



Network Flow Integration to Improve Generalization of Machine Learning-Based IDS

Kelson Carvalho Santos^{1,2}, Rodrigo Sanches Miani¹

¹Federal University of Uberlândia (UFU) – Uberlândia, MG – Brazil

²Federal Institute of Piauí (IFPI) – Teresina, PI – Brazil

{kelson,miani}@ufu.br, kelson@ifpi.edu.br

Abstract. *The growth of cyber threats and the evolution of attacks highlight the need for more robust machine learning-based IDS capable of operating in heterogeneous, dynamic network environments. Although some approaches presented in the literature improve the performance of these systems, their generalization capability remains limited in multi-domain data scenarios. This paper investigates supervised domain adaptation through controlled integration of labeled target-domain flows to diversify training and reduce generalization performance degradation in IDS. Results show improvements in specific scenarios, depending on the integrated flow type, with trade-offs among attack detection, class balance, and false alarms.*

1. Introduction

Intrusion Detection Systems (IDS) monitor, analyze, and generate alerts for suspicious or malicious behaviors in network traffic [Putra and Huang 2019]. In machine learning-based approaches, models build profiles of legitimate network behavior to identify deviations in the traffic [Mahfouz et al. 2020]. However, the effectiveness of these systems depends on their generalization capability, that is, their ability to maintain consistent performance in environments different from those used during training [Sudjana et al. 2024].

Despite recent advances, the adoption of machine learning-based IDS in heterogeneous environments still faces limitations in adapting across different operational contexts. Studies by [Kenyon et al. 2020] and [Molina-Coronado et al. 2020] highlight heterogeneity across datasets, both in statistical characteristics and attack types, making it difficult to apply intrusion detection models in scenarios involving multiple data domains. In [Zoppi et al. 2023], this weakness is identified as an obstacle to the operational deployment of such IDS in real and dynamic environments. Furthermore, evaluations conducted only in single-domain scenarios do not adequately represent heterogeneous environments. Results presented by [Layeghy and Portmann 2022], [Domingues et al. 2022], [Verkerken et al. 2021], and [D’hooge et al. 2020] show consistent performance degradation when intrusion detection models are evaluated on datasets different from those used during training. These findings highlight the need for approaches capable of handling the statistical and behavioral variability present across multiple data domains.

The adoption of machine learning-based IDS in real-world environments depends on the availability of large volumes of labeled network traffic containing both legitimate and anomalous behaviors. However, obtaining such data is challenging due to the difficulties of collecting, labeling, and sharing malicious traffic in operational environments.

As an alternative, public datasets have been used to train these models, enabling the development and experimental evaluation of different intrusion detection approaches. Nevertheless, such datasets may not fully reflect the operational conditions of real networks, which differ in infrastructure, devices, applications, and communication patterns, and which also undergo continuous changes in traffic behavior. Consequently, models trained exclusively on public data from a single domain may experience reduced generalization capability when deployed in environments different from those used during training.

In this context, this paper investigates a controlled supervised domain adaptation method that integrates labeled network flows from distinct traffic domains to increase training diversity and reduce performance degradation during the generalization of machine learning-based IDS. The proposed approach seeks to promote experimental reproducibility by using reusable public datasets across different scenarios. To achieve this, the UNSW-NB15 [Moustafa and Slay 2015], CIC-IDS2017, and CIC-IDS2018 [Sharafaldin et al. 2018] datasets were used. These datasets underwent a feature standardization process by extracting network flows from PCAP files using the NFStream tool [Aouini and Pekar 2022], which transforms raw packets into bidirectional flows enriched with statistical features. In addition to preprocessing and manually labeling the network flows, the supervised XGBoost algorithm was employed to build the models.

The main contributions of this work are threefold: (i) analyzing how the controlled integration of labeled target-domain flows affects the performance of machine learning-based IDS during generalization across distinct data domains; (ii) systematically evaluating supervised integration strategies involving benign, malicious, and mixed flows; and (iii) providing standardized versions of public datasets and experimental artifacts to support reproducible research on the generalization of machine learning-based IDS.

2. Related Work

The generalization of machine learning-based IDS has received attention over the last five years because of the difficulties models face in multi-domain data scenarios. Studies by [D’hooge et al. 2020], [Verkerken et al. 2021], and [Layeghy and Portmann 2022] showed that although supervised and unsupervised models perform well in single-domain scenarios, performance degrades when evaluated across different data domains. These results highlight limitations in the generalization capability of intrusion detection models when dealing with attacks and patterns not observed during training.

Several studies have sought to mitigate this problem by improving data representation. In [Marvi et al. 2021], feature selection was used to eliminate redundant features, thereby improving model performance. However, the evaluation was limited to a single dataset. In [Rocha et al. 2023], t-distributed Stochastic Neighbor Embedding (t-SNE) was used to analyze attack clusters and evaluate models’ ability to recognize unknown variants, but without exploring multi-domain data. In turn, [Sudyana et al. 2024] proposed an approach based on the attack lifecycle combined with deep learning models, achieving improvements in generalization, though difficulties in identifying benign traffic and increased false alarms remained.

More recently, [Santos and Miani 2025] investigated dimension reduction and feature selection techniques as intervention strategies to improve data quality at the feature level and, consequently, enhance the generalization of intrusion detection models in

multi-domain data scenarios. The results showed that although these approaches improve feature-level data representation, they do not independently address the generalization problem across different domains, indicating that this limitation is strongly related to the diversity of network flows encountered during training.

Given these limitations, this paper proposes a network flow integration method as an instance-level intervention strategy to increase the diversity of patterns used during training and, consequently, improve the performance and generalization capability of machine learning-based IDS. Table 1 presents a comparative summary of the analyzed studies. Unlike related work, our proposal combines interventions at both the feature and instance levels, details the process of extracting and structuring network flows, and evaluates intrusion detection models under different multi-domain data scenarios.

