Wadi Alshatti University Journal of Pure and Applied Sciences

مجلة جامعة وادي الشاطئ للعلوم البحتة والتطبيقية

Volume 4, No. 2, July-December 2026 Online ISSN: 3006-0877 المجلد 4، الاصدار 2، يوليو- ديسمبر 2026

RESEARCH ARTICLE COMPUTER SCIENCE AND CYBERSECURITY

Bidirectional Cross-Dataset Generalization and Label-Efficient Adaptation for 5G Network Intrusion Detection

Abraheem Sulayman Alsaedi Abraheem[1,*], Albussiry Alhussin Ali Edhirig[2]

[1]Libyan Center for Studies & Researches in Environmental Science & Technology, Brack Alshatii, Libya
[2]Almadar Aljadid Company, Libya

ARTICLE HISTORY
Received 29 May 2026
Revised 20 July 2026
Accepted 31 July 2026
Online 06 August 2026

KEYWORDS
5G intrusion detection;
Cross-dataset generalization;
Dataset fingerprint;
Domain shift;
Label-efficient adaptation;
Active learning;
Open RAN security.

A B S T R A C T

Machine-learning-based intrusion detection in 5G networks is often assessed using a single dataset, where strong test results may indicate regularities tied to that dataset rather than attack behavior that generalizes across settings. This study examines bidirectional cross-dataset generalization between 5G-NIDD and NetsLab-5GORAN-IDD through conservative semantic matching of denial-of-service, flooding, and scanning attacks. The assessment uses five-seed zero-shot transfer, a multivariate dataset-fingerprint audit, unsupervised correlation alignment, few-shot target adaptation, and class-blind target-label selection while varying attack prevalences in a controlled manner at 1%, 5%, and 10%. Zero-shot XGBoost performance was highly dependent on both the specific task and the direction of transfer. For matched aggregate tasks, balanced accuracy was 0.518 and 0.505 for application-layer denial-of-service, 0.807 and 0.500 for network/transport flooding, and 0.834 and 0.558 for scanning when transferring from 5G-NIDD to NetsLab and in the reverse direction, respectively. A dataset-source classifier reached balanced accuracy of 0.993 when using all 15 harmonized numerical features and maintained 0.993 after source-predictiveness-aware Top-6 filtering, indicating that dataset identity persisted across multivariate feature interactions. CORAL did not eliminate this fingerprint, whereas using limited target labels produced much larger improvements; in the exploratory adaptation experiment, the mean best gain with no more than 5% labeled target data was 0.346 balanced-accuracy points. Under deployment-like prevalence pressure, Random or Diversity-based querying provided the best overall strategy in 13 of 15 prevalence-budget conditions, while source-model uncertainty sampling was unstable. Overall, these findings indicate that realistic 5G intrusion detection depends on explicit cross-dataset evaluation and limited, target-specific supervision rather than on performance from a single dataset, feature filtering, or straightforward distribution alignment.

التعميم ثنائي الاتجاه عبر مجموعات البيانات والتكيف الكفؤ في استخدام التسميات لكشف التسلل في شبكات الجيل الخامس

إبراهيم سليمان الساعدي إبراهيم[1,*]، البصيري الحسين علي إدحيريج[2]

الكلمات المفتاحية
كشف التسلل في شبكات الجيل الخامس
التعميم عبر مجموعات البيانات
بصمة مجموعة البيانات
انزياح المجال
التكيف الكفؤ في استخدام التسميات
التعلم النشط
أمن شبكات الوصول الراديوي المفتوحة

الملخص

غالبًا ما يُقيَّم كشف التسلل بالتعلم الآلي في شبكات الجيل الخامس باستخدام مجموعة بيانات واحدة، مما قد يجعل النتائج المرتفعة تعكس خصائصها بدلًا من أنماط هجوم قابلة للتعميم. تبحث هذه الدراسة التعميم ثنائي الاتجاه بين 5 G-NIDD و NetsLab-5GORAN-IDD عبر مطابقة دلالية متحفظة لهجمات حجب الخدمة والإغراق والمسح. وشمل التقييم نقلًا صفريًا بخمس بذور، وتدقيقًا متعدد المتغيرات لبصمة البيانات، ومحاذاة ارتباط غير خاضعة للإشراف، وتكيفًا محدود العينات، واختيار تسميات محايدًا للفئات، مع ضبط انتشار الهجمات عند 1% و5% و10%. اعتمد أداء XGBoost في النقل الصفري على المهمة واتجاه النقل. ففي المهام التجميعية المتطابقة، بلغت الدقة المتوازنة 0.518 و0.505 لحجب الخدمة على طبقة التطبيق، و0.807 و0.500 لإغراق الشبكة أو النقل، و0.834 و0.558 للمسح، عند النقل من 5 G-NIDD إلى NetsLab وبالعكس. وحقق مصنف مصدر البيانات دقة متوازنة قدرها 0.993 باستخدام السمات العددية الخمس عشرة المنسقة، وحافظ عليها بعد ترشيح أفضل ست سمات وفق قدرتها على التنبؤ بالمصدر، مما يدل على بقاء هوية مجموعة البيانات ضمن التفاعلات متعددة المتغيرات. لم تُزل CORAL هذه البصمة، بينما أدى استخدام تسميات محدودة من الهدف إلى تحسينات أكبر؛ إذ بلغ متوسط أفضل تحسن باستخدام ما لا يزيد على 5% من بيانات الهدف الموسومة 0.346 نقطة. وفي ظروف تحاكي التشغيل الفعلي، كان الاستعلام العشوائي أو القائم على التنوع الأفضل في 13 من أصل 15 حالة انتشار–ميزانية، بينما كان أخذ العينات القائم على عدم يقين نموذج المصدر غير مستقر. وتؤكد النتائج أن الكشف الواقعي للتسلل يتطلب تقييمًا عبر مجموعات البيانات وإشرافًا محدودًا خاصًا بالهدف، بدلًا من الاعتماد على أداء مجموعة واحدة أو ترشيح السمات أو المحاذاة المباشرة للتوزيعات.

*Corresponding author
https://doi.org/10.63318/waujpasv4i2_24
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Introduction

Fifth-generation mobile networks integrate virtualization, edge computing, service-based core functions, network slicing, and, increasingly, open and programmable radio-access components. These capabilities enhance flexibility and enable clearer service differentiation, yet they also widen the security attack surface and raise the demand for data-driven monitoring. The 5G-NIDD dataset was first released in 2022 from a functional 5G test network [1]. A later descriptor characterized it as a labeled resource intended for machine-learning-based intrusion detection [2]. More recently, NetsLab-5GORAN-IDD was gathered from a physical 5G Open RAN testbed and delivers higher-layer network observations along with radio- and device-level measurements [3]. For 5G systems, the security architecture and associated procedures are standardized in 3GPP TS 33.501 [4]. Guidance for Open RAN further highlights risk-based threat modeling for open interfaces and disaggregated components [5]. Together, these resources and standards allow intrusion detection to be studied under contemporary mobile-network conditions.

Even so, the mere availability of realistic datasets does not, on its own, ensure realistic model evaluation. Most supervised intrusion-detection pipelines presume that the training and test records are drawn from the same underlying distribution. Random record-level splits fulfill this assumption by design, but real-world deployments typically break it. This mismatch constitutes dataset shift, where the joint or marginal distributions at deployment differ from those observed during model development [6]. Changes between the development and target networks may occur in traffic composition, user behavior, attack implementation, testbed configuration, collection point, flow exporter, and the preprocessing pipeline. Likewise, surveys of NIDS datasets indicate that the recording environment, traffic generation, labeling, and feature extraction differ substantially across benchmarks [7]. As a result, a model may reach very strong in-domain performance while learning shortcuts tied to a particular capture process rather than to stable attack characteristics.

Cross-dataset investigations have shown how serious this issue can be in conventional network-intrusion benchmarks. Cantone et al. reported near-perfect same-dataset results, but performance that was close to random guessing for many train-test pairings across independently collected datasets [8]. Their results suggest that overlapping semantic labels alone are insufficient when the underlying feature distributions and data-generation processes are not aligned. This concern is especially relevant for 5G intrusion detection because public datasets are relatively limited and their collection architectures are heterogeneous. In this study, 5G-NIDD flow fields are generated using an Argus-based pipeline, whereas the network-flow component of NetsLab-5GORAN-IDD is based on Zeek. Therefore, observed discrepancies should be understood as dataset or domain shift that jointly reflects environmental, collection, and extractor effects, rather than as a purely environmental shift.

Our earlier work on 5G-NIDD revealed a related internal-validity concern. Achieving near-perfect record-level results relied heavily on acquisition-order metadata, particularly sequence and offset variables; eliminating these features caused a substantial drop in holdout performance [9]. That finding demonstrated that leakage-aware preprocessing is necessary, but it did not determine whether the remaining flow features generalize to a separately collected 5G dataset. The subsequent methodological question is therefore external: once direct identifiers and known acquisition-order shortcuts are removed, do models learn attack behaviors that transfer, or do they instead learn a different set of dataset-specific fingerprints?

This question cannot be resolved by simply training on all attacks in one dataset and testing on all attacks in a different dataset. The attack taxonomies are not the same, so an unfavorable outcome could stem from label mismatch rather than from domain shift. We therefore build conservative semantic correspondences for attacks that can be compared directly, and we aggregate families, including application-layer denial-of-service, network/transport flooding, and scanning. Transfer is assessed in both directions because the source-to-target ordering itself is part of the difficulty. A detector may generalize from one dataset to another yet fail when applied in the reverse direction, particularly when class structure, coverage, or measurement conventions differ asymmetrically.

The study then disentangles diagnosis from mitigation. First, zero-shot transfer measures the external generalization gap. Second, a source-predictiveness audit tests whether the harmonized features encode the identity of the dataset they originate from. Third, CORrelation ALignment (CORAL), which matches source and target second-order statistics without using target labels [10], serves as a transparent unsupervised adaptation baseline. Fourth, small amounts of labeled target data are added to evaluate whether label-efficient adaptation is more effective than distribution alignment alone. Finally, target-label selection is examined under controlled attack prevalences of 1%, 5%, and 10%, where the task involves not only classification but also uncovering rare attacks under a limited annotation budget.

The frozen results show a consistent pattern. When semantics are matched, interpretability improves, but cross-dataset failure is not eliminated. For application-layer denial-of-service, aggregate XGBoost keeps balanced accuracy close to chance in both directions, whereas flooding and scanning exhibit substantial directional asymmetries. The identity of the dataset is recovered with near-perfect accuracy from the harmonized numerical features, including after feature filtering that accounts for source predictiveness. CORAL aligns covariance structure but does not remove the multivariate dataset fingerprint. In contrast, limited supervision on the target yields large improvements for multiple task-direction combinations. Under deployment-like prevalence stress, simple Random and Diversity-based querying outperform source-model uncertainty sampling, and increasing the complexity of adaptive pilot allocation does not provide a reliable benefit.

