Beyond Attack Success Rate: Representing Consistency in Evading the NIDS Feature-Space

Allan da S. Espindola[1,2], **Altair O. Santin**[1], **António Casimiro**[2]
Pedro M. Ferreira[2], **Eduardo K. Viegas**[1]

[1] Pontifícia Universidade Católica do Paraná (PUCPR)

{allan.espindola, santin, eduardo.viegas}@ppgia.pucpr.br

[2]LASIGE, Faculdade de Ciências, Universidade de Lisboa – Lisboa, Portugal

{adespindola, casim, pmf}@ciencias.ulisboa.pt

Abstract. *Feature-space attacks directly modify NIDS inputs. Their attack success rate (ASR) measures classifier evasion under vector access but does not by itself establish an end-to-end NIDS vulnerability. We evaluate FGSM and PGD attacks against an MLP-based NIDS using ASR, Pearson correlation changes across records, and flow-level checks of features within each flow. ASR and correlation change followed distinct trajectories. At the largest perturbation limit, almost all successful volume-only and mixed evasions introduced a tested violation. In the packet-derived reference, no strong clean pair reversed sign, and no re-extracted record violated a tested relationship. Together, these findings show that correlation change and flow-level checks provide complementary evidence about representation consistency beyond classifier evasion.*

1. Introduction

A flow-based network intrusion detection system (NIDS) classifies a feature vector generated through packet-to-flow extraction [Filho et al. 2025]. A flow exporter aggregates packets and derives temporal and volumetric features, which are subsequently transformed during preprocessing into the input representation used by the classifier [Simioni et al. 2025]. Feature-space attacks directly manipulate this input representation to induce a benign classification [Espindola et al. 2026b]. Their attack success rate (ASR) quantifies the ability of an attack to evade the classifier under direct access to its input representation, but it does not, by itself, establish an end-to-end vulnerability of the NIDS [Apruzzese et al. 2022, Espindola et al. 2026c].

Operational exploitability also depends on the transformations available to the adversary. Under network-only access, the adversary must preserve the malicious functionality of the attack while inducing the target representation through packet manipulation, flow extraction, and preprocessing [Pierazzi et al. 2020, Apruzzese et al. 2022]. Timing-based traffic reshaping exemplifies this attack path [Sharon et al. 2022]. In contrast, directly modifying a post-extraction feature vector assumes that the adversary can influence the resulting representation or its processing pipeline, thereby exposing a distinct attack surface. The flow representation provides a basis for examining this distinction. Totals, means, extrema, variances, and directional values derived from the same packets are related through structural dependencies in the evaluated extraction pipeline [Vormayr et al. 2020, Alhussien et al. 2024]. Changes in Pearson correlation

characterize structural deviations between aligned clean and attacked records, whereas flow-level checks assess relationships among features within individual flows. A failed check therefore provides record-level evidence that the attacked representation is incompatible with the evaluated flow representation.

Recent NIDS studies have shown that feature-space misclassification can coexist with traffic-class distortion [Catillo et al. 2025] or contradictions among traffic features [Catillo et al. 2026]. However, these studies do not jointly examine ASR, changes in correlations among flow features, and violations of feature relationships within individual flows as perturbation limits increase. It therefore remains unclear whether ASR and representation changes evolve together and whether similar evasion outcomes preserve the same relationships tested within the flow representation.

Contribution. We address this gap using two complementary first-order white-box attacks against an MLP trained on UNSW-NB15 [Moustafa and Slay 2015]. Both attacks operate under the same normalized L_∞ feature-space access model. FGSM applies gradient-sign updates [Goodfellow et al. 2015], whereas PGD performs iterative projected updates [Madry et al. 2018]. Across timing-only, volume-only, and mixed perturbations, we examine the relationship between ASR, changes in feature correlations, and newly introduced consistency violations as the perturbation limit increases. We further apply the structural diagnostics to records re-extracted after packet-level manipulations, with the aim of answering two questions:

RQ1. *How do the attack success rate and the correlations between flow features change as the perturbation limit increases?*

RQ2. *How often does direct feature perturbation introduce a new flow-level consistency violation under at least one tested relationship between features of the same flow?*

