Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable AI Audit of Temporal Signatures in Botnetand Multi-Vector Cyber-Attacks

Gizachew Teshome Habtamu

gizachew.teshome@ddu.edu.et

Dire Dawa University

Shahzad Ahmed

vCISO at VLBPO Limited [HQ at Jamaica]

Research Article

Keywords: Explainable AI, Intrusion detection, Botnet, SHAP, LIME, Temporal features, CIC-IDS2017, CTU-13

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10291522/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Explainable AI Audit of Temporal Signatures in Botnet and Multi-Vector Cyber-Attacks

Gizachew Teshome Habtamu* · **Shahzad Ahmed**

Received: date / Accepted: date

Abstract Machine learning based intrusion detection systems achieve excellent accuracy but remain black boxes, and it is unclear whether the evidence they rely on transfers across attack domains. This paper applies a three-layer explainable AI framework, comprising SHAP, LIME, and permutation importance, to seven classifiers trained on a nine-feature temporal representation harmonized across CIC-IDS2017 (multi-vector attack traffic) and CTU-13 (botnet traffic). All non-kernel detectors exceed 0.99 ROC-AUC using temporal features alone. Explanations, however, reveal domain-specific reliance: multi-vector attacks are dominated by volumetric and packet-size evidence, whereas botnets are exposed almost entirely by inter-arrival-time regularity, and cross-dataset importance rankings correlate weakly (mean Spearman ρ of $+0.01$). By taking the minimum importance across datasets, algorithms, and XAI methods, we distill a consensus set of robust temporal signatures, led by inter-arrival-time variability, source-byte ratio, and average packet size, that remains informative in both domains. The results demonstrate that XAI can convert black box detectors into auditable, transferable temporal evidence for trustworthy intrusion detection.

Keywords Explainable AI · Intrusion detection · Botnet · SHAP · LIME · Temporal features · CIC-IDS2017 · CTU-13

G. T. Habtamu*
Department of Computer Science, Dire Dawa University Institute of Technology, Dire Dawa, Ethiopia
E-mail: gizachew.teshome@ddu.edu.et
*Corresponding author.

S. Ahmed
vCISO at VLBPO Limited (HQ at Jamaica), Department of IT & Security, Karachi, Pakistan
E-mail: shahzad@vlbpo.com

1 Introduction

Modern networks are subjected to a continuous stream of increasingly automated attacks. Distributed denial-of-service (DDoS) campaigns, brute-force credential attacks, web exploitation, and large-scale botnets now co-exist on the same infrastructure, and the economic and societal cost of successful intrusions continues to grow. Machine learning (ML) based intrusion detection systems (IDS) have become the de facto response to this threat landscape: given labelled network flows, ensemble and deep models routinely report detection accuracies above 99% on benchmark datasets [1,5,11].

This success conceals two problems. First, the best-performing models—random forests, gradient-boosted trees, and neural networks—are black boxes. A security operations centre (SOC) analyst who receives an alert cannot see why a flow was flagged, cannot audit the model for shortcut learning (e.g., memorizing a port number used only in the capture environment), and cannot justify an automated response to stakeholders or regulators [5,6]. Second, and more subtly, near-perfect benchmark accuracy says nothing about whether the model has learned generalizable attack behavior or dataset-specific artefacts: recent cross-dataset audits show that detectors with ∼99% within-dataset accuracy can collapse to near-random performance when moved to another network [10].

Explainable AI (XAI) techniques such as SHAP [3] and LIME [4] promise to open the black box by attributing each prediction to input features. Yet most XAI-for-IDS studies stop at producing a feature-importance plot for a single model on a single dataset [6,7]. Whether the highlighted features constitute robust temporal signatures—behavioral timing and volume patterns that remain discriminative across different attack families, different learning algorithms, and different datasets—remains an open question. This question matters because botnet traffic (machine-driven, periodic, command-and-control oriented) and multi-attack traffic (volumetric floods, scans, brute force) may plausibly rely on very different evidence, and a signature is only operationally useful if we know where it holds.

This paper addresses that gap. We deliberately restrict all models to a compact set of nine harmonized temporal features—flow duration, volume, rate, directionality, packet size, and inter-arrival-time (IAT) statistics—that can be computed from any flow exporter, and we interrogate seven diverse classifiers with three independent XAI layers on two heterogeneous benchmarks: CIC-IDS2017 [1] (fourteen multi-attack families) and CTU-13 [2] (real botnet captures). The contributions are:

- A cross-dataset, cross-algorithm, cross-method XAI protocol that harmonizes CIC-IDS2017 and CTU-13 into an identical nine-dimensional temporal feature space, enabling like-for-like comparison of explanations between multi-vector and botnet attacks (Fig. 1).
- Empirical evidence that temporal features alone support near-perfect detection (ROC-AUC ≥ 0.99 for six of seven algorithms on both datasets),

demonstrating that behavioural timing is sufficient evidence for detection without payload or port information.
- A quantitative demonstration that global feature-importance rankings do not transfer between attack domains (mean cross-dataset Spearman $\rho = +0.01$; 9 of 21 algorithm–method pairs negative), which cautions against naive reuse of XAI findings.
- A consensus robustness-floor methodology—taking the minimum normalized importance across datasets and aggregating across algorithms and XAI methods—that distils genuinely transferable signatures: IAT standard deviation, source-byte ratio, and average packet size emerge as the most robust temporal signatures, while IAT statistics alone expose botnets.

The remainder of the paper is organized as follows. Section 2 surveys related work; Section 3 details the datasets, feature harmonization, models, and XAI pipeline; Section 4 presents results; Section 5 discusses implications and limitations; Section 6 concludes.

2 Literature Review

2.1 Machine Learning for Intrusion Detection

Flow-based ML-IDS research has matured around public benchmarks. Sharafaldin et al. introduced CIC-IDS2017 with more than eighty CICFlowMeter features spanning benign traffic and fourteen attack families [1], and García et al. released CTU-13, thirteen real botnet scenarios with NetFlow labels [2]; both remain the reference corpora for multi-vector attack and botnet detection respectively. Recent surveys confirm that ensemble methods (random forest, boosting) and deep architectures dominate the accuracy leaderboards on these benchmarks [5, 9, 15]. Hnamte et al. proposed a two-stage LSTM-autoencoder that reaches state-of-the-art accuracy on CIC-IDS2017-class traffic [11], and lightweight tree-based models have recently been shown to match deep baselines on CTU-13 at a fraction of the cost while retaining interpretability [16]. However, several audits caution that reported accuracy can be inflated by dataset artefacts, duplicated flows, label errors, and identifier features such as ports; Cantone et al. showed that classifiers with near-perfect within-dataset accuracy degrade to almost random performance in cross-dataset evaluation on CIC-IDS2017/2018 and their corrected LycoS variants [10]. This motivates both artefact-resistant feature sets and post-hoc auditing of what models actually learn.

2.2 Temporal and Behavioral Features

Timing behavior has long been considered a strong botnet discriminator: command-and-control (C&C) beaconing produces regular inter-arrival patterns that are hard for an adversary to remove without degrading the channel.