Accordingly, this paper presents cross-dataset 5G intrusion detection as a label-efficient adaptation problem in the presence of persistent dataset fingerprints. The goal is not to argue for a universally transferable classifier; instead, it is to offer a controlled evaluation protocol, determine where transfer breaks down, and measure how much target-specific information is needed to regain useful performance. This separation matters for both scientific communication and deployment planning: strong accuracy on a single benchmark indicates discrimination within the original domain, not robustness to a new 5G network or measurement pipeline. Figure 1 summarizes the distinction between conventional

same-dataset validation and the bidirectional cross-dataset evaluation adopted in this study.

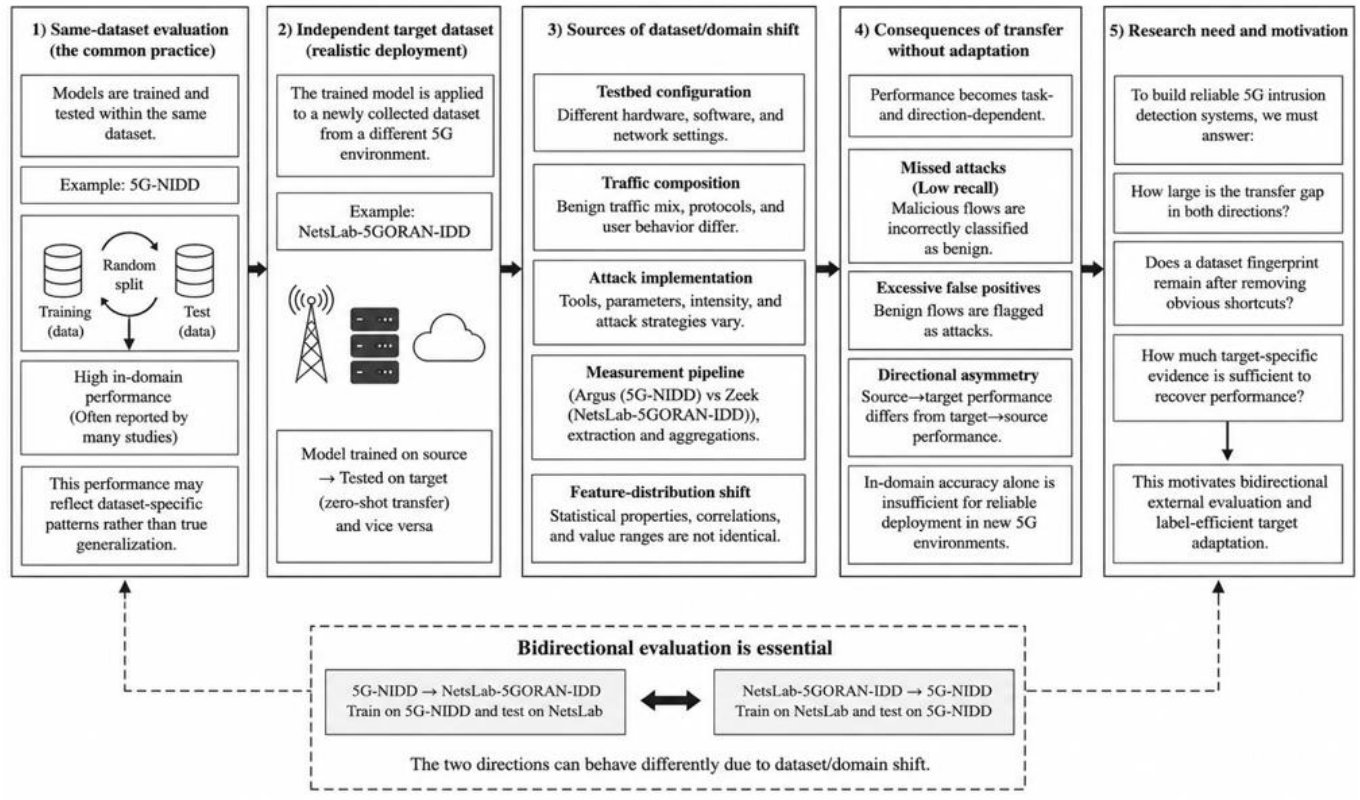

Figure 1: Conceptual framework for bidirectional cross-dataset evaluation in 5G intrusion detection.

Main Contributions

1. A conservative bidirectional cross-dataset benchmark for 5G intrusion detection. The study aligns comparable flow variables and verifies attack semantics before assessing transfer between 5G-NIDD and NetsLab-5GORAN-IDD in both source-to-target and target-to-source directions.

2. A direct multivariate audit of dataset fingerprints. Dataset-source classification and source-predictiveness-aware feature screening indicate that dataset identity remains strongly encoded even after excluding direct identifiers, sequence metadata, and highly source-predictive individual features.

3. **A transparent comparison of unsupervised and label-efficient adaptation.** CORAL is compared with direct transfer and weighted few-shot target adaptation, using disjoint target adaptation and test subsets across five random seeds.

4. **A deployment-oriented analysis of target-label selection.** Random, Diversity, uncertainty, and dual-stratum pilot strategies are evaluated at 1%, 5%, and 10% target attack prevalence, with annotation budgets spanning 20 to 500 records.

5. **A collection of negative and boundary results that limit defensible claims.** The study reports when feature filtering, covariance alignment, source-model uncertainty, and progressively more complex pilot allocation fail to deliver universal gains, and it explicitly separates dataset or domain shift from environmental shift alone.

Paper Organization

Section II reviews 5G intrusion datasets, cross-dataset generalization, domain adaptation, and label-efficient target selection. Section III details dataset provenance, attack matching, feature harmonization, sampling, models, software, and statistical procedures. Section IV presents zero-shot transfer and dataset-fingerprint findings. Section V evaluates CORAL and few-shot target adaptation. Section VI covers deployment-like prevalence and target-selection experiments. Section VII discusses implications, limitations, and threats to validity, and Section VIII concludes the paper. An unnumbered Declarations section and Appendix A appear before the references.

Related Work

A. 5G and Open RAN Intrusion Datasets

Moving beyond conventional enterprise benchmarks toward data gathered from operationally structured 5G testbeds is crucial for assessing mobile-network intrusion detection. 5G-NIDD was produced within a functional 5G network and includes benign traffic as well as scanning, flooding, application-layer denial-of-service, and credential attacks [1]. Its subsequent descriptor characterizes the dataset and its public flow representation [2]. NetsLab-5GORAN-IDD was later collected from a physical 5G Open RAN testbed and offers a multimodal perspective that includes higher-layer network traffic, radio measurements, and device-level observations [3]. This study uses the aggregated network-flow component of that dataset, whose fields are derived from Zeek. As a result, the datasets overlap semantically yet differ in topology, traffic generation, measurement location, extractor, feature definitions, and label taxonomy.

The benchmark analysis based on 5G-NIDD is mainly in-domain [1]. The published NetsLab-5GORAN-IDD benchmark is likewise evaluated within its own collection [3]. While these experiments support model development, they do not determine whether the classifier has learned portable attack behavior or reflects characteristics of a particular capture pipeline. In addition, even an attack family

that is nominally identical may be implemented differently across testbeds. Therefore, direct cross-dataset comparisons require both feature harmonization and conservative semantic matching; otherwise, domain shift and label mismatch become intertwined.

B. Cross-Dataset Generalization and Feature Standardization

The lack of standardized feature definitions has long constrained cross-dataset evaluation of network intrusion detection systems. Sarhan et al. suggested standardized NetFlow feature sets to enable fair comparison across different heterogeneous public datasets [11]. A later assessment indicated that shared feature representations can enhance comparability and, in some cases, improve generalization; however, they do not ensure transfer when the underlying distributions still differ [12]. Explainable cross-domain analysis has also demonstrated that the ordering of source and target can lead to substantial performance imbalances, and that features predictive in one dataset may fail to retain their class relationship in another [13].

Cantone et al. carried out a broad cross-dataset investigation and reported that high same-dataset scores frequently collapse when independently collected datasets are swapped as the source and the target [8]. This finding motivates two design principles used in this work. First, transfer should be assessed bidirectionally rather than relying on a single convenient source-to-target direction. Second, aggregate accuracy by itself is insufficient; attack recall, false-positive rate, MCC, and class-balanced measures are required, since a transferred detector may reduce to predicting only one class. Our earlier 5G-NIDD study adds an additional internal-validity perspective: acquisition-order metadata can yield nearly perfect record-level results while failing to reflect transferable behavior [9]. The current study extends this leakage-aware analysis from internal holdout evaluation to an independently collected 5G dataset.

C. Domain Adaptation and Target-Specific Supervision

Domain adaptation aims to reduce the mismatch between a labeled source distribution and a target distribution. Domain-adaptation theory connects target error to source error, domain divergence, and the existence of a hypothesis that performs well across both domains [14]. CORAL serves as a transparent unsupervised baseline that aligns the source and target covariance matrices without using target labels [10]. In network intrusion detection, more sophisticated methods have learned domain-invariant representations using adversarial training. DI-NIDS, for instance, integrates domain-adversarial feature extraction with one-class anomaly detection and reports improved cross-domain performance on standardized NetFlow datasets [15]. These approaches indicate that cross-domain NIDS is feasible, but their effectiveness depends on the specific domains, the availability of attack support, the way features are constructed, and the assumptions made about the target data.

An alternative strategy is to collect a small amount of target-specific supervision. Lawall and Zöller reported significant degradation across networks and examined whether target-network data could improve the performance of machine-learning-based IDS models [16]. Target labels can directly adjust class boundaries, rather than depending only on marginal distribution alignment. Nonetheless, annotation is expensive, and target traffic is frequently highly imbalanced. Consequently, the key operational question is not merely whether full target retraining succeeds, but instead how much labeled target data is needed and how those labeled records should be selected.

D. Active Learning and Label-Efficient Target Selection

Active learning has been used in intrusion detection to lessen analysts' labeling workload. Görnitz et al. proposed active learning for network intrusion detection and demonstrated that thoughtfully queried labels can enhance anomaly detection under constrained supervision [17]. Torres et al. subsequently studied active labeling of network-traffic datasets and showed that the query strategy influences how efficiently labels are acquired [18]. Beyond uncertainty sampling, core-set methods choose points that are geometrically representative, ensuring that the labeled subset covers the candidate pool [19]. BADGE offers a closely related batch-selection viewpoint by jointly modeling diversity and uncertainty through a gradient-based representation [20]. These findings support the Diversity baseline and the explicit comparison of acquisition criteria examined in this study.