2. Experimental Setup and Evaluation Methodology

We evaluate direct attacks against an MLP-based NIDS trained on UNSW-NB15 [Moustafa and Slay 2015]. We use the MLP for which complete and aligned results are available across attack methods, feature scopes, and perturbation limits. This controlled case study focuses on the structural interpretation of the attacks rather than on architectural comparisons. The analysis considers $4,937$ DoS records from day 17 of the test set and 49 classifier inputs, comprising 24 timing and 25 volume features. The classifier inputs were extracted using Go-Flows with a feature specification based on the IPFIX information model [Vormayr et al. 2020, Claise and Trammell 2013]. The recorded configuration consists of three hidden layers, each with 100 ReLU units, and a single output logit. Training was performed using Adam, binary cross-entropy loss, batch size 32, and early stopping within 100 epochs. The evaluated direct-attack artifacts and packet-derived datasets were generated using the D-MOOD experimental pipeline [Espindola et al. 2025].

Direct attacks modify the classifier input after feature extraction. The packet-derived reference modifies packet data before re-extraction. ASR, correlation change, and consistency checks respectively measure classifier evasion, structural changes across records, and contradictions among features within an individual flow.

2.1. Evaluation of Direct Feature-Space Attacks

We evaluate three perturbation scopes: timing features only, volume features only, and both feature groups (mixed). We consider the complete 14-value grid used in the attack experiments, with normalized L_∞ perturbation bounds ranging from $\epsilon = 10^{-5}$ to 10^{-1}. Each bound constrains the maximum change applied to any selected feature, although the attack may apply smaller perturbations. A min–max transformation fitted on the training set maps the features to the normalized ART input space. We inverse-transform the perturbed values so that correlation and consistency analyses are performed in the original features.

The attacks use their recorded ART 1.17.0 parameters [Nicolae et al. 2018] without retuning for this analysis. FGSM uses ART's minimal-perturbation mode with $\texttt{eps\_step} = \epsilon/30$, searching in those increments up to ϵ. PGD uses one 30-step run with step size $\epsilon/30$ and $\texttt{num\_random\_init} = 0$, without random starts or restarts. Its results therefore characterize this fixed search configuration.

Attack success rate (ASR). At the model's recorded operating threshold of 0.30, the classifier detects $4,788$ of the $4,937$ DoS records. These records form the initial set of detected records. For attack configuration a, the ASR is defined as the fraction of these records whose aligned perturbed counterparts cross the threshold:

$$\mathrm{ASR}(a) = \frac{|\{i : c_i > \tau \ \wedge \ c_{i,a} \leq \tau\}|}{|\{i : c_i > \tau\}|}, \qquad \tau = 0.30. \tag{1}$$

here $c_i = \sigma(z_i)$ and $c_{i,a} = \sigma(z_{i,a})$ are scores before and after attack. The numerator forms the successful-evasion subset. ASR thus measures classifier evasion when there is direct access to the input representation.

Pearson correlation change. We use Pearson's r to measure pairwise linear associations. For an aligned set S, let $R_0^{(S)}$ and $R_a^{(S)}$ denote the clean and perturbed Pearson correlation matrices over all classifier features. The set of feature pairs $\mathcal{P}$ in the strict upper triangle for which both correlations are finite, the mean absolute correlation change is:

$$D(a, S) = |\mathcal{P}|^{-1} \sum_{(j,k)\in\mathcal{P}} |R_{a,jk}^{(S)} - R_{0,jk}^{(S)}|. \tag{2}$$

a clean pair is strong when $|R_{0,jk}^{(S)}| \geq 0.50$ and reverses sign when $R_{0,jk}^{(S)} R_{a,jk}^{(S)} < 0$. The 0.50 cutoff affects only this descriptive sign-reversal count, while Equation 2 uses every finite pair. Correlation change describes relationships across records, not whether one record satisfies the flow representation.

Flow-level consistency violations. The audit defines the eight relationships summarized in Appendix A. Six checks apply to every record. The final applicable check is selected between two backward-direction definitions using the aligned clean record, with $a_i = \mathbf{1}[\mathrm{packetTotalCountBwd}(\mathbf{x}_i^{\mathrm{clean}}) > 0]$. If $a_i = 1$, the backward minimum, mean, and maximum must satisfy the expected ordering. Otherwise, the 13 backward-dependent timing and volume features must retain the encoded absence value of zero. This branch is determined by the clean value and remains fixed for the perturbed record. The clean backward-packet count is used only as auxiliary metadata for selecting this branch.

