Results in Engineering 31 (2026) 111405

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Multimodal contrastive learning-based network intrusion detection

Giampaolo Bovenzi [a,*], Idio Guarino [b], Antonio Pescapè [c]

[a] Department of Information Science and Technology, Pegaso Telematic University, Napoli, 80143, Italy
[b] Department of Computer Science and Engineering, University of Bologna, Bologna, 40136, Italy
[c] Department of Electrical Engineering and Information Technology, University of Napoli Federico II, Napoli, 80125, Italy

ARTICLE INFO

Keywords:
Network intrusion detection
Contrastive learning
Representation learning
Multimodal learning

ABSTRACT

The increasing adoption of the Internet has intensified cybersecurity risks such as data theft and service disruptions. Network intrusion detection is critical for identifying malicious activity, but traditional methods face limitations. Machine Learning (ML) and Deep Learning (DL) techniques often suffer from poor generalization, limiting their ability to transfer models across heterogeneous network datasets and reducing their empirical generalizability. To address this feature fragility, we propose a Multimodal Contrastive Learning approach for network intrusion detection. *The key design choice lies in a multi-step training orchestration designed to stabilize latent representations under significant dataset shift.* The framework integrates multimodal data for richer traffic modeling and contrastive learning to enhance latent distinction between benign and malicious traffic. To assess cross-dataset generalization, models are pre-trained on Edge-IIoT and adapted on IoT-NID and CIC-IoMT. We investigate two contrastive strategies—self-supervised and supervised—to analyze the impact of label-agnostic and label-aware learning under distribution shift.

Experiments show that contrastive multimodal learning matches transfer learning in saturated regimes (F1 Score > 90%) on both IoT-NID and CIC-IoMT. On IoT-NID, specific advantages emerge: improved Recall (up to +4%) on underrepresented classes; better short-flow performance (+3% F1 Score for biflows of three or fewer packets); and a more structured latent space (0.90 vs. 1.12 DB index). On CIC-IoMT, contrastive approaches outrank transfer learning, with supervised contrastive achieving the best discriminative performance and self-supervised yielding the most compact latent space. Within the evaluated source–target setups, our findings suggest that contrastive learning is a viable alternative to transfer learning, and that latent structural quality is a meaningful generalization indicator under domain shifts.

1. Introduction

The digital landscape has reached unprecedented scale, with about 68.7% of the world's population connected to the Internet by 2025.[1] In particular, ubiquitous connectivity has led to a massive surge in network traffic, further accelerated by the spread of Internet of Things (IoT) devices, which are expected to reach 39 billion by the end of 2030.[2] This increasing network activity, however, has widened the cyberattack surface, with home IoT networks facing an average of 29 attempted attacks every day,[3] contributing to data breach costs that reached approximately $4.44 million in 2025.[4]

Adopting effective countermeasures such as Network Intrusion Detection Systems (NIDSs) is crucial for efficiently defending against cyber threats and ensuring network security. NIDSs monitor traffic to detect *misuse* and *anomalies*, distinguishing legitimate (viz., benign) traffic from potentially malicious traffic—such as malware- or attack-related communications. They often provide further categorization of detected attacks (viz., *attack classification*). While early NIDSs relied on handcrafted rule- and signature-based approaches, requiring constant maintenance by security analysts, the scale and diversity of modern threats have necessitated a shift toward Machine Learning (ML) and Deep Learning (DL) solutions. Buczak and Guven [1], Ahmad et al. [2], Harshdeep et al. [3].

Nevertheless, learning-based NIDSs often experience performance drops when tested outside their training distribution. This stems from a fundamental *representation fragility*, as standard optimization overfits

* Corresponding author.
E-mail addresses: giampaolo.bovenzi@unina.it, giampaolo.bovenzi@unipegaso.it (G. Bovenzi), idio.guarino@unibo.it (I. Guarino), pescape@unina.it (A. Pescapè).

[1] https://datareportal.com/reports/digital-2025-july-global-statshot
[2] https://iot-analytics.com/number-connected-iot-devices
[3] https://www.bitdefender.com/en-au/blog/hotforsecurity/bitdefender-and-netgear-2025-iot-security-landscape-report-shows-alarming-rise-in-smart-home-threats
[4] https://blog.barracuda.com/2025/09/02/survey-global-decline-data-breach-costs

https://doi.org/10.1016/j.rineng.2026.111405
Received 22 December 2025; Received in revised form 30 May 2026; Accepted 4 June 2026
Available online 6 June 2026
2590-1230/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

to dataset-specific artifacts rather than capturing robust cross-dataset patterns. Current solutions for cross-domain adaptability are primarily based on the *Transfer Learning* (TL) paradigm, where a model pre-trained on a source network is adapted to a specific target domain [4,5]. However, effectiveness strongly depends on the similarity between the source and target networks. Given the significant domain shifts typical of heterogeneous IoT environments, standard TL often fails to preserve discriminative feature structures, resulting in unstable performance and reduced robustness.

Concurrently, *multimodal learning* has emerged as an approach to integrate complementary representations (viz., modalities) of traffic data, e.g., packet content and packet-level sequences, and improve detection under static conditions [6]. Yet, combining multimodal inputs with TL does not resolve *feature fragility* if training remains driven by standard discriminative objectives that overfits to source-specific traits.

More recently, Contrastive Learning (CL) offers a promising alternative by directly shaping the geometry of the latent space, enforcing intra-class compactness and inter-class separability without relying on decision-boundary optimization [7,8]. Despite its demonstrated benefits in other domains, the systematic application of CL to NIDS—particularly in a multimodal, cross-dataset setting—remains largely unexplored. In fact, existing works that combine multimodal and contrastive objectives [9] mostly focus on label efficiency rather than *representation stability under dataset shift*. Furthermore, their monolithic training procedures are susceptible to *gradient interference*, where task-specific updates override the structural organization of learned features.

To bridge this gap, we propose a *multi-step multimodal contrastive framework* that decouples representation learning from task-specific optimization, prioritizing *latent space structure* as the foundation for generalization. It is worth noting that the contribution of this work lies not in novel architectural components, but in the *orchestration* of established learning paradigms toward a specific engineering objective, namely representation stability under dataset shift. To this end, our empirical investigation focuses on assessing whether: (*i*) geometric constraints on the latent space—quantified via the Davies–Bouldin index—yield superior adaptation under distribution shifts compared to standard cross-entropy-based optimization; (*ii*) structurally organized latent representations mitigate the systematic misclassification of underrepresented attack classes, which typically suffer the most severe degradation under such shifts; (*iii*) self-supervised (SimCLR) and supervised (SupCon) strategies exhibit distinct generalization advantages, enabling a systematic comparison of label-agnostic versus label-aware representations within a multi-step orchestration that establishes stable representations before task-specific fine-tuning; and (*iv*) the contrastive optimization overhead remains restricted entirely to the offline phase, preserving a lightweight real-time inference latency suitable for resource-constrained IoT edge devices.

To assess cross-dataset generalizability, the framework is validated through a multi-target adaptation protocol. Specifically, models are pre-trained on `Edge-IIoT` and subsequently transferred (viz., adapted) to `IoT-NID` and `CIC-IoMT`, deliberately exposing them to heterogeneous traffic distribution and threat landscapes. Within the evaluated source–target setups, the contrastive strategy yields overall classification metrics on par with standard TL baselines, while consistently outperforming them in the structural organization of the latent space and the detection of underrepresented attack classes.

The **main contributions** of this work are summarized as follows:

- We propose a *multimodal contrastive learning framework* for NIDS, featuring a specific *multi-step training orchestration* designed to stabilize feature representations and enhance the distinction between benign and malicious traffic under significant dataset shift.
- We provide a *systematic investigation of two contrastive paradigms*—i.e., self-supervised (SimCLR) and supervised (SupCon)—analyzing the specific impact of label-agnostic versus label-aware strategies on the structural organization of the latent space.
- We evaluate the framework in a *multi-target cross-dataset adaptation scenario*: we utilize `Edge-IIoT` as the source dataset to pre-train the model and evaluate the transferability of the learned representations on *two* heterogeneous target datasets, `IoT-NID` and `CIC-IoMT`. This allows us to assess the robustness of the structural priors when transferred to target datasets with significantly different traffic distributions and threat models.
- We demonstrate that the contrastive paradigm achieves classification performance competitive with classical TL, while providing measurable advantages in latent space regularization (quantified via the Davies-Bouldin index) and in the detection of minority attack classes. We interpret these results as evidence that contrastive pre-training preserves the structural integrity of the learned representation, making it a viable alternative to standard TL in the evaluated cross-dataset generalization scenarios where representation quality is the primary concern.
- We conduct a *comprehensive ablation and architectural study* comparing single- and multi-modal configurations, as well as intermediate- and late-fusion strategies. Our findings demonstrate that: (*i*) the multi-modal approach consistently outperforms the single-modal counterparts; (*ii*) the performance trends, i.e., the superiority of contrastive learning and fine-tuning, are systematically mirrored in single-modal architectures; and (*iii*) feature-level synergy (viz., intermediate-fusion), fostered by our orchestration, is superior to decision-level consensus (viz., late-fusion) in preserving representation compactness, quantified via the DB index.
- We validate the *practical feasibility* of the proposed approach, demonstrating that the additional design complexity is confined to the offline pre-training phase. The resulting model maintains a lightweight inference profile equivalent to traditional baselines, ensuring suitability for real-time deployment in resource-constrained IIoT environments.

The paper is organized as follows. First, we cover the background on multimodal and contrastive learning in Section 2. Then, in Section 3, we outline related work addressing multimodal and contrastive learning for intrusion detection. Section 4 presents the proposed methodology that combines these learning paradigms, whose evaluation is presented in Section 5. Then, the main findings and limitations of our work are discussed in Section 6. Finally, Section 7 summarizes the key findings and discusses future directions.

2. Background

This section provides background information on the multimodal- and contrastive-learning paradigms in Sections 2.1 and 2.2, respectively, outlining the rationale and formalizations behind those paradigms.

2.1. Multimodal learning

The term multimodality refers to the ability to represent (and understand) a concept or object through multiple views or perspectives, which found analogies in the human process of object identification via multiple senses. Thus, to improve the understanding of complex phenomena, looking at network traffic from multiple modalities has shown potential in enhancing the generalization and detection capabilities of traffic classifiers [6] and intrusion detectors [10]. In the context of traffic classification, multimodality enables the inspection of flows from different perspectives, which can be examined individually to obtain a hierarchical representation of the data. Such multimodal datasets consist of data extracted from networking flows at different levels of abstraction. The goal of multimodal learning is to capitalize on such data in a complementary manner toward learning a complex classification task [11].

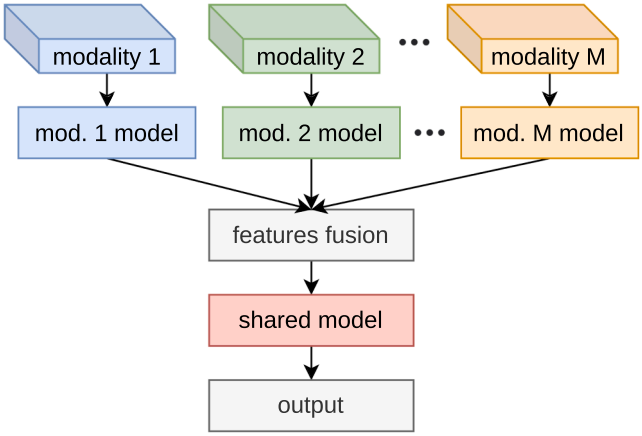


Fig. 1. Intermediate-fusion multimodal learning. Each modality is fed to its own model, then the fused features are passed to the shared model to produce an output.

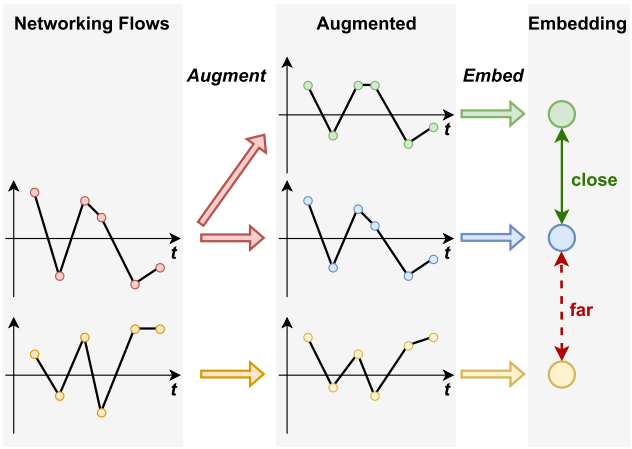


Fig. 2. Self-supervised contrastive learning. Sketch of self-supervised contrastive learning applied to time-series data like networking flows, where each point is a packet size (negative for downstream, positive for upstream) arrived at time t. Each flow is embedded into a latent space where contrastive loss separates similar and dissimilar flows.

Let $D_1, D_2, \ldots, D_M$ represent data from M different modalities or views that describe a given phenomenon. The objective of multimodal learning is to learn a function:

$$f(D_1, D_2, \ldots, D_M) \rightarrow \mathcal{Y}$$

which optimally combines datasets from all modalities to predict the output $\hat{y}$, thus minimizing a suitable loss function.

To effectively combine multiple modalities, two main choices should be made: the *data fusion strategy* and the *multimodal training procedure*. In particular, we can distinguish between *early, intermediate, and late fusions*, the first operating on data, the second on extracted features, and the latter on decisions [11]. In Fig. 1 the intermediate fusion or feature-level fusion is sketched, with the feature spaces of the per-modality models first fused and then fed to a shared-modality model. In addition, an intermediate or late fusion multimodal model can be trained in two ways, namely, in a *monolithic or multi-step training*. The monolithic train involves training all the single-modality models and the shared-modality one in a single pass. In contrast, the multi-phase training involves pre-training single-modality models standalone and then fine-tuning the shared model to perform data fusion.

On the one hand, the early-fusion strategy results in a single DL architecture that is suboptimal, since the model can easily overfit a single modality while underfitting on others. On the other hand, the late-fusion strategy results in a mixture of models, each dedicated to a specific modality: cross-modality interactions are not extracted by a single, shared model. Accordingly, in this paper, we leverage an *intermediate-fusion multimodal model trained in a multi-step fashion* that overcomes the limitations of other fusion strategies.

2.2. Contrastive learning

Contrastive Learning (CL) involves how DL models extract features from the (traffic) input, which is the process of translating samples into a latent space. The objective of CL is to enhance the latent space representation of the input data, thus falling under the representation learning strategies, obtaining a feature extractor that often exhibits improved robustness compared to purely supervised learning procedures. The CL process is similar to how humans learn by comparing similar and dissimilar objects. In particular, CL involves comparing samples in the latent space to bring similar samples closer and dissimilar samples farther away [8,12]. More formally, a reference sample, called *anchor*, is brought closer to similar samples, named *positive*, and distanced from dissimilar samples, referred to as *negative*.

In detail, there are three families of CL approaches that define the concepts of similarity and dissimilarity among samples: (*i*) *unsupervised*, (*ii*) *self-supervised*, and (*iii*) *supervised* [13]. The unsupervised approach exploits similarities and dissimilarities in the input space by looking at their spatial, temporal, or ordinal characteristics in the observations. Instead, both self-supervised and supervised approaches are based on transformations of the inputs, such as rotation or clipping for images. On the one hand, the self-supervised approach considers positive samples as those obtained via transformations from the anchor, while negative samples are transformations of other samples. On the other hand, the supervised approach exploits sample labels, with similar samples being those transformed that belong to the same class, and dissimilar ones being those transformed that belong to other classes. In Fig. 2 the behavior of the self-supervised contrastive loss function is represented graphically. The objective is to separate positive samples from negative ones to create clearly distinguishable clusters.

3. Related works

In this section, we present separately the works that tackle Multimodal Learning and Contrastive Learning for Network Intrusion Detection. The intertwining of such learning paradigms is practically unexplored in the literature, except for a recent work described in the positioning subsection.