Uncertainty sampling is appealing because it selects records close to a model's decision boundary, yet its data efficiency cannot be assumed [21]. Under cross-dataset shift, the boundary learned in the source domain may be miscalibrated or incorrectly positioned in the target domain; consequently, records judged uncertain may instead reflect atypical source-boundary artifacts rather than representative target instances. Large-scale studies of predictive uncertainty under dataset shift further indicate that confidence quality can decline as the target distribution diverges from the training distribution [22]. By contrast, diversity-driven coverage may overlook a rare attack class when its prevalence is extremely low. This study therefore assesses these competing considerations explicitly using controlled target attack prevalences and fixed annotation budgets.

E. Positioning of This Study

Previous work has independently shown that NIDS models generalize poorly, that standardized features improve comparability, that domain adaptation can alleviate some types of shift, and that active learning can reduce labeling requirements. The present gap lies in jointly evaluating these aspects in a 5G-specific, bidirectional scenario. We first build conservative semantic correspondences between two separately collected 5G datasets. We then decompose the problem into five linked questions: whether zero-shot transfer succeeds, whether dataset identity persists after harmonization, whether aligning with unlabeled target data removes that identity, whether limited target labels can restore performance, and whether added acquisition complexity improves class-blind target selection in the presence of imbalance. The objective is not to introduce an increasingly complex selection algorithm, but to determine the simplest strategy that is supported by reproducible evidence.

Materials and Methods

A. Study Design and Experimental Blocks

The study uses a staged design that distinguishes diagnosis from mitigation and focuses on low-prevalence deployment analysis. Block 1 runs bidirectional zero-shot transfer on conservatively matched tasks. Block 2 checks for residual dataset fingerprinting after harmonization and applies source-predictiveness-aware filtering. Block 3 evaluates CORAL and few-shot target adaptation. Block 4 examines class-blind target querying under deployment-like attack prevalences.

Block 5 brings together protocol freezing, statistical testing, and query-selection ablations. The full staged workflow is summarized in Figure 2.

The main research questions are: (RQ1) whether matched attack semantics transfer across the two datasets; (RQ2) whether harmonized flow variables include a recoverable dataset fingerprint; (RQ3) whether CORAL removes that fingerprint or improves transfer in a reliable manner; (RQ4) whether small target-label budgets recover meaningful performance; (RQ5) whether Random, Diversity, or pilot-based selection is preferable when target-class imbalance is present; and (RQ6) whether additional pilot-allocation complexity yields a stable benefit over simpler acquisition baselines. As shown in Figure 2, the workflow progresses from attack-family matching and zero-shot transfer to dataset-fingerprint diagnosis, CORAL and few-shot adaptation, low-prevalence target querying, and final statistical consolidation. Target-dependent fitting and query selection are performed without access to the held-out target test set.

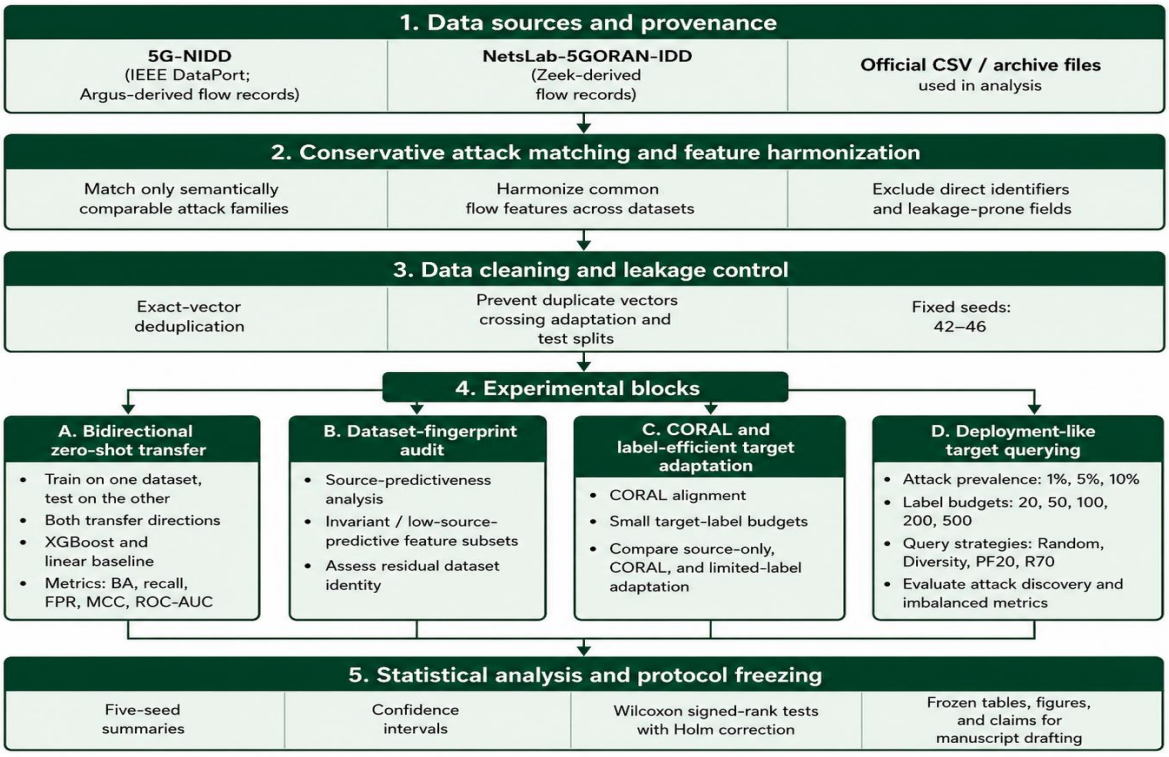

Figure 2: Unified workflow for bidirectional cross-dataset evaluation and label-efficient target adaptation

B. Dataset Provenance and Measurement Characteristics

In this study, the 5G-NIDD source is the dataset that was originally published in 2022 [1]. The specific IEEE DataPort version employed in the experiments was indicated as last updated on 1 January 2026 and was downloaded on 6 July 2026; the subsequent data descriptor is referenced in [2]. The original Combined.zip archive and the extracted Combined.csv file were kept along with SHA-256 checksums. The zero-shot campaign analysis further leveraged the individual base-station attack files to maintain campaign identities and corresponding captures. NetsLab-5GORAN-IDD was taken from the official dataset release described in [3] and was examined via its aggregated Network_Dataset.csv flow table. Table 1 summarizes the principal measurement characteristics of the two sources.

Table 1: Dataset Sources and Measurement Characteristics

Property	5G-NIDD	NetsLab-5GORAN-IDD
Collection setting	Functional 5G test network	Physical 5G Open RAN testbed
Flow extractor	Argus-derived flow records	Zeek-derived network-flow records
Files used	Individual campaign CSVs and Combined.csv	Aggregated Network_Dataset.csv
Primary labels used	Benign; DoS, flood, scan, and credential campaigns	benign; dos, ddos, probe, web, and attack_type labels
Role in this study	Source and target in both directions	Source and target in both directions
Interpretation boundary	Environment, collection, and extractor effects are entangled	Environment, collection, and extractor effects are entangled

C. Conservative Attack Matching

Attack matching was performed before model fitting. Only attacks with a defensible behavioral counterpart in both datasets were retained. NetsLab TCP-ACK flooding was excluded because 5G-NIDD has no direct equivalent, and the 5G-NIDD Torshammer campaign was excluded from the conservative application-layer mapping. Campaign names in the raw 5G-NIDD files and labels in Combined.csv are not identical: for example, GoldenEye is represented as HTTPFlood in the combined table, whereas Slowloris/slow-rate traffic is represented as SlowrateDoS. The experimental code therefore used explicit mapping rules rather than string similarity. Table 2 records the conservative semantic correspondences and exclusions.

Table 2: Conservative Semantic Correspondence Between Attack Labels

Matched task	5G-NIDD campaign / Combined label	NetsLab attack_type	Use
Slow-rate application DoS	Slowloris / SlowrateDoS	slowloris	Zero-shot and application-family adaptation
HTTP request flood	GoldenEye / HTTPFlood	http_flood	Zero-shot and application-family adaptation
SYN flooding	SYNFlood	syn	Zero-shot and flooding-family adaptation
UDP flooding	UDPFlood	udp	Zero-shot and flooding-family adaptation
ICMP flooding	ICMPFlood	icmp	Zero-shot and flooding-family adaptation
TCP scanning	SYNScan + TCPConnect / SYNScan + TCPConnectScan	portscan_tcp	Zero-shot; exploratory scanning adaptation
UDP scanning	UDPScan	portscan_udp	Zero-shot; exploratory scanning adaptation
Application-layer DoS family	HTTPFlood + SlowrateDoS	http_flood + slowloris	Primary adaptation family
Network/transport flooding family	SYNFlood + UDPFlood + ICMPFlood	syn + udp + icmp	Primary adaptation family
Scanning family	SYNScan + TCPConnect + UDPScan / SYNScan + TCPConnectScan + UDPScan	portscan_tcp + portscan_udp	Exploratory only

D. Feature Harmonization and Leakage Control

Direct identifiers, IP addresses, ports used as identifiers, timestamps, local exported indices, and known acquisition-order variables were excluded. In particular, Seq and Offset were not used because the preceding 5G-NIDD study showed that they dominated record-level discrimination [9]. Protocol was retained through harmonized indicator variables in the zero-shot campaign benchmark, but it was excluded from the 15-feature numeric adaptation representation to avoid dependence on incompatible categorical encodings.

Argus SrcBytes and DstBytes were mapped to Zeek src_ip_bytes and dst_ip_bytes, while SrcPkts and DstPkts were mapped to src_pkts and dst_pkts. Nonnegative cleaning was applied, and numerical variables and derived rates were transformed using log1p to reduce heavy-tailed scale differences:

$$x' = \log(1 + \max(x, 0)) \quad (1)$$

Rates used duration plus a small numerical constant in the denominator, and mean packet sizes used a minimum packet count of one. The final numeric representation contained the 15 variables in Table 3.

