AN EMPIRICAL EVALUATION FRAMEWORK FOR AUDITING EXPLAINABLE AI IN NETWORK INTRUSION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning Network Intrusion Detection Systems (NIDS) suffer from a transparency gap that hinders operational trust. While Explainable AI (XAI) methods aim to bridge this, they often exhibit domain blindness, generating semantically invalid network states, such as fractional TCP flags that distort evaluation metrics. This paper proposes a modular evaluation framework centered on a novel Protocol-Aware XAI (PA-XAI) engine. By enforcing protocol invariants and hierarchical constraints, PA-XAI ensures that explanations are grounded in physically realizable telemetry. Our framework employs four auditing modules: functional faithfulness, adversarial robustness, algorithmic consensus, and dimensionality reduction utility. Empirical results across four benchmark datasets reveal that standard XAI metrics often mask significant instability; for example, LIME's apparent stability is frequently an artifact of invalid neighborhood construction. Conversely, our PA-variants, specifically PA-DeepLIFT and PA-IG, demonstrate superior robustness against adversarial evasion. Furthermore, PA-XAI guides dimensionality reduction to achieve up to 90.5% feature compression with less than a 1% drop in detection fidelity. These findings demonstrate that domain-aware auditing is essential for transitioning XAI from a theoretical enhancement to a reliable component of operational network defense.

1 INTRODUCTION

The proliferation of internet-connected devices has reshaped the threat landscape facing modern network infrastructure. Every connected endpoint generates network telemetry governed by precisely specified protocols, which adversaries exploit to conduct sophisticated intrusions. Network Intrusion Detection Systems (NIDS) have progressively moved beyond signature-based heuristics toward deep learning anomaly detection models capable of learning high-dimensional packet-level data. Yet, this performance gain introduces a fundamental liability: the distributed, non-linear representations that enable accurate detection render these models inherently opaque. A NIDS that flags a flow as malicious but cannot explain why provides operators with a conclusion, not a basis for decision-making.

Explainable Artificial Intelligence (XAI) has emerged as the principal research framework for addressing this transparency deficit Barnard et al. (2022). Post-hoc techniques like SHAP Lundberg & Lee (2017), LIME Garreau & Luxburg (2020), Integrated Gradients (IG) Sundararajan et al. (2017), and DeepLIFT Shrikumar et al. (2017) allow practitioners to interrogate models without modifying their architectures Gaspar et al. (2024). While widely adopted in other security and non-security domains Ma et al. (2024); Manthena et al. (2025); Hernandes et al. (2021); Greco et al. (2023), their application to network intrusion detection remains understudied. Critically, most prior work treats NIDS XAI as a static visualization layer rather than an operational system component requiring rigorous auditing before its outputs can inform operational decisions.

A central obstacle to reliable XAI in NIDS is *domain blindness*: the failure of standard perturbation methods to respect semantic and structural constraints. Methods like LIME construct surrogate neighborhoods by sampling features with unconstrained Gaussian noise. While reasonable for independent tabular data, network traffic is governed by strict hierarchical bounds and discrete flag

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

co-occurrence rules. A sample where TCP SYN and RST flags are simultaneously set, or where maximum packet length is smaller than the mean, represents an impossible network state. Consequently, surrogates trained on invalid neighborhoods approximate classifier behavior on states the model was never trained to distinguish. As we demonstrate empirically, this distortion reverses faithfulness metrics, inflates stability scores, and misdirects analyst attention toward artifacts of generated noise rather than the classifier's true decision boundary.

This paper presents a modular, end-to-end empirical framework for auditing XAI in NIDS. Its primary architectural novelty is a *Protocol-Aware XAI (PA-XAI) engine* that enforces strict network invariants during generation, ensuring every perturbed sample represents a physically realizable state. Built around PA-XAI, the framework features four auditing modules: (1) Quantitative Benchmarking of faithfulness, complexity, and stability; (2) Adversarial Robustness and Scaffolding Evaluation under gradient-based evasion attacks; (3) Consensus and Stability Analysis across competing XAI methods; and (4) XAI-Driven Dimensionality Reduction. Together, these modules operationalize a central thesis: XAI outputs in NIDS cannot be treated as ground truth and require continuous, domain-aware auditing.

Our work delivers the following contributions:

1. The first end-to-end modular auditing framework for NIDS ML models.
2. A novel Protocol-Aware XAI (PA-XAI) engine.
3. Analysis of how model choice impacts explainability and evaluation metrics.
4. A rigorous empirical characterization of domain blindness.
5. An adversarial robustness taxonomy and scaffolding evaluation.