Modern deep-learning botnet detectors implicitly exploit this structure; Kundu et al. combine deep models with model-explanation layers to detect and classify botnet traffic and confirm that flow-timing attributes carry a large share of the decision evidence [8]. Reviews of IoT-botnet detection similarly identify inter-arrival time, flow duration, and packet-size statistics as recurrent top features across dozens of studies [9]. On the multi-attack side, feature-analysis studies of CIC-IDS2017 consistently rank packet-length and volumetric statistics among the most informative attributes [10,15]. Yet temporal features are usually mixed with dozens of protocol-specific features, making it difficult to isolate their contribution, and few works test whether the same temporal representation discriminates both botnet and multi-attacks under an identical schema.

2.3 Explainable AI in Cybersecurity

SHAP [3] provides additive, game-theoretic attributions with exact solutions for tree ensembles; LIME [4] fits sparse local surrogates; permutation importance measures the performance drop when a feature is destroyed. A rapidly growing body of work applies these tools to IDS. Moustafa et al. survey explainable intrusion detection for IoT defence and catalogue open challenges around trust and deployment [5]. Gaspar et al. evaluate LIME and SHAP on a multilayer-perceptron IDS and complement the attributions with perturbation analysis and an analyst survey, finding that explanations materially aid alert triage [7]. Arreche et al. propose E-XAI, an end-to-end framework that systematically benchmarks black-box XAI techniques (chiefly SHAP and LIME) for network intrusion detection across models and datasets [6]. Oseni et al. embed SHAP into a deep IDS for IoT-enabled transportation networks to provide resilient, human-auditable detection [12], Khediri et al. couple SHAP attributions with LLM-generated natural-language descriptions to make explanations accessible to non-expert operators [13], and federated IDS deployments now cite explainability as a first-class requirement at the network edge [14].

2.4 Research Gap

Across this literature three limitations recur. First, explanations are almost always produced for one model on one dataset, so their stability across learning paradigms is unknown [6,7]. Second, XAI studies rarely control the feature space across corpora, so differences in explanations confuse feature-schema differences with genuine behavioral differences. Third, while cross-dataset generalization of accuracy has begun to receive attention [10], the cross-dataset generalization of explanations—which features remain important when the attack domain changes—has, to our knowledge, not been quantified. This paper fills that gap: it (i) harmonizes a botnet corpus and a multi-attack corpus into an identical temporal feature space, (ii) explains seven heterogeneous classifiers

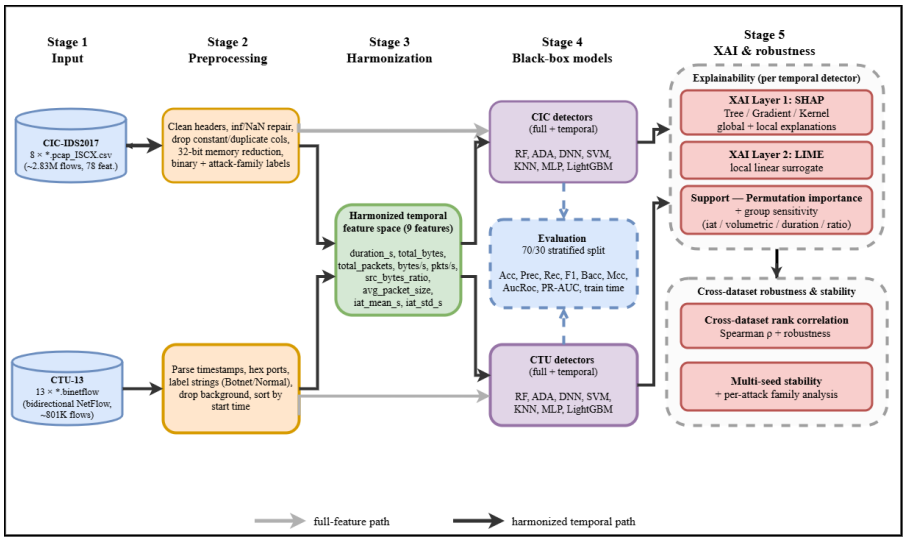


Fig. 1 Overview of the proposed five-stage methodology. Two heterogeneous corpora (CIC-IDS2017, CTU-13) are preprocessed and harmonized into a shared nine-feature temporal space; seven algorithms are trained per dataset (full + temporal variants) and evaluated with seven metrics; SHAP, LIME, and permutation/group-sensitivity layers are then applied, followed by cross-dataset rank-correlation, robustness-floor, and multi-seed stability analysis.

with three independent attribution methods, and (iii) quantifies cross-domain agreement of explanations to extract signatures that are robust by construction.

3 Methodology

Figure 1 summarizes the five-stage pipeline: raw corpora are ingested (Stage 1) and cleaned (Stage 2); a nine-feature harmonized temporal space is constructed (Stage 3); seven black-box detectors are trained per dataset on both the native full-feature path and the harmonized temporal path (Stage 4); and a three-layer XAI and robustness analysis interrogates every temporal detector (Stage 5).

3.1 Datasets and Preprocessing

CIC-IDS2017 (multi-vector attacks). All eight daily CICFlowMeter CSV files were concatenated (2,830,743 raw flows, 80 columns). Column headers were stripped of whitespace; $\pm\infty$ values in rate fields were median-imputed; non-numeric fields were coerced; exact duplicates and eight constant columns were dropped, yielding 2,574,264 flows. Labels were binarized ($y = 1$ for any attack) while the original family label (DoS Hulk, DDoS, PortScan, brute force, web attacks, Bot, etc.) was retained for per-family analysis. The class

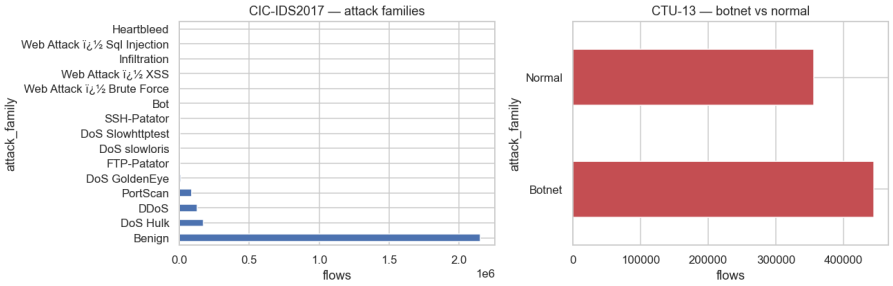


Fig. 2 Class composition of the two corpora. Left: CIC-IDS2017 attack families (heavily benign-dominated). Right: CTU-13 botnet vs. normal flows (near-balanced).

distribution is heavily imbalanced (2,148,386 benign flows; the largest attack family, DoS Hulk, has 172,849 flows).

CTU-13 (botnet). All thirteen `.binetflow` scenario files were parsed (801,132 flows after removing background traffic), keeping Botnet (444,699 flows, 55.5%) versus Normal (356,433 flows). Timestamps were parsed and flows sorted chronologically; hexadecimal ports were decoded and infinite and missing numerics were median-imputed. Figure 2 shows both label distributions.