Table 3: Harmonized Numeric Features Used for Adaptation

Feature	Definition before log1p
src_bytes_per_second	SrcBytes / (duration + epsilon)
src_mean_pkt	SrcBytes / max(SrcPkts, 1)
packets_per_second	(SrcPkts + DstPkts) / (duration + epsilon)
bytes_per_second	(SrcBytes + DstBytes) / (duration + epsilon)
byte_ratio	(SrcBytes + 1) / (DstBytes + 1)
pkt_ratio	(SrcPkts + 1) / (DstPkts + 1)
dst_mean_pkt	DstBytes / max(DstPkts, 1)
dst_bytes	DstBytes
src_bytes	SrcBytes
dst_pkts	DstPkts
src_pkts	SrcPkts
dst_bytes_per_second	DstBytes / (duration + epsilon)
total_bytes	SrcBytes + DstBytes
duration	Flow duration
total_pkts	SrcPkts + DstPkts

E. Deduplication, Sampling, and Target Splits

Exact duplicates were removed in feature-label space before sampling. This step prevents highly repeated vectors from dominating an annotation budget and avoids placing an identical feature-label vector in both the target adaptation and target test subsets. For the principal adaptation experiments, each dataset-task-seed condition used up to 8,000 unique benign and 8,000 unique attack records. The target sample was stratified into a 70% adaptation pool and a 30% held-out test set. The test set was not used for alignment, query selection, threshold tuning, or model fitting.

The deployment-like experiments used a balanced source training sample of 8,000 benign and 8,000 attack vectors. The target consisted of a 6,000-record selection/adaptation pool and a disjoint 3,000-record test set. Target pools were generated at controlled attack prevalences of 1%, 5%, and 10%. These prevalences are stress conditions, not estimates of production prevalence. Scanning was excluded from this block because only 341 unique 5G-NIDD scanning attack vectors remained in the 15-feature space, making budget comparisons unstable.

F. Zero-Shot Transfer Models

The zero-shot benchmark trained on one dataset and evaluated directly on the other without target labels or target-dependent threshold adjustment. For each task and seed, both datasets were sampled to the same benign and attack count, capped at 30,000 records per class. An approximately 70:30 split was constructed within each dataset, and identical common-feature vectors within a sampled class were kept on the same side of the split. The primary nonlinear model was XGBoost with 55 trees, maximum depth 5, learning rate 0.18, subsample 0.90, column subsample 0.90, and histogram-based tree construction. A linear SGD logistic classifier was included as a capacity comparator. Performance was evaluated in both directions across the five seeds.

The decision threshold was fixed at 0.5. Directional transfer was reported separately because averaging the two directions would conceal asymmetric failure. The linear model is interpreted as a comparative screening baseline; some fits reached the iteration cap and are not treated as fully optimized deployment models.

G. Dataset-Fingerprint Audit

A separate source-classification task tested whether a record originated from 5G-NIDD or NetsLab. The audit used balanced data from the matched tasks and the 15-feature numeric representation. Logistic and XGBoost classifiers were evaluated. For each feature, attack-label utility was quantified by the Kolmogorov–Smirnov (KS) divergence between benign and attack distributions in the training source, while source predictiveness was quantified by the KS divergence between benign distributions from the two datasets. Source-penalized subsets were then formed using attack KS minus benign source KS. High source-classification performance after filtering was interpreted as evidence that dataset identity is distributed across multivariate interactions rather than confined to one obvious feature.

H. CORAL and Label-Efficient Target Adaptation

CORAL was applied as an unsupervised, target-aware alignment baseline. Let X_s and X_t denote the source and target adaptation matrices, with means μ_s and μ_t, and let C_s and C_t denote their regularized covariance matrices. The aligned source representation was computed following [10]:

$$X_s^{\{CORAL\}} = (X_s - \mu_s)(C_s + \lambda I)^{\left\{-\frac{1}{2}\right\}(C_t+\lambda I)^{\left\{\frac{1}{2}\right\}}} + \mu_t \qquad (2)$$

with $\lambda = 10^{-4}$. CORAL used the unlabelled target adaptation pool but never the held-out target test set. It is therefore an unsupervised domain-adaptation condition, not a strict zero-shot condition.

Label-efficient adaptation compared source-only, CORAL-only, weighted few-shot, CORAL plus weighted few-shot, and a target-supervised ceiling. Nominal target-label fractions were 0.1%, 1%, and 5% per class, with a minimum of ten records per class. Source and labeled target records were pooled. To prevent the small target set from being numerically overwhelmed, each target record received weight n_s/n_t, where n_s is the number of source training records and n_t is the number of labeled target records. Thus the source and target domains contributed equal total training weight.

I. Deployment-Like Class-Blind Target Querying

The main class-blind query baselines were Random and Diversity. Random drew examples uniformly without replacement from the full unlabeled target pool. Diversity normalized the target pool, trained MiniBatchKMeans, and selected observed records that were closest to the resulting centroids; no class labels were used during selection. Source-model uncertainty—defined as the smallest absolute distance from 0.5—was kept as a negative-control strategy because preliminary experiments indicated marked instability under distribution shift.

Budgets were set to 20, 50, 100, 200, and 500 total target labels. During selection, the labels were concealed and then disclosed only after the query set had been finalized. The selected target records were merged with the source training data using equal domain-total weighting. In this block, XGBoost was configured with 30 trees, a maximum depth of 4, and a learning rate of 0.12.

Pilot-based strategies were considered ablations rather than the primary approach. The retained dual-stratum setup, PF20_R70, allocated a pilot equal to 20% of the total label budget; within this pilot, 70% was sampled uniformly to estimate target prevalence, and 30% used Diversity coverage. A Jeffreys-smoothed estimate, p_hat = (k + 0.5)/ (n + 1), was computed using only the random stratum, where k denotes the number of observed attacks and n denotes the size of the random stratum. The risk share assigned to high source-model attack scores in the remaining budget was:

$$r = \text{clip}\,[0.05 + 0.55\exp(-15\,p_{hat}), 0.05, 0.60] \qquad (3)$$

The remainder was allocated to continued diversity coverage. Sensitivity analysis varied pilot fractions of 10%, 20%, and 30% and random-stratum shares of 50%, 70%, and 90%.

J. Computational Environment, Software, and Model Configuration

All frozen experiments were executed with Python 3.13.5 in a 64-bit Linux 6.12.13 x86_64 virtualized environment. The compute instance exposed five virtual CPU cores from an AMD EPYC 9V74 processor and 5.9 GiB of RAM. These specifications document the analytical environment; the present paper does not report controlled training-time or deployment-latency comparisons.

pandas 2.2.3 and NumPy 2.3.5 were used for data ingestion and numerical operations. XGBoost 3.1.3 implemented the gradient-boosted tree models [23]. scikit-learn 1.8.0 provided splitting, standardization, the SGD-logistic comparator, MiniBatchKMeans, and performance metrics [24]. SciPy 1.17.0 supported matrix operations and statistical tests [25]. NumPy is described in [26], and Matplotlib 3.10.8 generated the figures [27]. The study analyzed the official flow tables rather than re-extracting both datasets through a common exporter: 5G-NIDD fields are Argus-derived, whereas NetsLab fields are Zeek-derived. This distinction is therefore treated as part of the dataset/domain shift. Table 4 summarizes the frozen implementations, model settings, and data-use rules.

K. Evaluation Metrics and Statistical Procedures

Balanced accuracy was calculated following the definition reported in [31]:

$$BA = 0.5\left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right] \qquad (4)$$

Macro-F1 and Matthews correlation coefficient summarized class-balanced classification quality. Attack recall measured sensitivity, and false-positive rate was defined as FP/ (FP + TN). ROC-AUC was reported for balanced experiments. Under 1%-10% target attack prevalence, attack precision and PR-AUC were emphasized because ROC-AUC and balanced accuracy alone can obscure low positive predictive value. Precision-recall analysis is particularly informative under severe class imbalance because precision reflects the prevalence-dependent false-positive burden [28]. We also recorded the number of attacks selected for labeling and the proportion of runs in which no attack was selected.

Results are reported as means across five seeds. Two-sided 95% confidence intervals used the Student t distribution with four degrees of freedom. Strategy comparisons used paired Wilcoxon signed-rank tests over matched task-direction-seed cases. Holm correction controlled the family-wise error rate within each prespecified comparison family. Statistical significance was assessed at $\alpha = 0.05$; descriptive improvements that did not survive correction are not presented as universal superiority.

L. Reproducibility and Interpretation Boundaries

Dataset versions, download dates, archive and extracted-file hashes, random seeds, label mappings, feature definitions, software versions, model settings, and per-seed outputs were

preserved. Appendix A provides the detailed SHA-256 values and frozen artifact identifiers. The experimental protocol was

Table 4: Computational Implementations and Frozen Model Settings

Experimental block	Python implementation	Frozen settings	Role and data-use rule
Zero-shot transfer	XGBClassifier 3.1.3	55 trees; depth 5; learning rate 0.18; subsample 0.90; column subsample 0.90; tree_method=hist; threshold 0.5	Primary nonlinear source-only benchmark; evaluated in both transfer directions.
Zero-shot comparator	StandardScaler + SGD logistic classifier in scikit-learn 1.8.0	Five seeds; fixed threshold 0.5; screening configuration; some fits reached the iteration cap	Capacity comparator, not an optimized deployment model.
Dataset-fingerprint audit	Logistic and XGBoost classifiers	Balanced source-classification sample; 15 numeric features; KS-based attack utility and source predictiveness	Diagnostic task predicting dataset origin.
CORAL and few-shot adaptation	NumPy/SciPy CORAL + XGBClassifier	CORAL regularization 10^{-4}; XGBoost 40 trees, depth 4, learning rate 0.10; equal domain-total weighting	Uses target adaptation data only; held-out target test remains excluded.
Diversity query selection	StandardScaler + MiniBatchKMeans	n_init=1; max_iter=60; batch size ≤1024; nearest observed record to each centroid; labels hidden during selection	Class-blind coverage baseline.
Deployment and PF20_R70	XGBClassifier + Random/Diversity/pilot logic	30 trees; depth 4; learning rate 0.12; subsample 0.85; column subsample 0.85; tree_method=hist; n_jobs=4; threshold 0.5	Final 1%, 5%, and 10% prevalence block with 20–500 target labels.
Inference and reporting	SciPy + Matplotlib	95% t intervals; paired Wilcoxon tests; Holm correction; 300-dpi figures	Computed from frozen per-seed outputs.

fixed before manuscript drafting, and the main tables and figures were generated from those frozen outputs, reducing the risk of protocol changes after preferred results had been observed.

Three constraints limit interpretation. First, the NetsLab aggregate flow file does not expose run or session identifiers; therefore, the external target split is record-disjoint through exact-vector deduplication rather than run-disjoint. Second, differences between Argus and Zeek extraction and between the independent testbeds are intertwined with environmental variation. The findings are therefore discussed as dataset/domain shift rather than pure environmental shift. Third, the controlled prevalence and annotation-budget scenarios are experimental stress tests and should not be interpreted as direct estimates of operational prevalence or labeling cost.