Our empirical evaluation across four benchmark datasets (CIC-IDS-2017, CSE-CIC-IDS2018, NSL-KDD, UNSW-NB15) reveals that domain blindness produces misleading metrics across all dimensions. Standard LIME reports artificially perfect stability (Jaccard of 1.0) that collapses to 0.39–0.86 under PA-generation, and its faithfulness frequently reverses sign on DNN classifiers. Under adversarial perturbation, PA-DeepLIFT and PA-IG maintain cosine similarity > 0.80, while standard LIME and DeepLIFT collapse to near-zero. Inter-explainer consensus does not exceed a Spearman correlation of 0.31 under standard generation, yet the PA-DeepLIFT/PA-IG pairing achieves correlations > 0.91. Finally, PA-XAI guides dimensionality reduction to achieve up to 90.5% feature compression with less than a 1% drop in detection fidelity, demonstrating that protocol-aware explanations translate directly to operational utility.

2 Background & Related Works

2.1 ML-NIDS Architecture and The Black-Box Predicament

To handle complex network telemetry, modern NIDS have increasingly adopted deep learning architectures Nikitenko & Bashkov (2024); Manocchio et al. (2024); Devendiran & Turukmane (2024); Wang et al. (2024); Hore et al. (2024); Ahmad et al. (2021); Orojo (2023). However, the distributed nature of neural representations renders these models inherently opaque. This transparency gap severely undermines operational trust in mission-critical environments Hasan et al. (2024). While Explainable AI (XAI) frameworks attempt to address this Islam et al. (2024), a growing body of research demonstrates that XAI attributions are highly sensitive to adversarial perturbations Galli et al. (2021); Ivanovs et al. (2021); Burger et al. (2023). Gradient and perturbation-based methods exhibit substantial variability under minor input manipulations Slack et al. (2020). This allows adversaries to systematically alter explanations while preserving classification outcomes, turning the explanation layer into an attack surface. Consequently, XAI outputs cannot be treated as ground truth without dedicated robustness auditing.

2.2 Evaluating XAI

The operational validity of XAI is quantified through four established metrics Arreche et al. (2024): *Faithfulness* (the Pearson correlation between feature attribution scores and the predictive probability drop upon feature ablation), *Complexity* (the Shannon entropy of the attribution vector, capturing

relevance spread), *Sparsity* (the fraction of scores below a relevance threshold to isolate minimal logic subsets), and *Stability* (the Jaccard Similarity of top-k features across independent algorithmic runs).

2.3 The Semantic Gap: Domain-Blindness in XAI

Perturbation-based XAI approaches (e.g., LIME, feature ablation) typically assume features can be modified independently and arbitrarily. In NIDS, this assumption is fundamentally flawed: network traffic is governed by strict protocol specifications, temporal dependencies, and causal relationships Orojo et al. (2025). Applying unconstrained perturbations to network data creates *domain blindness*, where algorithms operate on syntactic representations and ignore domain-specific bounds. The resulting semantic gap manifests as a disconnect between mechanistic XAI manipulations and the meaningful, physically realizable network behaviors analysts must actually interpret.

2.4 Assessment Frameworks for Explainable NIDS

Systematic evaluation of XAI within NIDS remains fragmented. Prior work highlights SHAP's inconsistencies Tritscher et al. (2023), DeepLIFT's operational superiority for LSTM models Kalakoti et al. (2025), and TRUSTEE's adversarial resilience compared to SHAP on state-of-the-art DL-NIDS Kumar & Thing (2024). However, fundamental gaps persist: prior efforts treat XAI as a static enhancement rather than a system-critical component, and none address domain blindness, where standard algorithms fabricate mathematically impossible network states (e.g., fractional TCP flags). Our framework directly addresses these gaps via a Protocol-Aware XAI (PA-XAI) engine, operationalizing rigorous, domain-constrained explanation auditing under active adversarial manipulation.

3 Architecture of the Evaluation Framework

We propose a modular, end-to-end evaluation framework (Figure 1) engineered to systematically measure XAI efficacy within NIDS. Operating as an independent layer wrapped around the core predictive engine, the framework processes normalized network telemetry, generates feature attributions, and routes them to an *Auditing Pipeline*. Rather than treating XAI outputs as ground truths, this pipeline stress-tests explanations across four independent modules to ensure they meet the reliability, security, and utility requirements necessary for operational deployment. The complete implementation is publicly available.[1]