3.1. Multimodal learning works

In Table 1, we present a comparative overview of multimodal intrusion detection approaches, categorizing them by *fusion type*, the number of *training phases*, and the number and type of *modalities*. Moreover, we detailed the leveraged DL model, the datasets used, and the intrusion detection task that was tackled. Notably, all the reviewed works propose a two-phase training strategy, employing either early ([14]), intermediate ([10,15,16]), or late ([14,17]) fusion approaches. The number of modalities leveraged varies from 2 ([15,16]) to 3 ([10,17]) or multiple ([14]). These modalities span diverse traffic characteristics, including time series of packet properties or header fields ([10,15–17]), sequences of TCP flags ([17]), and payload-related information, ranging from volumetric statistics ([17]) to raw payload bytes ([15,16]).

Regarding the intrusion detection task tackled, all the three main tasks are covered, namely anomaly detection (AD) [10,16,17]—i.e., outlier or novelty detection—, binary ([10,15,17]) and multiclass ([10,14,16]) Misuse Detection (bMD and mMD, respectively)—where the supervised objective is to distinguish between benign (viz., legitimate)

Table 1
Summary of multimodal intrusion detection research works.

Ref	Fusion Type			# Phases	DL Model	Modalities	*M*	Datasets	Task
	Early	*Inter.*	*Late*						
[10]	○	•	○	2	MDAE + LSTM	NF	3	NSL-KDD UNSW-NB15 CIC-IDS2017	AD bMD mMD
[17]	○	○	•	2	LSTM	NF (Flow, Header, Content)	3	IoT malware dataset	AD bMD
[15]	○	•	○	2	Transformer	PAY PS	2	real-world encrypted traffic	bMD
[14]	•	○	•	2	Fully-connected networks	NF	multiple	UNSW-NB15 NSL-KDD	mMD
[16]	○	•	○	2	MIMETIC (1D-CNN + GRU)	NET PSQ	2	IoT-23	AD mMD

Modalities Fusion Type: Data-level (**Early**) Feature-level (**Intermediate**), Decision-level (**Late**), Present (•), Absent (○). *Number of training phases* (**#Phases**). *Modalities*: Network flow features (**NF**), Transport Payload content (**PAY**), Sequence of Packet Sizes (**PS**), Network Payload content (**NET**), Sequence of header fields (**PSQ**); Flow-related features (e.g., flow duration, statistics on inter-arrival time and packet size), Header-related features (e.g., header and packet length, flags' count, number of transmitted/received packets), Content-related features (e.g., statistics on Segments' and Bytes' transmitted/received). *Number of Modalities* (**M**). *Intrusion Task*: Anomaly Detection (**AD**), binary Misuse Detection (**bMD**), multiclass Misuse Detection (**mMD**).

Table 2
Summary of contrastive learning approaches in intrusion detection.

Ref	Con. Loss			DL Model	Datasets	Task
	U	SS	S			
[13]	○	○	•	Prototype-label embeddings with max-margin loss	NSL-KDD UNSW-NB15	mMD
[22]	•	○	○	SEC + DE-GAT (Graph Neural Network)	Self-collected	AD
[18]	○	•	○	Co-attention Session Encoder with Triplet Loss	2× encrypted traffic	bMD
[21]	○	○	•	Dual-branch deep structure with contrastive cross-entropy loss	NSL-KDD UNSW-NB15	AD, mMD
[19]	○	•	○	Federated Contrastive Learning with ANN	NSL-KDD	bMD
[20]	○	•	○	Self-supervised contrastive learning with encoder	CIC-IDS2017 UNSW-NB15	AD

Contrastive Loss: Unsupervised (**U**), Self-Supervised (**SS**), Supervised (**S**), Present (•), Absent (○). *Intrusion Task*: multiclass Misuse Detection (**mMD**), Anomaly Detection (**AD**), binary Misuse Detection (**bMD**).

and malicious (viz., attack) traffic or its category (mMD). To accomplish these tasks, multiple DL architectures were leveraged, ranging from simple fully-connected networks ([14]), passing through convolutional ([16]) and recurrent ([10,16,17]) neural networks, up to multimodal deep autoencoders ([10]) and Transformers ([15]). The datasets used where primarily publicly available, such as NSL-KDD ([10,14]), UNSW-NB15 ([10,14]), CIC-IDS2017 ([10]), and IoT-23 ([16]), with 2 works ([15,17]) leveraging real-world traffic data.

3.2. Contrastive learning works

In Table 2 we summarized several proposals of CL solutions for intrusion detection. We detailed the nature of the contrastive loss used and also reported how this has been exploited for the DL model train. In addition, we highlight the intrusion detection task and the leveraged data. Regarding contrastive loss, the majority proposed self-supervised solutions ([18–20]), two works focused on supervised contrastive ([13,21]), and [22] exploits unsupervised contrastive. The tackled intrusion detection tasks span across anomaly detection ([20–22]), binary misuse detection ([18,19]), and also multiclass misuse detection ([13,21]). All the works (but [18,22]) leveraged publicly available datasets, such as NSL-KDD ([13,19,21]), UNSW-NB15 ([13,20,21]), and CIC-IDS2017 ([20]).

3.3. Positioning

Compared to the works reported in Tables 1 and 2, we propose integrating multimodal and contrastive learning specifically to address the structural fragility of NIDS under significant domain shifts. To this end, we evaluate our proposal in *multi-target cross-dataset scenarios* leveraging *three* recent heterogeneous network traffic datasets: Edge-IIoT, IoT-NID, and CIC-IoMT. Specifically, we utilize Edge-IIoT as the source dataset for pre-training, and IoT-NID and CIC-IoMT as separate target datasets for adaptation. While the intersection of these paradigms has been recently touched upon by Sun et al. [9], our work is fundamentally distinct in its objectives and execution. Specifically, while Sun et al. [9] focuses on improving label efficiency in constrained settings (few-shot learning) via inter-modal pre-training, our framework prioritizes *structural stabilization* through a specific *multi-step training procedure* designed to maintain representation robustness under distribution shift. Furthermore, compared to existing approaches that often rely on monolithic or simpler two-phase training, we define a decoupled strategy that separates modality-specific representation learning from cross-modal fusion and classification task optimization. This allows us to quantify and demonstrate that the *structural quality of the latent space* (as measured by the DB index) is a meaningful and consistent indicator of cross-dataset generalization. By evaluating our proposal against strong transfer learning baselines on heterogeneous targets, we provide a systematic assessment of how this orchestration preserves representation integrity. Finally, we examine the *intrusion detection task from multiple viewpoints*, including anomaly detection (AD) and multiclass (mMD) misuse detection.

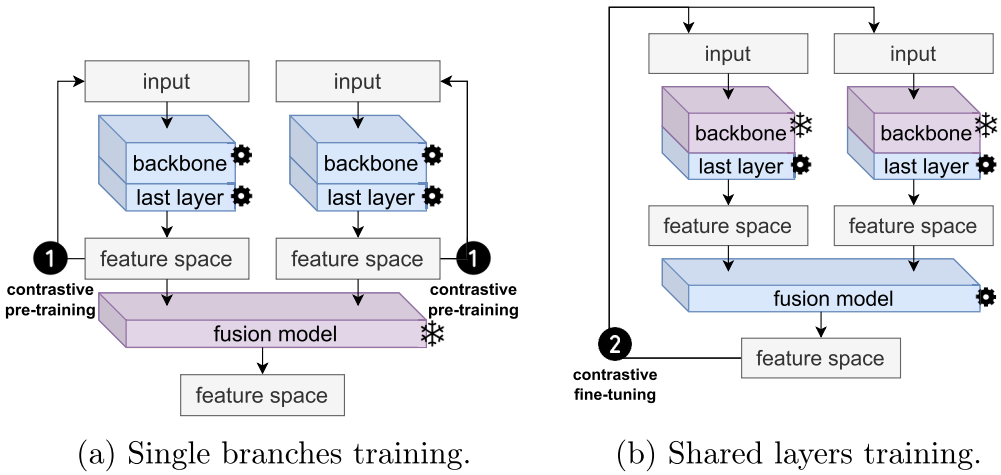
4. Methodology

In this section, we introduce and formalize our proposed solution based on multimodal and contrastive learning paradigms. Then, we also introduce the multimodal transfer learning baseline.

4.1. Multimodal contrastive learning

Framework rationale and design objectives. The proposed framework is strategically designed to exploit the synergy between multimodal representation learning and contrastive objectives, directly addressing the limitations of traditional discriminative models in handling significant domain shifts.

On the one hand, we leverage contrastive learning to enforce a structured embedding space that promotes intra-class compactness and inter-class separability. Unlike traditional cross-entropy, which typically focuses on minimizing classification error and often overfits domain-specific artifacts, our approach facilitates the extraction of transferable semantic features that remain stable across the evaluated network environments. This is critical in cross-dataset evaluation, where varying traffic collections can introduce dataset-specific noise that may negatively affect the detection of target traffic patterns. Specifically, by fostering a highly structured latent organization, our approach



(a) Single branches training. (b) Shared layers training.

Fig. 3. Multimodal contrastive training. Multi-step strategy with (a) single-branch contrastive pre-training and (b) shared-layer contrastive fine-tuning. In each phase, trainable (⚙) and frozen (❄) layers are highlighted in blue and violet, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

effectively widens the decision margins between representation clusters, thereby reducing the likelihood that samples from unseen traffic patterns are misclassified into neighboring regions. Consequently, the evaluation of the Davies-Bouldin (DB) index (see Section 5.1.5) serves as the primary geometric proxy for quantifying this structural organization. A lower DB value directly captures the trade-off between intra-cluster compactness and inter-cluster separation, providing a measurable indicator of the potential resilience of a model to representation collapse under significant dataset shifts.

On the other hand, the multimodal architecture leverages cross-modal consistency to filter domain-specific noise. By aligning diverse traffic views (viz., modalities), the contrastive objective prioritizes mutual information shared across modalities while discarding noise exclusive to a single view or network conditions (e.g., timing jitter). This produces distilled representations grounded in stable semantic features, significantly enhancing resilience to dataset-specific perturbations within the considered scenarios. This multi-perspective representation is supported by recent advances in the design of traffic classifiers robust to changing network conditions [23,24], underscoring the need to learn representations invariant to the intrinsic variability of encrypted traffic and demonstrating their effectiveness in mitigating the impact of network-specific perturbations.

Crucially, our multi-step strategy decouples representation learning from task-specific optimization, thereby preventing gradient interference. In monolithic training, the gradients from the classification head often dominate the process, forcing the model to prioritize decision boundaries over-fitted on the source domain [7,8]. In contrast, our approach enforces a structural baseline through contrastive pre-training before introducing the classification task. This serves as structural regularization, stabilizing the backbone weights by maximizing similarity among positive samples while pushing apart representations of negative ones. Furthermore, the separate pre-training of per-modality branches prevents inter-modal interference during the initial representation learning. By subsequently freezing the early layers during shared-layer training, we preserve the semantic integrity of each modality, ensuring that the fusion model stabilizes cross-modal features without distorting the underlying per-modality expertise. This ensures that the subsequent adaptation operates on a well-organized feature space that shows empirical robustness to cross-dataset variations rather than attempting to learn both representations and boundaries from scratch in the target domain.

Model parameterization and multi-step training orchestration. Formally, a generic DL classifier can be represented by a function $f_\theta : \mathcal{X} \to \mathcal{C}$ that transforms input data into output labels. In our multimodal framework, we decompose the global parameter set θ into three functionally distinct subsets enabling multi-step training: (i) $\theta_b = \{\theta_b^{(1)}, \dots, \theta_b^{(M)}\}$, representing the M modality-specific branches (viz., backbones) responsible for extracting per-modality features; (ii) θ_s, the parameters of the shared fusion model, which learns cross-modal interactions from the concatenated branch outputs; and (iii) θ_h, the parameters of the classification head, mapping the shared representation into the latent space to the output space $\mathcal{Y}$. Hence, the complete model can be expressed as $F(x;\theta)$, where $\theta = \{\theta_b, \theta_s, \theta_h\}$.

The designed multimodal DL architecture comprises $M = 2$ *modalities* on biflows, which we used as model input (viz., *Traffic Object*). In particular, we adopt an intermediate-fusion approach, with $M = 2$ single-modal branches, each fed with the transformed input of its modality, and a shared fusion model that combines the per-modality extracted features.

To mitigate feature fragility, the framework is optimized through a sequence of three training stages (see Fig. 3):

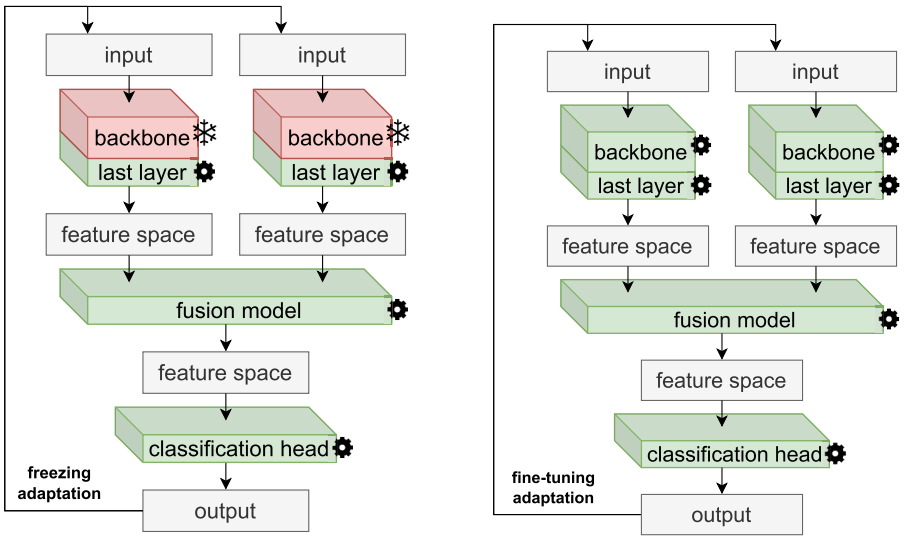
1. *Branch Pre-Training* (Fig. 3(a)), in which each modality branch $\theta_b^{(m)}$ is independently optimized on a source dataset $\mathcal{D}_{src}$ by minimizing the contrastive loss $\mathcal{L}_{CL}$ leveraging only its modality-specific input ($\mathcal{D}_{src}^{(m)}$):

$$\theta_b^{(m)*} = \arg\min_{\theta_b^{(m)}} \mathcal{L}_{CL}(\mathcal{D}_{src}^{(m)}; \theta_b^{(m)}) \tag{1}$$

2. *Fusion Model Training* (Fig. 3(b)), in which per-modality backbones are frozen and the fusion model θ_s is optimized on $\mathcal{D}_{src}$ using $\mathcal{L}_{CL}$. During this stage, the deeper backbone weights of each branch θ_b^* are frozen to preserve per-modality semantic integrity, while the final feature extraction layer of each branch is kept trainable to refine cross-modal alignment.

$$\theta_s^* = \arg\min_{\theta_s} \mathcal{L}_{CL}(\mathcal{D}_{src}; \theta_s | \theta_b^*) \tag{2}$$

   This stage acts as a structural bottleneck, preventing task-specific gradients from distorting the stabilized backbone weights and enforcing the model to extract deeper cross-modal features that are invariant to domain-specific noise.
3. *Downstream Adaptation* (Fig. 4), in which the classification head θ_h is trained on the target dataset $\mathcal{D}_{tgt}$ by minimizing the cross-entropy



(a) Freezing adaptation (FZ). (b) Fine-tuning adaptation (FT).

Fig. 4. Multimodal supervised adaptation. Supervised adaptation strategy with (a) freezing procedure and (b) fine-tuning procedure. For each approach, trainable (⚙) and frozen (❄) layers are highlighted in green and red, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

loss $\mathcal{L}_{CE}$. The pre-trained weights (θ_b^*, θ_s^*) act as a structural prior that can be either frozen (*freezing procedure*) or fine-tuned (*fine-tuning procedure*). Specifically, the freezing procedure (Fig. 4(a)) involves blocking (viz., "freezing") all single-modality layers except the final feature-extraction layers, which are updated alongside the fusion model and the classification head. Conversely, the fine-tuning procedure (Fig. 4(b)) involves a global update of all network weights, encompassing both the single-modal branches and the shared-modality components.

Contrastive objectives and data augmentation. The optimization of the pre-training stages relies on the definition of appropriate contrastive objectives and the generation of diverse traffic views through domain-specific augmentation.