3.2 Temporal Feature Engineering and Harmonization

Because the two corpora expose different native schemas (78 CICFlowMeter features vs. 8 usable NetFlow numerics), we engineered an identical nine-dimensional harmonized temporal representation computable from both: `duration_s`, `total_bytes`, `total_packets`, `bytes_per_s`, `packets_per_s`, `src_bytes_ratio` (directionality), `avg_packet_size`, `iat_mean_s`, and `iat_std_s`. For CIC-IDS2017 the IAT statistics derive from Flow IAT fields (converted to seconds); for CTU-13 they are derived from flow duration and packet counts. Safe division guarded zero denominators, and all rates were clipped of infinities.

Figures 3 and 4 show per-class $\log_{10}$ distributions. Two observations motivate the study. In CIC-IDS2017, attack flows ($y = 1$) concentrate in a narrow high mode of `iat_mean_s` and `iat_std_s` and a spiked `bytes_per_s` mode, reflecting machine-generated floods; benign traffic is broad and multimodal. In CTU-13, botnet flows form sharply peaked modes in `duration_s`, `bytes_per_s`, and especially `avg_packet_size` (a bimodal spike near 60-byte keep-alives and ∼1 kB transfers)—the visual fingerprint of automated command-and-control. The correlation heatmaps (Fig. 5) confirm the harmonized space is non-degenerate: only `total_bytes`/`total_packets` are strongly collinear in both corpora, while the IAT block correlates with duration, as expected.

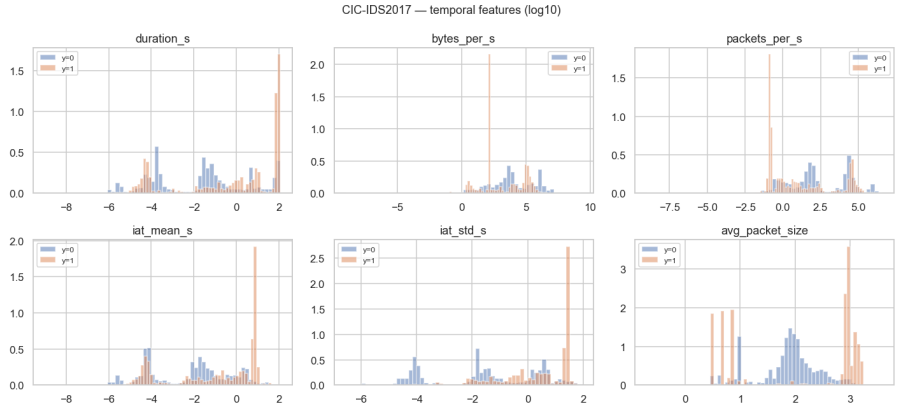


Fig. 3 CIC-IDS2017: per-class $\log_{10}$ distributions of the harmonized temporal features (benign $y = 0$ blue, attack $y = 1$ orange). Attacks concentrate in narrow high-IAT and fixed-rate modes characteristic of automated floods.

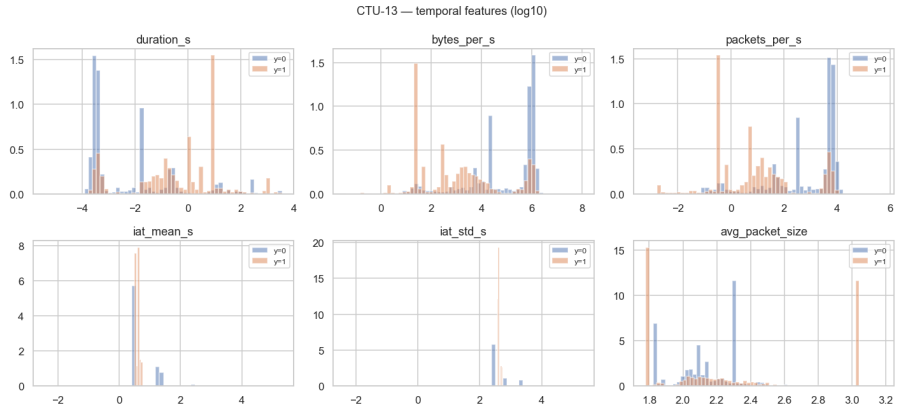


Fig. 4 CTU-13: per-class temporal distributions. Botnet flows exhibit spiked, low-entropy modes in duration, rate, and average packet size—the statistical shadow of periodic command-and-control beaconing.

3.3 Model Development

Seven classifiers spanning four learning paradigms serve as black boxes: Random Forest (RF, 300 trees, balanced class weights), AdaBoost (ADA, 200 depth-3 trees), a deep neural network (DNN; three dense layers of 128–64–32 units with batch normalization and dropout, falling back to an equivalent deep scikit-learn MLP when TensorFlow is unavailable), RBF-kernel SVM, k-nearest neighbors (KNN), a multilayer perceptron (MLP, 128–64), and LightGBM (histogram gradient boosting fallback). Distance- and gradient-based models (SVM, KNN, MLP, DNN) are wrapped in standardization pipelines; kernel models are capped at 40k–60k training rows for tractability. Each algorithm is trained on four variants {CIC-IDS2017, CTU-13} × {full native fea-

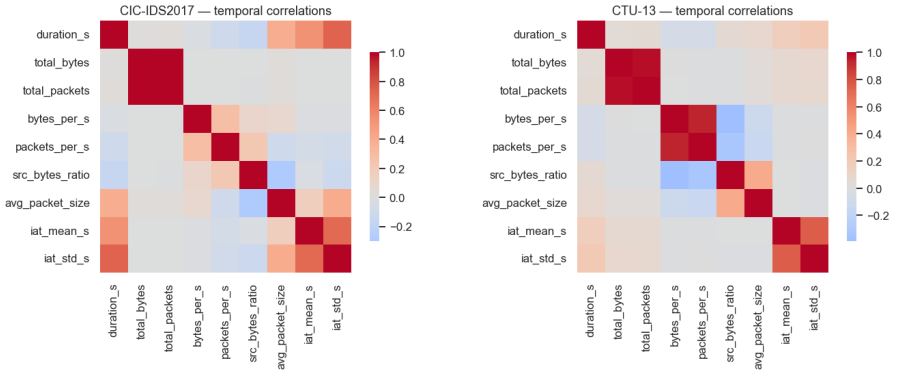


Fig. 5 Pearson correlations of the harmonized features. Apart from the expected `total_bytes/total_packets` pair, features are weakly correlated, so importance attributions are not dominated by redundancy.

tures, harmonized temporal features} with a stratified 70/30 split and fixed seed (42). The temporal variants share an identical feature space and are the ones whose explanations we compare; the full variants act as accuracy ceilings.