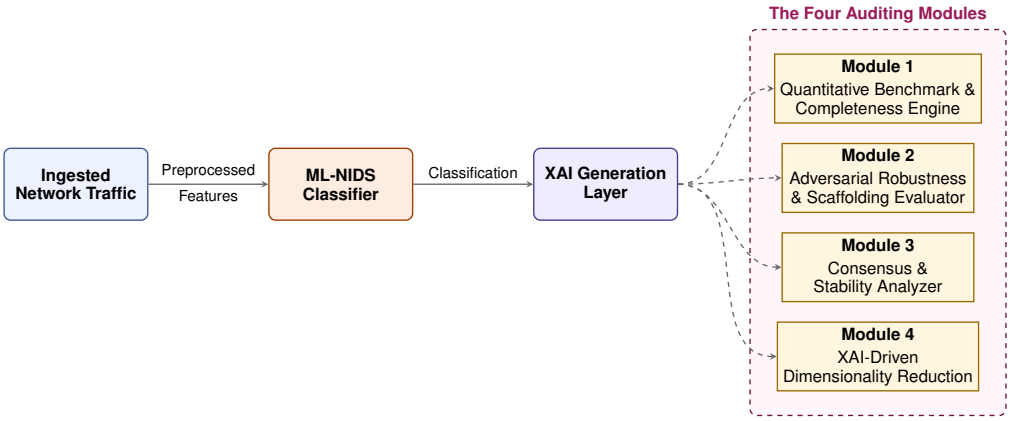


Figure 1: System architecture of the evaluation framework.

[1] https://github.com/BLIND_USERNAME/BLIND_REPO

3.1 Component Specifications & Baselines

The framework evaluates across four benchmark datasets (CIC-IDS-2017, CSE-CIC-IDS2018, NSL-KDD, UNSW-NB15) using three baseline classifiers (DNN, Random Forest, XGBoost). The architecture decouples the predictive engine from evaluation across three phases:

1. Classification Boundary: Preprocessed, normalized feature spaces are evaluated by the baseline NIDS.

2. XAI Generation Layer (PA-XAI Engine): Post-hoc methods (SHAP, LIME, IG, DeepLIFT) generate attributions via our Protocol-Aware engine (Section 3.2), which enforces strict protocol invariants to correct domain-blindness.

3. Auditing Modules: Four quantitative modules measure distinct XAI dimensions: (i) Faithfulness and complexity via iterative feature occlusion; (ii) Adversarial fragility and defensive scaffolding utility under constrained perturbations; (iii) Inter-method consensus via rank correlations; and (iv) Feature compression capability via XAI-guided dimensionality reduction.

3.2 The Protocol-Aware XAI (PA-XAI) Engine

To address domain-blindness, the PA-XAI engine algorithmically enforces physical network states, hierarchical boundaries, and discrete logic across explanations:

1. PA-LIME (Conditioned Marginal Sampling): Replaces LIME's unconstrained Gaussian neighborhoods Ribeiro et al. (2016) with Conditioned Marginal Sampling. Discrete features are drawn from protocol-filtered training subsets. A vectorized enforcer subsequently repairs cross-feature relationships, clamps standard deviations, and enforces hierarchical bounds (e.g., Packet Length Max $\geq$ Mean $\geq$ Min), ensuring every surrogate sample represents a realizable network flow.

2. PA-IG and PA-DeepLIFT (Decision Boundary Anchoring): Replaces the standard zero-vector baseline Sundararajan et al. (2017); Shrikumar et al. (2017) with *Decision Boundary Anchoring*, selecting the valid per-class sample nearest the decision threshold to highlight features actively crossing the boundary. To avoid interpolating through invalid discrete states, PA-DeepLIFT uses the Rescale Rule's multipliers to bypass intermediate states, while PA-IG uses a *Sequential Path* to step discrete features via exact finite-difference computations.

3. PA-SHAP (Constraint Game Theory): Replaces generic masking Lundberg & Lee (2017) with valid benign background samples across all protocols. For permutation-based variants (KernelSHAP), a Post-Hoc Coalition Wrapper enforces hierarchical ordering and protocol-gated feature zeroing prior to model evaluation.

4 Quantitative Benchmarking of Explanation Quality

The first phase of the experimental framework creates a quantitative evaluation process. This experiment deploys a standardized suite of metrics; See Sec 2; Faithfulness (Descriptive Accuracy), Sparsity, Complexity, Efficiency, Stability, and Completeness.

4.1 Results and Operational Implications

Table 1 reports the evaluation metrics averaged across all four benchmark datasets for both standard and protocol-aware XAI generation. The comparison is not framed as a contest between two competing approaches; rather, PA-XAI represents the corrected measurement in which the explanation algorithm only evaluates physically realizable network states. Where the standard and PA metrics diverge, the delta quantifies how much the unconstrained evaluation was distorted by the inclusion of impossible packet configurations in the explanation generation process. All model–explainer pairings achieve a completeness rate of 1.0 under both configurations; this metric is omitted.