To implement the sample separation process, we leverage two contrastive loss functions $\mathcal{L}_{CL}$ that have shown promising results for network traffic representation [25]. Specifically, we utilize the Normalized Temperature-Scaled Cross Entropy Loss (NT-Xent, shown in Eq. (3))—used in the SimCLR [7]—and the Supervised Contrastive Loss (SupCon, shown in Eq. (4))—which extends the InfoNCE for supervised settings [8].

$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f(x_i), f(x_i^+))/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp(\text{sim}(f(x_i), f(x_k^-))/\tau)} \tag{3}$$

$$\mathcal{L}_{\text{SupCon}} = \frac{1}{N}\sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\text{sim}(f(x_i), f(x_p^+))/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(f(x_i), f(x_k^-))/\tau)} \tag{4}$$

In these objectives, x_* represents the anchor, x_*^+ the positive sample, and x_*^- the negative samples, while $\text{sim}(\cdot,\cdot)$ denotes a similarity metric (e.g., cosine distance). Lastly, $f(\cdot)$ represents a traffic transformation function that generates alternative (viz.,"augmented") views of the input data. These losses compel the model to minimize the distance between the latent projection of positive pairs while maximizing the gap between dissimilar samples, thereby enforcing the structural regularization discussed in Section 4.1. The adoption of both self-supervised (viz., SimCLR) and supervised (viz., SupCon) contrastive losses leads to four multimodal contrastive learning variants, namely $\text{SimCLR}_{\text{FZ}}$, $\text{SimCLR}_{\text{FT}}$, $\text{SupCon}_{\text{FZ}}$, and $\text{SupCon}_{\text{FT}}$, where the subscript indicates the adaptation type, namely freezing (FZ) or fine-tuning (FT).

To generate the augmented views (x_*, x_*^+) required by the contrastive objectives, we employ transformations $(f(\cdot))$ tailored to network traffic [26]. Specifically, we utilize two transformations: *jittering* and *shuffling*. Notably, both transformations are applied independently per modality, ensuring that each branch of the multimodal architecture learns to extract robust features from its specific view (viz., modality) on traffic. These transformations introduce perturbations in two dimensions of each bidirectional flow (viz., *biflow*),[5] namely, in the value of attributes and their order (cf. Section 5.1.4), respectively.

Specifically, *jittering* introduces stochastic noise $(\rho \sim \mathcal{N}(0, \sigma))$ to all biflow attributes. In the networking domain, this simulates temporal and volumetric variability in real-world links, such as Inter-Arrival Times (IAT) fluctuations or subtle packet-size variations caused by different protocol implementations and transient network congestion. This transformation forces the model to ignore environment-specific measurement noise and focus on semantic-invariant traffic behaviors. Conversely, *shuffling* modifies the order of packets within a biflow. This is particularly relevant for networking data as it addresses the structural variability of traffic flows: by perturbing the packet sequence, the model is compelled to identify *functional correlations* (e.g., the relationship between early-handshake packets and subsequent data transfers) rather than overfitting to rigid arrival orders that may be altered by packet loss, retransmissions, or multi-path routing.

Notably, the sensitivity analysis reported in Appendix A confirms that the structural improvements are primarily driven by the contrastive objective rather than by the specific augmentation configuration.

[5] A biflow comprises network packets sharing the same 5-tuple: < `src_ip`, `dst_ip`, `src_port`, `dst_port`, `L4_proto` >, with interchangeable `src` and `dst`.

Table 3
Attack classes of public IoT datasets leveraged in this work. The benign class is present in all the datasets and has been omitted from the table. The leveraged devices are categorized on domain—smart home (🏠), industrial (🏭), and medical (🩺)—and on their nature— *physical testbed* (🖥) or *emulated* (🤖). Attacks with fewer than 20 samples are reported in *italic* and are not considered in experiments.

Dataset	Years	Attack Category					Leveraged Devices		
		(D)DoS	*Info Gathering*	*MITM*	*Injection*	*Malware*	*Domain*	*Nature*	*Description*
IoT-NID	2019	DoS SYN ACK Flood HTTP Flood UDP Flood	Host Discovery OS/Vers. Scan Port Scan	ARP Spoof	–	Mirai Telnet Bruteforce	🏠	🖥	2 Smart Home Some General Purpose
Edge-IIoT	2021–22	DDoS HTTP DDoS ICMP DDoS TCP SYN DDoS UDP	OS Fingerprint Port Scan Vulns Scan	MITM	SQL Inject Upload XSS	Backdoor Password Ransomware	🏭	🖥🤖	13 Sensors
CIC-IoMT	2024	(D)DoS ICMP (D)DoS TCP (D)DoS TCP SYN (D)DoS UDP MQTT (D)DoS Connect MQTT (D)DoS Publish MQTT Malformed Data	OS Fingerprint Port Scan Vulns Scan *Ping Sweep*	ARP Spoof	–	–	🩺	🖥🤖	15 Health-care 7 Smart Home 17 Bluetooth Low-Energy

4.2. Multimodal transfer learning

To assess the effectiveness of the devised methodology, which involves multimodal and contrastive learning paradigms, we compare it with transfer learning-based training variants. Transfer Learning (TL) enables solving a target task using a model previously trained to solve a (similar) task, named the source task. In other words, a model trained on the traffic collected on a source network is adapted to classify the traffic of a target network, involving or not the same set of classes [27].

In this work, we compared our proposal with two multimodal TL solutions, drawing on the cases defined for contrastive training. Accordingly, the two baselines are the combination of multi-step supervised training procedure with fine-tuning-based and freezing-based adaptation phases. We will refer to these two variants as $\texttt{TL}_{\texttt{FZ}}$ and $\texttt{TL}_{\texttt{FT}}$, where the approach is Transfer Learning (TL) and the subscripts indicate the adaptation type, namely freezing (FZ) or fine-tuning (FT).

5. Results

In this section, we describe the experimental setup, indicating all the information required to reproduce the experiments—fostering reproducibility of our proposal—including the leveraged datasets, the preprocessing operations, the evaluation metrics, the hardware resources, the model composition, and training hyperparameters. Then, we report on the results obtained, first in terms of detection and classification effectiveness—looking at anomaly detection and binary- and multiclass-misuse detection. In addition, the generalizability of the proposal and baselines is evaluated. Finally, each approach is evaluated based on its empirical temporal complexity.

5.1. Experimental setup

This section describes the experimental setup adopted to evaluate the proposed framework, providing all the information required to reproduce the experiments and foster reproducibility. Specifically, we describe the leveraged source and target datasets (Section 5.1.1), the detection model employed (Section 5.1.2), the training hyperparameters and hardware configuration (Section 5.1.3), the data preprocessing pipeline (Section 5.1.4), and the evaluation metrics used to assess both classification performance and representation quality (Section 5.1.5).

5.1.1. Datasets

In this work, we evaluated our proposal leveraging three datasets containing attack network traffic, namely Edge-IIoT [28], IoT-NID [29], and CIC-IoMT [30]. In detail, Edge-IIoT is leveraged as a source dataset (D_{src}) for the (contrastive) training phases (viz., pre-training and fine-tuning), and IoT-NID and CIC-IoMT are used separately as a *target dataset* (D_{tgt}) for the adaptation phase. The final goal is to classify the attacks contained in the IoT-NID and CIC-IoMT datasets. Their composition in terms of the number of biflows, alongside class encodings, is reported in Fig. 5(a) and 5(b), respectively.

All datasets are detailed in Table 3, where we taxonomize them according to the collection period, the attack classes included (grouped into categories defined in [28]), the IoT domain, the nature of the collection setup (i.e., physical testbed or emulated), and the specific types of devices employed.

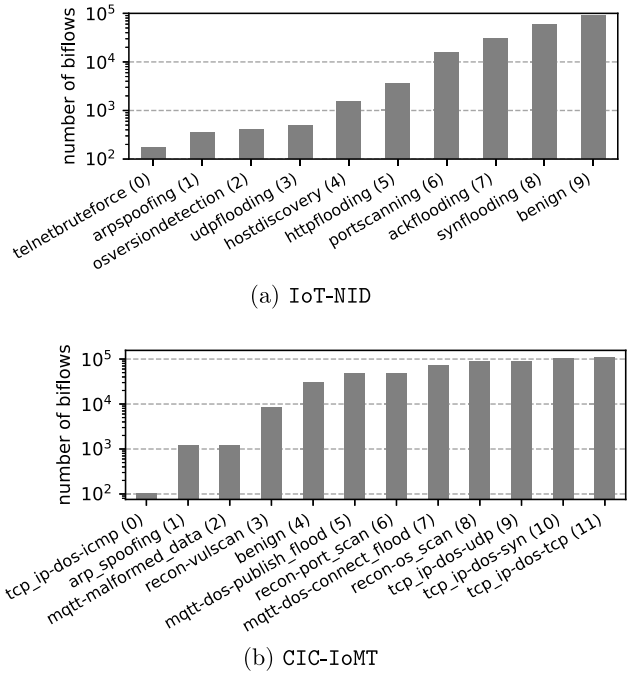


Fig. 5. Biflows distribution for (a) IoT-NID and (b) CIC-IoMT. Classes are arranged by increasing biflow count. Values on the y-axis are in *Log* scale. Labels in parentheses encode class names for subsequent analyses.

In detail, we divided attacks into 5 different categories that are:

(*i*) (Distributed) Denial of Service—(D)DoS—including attacks aiming at causing service disruptions;
(*ii*) Info Gathering, covering attacks focused on obtaining information from the victim;
(*iii*) Man-in-the-Middle—MITM—addressing attacks involving the interception of communication;
(*iv*) Injection, targeting the exploitation of vulnerabilities in applications or systems to introduce malicious data or code;
(*v*) Malware, encompassing generic attacks that do not fit into the other defined categories.

5.1.2. Detection models

The proposal of a novel detection model goes beyond the scope of this work. Instead, we relied on the MIMETIC framework [6] and derived a multimodal embedding function comprising $M = 2$ parallel CNN-based branches (i.e., one for each modality). The first modality is fed with the NET input and comprises two 1D-convolutional layers—each followed by a 1D max-pooling layer—and a final dense layer. Conversely, the second branch is fed with the PSQ input and consists of two 2D-convolutional layers, interleaved with 2D max-pooling and batch normalization, and a dense layer. The intermediate features extracted by each modality are first concatenated and then elaborated via a *shared dense layer* before the model head, which consists of a softmax-activated dense layer. We have selected these two complementary modalities (viz., NET and PSQ) to reflect a realistic deployment scenario in which payload availability may be partial, and feature extraction complexity must remain bounded. It is worth noting that the selected features are payload-length-aware but not payload-content-dependent for the PSQ modality, while NET relies on raw bytes that may be unavailable under strict deep packet inspection restrictions. In such encrypted traffic scenarios, the framework can fall back to the PSQ modality alone, potentially introducing a degradation of classification performance. Nonetheless, while currently limited to these inputs, the leveraged architecture is inherently modular and can be extended to incorporate additional traffic modalities. In fact, the work in [31] has effectively extended the MIMETIC architecture by integrating *Contextual Inputs* that characterize the aggregate traffic context of the target biflow (e.g., in terms of volume and packet counts). Furthermore, the multimodal paradigm discussed in [32] highlights the feasibility of integrating diverse traffic representations, such as *Flow Statistics* (e.g., maximum, minimum, mean, and standard deviation of packet length), to address the challenges posed by increasingly encrypted environments. Furthermore, the contrastive paradigm ensures modality balance by prioritizing shared mutual information, preventing any single view from dominating the representation as modality coverage increases. Despite this variety of choices, we opted for NET and PSQ because they are the most popular in the literature, are native to the MIMETIC framework, and require significantly lower computational overhead for real-time feature extraction than higher-level statistics or contextual information. By focusing on these two fundamental modalities, *we aim to isolate the effects of the contrastive strategy on the structure of the learned features in the latent space, ensuring that any observed enhancement is attributed to the learning methodology rather than to increased architectural complexity or heterogeneous data enrichment.*

5.1.3. Training hyperparameters and hardware setup

The models were evaluated over 10 independent runs using a randomized hold-out strategy. For each run, the dataset was partitioned into 70% for training and 30% for testing, with 10% of the training data used for validation. Each training phase—i.e., (*i*) contrastive/supervised pre-training, (*ii*) contrastive/supervised fine-tuning, and (*iii*) adaptation—has been performed over 50 epochs using a batch size of 64 and an early-stopping of 20 patience on validation loss. We have employed AdamW as an optimizer, with an initial learning rate of 10^{-3} and "Cosine Annealing" rate scheduler ($T_{\max} = 50$ and $lr_{\min} = 10^{-5}$). The same hyperparameter configuration is applied uniformly across all compared strategies. The selected values follow established best practices for contrastive and transfer learning on network traffic data [6,25,26], and were validated empirically on the source dataset validation split prior to the main evaluation.

All the analyses—including model training and inference—have been executed on a machine with 2 Intel(R) Xeon(R) Gold 6252N CPU @ 2.30 GHz and 378 GB of memory. Inference feasibility (Section 5.2.5) was assessed on an emulated Raspberry Pi 4 Model B (RPi4B) environment, based on a QEMU Virtual Machine (VM) configured to approximate the hardware characteristics of the physical device: a 64-bit ARMv8-A architecture with four Cortex-A72 cores, 4 GB of RAM, and Ubuntu 22.04.5 LTS. The virtualized environment is configured as a single NUMA node, with each vCPU pinned to a dedicated host core to reduce scheduling variability and improve resource determinism.

5.1.4. Data preprocessing

We processed PCAP traces and segmented traffic into biflows, where for each, we extracted the NET and PSQ inputs corresponding to the two modalities. Specifically, NET is obtained by concatenating the network/IP-layer header and payload. Following prior work [16], we limit NET to the first $N_b = 576$ bytes. Additionally, IP addresses and ports are intentionally obfuscated to prevent unintended bias. Conversely, PSQ comprises a sequence of six header fields extracted from the first $N_p = 10$ packets of the biflow. These fields include transport-level payload-length (PL), packet direction[6] (DIR), inter-arrival time (IAT), TCP-window size (WIN),[7] Time-to-Live (TTL), and TCP Flags (FLG).[8] We applied zero-padding to handle biflows containing fewer than N_b bytes or N_p packets. Moreover, to ensure a fair comparison across datasets, the maximum PL is uniformly limited to 1500 bytes (MTU) for all datasets, and a 60-second timeout is applied to the IAT [33]. Finally, we applied a per-feature *Min-Max* scaling across both modalities to bound numerical values to $[0, 1]$ and facilitate model convergence.

5.1.5. Evaluation metrics

To assess the effectiveness of our proposal, we leverage two sets of evaluation metrics, divided according to their evaluation objective: (*i*) assessing detection effectiveness and (*ii*) evaluating the learned representation.