3.4 Explainable AI Application

Three independent attribution layers unveil each black box. (1) SHAP [3]: exact TreeExplainer for RF/LightGBM, GradientExplainer for the Keras DNN, and KernelExplainer (with capped background sets) for ADA, SVM, KNN, and MLP; 1,500 test samples per model. Global importance is the mean absolute SHAP value per feature. (2) LIME [4]: sparse local linear surrogates on 60 test instances per model (400 perturbations each); global importance is the mean absolute local weight. (3) Permutation importance: mean ROC-AUC drop over three shuffles of each feature on 4,000 test rows, plus a group-sensitivity pass that shuffles whole behavioral groups (IAT, volumetric, duration, ratio) simultaneously. All importances are normalized to sum to one per (algorithm, dataset, method) so they are comparable across settings.

3.5 Robust Temporal Signature Identification

For each algorithm a and method m we obtain importance vectors w^{CIC} and w^{CTU} over the nine shared features. Two complementary statistics quantify robustness:

$$\rho(a,m) = \mathrm{Spearman}\big(w^{\mathrm{CIC}}_{a,m},\ w^{\mathrm{CTU}}_{a,m}\big)\,, \tag{1}$$

which measures whether the ranking of evidence transfers between domains, and the per-feature robustness floor

$$r(f;a,m) = \min\big(w^{\mathrm{CIC}}_{a,m}(f),\ w^{\mathrm{CTU}}_{a,m}(f)\big)\,, \tag{2}$$

Table 1 Performance of the Harmonized Temporal Detectors

Dataset	Alg.	Acc	Prec	Rec	F1	Bacc	Mcc	AUC
CIC-IDS2017	RF	0.997	0.991	0.990	0.990	0.994	0.988	0.999
	ADA	0.991	0.983	0.965	0.974	0.981	0.968	0.999
	DNN	0.992	0.974	0.978	0.976	0.986	0.971	0.999
	SVM	0.923	0.781	0.746	0.763	0.852	0.718	0.962
	KNN	0.990	0.968	0.975	0.971	0.984	0.965	0.995
	MLP	0.991	0.979	0.968	0.973	0.982	0.968	0.999
	LightGBM	0.995	0.990	0.981	0.985	0.989	0.983	0.999
CTU-13	RF	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	ADA	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DNN	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SVM	0.872	0.907	0.857	0.882	0.874	0.744	0.928
	KNN	0.993	0.993	0.994	0.993	0.993	0.985	1.000
	MLP	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	LightGBM	1.000	1.000	1.000	1.000	1.000	1.000	1.000

which is high only if feature f is important in both domains. Averaging $r(f)$ over the seven algorithms and three XAI methods yields a consensus score per feature; features with high mean and low variance constitute the robust temporal signatures. Stability across five random seeds was additionally verified for the tree ensembles.

4 Results and Analysis

4.1 Model Performance

Table 1 reports the performance of all seven detectors on the harmonized temporal features. Six of seven algorithms exceed 0.99 ROC-AUC on both corpora using only nine timing/volume statistics; the tree ensembles are essentially perfect on CTU-13 (RF, ADA, LightGBM: 1.000 across all metrics). Only the capped RBF-SVM lags (AUC 0.962/0.928), a known consequence of subsampled kernel training rather than of the feature space. Figure 6 summarizes ROC-AUC across all algorithm × dataset × mode combinations: moving from ∼78 native features to 9 temporal features costs at most 0.001–0.005 AUC for the non-kernel models, i.e., the temporal representation preserves virtually all detectable signal. The ROC curves (Fig. 7) and confusion matrices (Fig. 8) corroborate this: RF misclassifies only 1,177 of 644,516 benign and 1,317 of 127,764 attack flows on CIC-IDS2017 (false-positive rate 0.18%, recall 98.97%), and LightGBM performs comparably (1,251 false alarms). These results establish that the black boxes are competent, a precondition for their explanations to be meaningful.

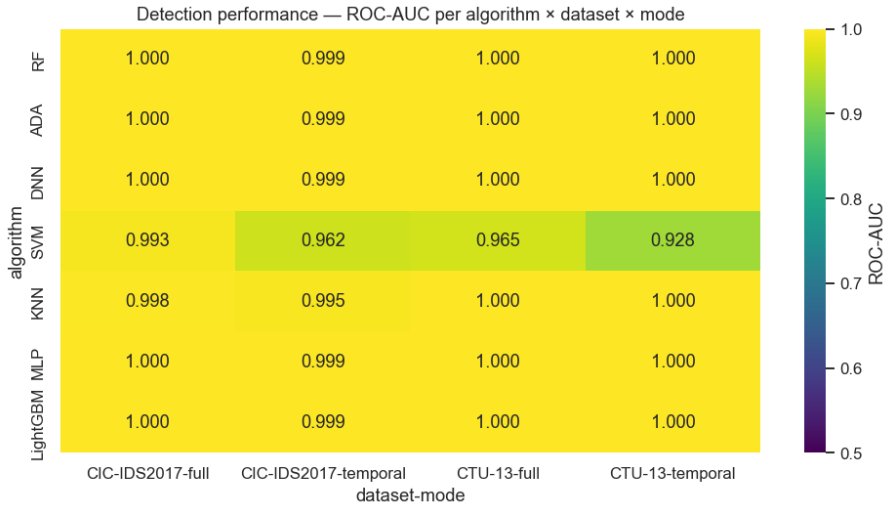


Fig. 6 ROC-AUC for every algorithm × dataset × feature mode. Reducing ∼78 native features to 9 harmonized temporal features costs almost no discriminative power for non-kernel models.

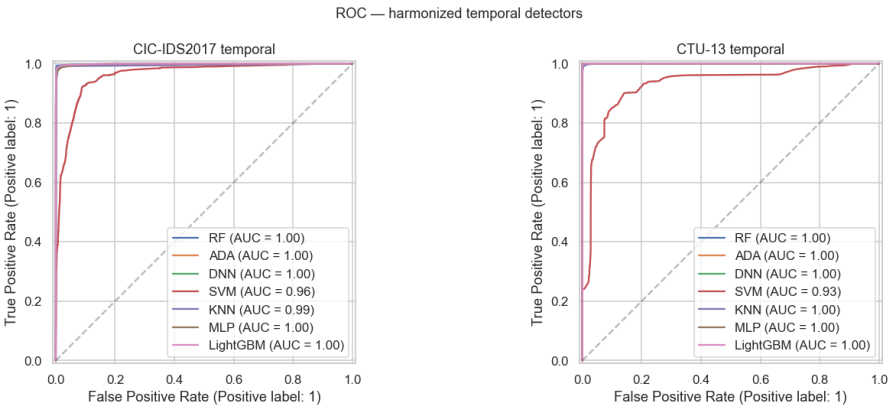


Fig. 7 ROC curves of the harmonized temporal detectors. All curves except the capped SVM hug the top-left corner on both corpora.