Cross-Dataset Generalization and Dataset-Fingerprint Results

A. Bidirectional Zero-Shot Transfer

The bidirectional zero-shot benchmark exposed substantial external generalization failure even after conservative semantic matching. Across the ten matched tasks and two directions, only three of the twenty XGBoost transfers reached a balanced accuracy of at least 0.80, while ten were below 0.56. The matched application-layer DoS family remained close to chance in both directions (0.5178 for 5G-NIDD to NetsLab and 0.5052 in the reverse direction). The network/transport flooding family transferred asymmetrically: balanced accuracy was 0.8075 from 5G-NIDD to NetsLab but 0.4999 from NetsLab to 5G-NIDD. Scanning showed the same directional pattern, with 0.8345 and 0.5579, respectively.

Attack-level results confirm that the asymmetry was not confined to one family or one transfer direction. In some tasks, the detector retained partial discrimination, whereas in others, the same source model collapsed toward near-chance balanced accuracy or exhibited large false-positive rates. Table 5 and Figure 3 summarize the XGBoost transfer results.

B. Operational Failure Modes

Balanced accuracy alone did not fully describe the transferred detectors. Several models collapsed toward one class. For HTTP flooding, attack recall was approximately zero in both directions. The NetsLab-to-5G SYN-flood model likewise produced a recall below 0.0001. Conversely, the reverse UDP- and ICMP-flood transfers achieved attack recall of 1.0 but incurred false-positive rates of 0.7182 and 0.9008. The family-level NetsLab-to-5G flooding model detected no attacks at the fixed threshold. These outcomes show why attack recall, false-positive rate, and MCC must accompany aggregate performance when assessing external transfer.

C. Residual Dataset Fingerprints

The source-classification audit showed that the harmonized r

Table 5: Bidirectional Zero-Shot Xgboost Transfer On Conservatively Matched Attacks

Matched task	BA 5G→Nets (95% CI)	Recall	FPR	BA Nets→5G (95% CI)	Recall	FPR
Slowloris	0.5130 [0.4801, 0.5459]	0.0788	0.0528	0.7719 [0.6010, 0.9427]	0.8158	0.2720
HTTP Flood	0.4821 [0.4644, 0.4998]	0.0001	0.0359	0.4997 [0.4993, 0.5002]	0.0000	0.0006
SYN Flood	0.9238 [0.8552, 0.9923]	0.9244	0.0769	0.4991 [0.4966, 0.5016]	0.0001	0.0019
UDP Flood	0.7995 [0.7572, 0.8418]	0.6792	0.0802	0.6409 [0.5652, 0.7166]	1.0000	0.7182
ICMP Flood	0.5188 [0.5023, 0.5353]	0.0457	0.0081	0.5496 [0.5087, 0.5905]	1.0000	0.9008
TCP Scanning	0.7715 [0.5710, 0.9719]	0.5559	0.0130	0.7238 [0.6403, 0.8074]	0.4514	0.0038
UDP Scanning	0.6547 [0.3733, 0.9361]	0.3147	0.0053	0.6542 [0.5426, 0.7658]	0.3303	0.0219
Matched Application-layer DoS	0.5178 [0.4629, 0.5728]	0.0626	0.0270	0.5052 [0.4897, 0.5206]	0.1666	0.1562
Matched Network/transport Flooding	0.8075 [0.6849, 0.9300]	0.6901	0.0752	0.4999 [0.4995, 0.5002]	0.0000	0.0003
Matched Scanning	0.8345 [0.7861, 0.8829]	0.6821	0.0131	0.5579 [0.4629, 0.6530]	0.1385	0.0226

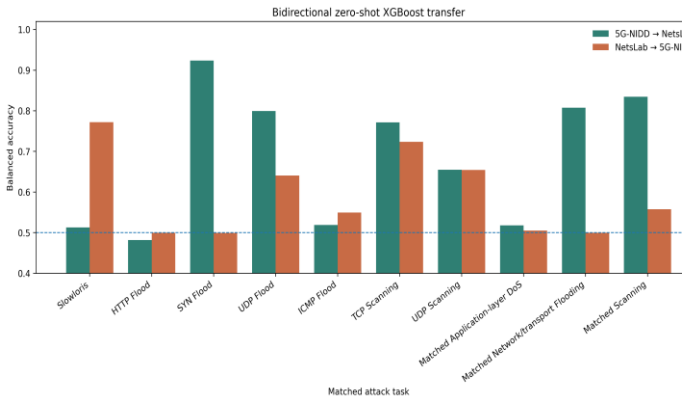

Figure 3: Bidirectional zero-shot XGBoost balanced accuracy for conservative matched attack tasks

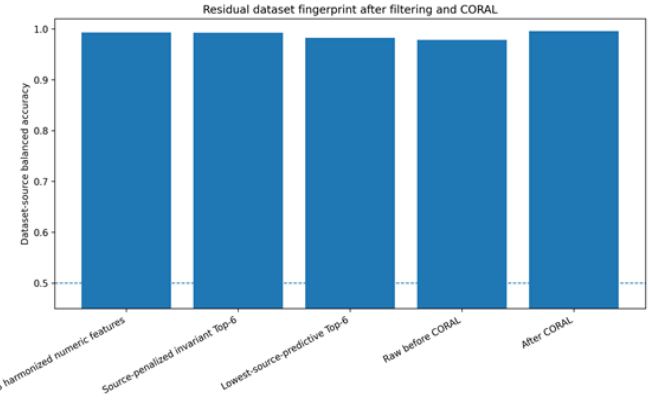

Figure 4: Residual dataset-source classification after feature filtering and CORAL

epresentation retained a strong multivariate dataset fingerprint. Using all fifteen numeric features, XGBoostdistinguished 5G-NIDD from NetsLab with a mean balanced accuracy of 0.9933. Penalizing features for source predictiveness did not remove this signal: the source-penalized Top-6 achieved 0.9928, and the six least source-predictive features still achieved 0.9823. This result indicates that dataset identity was distributed across the representation rather than concentrated in one obvious field.

The CORAL diagnostic led to the same conclusion. Mean dataset-source balanced accuracy remained high after alignment, indicating that first- and second-order covariance matching did not eliminate the source-domain fingerprint. The implication is that distributional mismatch extends beyond obvious direct shortcuts and cannot be assumed to vanish after simple covariance alignment. Table 6 and Figure 4 report the residual fingerprint results.

Label-Efficient Target Adaptation Results

A. CORAL Versus Limited Target Supervision

Across the exploratory three-family adaptation experiment, mean source-only transfer balanced accuracy was 0.5852. CORAL increased the mean to 0.6371, but its effect varied by family and direction and did not remove the dataset fingerprint. In contrast, the best condition using no more than 5% labeled target data improved balanced accuracy by an average of 0.3462 relative to source-only transfer. The magnitude of the gain ranged from 0.0528 to 0.5509, showing that a small amount of target-specific supervision corrected decision boundaries more effectively than marginal alignment alone.

Improvements in application-layer DoS were observed, reaching 0.9330 in the 5G-NIDD-to-NetsLab direction and 0.9749 in the opposite direction. Flooding performance also improved, achieving 0.9085 and 0.8107 for the respective directions. Despite high scanning results, these were retained as exploratory evidence due to the limited representation in 5G-NIDD scanning, which included only 341 distinct attack vectors. This led to higher effective fractions after applying the minimum-per-class rule. Detailed settings for the best task-direction adaptation can be found in Table 7.

B. Class-Blind Selection on a Balanced Target Pool

The balanced-pool experiment isolated the effect of acquisition strategy from target-class imbalance. Random and Diversity improved consistently as the label budget increased, whereas source-model uncertainty sampling lagged significantly. With 20 labels, Random achieved a

Table 6: Residual dataset-source predictability after feature filtering

Feature representation	Source BA	95% CI	MCC	ROC-AUC
All 15 harmonized numeric features	0.9933	[0.9923, 0.9943]	0.9867	0.9996
Source-penalized invariant Top-6	0.9928	[0.9920, 0.9935]	0.9856	0.9994
Lowest-source-predictive Top-6	0.9823	[0.9814, 0.9833]	0.9647	0.9985

Table 7: Best label-efficient adaptation result by task and transfer direction

Matched task	Direction	Best method	Target labels	Best BA	Δ vs source-only	Recall FPR
Matched Application-layer DoS	5G-NIDD → NetsLab	Weighted_fewshot	5.0%	0.9330	0.5509	0.9248 0.0589
Matched Application-layer DoS	NetsLab → 5G-NIDD	Coral_weighted_fewshot	5.0%	0.9749	0.4388	0.9774 0.0276
Matched Network/transport Flooding	5G-NIDD → NetsLab	Coral_weighted_fewshot	5.0%	0.9085	0.1395	0.9119 0.0948
Matched Network/transport Flooding	NetsLab → 5G-NIDD	Coral_weighted_fewshot	5.0%	0.8107	0.0528	0.8698 0.2484
Matched Scanning	5G-NIDD → NetsLab	Coral_weighted_fewshot	5.0%	0.9845	0.4470	0.9733 0.0042
Matched Scanning	NetsLab → 5G-NIDD	Coral_weighted_fewshot	1.0%	0.9771	0.4482	0.9784 0.0243

Table 8: Class-Blind Target-Selection Results On The Balanced Target Pool

Total labels	Random BA	Uncertainty BA	Diversity BA	Diversity − Random	Holm-adjusted p (Diversity vs Random)	Holm-adjusted p (Uncertainty vs Random)
20	0.8148	0.6378	0.8288	0.0139	0.8983	0.0002
50	0.8477	0.6534	0.8765	0.0288	0.0401	0.0010
100	0.8776	0.6597	0.8902	0.0126	0.1200	<0.0001
200	0.8879	0.6670	0.8965	0.0086	0.0483	<0.0001
500	0.9001	0.6994	0.9041	0.0040	0.1650	<0.0001

mean balanced accuracy of 0.8148, Diversity reached 0.8288, and uncertainty achieved only 0.6378. With 500 labels, the corresponding values were 0.9001, 0.9041, and 0.6994. Diversity exceeded Random by 0.0288 at 50 labels and by 0.0086 at 200 labels; both differences remained significant after Holm correction, with adjusted p-values of 0.0401 and 0.0483, respectively. Other Diversity-Random differences were smaller and not significant after correction. Random outperformed uncertainty sampling at every budget, underscoring the unreliability of source-model decision boundaries under pronounced cross-dataset shift. Table 8 summarizes the balanced-pool results.