The *detection effectiveness* metrics that we focus on are: the macro-averaged Area Under Curve (AUC) in its partial formulation, namely the partial AUC or pAUC, for the binary Misuse Detection (bMD) task, and the Recall, the Precision, and F1-score (F1) averaged macro, for the multiclass Misuse Detection (mMD) task. In detail, the pAUC measures the (partial) area under the Receiver Operating Characteristic (ROC) curve; partial means that the area under the curve is computed for a fixed False Positive Rate (FPR) interval (see Eq. (5)), which we fixed in $[0 - 1\%]$. The macro-averaged version of the pAUC is obtained by considering, in turn, each attack class versus the benign class. Regarding the metrics used for the mMD task, the Recall is the number of correctly classified samples of a class divided by the total number of the samples that belong to that class, the Precision is the number of correctly classified samples of a class divided by the number of samples that has been classified as that class, and the macro F1 is the average of harmonic means of per-class precisions and recalls (see Eq. (6)). All these metrics, namely pAUC, Recall, Precision, and F1, reads as: the higher the better.

$$\text{macro-pAUC} = \frac{1}{|C|} \sum_{i \in C} \int_{\alpha}^{\beta} \text{TPR}_i(\text{FPR})\, d\text{FPR} \tag{5}$$

$$\text{macro-F1} = \frac{1}{|C|} \sum_{i \in C} \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{6}$$

[6] Represented as 0 for downstream and 1 upstream.
[7] WIN is set to 0 for UDP packets.
[8] It refers to the decimal representation of the 8 bits composing the TCP flags

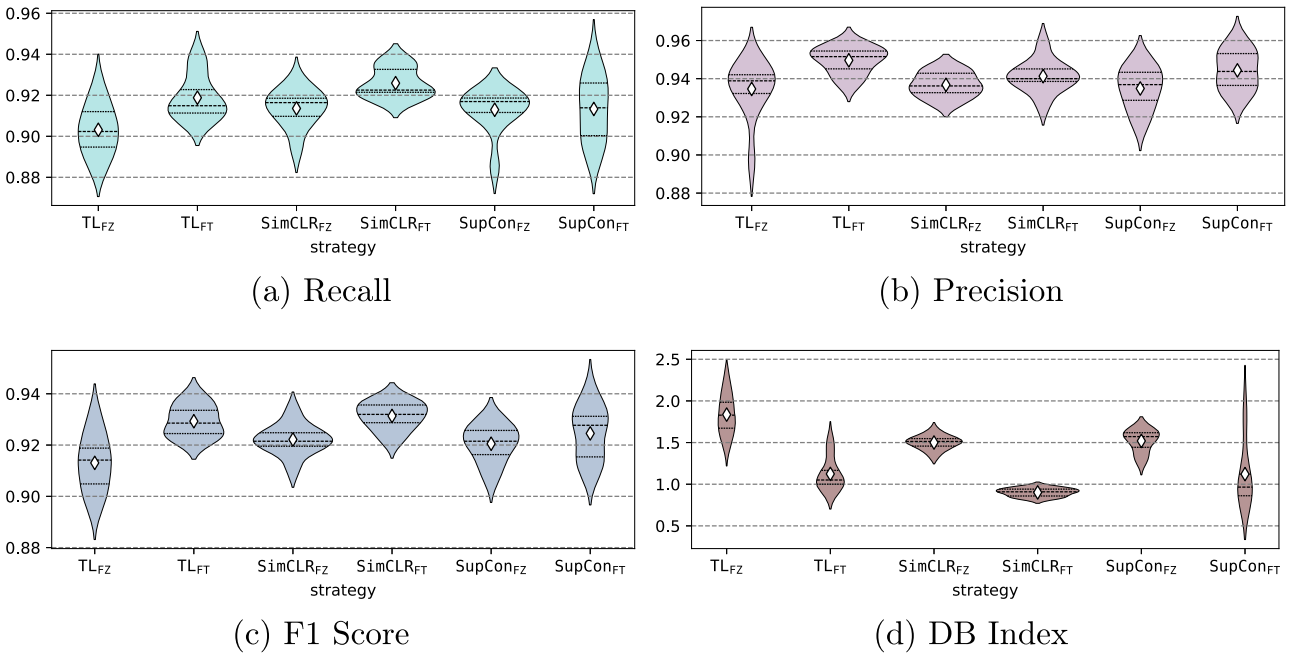


Fig. 6. Comparison with baselines. Comparison of the proposed solutions against Transfer-Learning baselines (TL_{FT} and TL_{FZ}) in terms of (a) Recall, (b) Precision, (c) F1 Score, and (d) Davies-Bouldin (DB) index. Results are averaged over 10 runs.

Regarding the assessment of the *learned representations*, we analyze the structure of the multimodal backbone latent space using the macro-averaged Davies-Bouldin (DB) index. The DB index provides an indication of class separability in the latent space by quantifying the ratio between intra-class compactness and inter-class separation (see Eq. (7)). Importantly, the DB index is not used as a direct measure of model generalization performance, but rather as an indicator of the quality and structure of the representations learned by the adapted model. Since the primary goal of this work is to assess the effectiveness of contrastive learning in yielding models with improved performance after adaptation, the DB index is computed on the adapted model, i.e., after the adaptation phase on the target dataset. Lower DB values indicate more compact and well-separated class clusters in the latent space.

$$\text{macro-DB} = \frac{1}{|C|}\sum_{i\in C}\max_{j\neq i}\frac{\sigma_i+\sigma_j}{d_{i,j}}, \tag{7}$$

where σ_i is the intra-cluster distance of a cluster i and $d_{i,j}$ is the inter-cluster distance between clusters i and j. In the following (cf. Section 5.2), these metrics are referred (for brevity) without the macro- prefix.

To assess the statistical significance of performance differences across strategies, we conduct a pairwise comparison using the Conover post-hoc test [34], applied after a Friedman test [35] to account for the non-parametric, multi-measurement nature of our evaluation. For each macro-averaged metric—i.e., *recall*, *precision*, *F1 score*, *DB index*, and *pAUC*—, we compute the average rank of each strategy across all train-test splits and report the results through Critical Difference (CD) diagrams [36], which visually group strategies that are not statistically different from one another ($p > .05$). A summary radar plot of average rankings across all metrics provides an overall view of each strategy's relative performance profile. Detailed per-metric CD diagrams are provided in Appendix B.

5.2. Experimental evaluation

This section comprehensively evaluates the multimodal contrastive learning solution for NIDS. Our experimental evaluation assesses the effectiveness in terms of both (binary and multiclass) classification performance and feature extraction (see Section 5.2.1). Then, in Section 5.2.2 we provide an ablation study that contrasts multimodal contrastive learning with its singlemodal counterparts, showcasing the importance of the interplay between modalities. Building on the outcome of these analyses, in Section 5.2.3 we delve deeper into the classification performance at a finer granularity, i.e., looking at confusion matrices and showing the detailed detection behavior for a specific attack class (viz., `ARP spoofing`). Then, we present a breakdown of performance in terms of the number of packets composing the biflows (Section 5.2.4). Finally, we measure the resource footprint of multimodal contrastive learning compared to baselines (Section 5.2.5).

5.2.1. Comparison with baselines

In this section, we provide an overview of performance figures of the proposed multimodal contrastive learning solutions, in the self-supervised (SimCLR) and supervised (SupCon) variants, and compare them with those of the classical multimodal transfer learning. We recall that all techniques require a two-phase training, that is, training on the source dataset ($\mathcal{D}_{src}$), i.e., `Edge-IIoT`, and the adaptation to the target dataset ($\mathcal{D}_{tgt}$), i.e., `IoT-NID`. TL baselines involve supervised training on ($\mathcal{D}_{src}$), whereas contrastive-learning-based solutions perform contrastive training on the same $\mathcal{D}_{src}$.[9] After learning involving the source dataset, all the solutions undergo adaptation, either by freezing the single-modality branches and fine-tuning the other layers (i.e., `FZ`), or by fine-tuning the entire multimodal model (i.e., `FT`). This pipeline results in network intrusion detectors capable of performing multiclass misuse detection, whose performance is analyzed for multiclass and binary detection, that is, benign vs. malicious. Binary detection is obtained by looking at the not-being-benign probability—the sum of probabilities associated with attack classes.

The results are summarized in Fig. 6. where we evaluate three metrics for the multiclass classification task, namely recall (Fig. 6(a)), precision(Fig. 6(b)), and their harmonic mean, i.e., the F1 score (Fig. 6(c)). In addition, we also provide an evaluation of the backbone via the DB index computed on the adapted model (Fig. 6(d)). As a general comment,

[9] The sensitivity analysis for the σ parameter of the jittering transformation has been reported in Appendix A.

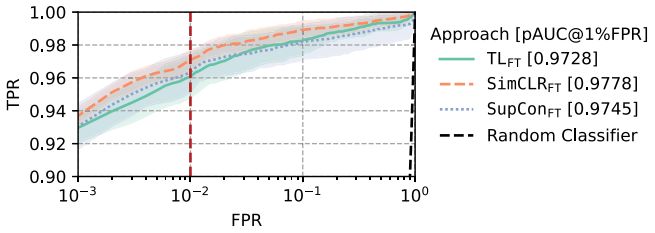


Fig. 7. ROC curves. ROC curves of all approaches, separating malicious classes from benign. Partial-AUC at 1% FPR is reported in brackets. Vertical red dashed lines indicate 1% FPR. Results shown as $mean \pm std.dev.$ over 10 runs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

all the approaches reach a satisfactory average (multiclass) classification performance, with a recall higher than 0.90, a precision higher than 0.93, and an F1 higher than 0.91. Regarding the structure of the learned feature space, only FT adaptation-based solutions obtain well-separated clusters. As a first outcome, we found that all freezing adaptation procedures underperform compared to the respective fine-tuning solutions. Another insight is that, among contrastive variants, the self-supervised procedure achieves performance figures slightly higher than those of the supervised one (e.g., +1.26% recall for fine-tuning adaptation). This outcome can be explained by the misalignment between the source and the target dataset, which makes the self-supervised solution less bounded to the source dataset when compared to the supervised training. In addition, we found that the best contrastive approach, i.e., $\texttt{SimCLR}_{\texttt{FT}}$, slightly improves the best baseline, i.e., $\texttt{TL}_{\texttt{FT}}$, by obtaining a higher recall that is balanced by a slightly lower precision.

In terms of classification performance, $SimCLR_{FT}$ achieves results comparable to the best baseline (TL_{FT}) in global metrics. However, while the $F1$ gain appears modest ($< 1\%$), it represents a consistent trend within a saturated metric regime, where the primary advantage lies in the superior stability of the latent organization. This is highlighted by the reduction in the DB index (0.90 vs. 1.12), indicating a more robust class separation that serves as a geometric proxy for higher reliability under significant domain network shifts.

Regarding binary classification performance, we show in Fig. 7 the macro-averaged ROC curves along with the respective macro-pAUC for the best strategies, those based on the fine-tuning adaptation. Notably, all solutions achieve a commendable pAUC higher than 0.97. In detail, the best performance figures are shown by the $\texttt{SimCLR}_{\texttt{FT}}$, which slightly improves the best baseline (i.e., $\texttt{TL}_{\texttt{FT}}$) by +1% TPR at 1% FPR.

To provide a holistic view of the relative performance of the strategies across all metrics, Fig. 8(a) reports the radar plot of average rankings, while Fig. 8(b) shows the Critical Difference diagram computed on the average rank across metrics. Overall, $\texttt{SimCLR}_{\texttt{FT}}$ achieves the best average ranking across most metrics, with the only exception of precision, where $\texttt{TL}_{\texttt{FT}}$ ranks first. The CD diagram confirms that fine-tuning-based approaches (FT) are statistically distinguishable from freezing-based ones (FZ), regardless of the pretraining approach ($p \leq 0.05$), while no significant difference emerges within the two groups. As shown in Appendix B, this separation is most pronounced for the DB index, where $\texttt{SimCLR}_{\texttt{FT}}$ ranks first with no overlap with any other approach.

5.2.2. Impact of multimodal fusion

In this section, we conduct an ablation study isolating the contribution of each traffic modality within the contrastive learning pipeline. It is worth noting that our goal is not to benchmark multi-modal and single-modal learning but rather to verify how the contrastive strategy behaves when limited to specific traffic views. Accordingly, in Table 4, we report the performance of the four contrastive strategies alongside the TL baselines when IoT-NID is used as the target dataset. Specifically, we compare results obtained by training on each modality independently (i.e., NET and PSQ) against the full multimodal configuration (NET + PSQ). This comparison allows us to quantify the information gain from the joint traffic representation and verify whether the contrastive objectives retain their regularization properties even in single-modal scenarios.

In general, NET modality consistently outperforms PSQ across all strategies and metrics, suggesting that raw bytes carry more discriminative information for intrusion detection. Nonetheless, multimodal fusion achieves the best performance in the vast majority of cases, confirming that the two modalities provide complementary information that the model can effectively exploit. The benefit of fusion is particularly evident for the DB index, where the multimodal model consistently yields the most compact and well-separated feature space across all strategies, regardless of the adaptation procedure. This suggests that intertwining the two modalities leads to richer and more structured representations, further corroborating the effectiveness of the proposed multimodal contrastive learning approach.

Notably, the trends observed in the multimodal setup are also mirrored in the single-modal one. Specifically, contrastive-based models maintain a marginal yet consistent advantage over TL in both classification accuracy and latent-space regularization, even when considering a single traffic view. Furthermore, FT adaptation continues to outperform FZ, regardless of the underlying training strategy (i.e., Contrastive or TL-based). This underscores the need for task-specific weight adjustments to fully leverage the pre-trained feature space, confirming that global optimization of θ is essential for maximizing performance under significant domain shift.

Finally, we evaluate the impact of different fusion strategies as described in Section 2.1. As a result, in Table 5 we compare our proposed intermediate fusion strategy (NET + PSQ) against a late-fusion baseline (NET | PSQ) across both contrastive-based (viz., $\texttt{SimCLR}_{\texttt{FT}}$ and $\texttt{SupCon}_{\texttt{FT}}$) and TL (viz., $\texttt{TL}_{\texttt{FT}}$) approaches using FT adaptation.

In particular, in the late-fusion configuration, each pre-trained branch is extended with a dedicated classification head and independently optimized on the target dataset. The final classification outcome is then derived via an ensemble aggregation by averaging the soft-outputs produced by the branches. This comparison allows us to assess whether the structural regularization induced by contrastive learning is better preserved by joint feature-level fusion (NET + PSQ) or by decision-level consensus (NET | PSQ).

As reported, both fusion strategies achieve comparable performance across all classification metrics (i.e., Recall, Precision, F1-score, and pAUC@1%FPR). This suggests that, under stable conditions, they can achieve similar detection accuracy. However, a distinct trend emerges by examining the geometric properties of the learned latent space. In particular, the intermediate fusion strategy (NET + PSQ) consistently yields a significantly lower DB index than late fusion (NET | PSQ) across both contrastive- and transfer-learning based approaches. This result indicates that although both fusion strategies achieve similar accuracy under stable conditions, intermediate fusion proves superior for cross-modal feature integration. In fact, the late-fusion configuration operates as a shallow consensus among independent branches, whereas our intermediate fusion compels the model to learn synergistic correlations between NET and PSQ modalities. Such structural superiority (e.g., 0.90 vs. 1.68 of DB index for $\texttt{SimCLR}_{\texttt{FT}}$) serves as a key indicator of model robustness, suggesting that the model has learned a deeper, more cohesive representation from multimodal traffic rather than relying on decision-level ensemble aggregation.

5.2.3. Fine-grained analysis

To gain deeper insights into the decision process of the detectors, in this section, we analyze performance figures at a finer granularity. The results are shown in Fig. 9, where the confusion matrices for the best baseline and the best multimodal contrastive approaches are shown, namely, we compare decisions by $\texttt{TL}_{\texttt{FT}}$ and $\texttt{SimCLR}_{\texttt{FT}}$. Notably, looking at the main diagonal of such matrices, that is, the per-class recalls, the contrastive solution improves over minority classes, namely classes

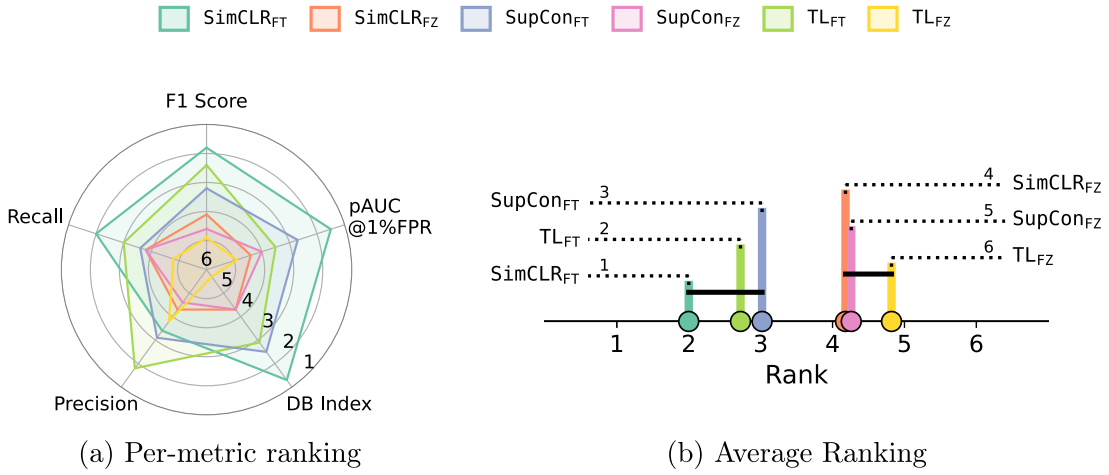


Fig. 8. Strategy ranking comparison. (a) Radar chart showing the average Friedman rank per metric across 10 runs for each strategy; outer positions indicate better rank. (b) Critical Difference (CD) diagram based on the average rank across all metrics (Recall, Precision, F1 Score, Davies-Bouldin index, and pAUC@1%FPR); strategies connected by a horizontal bar are not significantly different (Conover post-hoc, $\alpha = 0.05$).