4.1.1 Standard LIME Produces Misleading Stability Assessments.

The most operationally consequential finding concerns LIME's stability metric. Under standard generation, LIME reports perfect deterministic stability ($J = 1.0$) on both UNSW-NB15 and CSE-

CIC-IDS2018 across all three classifiers (Table 2). A practitioner relying on this metric would conclude that LIME provides consistent, repeatable explanations on these datasets, a key requirement for operational trust. Protocol-aware evaluation reveals this assessment to be an artifact of invalid neighborhood construction. When PA-LIME restricts the surrogate neighborhood to physically realizable packets, stability drops to 0.43–0.69 on UNSW-NB15 and 0.39–0.86 on CSE-CIC-IDS2018. The apparent stability under standard generation arises because the unconstrained Gaussian neighborhoods are dominated by impossible packet states; the NIDS model assigns these synthetic inputs to a narrow region of its output space (since they bear no resemblance to any learned traffic pattern), producing artificially consistent surrogate fits. When the neighborhood is corrected to contain only valid network flows, the true variability of the local decision boundary is exposed.

4.1.2 Standard Faithfulness Metrics Mask Explainer Failure on DNNs.

Standard LIME reports moderate faithfulness for the DNN on CIC-IDS-2017 ($\rho = 0.50$) and CSE-CIC-IDS2018 ($\rho = 0.49$). Under protocol-aware generation, these values reverse sign to -0.05 and -0.30 respectively. A negative faithfulness correlation indicates that the features the explainer identifies as most important are *less* influential on the classifier's output than those it ranks as unimportant. The explanation is inversely aligned with the model's logic. The standard evaluation obscured this failure because the surrogate was fitting a neighborhood of invalid packets, producing a correlation that reflected the structure of the noise rather than the model's decision boundary. The corrected PA measurement reveals that LIME on the DNN is misleading on these datasets. This pattern does not hold uniformly: on NSL-KDD, PA-LIME corrects a negative standard faithfulness (-0.05) to a positive value (0.20), and on tree-based models the standard and PA faithfulness values are broadly consistent (RF/LIME diverges by ≤ 0.10 on three of four datasets). The DNN's distributed representations are more sensitive to neighborhood composition than the axis-aligned splits of RF and XGB, making the DNN–LIME pairing particularly vulnerable to domain-blind evaluation.

4.1.3 Deterministic Methods Are Less Affected by Domain Blindness.

SHAP, IG, and DeepLIFT maintain $J = 1.0$ stability under both standard and PA generation, with correspondingly small faithfulness deltas (SHAP: DNN -0.01, RF $+0.00$, XGB $+0.13$; DeepLIFT: $+0.03$), confirming that deterministic computation paths are not susceptible to the neighborhood-dependent artifacts affecting LIME. The one notable shift is DNN/IG, where faithfulness decreases from 0.57 to 0.43 under PA, reflecting the replacement of the zero-vector baseline with a decision-boundary prototype that captures the features pushing a packet across the decision boundary rather than from an arbitrary reference. This shift alters the DNN explainer ranking from SHAP (0.63) $>$ IG (0.57) $>$ DeepLIFT (0.51) $>$ LIME (0.28) to SHAP (0.62) $>$ DeepLIFT (0.54) $>$ IG (0.43) $>$ LIME (-0.07). The inversion of DeepLIFT and IG under corrected evaluation carries practical implications: standard benchmarks would lead a practitioner to select IG over DeepLIFT, while the domain-corrected measurement favors DeepLIFT.

PA-XAI introduces higher latency for perturbation-based methods: PA-LIME averages 1,690–1,738 ms per sample, and PA-SHAP on tree models ranges from 248.5 ms (RF) to 1,990.7 ms (XGB).

Table 1: Evaluation metrics averaged across four benchmark datasets under Standard (Std) and Protocol-Aware (PA) generation.

		Faithfulness (ρ)			**Stability (Jaccard)**			**Eff. (ms)**	
Model	**Method**	Std	PA	Δ	Std	PA	Δ	Std	PA
DNN	SHAP	0.63	0.62	−0.01	1.00	1.00	0.00	65.1	828.8
DNN	LIME	0.28	−0.07	−0.36	0.63	0.46	−0.17	92.2	1690.6
DNN	IG	0.57	0.43	−0.14	1.00	1.00	0.00	0.7	54.6
DNN	DeepLIFT	0.51	0.54	+0.03	1.00	1.00	0.00	0.01	6.7
RF	SHAP	0.79	0.80	+0.00	1.00	1.00	0.00	13.9	248.5
RF	LIME	0.38	0.49	+0.12	0.69	0.74	+0.05	127.3	1738.0
XGB	SHAP	0.58	0.71	+0.13	1.00	1.00	0.00	0.1	1990.7
XGB	LIME	0.23	0.18	−0.05	0.65	0.54	−0.11	123.4	1737.4

Table 2: Per-dataset LIME (ρ) and (Jaccard) under standard and PA generation.