4.2 Global XAI Feature Importance

Before comparing domains, we validate the SHAP layer on the strongest full-feature model. Figure 9 shows the global mean |SHAP| ranking of the RF trained on all native CIC-IDS2017 features. Destination Port ranks first a classic dataset artefact, since attack services in the testbed occupy characteristic ports followed by a coherent block of backward-direction size features Init_Win_Bytes_Backward, Bwd_Packet_Length_Std, Subflow_Bwd_Bytes, Average_Packet_Size. Two insights follow. First, XAI immediately exposes

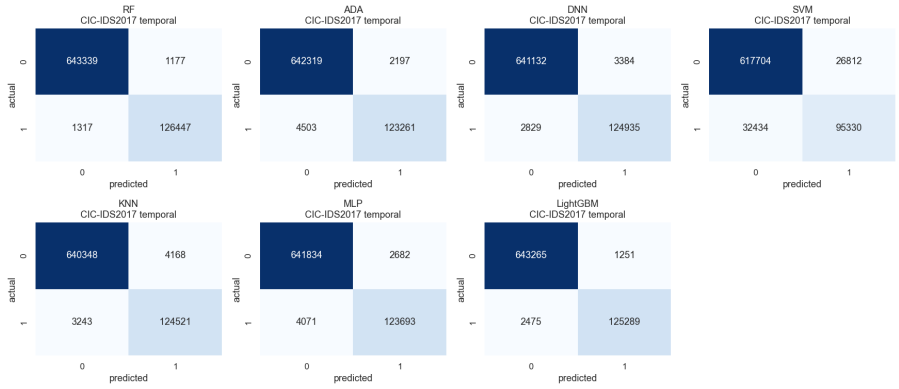


Fig. 8 Confusion matrices of the seven temporal detectors on CIC-IDS2017 (193k test attack flows, 645k benign). RF and LightGBM keep both error types below 0.2 % / 2 %, confirming that temporal evidence alone supports operational-grade detection.

the port shortcut that inflates full-feature benchmarks and has been independently blamed for poor cross-dataset generalization [10], justifying our decision to exclude identifiers from the harmonized space. Second, even in the unconstrained model, size/volume dispersion features—which our harmonized space retains as `avg_packet_size` and volumetric features—carry most of the legitimate behavioral signal: nine of the top twelve features are packet-length statistics.

4.3 Temporal Signature Analysis Across Algorithms and Attacks

Figure 10 presents the central evidence: normalized SHAP importance of each temporal feature for all seven algorithms on both corpora. The two panels tell strikingly different stories. On CIC-IDS2017 (multi-vector attacks), every algorithm concentrates mass on `avg_packet_size` (0.17–0.34) and `src_bytes_ratio` (0.15–0.24): volumetric floods and scans manipulate packet size and are heavily asymmetric (a DoS victim answers little; a scanner receives little). On CTU-13 (botnet), the same algorithms pivot almost entirely to the IAT block—`iat_mean_s` reaches 0.44–0.99 for RF, ADA, and LightGBM, and `iat_std_s` reaches 0.35–0.97 for DNN, KNN, and MLP—while packet-size features collapse below 0.13. Permutation importance (Fig. 11), computed by a completely different mechanism (AUC destruction rather than attribution), reproduces the same dichotomy: on CTU-13, shuffling `iat_mean_s` alone destroys 70–99.9% of the recoverable AUC for the boosting models, whereas on CIC-IDS2017 importance spreads over `total_bytes`, `src_bytes_ratio`, and `avg_packet_size`. The agreement of two independent XAI layers across seven model families is strong evidence that this is a property of the attacks, not of any model.

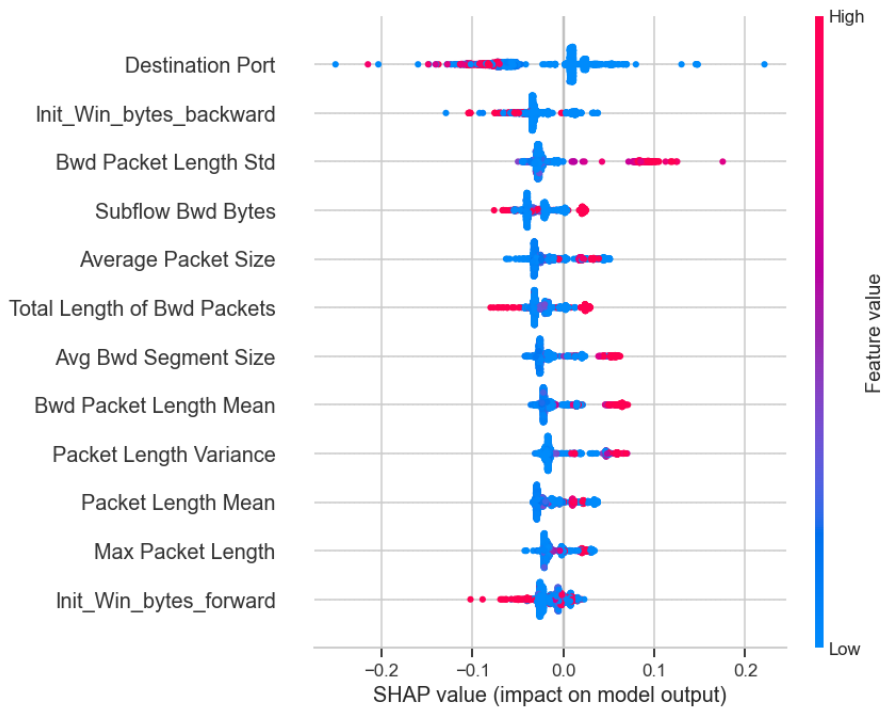


Fig. 9 Global SHAP importance (mean |SHAP|) of the full-feature RF on CIC-IDS2017. XAI exposes the Destination-Port shortcut and shows that legitimate signal concentrates in packet-size dispersion features.

The behavioral-group sensitivity test (Fig. 12) makes the contrast operational. On CTU-13, shuffling the IAT group costs every algorithm 0.06–0.54 AUC—for LightGBM the drop is 0.54, i.e., the detector falls to near-random—while destroying the volumetric, duration, or ratio groups barely moves most models. On CIC-IDS2017 the profile inverts: the volumetric group is the most damaging for all seven algorithms (drops up to 0.49 for MLP), with IAT and duration mattering for some models only. In other words, timing regularity is the botnet signature; volume/size anomaly is the multi-attack signature.