Table 4
Singlemodal vs. multimodal learning results. Classification performance of both contrastive (SimCLR$_{FT}$, SimCLR$_{FZ}$, SupCon$_{FT}$, SupCon$_{FZ}$) and transfer learning (TL$_{FT}$ and TL$_{FZ}$) strategies across singlemodal (NET, PSQ) and multimodal (NET + PSQ) models on the IoT-NID dataset. Metrics are averaged over 10 runs (mean ± std). For each strategy and metric, the best modality is highlighted in **bold**. The DB Index measures latent class separability (lower is better).

Approach	Modality	Recall	Precision	F1 Score	DB Index	pAUC@1%FPR
SimCLR$_{FT}$	NET	0.9226 ±0.0077	0.9382 ±0.0063	0.9278 ±0.0055	1.1239 ±0.0506	0.9772 ±0.0023
	PSQ	0.9047 ±0.0126	0.9317 ±0.0064	0.9138 ±0.0088	1.7856 ±0.0580	0.9702 ±0.0038
	NET + PSQ	**0.9259** ±0.0071	**0.9413** ±0.0083	**0.9314** ±0.0054	**0.9015** ±0.0476	**0.9778** ±0.0026
SimCLR$_{FZ}$	NET	0.9088 ±0.0087	0.9337 ±0.0073	0.9179 ±0.0069	1.9251 ±0.0974	0.9680 ±0.0053
	PSQ	0.8731 ±0.0227	0.9063 ±0.0186	0.8834 ±0.0207	2.6700 ±0.3579	0.9528 ±0.0112
	NET + PSQ	**0.9136** ±0.0092	**0.9368** ±0.0063	**0.9222** ±0.0058	**1.5005** ±0.0809	**0.9700** ±0.0043
SupCon$_{FT}$	NET	**0.9219** ±0.0073	0.9402 ±0.0087	**0.9272** ±0.0045	1.1321 ±0.0565	**0.9765** ±0.0022
	PSQ	0.9032 ±0.0148	0.9327 ±0.0120	0.9137 ±0.0112	1.6695 ±0.1007	0.9691 ±0.0051
	NET + PSQ	0.9133 ±0.0155	**0.9444** ±0.0103	0.9245 ±0.0103	**1.1213** ±0.3905	0.9745 ±0.0063
SupCon$_{FZ}$	NET	0.9058 ±0.0065	0.9314 ±0.0059	0.9148 ±0.0051	1.9422 ±0.0556	**0.9710** ±0.0027
	PSQ	0.8794 ±0.0149	0.9112 ±0.0290	0.8882 ±0.0219	2.6707 ±0.1401	0.9543 ±0.0049
	NET + PSQ	**0.9129** ±0.0105	**0.9349** ±0.0109	**0.9205** ±0.0073	**1.5196** ±0.1323	0.9708 ±0.0057
TL$_{FT}$	NET	**0.9210** ±0.0050	0.9349 ±0.0073	0.9258 ±0.0053	1.1432 ±0.0474	**0.9746** ±0.0027
	PSQ	0.8939 ±0.0302	0.9209 ±0.0222	0.8989 ±0.0410	1.7789 ±0.1223	0.9675 ±0.0096
	NET + PSQ	0.9188 ±0.0104	**0.9496** ±0.0070	**0.9293** ±0.0061	**1.1237** ±0.1934	0.9728 ±0.0057
TL$_{FZ}$	NET	**0.9046** ±0.0057	0.9243 ±0.0053	0.9122 ±0.0045	1.9728 ±0.1396	**0.9669** ±0.0070
	PSQ	0.8666 ±0.0176	0.8952 ±0.0118	0.8740 ±0.0135	2.4845 ±0.2132	0.9482 ±0.0089
	NET + PSQ	0.9032 ±0.0122	**0.9346** ±0.0147	**0.9131** ±0.0110	**1.8358** ±0.2339	0.9658 ±0.0065

Table 5
Intermediate- vs late-fusion multimodal learning results. Classification performance of both contrastive (SimCLR$_{FT}$ and SupCon$_{FT}$) and transfer learning (TL$_{FT}$) strategies across intermediate- and late-fusion multimodal (resp., NET + PSQ and NET | PSQ) models on the IoT-NID dataset. Metrics are averaged over 10 runs (mean ± std). For each strategy and metric, the best fusion strategy is highlighted in **bold**. The DB Index measures latent class separability (lower is better).

Approach	Modality	Recall	Precision	F1 Score	DB Index	pAUC@1%FPR
SimCLR$_{FT}$	NET | PSQ	0.9151 ±0.0084	**0.9627** ±0.0048	0.9300 ±0.0055	1.6827 ±0.0498	**0.9796** ±0.0028
	NET + PSQ	**0.9259** ±0.0071	0.9413 ±0.0083	**0.9314** ±0.0054	**0.9015** ±0.0476	0.9778 ±0.0026
SupCon$_{FT}$	NET | PSQ	**0.9154** ±0.0094	**0.9658** ±0.0054	**0.9310** ±0.0059	1.5663 ±0.0888	**0.9790** ±0.0032
	NET + PSQ	0.9133 ±0.0155	0.9444 ±0.0103	0.9245 ±0.0103	**1.1213** ±0.3905	0.9745 ±0.0063
TL$_{FT}$	NET | PSQ	0.9168 ±0.0041	**0.9617** ±0.0049	**0.9315** ±0.0041	1.6292 ±0.1224	**0.9784** ±0.0031
	NET + PSQ	**0.9188** ±0.0104	0.9496 ±0.0070	0.9293 ±0.0061	**1.1237** ±0.1934	0.9728 ±0.0057

N.B.: For the late-fusion configuration (NET | PSQ), the DB Index is computed on the concatenation of the per-modality latent spaces, as no shared representation exists.

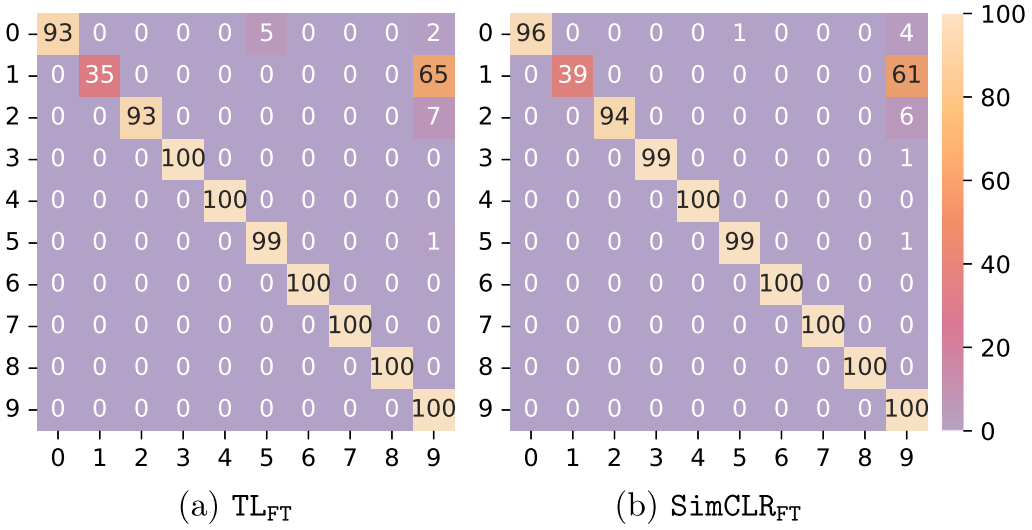


Fig. 9. Confusion matrices. Confusion matrices for the best baseline (TL_{FT}, a) and best contrastive solution ($SimCLR_{FT}$, b). Classes are sorted by increasing sample count, using the same encoding as Fig. 5(a). Matrices are averaged over 10 runs.

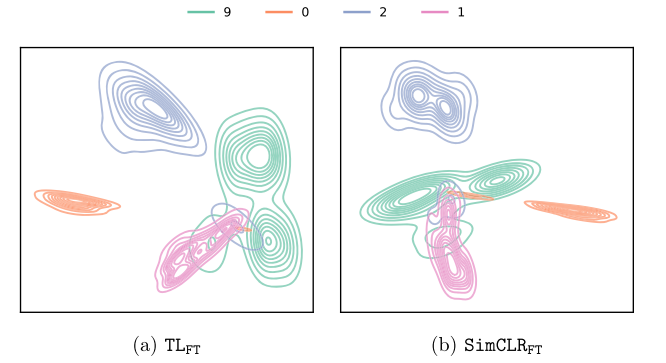


Fig. 10. t-SNE projections. t-SNE projections of the learned feature space for the best baseline (TL_{FT}, a) and best contrastive solution ($SimCLR_{FT}$, b), restricted to the four most confused classes—indices 9 (i.e., benign), 0, 2, and 1—sorted by decreasing recall score. A single representative run (seed = 1) is shown.

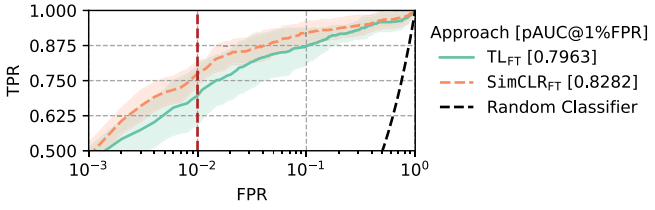


Fig. 11. Class-specific ROC. ROC curves for FT-based approaches on ARP Spoofing versus benign traffic. Partial-AUC at 1% FPR in brackets, vertical red dashed lines show 1% FPR. Results are reported as *mean* ± *std.dev.* over 10 runs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

0 (viz., `Telnet bruteforce`, ≈ +3%) and 1 (viz., `ARP spoofing`, ≈+4%). While in the first case, namely for `Telnet bruteforce`, we can notice an increase of ≈+2% in malicious biflows classified as `benign` (index 9), for the `ARP spoofing` attack $SimCLR_{FT}$ reduces the percentage of false negatives by ≈−4%.

To further investigate these findings, Fig. 10 presents the t-SNE projections of the feature spaces learned by TL_{FT} and $SimCLR_{FT}$. The analysis focuses on the four most confused classes: `Telnet bruteforce` (0), `ARP spoofing` (1), `OSversion detection` (2), and `Benign` (9). A visual inspection reveals that while TL_{FT} exhibits a high-density region where samples from all categories converge, $SimCLR_{FT}$ effectively deconstructs this entanglement. In fact, with $SimCLR_{FT}$ clusters appear more compact, particularly for classes 0 and 2. In addition, $SimCLR_{FT}$ reduces the overlap between the `Benign` traffic (class 9) and class 2, with the majority of class 2 samples displaced away from the former, which is also the most represented class. Although some overlap between classes 0 and 9 persists, in $SimCLR_{FT}$ it is confined to low-density peripheral areas rather than their respective semantic cores. This increased inter-class distance and geometric class separation confirm that the proposed multi-step strategy fosters a more robust representation space, directly improving the detection of minority attack clusters.

Finally, in Fig. 11, we provide an overview of the detection performance for the `ARP spoofing` class, showing the ROC curves and the pAUC of the TL_{FT} and $SimCLR_{FT}$ approaches. This analysis aims to gain deeper insight into the improvements provided by $SimCLR_{FT}$ over TL_{FT} in a binary misuse detection scenario. Notably, although no detector can obtain effective multiclass classification performance—with the higher recall that is ≈ 39% by $SimCLR_{FT}$ (see Fig. 9)—in a binary setup, the $SimCLR_{FT}$ obtains (at 1% FPR) up to 77.55% TPR and a pAUC of 0.83. Notably, this contrastive solution based on $SimCLR_{FT}$ significantly improves the TL_{FT} baseline of +7.93% TPR and +0.0319 pAUC at 1% FPR.

5.2.4. *Sensitivity to biflow-length*

Herein, we assess the impact of the biflow length (hereafter denoted as BF_{len})—i.e., the number of packets composing it—on classification performance. Hence, Fig. 12 shows the F1-score as the biflow length increases for TL_{FT}, $SimCLR_{FT}$, and $SupCon_{FT}$ where, for each length $L \in \{1, N_p\}$, we compute the F1-score by aggregating all biflows $\{BF^i \in D_{tgt} : BF^i_{len} \leq L\}$.

The F1-score increases as the biflow-length increases, suggesting that larger biflows provide richer information, leading to improved model performance. This trend is particularly evident when classifying biflows with up to 5 packets, where we observe an improvement of up to +12.2% for TL_{FT} compared with the performance of single-packet biflows.

Notably, biflows comprising only 1 packet represent most of the dataset (i.e., 82.67%), with an F1-score in the range 75%–79%. The performance obtained is relatively high, demonstrating the robustness of the model in scenarios characterized by limited availability of useful information. In contrast, we note a plateau starting from ≈ 9 packets, suggesting that adding more packets does not lead to significant improvements. This indicates that the most relevant information falls within the first 9 packets, while further packets contribute less discrim-

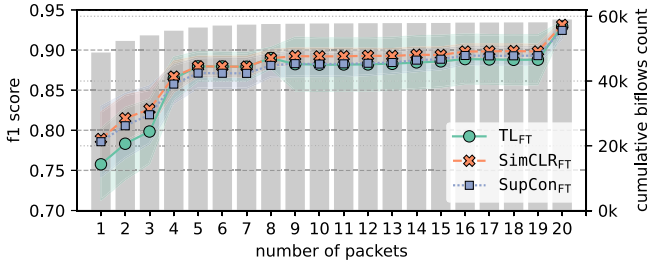


Fig. 12. F1 score vs. biflow length. F1 Score as the maximum biflow length (number of packets) increases. Values shown as $mean \pm std.dev.$ over 10 runs.

Table 6
Computational complexity of the evaluated approaches. Training and inference costs for the transfer learning baseline (TL_{FT}, TL_{FZ}) and all contrastive strategies ($SimCLR_{FT}$, $SimCLR_{FZ}$, $SupCon_{FT}$, $SupCon_{FZ}$), across singlemodal (NET, PSQ) and multimodal (NET + PSQ) configurations. Training time (**tTime**) is reported per epoch; the Pre-Train column shows t_{NET}/t_{PSQ} with only the slower modality contributing to Total. Inference time (**iTime**) is reported per sample. **FP** and **TP** denote FLOPs and trainable parameters, respectively; **DS** is the maximum model size on disk across all training phases. All values are averaged over 30 runs (mean ± std) of 10 epochs each.

Approach	tTime [min/epoch]			
	Pre-Train	Fine-Tune	Adaptation	Total
TL_{FT}	1.35 ±0.29 / 1.19 ±0.22	1.46 ±0.21	1.98 ±0.40	4.80 ±0.54
TL_{FZ}			1.48 ±0.19	4.30 ±0.40
$SimCLR_{FT}$	2.43 ±0.37 / 2.22 ±0.16	2.45 ±0.22	2.26 ±0.37	7.15 ±0.56
$SimCLR_{FZ}$			1.70 ±0.13	6.58 ±0.45
$SupCon_{FT}$	2.49 ±0.43 / 2.23 ±0.16	2.48 ±0.20	2.36 ±0.35	7.33 ±0.59
$SupCon_{FZ}$			1.71 ±0.12	6.68 ±0.48

Table 7
Inference-time resource consumption on the emulated RPi 4B. iTime is reported per sample. **FP** and **TP** denote FLOPs and trainable parameters, respectively; **DS** is the model size on disk after PTQ (when enforced). All values are averaged over 3 runs (mean ± std).

Approach	PTQ	iTime [ms/sample]	FP [M]	TP [K]	DS [MB]
TL_{FT}	–	51.53 ±26.78	7.26	780.6	2.99
TL_{FT}	int8	59.62 ±21.02	7.26	780.6	0.85
TL_{FT}	int8wo	85.08 ±54.84	7.26	780.6	0.85
$SimCLR_{FT}$	–	52.30 ±24.80	7.26	780.6	2.99
$SimCLR_{FT}$	int8	60.16 ±21.24	7.26	780.6	0.85
$SimCLR_{FT}$	int8wo	87.10 ±52.35	7.26	780.6	0.85

iTime: inference time per sample. **FP**: FLOPs per sample. **TP**: trainable parameters. **DS**: model size on disk after PTQ (if enforced).

inative information. Notably, including biflows with at least 20 packets leads to an improvement of ≈+4%.