Table 1. Summary of related works

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
[D’hooge et al. 2020]	7 Tree-Based 2 neighborhood-based 2 vector-based 1 regression-based	no	no	CIC-IDS2017 CIC-IDS2018	CICFlowMeter	no	yes	yes
[Verkerken et al. 2021]	Autoencoder oSVM iForest	Dimension Reduction	no	CIC-IDS2017 CIC-IDS2018	no	yes	yes	yes
[Marvi et al. 2021]	LGBM	Feature Selection	no	CIC-DDoS-2019	no	yes	yes	no
[Layeghy and Portmann 2022]	Feed Forward LSTM Random Forest Extra-Tree iForest oSVM SGD-oSVM	no	no	UNSW-NB15 CIC-IDS2018 ToN-IoT Bot-IoT	nProbe	no	yes	yes
[Rocha et al. 2023]	Decision Tree Random Forest SVM Logistic Regression Extra-Tree	Analysis of Attack Groups	no	CIC-IDS2017	no	no	yes	no
[Sudyana et al. 2024]	CNN XGBoost	Dimension Reduction	no	CREMEv2 CIC-IDS2017 CIC-IDS2018 CIC-DDoS-2019 CCU Mirai HTTP	nProbe	no	yes	yes
[Santos and Miani 2025]	XGBoost iForest	Dimension Reduction Feature Selection	no	UNSW-NB15 CIC-IDS2017 CIC-IDS2018	NFStream	yes	yes	yes
Our proposed approach	XGBoost	Dimension Reduction Feature Selection Network Flow Integration	yes	UNSW-NB15 CIC-IDS2017 CIC-IDS2018	NFStream	yes	yes	yes

[1] Reference; [2] Algorithms; [3] Implementation of Intervention Strategies in the Datasets; [4] Comparison of the Intervention Strategies; [5] Dataset; [6] Tool for Network Flow Extraction; [7] Description of the Features Extracted and Integrated into the Datasets; [8] Performance Evaluation in a Single Data Domain; [9] Performance Evaluation Across Multiple Data Domains.

3. Dataset Intervention Method

The workflow for building machine learning-based intrusion detection models involves a sequence of fundamental steps: dataset selection; data analysis and preprocessing to improve data quality; algorithm selection for model building; and performance evaluation. The proposed approach extends this workflow by incorporating supervised domain adaptation strategies applied to the datasets and validation scenarios composed of distinct data domains. Figure 1 illustrates this new workflow.

The intervention strategies, as represented in Step 3, are applied to the datasets at the instance level via controlled integration of network flows. These strategies can be used independently or combined with feature-level interventions [Santos and Miani 2025].

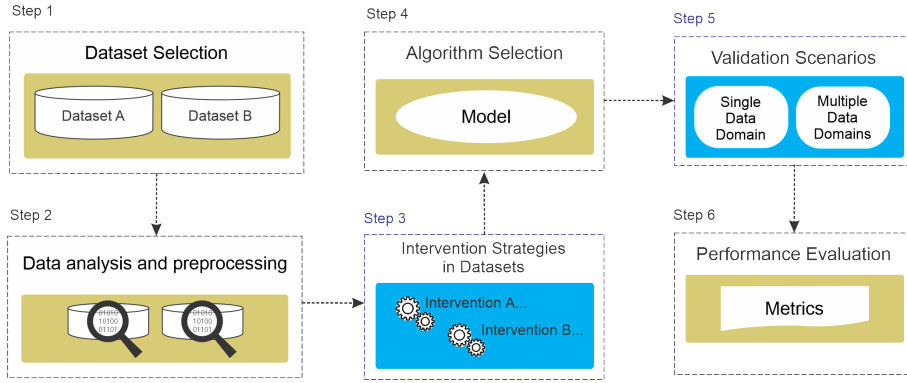


Figure 1. Workflow of the proposed approach

The objective is to increase the diversity of patterns observed during training and to analyze how this diversity influences the models' ability to maintain performance across distinct domains. The validation scenarios, represented in Step 5, consider both single-domain environments, in which the model is trained and tested on the same dataset, and multi-domain scenarios, in which training and testing are performed on different datasets.

3.1. Problem Formalization

Consider a network traffic dataset as a domain $D = (X, Y, P)$, where $X \subseteq \mathbb{R}^d$ represents the feature space of network flows, Y represents the set of classes, and $P(X, Y)$ denotes the joint data distribution. In this study, UNSW-NB15, CIC-IDS2017, and CIC-IDS2018 are treated as distinct domains because they were generated in different periods, network infrastructures, traffic configurations, application environments, and attack scenarios. Although NFStream provides a common feature space, it does not eliminate differences in the distributions and behavioral patterns represented by the network flows. Given a source domain D_A and a target domain D_B , the objective is to evaluate the ability of a classifier $f_\theta : X \rightarrow Y$, trained on D_A , to maintain predictive performance when evaluated on D_B .

In this work, generalization refers to the evaluation objective of maintaining predictive performance across distinct network traffic domains. Three learning settings can be distinguished according to the availability of target-domain data: (i) cross-domain generalization, in which the model is trained exclusively on the source domain and evaluated on an unseen target domain; (ii) unsupervised domain adaptation, in which unlabeled target-domain data are available during training; and (iii) supervised domain adaptation, in which labeled target-domain data are incorporated into the training process. The baseline adopted in this study corresponds to the first setting. Thus, for an evaluation metric M , its cross-domain performance is expressed as $G(f_\theta, D_A, D_B) = M(f_\theta, D_B^{test})$, where f_θ is trained exclusively on D_A^{train} and evaluated on D_B^{test} . Generalization degradation is estimated as $\Delta G = M(f_\theta, D_A^{test}) - M(f_\theta, D_B^{test})$, with higher values indicating greater performance loss when the model is evaluated on a distinct domain.

In contrast to the baseline, the proposed interventions correspond mainly to supervised domain adaptation because controlled partitions of labeled benign or malicious flows from the target domain are incorporated into the source-domain training set. The interventions are modeled as operators I_k^ρ , where k represents the integration strategy and $\rho \in \{20\%, 40\%, 60\%, 80\%\}$ represents the proportion of target-domain flows used during

training. Each operator produces a modified training set, $D_{A,k}^{train,\rho} = I_k^\rho(D_A^{train}, D_B)$. The incorporated flows are removed from the corresponding target-domain test set, satisfying $D_{A,k}^{train,\rho} \cap D_{B,k}^{test,\rho} = \emptyset$. This constraint prevents overlap between training and testing data. Therefore, generalization is the evaluation objective, whereas supervised domain adaptation characterizes the learning setting adopted by the proposed interventions.

3.2. Network Flow Integration

Network flow integration involves the controlled incorporation of flows across datasets to increase the diversity of patterns during training and analyze their impact on model generalization. For this purpose, the selected flows are extracted from dataset B in their original order and incorporated into the training set A . This strategy seeks to introduce patterns present in one domain but absent or underrepresented in another, promoting the building of more robust representations during the learning process.

Let B_N and B_M be the subsets of benign and malicious flows from the target domain D_B , respectively. For a proportion ρ , we define $S_N^\rho \subset B_N$ and $S_M^\rho \subset B_M$ as the subsets selected for integration while preserving the original order of the flows. The three evaluated strategies are defined as: i) $I_N^\rho(D_A^{train}, D_B) = D_A^{train} \cup S_N^\rho$, ii) $I_M^\rho(D_A^{train}, D_B) = D_A^{train} \cup S_M^\rho$, and iii) $I_{NM}^\rho(D_A^{train}, D_B) = D_A^{train} \cup S_N^\rho \cup S_M^\rho$.