	NSL-KDD		CIC-IDS-2017		UNSW-NB15		CSE-CIC-2018	
	Std	PA	Std	PA	Std	PA	Std	PA
Faithfulness (ρ)								
DNN/LIME	−0.05	0.20	0.50	−0.05	0.19	−0.14	0.49	−0.30
RF/LIME	0.18	0.66	0.31	0.39	0.66	0.56	0.36	0.36
XGB/LIME	0.14	0.38	0.21	0.29	0.25	0.10	0.31	−0.07
Stability (Jaccard)								
DNN/LIME	0.17	0.52	0.37	0.45	1.00	0.48	1.00	0.39
RF/LIME	0.23	0.74	0.52	0.67	1.00	0.69	1.00	0.86
XGB/LIME	0.15	0.62	0.44	0.50	1.00	0.43	1.00	0.60

These figures reflect a research prototype without vectorized batch processing, caching, or compiled-kernel optimizations; the constraint enforcement loop currently operates per-coalition in Python, whereas production deployment would amortize this cost across batches. Gradient-based methods already operate at practical latencies (PA-IG: 54.6 ms; PA-DeepLIFT: 6.7 ms), as they require minimal modification to their standard computation paths. The reported latencies should therefore be interpreted as upper bounds rather than intrinsic limitations of the PA-XAI approach; reducing perturbation-based overhead through implementation-level optimization is a target for future work.

5 Adversarial Robustness and Threat Modeling

Module 2 subjects the NIDS classifier and its XAI outputs to gradient-based evasion attacks to evaluate explanation fragility and defensive scaffolding utility.

5.1 Threat Model and Attack Formulation

We adopt a white-box threat model (MITRE ATLAS **AML.T0040**) deploying Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015) and Projected Gradient Descent (PGD) Madry et al. (2018) attacks across escalating budgets (ε). Attacks are evaluated under two regimes: *unconstrained* (features perturbed arbitrarily) and *constrained* (PA-XAI enforcer repairs post-perturbation protocol violations, modeling adversaries who must bypass boundary validation).

Table 3 shows that constrained perturbations successfully reduce classifier evasion rates, particularly at moderate budgets ($\varepsilon = 0.1$), by restricting the adversary's attack surface to physically realizable network flows.

5.2 Explanation Fragility Under Adversarial Perturbation

Robustness is quantified via Cosine Similarity (CosSim), Lipschitz ratio (Lip_{mean}), and Class-Equivalent Jaccard (ClsEq_Jacc) to isolate explanation instability from legitimate decision boundary crossings. Figure 2 plots CosSim under constrained FGSM.

Standard methods degrade rapidly: Standard LIME is fundamentally fragile, exhibiting CosSim below 0.19 and immense Lip_{mean} shifts at minimal budgets ($\varepsilon = 0.01$), driven by unconstrained Gaussian neighborhoods amplifying adversarial noise. Standard DeepLIFT collapses to negative CosSim by $\varepsilon = 0.05$ due to its unstable zero-vector baseline.

Table 3: Evasion success rates (%) averaged across datasets

	FGSM (ε)					PGD (ε)		
Regime	0.01	0.05	0.1	0.2	0.3	0.01	0.05	0.1
Unconstrained	20.1	46.9	63.7	74.9	76.1	21.1	54.9	74.6
Constrained	18.3	45.9	51.5	67.2	75.3	18.5	49.1	58.8

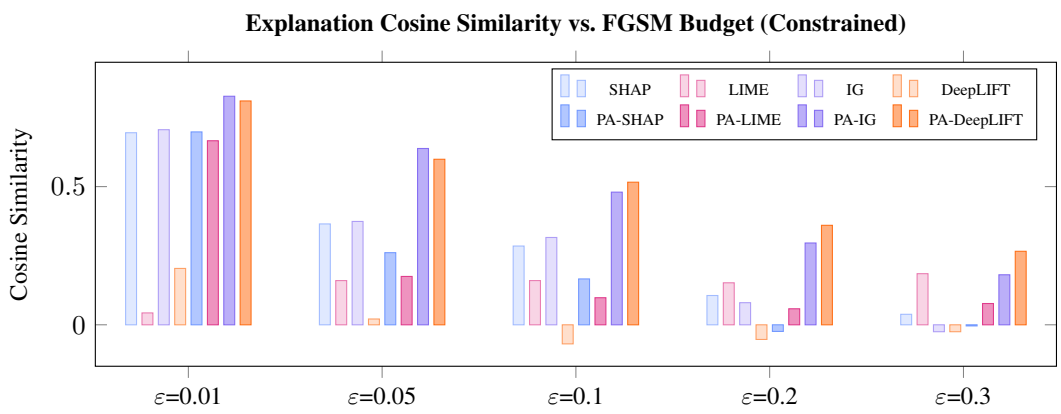


Figure 2: Cosine similarity between original and adversarial attribution vectors under constrained FGSM at escalating ε.