On the other hand, comparing the different approaches, we note that $SimCLR_{FT}$ consistently outperforms the other methods across all biflow lengths. Moreover, contrastive learning-based approaches (viz., $SimCLR_{FT}$ and $SupCon_{FT}$) are more effective than TL_{FT} when having only a limited number of packets (≤3), achieving a consistent gap ranging from +2.78% to +3.15%. However, as the biflow length increases (>3), the performance gap narrows, and all models perform similarly. To further investigate the reasons behind the enhanced contrastive effectiveness on short biflows, in Fig. 13 we report the confusion matrices for TL_{FT} and $SimCLR_{FT}$ considering only biflows with 3 or fewer packets. This analysis shows that the higher performance enhancement attained by the contrastive solution involves the least populous class, i.e., Telnet bruteforce (0), with a gain of +30% in recall. These findings suggest that contrastive learning approaches are particularly beneficial when only a limited number of packets are available.

Table 8
Per-biflow average time-to-feature on emulated RPi4B. Tra t_{out}: biflow timeout. t_{extr}: feature extraction. t_{disp}: dispatcher overhead. t_{TTF}: end-to-end preprocessing wall-clock (excludes model inference). Results are shown as mean ± std over 10 runs.

t_{out} [ms]	t_{disp} [ms]	t_{extr} [ms]	t_{TTF} [ms]
10	10.026 ± 0.003	0.266 ± 0.022	10.292 ± 0.022
50	49.995 ± 0.004	0.270 ± 0.022	50.266 ± 0.020
150	149.376 ± 0.023	0.273 ± 0.022	149.649 ± 0.025
300	297.834 ± 0.080	0.272 ± 0.021	298.106 ± 0.078
600	594.330 ± 0.184	0.273 ± 0.021	594.603 ± 0.180

5.2.5. *Computational overhead and edge deployment feasibility*

In this section, we assess the computational overhead of the proposed contrastive approaches compared to the transfer learning baselines, in terms of training time, inference latency, model complexity, and memory footprint. The objective is twofold: (*i*) verifying whether the training overhead remains confined to the offline phase without affecting inference complexity and (*ii*) assessing the end-to-end feasibility—from the raw traffic to the inference—when the pipeline is deployed on an edge-grade device. In doing this, we have trained the end-to-end intrusion detection pipeline on a server-grade machine and have deployed it, for inference, on a resource-constrained emulated RPi4B device (see Section 5.1.3 for hardware details), applying Post-Training Quantization (PTQ) to assess model compression. Results are reported in Table 6 for the training phase, and in Tables 8 and 7, reporting the Time-To-Feature (TTF)—the time required to obtain the input for the intrusion detection model as the raw traffic arrives—and the inference time and model disk footprint, respectively.

Regarding training time, contrastive-based approaches exhibit a higher per-epoch cost compared to the TL baselines, with a Pre-Train phase approximately 1.8× longer (e.g., 2.43/2.49 min/epoch for $SimCLR_{FT}$/$SupCon_{FT}$ vs. 1.35 min/epoch for TL_{FT}). This overhead is intrinsic to the contrastive pre-training objective, which necessitates processing multiple augmented views per sample—thereby increasing the effective batch size per iteration—and involves the computational cost of evaluating pairwise similarity metrics (e.g., cosine distance) across all augmented views within each batch—which is partly implementation-dependent.

The Fine-Tune phase follows a similar trend, with contrastive approaches requiring approximately 2.45/2.48 ($SimCLR_{FT}$/$SupCon_{FT}$) min/epoch against 1.46 min/epoch for TL. On the contrary, the Adaptation phase is comparable across all approaches, ranging from 1.48 to 2.36 min/epoch, with fine-tuning adaptation (FT) consistently requiring slightly more time than freezing (FZ), due to the larger number of trainable parameters. Overall, the total training time of contrastive approaches is approximately 1.5× that of TL (e.g., 7.15/7.33 vs. 4.80 min/epoch), a cost that is amortized by the improved representation quality and minority-class detection performance reported in Sections 5.2.1, 5.2.3, and 5.2.4.

Regarding inference, all approaches share the same backbone architecture and thus exhibit comparable per-sample latency. As reported in Table 7, when deployed on the emulated RPi4B without quantization, inference time is approximately 51–52ms/sample for both TL_{FT} and $SimCLR_{FT}$, with large standard deviations attributable to emulation-induced variability rather than structural differences between strategies. FLOPs per sample (7.26M), trainable parameters (780.6K), and model size on disk (2.99MB) are identical across all strategies, since all approaches ultimately converge to the same architecture for inference. Applying PTQ reduces the on-disk footprint from 2.99MB to 0.85MB (≈ 3.5× compression) for both int8 and int8wo schemes; however, this compression does not translate into a latency reduction on the target platform—int8 marginally increases inference time to ≈ 60ms/sample, while int8wo raises it further to ≈ 85–87ms/sample, likely due to the

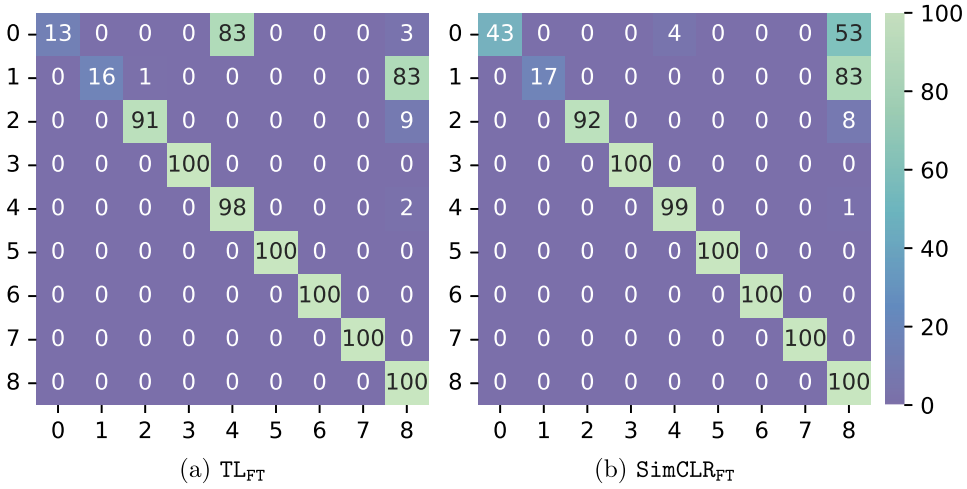


Fig. 13. Short biflows confusion matrices. Confusion matrices on the biflows with less than 4 packets for the best baseline ($\texttt{TL}_{\texttt{FT}}$, a) and best contrastive solution ($\texttt{SimCLR}_{\texttt{FT}}$, b). Classes are sorted by increasing sample count, using the same encoding as Fig. 5(a). Matrices are averaged over 10 runs.

Table 9
Contrastive and transfer learning results on CIC-IoMT. Classification performance on CIC-IoMT for the four contrastive strategies ($\texttt{SimCLR}_{\texttt{FT}}$, $\texttt{SimCLR}_{\texttt{FZ}}$, $\texttt{SupCon}_{\texttt{FT}}$, $\texttt{SupCon}_{\texttt{FZ}}$) and the two transfer-learning counterparts ($\texttt{TL}_{\texttt{FT}}$, $\texttt{TL}_{\texttt{FZ}}$). Metrics are averaged over 10 runs (mean ± std). The best approach is highlighted in **bold**. The DB Index measures feature space compactness; lower is better.

Approach	Recall	Precision	F1 Score	DB Index	pAUC@1%FPR
$\texttt{SimCLR}_{\texttt{FT}}$	0.9607 ±0.0044	0.9696 ±0.0031	0.9647 ±0.0021	**1.1803** ±0.0498	**0.9878** ±0.0017
$\texttt{SimCLR}_{\texttt{FZ}}$	0.9585 ±0.0064	0.9650 ±0.0055	0.9614 ±0.0031	2.1695 ±0.0859	0.9877 ±0.0022
$\texttt{SupCon}_{\texttt{FT}}$	**0.9608** ±0.0063	**0.9708** ±0.0018	**0.9653** ±0.0031	1.2576 ±0.0372	**0.9878** ±0.0025
$\texttt{SupCon}_{\texttt{FZ}}$	0.9590 ±0.0058	0.9649 ±0.0040	0.9617 ±0.0036	2.1870 ±0.1345	0.9870 ±0.0018
$\texttt{TL}_{\texttt{FT}}$	0.9598 ±0.0073	0.9680 ±0.0050	0.9634 ±0.0034	1.3805 ±0.1301	0.9859 ±0.0030
$\texttt{TL}_{\texttt{FZ}}$	0.9440 ±0.0187	0.9547 ±0.0158	0.9471 ±0.0193	2.5489 ±0.1339	0.9783 ±0.0110

lack of hardware-accelerated integer arithmetic on the emulated device. These results confirm that the additional computational cost of contrastive learning is entirely confined to the training phase and does not translate into any inference overhead.

Beyond model inference, Table 8 reports the end-to-end Time-To-Feature (TTF)—the wall-clock time elapsed from the arrival of the first packet of a biflow to the availability of the extracted feature vector. The results show that the feature extraction step itself introduces a negligible and stable overhead of $t_{\text{extr}} \approx 0.27$ms regardless of the configured biflow timeout, while the dispatcher overhead t_{disp} closely tracks t_{out}, confirming that TTF is dominated by the biflow collection window rather than by processing latency. Even at the shortest tested timeout of 10ms, the full preprocessing pipeline completes in ≈10.3ms, well within the latency budget of a near-real-time NIDS. Taken together, these findings confirm that the proposed approach is practically viable for deployment in resource-constrained, edge-grade NIDS environments, with the preprocessing and inference pipelines jointly contributing a bounded and predictable latency overhead.

5.2.6. Generalization to the internet of medical things

To assess the cross-domain generalizability of the proposed framework, we extended our evaluation to the Internet of Medical Things (IoMT) domain. To this end, we employed the CIC-IoMT dataset as a "target" downstream task to verify whether the representation (θ_b^*, θ_s^*) learned during pre-training on Edge-IIoT remains effective when transferred to a domain with significantly different traffic distributions and threat models.

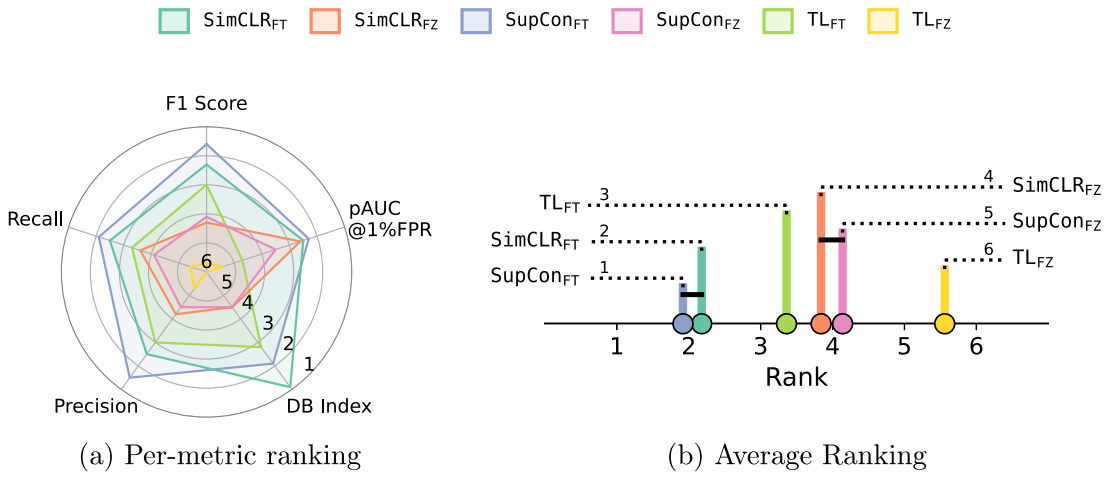
Table 9 reports the results on CIC-IoMT for both the contrastive-based variants (i.e., $\texttt{SimCLR}_{\texttt{FT}}$, $\texttt{SimCLR}_{\texttt{FZ}}$, $\texttt{SupCon}_{\texttt{FT}}$, $\texttt{SupCon}_{\texttt{FZ}}$) and the TL baselines (i.e., $\texttt{TL}_{\texttt{FT}}$ and $\texttt{TL}_{\texttt{FZ}}$). All strategies achieve high classification performance, with F1 Score in the range $[0.947, 0.965]$ and pAUC@1%FPR consistently above 0.978. $\texttt{SupCon}_{\texttt{FT}}$ attains the best discriminative performance (F1 $= 0.9653 \pm 0.0031$, pAUC $= 0.9878 \pm 0.0025$), while $\texttt{SimCLR}_{\texttt{FT}}$ achieves the most compact feature space (DB $= 1.1803 \pm 0.0498$). Notably, the FT projection consistently outperforms its FZ counterpart across all strategies and metrics, a pattern visible in both the per-metric radar chart and the average ranking (Fig. 14(a)–(b)). Transfer learning ($\texttt{TL}_{\texttt{FT}}$, $\texttt{TL}_{\texttt{FZ}}$) trails behind the contrastive strategies in the average ranking, with $\texttt{TL}_{\texttt{FZ}}$ showing the largest performance gap (F1 $= 0.9471 \pm 0.0193$). Notably, while $\texttt{SimCLR}_{\texttt{FT}}$ consistently achieves superior latent space quality on both targets, $\texttt{SupCon}_{\texttt{FT}}$ achieves higher discriminative performance on CIC-IoMT. This outcome, which is a reversal with respect to IoT-NID, may be attributed to a closer semantic alignment between the CIC-IoMT attack taxonomy and the Edge-IIoT source domain, which reduces the label-anchoring penalty of supervised contrastive objectives under this smaller distribution shift.

6. Discussion

This section discusses the key findings of our experimental evaluation (Section 6.1), contextualizes them with respect to the existing literature (Section 6.2), and reflects on the limitations and future directions of the proposed approach—Sections 6.3 and 6.4, respectively.

6.1. Key findings and interpretation

The experimental results demonstrate that multimodal contrastive learning constitutes a viable alternative to classical transfer learning (TL) for cross-dataset adaptation in IoT network intrusion detection. Overall, the proposed approach achieves performance comparable to TL baselines in multiclass misuse detection, while offering specific and consistent advantages in latent space structuring, minority-class detection, and short-biflow classification.



(a) Per-metric ranking (b) Average Ranking

Fig. 14. Strategy ranking comparison on `CIC-IoMT`. (a) Radar chart showing the average Friedman rank per metric across 10 runs for each strategy; outer positions indicate better rank. (b) Critical Difference (CD) diagram based on the average rank across all metrics (Recall, Precision, F1 Score, Davies-Bouldin index, and pAUC@1%FPR); strategies connected by a horizontal bar are not significantly different (Conover post-hoc, $\alpha = 0.05$).

The first key insight concerns the role of the adaptation strategy. Across all evaluated approaches—both contrastive and TL-based—fine-tuning adaptation (`FT`) consistently and significantly outperforms its freezing counterpart (`FZ`), as confirmed by the Conover post-hoc test ($p \leq 0.05$)—see Figs. 8(b) and 14(b). This finding suggests that, unsurprisingly, updating the full model during adaptation is critical for achieving high classification performance under a significant traffic shift, regardless of the pre-training strategy adopted.

A second insight concerns the relative behavior of self-supervised (SimCLR) and supervised (SupCon) contrastive objectives across heterogeneous target datasets (i.e., `IoT-NID` and `CIC-IoMT`). While on the `IoT-NID` dataset `SimCLR`$_{FT}$ achieves slightly higher recall than `SupCon`$_{FT}$ (e.g., +1.26% for fine-tuning adaptation), we observe a reversal in discriminative performance on the `CIC-IoMT` dataset, where `SupCon`$_{FT}$ attains the best results (e.g., 96.5% of F1 Score).