The first strategy considers only benign flows, the second only malicious flows, and the third combines both. In all cases, the integrated flows are removed from the corresponding test set. The progressive integration process between datasets A and B considers different proportions, such as 20%, 40%, 60%, and 80%. As an example, when integrating 20% of the malicious flows from B into A , these flows are removed from B before evaluation. In addition, the same proportion is removed from A , preserving the original dataset dimensions and preventing performance gains from being attributed to increased training data volume. Table 2 summarizes the practical applicability, evaluated hypotheses, and methodological constraints of three network flow integration strategies: benign, malicious, and mixed-flow integration. Together, these strategies enable separate analysis of how legitimate, malicious, and combined target-domain patterns affect model performance during generalization across data domains.

Table 2. Practical applicability, hypotheses, and constraints of the network flow integration strategies

Strategy	Practical Applicability	Evaluated Hypothesis	Methodological Constraint
Benign Flow Integration	Scenarios with labeled benign target-domain traffic.	Increasing the diversity of legitimate patterns may improve adaptation to the target domain.	Does not introduce malicious patterns from the target domain.
Malicious Flow Integration	Scenarios with labeled target-domain attacks.	Introducing malicious patterns may improve the detection of attacks shared across domains.	Depends on the availability and correct labeling of target-domain attacks.
Mixed Integration	Scenarios with labeled benign and malicious target-domain traffic.	Combining both flow types may provide a more representative target-domain sample.	Requires labeled data from both classes and is sensitive to their proportions.

In an operational scenario, a labeled public dataset may represent the source domain used for model training, while flows collected from a private network represent the target domain. Unlike the controlled experimental setting, reliable labels are unavailable for target-domain traffic, particularly for malicious flows. Therefore, interventions that

integrate labeled target-domain flows represent a supervised adaptation condition whose operational application depends on the prior collection, validation, and labeling of network traffic. The results quantify the potential benefits of these interventions under controlled data availability and motivate evaluations using real network traffic.

3.3. Validation Scenarios

The models are evaluated under two distinct scenarios. In the Intraset scenario, training and testing are performed on different partitions of the same dataset, allowing the evaluation of model performance within a single network traffic domain: $f_\theta \leftarrow \text{train}(D_A^{\text{train}}), \text{eval}(f_\theta, D_A^{\text{test}})$. In the Intersect scenario, the models are trained on a source domain and evaluated on a distinct target domain, making it possible to analyze their generalization capability in the presence of variations in data distribution, traffic patterns, and threat types: $f_\theta \leftarrow \text{train}(D_A^{\text{train}}), \text{eval}(f_\theta, D_B^{\text{test}})$. After applying the interventions, the training considers the modified dataset: $f_\theta \leftarrow \text{train}(D_{A,k}^{\text{train},\rho}), \text{eval}(f_\theta, D_{B,k}^{\text{test},\rho})$.

Figure 2 illustrates both scenarios. Initially, a partition of dataset A undergoes an intervention and is used for model training. In the Intraset scenario, testing is performed using another partition of the same dataset. In the Intersect scenario, the model is evaluated using dataset B , which belongs to a different data domain and is subject to the constraints defined for each intervention. Thus, the Intersect baseline evaluates cross-domain generalization without access to target-domain data, whereas the intervention scenarios evaluate whether controlled supervised adaptation reduces performance degradation in the target domain.

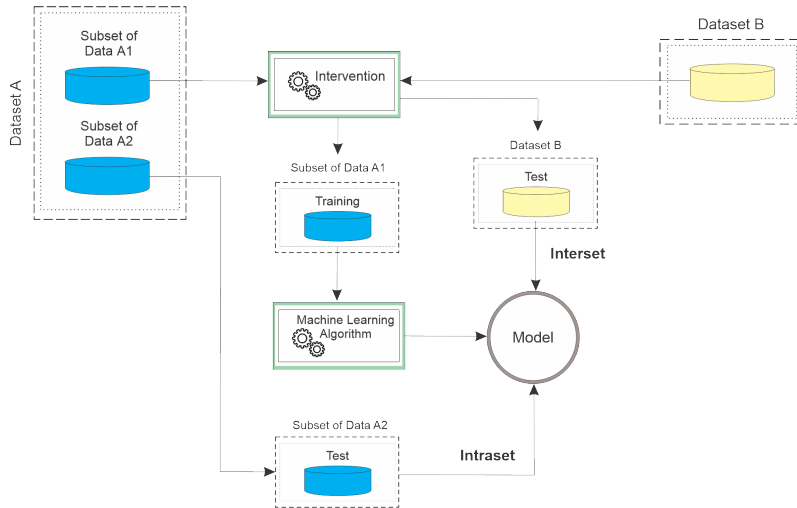


Figure 2. Model performance evaluation scenarios

4. Experiments and Results

The Intraset experiments confirmed that intrusion detection models perform well when trained and tested within the same data domain. Therefore, due to space limitations, the analysis focuses on the more challenging Intersect scenarios, in which training and testing use different network traffic datasets.

The experiments were divided into three intervention groups. IN1 integrated benign, malicious, or mixed flows; IN2 combined these integration strategies with dimensionality reduction; and IN3 combined them with feature selection. The UNSW-NB15, CIC-IDS2017, and CIC-IDS2018 datasets were used in different training and testing combinations to assess cross-domain generalization. Since only XGBoost was evaluated, the conclusions are limited to this algorithm.

All datasets were standardized and preprocessed through normalization, encoding, and manual labeling, following the guidelines presented in [Santos et al. 2024]. Malicious flows were grouped into two classes: (D)DoS, which includes DoS and DDoS attacks common to the three datasets, and Background, which comprises the remaining heterogeneous attacks. This separation enabled the evaluation of both shared attack patterns and attacks not directly integrated during the interventions.

Although NFStream provides a common feature space, the datasets remain distinct domains because of differences in capture environments, applications, traffic composition, temporal characteristics, and attack-generation procedures. The projections and centroid distances presented in Figures 3, 4, and 5 illustrate scenario-specific changes in class distribution and separability, but should be interpreted as exploratory evidence rather than as statistical tests of distributional equivalence.

4.1. Experiments 1, 2, and 3 (Intervention IN1)

In Experiment 1, benign flows were exclusively integrated using a binary classification approach. The results presented in Table 3 show that, although the integration increases the AUC-ROC in some scenarios, degradation occurs in attack detection, as evidenced by the low Recall values for the malicious class. This behavior indicates that integrating benign flows reduces the separability between benign and anomalous behaviors, thereby increasing performance degradation during generalization across data domains.