PA methods exhibit superior robustness: PA-IG and PA-DeepLIFT achieve the highest resilience (CosSim > 0.81 at $\varepsilon = 0.01$ and > 0.48 at $\varepsilon = 0.1$). Decision boundary anchoring provides a geometrically stable reference that induces proportionally small attribution shifts. PA-LIME recovers significantly from baseline fragility (CosSim 0.666 at $\varepsilon = 0.01$) by conditioning the surrogate on valid local behavior.

5.3 Explanation-Aware Defensive Scaffolding

Module 2 also evaluates if attributions identify truly perturbed features to aid signature generation (Table 4). Standard SHAP fails entirely, and standard LIME ranks perturbed features far outside practical attention windows (mean rank 73.0). PA-LIME achieves the strongest performance, doubling top-3 precision to 0.384 and halving mean rank to 41.2. However, even the best top-1 precision (0.152) confirms XAI scaffolding should serve as an investigative lead rather than a fully automated defense mechanism.

Table 4: Scaffolding precision averaged across datasets. Top-1/Top-3: overlap between most-attributed and truly perturbed features. Mean Rank: position of perturbed feature in attribution ordering (lower is better).

Method	Top-1	Top-3	Mean Rank
LIME	0.064	0.200	73.0
SHAP	0.000	0.000	123.0
PA-LIME	0.152	0.384	41.2
PA-SHAP	0.038	0.122	87.5

6 Algorithmic Consensus and Feature Interactions

Security teams often consult multiple explainers to corroborate alerts. Module 3 evaluates whether different XAI methods agree on NIDS classifications and align with underlying non-linear feature interactions.

6.1 Pairwise Consensus Analysis

Consensus was measured across 66 explainer pairs per dataset using rank correlations and top-k overlaps (Table 5). Standard inter-explainer agreement is remarkably weak: mean Spearman correlation (ρ_s) does not exceed 0.31, and Wilcoxon tests reject equivalent attributions in $> 84\%$ of pairings.

Table 5: Pairwise consensus across 66 explainer pairs per dataset.

Dataset	ρ_s (mean $\pm$ std)	τ_b (mean $\pm$ std)	Top-5	Top-10	Wilcoxon (%)
NSL-KDD	0.262 ± 0.259	0.220 ± 0.229	0.226	0.300	84.8
CIC-IDS-2017	0.256 ± 0.207	0.196 ± 0.169	0.168	0.240	89.4
UNSW-NB15	0.309 ± 0.132	0.249 ± 0.110	0.322	0.328	84.8
CSE-CIC-IDS2018	0.256 ± 0.122	0.206 ± 0.098	0.213	0.262	84.8

However, stratifying by generation mode reveals a powerful regularization effect. Within-PA pairs consistently achieve higher consensus than standard pairs (e.g., UNSW-NB15: $\rho_s = 0.414$ vs. 0.231). By restricting explainers to protocol-valid perturbation spaces, PA-XAI eliminates inter-method variance caused by handling invalid network states. The PA-DeepLIFT $\leftrightarrow$ PA-IG pair achieves exceptional alignment ($\rho_s > 0.91$ on three datasets) due to shared boundary anchoring.

6.2 Feature Interaction Analysis

The consensus deficit partly stems from LIME assuming feature independence, whereas NIDS classifiers exploit complex interactions. SHAP interaction values (Table 6) confirm DNNs exhibit 1.5–$3.3\times$ stronger interactions than Random Forests.

Table 6: Feature interaction strength: DNN-to-RF ratio and LIME's correlation with interaction importance. Positive r indicates LIME *under*-weights high-interaction features.

	Interaction Ratio		**LIME–Interaction r (DNN)**	
Dataset	DNN	RF	r	p
NSL-KDD	2.10	1.31	-0.535	2.1×10^{-10}
CIC-IDS-2017	3.06	1.95	0.415	1.4×10^{-4}
UNSW-NB15	1.50	1.47	0.483	8.8×10^{-6}
CSE-CIC-IDS2018	3.30	2.24	0.280	1.3×10^{-2}

On high-interaction datasets (DNN ratios ≥ 3.0), LIME's rankings diverge severely from interaction-derived ground truth. For instance, critical protocol-semantic features like *Protocol* or *Bwd Pkt Len Std* rank in the top 10 for interaction strength but fall below rank 70 under LIME. LIME systematically suppresses features critical to the model's actual reasoning, directing analysts toward individually predictive but interaction-poor artifacts.