For equality, ordering, and absence checks, the corresponding residuals are $|l_i - r_i|$, $\max(m_i - \mu_i, \mu_i - M_i, 0)$, and $\max_{j \in \mathcal{B}} |x_{i,j}|$, respectively. Both evaluation routes use the same scale-aware tolerance. A check fails when $\text{residual} > 10^{-3} + 10^{-5}\text{scale}$, where the scale is the largest absolute operand or 1. The absolute and relative terms accommodate numerical differences near zero and across different operand magnitudes. A new violation occurs when the perturbed record fails a check that its clean counterpart passes. Thus, a failure provides evidence of a tested contradiction, whereas passing indicates that no such contradiction was detected.

2.2. Packet-Derived Structural Reference

We analyze packet-derived datasets generated using native nominal factors of 10, 50, and 100 for timing, packet length, and combined changes, respectively. We report these factors separately from the direct-attack ϵ values. For each packet-derived dataset, we compare Pearson correlations among its flow features with those of the clean reference and apply the same strong-pair threshold, consistency checks, and tolerance. The re-extracted backward-packet count determines the applicable conditional check.

Because timestamp modifications can alter flow grouping [Vormayr et al. 2020], and stable cross-dataset identifiers are unavailable, the clean and re-extracted sets cannot be aligned on a record-by-record basis. We therefore compare correlations across the two sets and apply the consistency checks independently to individual records. These data provide a structural reference for an attack path that includes re-extraction rather than an equal-strength comparison with the direct attacks. Equation 1 therefore applies only to the aligned direct attacks.

3. Results

All $4,937$ clean records passed every applicable flow-level consistency check using the residual threshold defined in Section 2.1.

3.1. RQ1: Attack Success Rate and Correlation Change

Figure 1 summarizes the direct feature-space results: panels (a) and (b) answer RQ1, and panel (c) is discussed under RQ2. Across configurations, the mean $|\Delta r|$ generally increased with ϵ, whereas ASR did not consistently increase as ϵ grew. Mixed FGSM illustrates this distinction clearly: beyond $\epsilon = 5 \times 10^{-2}$, the representation deviated further from its clean correlation structure even as fewer records were successfully evaded. A larger perturbation budget can increase structural deviation without improving attack effectiveness.

The comparison at the largest evaluated perturbation limit illustrates why ASR alone cannot rank representation effects. Volume and mixed PGD achieved nearly identical ASR values of 81.35% and 80.85%, respectively, yet mixed PGD produced more than four times the correlation change, with $|\Delta r|$ values of 0.2668 and 0.0628, respectively. FGSM exhibited the same ordering despite the absence of similarly close ASR values: volume achieved a higher ASR, whereas mixed FGSM again produced more than four times the correlation change. Thus, the similar-ASR comparison applies specifically to PGD, while for neither attack method did ASR rank structural departure. Sign reversals occurred only when timing features were included, indicating that the perturbation scope influenced the nature of the structural change, rather than merely its magnitude. RQ1 therefore shows that ASR and changes in linear correlations did not evolve across the evaluated configurations.