We attribute the relative advantage of the "label-agnostic" objective (viz., SimCLR) on `IoT-NID` to the inherent misalignment between the source (`Edge-IIoT`) and target taxonomies. In such cases, a "label-aware" objective (viz., SupCon) trained on source-specific class boundaries may introduce a bias that partially conflicts with the target label structure, whereas SimCLR learns more agnostic features that generalize more freely under network domain shifts. Conversely, the superior performance of `SupCon`$_{FT}$ on `CIC-IoMT` suggests a closer semantic alignment with the source domain, which mitigates this penalty and allows the supervised contrastive priors to generalize more effectively under a smaller distribution shift. This observation is consistent with prior work on the trade-off between label-aware and label-agnostic representation learning under distribution shift [25].

A third and arguably most significant finding concerns the structure of the learned latent space. We argue that the contrastive objective establishes a robust geometric regularizer that stabilizes representations against domain-specific noise. This leads to the emergence of well-defined representation clusters during adaptation, which proves more resilient to domain shift compared to traditional discriminative training. The DB index reveals that contrastive approaches—especially `SimCLR`$_{FT}$—yield markedly more compact and well-separated class clusters compared to TL baselines across both target datasets. On `IoT-NID`, DB index = 0.90 vs. 1.12 for `SimCLR`$_{FT}$ and `TL`$_{FT}$, respectively; on `CIC-IoMT`, the same ordering is preserved, with DB = 1.18 vs. 1.38, confirming the structural advantage of contrastive pre-training regardless of the target domain. Crucially, the t-SNE analysis (see Fig. 10) provides a direct visual explanation of how this structural advantage translates into downstream performance gains: while `TL`$_{FT}$ exhibits a high-density entanglement region where traffic from different classes converges, `SimCLR`$_{FT}$ effectively deconstructs this overlap, displacing minority-class samples away from the dominant `Benign` cluster. This geometric reorganization of the latent space directly accounts for the improved recall on the minority attack classes—i.e., `ARP spoofing` (+4%) and `Telnet bruteforce` (+3%)—and confirms that latent space quality, as measured by the DB index, is a meaningful and interpretable indicator of cross-dataset generalization performance. To further corroborate this finding, we complement the DB index with the Silhouette score [37] computed on the target domain embeddings (cosine similarity). Unlike the DB index, the Silhouette score evaluates each sample individually against its nearest neighboring cluster, providing a complementary, sample-level view less sensitive to class-size imbalance. Consistently, `SimCLR`$_{FT}$ achieves the best—higher is better— Silhouette score (0.84 ± 0.03), compared to `TL`$_{FT}$ (0.67 ± 0.05) and `SupCon`$_{FT}$ (0.81 ± 0.04), confirming that the self-supervised contrastive objective yields more cohesive and well-separated class clusters after adaptation. Although classification metrics are influenced by the attack taxonomy and semantic alignment of the dataset, the structural advantage of `SimCLR`$_{FT}$ in latent compactness is consistent across all scenarios.

Finally, the improvement on short biflows (≤ 3 packets) further corroborates this interpretation, as evaluated on the `IoT-NID` dataset. When only limited temporal information is available, the decision boundaries learned by TL are more susceptible to misclassification, as the model cannot rely on the accumulation of discriminative packet-level evidence. In contrast, the intra-class compactness enforced by contrastive pre-training reduces the likelihood that sparse representations of minority-class biflows are absorbed into neighboring, more densely populated regions of the latent space—an effect particularly pronounced for the least populous class, i.e., `Telnet bruteforce`, with a gain of +30% recall on biflows with fewer than 4 packets.

Taken together, these findings suggest that the observed gains arise from the *interplay* between the three components of the proposed framework rather than from any single element in isolation. Multimodality provides complementary traffic views that enrich the representational substrate; the contrastive objective enforces geometric structure on that substrate, promoting intra-class compactness and inter-class separability; and the multi-step orchestration ensures that this structure is established *before* task-specific gradients are introduced, preventing the representation collapse that characterizes monolithic training. The ablation study (Section 5.2.2) corroborates this interpretation: the struc-

tural advantage of $\texttt{SimCLR}_{\texttt{FT}}$ over $\texttt{TL}_{\texttt{FT}}$—quantified via the DB index—is consistently replicated in single-modal configurations, confirming that the gains stem from the training methodology rather than from architectural choices.

6.2. Contextualization with respect to the literature

The experimental results allow for a more nuanced reading of how the proposed approach relates to existing work, beyond the methodological distinctions drawn in Section 3.

The DB index advantage of $\texttt{SimCLR}_{\texttt{FT}}$ over $\texttt{TL}_{\texttt{FT}}$ (0.90 vs. 1.12) extends the qualitative claims of Lopez-Martin et al. [13], who report well-separated embedding clusters in single-dataset contrastive NIDS, to a cross-dataset adaptation scenario where such structural advantage is statistically confirmed after domain shift. The relative behavior of SimCLR and SupCon is consistent with findings from [25] in encrypted traffic classification, where label-agnostic contrastive objectives yield more transferable representations than supervised ones under source-target distribution mismatch—an effect attributable to the anchoring of supervised losses to source-specific class boundaries. Regarding minority-class detection, the recall gains on `ARP spoofing` and `Telnet bruteforce` align with the class-imbalance robustness reported in [13], and with the cross-dataset generalization demonstrated by Golchin et al. [20] for self-supervised contrastive NIDS, albeit in an anomaly detection rather than a multiclass setting. Compared to CoMDet ([9]), which targets few-shot label efficiency via inter-modal contrastive pre-training, our evaluation operates in a full-data adaptation regime and includes latent space quality as an assessment criterion, offering a complementary perspective on the trade-offs of multimodal contrastive learning for intrusion detection.

6.3. Limitations

Despite the promising results, the proposed framework is subject to several limitations that should be acknowledged.

Dataset diversity. The current evaluation covers two source-to-target adaptation scenarios, both using `Edge-IIoT` as the source and evaluating on `IoT-NID` and `CIC-IoMT` as independent targets. While the three datasets differ substantially in terms of topology, traffic composition, attack taxonomy, and collection conditions—making the adaptation scenarios genuinely challenging—broader validation across multiple source datasets and additional IoT environments, including real-world traffic traces, would further strengthen generalizability claims. We acknowledge this as a primary direction for future work.

Modality availability and privacy constraints. A significant challenge for multimodal frameworks is their reliance on the concurrent availability of both `NET` and `PSQ` inputs. In real-world NIDS deployments, specific modalities might be inaccessible due to technical failures or, increasingly, privacy-driven constraints. For instance, pervasive end-to-end encryption or strict data protection regulations (e.g., GDPR) may preclude payload inspection, effectively rendering the `NET` modality unavailable. To address this, our framework leverages the `PSQ` branch as a robust, encryption-agnostic alternative that relies exclusively on packet header metadata. As demonstrated in our ablation study (Section 5.2.2), even when relying solely on `PSQ`, the model maintains high detection performance, with only a marginal reduction in performance (e.g., $\leq 1\%$ in F1-Score for $\texttt{SimCLR}_{\texttt{FT}}$) compared to the multimodal setup. This ensures operational continuity in environments where content-level features are restricted, without significantly compromising detection efficacy. Nonetheless, the explicit handling of missing modalities—e.g., through masking or modality dropout during training—remains an interesting direction for future work.

Computational overhead of pre-training. As reported in Section 5.2.5, contrastive approaches incur approximately 1.5× more training overhead than TL baselines. This is primarily due to the processing of augmented pairs required by the contrastive objective, as well as the computational cost of evaluating similarity metrics (e.g., cosine distance, whose cost is partly implementation-dependent) across all pairs of augmented views within each batch. It is worth noting that this overhead is a *one-time*, *offline* cost confined to the pre-training phase and does not affect inference, making it identical to classical TL baselines in terms of latency, FLOPs, and model size. Consequently, the investment in contrastive pre-training is justified by improved structural stability and a +30% increase in recall for short-flow attacks and up to a +4% gain for minority attack classes (e.g., ARP spoofing). Nonetheless, in deployment scenarios where the pre-training budget is severely constrained, this overhead should be carefully weighed against the representation-quality gains it provides.

Real-time deployment feasibility. As demonstrated in Section 5.2.5, the inference pipeline is practically viable on edge-grade hardware, with per-sample latency of ≈ 51 ms and negligible preprocessing overhead. The primary residual constraint concerns the gap between the emulated environment and physical IoT gateways: whether hardware-accelerated integer arithmetic on purpose-built devices would translate PTQ compression into a latency reduction remains an open question, left to future work.

6.4. Future directions

Building on the above, several directions for future work are identified. First, the evaluation should be extended to additional cross-dataset scenarios involving diverse IoT environments and, where available, real-world traffic traces, in order to probe the robustness of the learned representations across a wider range of domain shifts. Second, exploring alternative contrastive loss functions—such as those operating across modalities rather than within each branch independently—could further exploit cross-modal consistency as a supervisory signal. Third, the explicit handling of missing modalities—e.g., through masking or modality dropout during training—would enable graceful degradation in deployments where payload inspection is restricted, extending the operational continuity currently provided by the `PSQ` fallback. Furthermore, while the current framework operates on a fixed two-modality setup, the leveraged multimodal architecture is inherently modular, and the contrastive objective is agnostic to the number of modalities. Hence, extending the framework to incorporate additional traffic views (e.g., flow statistics or contextual counters [31]) is straightforward and may yield further gains in representation richness. Finally, while the inference feasibility of the proposed framework has been assessed on an emulated edge environment, profiling on physical IoT gateways and purpose-built hardware accelerators—particularly in conjunction with post-training quantization—remains an important step toward concrete deployment validation.

7. Conclusions

In this work, we presented a *multimodal contrastive learning* approach to mitigate some of the robustness challenges encountered by NIDSs under cross-dataset adaptation scenarios. While ML- and DL-based NIDSs exhibit promising detection capabilities, their limited ability to generalize across heterogeneous network environments reduces their real-world applicability. To mitigate this limitation, the proposed framework integrates *multimodal learning* for a richer traffic representation and *contrastive learning* to improve the structural organization of the latent space, promoting intra-class compactness and inter-class separability.

The approach was validated through an extensive experimental evaluation, leveraging publicly available datasets `Edge-IIoT`—for source

training—and two independent target datasets, IoT-NID and CIC-IoMT, covering distinct IoT and IoMT domains respectively. Overall, the results suggest that contrastive multimodal training represents a viable alternative to classical transfer learning when dealing with cross-dataset adaptation in IoT intrusion detection, particularly in challenging conditions such as minority attacks and short biflows. While the proposed approach does not aim to replace transfer learning, it provides complementary benefits in terms of representation robustness and decision stability.

On IoT-NID, contrastive learning achieves performance figures comparable to those of transfer learning in multiclass misuse detection, with both approaches exhibiting high macro F1 (e.g., 93.14% for $\texttt{SimCLR}_{\texttt{FT}}$ vs. 92.93% for $\texttt{TL}_{\texttt{FT}}$). However, contrastive training results in a more structured latent space, with a lower Davies-Bouldin (DB) index (0.90 for $\texttt{SimCLR}_{\texttt{FT}}$ vs. 1.12 for $\texttt{TL}_{\texttt{FT}}$), improving class separability of the adapted model backbone. This structural advantage is statistically confirmed by the Friedman and Conover post-hoc tests, which show that $\texttt{SimCLR}_{\texttt{FT}}$ achieves a statistically significant advantage in the latent space quality, ranking first in the DB index with no overlap with any other strategy ($p \leq 0.05$). The ablation study further corroborates the effectiveness of multimodal fusion, showing that the combination of NET and PSQ modalities consistently outperforms either modality in isolation across all contrastive strategies, with the benefit of fusion being most pronounced in terms of latent space compactness. Furthermore, intermediate fusion consistently yields a lower DB index than late-fusion consensus (e.g., 0.90 vs. 1.68 for $\texttt{SimCLR}_{\texttt{FT}}$), confirming that feature-level integration produces more cohesive representations than decision-level ensemble aggregation. Performing fine-grained analyses on IoT-NID, we found that contrastive learning obtains better detection for minority attack classes, namely ARP spoofing and Telnet bruteforce, showing a Recall gain of +4% and +3%, respectively, while also reducing false negatives. Furthermore, contrastive learning is more effective than transfer learning in scenarios with a limited number of packets, achieving improvements of +2.78% to +3.15% for biflows composed of three or fewer packets, with a gain of +30% Recall on Telnet bruteforce in the short-biflow regime. On CIC-IoMT, contrastive approaches consistently outrank transfer learning in average performance across all metrics, with $\texttt{SupCon}_{\texttt{FT}}$ achieving the best discriminative performance (F1 = 96.53%) and $\texttt{SimCLR}_{\texttt{FT}}$ yielding the most compact latent space (DB = 1.18 vs. 1.38 for $\texttt{TL}_{\texttt{FT}}$)—confirming the structural advantage of contrastive pre-training across heterogeneous IoT domains. Finally, the computational analysis confirms that the additional overhead of contrastive pre-training—approximately 1.5× the training time of TL baselines—is confined to the training phase and does not translate into any inference overhead, as all approaches converge to the same architecture at deployment time (780.6 K parameters, 7.26 M FLOPs/sample, 2.99 MB on disk).

Future directions include the exploration of *alternative contrastive loss functions*, possibly enhancing feature extraction and separation, the extension of the framework to *additional traffic modalities* enabled by the modular nature of the multimodal architecture, the explicit *handling of missing modalities* (e.g., via modality dropout) for graceful degradation under payload unavailability, and the validation of the proposed approach on *additional source-target combinations across diverse IoT environments*, to further stress-test generalizability under varying network conditions and distribution shifts.

CRediT authorship contribution statement

Giampaolo Bovenzi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Idio Guarino:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Antonio Pescapè:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”). Additionally, this work was partially carried out within the “xInternet” Project supported by the MUR PRIN 2022 program (D.D.104—02/02/2022) funded by the NextGenerationEU.

Data availability

Data will be made available on request.

Appendix A. Sensitivity analysis for jittering

To assess the impact of transformations’ hyperparameters, we perform a single-transformation contrastive training, in order to isolate the contribution of each transformation. Accordingly, we evaluate all four contrastive strategies, i.e., $\texttt{SupCon}_{\texttt{FT}}$, $\texttt{SupCon}_{\texttt{FZ}}$, $\texttt{SimCLR}_{\texttt{FT}}$, $\texttt{SimCLR}_{\texttt{FZ}}$ by considering jittering and shuffling one at a time. In detail, we perform a sensitivity analysis to the jittering augmentation parameter σ by testing five values $\sigma \in \{0.01, 0.03, 0.05, 0.1, 0.2\}$, spanning an order of magnitude around the default value $\sigma = 0.03$ adopted from [26]. Regarding the shuffling strategy, we tested three values of its amplitude hyperparameter, namely $a \in \{0.5, 0.75, 1.0\}$: it controls the number of segments to be shuffled—e.g., with $a = 1.0$ each biflow is split into $N \sim U(2, 4)$ segments that are then shuffled (see [26] for details). For each configuration, we report the distribution of F1 Score and Davies-Bouldin index over ten independent runs in Figs. A.1 and A.2, respectively for jittering and shuffling transformations.

Regarding the jittering transformation (Fig. A.1), the results show no meaningful variation visually in either metric as σ varies. F1 Score shows no monotonic trend observable as σ increases. Similarly, the Davies-Bouldin index exhibits stable distributions for each strategy. These observations indicate that the learned representations and the downstream classification performance are not sensitive to the precise choice of σ, and that the default value $\sigma = 0.03$ [26] constitutes a safe and well-supported operating point.

Regarding the shuffling transformation (Fig. A.2), a clear trend emerges: the higher the a hyperparameter, the better the median performance figures for both the F1 Score and the Davies-Bouldin index. Accordingly, we selected $a = 1.0$ for our experiments.