Per-attack-family resolution (Fig. 13, RF SHAP on CIC-IDS2017) explains why each feature matters and grounds the attributions in attack mechanics. `avg_packet_size` peaks for FTP-Patator (0.32) and DDoS (0.27), whose tooling emits uniform-size packets; `total_bytes` dominate PortScan (0.29) and DoS Hulk (0.27), where flows are extremely small or extremely large respectively; `iat_mean_s` and `packets_per_s` rise precisely for the slow-rate DoS family (slowloris 0.18, Slowhttptest 0.22/0.19), which is defined by pathological timing; and `iat_std_s` peaks for web brute force (0.20), where scripted retries produce low-variance rhythms. Notably, the intra-CIC Bot family already shifts weight toward `src_bytes_ratio` (0.18) and `avg_packet_size` (0.20), part-way toward the CTU-13 botnet profile. The importance pattern

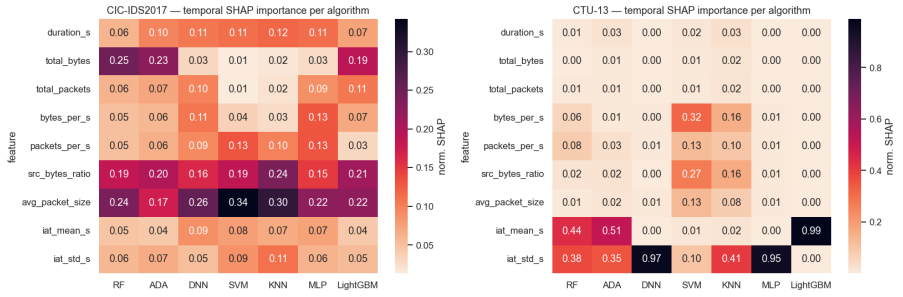


Fig. 10 Normalized SHAP importance per algorithm. Multi-vector attacks (left) are explained by packet size and directionality; botnets (right) are explained almost exclusively by inter-arrival-time statistics—the statistical trace of command-and-control periodicity.

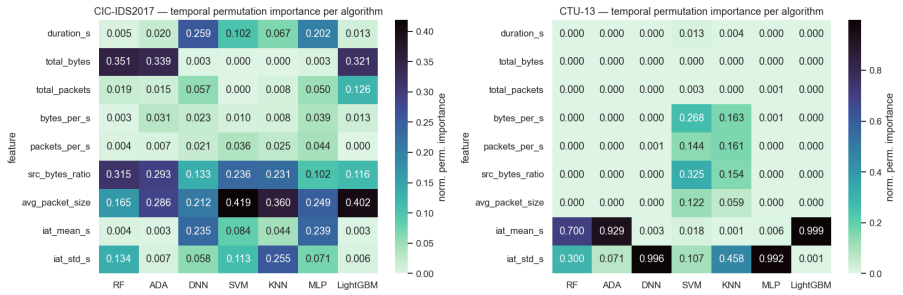


Fig. 11 Normalized permutation importance (AUC drop). An independent attribution mechanism reproduces the SHAP dichotomy: IAT features are single points of failure for botnet detection.

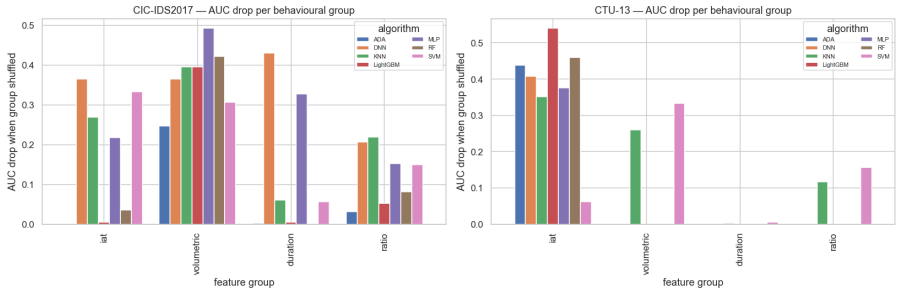


Fig. 12 Group-sensitivity: AUC drop when an entire behavioral feature group is shuffled. CIC-IDS2017 detectors depend on volumetric evidence; CTU-13 detectors depend critically on the IAT group.

therefore tracks attack semantics feature-by-feature, which is exactly the property required of a signature.

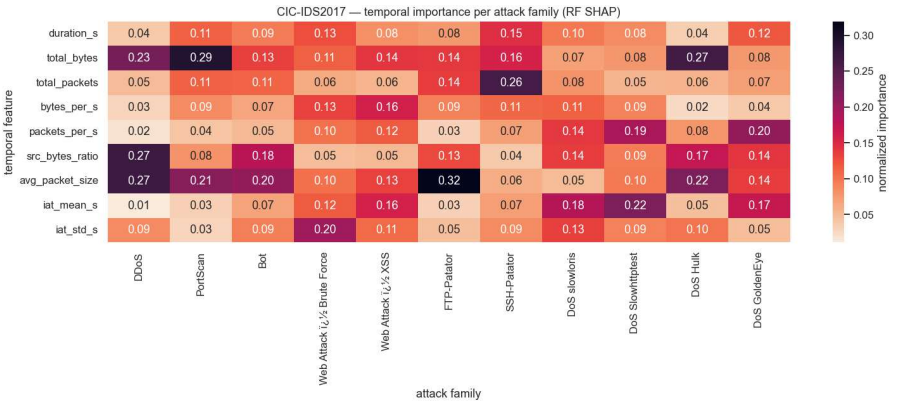


Fig. 13 RF SHAP importance resolved per CIC-IDS2017 attack family. Each family's dominant feature matches its mechanism: volumetric floods → bytes/packet size; slow-rate DoS → IAT mean and packet rate; scripted brute force → IAT regularity.

4.4 Cross-Dataset Robustness and Consensus Signatures

Do importance rankings transfer between domains? Figure 14 answers negatively: the cross-dataset Spearman correlation averages only +0.010 over the 21 algorithm method pairs, and 9 of 21 pairs are negative (RF–SHAP: −0.78; ADA–Permutation: −0.61). This mirrors, at the level of explanations, the accuracy-generalization failures recently reported for cross-dataset NIDS evaluation [10]: a feature ranking harvested on one benchmark can be actively misleading on another, because botnet and multi-vector attacks invert the importance ordering of the very same feature space. Only the instance-based KNN retains moderately positive transfer ($\rho = 0.50$–0.65), consistent with its reliance on broad geometric structure rather than sharp decision boundaries.

Rank transfer failing does not mean no feature transfers. Figure 15 scatters each feature's CIC importance against its CTU importance for the three methods: robust features lie near the diagonal with high values, and the off-diagonal outliers (`iat_std_s`, `iat_mean_s` high in CTU only; `avg_packet_size` high in CIC only) are exactly the domain specialists identified above. Applying the robustness floor of (2) and averaging over all 21 settings yields the consensus ranking of Fig. 16: `iat_std_s` (mean floor ≈ 0.09) leads under every individual method (LIME 0.103, permutation 0.090, SHAP 0.063), followed by `src_bytes_ratio` (≈ 0.06) and `avg_packet_size` (≈ 0.04), with `packets_per_s` and `iat_mean_s` close behind. These constitute the robust temporal signatures: IAT variability is the only feature that is simultaneously the botnet detectors' primary evidence and a meaningful contributor to multi-vector attack detection (slow-rate DoS, scripted brute force), while directional byte asymmetry and packet-size uniformity remain informative in both domains. They are also behaviorally hard to evade an adversary must inject ir-