Figure 3 presents the bidimensional projection of the network flows and the distance between the class centroids in the scenario where the model is trained using CIC-IDS2017 and tested on CIC-IDS2018, with the integration of 40% of benign flows, as this scenario exhibits a pronounced degradation in attack detection after the intervention, allowing its impacts on the model generalization capability to be more clearly observed. In the baseline, Figure 3(a), partial overlap between benign and malicious flows can be observed, indicating low separability between the classes. After the intervention, Figure 3(b), a reorganization of the flows is observed, with greater concentration in specific regions and the presence of possible outliers. In addition, Figure 3(c) shows a reduction in the distance between the centroids, indicating greater proximity between benign and malicious behaviors. These results suggest that integrating benign flows reduces class separability and degrades model performance during generalization across data domains.

In Experiment 2, priority was given to integrating malicious flows using a multi-class approach, considering only the (D)DoS category. Due to the high dimensionality of the results, only the obtained gains are presented. The results in Table 4 show a significant increase in the Recall of this class, indicating a greater ability to identify the integrated attacks. However, this gain is accompanied by degradation in the Background class, a reduction in the Macro F1-Score, and an increase in the FAR in some scenarios, indicating a bias toward detecting specific attacks.

Table 3. Comparison between the baseline and the integration of benign flows

Training	Test	Method	Integration Percentage	Attack Recall	Macro			Δ (Gain)
					F1-Score	FAR	AUC-ROC	
UNSW-NB15	CIC-IDS2017	Baseline	—	8,57	49,66	8,71	52,37	—
		IN1 (Benign)	20	0,00	46,10	0,00	81,98	-8,57 / -3,56 / -8,71 / +29,61
			40	0,00	45,89	0,00	76,61	-8,57 / -3,77 / -8,71 / +24,24
			60	0,00	45,65	0,00	65,17	-8,57 / -4,01 / -8,71 / +12,80
			80	0,00	45,39	0,00	72,34	-8,57 / -4,27 / -8,71 / +19,97
UNSW-NB15	CIC-IDS2018	Baseline	—	0,24	44,27	5,65	60,77	—
		IN1 (Benign)	20	0,00	45,07	0,00	50,51	-0,24 / +0,80 / -5,65 / -10,26
			40	0,00	44,81	0,00	58,40	-0,24 / +0,54 / -5,65 / -2,37
			60	0,00	44,52	0,00	60,38	-0,24 / +0,25 / -5,65 / -0,39
			80	0,00	44,20	0,00	72,24	-0,24 / -0,07 / -5,65 / +11,47
CIC-IDS2017	UNSW-NB15	Baseline	—	0,00	49,01	0,01	65,21	—
		IN1 (Benign)	20	0,00	48,77	0,00	87,69	0,00 / -0,24 / -0,01 / +22,48
			40	0,00	48,37	0,00	95,29	0,00 / -0,64 / -0,01 / +30,08
			60	0,00	47,60	0,00	95,29	0,00 / -1,41 / -0,01 / +30,08
			80	0,00	45,41	0,00	95,16	0,00 / -3,60 / -0,01 / +29,95
CIC-IDS2017	CIC-IDS2018	Baseline	—	50,30	79,10	1,54	89,26	—
		IN1 (Benign)	20	0,00	44,52	0,00	44,52	-50,30 / -34,58 / -1,54 / -44,74
			40	0,00	42,95	0,00	85,44	-50,30 / -36,15 / -1,54 / -3,82
			60	0,00	40,12	0,00	81,90	-50,30 / -38,98 / -1,54 / -7,36
			80	0,00	33,51	0,00	87,60	-50,30 / -45,59 / -1,54 / -1,66
CIC-IDS2018	UNSW-NB15	Baseline	—	0,00	48,08	3,65	58,02	—
		IN1 (Benign)	20	0,00	48,77	0,00	94,46	0,00 / +0,69 / -3,65 / +36,44
			40	0,00	48,37	0,00	95,90	0,00 / +0,29 / -3,65 / +37,88
			60	0,00	47,60	0,00	94,89	0,00 / -0,48 / -3,65 / +36,87
			80	0,00	45,41	0,00	94,71	0,00 / -2,67 / -3,65 / +36,69
CIC-IDS2018	CIC-IDS2017	Baseline	—	0,55	46,99	0,11	80,51	—
		IN1 (Benign)	20	0,00	45,65	0,00	88,88	-0,55 / -1,34 / -0,11 / +8,37
			40	0,00	44,37	0,00	79,55	-0,55 / -2,62 / -0,11 / -0,96
			60	0,00	42,00	0,00	86,31	-0,55 / -4,99 / -0,11 / +5,80
			80	0,00	36,21	0,00	79,05	-0,55 / -10,78 / -0,11 / -1,46

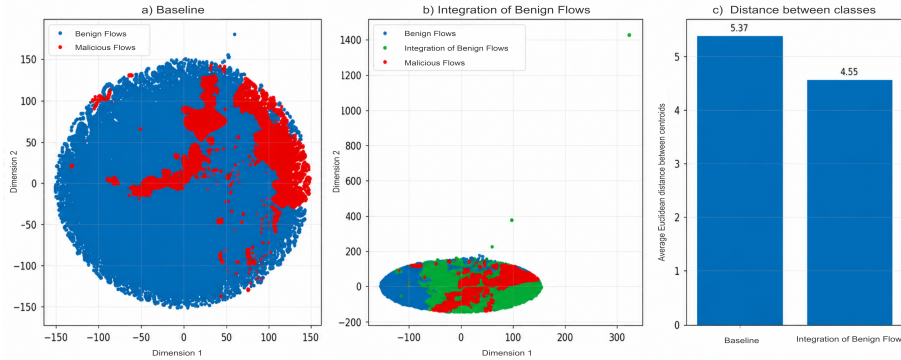


Figure 3. Class separation in benign flow integration

Figure 4 presents the bidimensional projection in the scenario where the model was trained using the CIC-IDS2018 dataset and tested on CIC-IDS2017, using the integration of 60% malicious flows, as this scenario simultaneously presents high Recall for (D)DoS attacks, partial recovery of the Background class, and improvements in the global metrics. In the baseline, Figure 4(a), overlap between normal and (D)DoS flows can be observed, in addition to the dispersion of the Background class, indicating low separability among the patterns. After the intervention, Figure 4(b), a reorganization of the flows in the projected space is observed, with more defined regions associated with (D)DoS attacks and lower overlap with benign flows. Furthermore, Figure 4(c) shows an increase in the distance between the centroids, indicating greater separability between benign and malicious behaviors. These results suggest that integrating malicious flows enhances the