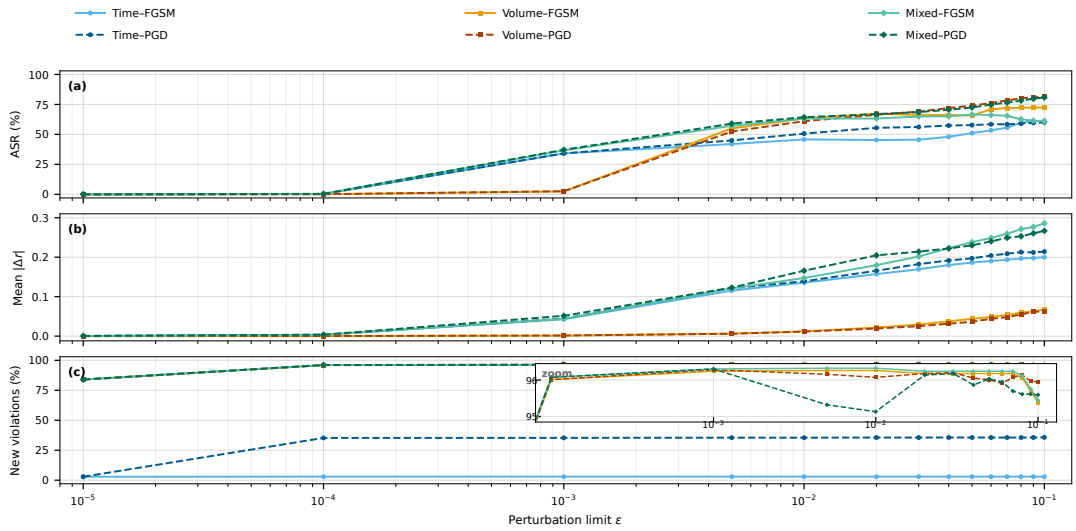

Figure 1. ASR, Pearson correlation change, and flow-level consistency violations.

3.2. RQ2: Flow-Level Consistency Violations

Panel (c) reports the percentage of initially detected records with at least one new violation. Table 1 additionally reports the corresponding rates among successful evasions. The plotted violation rates did not track ASR, highlighting the overlapping volume and mixed trajectories. At $\epsilon = 10^{-1}$, timing-only and mixed FGSM achieved similar ASR values of 61.45% and 60.78%, respectively, but new violations affected only 2.94% and 95.45% of the initially detected records, respectively. Thus, similar classifier outcomes can be accompanied by substantially different effects on the tested relationships within a flow. At the same perturbation limit, all four volume and mixed configurations produced new-violation rates of at least 95.38% among initially detected records and 98.52% among successful evasions. Almost every successful evasion under volume-only and mixed perturbations contradicted at least one tested relationship in the evaluated flow representation.

Table 1 summarizes the three outcomes at $\epsilon = 10^{-1}$. The classifier-evasion columns report the number of successful evasions and the corresponding ASR over the initially detected set. The correlation columns report the structural changes observed for the same aligned set. The final two columns report the number of records with at least one new violation among the initially detected records and among the successfully evaded records, respectively. Timing-only violation rates differed substantially between the two methods: 2.94% for FGSM and 35.74% for PGD among initially detected records. All such violations resulted from assigning nonzero values to timing features whose backward direction was encoded as absent in the corresponding clean record.

Packet-derived structural reference. At native factor 100, the mean $|\Delta r|$ across the re-extracted sets was 0.01597 for timing changes, 0.01650 for volume changes, and 0.03438 for combined changes, with combined changes producing the largest structural departure. Across all evaluated factors, none of the 219 strong clean pairs reversed sign, and every re-extracted record passed the shared consistency checks. Modifications applied before re-extraction altered the correlation structure without producing any tested contradictions. The result contrasts with the frequent violations observed in direct volume-only and mixed perturbations at the largest direct perturbation limit.

Table 1. Feature-space attack results at $\epsilon = 10^{-1}$.

		Classifier evasion		Correlation change		New flow-level violations	
Feature scope	Method	Successes /4,788	ASR (%)	mean $\lvert\Delta r\rvert$	Sign reversals /219	Initially detected	Successful evasions
Time	FGSM	2,942/4,788	61.45	0.1998	66/219	141/4,788	58/2,942
Time	PGD	2,872/4,788	59.98	0.2142	60/219	1,711/4,788	1,618/2,872
Volume	FGSM	3,467/4,788	72.41	0.0677	0/219	4,567/4,788	3,422/3,467
Volume	PGD	3,895/4,788	81.35	0.0628	0/219	4,594/4,788	3,894/3,895
Mixed	FGSM	2,910/4,788	60.78	0.2859	65/219	4,570/4,788	2,867/2,910
Mixed	PGD	3,871/4,788	80.85	0.2668	63/219	4,577/4,788	3,865/3,871

4. Discussion

The direct results distinguish classifier sensitivity from representation compatibility. Similar ASR values were accompanied by substantially different changes in feature correlations, showing that ASR alone does not indicate the degree of structural departure. The high prevalence of violations under volume-only and mixed perturbations further shows that the classifier can accept feature vectors that contradict relationships satisfied by their clean counterparts. For this model instance, the observed representation effects varied across attack methods and perturbation scopes.

