

Supplementary Information

A Multi-Seed Evaluation of Explanation Faithfulness and Stability under Adversarial Evasion in IoT/IIoT Intrusion Detection

S. M. Nihal Ahmed, Afrim Hossen Khan, Sabikun Nahar Sinthia

Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

This document reports the experimental configuration and four secondary analyses that support the main text: the effect of individual preprocessing choices (Section S2), an adversarial-training sweep (Section S3), a multiclass attack-family analysis (Section S4) and resource profiling (Section S5). Unless stated otherwise, notation, datasets, detectors and attacks are as in the main text. All values are computed from the archived seed-level results.

S1 Experimental configuration

Table S1: Experimental configuration.

Item	Setting
Training seeds	20 per detector on the primary partition (17, 29, 41, 53, 67, 83, 97, 101, 113, 127, 131, 149, 163, 179, 191, 211, 223, 239, 251, 269); the first 5 for the sensitivity analyses
Partitions	family-stratified 70/15/15 with seeds 2026 (primary), 2027 and 2028; label conflicts retained (seed 2026); global and within-family temporal 70/15/15 (frame time for Edge-IIoTset, row order for ToN-IoT)
MLP	128–64–32, ReLU, dropout 0.2, Adam 10^{-3} , batch 2048, ≤ 30 epochs, patience 5
FGSM-AT / PGD-AT	$\epsilon_t = 0.01$, $\lambda = 1 / \epsilon_t = 0.03$, 5 steps, $\lambda = 0.5$
XGBoost	400 trees, depth 6, $\eta = 0.1$, subsample 0.8, colsample 0.8, early stopping 30
Thresholds	validation macro-F1 optimum; 0.5; validation false-positive budgets 0.1, 0.5, 1, 2, 5%
Attack budgets	$\epsilon \in \{0.01, 0.03, 0.05\}$ (normalised ℓ_∞ , all threat models); constrained padding additionally $\epsilon \in \{0.1, 0.25, 0.5, 1\}$
PGD / FGSM	10 steps, step $2.5\epsilon/10$, random start / one step of size ϵ
Black-box	random search on box vertices, 300 queries with early stop; 3,000 queries without early stop (XGBoost)
Coordinate search	split thresholds of the trained ensemble (XGBoost) or a 33-point grid (MLP), up to 3 passes
Attack subset	2,000 malicious test records, family-stratified
Explanation subset	300 benign + 300 malicious test records; attacked versions of the 300 malicious
SHAP	Deep SHAP (MLP), interventional Tree SHAP (XGBoost); 100 training records as background; KernelSHAP with 50 background records and 2,048 samples (configuration analysis)
LIME	tabular, no discretisation, sampling around the instance, 1,000 samples, kernel width $0.75\sqrt{p}$, 2,000 training records for feature statistics; configuration analysis: 1,000 or 5,000 samples, kernel width 0.25, 0.75 or $2\sqrt{p}$, quartile discretisation
Faithfulness	$k \in \{10, 20, 30, 40, 50\}\%$ of groups; references median, marginal (16 draws), k -NN in-fill ($k=8$, pool 5,000); random control (3 orderings)
Stability	Jaccard@3, Jaccard@5, Jaccard@[$0.2G + 0.5$], Spearman ρ , signed cosine, top-1 agreement
Repeated calls	first 5 seeds, 100 records: LIME $\times 5$ sampling seeds; SHAP $\times 3$ background draws
Hardware	Kaggle, NVIDIA Tesla T4 GPU, 4 virtual CPUs

S2 Effect of individual preprocessing choices

Table S2 starts from a simpler reference pipeline—a narrow numeric feature list (8 fields for ToN-IoT, 22 for Edge-IIoTset), linear scaling clipped at the 1st and 99th training percentiles, and no de-duplication—and changes one factor at a time until the pipeline of the main text is reached. Each step uses five seeds.

On Edge-IIoTset, the reference pipeline gives an MLP macro-F1 of 0.673 ± 0.007 at threshold 0.5, with benign precision 0.370. Replacing the clipped linear scaling by log/min–max scaling (P1) raises it to 0.905. The subsequent changes (feature set, categorical fields, de-duplication) change it by less than 0.002. XGBoost, whose splits are invariant to monotone rescaling, scores 0.900 under both scalings, which confirms that the deficit is a preprocessing effect specific to the neural model: many Edge-IIoTset fields are heavy-tailed, and their discriminative range lies above the training 99th percentile, so clipping maps it onto a single value. Threshold tuning adds a further 0.04–0.05 macro-F1, at the cost of the false-positive rate discussed in the main text. On ToN-IoT, de-duplication (P3→P4) lowers MLP macro-F1 from 0.983 to 0.967 and raises the false-positive rate at threshold 0.5 from 0.019 to 0.043, which shows that duplicated records spanning the partitions inflate clean scores.

Table S2: Macro-F1 from a reference pipeline to the pipeline of the main text, changing one factor per step (5 seeds; macro-F1 at threshold 0.5 / at the validation-tuned threshold).