7 XAI-Driven Dimensionality Reduction

Module 4 evaluates feature compression by comparing XAI attributions against statistical baselines (Chi-Squared, Info Gain, PCA, Spearman). Features are ranked via mean absolute attribution over 5,000 samples and iteratively pruned using a Random Forest proxy. Pruning halts when the F1 score degrades by $\Delta\text{F1} > 0.01$. The trade-off between compression and detection fidelity is evaluated using the median matched-classifier ΔF1_w across 12 evaluation pairs (Figure 3).

PA-IG and PA-SHAP define the Pareto frontier. PA-IG-DNN maximizes compression (90.5% reduction, -0.89 pp median F1 drop), successfully stripping complex datasets down to 6–8 features. PA-SHAP-DNN optimally balances compression and fidelity (85.3% reduction, -0.68 pp drop). These gradient-based PA methods excel by directly encoding the classifier's learned boundaries, isolating the exact feature core required for discriminative logic. Conversely, conventional methods are competitive but hit a hard plateau; PCA achieves a -0.69 pp drop but cannot push past an 80.6% reduction limit without the model-aware ranking provided by XAI.

LIME partially defeats the reduction objective. Standard LIME reduces dimensionality by only 54–59% (retaining $\sim$45 features). Because its locally linear surrogates distribute importance diffusely, the pruning threshold shifts before meaningful reduction occurs. PA-LIME-DNN performs even worse (-4.27 pp drop), confirming that LIME's instability under protocol-aware generation severely undermines feature selection.

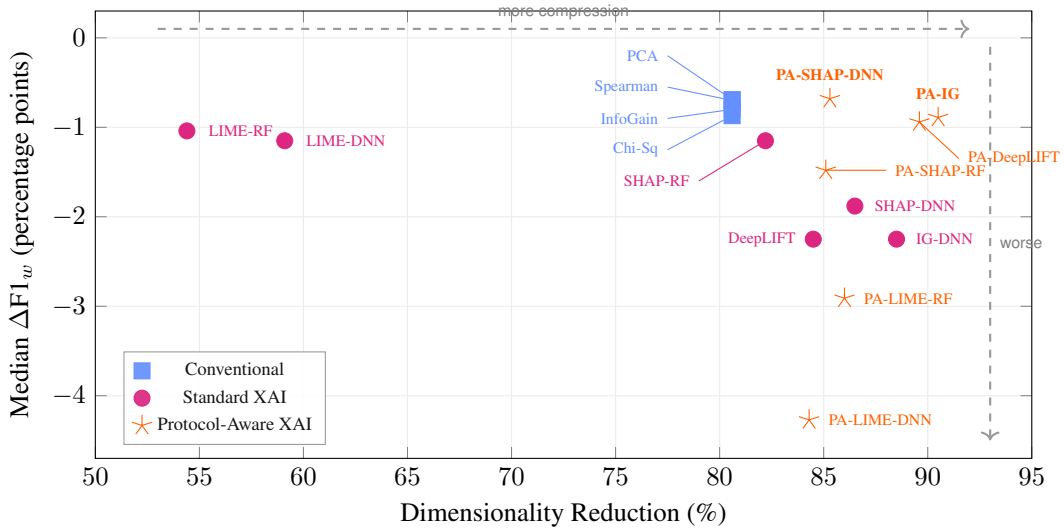


Figure 3: Trade-off between dimensionality reduction and detection fidelity. Each point reports the *median* matched-classifier $\Delta F1_w$ across 12 evaluation pairs (4 classifiers $\times$ 4 datasets), where each method is compared against its full-feature baseline.

Operational Implications: These results support a tiered deployment strategy: PCA for rapid prototyping (80% reduction in seconds), PA-SHAP-DNN for robust periodic retraining (85% reduction), and PA-IG-DNN or PA-DeepLIFT-DNN for resource-constrained edge deployments requiring extreme compression (90%) with minimal detection loss.

8 Conclusion

This paper introduces a domain-aware auditing framework for evaluating Explainable AI (XAI) in Network Intrusion Detection Systems (NIDS). Our findings demonstrate that standard XAI evaluation is severely distorted by "domain blindness," yielding misleading metrics and artificially inflated stability. By contrast, our Protocol-Aware (PA) framework enforces physically realizable network states. We found that PA-DeepLIFT and PA-IG remain robust under adversarial attacks, whereas standard methods degrade rapidly. Furthermore, while inter-explainer consensus is generally weak, PA methods significantly improve alignment and enable 85–90% dimensionality reduction with negligible performance loss (¡1%). Ultimately, XAI outputs cannot be treated as ground truth without continuous, domain-aware auditing. Future work will focus on optimizing computational overhead, enabling real-time evaluation, and expanding semantic constraints to broader protocol families.

References

Zeeshan Ahmad, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, and Farhan Ahmad. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1):e4150, 2021.

Osvaldo Arreche, Tanish R. Guntur, Jack W. Roberts, and Mustafa Abdallah. E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection. *IEEE Access*, 12:23988–24003, 2024. doi: 10.1109/ACCESS.2024.3365140.

