score-decomposition vectors is 0.870 ± 0.021, indicating that the overall feature-ranking structure is largely consistent across seeds. The single top-ranked feature can alternate among closely scored features when several features have similar decomposition magnitudes, while the broader high-contribution feature set remains similar. These per-attack differences suggest that covariance weighting highlights attack-class-specific residual structure that is less visible under scalar MAE aggregation. Consequently, each alarm is accompanied by a ranked list of score-decomposition terms that can support operator triage without post-hoc explanation models.

4. Conclusion

We present a covariance-aware conformal framework for autoencoder-based intrusion detection in O-RAN-based networks. The proposed method combines the Mahalanobis score in reconstruction-error space and split conformal threshold calibration to replace heuristic thresholds with a target FAR calibration rule. The Mahalanobis score accounts for residual dependencies, and split conformal calibration ties the alarm threshold to a FAR budget. Experiments on 5G-NIDD demonstrate that the method improves the detection rate over conformal baselines while satisfying the latency requirement of Near-RT RIC. Experiments on 5G-NIDD demonstrate that the method improves detection performance over conformal baselines while satisfying the latency requirement for Near-RT RIC. The quadratic score also provides an exact feature-wise decomposition, enabling low-cost per-alarm diagnostics without post-hoc explanation models. Future work will study how to maintain calibration under traffic drift in non-stationary O-RAN-based networks.

References

3GPP. Architecture enhancements for 5G system (5GS) to support network data analytics services. Technical Specification TS 23.288, 3rd Generation Partnership Project (3GPP), 2024a. Release 18.

3GPP. Management data analytics (MDA). Technical Specification TS 28.104, 3rd Generation Partnership Project (3GPP), 2024b. Release 18.

Amachaghi, E. N., Shojafar, M., Foh, C. H., and Moessner, K. A survey for intrusion detection systems in open RAN. *IEEE Access*, 12:88146–88173, 2024.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Borghesi, A., Bartolini, A., Lombardi, M., Milano, M., and Benini, L. Anomaly detection using autoencoders in high performance computing systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9428–9433, 2019.

Cai, F., Ozdagli, A. I., and Koutsoukos, X. Variational autoencoder for classification and regression for out-of-distribution detection in learning-enabled cyber-physical systems. *Applied Artificial Intelligence*, 36(1):2131056, 2022.

Chinnasamy, R., Subramanian, M., Easwaramoorthy, S. V., and Cho, J. Deep learning-driven methods for network-based intrusion detection systems: A systematic review. *ICT Express*, 11(1):181–215, 2025.

Choi, D., Rheey, J., and Park, H. Attack-specific feature analysis framework for NetFlow IoT datasets. *Computers & Security*, 157:104536, 2025.

Gao, S., Lin, R., Fu, Y., Li, H., and Cao, J. Security threats, requirements and recommendations on creating 5G network slicing system: A survey. *Electronics*, 13(10):1860, 2024.

Hung, C.-F., Tseng, C.-H., and Cheng, S.-M. Anomaly detection for mitigating xapp and e2 interface threats in o-ran near-rt ric. *IEEE Open Journal of the Communications Society*, 2025.

Kye, H., Kim, M., and Kwon, M. Hierarchical detection of network anomalies: A self-supervised learning approach. *IEEE Signal Processing Letters*, 29:1908–1912, 2022.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of Advances in Neural Information Processing Systems*, volume 31, 2018.

Lekidis, A. Towards 5G advanced network slice assurance through isolation mechanisms. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pp. 1–7, 2024.

Montgomery, D. C. *Introduction to statistical quality control*. John wiley & sons, 2020.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Samarakoon, S., Siriwardhana, Y., Porambage, P., Liyanage, M., Chang, S.-Y., Kim, J., Kim, J., and Ylianttila, M. 5G-NIDD: A comprehensive network intrusion detection dataset generated over 5G wireless network, 2022.

Umsonst, D., Ruths, J., and Sandberg, H. Finite sample guarantees for quantile estimation: An application to detector threshold tuning. *IEEE Transactions on Control Systems Technology*, 31(2):921–928, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Boston, MA: Springer, 2005.

Wen, H., Porras, P. A., Yegneswaran, V., Gehani, A., and Lin, Z. 5G-SPECTOR: An O-RAN compliant layer-3 cellular attack detection service. In *Proceedings of the 31st Annual Network and Distributed System Security Symposium*, 2024a.

Wen, H., Sharma, P., Yegneswaran, V., Porras, P., Gehani, A., and Lin, Z. 6G-XSec: Explainable edge security for emerging OpenRAN architectures. In *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, pp. 77–85, 2024b.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *Proceedings of International Conference on Learning Representations*, 2018.