VI. Deployment-Like Target-Selection Results

A. Effect of Attack Prevalence and Label Budget

In a final deployment-like rerun, Random, Diversity, and the PF20_R70 dual-stratum ablation were evaluated under target attack prevalence levels of 1%, 5%, and 10%. The relative effectiveness of each strategy varied depending on prevalence and budget, with Random or Diversity leading in thirteen out of fifteen conditions. This suggests that increasing pilot-allocation complexity does not necessarily yield better results. Specifically, at 1% prevalence, Diversity was the most effective strategy across label counts ranging from 20 to 200, with balanced accuracy increasing from 0.5992 to 0.6596. At 500 labels, PF20_R70 reached 0.6561, only 0.0022 above Diversity. At 5% prevalence, Diversity led from 50 through 500 labels and reached 0.8000 at the largest budget. At 10% prevalence, Random was strongest at 50, 100, and 200 labels, while Diversity led at 20 and 500 labels. These changes show that the preferred query policy depends on both target prevalence and the available annotation budget (Table 9 and Figures 5–7).

Table 9: Balanced Accuracy Under Deployment-Like Target Attack Prevalence

Attack prevalence	Total labels	Random BA	Diversity BA	PF20_R70 BA	Best strategy
1%	20	0.5138	0.5992	0.5488	Diversity
1%	50	0.5370	0.6523	0.6161	Diversity
1%	100	0.5434	0.6613	0.6160	Diversity
1%	200	0.6214	0.6596	0.6281	Diversity
1%	500	0.6354	0.6539	0.6561	PF20_R70
5%	20	0.5817	0.5953	0.5991	PF20_R70
5%	50	0.6324	0.6696	0.6605	Diversity
5%	100	0.6935	0.7081	0.6880	Diversity
5%	200	0.7517	0.7581	0.7062	Diversity
5%	500	0.7907	0.8000	0.7431	Diversity
10%	20	0.6179	0.6351	0.6063	Diversity
10%	50	0.7566	0.7126	0.7024	Random
10%	100	0.7921	0.7273	0.7203	Random
10%	200	0.8267	0.7726	0.7814	Random
10%	500	0.8544	0.8670	0.8626	Diversity

B. Rare-Attack Discovery and Imbalanced Metrics

At 1% attack prevalence, target querying becomes a rare-class discovery problem. Across the 20-label trials, Random selected no attacks. Diversity selected an average of 0.85 attacks, but 50% of its trials still contained no attack. PF20_R70 performed slightly better, selecting an average of 1.05 attacks, while 45% of trials still contained no attack. At a 50-label budget, the mean attack yield was 0.20 for Random, 1.70 for Diversity, and 3.00 for PF20_R70. a trade-

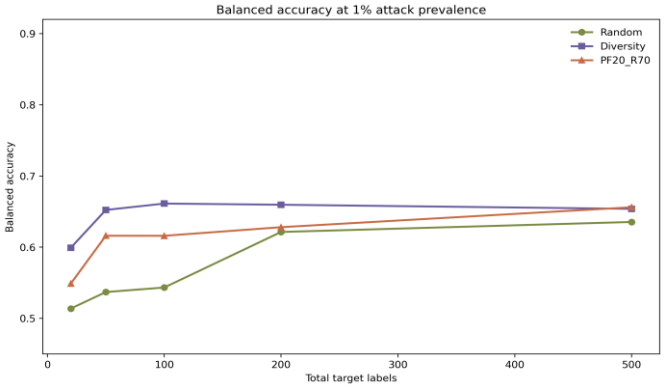

Figure 5: Balanced accuracy versus total target-label budget at 1% attack prevalence.

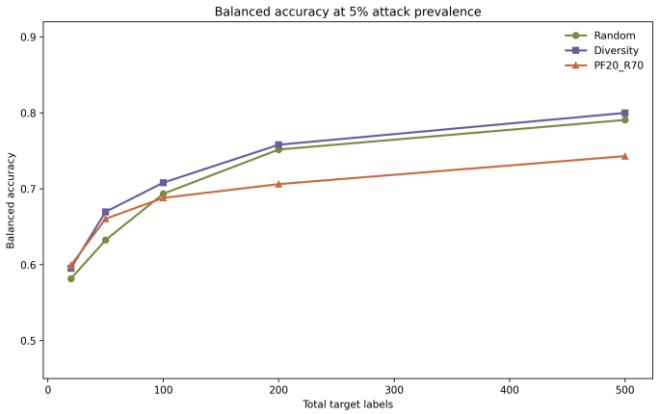

Figure 6: Balanced accuracy versus total target-label budget at 5% attack prevalence.

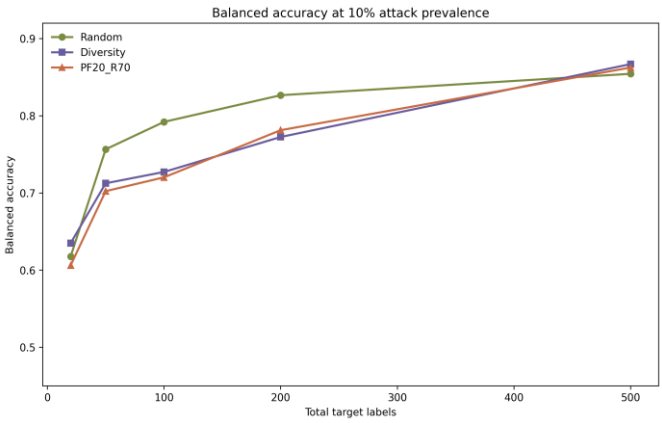

Figure 7: Balanced accuracy versus total target-label budget at 10% attack prevalence.

off between risk enrichment and coverage of the target feature space.

Although PF20_R70 yielded more attacks, it did not consistently achieve the highest balanced accuracy, indicating Precision and PR-AUC serve as complementary metrics to balanced accuracy by indicating if enhanced discrimination correlates with effective rare-attack recovery. At a prevalence of 1%, the mean number of attacks identified from queried records remained low with small budgets but showed consistent growth as the label budget increased. Diversity often outperformed under smaller budgets, while PF20_R70 provided a balanced approach that was not statistically superior. Figure 8 illustrates the queried attack yield at this prevalence.

C. Pilot-Allocation Sensitivity

A sensitivity analysis on pilot allocation adjusted pilot

fractions of 10%, 20%, and 30% with random-stratum shares of 50%, 70%, and 90%. The highest overall mean balanced accuracy (0.6757) was achieved by PF20_R70, while PF30_R90 had the lowest prevalence-estimation error (0.0359). However, the range of mean balanced accuracy across the five configurations was minimal at just 0.0030.

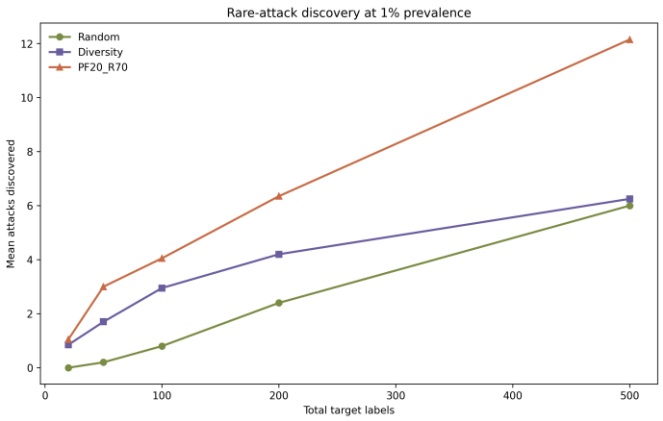

Figure 8: Mean number of attacks discovered in the queried target set at 1% prevalence.

After adjusting for Holm correction, none of the sixty internal comparisons involving pilot size or random share maintained statistical significance. The only exception was PF20_R90, which underperformed compared to Diversity at a 5% prevalence with 200 labels (mean difference −0.0746; adjusted p = 0.0236). As a result, PF20_R70 remains a transparent and balanced choice rather than a statistically optimal one, with Random and Diversity as the primary operational benchmarks. Table 10 provides a summary of the retained pilot configurations.

Table 10: Pilot-Allocation Sensitivity Summary

Configuration	Pilot fraction	Random share	Mean BA	Posterior MAE	Selected attacks	Pareto efficient
PF20_R70	20%	70%	0.6757	0.0533	14.647	Yes
PF10_R90	10%	90%	0.6749	0.0503	14.940	Yes
PF20_R90	20%	90%	0.6741	0.0430	14.717	Yes
PF30_R90	30%	90%	0.6739	0.0359	14.373	Yes
PF20_R50	20%	50%	0.6727	0.0598	14.270	No

D. Consolidated Result

Collectively, the findings point to an interpretation focused on diagnosis and adaptation. The performance gap across different datasets persists despite techniques like semantic matching, source-aware feature filtering, and moment alignment, indicating strong domain-specific regularities within the datasets. With limited target supervision, significant performance improvements are achievable, yet no single acquisition strategy excels at every prevalence level and budget. Therefore, the most prudent operational approach is to use Random as a robustness baseline, Diversity as a coverage-focused alternative, and PF20_R70 as an ablation reflecting the compromise between prevalence estimation, rare-attack enrichment, and target-space coverage.

VII. Discussion and Threats to Validity

A. Cross-Dataset Transfer Is a Conditional Property

The conservative alignment reduces one clear source of error—incompatible labels—but does not make the datasets interchangeable. Transfer performance depends on the attack family, the source-to-target direction, and the model. For example, SYN flooding transferred effectively from 5G-NIDD to NetsLab but failed in the reverse direction, whereas Slowloris showed the opposite pattern. No dataset can therefore be described as universally easier or harder; the learned decision boundary interacts with the way the target data are collected and the attacks are represented. This result is consistent with Cantone et al. [8], who reported strong within-dataset performance but unstable transfer across independently collected NIDS datasets. Explainable cross-domain analysis likewise shows that reversing the source-target order can change the apparent usefulness of individual features [13].

Operational metrics further underscore why relying solely on aggregate scores is inadequate. A balanced accuracy near 0.5 sometimes signifies almost no attack recall, while in other instances, it indicates a high rate of false alarms. Both outcomes are unacceptable operationally, but they call for different corrective actions. Therefore, thorough reporting, including attack-level results, attack recall, false-positive rate, MCC, and metrics sensitive to prevalence should be essential parts of external NIDS evaluation rather than optional diagnostics.

B. Dataset Fingerprints Are Distributed Across the Representation

A previous 5G-NIDD study demonstrated how variables such as Seq and Offset can dominate record-level evaluation [9]. These variables and other direct identifiers were excluded here, yet dataset-source classification remained nearly perfect. XGBoost achieved mean balanced accuracy of 0.9933 with all fifteen numeric features and 0.9928 with the source-penalized Top-6. Source identity was therefore not tied to a single removable field, but was distributed across combinations of duration, byte and packet counts, mean packet sizes, ratios, and rate variables. Removing obvious metadata shortcuts is necessary, but it does not ensure cross-dataset portability.