Step	Edge-IIoTset		ToN-IoT	
	MLP	XGBoost	MLP	XGBoost
P0: reference pipeline (narrow feature list, 1–99% clipped scaling, duplicates kept)	0.673 / 0.761	0.900 / 0.946	0.968 / 0.976	0.984 / 0.985
P1: + log/min–max scaling	0.905 / 0.946	0.900 / 0.946	0.974 / 0.976	0.985 / 0.985
P2: + full numeric feature set	0.905 / 0.947	0.898 / 0.947	0.974 / 0.975	0.995 / 0.996
P3: + categorical protocol fields	0.905 / 0.946	0.900 / 0.947	0.983 / 0.985	0.997 / 0.997
P4: + de-duplication (main pipeline)	0.904 / 0.951	0.909 / 0.953	0.967 / 0.969	0.996 / 0.995

S3 Adversarial-training sweep

Figure S1 varies the PGD adversarial-training budget $\epsilon_t \in \{0.01, 0.03, 0.05\}$ and the clean–adversarial loss weight $\lambda \in \{0.5, 1\}$ (five seeds). On ToN-IoT, clean macro-F1 stays at 0.965–0.971 for all six settings, while unconstrained PGD success at $\epsilon = 0.05$ falls from 0.305 (no adversarial training) to 0.011–0.041. On Edge-IIoTset, macro-F1 stays at 0.947–0.952 except at the strongest setting ($\epsilon_t = 0.05$, $\lambda = 1$: 0.897), where unconstrained success falls from 0.291 to 0.044. Constrained success is near zero for every setting at this budget. In this pipeline, adversarial training therefore costs little or no clean accuracy.

S4 Multiclass attack-family analysis

An MLP and an XGBoost model with softmax outputs over all families are trained with five seeds. The MLP is attacked on the 300 explained malicious records with untargeted PGD (maximising the loss of the true family) and with targeted PGD toward the benign class, under all three threat models at $\epsilon = 0.05$. Success is reported per family, and the stability of the SHAP attribution for the true family is computed between clean and attacked records.

Table S3 summarises the results. Family-level macro-F1 is 0.74 (MLP) and 0.95 (XGBoost) on ToN-IoT, and 0.75 and 0.86 on Edge-IIoTset. Collapsing the multiclass predictions to benign/malicious recovers 0.91–0.996, so a binary score hides substantial confusion between families, notably password, uploading and XSS on Edge-IIoTset (per-family F1 0.37–0.57) and MITM and backdoor on ToN-IoT (0.36, 0.48). Under the unconstrained attack, 90% (ToN-IoT) and 65% (Edge-IIoTset) of correctly classified attacks can be moved to some other family, and 89% and 46% specifically to the benign class. Susceptibility differs sharply between families: on Edge-IIoTset, DDoS-UDP, DDoS-ICMP and MITM records are never redirected to benign,