Pearson's r characterizes linear structure across records, whereas the consistency checks identify declared contradictions among features within individual flows. For network-facing exploitability, an attacker must induce a compatible representation through packet transformations while preserving the malicious functionality [Pierazzi et al. 2020, Apruzzese et al. 2022]. The packet-derived reference preserves re-extraction in this attack path and characterizes the resulting structure after pre-extraction modifications, rather than providing a direct ranking of attack strength against the direct ϵ budget.

This case study considers a single MLP instance, UNSW-NB15 DoS records, two first-order attacks, Pearson's linear perspective, and a finite catalog of consistency checks. Passing every check establishes only that no contradiction covered by the catalog was detected. Future work should evaluate additional datasets, traffic categories, training seeds, attack methods, and NIDS architectures, including multi-view ensembles [Espindola et al. 2026a]. It should also quantify uncertainty and nonlinear dependence, evaluate multi-start PGD, and derive consistency checks from flow-exporter specifications.

5. Conclusion

RQ1 showed that ASR and changes in Pearson correlation did not evolve together, indicating that ASR alone does not characterize the departure from the clean representation. RQ2 showed that nearly every successful volume-only and mixed evasion at the largest perturbation limit introduced a tested contradiction, whereas the timing-only results differed substantially between FGSM and PGD. These findings establish classifier sensitivity, linear structure across records, and feature compatibility within individual flows as distinct evaluation dimensions. An end-to-end vulnerability claim should therefore combine evidence of classifier evasion with evidence that modified traffic can produce a compatible representation while preserving its malicious functionality. Broader evaluations across models, datasets, attack methods, and automatically derived consistency checks are needed to determine the extent to which this separation holds.

Acknowledgment

This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT) through the PhD Scholarship 2023.00446.BD, the LASIGE Research Unit (ref. UID/00408/2025; DOI 10.54499/UID/00408/2025 - LASIGE), and by the Brazilian National Council for Scientific and Technological Development (CNPq), under grants 302937/2023-4, 407879/2023-4, 442262/2024-8, 406603/2025-1 and 307706/2025-7.

A. Flow-Level Consistency Checks

Table 2. Flow-level checks, feature relationships, and semantic basis.

Check	Features and relationship	Semantic basis
Total-byte agreement	$L = O$	Both fields aggregate the same IP-byte quantity over the flow.
Forward-byte agreement	$L_F = O_F$	The same byte total is represented independently for forward packets.
Backward-byte agreement	$L_B = O_B$	The same byte total is represented independently for backward packets.
Dispersion agreement	$V = S^2$	Variance and standard deviation encode the same packet-length dispersion.
Overall length range	$m \leq \mu \leq M$	An arithmetic mean lies between the observed minimum and maximum.
Forward length range	$m_F \leq \mu_F \leq M_F$	The same ordering must hold among forward packet lengths.
Backward length range	$m_B \leq \mu_B \leq M_B, a_i = 1$	The same ordering applies when the flow contains backward packets.
Encoded backward absence	$x_j = 0 \; \forall j \in \mathcal{B}, a_i = 0$	When a flow has no backward packets, the evaluated extraction and preprocessing pipeline maps the backward-dependent statistics to zero.

L, O, μ, m, and M denote `ipTotalLength`, `octetTotalCount`, `ipTotalLengthMean`, `minimumIpTotalLength`, and `maximumIpTotalLength`. Subscripts F and B add the `Fwd` and `Bwd` suffixes. V = `ipTotalLengthVar` and S = `ipTotalLengthStdev`. Here $a_i = \mathbf{1}[\texttt{packetTotalCountBwd} > 0]$ and $\mathcal{B}$ contains:

`bwdBytesAvg`	`interPacketTimeSecondsSumBwd`
`bwdBytesPerMicroseconds`	`ipTotalLengthBwd`
`bwdJitterMilliseconds`	`ipTotalLengthMeanBwd`
`interPacketTimeSecondsMaxBwd`	`ipTotalLengthStdevBwd`
`interPacketTimeSecondsMeanBwd`	`maximumIpTotalLengthBwd`
`interPacketTimeSecondsStdevBwd`	`minimumIpTotalLengthBwd`
	`octetTotalCountBwd`

The first six checks always apply, while backward activity selects exactly one of the final two. All operands belong to the classifier inputs. `packetTotalCountBwd` is auxiliary metadata used only to select the branch. Every check uses the common tolerance defined in Section 2.1.