Table 4. Comparison between the baseline and the Integration of malicious flows

Training	Test	Method	Integration Percentage	Δ (Gain)					
				(D)DoS		Background		Macro	
				F1-Score	Recall	F1-Score	Recall	F1-Score	AUC-ROC
UNSW-NB15	CIC-IDS2017	IN1 (Malicious)	20%	+13,64	+100,00	-1,42	-6,18	-18,94	-8,61
			40%	+10,49	+100,00	-1,42	-6,18	-20,70	-8,27
			60%	+7,25	+100,00	-1,42	-6,18	-21,78	-8,27
			80%	+3,76	+100,00	-1,42	-6,18	-22,95	-8,27
UNSW-NB15	CIC-IDS2018	IN1 (Malicious)	20%	+14,63	+100,00	-49,99	-63,86	-42,25	+11,52
			40%	+11,35	+100,00	-49,99	-63,86	-43,61	+11,64
			60%	+7,86	+100,00	-49,99	-63,86	-44,77	+11,64
			80%	+4,09	+100,00	-49,99	-63,86	-46,03	+11,64
CIC-IDS2017	UNSW-NB15	IN1 (Malicious)	20%	+0,97	+99,43	-0,02	-0,01	-27,05	-1,83
			40%	+0,70	+99,96	-0,02	-0,01	-30,98	-2,00
			60%	+0,46	+100,00	-0,02	-0,01	-31,07	-0,01
			80%	+0,23	+100,00	-0,02	-0,01	-31,16	0,00
CIC-IDS2017	CIC-IDS2018	IN1 (Malicious)	20%	+2,46	+86,27	-60,11	-63,63	-28,17	-0,13
			40%	-3,99	+86,28	-60,14	-63,65	-31,44	+1,16
			60%	-1,66	+86,28	-60,13	-63,65	-23,71	-7,04
			80%	-16,29	+86,28	-60,13	-63,65	-34,22	-0,67
CIC-IDS2018	UNSW-NB15	IN1 (Malicious)	20%	+0,92	+100,00	0,00	0,00	-30,20	-1,41
			40%	+0,70	+100,00	0,00	0,00	-30,29	-1,41
			60%	+0,46	+100,00	0,00	0,00	-30,37	-1,41
			80%	+0,23	+100,00	0,00	0,00	-30,44	-1,41
CIC-IDS2018	CIC-IDS2017	IN1 (Malicious)	20%	+43,42	+100,00	+0,55	+0,29	+12,29	-14,38
			40%	+36,29	+100,00	+2,77	+2,39	+9,93	-11,47
			60%	+25,34	+100,00	+48,61	+86,20	+20,60	-21,86
			80%	+11,71	+100,00	+48,33	+86,77	+14,31	-20,00

model's discriminative capability and improves the detection of specific attacks.

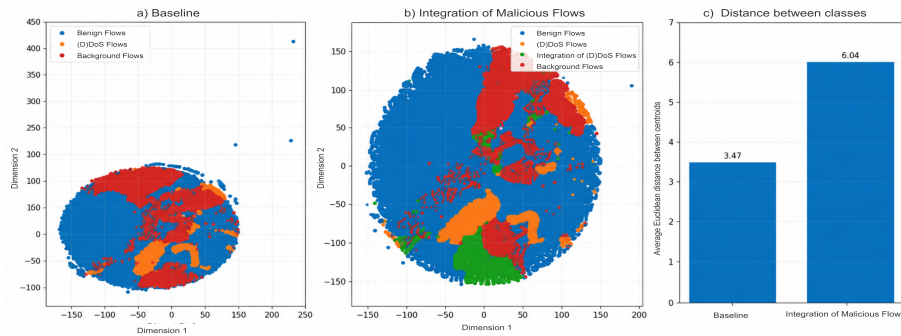


Figure 4. Class separation in malicious flow integration

In Experiment 3, the mixed integration of flows was evaluated while maintaining the same multiclass approach adopted previously. The results presented in Table 5 indicate an increase in the Recall of (D)DoS attacks. However, this improvement is accompanied by degradation in the Background class, evidencing a trade-off between sensitivity and overall class balance. Nevertheless, some scenarios presented simultaneous gains in the Recall of (D)DoS attacks and global metrics, suggesting that the mixed integration of flows partially increases the diversity of the patterns observed during training.

Figure 5 presents the bidimensional projection in the scenario where the model is trained on CIC-IDS2018 and tested on CIC-IDS2017, with 20% mixed flow integration, as this scenario provides better class balance and a better representation of different flow types in the projected space. In the baseline, Figure 5(a), partial overlap among benign, (D)DoS, and Background flows can be observed, indicating limited separability between the classes. After the joint integration of mixed flows, Figure 5(b), a reorganization of

Table 5. Comparison between the baseline and mixed flow integration

Training	Test	Method	Integration Percentage	Δ (Gain)					
				(D)DoS		Background		Macro	
				F1-Score	Recall	F1-Score	Recall	F1-Score	AUC-ROC
UNSW-NB15	CIC-IDS2017	IN1 (Benign and Malicious)	20%	+18,44	+100,00	-1,42	-6,18	-9,18	-12,85
			40%	+16,20	+100,00	-1,42	-6,18	-17,57	-8,35
			60%	+96,44	+93,47	-1,42	-6,18	+39,22	-22,34
			80%	+95,29	+100,00	-1,42	-6,18	+37,08	-17,86
UNSW-NB15	CIC-IDS2018	IN1 (Benign and Malicious)	20%	+78,03	+99,96	-49,99	-63,86	+11,34	-14,04
			40%	+72,76	+99,98	-49,99	-63,86	+9,39	-12,59
			60%	+99,98	+99,99	-49,99	-63,86	+16,19	-0,41
			80%	+47,07	+99,97	-49,99	-63,86	-0,66	-6,10
CIC-IDS2017	UNSW-NB15	IN1 (Benign and Malicious)	20%	+8,92	+89,22	-0,02	-0,01	+3,45	-15,81
			40%	+5,20	+97,89	+0,01	0,00	+1,93	-19,96
			60%	+4,48	+97,97	-0,02	-0,01	+0,37	-14,02
			80%	+4,60	+95,94	-0,02	-0,01	-3,18	-5,72
CIC-IDS2017	CIC-IDS2018	IN1 (Benign and Malicious)	20%	+53,97	+86,27	-60,18	-63,68	-0,89	-10,61
			40%	+48,66	+86,27	-60,18	-63,68	-2,86	-9,16
			60%	+39,94	+86,25	-60,18	-63,68	-6,15	-6,92
			80%	+23,01	+86,26	-60,18	-63,68	-12,90	-2,66
CIC-IDS2018	UNSW-NB15	IN1 (Benign and Malicious)	20%	+18,30	+45,48	0,00	0,00	+6,68	-5,62
			40%	+14,35	+52,43	0,00	0,00	+4,75	-6,03
			60%	+11,52	+58,15	0,00	0,00	+2,59	-5,59
			80%	+5,99	+43,60	0,00	0,00	-2,48	-4,16
CIC-IDS2018	CIC-IDS2017	IN1 (Benign and Malicious)	20%	+71,90	+58,43	0,00	0,00	+24,66	-12,26
			40%	+56,77	+41,26	0,00	0,00	+19,02	-7,54
			60%	+77,74	+64,85	0,00	0,00	+25,85	-10,03
			80%	+91,92	+88,01	0,00	0,00	+29,21	-8,81

the flows in the projected space is observed, with expansion of the regions occupied by the integrated patterns and partial reduction of overlap between the classes, despite the existence of isolated points associated with possible outliers. In addition, Figure 5(c) shows an increase in the distance between the centroids, indicating greater separation. The results suggest that mixed integration favors a more balanced distribution of patterns and reduces performance degradation during generalization across data domains.

