Producing and Validating a Synthetic 5G DDoS Dataset

RQ: How to validate dataset quality for DDoS detection task in 5G network and does dataset quality metrics actually predict detection performance?

Karim Khalil

Reference Datasets

Three real 5G DDoS datasets exist. Two are the main targets:

Dataset	Infrastructure	Feature format	DDoS coverage
5G-NIDD [1]	Univ. of Oulu 5G testnet, 2 base stations	PCAP + Argues flows CSV	UDP/SYN/HTTP flo Slowrate, scans — **primary reference**
NCSRD-DS-5GDDoS [2]	Amarisoft Callbox, 3 cells, 9 UEs, mixed devices	Tabular CSV (radio/RAN counters)	SYN/UDP/ICMP/DN flood, GTP-U flood — **secondary reference**
5GAD-2022 [3]	free5GC software core, 4 interfaces	PCAP	Mostly 5G-core AP level attacks (NF impersonation, da exfiltration) — **limited DDoS scope**

[1] Siriwardhana, Yushan, et al. "Descriptor: 5G Wireless Network Intrusion Detection Dataset (5G-NIDD)." IE Descriptions (2025).
[2] Christopoulou, Maria, et al. "User terminals as attackers: An open dataset analysis of DDoS attacks in 5G networks." 2024 IEEE Conference on Standards for Communications and Networking (CSCN). IEEE, 2024.
[3] Coldwell, Cooper, et al. "Machine learning 5g attack detection in programmable logic." 2022 IEEE Globec Workshops (GC Wkshps). IEEE, 2022.

Bad Design Smell Analysis[4] (WhiffSuite)

- Before replication, we audit the existing reference output dataset against known structural flaws using WhiffSuite [4].
  - PortSniff - Is port alone a shortcut?: The model learns topology, not behaviour.
  - BackwardPacketsSniff - Does the server ever reply?:
  - CosineSniff - Is there any structure in the feature space?: Flows from different classes are geometrically indistinguishable. No learnable decis boundary exists.
  - NearestNeighboursSniff - Are labels locally consistent?: Minority class i surrounded by the majority class. Symptom of class imbalance
  - SingleFeatureEfficacySniff — Can any single feature classify alone?: No feature should dominate this strongly in a realistic dataset.

[4] Flood, Robert, et al. "Bad design smells in benchmark nids datasets." 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P). IEEE, 2024.

GAN-based Evaluation metric: TRTR / TSTR / TRTS

- Three evaluation protocols using Random Forest, XGBoost, and MLP — evaluated with 5-fold stratified cross-validation, reported as macro-F1.[1][2]:
  - TRTR — Train Real, Test Real: baseline detection ceiling on each reference dataset
  - TSTR — Train Synthetic, Test Real: does synthetic data substitute for real training data?
  - TRTS — Train Real, Test Synthetic: does the synthetic data cover the same attack space as real data?

Key quantity: F1 gap - $\Delta F1 = F1(TRTR) - F1(TSTR)$. Lower is better; this is th ground-truth quality score for each synthetic variant.

[1] **5-fold stratified CV**: Dataset split into 5 folds; each fold tested once while training on the rest. Stratified = class ratio preserved in every fold. Average score reported.
[2] **Macro-F1**: F1 computed per class then averaged equally — penalises models that ignore the minority class.

Dataset Replications

Goal: Produce synthetic datasets that mirror the statistical properties of reference datasets

- Target 1 - 5G-NIDD (full replication)
  - Match Bening traffic distribution
  - Match attack families: UDP flood, SYN flood, HTTP flood, Slowrate DoS
  - Feature schema: CICFlowMeter — directly comparable, no alignment needed
- Target 2 - NCSRD-DS-5GDDoS (full replication)
  - Match Bening traffic distribution
  - Match attack families: SYN/UDP/ICMP/DNS flood, GTP-U flood
  - Feature schema: radio/RAN counters - requires separate feature extraction pipeline (poss
- Target 3 - 5GAD-2022 (partial — benign traffic only)
  - 5GAD's attacks are 5G-core API exploits outside DDoSimu5G's scope
  - Replicate the benign traffic profiles and overlay DDoSimu5G volumetric DDoS attacks
  - Serves as a cross-topology generalisation test rather than a full replication

Replication validation: for each target, TSTR and TRTS F1 gap ≤ μ, relative to the TRT ceiling of that reference dataset.

Controlled Quality Gradient

DDoSimu5G's controllability is the experimental lever. We generate ~10-20 variants by changing one dimension at a time:

- **Benign diversity**: Vary the benign traffic profile mix.
- **Temporal realism**: Vary the attack temporal profile.
- **Attack coverage**: Vary the number of DDoS families.
- **Traffic protocol mix**: Vary the ratio of UDP, TCP, and HTTP flows. Changes inter-feature covariance structure (e.g. TCP flag correlations, flow duration patterns) wh keeping individual feature ranges similar.
- **Artificially injected shortcut feature:** (tests whether metrics catch leakage) We a synthetic column to the dataset that partially/perfectly class label.
- **Duplicate injection** (0.5%, 1%, 5%, 10%)
- **Label noise** (1%, 5%, 10%, 20% flipped labels)

Dataset Quality Metrics

Some metrics could be used to judge dataset quality —but are they actually predictive of detection performance?