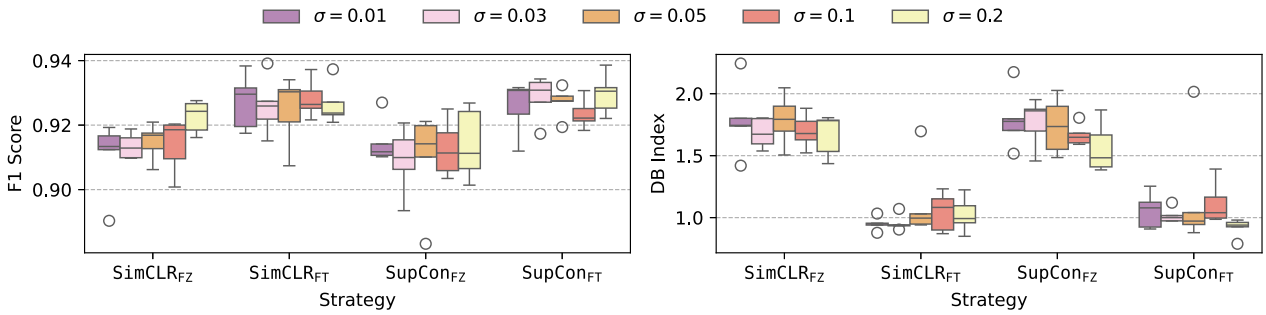


Fig. A.1. Sensitivity analysis of the jittering σ parameter. Box plots of *F1 Score* and *Davies-Bouldin index* across five values of the jittering standard deviation $\sigma \in \{0.01, 0.03, 0.05, 0.1, 0.2\}$, evaluated over 10 runs for each of the four contrastive strategies ($\texttt{SupCon}_{\texttt{FZ}}$, $\texttt{SupCon}_{\texttt{FT}}$, $\texttt{SimCLR}_{\texttt{FZ}}$, $\texttt{SimCLR}_{\texttt{FT}}$).

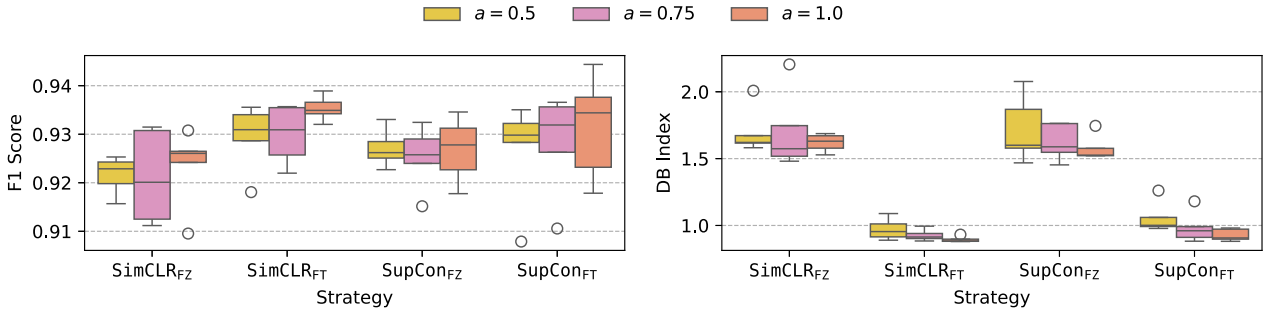


Fig. A.2. Sensitivity analysis of the shuffling a parameter. Box plots of *F1 Score* and *Davies-Bouldin index* across three values of the shuffling amplitude $a \in \{0.5, 0.75, 1.0\}$, evaluated over 10 runs for each of the four contrastive strategies ($\texttt{SupCon}_{\texttt{FZ}}$, $\texttt{SupCon}_{\texttt{FT}}$, $\texttt{SimCLR}_{\texttt{FZ}}$, $\texttt{SimCLR}_{\texttt{FT}}$).

Appendix B. Critical distance diagrams per metric

Fig. B.1 reports the Critical Distance (CD) diagrams for each individual evaluation metric. Across all five metrics, the separation between fine-tuning- and freezing-based strategies is consistent. This pattern is most pronounced for the Davies-Bouldin index, where $\texttt{SimCLR}_{\texttt{FT}}$ achieves the best average rank with no overlap with any other strategy, indicating that full fine-tuning under the SimCLR objective yields substantially more compact and well-separated clusters than any freezing-based alternative. For the remaining metrics—*Recall*, *Precision*, *F1 Score*, and *pAUC@1%FPR*—the FT/FZ partition is preserved, though the absolute rank gaps are narrower, reflecting that freezing-based strategies remain competitive on discriminative metrics despite their weaker cluster structure.

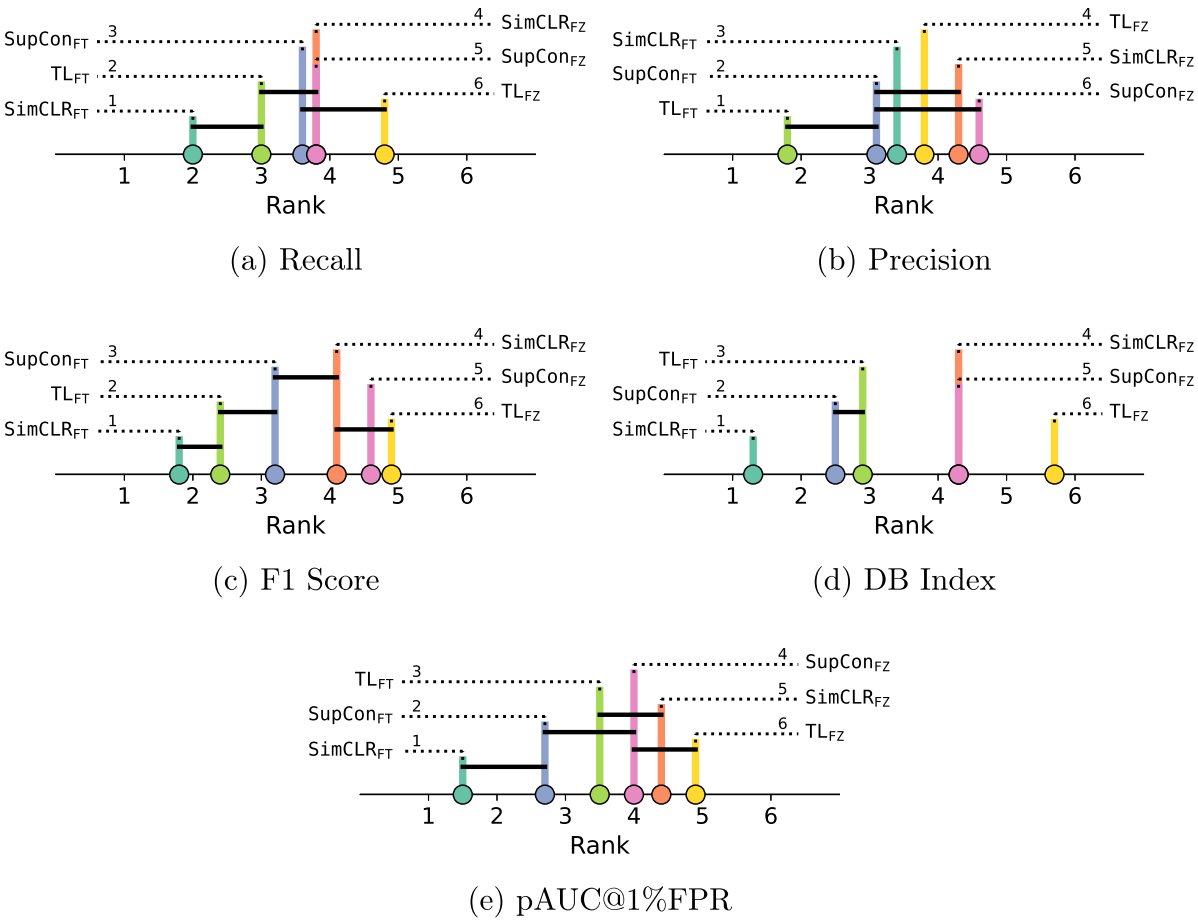


Fig. B.1. Per-metric ranking comparison. CD diagrams for each individual metric—*Recall*, *Precision*, *F1 Score*, *Davies-Bouldin index*, and *pAUC@1%FPR*—based on the average rank across 10 runs; strategies connected by a horizontal bar are not significantly different (Conover post-hoc, $\alpha = 0.05$).

References

[1] A.L. Buczak, E. Guven, A survey of data mining and machine learning methods for cyber security intrusion detection, IEEE Commun. Surv. Tutorials 18 (2) (2016) 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[2] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, F. Ahmad, Network intrusion detection system: a systematic study of machine learning and deep learning approaches, Trans. Emerg. Telecom. Technol. 32 (1) (2021) e4150. https://doi.org/10.1002/ett.4150

[3] K. Harshdeep, K. Sumalatha, R. Mathur, DeepTransIDS: transformer-based deep learning model for detecting DDoS attacks on 5G NIDD, Results Eng. 26 (2025) 104826. https://doi.org/10.1016/j.rineng.2025.104826

[4] H. Kheddar, Y. Himeur, A.I. Awad, Deep transfer learning for intrusion detection in industrial control networks: a comprehensive review, J. Netw. Comput. Appl. 220 (2023) 103760. https://doi.org/10.1016/j.jnca.2023.103760

[5] Y.K. Saheed, J.E. Chukwuere, XAIEnsembleTL-IoV: a new explainable artificial intelligence ensemble transfer learning for zero-day botnet attack detection in the internet of vehicles, Results Eng. 24 (2024) 103171. https://doi.org/10.1016/j.rineng.2024.103171

[6] G. Aceto, D. Ciuonzo, A. Montieri, A. Pescapè, MIMETIC: mobile encrypted traffic classification using multimodal deep learning, Comput. Netw. 165 (2019) 106944. https://doi.org/10.1016/j.comnet.2019.106944

[7] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020, pp. 1597–1607. https://doi.org/10.5555/3524938.3525087

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 18661–18673. https://doi.org/10.5555/3495724.3497291

[9] J. Sun, X. Zhang, Y. Wang, S. Jin, CoMDet: a contrastive multimodal pre-training approach to encrypted malicious traffic detection, in: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 2024, pp. 1118–1125. https://doi.org/10.1109/COMPSAC61105.2024.00151

[10] H. He, X. Sun, H. He, G. Zhao, L. He, J. Ren, A novel multimodal-sequential approach based on multi-view features for network intrusion detection, IEEE Access 7 (2019) 183207–183221. https://doi.org/10.1109/ACCESS.2019.2959131

[11] D. Ramachandram, G.W. Taylor, Deep multimodal learning: a survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108. https://doi.org/10.1109/MSP.2017.2738401

[12] O.A. van den, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv:1807.03748, https://doi.org/10.48550/arXiv.1807.03748

[13] M. Lopez-Martin, A. Sanchez-Esguevillas, J.I. Arribas, B. Carro, Supervised contrastive learning over prototype-label embeddings for network intrusion detection, Inf. Fusion 79 (2022) 200–228. https://doi.org/10.1016/j.inffus.2021.09.014

[14] A. Ayantayo, A. Kaur, A. Kour, X. Schmoor, F. Shah, I. Vickers, P. Kearney, M.M. Abdelsamea, Network intrusion detection using feature fusion with deep learning, J. Big Data 10 (1) (2023) 167. https://doi.org/10.1186/s40537-023-00834-0

[15] P. Lin, K. Ye, Y. Hu, Y. Lin, C.-Z. Xu, A novel multimodal deep learning framework for encrypted traffic classification, IEEE/ACM Trans. Netw. 31 (3) (2023) 1369–1384. https://doi.org/10.1109/TNET.2022.3215507

[16] A. Nascita, F. Cerasuolo, D. Di Monda, J.T.A. Garcia, A. Montieri, A. Pescape, Machine and deep learning approaches for IoT attack classification, in: IEEE INFOCOM Workshops (INFOCOM WKSHPS), 2022, pp. 1–6. https://doi.org/10.1109/INFOCOMWKSHPS54753.2022.9797971

[17] S. Ali, O. Abusabha, F. Ali, M. Imran, T. Abuhmed, Effective multitask deep learning for IoT malware detection and identification using behavioral traffic analysis, IEEE Trans. Netw. Serv. Manage. 20 (2) (2023) 1199–1209. https://doi.org/10.1109/TNSM.2022.3200741

[18] X. Yang, S. Ruan, J. Li, Y. Yue, B. Sun, TrafCL: robust encrypted malicious traffic detection via contrastive learning, in: Proc. of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), 2024, pp. 2910–2919. https://doi.org/10.1145/3627673.3679839

[19] N. Wang, Y. Chen, Y. Hu, W. Lou, Y.T. Hou, FeCo: boosting intrusion detection capability in IoT networks via contrastive learning, in: IEEE INFOCOM 2022-IEEE Conference on Computer Communications, 2022, pp. 1409–1418. https://doi.org/10.1109/INFOCOM48880.2022.9796926

[20] P. Golchin, N. Rafiee, M. Hajizadeh, A. Khalil, R. Kundel, R. Steinmetz, SSCL-IDS: enhancing generalization of intrusion detection with self-supervised contrastive learning, in: 2024 IFIP Networking Conference (IFIP Networking), 2024, pp. 404–412. https://doi.org/10.23919/IFIPNetworking62109.2024.10619725

[21] Y. Yue, X. Chen, Z. Han, X. Zeng, Y. Zhu, Contrastive learning enhanced intrusion detection, IEEE J. Netw. Serv. Manage. 19 (4) (2022) 4232–4247. https://doi.org/10.1109/TNSM.2022.3218843

[22] X. Han, S. Cui, J. Qin, S. Liu, B. Jiang, C. Dong, Z. Lu, B. Liu, ContraMTD: an unsupervised malicious network traffic detection method based on contrastive learning, in: Proceedings of the ACM on Web Conference 2024, 2024, pp. 1680–1689. https://doi.org/10.1145/3589334.3645479

[23] Y. Jiang, X. Wang, Y. Lai, Y. Wang, A packet sequence permutation-aware approach to robust network traffic classification, IEEE Netw. Lett. 6 (3) (2024) 203–207. https://doi.org/10.1109/LNET.2024.3435723
[24] I. Guarino, G. Bovenzi, A. Nascita, D. Ciuonzo, D. Carra, A. Pescapè, A multimodal and perturbation-aware learning approach for robust traffic classification, Comput. Netw. (2026) 112090. https://doi.org/10.1016/j.comnet.2026.112090
[25] I. Guarino, C. Wang, A. Finamore, A. Pescape, D. Rossi, Many or few samples?: comparing transfer, contrastive and meta-learning in encrypted traffic classification, in: 2023 7th Network Traffic Measurement and Analysis Conference (TMA), 2023, pp. 1–10. https://doi.org/10.23919/TMA58422.2023.10198965
[26] C. Wang, A. Finamore, P. Michiardi, M. Gallo, D. Rossi, Data augmentation for traffic classification, in: Springer Nature Passive and Active Measurement (PAM), 2024, pp. 159–186. https://doi.org/10.1007/978-3-031-56249-5_7
[27] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359. https://doi.org/10.1109/TKDE.2009.191
[28] M.A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, H. Janicke, Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning, IEEE Access 10 (2022) 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809
[29] H. Kang, D.H. Ahn, G.M. Lee, J.D. Yoo, K.H. Park, H.K. Kim, IoT network intrusion dataset, IEEE Dataport Dataset (2019). https://doi.org/10.21227/q70p-q449
[30] S. Dadkhah, E.C.P. Neto, R. Ferreira, R.C. Molokwu, S. Sadeghi, A.A. Ghorbani, CI-CIoMT2024: a benchmark dataset for multi-protocol security assessment in IoMT, Internet Things 28 (2024) 101351. https://doi.org/10.1016/j.iot.2024.101351
[31] I. Guarino, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, A. Pescapè, Contextual counters and multimodal deep learning for activity-level traffic classification of mobile communication apps during COVID-19 pandemic, Elsevier Comput. Netw. 219 (2022) 109452. https://doi.org/10.1016/j.comnet.2022.109452
[32] I. Akbari, M.A. Salahuddin, L. Aniva, N. Limam, R. Boutaba, B. Mathieu, S. Moteau, S. Tuffin, A look behind the curtain: traffic classification in an increasingly encrypted web, Proc. ACM Meas. Anal. Comput. Syst. 5 (1) (2021). https://doi.org/10.1145/3447382
[33] G. Apruzzese, L. Pajola, M. Conti, The cross-evaluation of machine learning-based network intrusion detection systems, IEEE Trans. Netw. Serv. Manage. 19 (4) (2022) 5152–5169. https://doi.org/10.1109/TNSM.2022.3157344
[34] W.J. Conover, Practical Nonparametric Statistics, John Wiley & Sons, New York, NY, USA, 3rd edition, 1999.
[35] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (200) (1937) 675–701. https://doi.org/10.1080/01621459.1937.10503522
[36] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30. https://doi.org/10.5555/1248547.1248548
[37] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.