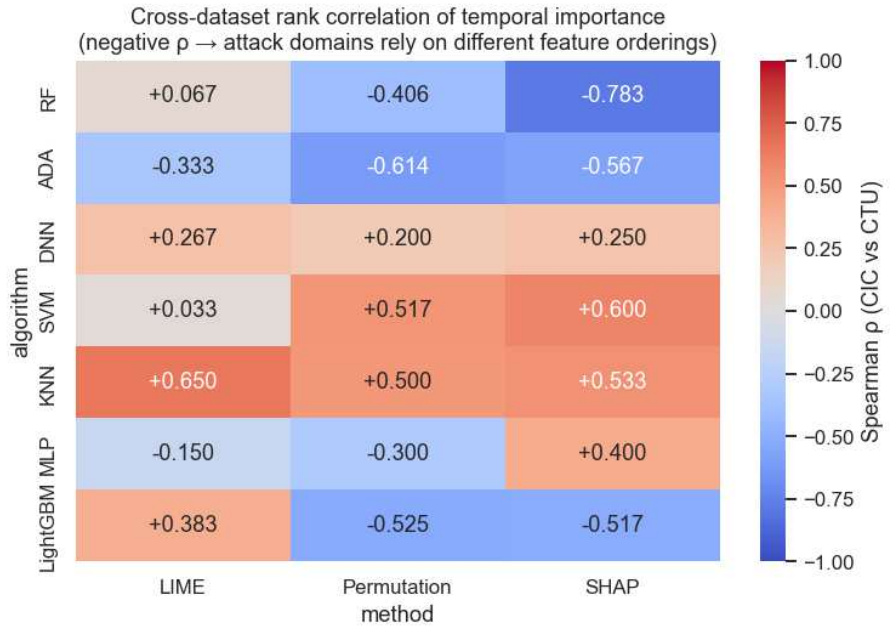


Fig. 14 Cross-dataset Spearman rank correlation of temporal importance (CIC vs. CTU) per algorithm and XAI method. Mean $\rho = +0.01$; 9/21 pairs negative: global rankings do not transfer between attack domains.

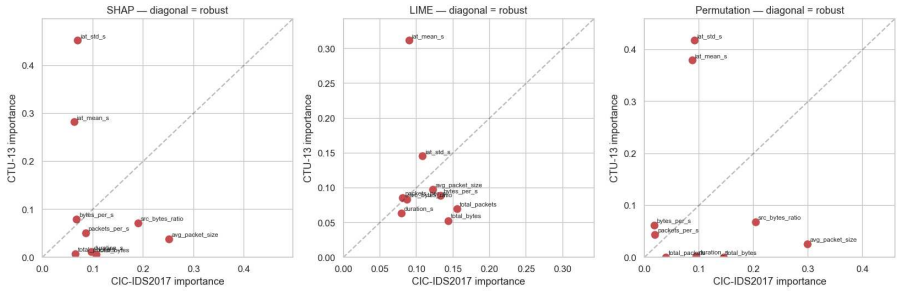


Fig. 15 Per-feature importance, CIC-IDS2017 vs. CTU-13, for each XAI method. The diagonal marks perfect robustness; IAT features specialize to botnets and packet size to multi-vector attacks, while `src_bytes_ratio` sits nearest the diagonal.

regular timing jitter (degrading C&C reliability), rebalance upload/download volume, and randomize packet sizes simultaneously to suppress all three.

4.5 Case Study: Local Interpretability

Global scores summarize; analysts triage individual alerts. Figure 17 shows the SHAP of the full-feature RF on CIC-IDS2017 test flows, which reveals

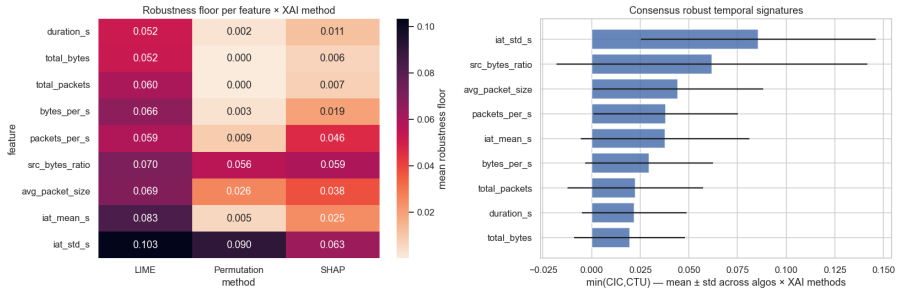


Fig. 16 Left: robustness floor min(CIC, CTU) per feature and XAI method. Right: consensus robust temporal signatures (mean ± std over 7 algorithms × 3 methods). IAT standard deviation, source-byte ratio, and average packet size are robust across domains, algorithms, and attribution methods.

the direction of evidence and its failure modes. High values (red) of Bwd Packet Length Std, Avg Bwd Segment Size, and Bwd Packet Length Mean push predictions firmly toward attack (SHAP up to +0.18), matching flood responses, while Destination Port shows a wide bidirectional spread: many low-value (blue) ports push toward attack, confirming that port evidence is ambiguous and context-dependent. For a borderline flow, an analyst can read exactly which measurements drove the score (e.g., a flow flagged at probability 0.999 by the temporal LightGBM was driven by elevated `iat_std_s` and `avg_packet_size` contributions)—turning an opaque alert into a verifiable claim about observable behaviour, in line with the analyst-facing benefits reported for LIME/SHAP explanations in operational studies [7,13]. LIME local surrogates on the same instances agreed on sign for the top features, adding a second, model-agnostic layer of diagnostic trust.

5 Discussion

The XAI consensus quantifies a long-standing intuition: botnets betray themselves in time, multi-vector attacks in volume and shape. Command-and-control channels must poll on a schedule, so their IAT distribution collapses to low-variance spikes (Fig. 4) that no single detector can ignore—shuffling IAT costs up to 0.54 AUC (Fig. 12). Floods and scans, by contrast, are free to randomize timing but cannot hide their purpose: they must move many bytes (DoS), touch many ports with tiny flows (scan), or repeat fixed-size requests (brute force), which is why packet-size and directionality features dominate CIC-IDS2017. The per-family heatmap (Fig. 13) shows the mapping is fine-grained and mechanistic, not statistical accident.

Three practical lessons follow. (i) A nine-feature temporal detector reaches $\geq$ 0.99 AUC, so privacy-preserving, payload-free, exporter-agnostic IDS is feasible—valuable where deep packet inspection is impossible (encrypted traffic) or disproportionate, and attractive for resource-constrained and federated edge deployments where compact feature sets reduce communication and