Metric	What it measures
Hurst exponent gap (ΔH)	Self-similarity of traffic ti series (Does traffic temporal stru matter) [5]
Mean per-feature JSD	Per-feature marginal distribution match (if featur distributions match well) [1]
MI-matrix Frobenius distance	Inter-feature dependency structure match (does prese inter-feature relationships matter)
Max single-feature AUC	Shortcut / artefact detect
Cross-class duplicate rate	Data leakage detection [4]
C2ST accuracy	Global distributional indistinguishability [2]

Q: Which of these correlate scores (ρ) with the TSTR F1 ga across a controlled quality gradient of synthetic variants? the first empirical test of these metrics against a real dete task.

[1] Cover & Thomas, Elements of Information Theory, 2nd ed., Wiley, 2006.
[2] Lopez-Paz & Oquab, "Revisiting Classifier Two-Sample Tests," ICLR, 2017.
[4] Flood, Robert, et al. "Bad design smells in benchmark nids datasets." 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P). IEEE, 2024.
[5] Goldschmidt, Patrik, and Daniela Chudá. "Network intrusion datasets: A survey, limitations, and recommendations." Computers & Security 156 (2025): 104510.

Measure Metrics Score vs Predictio Performance

- For each of the ~10-20 variants we have:
  - A metric score (JSD, Hurst, AUC, duplicate rate, C2ST...)
  - A ground-truth quality score: $\Delta F1 = F1(TRTR) - F1(TSTR)$
- What we do:
  Rank all variants by metric score. Rank all variants by ΔF1.
  Compute correlation score (ρ) between the two rankings.
- What the result tells us:

ρ close to 1	Metric reliably predicts detection performance ✓
ρ near 0	Metric has no predictive value ✗

Methodology Overview

- **Step 1 - Audit reference datasets**
  Apply bad design smell tests (PortSniff, BackwardPacketsSniff, CosineSniff, NearestNeighboursSniff, SingleFeatureEfficacySniff) to 5G-NIDD, NCSRD, and 5GAD. Understand what structural properties real data have before attempting replication.
- **Step 2 - Replicate reference datasets (DDoSimu5G)**
  Configure DDoSimu5G to match each reference: Attack families, benign traffic profiles. Extract flows with CICFlowMeter. Produce one synthetic dataset D* per reference R.
- **Step 3 - Establish detection baselines**
  For each pair (D*, R): measure TRTR on R (detection ceiling), TSTR (train on D*, test on R), TRTS (train on R on D*). Compute ΔF1 = F1(TRTR) − F1(TSTR) for each D*.
- **Step 4 - Generate controlled quality gradient**
  Using DDoSimu5G's parameter controllability, produce 10–20 synthetic variants per reference, each delibe degrading or altering one quality dimension. Each variant has a known ΔF1.
- **Step 5 - Compute quality metrics on all datasets**
  For every D* and every gradient variant: compute Hurst exponent gap, per-feature JSD, MI-matrix distance single-feature AUC, duplicate rate, C2ST accuracy.
- **Step 6 - Correlate metrics against detection performance**
  For each metric, compute score (ρ) against ΔF1 across all variants. Answer: which metrics actually predict detection performance — and which do not?

Preliminary Results - WhiffSuite Bad-Smell Audit

- **5G-NIDD** *(1,215,890 flows, Argus features)*
  - Class balance: 61% malicious / 39% benign. Moderate, manageable ✓
  - Port: 80% of malicious flows use background ports. Port is a weak separator, model must learn behaviour ✓
  - Backward packets: 72% malicious / 90% benign have zero. Affects both classes equ not a shortcut but a capture artefact ✗
  - Zero-variance: 28/69 features (41%). Argus fields never populated in this 5G deployment ✗
  - 0 duplicate rows ✓
- **DDoSimu5G** *(SoftwareX - our own output) (18,314 flows, CICFlowMeter featur*
  - Class balance: 97.5% malicious / 2.5% benign (caused by TCP Syn). Severe 40:1 imbalance ✗
  - Port: dst_port alone achieves F1 = 0.997. Trivial shortcut ✗
  - Backward packets: 99.96% of malicious flows have zero. Simulation artefact (echoM = false + SYN flood) ✗
  - Feature space: 99.98% of flows in one cluster (Due to imbalance). No learnable geometry ✗
  - Zero-variance: 20/83 features. SYN flood never exercises TCP session mechanics ✗
- **Takeaway:** 5G-NIDD is structurally clean. Our current simulator output has 5 confirmed bad smells. All traced to SYN flood dominance and simulation configuration. These are the targets for the replication fixes.