Pieter Barnard, Nicola Marchetti, and Luiz A DaSilva. Robust network intrusion detection through explainable artificial intelligence (xai). *IEEE Networking Letters*, 4(3):167–171, 2022.

Christopher Burger, Lingwei Chen, and Thai Le. "are your explanations reliable?" investigating the stability of lime in explaining text classifiers by marrying xai and adversarial attack. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12831–12844, 2023.

Ramkumar Devendiran and Anil V Turukmane. Dugat-lstm: Deep learning based network intrusion detection system using chaotic optimization strategy. *Expert Systems with Applications*, 245: 123027, 2024.

Antonio Galli, Stefano Marrone, Vincenzo Moscato, and Carlo Sansone. Reliability of explainable artificial intelligence in adversarial perturbation scenarios. In *International Conference on Pattern Recognition*, pp. 243–256. Springer, 2021.

Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International conference on artificial intelligence and statistics*, pp. 1287–1296. PMLR, 2020.

Diogo Gaspar, Paulo Silva, and Catarina Silva. Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access*, 12:30164–30175, 2024.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Francesco Greco, Giuseppe Desolda, Andrea Esposito, et al. Explaining phishing attacks: An xai approach to enhance user awareness and trust. In *ITASEC*, 2023.

M.K. Hasan, R.A. Abdulkadir, S. Islam, T.R. Gadekallu, and N. Safie. A review on machine learning techniques for secured cyber-physical systems in smart grid networks. *Energy Reports*, 11:1268–1290, 2024. doi: 10.1016/j.egyr.2023.12.040.

Paulo R Galego Hernandes, Camila P Floret, Katia F Cardozo De Almeida, Vinícius Camargo Da Silva, João Paulo Papa, and Kelton A Pontara Da Costa. Phishing detection using url-based xai techniques. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 01–06. IEEE, 2021.

Soumyadeep Hore, Jalal Ghadermazi, Ankit Shah, and Nathaniel D Bastian. A sequential deep learning framework for a robust and resilient network intrusion detection system. *Computers & Security*, 144:103928, 2024.

Md Tohidul Islam, Md Khalid Syfullah, Md Golam Rashed, and Dipankar Das. Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable ai. *International Journal of Machine Learning and Cybernetics*, 15(11):5337–5360, 2024.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

Rajesh Kalakoti, Risto Vaarandi, Hayretdin Bahsi, and Sven Nõmm. Evaluating explainable ai for deep learning-based network intrusion detection system alert classification. *arXiv preprint arXiv:2506.07882*, 2025.

Ayush Kumar and Vrizlynn LL Thing. Evaluating the explainability of state-of-the-art deep learning-based network intrusion detection systems. *arXiv preprint arXiv:2408.14040*, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ki-Pyoung Ma, Dong-Ju Ryu, and Sang-Joon Lee. Reverse analysis method and process for improving malware detection based on xai model. *Computers, Materials & Continua*, 81(3), 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Liam Daly Manocchio, Siamak Layeghy, Wai Weng Lo, Gayan K Kulatilleke, Mohanad Sarhan, and Marius Portmann. Flowtransformer: A transformer framework for flow-based network intrusion detection systems. *Expert Systems with Applications*, 241:122564, 2024.

Harikha Manthena, Shaghayegh Shajarian, Jeffrey Kimmell, Mahmoud Abdelsalam, Sajad Khorsandroo, and Maanak Gupta. Explainable artificial intelligence (xai) for malware analysis: A survey of techniques, applications, and open challenges. *IEEE Access*, 2025.

Andrii Nikitenko and Yevhen Bashkov. Construction of a network intrusion detection system based on a convolutional neural network and a bidirectional gated recurrent unit with attention mechanism. *Eastern-European Journal of Enterprise Technologies*, 129(9), 2024.

Abanisenioluwa Orojo, Ehis El-Mahmoud, Eric Leal, and Pablo Rivas. Byteflow: A byte-level llm for deep packet inspection and network intelligence. In *2025 Annual Computer Security Applications Conference Workshops (ACSAC)*, 2025.

Abanisenioluwa K. Orojo. Advancing network intrusion detection with a sliding window-based cnn and cfs-ba feature selection, May 2023. Preprint, ResearchGate.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Julian Tritscher, Maximilian Wolf, Andreas Hotho, and Daniel Schlör. Evaluating feature relevance xai in network intrusion detection. In *World Conference on Explainable Artificial Intelligence*, pp. 483–497. Springer, 2023.

Xiao Wang, Lie Dai, and Guang Yang. A network intrusion detection system based on deep learning in the iot: X. wang et al. *The Journal of Supercomputing*, 80(16):24520–24558, 2024.