References

Alhussien, N., Aleroud, A., Melhem, A., and Khamaiseh, S. Y. (2024). Constraining adversarial attacks on network intrusion detection systems: Transferability and defense analysis. *IEEE Transactions on Network and Service Management*, 21(3):2751–2772.

Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., and Colajanni, M. (2022). Modeling realistic adversarial attacks against network intrusion detection systems. *Digital Threats: Research and Practice*, 3(3):1–19.

Catillo, M., Pecchia, A., Repola, A., and Villano, U. (2025). A critique on the (mis)use of feature-space attacks for adversarial machine learning in NIDS. In *2025 20th European Dependable Computing Conference (EDCC)*, pages 110–115. IEEE.

Catillo, M., Pecchia, A., and Villano, U. (2026). Similarity is not enough: Issues with adversarial perturbations of traffic features against intrusion detection systems. In *Proceedings of the 12th International Conference on Information Systems Security and Privacy (ICISSP)*, volume 1, pages 325–332. SciTePress.

Claise, B. and Trammell, B. (2013). Information model for IP flow information export (IPFIX). RFC 7012, Internet Engineering Task Force.

Espindola, A. d. S., Casimiro, A., Santin, A. O., Ferreira, P. M., and Viegas, E. K. (2026a). Enhancing intrusion detection generalization via diversity-driven multi-view ensemble learning in industrial systems. *Future Generation Computer Systems*, 182:108458.

Espindola, A. d. S., Santin, A. O., Casimiro, A., Ferreira, P. M., and Viegas, E. K. (2026b). Understanding the adversary: A survey of adversarial machine learning in network intrusion detection. *Computer Science Review*, 62:100995.

Espindola, A. d. S., Santin, A. O., Viegas, E. K., Ferreira, P. M., and Casimiro, A. (2026c). Diversity as a security primitive for ML-based network intrusion detection. In *2026 IEEE 12th International Conference on Network Softwarization (NetSoft)*, pages 409–414. IEEE.

Espindola, A. d. S., Viegas, E. K., Casimiro, A., Santin, A. O., and Ferreira, P. M. (2025). D-MOOD: Diversity-based multi-objective optimization defense. GitHub software repository, `https://github.com/espindolaallan/d-mood`. Revision 483af25.

Filho, A. G., Viegas, E. K., Santin, A. O., and Geremias, J. (2025). A dynamic network intrusion detection model for infrastructure as code deployed environments. *Journal of Network and Systems Management*, 33(4).

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR 2015), Conference Track Proceedings*, San Diego, CA, USA.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations (ICLR 2018), Conference Track Proceedings*, Vancouver, BC, Canada.

Moustafa, N. and Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*, pages 1–6. IEEE.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I. M., and Edwards, B. (2018). Adversarial robustness toolbox v1.0.0. *CoRR*, abs/1807.01069.

Pierazzi, F., Pendlebury, F., Cortellazzi, J., and Cavallaro, L. (2020). Intriguing properties of adversarial ML attacks in the problem space. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1332–1349. IEEE.

Sharon, Y., Berend, D., Liu, Y., Shabtai, A., and Elovici, Y. (2022). TANTRA: Timing-based adversarial network traffic reshaping attack. *IEEE Transactions on Information Forensics and Security*, 17:3225–3237.

Simioni, J., Viegas, E. K., Santin, A., and Horchulhack, P. (2025). An early exit deep neural network for fast inference intrusion detection. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC '25, page 730–737. ACM.

Vormayr, G., Fabini, J., and Zseby, T. (2020). Why are my flows different? a tutorial on flow exporters. *IEEE Communications Surveys & Tutorials*, 22(3):2064–2103.