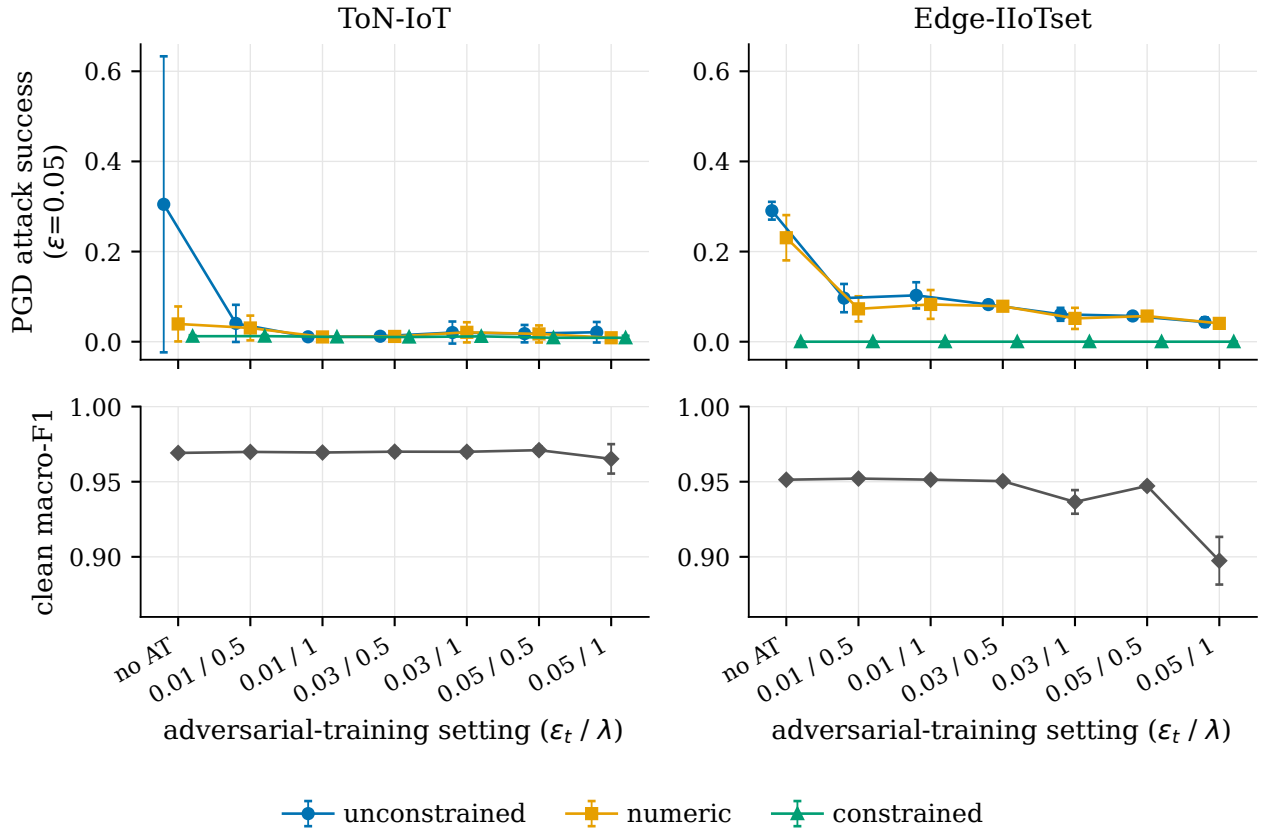


Figure S1: Robustness–accuracy trade-off of PGD adversarial training (5 seeds; mean $\pm$ SD). Top: PGD attack success at $\epsilon = 0.05$ under each threat model; bottom: clean macro-F1. The horizontal axis gives ϵ_t and λ ; “no AT” is the clean MLP. Points are offset horizontally so that overlapping series remain visible.

whereas 98% of uploading records are. Under the constrained model, no record of any family is redirected to benign on either dataset. On ToN-IoT, however, 31% of attacks are moved to a *different attack family* (95% of backdoor and 93% of password records). Where the predicted family selects the response playbook, this is an evasion of the response rather than of the alert. The SHAP attribution for the true family is as stable under the constrained attack as in the binary task (Spearman $\rho = 0.98$ on ToN-IoT, 1.00 on Edge-IIoTset) and changes more under the unconstrained attack ($\rho = 0.39$ – 0.48).

Table S3: Multiclass analysis (5 seeds). Family-level macro-F1, and attack success of untargeted PGD (any family change) and targeted PGD toward the benign class on the 300 explained malicious records (MLP, $\epsilon = 0.05$).

Dataset	Macro-F1		Unconstrained		Numeric-only		Constrained	
	MLP	XGBoost	untarg.	benign	untarg.	benign	untarg.	benign
ToN-IoT	0.737 ± 0.008	0.954 ± 0.001	0.90	0.89	0.71	0.02	0.31	0.00
Edge-IIoTset	0.752 ± 0.022	0.862 ± 0.000	0.65	0.46	0.58	0.28	0.01	0.00

S5 Runtime and memory

Table S4 reports per-record cost on a single CPU thread, used as a proxy for an edge gateway, and on the T4 GPU. Inference takes 0.2–0.6 ms per record. SHAP costs 12–26 ms per record and LIME 30–61 ms on one CPU thread; in the batched main runs on the GPU, which explain many records per call, SHAP took 5–19 ms and LIME 9–23 ms per record. Explaining 100 records increased process memory by at most 22 MB (memory allocated by the explainers themselves ≤ 7.5 MB; GPU memory ≤ 19 MB). The trained MLPs occupy 68–85 KB and the XGBoost models 0.66–1.1 MB. The process footprint of about 2 GB is dominated by the software libraries and the loaded data. These measurements are taken on a cloud machine with recorded traffic; they bound the cost of explanation per alert but do not establish throughput on a specific embedded device or under continuous traffic.

Table S4: Runtime (ms per record) and memory for 100 records. CPU = one thread; GPU = NVIDIA T4 (MLP only). BB / PGD = constrained black-box and PGD attack cost. Δ Mem = peak increase of process memory during SHAP/LIME (MB); Expl. = peak memory allocated by the explainer (MB); GPU = peak GPU memory (MB).

Dataset	Detector	CPU (1 thread)			GPU	Attack	Memory (MB)			Model (KB)
		Infer.	SHAP	LIME			Δ Mem	Expl.	GPU	
ToN-IoT	MLP	0.23	15.7	57.3	15.2 / 52.6	8.5 / 0.16	22.4	5.1	19.0	85.0
	XGBoost	0.57	25.8	61.3	–	11.1 / –	0.0	7.5	–	1132
Edge-IIoTset	MLP	0.20	13.4	29.8	14.6 / 29.7	7.1 / 0.15	0.0	3.1	18.8	67.5
	XGBoost	0.46	11.7	33.3	–	8.4 / –	1.9	4.4	–	661