Various factors contribute to dataset fingerprinting, such as flow-exporter behavior, timeout rules, capture location, traffic generation, benign-traffic composition, protocol implementation, attack tools, and preprocessing. The current design cannot pinpoint the contribution of each factor since 5G-NIDD and NetsLab differ in their collection environments and measurement pipelines between Argus and Zeek. CORAL attempts to align the first two moments of the source and target representations but high predictability remains after alignment [10]. This indicates higher-order discrepancies or modality-specific structures beyond simple covariance matching.

C. Why Limited Target Supervision Is More Effective

While CORAL increased mean transfer balanced accuracy from 0.5852 to 0.6371 in the three-family adaptation experiment, its effectiveness varied across families and transfer directions. Limited target supervision proved more influential: using no more than 5% labeled target data led to an average best improvement of 0.3462 in balanced accuracy. Labels clearly delineate benign vs. malicious records in the target dataset, whereas unsupervised methods only aim to align feature distributions more closely. This explains why direct target supervision can adjust a misaligned class boundary even with remaining source-domain fingerprints.

Few-shot results should be viewed as adaptation curves rather than assurances that a fixed percentage of labels always

ensures success. The target samples were carefully managed, duplicate vectors were excluded, and the labels assumed correct. Moreover, reverse flooding was significantly harder than application-layer DoS attacks, and scanning had limited support in 5G-NIDD. Practically, annotation budgets should be assessed for each attack family and transfer direction separately, using disjoint target test sets, rather than relying on a universal percentage.

D. Target Selection Under Class Imbalance

Class-blind selection experiments highlight both the advantages and limitations of acquisition strategies. Diversity typically peaks at lower budgets, covering various areas within the target feature space. Random sampling emerged as a solid benchmark and, in numerous medium-prevalence scenarios, either matched or outperformed others. In contrast, relying on uncertainty from the source model often yielded poorer results. During domain shifts, a probability near 0.5 does not always correspond to an informative target point; it might instead result from a flawed or poorly calibrated source boundary. As a result, the chosen records may showcase boundary artifacts rather than genuine exemplars. Even in standard active-learning contexts, the efficiency of uncertainty sampling is not guaranteed. Under dataset shifts, predictive confidence may decrease further. Hence, this negative outcome suggests that source-model uncertainty should be used more as a diagnostic check than as a standard acquisition strategy.

At an attack prevalence of 1%, selection also turned into a discovery challenge. A small query batch might lack any attacks, preventing any classifier from learning the target attack boundary regardless of the downstream algorithm used. For this reason, metrics such as precision, PR-AUC, attack yield, and the proportion of query sets without attacks are reported alongside balanced accuracy. When positives are infrequent, precision-recall analysis provides more insights than ROC analysis. The pilot experiments aimed to gauge target rarity and reserve labels for high-risk records; however, the configuration differences were minimal and statistically inconsistent. Therefore, PF20_R70 is maintained as an ablation rather than promoted as a new primary algorithm. Overall, the findings advocate for using simple and transparent Random and Diversity baselines instead of complicating allocation with additional complexity.

E. Implications and Impacts for 5G IDS Evaluation and Deployment

In designing benchmarks, the findings advocate for a streamlined external-evaluation protocol. Key steps include removing direct identifiers and acquisition-order variables, aligning attack semantics prior to transfer, reporting bidirectional source-target data, evaluating attack-level recall and false-positive rates, and assessing whether the dataset identity is still discernible from the retained representation. When using target labels, results should reflect the absolute annotation budget and explicit conditions of attack prevalence. These measures are intended to mitigate the risk of misinterpreting a high internal score as an indication of transferable attack recognition.

For deployment, source-trained models should be treated as starting points rather than complete detectors. A small, well-audited target-labeling program can support recalibration or retraining, while Random or Diversity batches provide transparent initial acquisition strategies. The target test set must remain separate from all adaptation processes, and ongoing monitoring should track false alarms, missed attacks, PR-AUC, and feature-distribution shift. The evidence supports local adaptation but cautions against assuming that a static cross-dataset model will remain reliable as network, software, extractor, or traffic conditions change.

The prospective impacts extend beyond model accuracy. Economically, external validation before deployment can reduce the costs associated with brittle detectors and repeated false-alarm investigation, while label-efficient adaptation limits expert annotation demands. Societally, more dependable intrusion detection can support service continuity and user trust in 5G-enabled public and commercial services. Environmental effects were not measured; however, favoring small target-label budgets and simpler acquisition baselines may avoid unnecessary retraining and computational overhead. These statements are deployment implications rather than quantified impact estimates.

These deployment recommendations are consistent with complementary studies that emphasize empirical security assessment and edge-deployment constraints. Masoud et al. used Nessus to assess vulnerabilities in Libyan banking websites [29], while Osman et al. compared quantized autoencoders, LSTMs, and lightweight Transformers on RT-IoT2022 and reported the strongest accuracy-latency-energy trade-off for a quantized autoencoder on a Raspberry Pi 4 [30]. Together with the earlier 5G-NIDD leakage-aware study [9], these works reinforce the need to evaluate security models in terms of measurement validity and deployment constraints rather than predictive accuracy alone.

F. Threats to Validity

Construct validity: Attack labels were matched cautiously, but sharing a family name does not necessarily indicate identical behavior at the packet level, nor does it reflect the intensity, tooling, or network context. Metrics like balanced accuracy, MCC, ROC-AUC, precision, and PR-AUC each highlight different facets of detection. However, they do not address analyst workload, the seriousness of incidents, or the impact of detection delays directly. The controlled prevalence experiments serve as stress tests rather than precise measures of real-world prevalence.

Internal validity: Exact feature-label duplicates were eliminated, and target adaptation and test sets did not overlap, but the NetsLab aggregate file lacks identifiers for runs and sessions. As such, related records from the same underlying activities might span different splits even though there are no exact vector matches. Hyperparameters and probability thresholds were kept consistent for comparability rather than independently adjusted for each source-target comparison. Equal weighting for source-target domains was chosen as one of several viable adaptation strategies.

External validity: The study compares two independently collected 5G datasets and a conservative selection of similar attacks. Results may vary with different 5G infrastructures, radio settings, user populations, data exporters, attack methods, or enterprise networks. Any disparities between Argus and Zeek are intertwined with environmental differences, and the study does not account for temporal shifts, the evolution of encrypted traffic, online feedback, or live incident response. Scanning was omitted from the deployment-like block due to limited unique attack vectors in 5G-NIDD.

Statistical conclusion validity: Using five random seeds provided repeated estimates but limited power for detecting small effects. Confidence intervals and paired Wilcoxon tests were applied, along with Holm correction for comparison

families. Several adaptation strategies were tested before finalizing the protocol; thus, the manuscript clearly distinguishes core results from supplementary diagnostics and refrains from treating the most successful exploratory setup as definitively optimal.

Reproducibility: Versions of datasets, download dates, hashes, label mappings, feature definitions, seeds, outputs per seed, and the frozen result tables are all documented. Nevertheless, numerical replication might depend on variations in software-library versions, floating-point calculations, and XGBoost dynamics. While provided artifacts mitigate this risk, they do not entirely eliminate platform-related inconsistencies.

VIII. Conclusion and Future Work

This study explored the ability of intrusion detectors trained on one 5G dataset to recognize attacks in another dataset with similar semantics. The findings were conditional and not consistently favorable. Zero-shot transfer heavily depended on specific attacks and exhibited strong directionality; several transferred models either missed attacks or generated too many false positives. Even after removing direct identifiers, sequence metadata, and highly predictive individual features associated with the source, the remaining harmonized numeric flow features maintained a nearly perfect multivariate dataset fingerprint. However, the CORAL technique failed to eliminate this fingerprint and only occasionally improved transferability.

The use of limited labeled target data emerged as the most reliable mitigation strategy. Diversity was often effective at lower budgets. Random selection remained a solid baseline for robustness, while the source model's uncertainty was not dependable when distributions shifted. In scenarios where attack prevalence ranged from 1% to 10%, the task involved both identifying rare attacks and training a classifier, necessitating precision, PR-AUC, and attack yield alongside balanced accuracy. More sophisticated pilot-allocation strategies failed to consistently provide benefits. Consequently, the main takeaway is not the creation of a universal cross-dataset 5G detector, but rather that successful transfer evaluation hinges on explicit domain diagnosis, and that moderate target supervision can enhance an initially unreliable source model into a more effective target-specific detector.

In terms of future work, the priority is to conduct an evaluation that is disjoint across runs, campaigns, and time. When raw packet captures are available, both datasets should be reprocessed using the same flow-export pipeline to distinguish effects originating from the extractor versus those from the environment. Additional independently gathered 5G and Open RAN datasets should be incorporated to determine if the observed fingerprints are unique to a specific dataset pair or indicate more general measurement patterns.

Methodologically, upcoming studies should explore calibrated probability estimates, policies for abstention and human review, open-set attacks, and sequential adaptation in the face of concept drift. Label acquisition should consider analyst costs, delays in labeling, and asymmetric attack severity rather than solely optimize classification metrics. Access to real target prevalence and chronological traffic streams is crucial to validate findings related to rare attacks. These methodological extensions aim to transition evaluation from controlled record-level adaptation to a reproducible, continuously monitored cross-environment protocol for 5G intrusion detection.

Author Contributions: Abraheem Sulayman Alsaedi Abraheem: Conceptualization; methodology; software; validation; formal analysis; data curation; visualization; writing—original draft; writing—review and editing; supervision; and project administration. ALBUSSIRY Alhussin Ali Edhirig: methodology; investigation; validation; and writing—review and editing. Both authors read and approved the final manuscript.

Funding: This research received no external funding.

Data Availability Statement: 5G-NIDD is available through IEEE DataPort and was originally released in the study cited in [1]. Its later descriptor is cited in [2]. The experiments used the dataset originally released in 2022, specifically the IEEE DataPort version listed as last updated on 1 January 2026 and accessed on 6 July 2026. NetsLab-5GORAN-IDD is available through the official NetsLab release described in [3]. Exact archive and extracted-file SHA-256 values, label mappings, preprocessing rules, and frozen result identifiers are provided in Appendix A.

Code and Materials Availability: The versioned publication-companion package (v1.0.1) contains the archived per-seed outputs, feature mappings, publication figures, environment information, SHA-256 manifests, and the exact frozen workbook and results-package identifiers listed in Appendix A. It reproduces the manuscript tables and figures exactly from the archived frozen experimental outputs. The package also includes clearly labeled reconstructed executable implementations of the documented raw-data protocols; these implementations are provided for transparency and further analysis and are not claimed to reproduce every reported numerical value exactly from raw data across software platforms. The public repository is available at https://github.com/Abraheem2026/5G-Cross-Dataset-IDS-Adaptation, the fixed GitHub release v1.0.1 is available at https://github.com/Abraheem2026/5G-Cross-Dataset-IDS-Adaptation/releases/tag/v1.0.1, and the archived software record is available under https://doi.org/10.5281/zenodo.21678601.