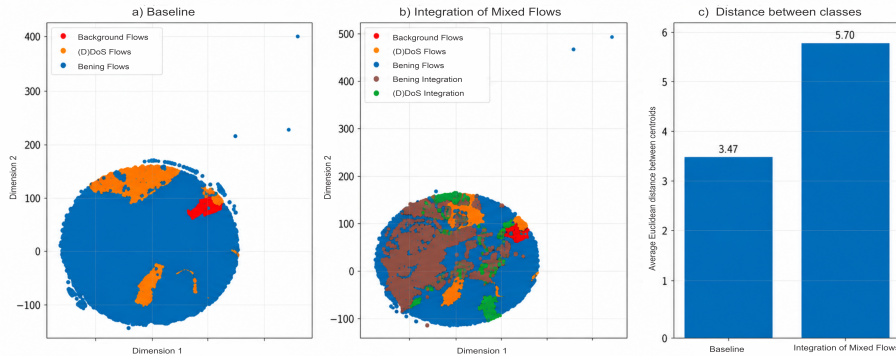


Figure 5. Class separation in mixed flow integration

4.2. Experiments 4, 5, and 6 (Intervention IN2)

Experiments 4, 5, and 6 were conducted using simultaneous interventions at the feature and instance levels. For this purpose, Intervention IN2 was defined, combining dimension reduction with network flow integration. The evaluated configurations considered: integration of benign flows (Experiment 4), integration of malicious flows (Experiment 5), and mixed integration of benign and malicious flows (Experiment 6). For dimension reduction, Principal Component Analysis (PCA) was applied to transform the features

into a lower-dimensional representation, as illustrated in Figure 6. The reduction corresponded to approximately 65% of the features, decreasing from 70 to 25 features while preserving explained variances of approximately 99%, 96%, and 98% for the UNSW-NB15, CIC-IDS2017, and CIC-IDS2018 datasets, respectively.

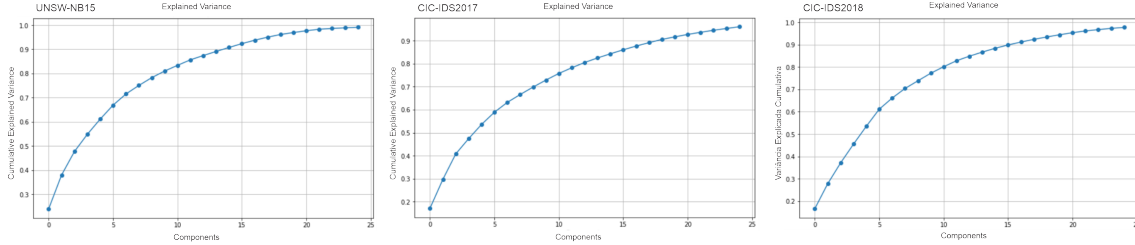


Figure 6. Explained variance by PCA

In Experiment 4, Table 6 presents the results of combining dimension reduction with the integration of benign flows in the CIC-IDS2017 → CIC-IDS2018 scenario, using 40% integration. It was observed that the strategy does not outperform the baseline, resulting in losses in Recall and global metrics, which compromise attack detection capability. Although it offers slight improvements over isolated benign-flow integration, the results indicate performance degradation during generalization to the target domain.

Table 6. Comparison of (IN2) benign flow integration with the baseline and IN1

Training	Test	Method	Attack Recall	Macro			Δ (Gain)
				F1-Score	FAR	AUC-ROC	
CIC-IDS2017	CIC-IDS2018	Baseline	50,30	79,10	1,54	89,26	—
		IN1	0,00	42,95	0,00	85,44	Baseline: -50,30 / -36,15 / -1,54 / -3,82
		IN2	15,65	57,45	0,02	81,84	Baseline: -34,65 / -21,65 / -1,52 / -7,42 IN1: +15,65 / +14,50 / +0,02 / -3,60

In Experiment 5, Table 7 presents the results of combining dimension reduction with integration of malicious (D)DoS flows in the CIC-IDS2018 → CIC-IDS2017 scenario, using 60% integration. The strategy achieved gains in the Recall and F1-Score of the (D)DoS class compared with the baseline, indicating greater capability to detect this attack. However, degradation in the Background class reveals a loss of class balance and limited performance across different types of malicious traffic in the target domain.

Table 7. Comparison of (IN2) malicious flow integration with the baseline and IN1

Training	Test	Method	Comparison	Δ (Gain)					
				(D)DoS		Background		Macro	
				F1-Score	Recall	F1-Score	Recall	F1-Score	FAR AUC-ROC
CIC-IDS2018	CIC-IDS2017	Baseline	—	0,00	0,00	0,00	0,00	30,94	33,39 85,08
		IN1	Baseline	+25,35	+100,00	+48,61	+86,20	+20,60	-21,86 +4,83
		IN2	Baseline	+31,48	+95,53	0,00	0,00	+9,03	-9,51 -10,74
			IN1	+6,13	-4,74	-48,61	-86,20	-11,57	+12,35 -15,57

In Experiment 6, Table 8 presents the results of combining dimension reduction with mixed integration of benign and malicious flows in the CIC-IDS2018 → CIC-IDS2017 scenario, using 20% integration. The results show gains in (D)DoS class metrics and global metrics compared with the baseline. Compared with isolated integration of mixed flows, the strategy showed similar performance while reducing the false-alarm rate. These results indicate that the combination of dimension reduction and mixed integration

favors a more balanced representation, though it still has limitations in comprehensively detecting different attack categories.

Table 8. Comparison of (IN2) mixed flow integration with the baseline and IN1

Training	Test	Method	Δ (Gain)							
			Comparison	(D)DoS		Background		Macro		
				F1-Score	Recall	F1-Score	Recall	F1-Score	FAR	AUC-ROC
CIC-IDS2018	CIC-IDS2017	Baseline	–	0.00	0.00	0.00	0.00	30.94	33.39	85.08
		IN1	Baseline	+71.90	+58.43	0.00	0.00	+24.66	-12.26	+11.08
		IN2	Baseline	+70.13	+55.45	+0.02	+0.01	+24.03	+0.77	+11.24
			IN1	-1.77	-2.89	+0.02	+0.01	-0.63	-11.49	+0.16

4.3. Experiments 7, 8, and 9 (Intervention IN3)

Experiments 7, 8, and 9 follow the same configuration adopted in interventions IN1 and IN2, maintaining the training and testing scenarios, as well as the percentages and types of flow integration. The main difference lies in replacing dimension reduction with feature selection, combined with the integration strategies of benign flows (Experiment 7), malicious flows (Experiment 8), and mixed flows (Experiment 9). For feature selection, the Chi-Square test was used to assess the dependence between each feature and the target class, identifying the features most relevant to the classification task. Similar to dimension reduction, the reduction was close to 65%, decreasing from 70 to 25 features.