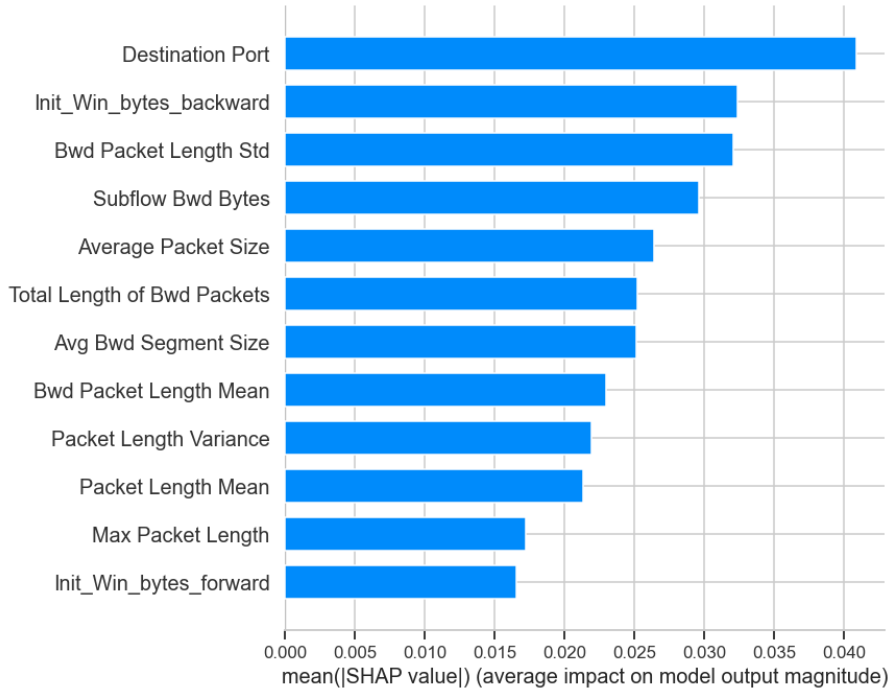


Fig. 17 SHAP (full-feature RF, CIC-IDS2017). Color encodes feature value; horizontal position its push toward attack (> 0) or benign (< 0). High backward packet-size dispersion is unambiguous attack evidence, whereas port evidence is bidirectional and thus unreliable.

computation costs [14,16]. (ii) XAI must be part of the MLOps loop: it exposed the Destination-Port shortcut instantly (Fig. 9) and would flag similar artefacts before deployment, complementing systematic XAI-evaluation frameworks such as E-XAI [6]. (iii) Because rankings do not transfer (mean $\rho = 0.01$), SOCs should not import feature-importance dashboards across environments; instead, the robustness-floor procedure gives a principled way to build multi-environment consensus signatures, and monitoring drift of the IAT/volumetric balance can even suggest which kind of attack pressure a network faces.

Explanations on CTU-13 are more concentrated (one or two features absorb up to 99% of importance) and more consistent across methods, reflecting the near-separable structure of beaconing traffic; explanations on CIC-IDS2017 are flatter and vary more across algorithms because fourteen heterogeneous families share one binary label. Concentrated explanations are easier to audit but also identify a single evasion target, which is why the multi-feature consensus signature matters.

The limitations: both corpora are testbed/laboratory captures; absolute metrics will degrade on modern encrypted, cloud-mediated traffic, and forward-temporal drift of the kind documented in recent generalization stud-

ies [10] remains to be measured for the signatures themselves. CTU-13 IAT statistics are derived from flow aggregates rather than raw packet timestamps. KernelSHAP and LIME operate on sampled subsets (1,500/60 instances), introducing estimation variance, and near-saturated AUC on CTU-13 compresses permutation scores. Finally, robustness is demonstrated across two domains; adding IoT botnet corpora [9] and newer enterprise captures would strengthen the consensus, and we did not yet test adaptive adversaries that actively perturb timing.

6 Conclusion and Future Work

This paper set out to unveil what black-box intrusion detectors actually learn, and whether that knowledge constitutes robust temporal signatures. Using seven classifiers, two heterogeneous benchmarks harmonized into one nine-feature temporal space, and three independent XAI layers, we showed that (i) temporal behavior alone suffices for near-perfect detection; (ii) explanations are domain-structured—IAT regularity exposes botnets, volumetric/size anomaly exposes multi-vector attacks—and importance rankings do not transfer naively (mean $\rho = +0.01$); and (iii) a cross-dataset robustness floor distils a consensus signature set—led by IAT standard deviation, source-byte ratio, and average packet size—that is stable across algorithms, attribution methods, and attack domains. XAI thus does more than decorate an IDS with plots: it audits shortcuts, grounds alerts in verifiable behavior, and provides a quantitative path to transferable detection knowledge, making ML-based defence more transparent and trustworthy.

Future work will (a) validate the consensus signatures on live and encrypted traffic (TLS); (b) integrate the local SHAP/LIME layer into SOAR playbooks so that automated responses carry machine-readable justifications, potentially paired with LLM-generated analyst narratives [13]; (c) stress-test the signatures against adaptive adversaries that inject timing jitter and padding, quantifying the cost of evasion; and (d) extend the harmonization to IoT botnet corpora [9] and concept-drift settings with online recalibration of the robustness floor, including federated variants for edge networks [14].

Statements and Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Gizachew Teshome Habtamu designed the study, implemented the pipeline and experiments, and drafted the manuscript. Shahzad Ahmed contributed to the security framework, analysis of results, and review of the manuscript. Both authors read and approved the final manuscript

References

1. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Security Privacy (ICISSP)*, Funchal, Portugal, 2018, pp. 108–116.
2. S. García, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Comput. Security*, vol. 45, pp. 100–123, Sep. 2014.
3. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.
4. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.
5. N. Moustafa, N. Koroniotis, M. Keshk, A. Y. Zomaya, and Z. Tari, "Explainable intrusion detection for cyber defences in the Internet of Things: Opportunities and solutions," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 3, pp. 1775–1807, 2023.
6. O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," *IEEE Access*, vol. 12, pp. 23954–23988, 2024.
7. D. Gaspar, P. Silva, and C. Silva, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024.
8. P. P. Kundu, T. Truong-Huu, L. Chen, L. Zhou, and S. G. Teo, "Detection and classification of botnet traffic using deep learning with model explanation," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 4, pp. 3067–3084, 2024.
9. T. Al-Shurbaji et al., "Deep learning-based intrusion detection system for detecting IoT botnet attacks: A review," *IEEE Access*, vol. 13, pp. 11792–11822, 2025.
10. 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.
11. V. Hnamte, H. Nhung-Nguyen, J. Hussain, and Y. Hwa-Kim, "A novel two-stage deep learning model for network intrusion detection: LSTM-AE," *IEEE Access*, vol. 11, pp. 37131–37148, 2023.
12. A. Oseni et al., "An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1000–1014, 2023.
13. A. Khediri, H. Slimi, A. Yahiaoui, M. Derdour, H. Bendjenna, and C. E. Ghenai, "Enhancing machine learning model interpretability in intrusion detection systems through SHAP explanations and LLM-generated descriptions," in *Proc. 6th Int. Conf. Pattern Anal. Intell. Syst. (PAIS)*, El Oued, Algeria, 2024, pp. 1–6.
14. M. Bhavsar, Y. Bekele, K. Roy, J. Kelly, and D. Limbrick, "FL-IDS: Federated learning-based intrusion detection system using edge devices for transportation IoT," *IEEE Access*, vol. 12, pp. 52215–52226, 2024.
15. L. Diana, P. Dini, and D. Paolini, "Overview on intrusion detection systems for computers networking security," *Computers*, vol. 14, no. 3, p. 87, 2025.
16. S. Gurappa, Y. Hariprasad, S. S. Iyengar, and N. K. Chaudhary, "Botnet detection on CTU-13 using lightweight machine learning models," *arXiv preprint* arXiv:2605.23004, 2026.