Ethics Statement: This study used publicly released network-traffic datasets and did not involve human participants, identifiable personal data, animals, or clinical interventions. Institutional ethics approval was therefore not required.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration of Generative AI and AI-Assisted Technologies: During the preparation of this manuscript, the authors used OpenAI ChatGPT to support manuscript organization, English-language refinement, code development, review, debugging, and consistency checking. All computational procedures and AI-assisted suggestions were critically reviewed by the authors and verified against the archived datasets, documented protocols, and frozen experimental outputs. The authors take full responsibility for the accuracy, integrity, and final content of the work.

References

[1] S. Samarakoon, et al. “5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network.” arXiv:2212.01298, 2022

[2] Y. Siriwardhana, et al. “Descriptor: 5G Wireless Network Intrusion Detection Dataset (5G-NIDD*).” IEEE Data Descriptions*, vol. 2, pp. 358–369, 2025. https://doi.org/10.1109/IEEEDATA. 2025.3592888

[3] F. Zadeh, A. Civciss, V. Ravihansa, C. Sandeepa, and M. Liyanage. "Descriptor: 5G Open Radio Access Network Multi-Modal Intrusion Detection Dataset (NetsLab-5GORAN-IDD)." *IEEE Data Descriptions*, vol. 2, pp. 348–357, 2025. https://doi.org/10.1109/IEEEDATA.2025.3614167

[4] 3rd Generation Partnership Project. "Security Architecture and Procedures for 5G System." 3GPP TS 33.501, ver. 18.9.0, Release 18, Apr. 2025.

[5] Open RAN MoU Group. "Open RAN Security White Paper." Mar. 2022. [Online]. Available: https://www.o-ran.org/oran-ecosystem-resources/open-ran-security-white-paper-march-2022.

[6] J. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. Chawla, and F. Herrera. "A Unifying View on Dataset Shift in Classification." *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012, https://doi.org/10.1016/j.patcog.2011.06.019

[7] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho. "A Survey of Network-Based Intrusion Detection Data Sets." *Computers & Security*, vol. 86, pp. 147–167, 2019, https://doi.org/10.1016/j.cose.2019.06.005

[8] M. Cantone, C. Marrocco, and A. Bria. "Machine Learning in Network Intrusion Detection: A Cross-Dataset Generalization Study." *IEEE Access*, vol. 12, pp. 144489–144508, 2024, https://doi.org/10.1109/ACCESS.2024.3472907

[9] A. Abraheem, A. Arhoumah, H. Fokla, and M. Basheer. "Feature-Selected Machine Learning for 5G Intrusion Detection: Leakage-Aware Evaluation and Sensitivity to Sequence Metadata in 5G-NIDD." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 2, pp. 121–135, 2026. https://doi.org/10.63318/waujpasv4i2_14

[10] B. Sun, J. Feng, and K. Saenko. "Return of Frustratingly Easy Domain Adaptation." In Proc. 30th AAAI Conf. Artificial Intelligence, 2016, pp. 2058–2065. https://doi.org/10.1609/aaai.v30i1.10306

[11] M. Sarhan, S. Layeghy, and M. Portmann. "Towards a Standard Feature Set for Network Intrusion Detection System Datasets." *Mobile Networks and Applications*, vol. 27, pp. 357–370, 2022, https://doi.org/10.1007/s11036-021-01843-0

[12] M. Sarhan, S. Layeghy, and M. Portmann. "Evaluating Standard Feature Sets Towards Increased Generalisability and Explainability of ML-Based Network Intrusion Detection." *Big Data Research*, vol. 30, Art. no. 100359, 2022. https://doi.org/10.1016/j.bdr.2022.100359

[13] S. Layeghy and M. Portmann. "Explainable Cross-Domain Evaluation of ML-Based Network Intrusion Detection Systems." *Computers & Electrical Engineering*, vol. 108, Art. no. 108692, 2023, https://doi.org/10.1016/j.compeleceng.2023.108692

[14] S. Ben-David, et al. "A Theory of Learning from Different Domains." *Machine Learning*, vol. 79, pp. 151–175, 2010, https://doi.org/10.1007/s10994-009-5152-4

[15] S. Layeghy, M. Baktashmotlagh, and M. Portmann. "DI-NIDS: Domain Invariant Network Intrusion Detection System." *Knowledge-Based Systems*, vol. 273, Art. no. 110626, 2023. https://doi.org/10.1016/j.knosys.2023.110626

[16] A. Lawall and T. Zöller. "Impact of Target Network-Specific Data on the Performance of Machine-Learning-Based Intrusion Detection System Models." In Proc. 2025 5th Intelligent Cybersecurity Conference (ICSC), Tampa, FL, USA, 2025, pp. 290–297. https://doi.org/10.1109/ICSC65596.2025.11140289

[17] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld. "Active Learning for Network Intrusion Detection." In Proc. 2nd ACM Workshop on Security and Artificial Intelligence, 2009, pp. 47–54, https://doi.org/10.1145/1654988.1655002

[18] J. L. Guerra Torres, C. A. Catania, and E. Veas. "Active Learning Approach to Label Network Traffic Datasets." *Journal of Information Security and Applications*, vol. 49, Art. no. 102388, 2019, https://doi.org/10.1016/j.jisa.2019.102388

[19] O. Sener and S. Savarese. "Active Learning for Convolutional Neural Networks: A Core-Set Approach." In Proc. 6th International Conference on Learning Representations (ICLR), 2018

[20] J. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. "Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds." In Proc. 8th International Conference on Learning Representations (ICLR), 2020

[21] S. Mussmann and P. Liang. "On the Relationship Between Data Efficiency and Error for Uncertainty Sampling." In Proc. 35th International Conference on Machine Learning, PMLR, vol. 80, pp. 3674–3682, 2018

[22] Y. Ovadia, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." In Advances in Neural Information Processing Systems 32, 2019

[23] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." In Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785

[24] F. Pedregosa, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011

[25] P. Virtanen, et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, vol. 17, pp. 261–272, 2020, https://doi.org/10.1038/s41592-019-0686-2

[26] C. R. Harris, et al. "Array Programming with NumPy." *Nature*, vol. 585, pp. 357–362, 2020, https://doi.org/10.1038/s41586-020-2649-2

[27] J. D. Hunter. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, https://doi.org/10.1109/MCSE.2007.55

[28] T. Saito and M. Rehmsmeier. "The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015. https://doi.org/10.1371/journal.pone.0118432

[29] R. Masoud, A. Ahmed, and M. Alghali. "Security Assessment of Some Libyan Banks Websites." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 3, no. 1, pp. 6–10, 2025. [Online]. Available: https://waujpas.com/index.php/journal/article/view/96

[30] M. Osman, F. Elghaffi, L. Ben Dalla, Ö. Karal, and T. Rashid. "A New Approach for Low-Latency, High-Accuracy Anomaly Detection at the Edge: Benchmarking Quantized Autoencoders, LSTMs, and Lightweight Transformers on RT-IoT2022 Time-Series Traffic." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 110–121, 2026, https://doi.org/10.63318/waujpasv4i1_12

[31] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. "The Balanced Accuracy and Its Posterior Distribution." Proceedings of the 20th International Conference on Pattern Recognition (ICPR), pp. 3121–3124, 2010, https://doi.org/10.1109/ICPR.2010.764

Appendix A. Reproducibility Checklist

Table A1 records the exact data files, software environment, sampling rules, statistical procedures, and frozen artifacts used for the reported experiments.

Table A1: Reproducibility Checklist

Item	Frozen value
5G-NIDD source archive	Combined.zip; SHA-256 3389b947b21a0106c359ca887de36317305f32a4f20c0b72fbd3ec1a513af240.
5G-NIDD extracted file	Combined.csv; SHA-256 fa36f80859585f474504ca69eb951a079b35145537976c86b00aae3aab46ee59.

Item	Frozen value
NetsLab source archive	Network_Dataset.zip; SHA-256 520d5e4db85f320ab7bca35a2a879969f014b33645a719bf82fa6f29c37844b3.
NetsLab extracted file	Network_Dataset.csv; SHA-256 656ff0fe4a5b2a4234a6916d2a24e57ba96a1302fb01f5ea3d32aea3090420f9.
Dataset versioning	5G-NIDD original release year 2022; IEEE DataPort version listed as updated 1 January 2026; downloaded 6 July 2026.
Measurement lineage	5G-NIDD flow fields are Argus-derived; NetsLab flow fields are Zeek-derived. No common-exporter re-extraction was performed.
Software environment	Python 3.13.5; pandas 2.2.3; NumPy 2.3.5; scikit-learn 1.8.0; XGBoost 3.1.3; SciPy 1.17.0; Matplotlib 3.10.8.
Compute environment	64-bit Linux 6.12.13 x86_64; KVM virtualized; AMD EPYC 9V74; five virtual CPU cores; 5.9 GiB RAM.
Randomization and threshold	Seeds 42, 43, 44, 45, and 46; fixed classification threshold 0.5.
Zero-shot sampling	Matched benign and attack counts capped at 30,000 per class; approximately 70:30 split; identical feature vectors grouped to one side.
Few-shot adaptation	Up to 8,000 benign and 8,000 attack vectors; target 70% adaptation pool and 30% held-out test; nominal 0.1%, 1%, and 5% labels per class, minimum 10.
Deployment-like querying	Source 8,000 records per class; target pool 6,000; target test 3,000; prevalence 1%, 5%, and 10%; budgets 20, 50, 100, 200, and 500.
Primary acquisition baselines	Random and Diversity; source-model uncertainty retained only as a negative control; PF20_R70 retained as an ablation.
Statistical inference	Two-sided 95% t intervals across five seeds; paired Wilcoxon signed-rank tests; Holm family-wise correction.
Frozen results workbook	Paper2_Frozen_Protocol_and_Manuscript_Results_v1.6.xlsx; SHA-256 c6086735819d660309ce06d25412de2dff0ac74438c261b7d4123acc5dda0573.
Frozen results package	Paper2_Frozen_Results_Package_v1.6.zip; SHA-256 26afbcbe46caea95cd38f2f19406f7da12c34d3f28ff5427ba2d72ee7a0164e3.
Public code release	Publication-companion package v1.0.1; reproduces the frozen manuscript tables and figures exactly and includes clearly labeled reconstructed raw-data implementations. GitHub repository: https://github.com/Abraheem2026/5G-Cross-Dataset-IDS-Adaptation. Fixed release: https://github.com/Abraheem2026/5G-Cross-Dataset-IDS-Adaptation/releases/tag/v1.0.1. Archived software record: https://doi.org/10.5281/zenodo.21678601.