In Experiment 7, Table 9 presents the results of combining feature selection with the integration of benign flows, considering the same scenario adopted in Experiment 4. It was observed that the strategy neither outperforms the baseline nor presents relevant gains compared with the previous interventions. The results indicate that the model retains low attack detection capability, increasing performance degradation during generalization across network traffic domains.

Table 9. Comparison of (IN3) benign flow integration with baseline, IN1, and IN2

Training	Test	Method	Attack Recall	Macro			Δ (Gain)
				F1-Score	FAR	AUC-ROC	
CIC-IDS2017	CIC-IDS2018	Baseline	50,30	79,10	1,54	89,26	–
		IN1	0,00	42,95	0,00	85,44	Baseline: -50,30 / -36,15 / -1,54 / -3,82
		IN2	15,65	57,45	0,02	81,84	Baseline: -34,65 / -21,65 / -1,52 / -7,42 IN1: +15,65 / +14,50 / +0,02 / -3,60
							Baseline: -50,30 / -36,15 / -1,54 / -3,51 IN1: 0,00 / 0,00 / 0,00 / +0,31
		IN3	0,00	42,95	0,00	85,75	IN2: -15,65 / -14,50 / -0,02 / +3,91

In Experiment 8, Table 10 presents the results of combining feature selection with the integration of malicious (D)DoS flows, using the same scenarios as in Experiment 5. The strategy proved effective at maximizing detection of this attack class, yielding gains in metrics associated with (D)DoS compared with the baseline. However, an inability to identify the Background class was observed, along with an increase in FAR and a reduction in global metrics compared with the previous interventions. These results indicate specialization and limited performance during generalization in multiclass scenarios.

In Experiment 9, Table 11 presents the results of combining feature selection with mixed integration of benign and malicious flows, considering the same scenario adopted in Experiment 6. The best performance among the evaluated approaches was observed, with significant gains in metrics associated with (D)DoS attacks and improvements in

Table 10. Comparison of (IN3) malicious flow integration with baseline, IN1, IN2

Training	Test	Method	Δ (Gain)							
			Comparison	(D)DoS		Background		Macro		
				F1-Score	Recall	F1-Score	Recall	F1-Score	FAR	AUC-ROC
CIC-IDS2018	CIC-IDS2017	Baseline	–	0,00	0,00	0,00	0,00	30,94	33,39	85,08
		IN1	Baseline	+25,35	+100,00	+48,61	+86,20	+20,60	-21,86	+4,83
		IN2	Baseline	+31,48	+95,53	0,00	0,00	+9,03	-9,51	-10,74
			IN1	+6,13	-4,74	-48,61	-86,20	-11,57	+12,35	-15,57
		IN3	Baseline	+26,72	+100,00	+0,03	+0,02	+6,15	+42,19	-10,98
			IN1	+1,37	0,00	-48,58	-86,18	-14,45	+64,05	-15,81
			IN2	-4,76	+4,47	+0,03	+0,02	-2,88	+51,70	-0,24

global metrics relative to the baseline and interventions IN1 and IN2. Although limitations in identifying the Background class persist, combining feature selection with mixed integration produces a more discriminative representation, greater class balance, and reduced performance degradation for specific attacks in the target domain.

Table 11. Comparison of (IN3) mixed flow integration with baseline, IN1, and IN2

Training	Test	Method	Δ (Gain)							
			Comparison	(D)DoS		Background		Macro		
				F1-Score	Recall	F1-Score	Recall	F1-Score	FAR	AUC-ROC
CIC-IDS2018	CIC-IDS2017	Baseline	–	0,00	0,00	0,00	0,00	30,94	33,39	85,08
		IN1	Baseline	+71,90	+58,43	0,00	0,00	+24,66	-12,26	+11,08
		IN2	Baseline	+70,13	+55,45	+0,02	+0,01	+24,03	+0,77	+11,24
			IN1	-1,77	-2,89	+0,02	+0,01	-0,63	-11,49	+0,16
		IN3	Baseline	+97,28	+99,58	0,00	0,00	+33,77	-20,83	+13,69
			IN1	+25,38	+41,15	0,00	0,00	+9,11	-8,58	+2,61
			IN2	+27,15	+44,13	-0,02	-0,01	+9,74	-9,35	+2,45

4.4. Comparative Analysis of Network Flow Integration Interventions

Based on the results, Table 12 summarizes the intervention strategies (IN1, IN2, and IN3), highlighting their predominant effects, main limitations, and performance outcomes. Overall, model behavior varies according to the type of integrated flow, indicating that data composition influences performance during generalization across data domains.

Table 12. Comparative summary of network flow integration strategies

Experiment	Strategy	Predominant Effect	Limitation	Highlight
1, 4, and 7	Benign Flow Integration.	Greater representation of benign target-domain patterns.	Low attack detection capability.	Improved representation of benign traffic, but with loss of class separability.
2, 5, and 8	Malicious Flow Integration.	Specialization in the detection of specific attacks.	Loss of class balance and increase in the FAR.	Greater sensitivity to (D)DoS attacks, with bias toward the malicious class.
3, 6, and 9	Mixed Integration of Benign and Malicious Flows.	Greater metric balance and reduced generalization performance degradation.	Limitations in identifying the Background class.	Best overall performance among the evaluated strategies.

Strategies based solely on integrating benign flows tend to reduce class separability, compromising attack detection. In contrast, integrating only malicious flows favors identifying specific attacks, such as the (D)DoS class, but at the expense of balance across the remaining classes and an increased false positive rate. In turn, mixed integration yielded the most balanced results, with simultaneous gains in class-specific and global metrics in some target-domain scenarios. However, persistent limitations in detecting the Background class show that these gains were not uniform across attack categories.

Therefore, the interventions should be assessed jointly using class-specific F1-Score and Recall and the Macro F1-Score, FAR, and AUC-ROC.

Figure 7 conceptually illustrates the effects of these strategies on data distribution and class separability. In Figure 7(a), the integration of benign flows brings normal and malicious patterns closer together, reducing class separability. In Figure 7(b), the integration of malicious flows increases the separability of the (D)DoS class, but reduces the balance among the remaining classes. Meanwhile, Figure 7(c) shows that mixed integration promotes a more balanced data distribution, favoring overall separability and reducing performance degradation during generalization. At the feature level, strategies based on feature selection (IN3) showed greater sensitivity to the composition of the integrated flows, whereas dimension reduction approaches (IN2) exhibited more stable behavior. However, IN3 achieved the most promising results under mixed-flow integration, suggesting that feature selection may produce more discriminative representations when training includes greater pattern diversity.

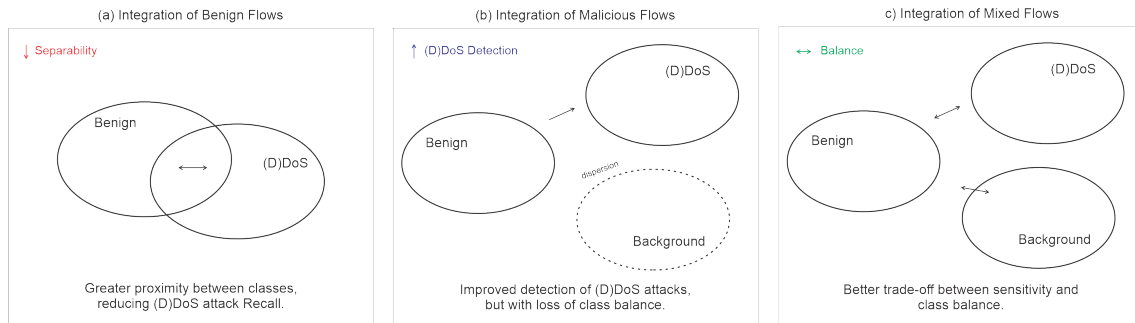


Figure 7. Conceptual comparison between network flow integration strategies

5. Conclusion and Future Work

This work analyzed domain adaptation strategies using controlled integration of labeled target-domain flows to increase training diversity and reduce performance degradation in the generalization of machine learning-based IDS. Results showed that effectiveness depends on the type of integrated flow. Benign-flow integration compromised attack detection, whereas malicious-flow integration favored specialization in attacks. Mixed-flow integration produced the most balanced results, particularly when combined with feature-level interventions. However, persistent difficulty in detecting the Background class indicates the performance gains were not uniform across attack categories. These findings show that supervised integration can improve performance in specific target domains, but should not be interpreted as generalization to completely unseen domains.

Despite its contributions, this work is limited by the use of a single learning algorithm, the limited representation of some attack categories, and the exclusive use of public datasets, which may not reflect real-world operational environments. Moreover, the results were not evaluated using confidence intervals or significance tests. Future work will integrate real network traffic, evaluate additional learning algorithms and intervention strategies, and apply statistical tests, confidence intervals, and distribution-distance measures to assess the robustness of the results and quantify differences between domains.

Artifact Availability

The supplementary materials of this paper, including source code, feature extracted using the NFStream tool, and additional experimental results (tables and figures), are available in the following GitHub repository: <https://github.com/kelsonc/paper-sbseg2026.git>

Acknowledgments

This work was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grants 409743/2025-9 and 408432/2024-1.

References

- Aouini, Z. and Pekar, A. (2022). Nfstream: A flexible network data analysis framework. *Computer Networks*, 204:1–8.
- Domingues, M., Bertoli, G., Melo, L., Saotome, O., Santos, A., and Pereira, L. (2022). Avaliação da capacidade de generalização de ids stateful utilizando aprendizado de máquina. In *Anais do XXII SBSEG (2022)*, pages 236–249. SBC. Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSEG), Santa Maria, RS, 12-15 Set 2023.
- D’hooge, L., Wauters, T., Volckaert, B., and De Turck, F. (2020). Inter-dataset generalization strength of supervised machine learning methods for intrusion detection. *Journal of Information Security and Applications*, 54:1–13.
- Kenyon, A., Deka, L., and D., E. (2020). Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets. *Computers & Security*, 99:1–26.
- Layeghy, S. and Portmann, M. (2022). On generalisability of machine learning-based network intrusion detection systems. *arXiv preprint arXiv:2205.04112*, [s.n.]:1–12.
- Mahfouz, A., Abuhussein, A., Venugopal, D., and Shiva, S. (2020). Ensemble classifiers for network intrusion detection using a novel network attack dataset. *Future Internet*, 12(11):1–19.
- Marvi, M., Arfeen, A., and Uddin, R. (2021). A generalized machine learning-based model for the detection of ddos attacks. *International Journal of Network Management*, 31(6):1–22.
- Molina-Coronado, B., Mori, U., Mendiburu, A., and Miguel-Alonso, J. (2020). Survey of network intrusion detection methods from the perspective of the knowledge discovery in databases process. *IEEE Transactions on Network and Service Management*, 17(4):2451–2479.
- Moustafa, N. and Slay, J. (2015). A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 10-12 Nov 2015.
- Putra, W. and Huang, J. J. (2019). A survey of intrusion detection system. *International Journal of Informatics and Computation*, 1(1):1–19.

- Rocha, M. S., Bernardo, G. D., Mundim, L., Zarpelão, B. B., and Miani, R. S. (2023). Supervised machine learning and detection of unknown attacks: An empirical evaluation. AINA 2023: International Conference on Advanced Information Networking and Applications, Juiz de Fora, MG, Brazil, 29-31 Mar 2023.
- Santos, K. and Miani, R. (2025). Impacto da redução de dimensão e seleção de atributos na generalização de modelos de detecção de intrusão. In *Anais do XLIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 728–741, Porto Alegre, RS, Brasil. SBC.
- Santos, K. C., Miani, R. S., and de Oliveira Silva, F. (2024). Evaluating the impact of data preprocessing techniques on the performance of intrusion detection systems. *Journal of Network and Systems Management*, 32(36):1–54.
- Sharafaldin, I., Lashkari, A. H., and Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. 4th International Conference on Information Systems Security and Privacy (ICISSp), Funchal, Madeira, Portugal, 22-24 Jan 2018.
- Sudjana, D., Lin, Y., Verkerken, M., Hwang, R., Lai, Y., D’Hooge, L., Wauters, T., Volckaert, B., and De Turck, F. (2024). Improving generalization of ml-based ids with lifecycle-based dataset, auto-learning features, and deep learning. *IEEE Transactions on Machine Learning in Communications and Networking*, 2:645–662.
- Verkerken, M., D’hooge, L., Wauters, T., Volckaert, B., and De Turck, F. (2021). Towards model generalization for intrusion detection: Unsupervised machine learning techniques. *Journal of Network and Systems Management*, 30(1):1–25.
- Zoppi, T., Ceccarelli, A., Puccetti, T., and Bondavalli, A. (2023). Which algorithm can detect unknown attacks? comparison of supervised, unsupervised and meta-learning algorithms for intrusion detection. *Computers & Security*, 127:1–12.